UC Berkeley
UC Berkeley Previously Published Works

Title
A Reference Genome Assembly of Hybrid-Derived California Wild Radish (Raphanus sativus × raphanistrum).

Permalink
https://escholarship.org/uc/item/2ns4d4zt

Journal
Journal of Heredity, 113(2)

ISSN
0022-1503

Authors
Alexandre, Nicolas M
Haji, Diler
Bakhtiari, Moe
et al.

Publication Date
2022-05-16

DOI
10.1093/jhered/esab076

Peer reviewed

eScholarship.org Powered by the California Digital Library
University of California

Journal of Heredity, 2022, **113**, 197–204
https://doi.org/10.1093/jhered/esab076
Advance access publication 16 December 2021
Genome Resources

American Genetic Association
OXFORD

Genome Resources

A Reference Genome Assembly of Hybrid-Derived California Wild Radish (*Raphanus sativus × raphanistrum*)

Nicolas M. Alexandre, Diler Haji, Moe Bakhtiari, Kamalakar Chatla, Jessica M. Aguilar, Ksenia Arzumanova, and Noah K. Whiteman

From the Department of Integrative Biology, University of California, Berkeley, 3040 Valley Life Sciences Building, Berkeley, CA 94720, USA (Alexandre, Haji, Bakhtiari, Chatla, Aguilar, and Arzumanova) and Department of Molecular and Cell Biology, University of California, Berkeley, 142 Weill Hall #3200, Berkeley, CA 94720, USA (Whiteman).

Address correpondence to N. M. Alexandre and N. K. Whiteman at the address above, or e-mail: nalexandre@berkeley.edu and whiteman@berkeley.edu

Corresponding Editor: Beth Shapiro

Abstract

For agriculturally important plants, pollination and herbivory are 2 ecological factors that play into the success of crop yields. Each is also important in natural environments where invasive plants and their effect on species interactions may alter the native ecology. The California Wild Radish (*Raphanus sativus × raphanistrum*), a hybrid derived from an agriculturally important crop and a nonnative cultivar, is common in California. Remarkably, it has recently replaced wild populations of both progenitor species. Experiments on phenotypic variation for petal color and antiherbivore defenses suggest both pairs of polymorphisms are maintained as a result of pollinator- and herbivore-mediated natural selection. This species provides an opportunity to understand how natural selection shapes the evolution of ecologically important traits when traits are constrained by 2 opposing forces. Here we provide the genome assembly of the California Wild Radish displaying improvement to currently existing genomes for agronomically important crucifers. This genome sequence provides the tools to dissect the genomic architecture of traits related to herbivory and pollination using natural variation in the wild as well as the ability to infer demographic and selective history in the context of hybridization. Study systems like these will improve our understanding and predictions of evolutionary change for correlated traits.

Keywords: California Wild Radish, herbivory, hybrid, invasive, pollination, speciation

Introduction

Radish is a major root vegetable crop derived from wild *Raphanus* species in the agronomically important Brassicaceae. Approximately 100 years ago, wild radish (*Raphanus raphanistrum*) was introduced into California from Europe and began to hybridize with cultivated radish (*R. sativus*), which was grown in California by European settlers and their descendants for hundreds of years (Panetsos and Baker 1967). By the early 20th century, just a few morphological hybrids were observed in the wild, but by the 1960s, some populations were completely composed of hybrid plants with parental morphological traits in proportions that were correlated with whether they grew inland or on the coast (Panetsos and Baker 1967). Hybridization between wild radish and cultivated radish resulted in the origin of a hybrid swarm, known as California Wild Radish, that replaced both parent species in the wild (Hegde et al. 2006). Analyses of chloroplast genomes revealed that wild radish may have evolved following bilateral genetic assimilation between *R. sativus* and *R. raphanistrum* (Ridley et al. 2008); however, little is known about the nature of the hybridization events that produced the hybrid-derived lineage because the whole genome assembly of California Wild Radish was unavailable before this study.

The unique combination of parental traits in California Wild Radish likely contributed to its ecological success in western North America, making it a useful system for studying evolutionary and functional genomics of invasive species. In addition, the biology of California Wild Radish has garnered broad interest (Ridley and Ellstrand 2008). California Wild Radish segregates in nearly all populations as 4 petal color morphs determined by the presence or absence of anthocyanin pigments (phenylpropanoids). Both pollinators and herbivores prefer the morph lacking anthocyanins due to its decreased production of inducible indole glucosinolates (Stanton 1987, Irwin et al. 2003, Strauss et al. 2004). Petal color and antiherbivore defense polymorphisms in California Wild Radish are maintained in naturally occurring populations as a result of antagonistic pollinator- and herbivore-mediated natural selection (Irwin and Strauss 2005). In the presence of pollinator discrimination within the parent species *R. raphanistrum*, the color composition of populations are expected to shift toward higher yellow petal frequencies (Stanton et al. 1989). However, color frequencies in California Wild Radish are relatively constant over time (Irwin and Strauss 2005). Considering that variation in petal color and inducible indole glucosinolate levels are heritable traits in

Received September 21, 2021; Accepted November 24, 2021

© The Author(s) 2022. Published by Oxford University Press on behalf of The American Genetic Association. All rights reserved. For permissions, please e-mail: journals.permissions@oup.com

this species (Carlson et al. 1985, Ishii et al. 1989), conflicting selection pressures exerted by pollinators and herbivores, coupled with pleiotropy, tight linkage, and/or long-range linkage disequilibrium (coselection) among color and defense loci, could maintain variation in both traits (Fineblum and Rausher 1997, Armbruster 2002, Brachi et al. 2015).

Despite the large-scale genomics and genetics resources that have been accumulated for its cultivated progenitor, which serves as a genomic and molecular genetics reference (Shen et al. 2013), no study has investigated the genomic basis of such complex phenotypes such as petal color and indole glucosinolate production in California wild radish due to the lack of genome assemblies of this hybrid species. Thus, development of genomic resources, especially a high-quality full genome sequence is necessary to 1) illuminate the role of natural selection in shaping the evolution of ecologically important traits and 2) the extent to which adaptation is constrained by genetic correlations among traits, of which many mediate mutualistic and antagonistic interactions between species. These biotic interactions are important components in predicting evolutionary responses to changing environments, particularly under global climate change. This has implications for both applied (crop breeding) and basic (evolutionary genomics) perspectives. Moreover, the high-quality assembly of the California Wild Radish genome sequence would provide a valuable resource for comparative genomics analyses with its progenitors as well as other related crucifers.

In this study, we performed de novo assembly of the California Wild Radish (*Raphanus raphanistrum* × *sativus*) genome using a combination of short- and long-read technology. We provide this genome as a resource for future studies on the evolutionary outcomes of hybridization as well as the genomic architecture of pollination and herbivore defense traits in the California Wild Radish hybrid swarm.

Methods

Biological Materials

We collected seeds from a single California Wild Radish plant for use as a maternal line from McLaughlin Eastshore State Park in Berkeley, CA and then germinated seeds from this individual at the University of California, Berkeley Oxford Tract greenhouse. We then collected young, fully extended radish leaves in liquid nitrogen, which were transported on dry ice, and stored at −80 °C.

Nucleic Acid Library Preparation

We extracted genomic DNA for genome sequencing from leaves of a single individual of California Wild Radish using a phenol extraction and 2 Salt:PCI cleanups. We then prepared a single library using the SMRTbell Express Template Preparation Kit 2.0 and generated sequence data from a single cell on the PacBio Sequel II. We used a subset of the same DNA extract used for PacBio libraries to generate paired-end reads on the NovaSeq 6000, which were used to error-correct the PacBio-based genome assembly. We prepared a single library for Illumina sequencing using the Roche Kapa Hyper Prep Kit.

DNA Sequencing and Genome Assembly

Based on the 6 assemblies of *R. sativus* in GenBank (GCA_000801105.2, GCA_902824885.1, GCA_002197605.1, GCA_000715565.1, GCA_001047155.1, GCA_010725405.1) (Pellicer and Leitch 2020), we expected a genome size of between 0.6 and 1.5 pg or 0.4 and 0.7 Gb. Using the conversion of 1 pg ~ 0.978 Gb, this is ca. 0.54–1.46 Gb (mean = 1 Gb, standard deviation = ±0.46 Gb). We used raw reads from Illumina data to plot a read count distribution for kmers of size 19 using Jellyfish2 and estimated genome size as 0.515 Gb (Marçais and Kingsford 2011). We then assembled the raw PacBio reads using Falcon Unzip 1.3.7 with the expected genome size set to 0.515 Gb (Chin et al. 2016). Because regions of high heterozygosity in eukaryotic assemblies often produce adjacent rather than collapsed copies of sequence (haplotigs), we used purge_dups v.1.2.3 pipeline to identify and remove haplotigs from the assembly (Guan et al. 2020). We used Pbmm2 (PacBio SMRT Tools) to index the resulting purged assembly and align raw pacbio reads. We then used variantCaller (PacBio SMRT Tools) to polish the assembly using the arrow algorithm, which calls the maximum likelihood base at each position given the reference and the raw PacBio reads (GenomicConsensus v.2.3.3) (Alexander 2018). We further polished the assembly using paired-end read data using a modified pipeline based on Freebayes v.1.3.2 (Garrison and Others 2010, Garrison and Marth 2012).

As previously described (Mudd et al. 2020), we identified putative archaeal, bacterial, viral, and vector contamination with BLAST+ (v2.7.1) (Camacho et al. 2009) against the respective RefSeq and UniVec databases using general_decon.sh (V1.0). The pipeline General_decon.sh did not suggest any removals after querying the assembly, but did provide several candidates for further inspection. The general_decon.sh works in 2 iterative steps. The first step identifies scaffolds with greater than or equal to 95% identity and an e-value of 1E-10 for positive blast hits. The second step then assigns a scaffold to the category of "Scaffolds to Check" or "Scaffolds to Remove". A scaffold is assigned in the first category if a scaffold with a positive blast hit aligns to more than half of the scaffold or more than 200 bp. Scaffolds are recommended for removal if a subset of the "Scaffolds to Check" category align to 98% or more of the scaffold length (Mudd et al. 2020). Because no scaffolds were flagged for removal, we decided to focus on the "Scaffolds to Check" category. Because different putative contaminants occasionally had positive blast hits on overlapping segments, we wrote a custom script to calculate the total contiguous length of overlapping blast hits in each contig via start and endpoint to calculate the true size of putative contaminants aligning on the contig. We found that the sum of all blast hits along a scaffold contributed to less than 1% of the scaffold length in almost all cases with the exception of 2 scaffolds with 2 and 19% of the scaffold length, respectively. Because this coverage was lower than half of the scaffold length for all putatively contaminated scaffolds, no contigs were excluded from the assembly and all were considered endogenous. All contigs were then aligned to the publicly available mitochondrial and chloroplast genomes of *R. sativus* (Jeong et al. 2014, 2016) using Minimap2 (Li 2018). Because percent contig coverage was typically below ~1% for all blast hits, another custom script incorporating bedtools was used to calculate the percent of summed chloroplast or mitochondrial genome alignments represented on each genomic contig (Quinlan and Hall 2010).

The output of Falcon Unzip was assessed for quality using countFasta.pl (Gupta n.d.). We assessed the quality of the final assembly using N50 values and BUSCO scores. We determined the genome assembly size, N50, and contig size distribution using QUAST v.5.0.2 (Gurevich et al. 2013). We

assessed genome completeness on the basis of single-copy orthologs present in Brassicales (Brassicales_odb10 database) using the BUSCO v.4.1.4 pipeline (Seppey et al. 2019). We ran Quast v.5.0.2 and BUSCO v.4.1.4 analyses for both the California Wild Radish and the *R. sativus* V1.0 genome assembly and then calculated the number of gaps in the final assembly using Genome Assembly Annotation Services (GAAS: Genome Assembly and Annotation Service code n.d.). Gaps were counted using the faCount function in the UCSC Genome Browser for each step of the assembly process including the outputs of Falcon Unzip, Purge_dups, Arrow, and Freebayes (Karolchik et al. 2003). A custom script get-N-sizes.sh from ("Visualize gaps in the genome" 2022) was then used to count and measure the sizes of these gaps, which were then plotted in R (Computing and Others 2013). We list all software used locally in Supplementary Table 1.

Genome Annotation

We used ancestral shared repeats from Repbase for Viridiplantae (Bao et al. 2015) and RepeatModeler v2.0.1 to build a database of ancestral repeats in our current assembly prior to the detection of long terminal repeats (LTR) with LTRHarvest (Genometools v.1.6.1), LTR_retreiver v.2.9.0, and LTR_FINDER v.1.1 (Xu and Wang 2007, Ellinghaus et al. 2008, Ou and Jiang 2018, Flynn et al. 2020). The output sequences from Repeatmodeler were then concatenated with redundant LTRs from the other software prior to softmasking with Repeatmasker v. 4.1.1 (Tarailo-Graovac and Chen 2009).

We conducted annotation and gene prediction using the Comparative Annotation Toolkit (CAT) (Fiddes et al. 2018) and used the *R. sativus* V1.0 genome as input for gene prediction. CAT uses an end-to-end pipeline that takes a HAL-format multiple whole genome alignment and annotations from one related genome used in the alignment to produce an annotation of the target genome (Fiddes et al. 2018). The genePred table output from CAT was then converted into gtf format with the genePredToGtf (Karolchik et al. 2003). We then assessed the quality of the annotation with OrthoVenn2 using the annotated proteome from *Arabidopsis thaliana*, *Brassica rapa*, and *R. sativus* V1.0 for comparison (Xu et al. 2019). OrthoVenn2 takes protein fastas as input and follows 2 steps, first identifying orthologous genes based on a graph-based method, and then computing the percentage of shared orthologous genes (shared elements) among input proteomes (Xu et al. 2019).

Genome Alignment and Phylogenetic Relationships

We used Minimap2 to align the *R. sativus* (V1.0) genome assembly against our California Wild Radish assembly and then visualized contiguity between California Wild Radish and *R. sativus* on each of the *R. sativus* chromosomes. The *R. sativus* genome used for this analysis was a pseudochromosome-scale genome assembly from the Radish Genome Database (Yu et al. 2019). We used BUSCO v5.1.2 (Simão et al. 2015) to obtain 640 full-length single-copy orthologous genes from California Wild Radish, *R. sativus* V1.0 (GCA_000801105.2), *R. raphanistrum* (GCA_000769845.1), and *B. napus* (Song et al. 2020) genome assemblies. We then generated alignments for each gene using MaFFT v7 (Katoh et al. 2002) and inferred tree topologies using the neighbor-joining method in PAUP* v4 (Swofford 2003). The resulting trees were binned into 3 topologies, of which 2 topologies show monophyly between either *R. sativus* and California Wild Radish or *R. raphinastrum* and California Wild Radish. We then estimated chromosome-specific maximum likelihood phylogenetic trees of concatenated BUSCO genes using RAxML v8 (GTR+GAMMA nucleotide substitution model; Stamatakis 2014).

Results

The California Wild Radish Genome Assembly Is More Contiguous and Complete Than Previous Radish Genome Assemblies

Coverage determined with BBMap v.38.76 indicated 20× coverage for paired-end Illumina data generated from NovaSeq6000 and 164× coverage for PacBio data from PacBio Sequel II (Bushnell 2014). The pre-assembly output of the Falcon assembler is listed in Supplementary Table 2. Genome size was estimated to be roughly 0.7 Gb using the initial output of Falcon with a total of 1859 contigs. N50 values place 50% of the total sequence length in 95 contigs above ~2 Mb. These statistics correspond to the output of the second step of the Falcon Assembler after generation of the string graph assembly. The genome size reported in Supplementary Table 2 was used in the calculation of coverage based on dividing the number of PacBio and Illumina reads by the genome size. The next step in the Falcon Assembler (Falcon Unzip) phased and polished the assembly, improving contiguity and assigning contigs as either primary contigs or haplotigs. Statistics for this output can also be found in Supplementary Table 2. The number of contigs was reduced to 1702 and N50 was improved by about 15 Kb. Highly heterozygous assemblies form haploid partial paths during the initial assembly steps which results in higher rates of contig fragmentation lowering the N50 relative to similar, more homozygous assemblies (Asalone et al. 2020). Because this is a hybrid-derived genome, heterozygosity is expected to be high; thus, N50 is a less reliable metric of completeness than other metrics such as BUSCO scores.

The final consensus assembly showed a significant improvement in the contig N50 with a lower frequency of internal gaps compared to the representative *R. sativus* genome (V1.0; Rs1.0) (Table 1). The output of faCount via UCSC Genome Browser indicated that gaps were introduced by polishing with long reads using Arrow (Alexander 2018) after the Purge_dups step where the number of gaps were modestly reduced in subsequent steps with short-read polishing (Supplementary Figure 1). We find that the distribution of gap sizes introduced by Arrow was comprised of gaps that were not equal in length indicating estimable sizes. Additionally, the purging of haplotigs in addition to polishing with PacBio and Illumina reads improved our initial assembly by reducing size from roughly 0.7 to 0.45 Gb. BUSCO scores comparing the 2 assemblies show a similar completeness to the *R. sativus* assembly at 97.3% completeness, and a lower fraction of fragmented BUSCOs (Table 2). Mapping both assemblies against each other using Minimap2 shows high homology and contiguity between the 2 assemblies (Figure 1).

Genome–genome alignments between the reference assembly with the mitochondrial genome suggest that all contigs, with the exception of 3, have alignments composed of 1% or less of the total contig length (Supplementary Table 3). When calculating the percentage of each respective organelle

Table 1. Quast output comparing final assembly of *R. sativus* × *raphistrum* and *R. sativus* V1.0

Assembly contigs	*R. sativus* × *raphanistrum*	*R. sativus*	Assembly scaffolds	*R. sativus*
# contigs (= 0 bp)	1203	—	# contigs (= 0 bp)	10 676
# contigs (= 1000 bp)	1203	—	# contigs (= 1000 bp)	10 676
# contigs (= 5000 bp)	1203	—	# contigs (= 5000 bp)	3140
# contigs (= 10 000 bp)	1203	—	# contigs (= 10 000 bp)	1501
# contigs (= 25 000 bp)	1187	—	# contigs (= 25 000 bp)	426 614 037
# contigs (= 50 000 bp)	901	—	# contigs (= 50 000 bp)	404 244 259
Total length (= 0 bp)	450 789 643	—	Total length (= 0 bp)	426 614 037
Total length (= 1000 bp)	450 789 643	—	Total length (= 1000 bp)	426 614 037
Total length (= 5000 bp)	450 789 643	—	Total length (= 5000 bp)	404 244 259
Total length (= 10 000 bp)	450 789 643	—	Total length (= 10000 bp)	392 962 934
Total length (= 25 000 bp)	450 459 859	—	Total length (= 25 000 bp)	378 297 583
Total length (= 50 000 bp)	43 8871 384	—	Total length (= 50 000 bp)	368 214 092
# contigs	1203	35 745	# contigs	10 676
Largest contig	10 718 699	—	Largest contig	53 636 577
Total length	450 789 643	—	Total length	426 614 037
GC (%)	36.76	—	GC (%)	35.29
N50	2 564 273	19 900	N50	38 354 807
N75	618 303	—	N75	26 309 735
L50	47	5028	L50	5
L75	129	—	L75	9
# *N*'s per 100 kbp	0.67	—	# *N*'s per 100 kbp	12 797.17

Table 2. BUSCO scores comparing *R. sativus* × *raphanistrum* assembly with *R. sativus* V1.0

BUSCO type	*Raphanus sativus* × *raphanistrum* %	*Raphanus sativus* × *raphanistrum n*	*Raphanus sativus* %	*Raphanus sativus* n
Complete	97.30	4475	97.60	4484
Complete and single-copy	90.90	4180	87.30	4012
Complete and duplicated	6.40	295	10.30	472
Fragmented	0.50	24	0.70	34
Missing	2.20	97	1.70	78
Total BUSCOs searched	4596	4596	4596	4596

sequence represented on each contig, we found different patterns for mitochondria and chloroplast. For alignments to the chloroplast genome, we found 5 contigs containing chloroplast sequence, each of which has a composition ranging between 11 and 53% of the chloroplast genome, and were thus masked using bedtools maskFasta (Quinlan and Hall 2010). For mitochondria, all contigs with alignments represent 1–3% or less of the mitochondrial genome size. Therefore, all mitochondrial sequences were treated as numts and were not masked.

The California Wild Radish Genome Sequence Is Repeat Rich With Localized Repeat Elements in Narrow Genomic Windows

Repeat content of the assembly was high: the total number of bases masked at around 0.263 Gb, ca. 58% of the assembly (Supplementary Table 4). The majority of this content was composed of long tandem repeats (LTRs) or unclassified repeat elements (Supplementary Table 5, Supplementary Figure 2). Similar to other radish genome sequences, these are primarily long tandem repeats (LTR) of the Gypsy or Copia class, although the California Wild Radish contains a high proportion of Retroelements (~25%) (Zhang et al. 2015). LTR-type repeats are localized (Figure 1: Chromosomes R2, R2, and R5) based on homology to the *R. sativus* reference genome.

The California Wild Radish Genome Annotation Improves Upon Previous Radish Genome Annotations

The output of OrthoVenn2 displays the distribution of shared orthologous clusters among the 4 input species as overlap in gene content among pairwise and multi-wise comparisons (Figure 2). The California Wild Radish genome assembly shows an improvement over the counts of specific or shared

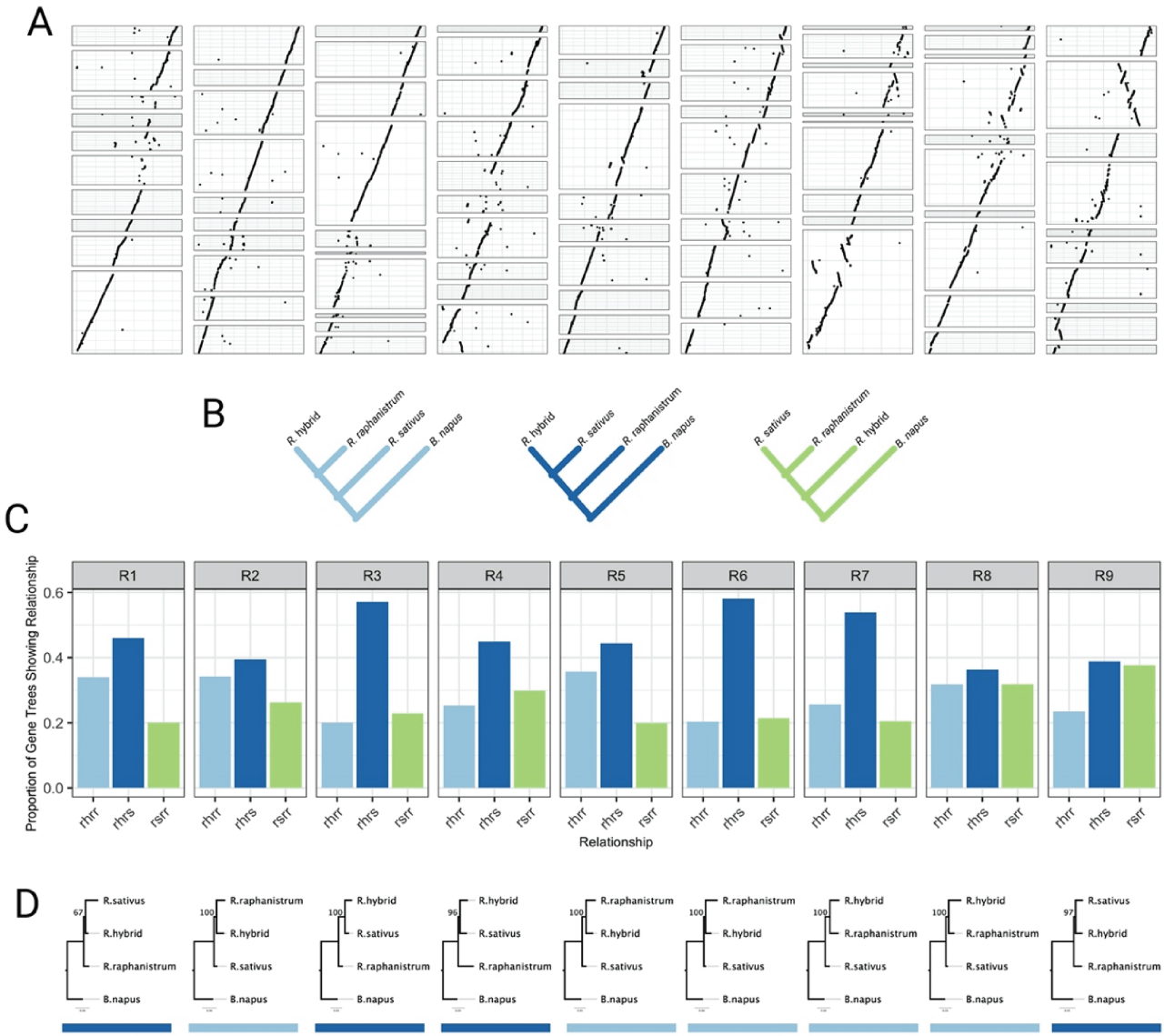

Figure 1. California Wild Radish genome assembly mapped with minimap2 against cultivated *Raphinus sativus* reference genome Rs. 1 chromosomes R1–R9 (A). Genomic contigs are oriented to the same mapping direction and only the top 10 contigs with the highest number of mappings are shown. Proportion of neighbor-joining tree topologies inferred from alignments of 640 BUSCOs (B) that match each of 3 rooted topologies (C). Trees were rooted using *Brassica napus* as an outgroup.

orthologous gene clusters in *A. thaliana*, and *B. rapa* with a comparable orthologous gene cluster count to the *R. sativus* genome sequence.

Parental Contributions to the Hybrid-Derived California Wild Radish Genome

Using BUSCO genes extracted from the *R. sativus*, *R. raphanistrum*, *B. napus*, and California Wild Radish reference genome assemblies, we found an enrichment of neighbor-joining tree topologies supporting a monophyletic relationship between California Wild Radish and the cultivar *R. sativus* as opposed to the wild progenitor *R. raphanistrum* (Figure 1C). However, this enrichment is likely due to differences in assembly quality between genomes rather than actual asymmetry in parentage. In addition, we inferred maximum likelihood trees using concatenated alignments of BUSCOs and found that 4 chromosomes supported a monophyletic relationship between California Wild Radish and the cultivar *R. sativus* and the other 5 chromosomes supported a monophyletic relationship between California Wild Radish and the wild progenitor *R. raphanistrum* (Figure 1D).

Discussion

The genome assembly of the California Wild Radish (*Raphanus sativus* × *raphanistrum*) reported here is a distinct improvement over most existing genome assemblies with representatives in Brassicaceae, a diverse plant family that has played important roles in genetics and agriculture (Bailey et al. 2006). Although our genome assembly is not chromosome level, our contig N50 is higher than that of the *Raphanus sativus* contig N50 assembly, where an N50 above 1 Mb is generally considered useful ("Beyond contiguity – assessing

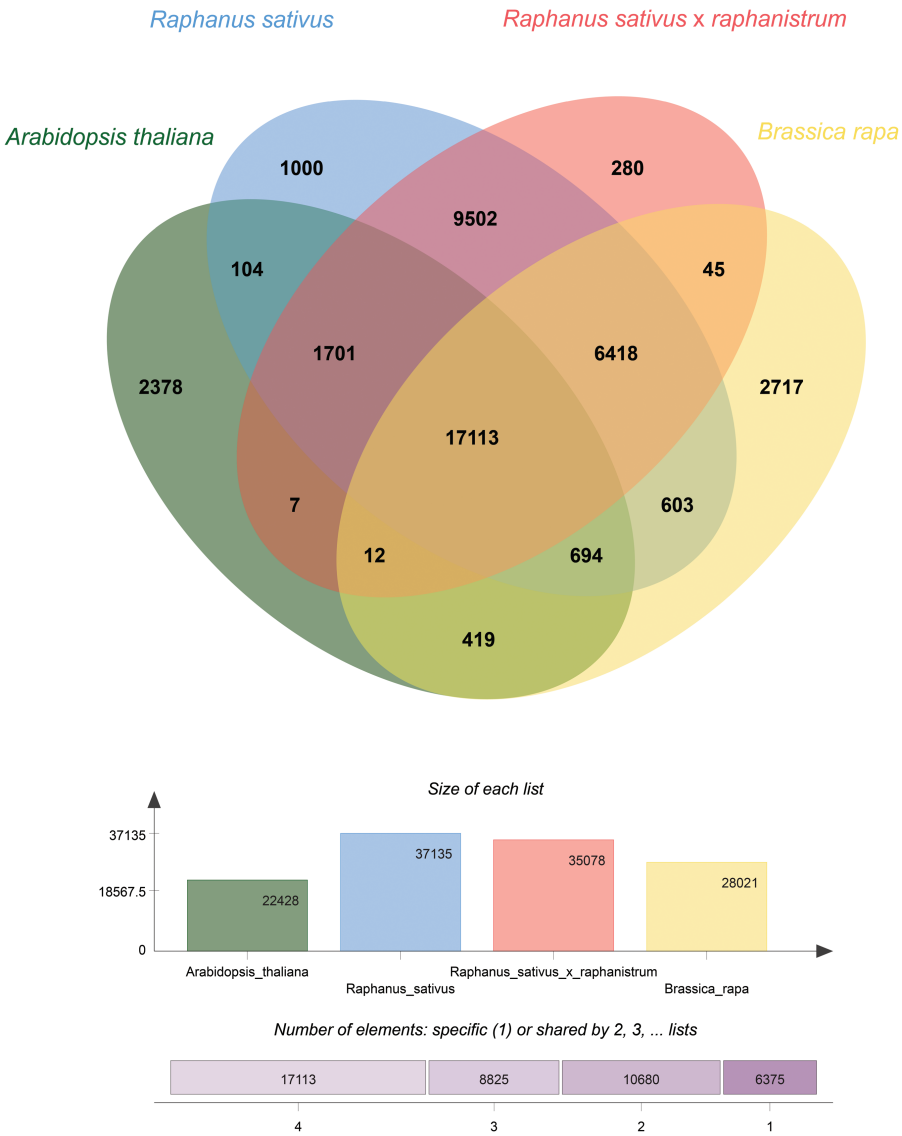

Figure 2. Output of OrthoVenn2 depicting overlap in shared elements among species used for comparison.

the quality of genome assemblies with the 3 C's" 2020). Rather, BUSCO is a more accurate metric of completeness because it ascertains the presence or absence of highly conserved genes in an assembly (Seppey et al. 2019). Therefore, this genome represents a highly comparable completeness to the representative *R. sativus* genome by the metric of conserved gene content. While protein content can be compared between annotated assemblies with software like OrthoVenn2, protein predictions in this case have been directly estimated using a transcriptome-free approach with the Comparative Annotation Toolkit and are therefore an underestimate of gene content (Fiddes et al. 2018).

The Brassicaceae is of broad interest given that it has been well studied in the context of plant-herbivore interactions, vis-a-vis the production of secondary metabolites known as glucosinolates, which are hydrolyzed into toxic molecules such as isothiocyanates (Rask et al. 2000). Cultivated radish (*Raphanus sativus*) is one such mustard species, used worldwide as an important root vegetable crop, the genome sequence of which has been used to identify the molecular mechanisms underlying formation of *Raphanus*-specific glucosinolates (Kitashiba et al. 2014). While California Wild Radish is derived from an invasive parental species in California, its fitness exceeded that of its parent species, eventually replacing populations of both in a hybrid swarm (Hegde et al. 2006). These hybrids are both invasive and have displayed evidence of rapid evolution via local adaptation to coastal and inland habitats following its recent spread in California (Ridley and Ellstrand 2010). Additionally, natural selection from both mutualistic pollinators and antagonistic herbivores interacts antagonistically in this species to maintain phenotypic variation in both plant fitness via pollination and plant secondary chemistry (Irwin et al. 2003). Therefore, this California Wild Radish genome assembly will provide a platform for future comparative genomics and population resequencing efforts that aim to understand the genetic bases of herbivore resistance, plant fitness, and root yield. Additionally, these resources will enable a clearer picture of the demographic and selective history of this invasive species, which may inform management decisions for its control.

The improvement in annotation completeness over the *A. thaliana* genome sequence using OrthoVenn2 in addition to a reduced number of contigs relative to the *R. sativus* V1.0

assembly demonstrate that our assembly is of relatively high quality. Our analysis of BUSCOs suggests that the genome contribution of parental genomes to the hybrid genome are roughly 50% on the basis of phylogenetic relationships among lineages represented by each chromosome, a pattern supported by the recency of the hybridization in this group (Hegde et al. 2006). This genome assembly provides a unique opportunity to study the genetic architecture of herbivory and agriculturally important phenotypes through the context of hybridization. California Wild Radish contains a mix of traits unique to both parental species which has resulted in transgressive fruit weight, aggressive colonizing behavior, and is reproductively isolated from both parental species following their local extinction (Hegde et al. 2006). This is interesting given the low values of F_{st} between cultivar and parent species based on allozyme polymorphism data, suggesting genetic uniformity (Hegde et al. 2006; Heredia and Ellstrand 2014; Heredia and Ellstrand 2014).

Crop yields worldwide are projected to be insufficient to feed the world's growing population by 2050 (Ray et al. 2013). Therefore, improvements in yields by reducing herbivory, increasing biomass, and improving resilience to climate fluctuations are becoming increasingly important (Bailey-Serres et al. 2019). The public availability of a wide diversity of genome sequences, including those of hybrid origin or naturally occurring relatives, is integral to the identification of novel genetic architectures. Local adaptation and vast phenotypic variation in California Wild Radish permits leveraging natural variation that may prove useful for the management of invasive species or the improvement of crop yields. The reference genome assembly reported here will facilitate basic and applied research on this fascinating and problematic species.

Supplementary Material

Supplementary data are available at *Journal of Heredity* online.

Supplementary Table 1. All installed software packages used along with the relevant versions and parameters.

Supplementary Table 2. Pre- and Post-Assembly Statistics generated 1. after Falcon Assembler and prior to Falcon Unzip step and 2. after the Falcon Unzip step.

Supplementary Table 3. Candidate contigs containing mitochondrial or chloroplast sequences with alignments of at least 1% of the contig length.

Supplementary Table 4. Output of RepeatMasker showing the total number of bases masked along with GC content.

Supplementary Table 5. Output of RepeatMasker showing the total number of repeats by repeat class.

Supplementary Figure 1. Gap sizes for outputs of Arrow (Long-Read Polishing) and Freebayes (Short-Read Polishing) (Alexander 2018.) showing the distribution of gap sizes.

Supplementary Figure 2. Distribution of repeat content in California Wild Radish. Genomic positions are based on minimap2 alignments to *R. sativus*. Unknown repeats were excluded. Contigs mapping to *R. sativus* reference assembly contigs without chromosome designations were excluded.

Funding

N.K.W. was supported by the National Institute for General Medical Sciences of the National Institutes of Health award [R35GM119816]. M.B. was supported by the Swiss National Science Foundation grant [P2NEP3_191665]. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Acknowledgments

Seeds were collected according to permit stipulations on September 29, 2017 at McLaughlin Eastshore State Park under East Bay Regional Park District permit number 910. Special thanks to Kirsten Isabel Verster for permit and collection assistance, Heather David and Pam Beitz for permit assistance, McLaughlin Eastshore park supervisor, Scott Possin, and Dr. Rohit Kolora for help on polishing the genome assembly.

Data Availability

All metadata associated with this assembly can be found in the main text. Raw sequence data and the genome fasta were deposited to Genbank. The final DNA sequence assembly can be found under the following bioproject PRJNA746262, biosample SAMN20198386, and accession JAIFRN000000000. Raw PacBio reads can be found under accession SRR15346202 and paired-end Illumina reads under accession PRJNA746262. The annotation file in gtf format can be found on our Github repository at https://github.com/nicolasalexandre21/Radish.

References

Alexander DH. 2018. *GenomicConsensus*. Available from: https://github.com/PacificBiosciences/GenomicConsensus.

Armbruster WS. 2002. Can indirect selection and genetic context contribute to trait diversification? A transition-probability study of blossom-colour evolution in two genera. *J Evol Biol*. 15:468–486.

Asalone KC, Ryan KM, Yamadi M, Cohen AL, Farmer WG, George DJ, Joppert C, Kim K, Mughal MF, Said R, et al. 2020. Regional sequence expansion or collapse in heterozygous genome assemblies. *PLoS Comput Biol*. 16:e1008104.

Bailey CD, Koch MA, Mayer M, Mummenhoff K, O'Kane SL Jr, Warwick SI, Windham MD, Al-Shehbaz IA. 2006. Toward a global phylogeny of the Brassicaceae. *Mol Biol Evol*. 23:2142–2160.

Bailey-Serres J, Parker JE, Ainsworth EA, Oldroyd GED, Schroeder JI. 2019. Genetic strategies for improving crop yields. *Nature*. 575:109–118.

Bao W, Kojima KK, Kohany O. 2015. Repbase Update, a database of repetitive elements in eukaryotic genomes. *Mob DNA*. 6:11.

Beyond Contiguity—Assessing the Quality of Genome Assemblies with the 3 C's. 2020. Available from: https://www.pacb.com/blog/beyond-contiguity/.

Brachi B, Meyer CG, Villoutreix R, Platt A, Morton TC, Roux F, Bergelson J. 2015. Coselected genes determine adaptive variation in herbivore resistance throughout the native range of *Arabidopsis thaliana*. *Proc Natl Acad Sci USA*. 112:4032–4037.

Bushnell B. 2014. *BBMap: a fast, accurate, splice-aware aligner*. Berkeley (CA): Lawrence Berkeley National Lab. (LBNL).

Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL. 2009. BLAST+: architecture and applications. *BMC Bioinf*. 10:421.

Carlson DG, Daxenbichler ME, VanEtten CH, Hill CB, Williams PH. 1985. Glucosinolates in radish cultivars. *J Am Soc Hortic Sci*.110:634–638.

Chin CS, Peluso P, Sedlazeck FJ, Nattestad M, Concepcion GT, Clum A, Dunn C, O'Malley R, Figueroa-Balderas R, Morales-Cruz A, et al. 2016. Phased diploid genome assembly with single-molecule real-time sequencing. *Nat Methods*. 13:1050–1054.

Computing, R., and Others. 2013. *R: a language and environment for statistical computing*. Vienna (Austria): R Core Team.

Ellinghaus D, Kurtz S, Willhoeft U. 2008. LTRharvest, an efficient and flexible software for de novo detection of LTR retrotransposons. *BMC Bioinf*. 9:18.

Fiddes IT, Armstrong J, Diekhans M, Nachtweide S, Kronenberg ZN, Underwood JG, Gordon D, Earl D, Keane T, Eichler EE, et al. 2018. Comparative Annotation Toolkit (CAT)-simultaneous clade and personal genome annotation. *Genome Res*. 28:1029–1038.

Fineblum WL, Rausher MD. 1997. Do floral pigmentation genes also influence resistance to enemies? The W locus in *Ipomoea purpurea*. *Ecology*. 78:1646–1654.

Flynn JM, Hubley R, Goubert C, Rosen J, Clark AG, Feschotte C, Smit AF. 2020. RepeatModeler2 for automated genomic discovery of transposable element families. *Proc Natl Acad Sci USA*. 117:9451–9457.

GAAS: Genome Assembly and Annotation Service code. n.d. Github.

Garrison E, Marth G. 2012. Haplotype-based variant detection from short-read sequencing.

Garrison E, and Others. 2010. FreeBayes. Marth Lab.

Guan D, McCarthy SA, Wood J, Howe K, Wang Y, Durbin R. 2020. Identifying and removing haplotypic duplication in primary genome assemblies. *Bioinformatics*. 36:2896–2898.

Gupta V. n.d. CountFasta.pl at master vikas0633/perl. Github.

Gurevich A, Saveliev V, Vyahhi N, Tesler G. 2013. QUAST: quality assessment tool for genome assemblies. *Bioinformatics*. 29:1072–1075.

Hegde SG, Nason JD, Clegg JM, Ellstrand NC. 2006. The evolution of California's wild radish has resulted in the extinction of its progenitors. *Evolution*. 60:1187–1197.

Heredia SM, Ellstrand NC. 2014. Novel seed protection in the recently evolved invasive, California wild radish, a hybrid *Raphanus* sp. (Brassicaceae). *Am J Bot*. 101:2043–2051.

Irwin RE, Strauss SY. 2005. Flower color microevolution in wild radish: evolutionary response to pollinator-mediated selection. *Am Nat*. 165:225–237.

Irwin RE, Strauss SY, Storz S, Emerson A, Guibert G. 2003. The role of herbivores in the maintenance of a flower color polymorphism in wild radish. *Ecology*. 84:1733–1743.

Ishii G, Saijo R, Nagata M. 1989. The difference of glucosinolate content in different cultivar of Daikon roots (*Raphanus sativus* L.). *Japanese Society for Food Science and Technology*. 36:739–742.

Jeong YM, Chung WH, Choi AY, Mun JH, Kim N, Yu HJ. 2016. The complete mitochondrial genome of cultivated radish WK10039 (*Raphanus sativus* L.). *Mitochondrial DNA A DNA Mapp Seq Anal*. 27:941–942.

Jeong YM, Chung WH, Mun JH, Kim N, Yu HJ. 2014. De novo assembly and characterization of the complete chloroplast genome of radish (*Raphanus sativus* L.). *Gene*. 551:39–48.

Karolchik D, Baertsch R, Diekhans M, Furey TS, Hinrichs A, Lu YT, Roskin KM, Schwartz M, Sugnet CW, Thomas DJ, et al.; University of California Santa Cruz. 2003. The UCSC Genome Browser Database. *Nucleic Acids Res*. 31:51–54.

Katoh K, Misawa K, Kuma K, Miyata T. 2002. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res*. 30:3059–3066.

Kitashiba H, Li F, Hirakawa H, Kawanabe T, Zou Z, Hasegawa Y, Tonosaki K, Shirasawa S, Fukushima A, Yokoi S, et al. 2014. Draft sequences of the radish (*Raphanus sativus* L.) genome. *DNA Res*. 21:481–490.

Li H. 2018. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics*. 34:3094–3100.

Marçais G, Kingsford C. 2011. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics*. 27:764–770.

Mudd AB, Bredeson JV, Baum R, Hockemeyer D, Rokhsar DS. 2020. Analysis of muntjac deer genome and chromatin architecture reveals rapid karyotype evolution. *Commun Biol*. 3:480.

Ou S, Jiang N. 2018. LTR_retriever: A highly accurate and sensitive program for identification of long terminal repeat retrotransposons. *Plant Physiol*. 176:1410–1422.

Panetsos CA, Baker HG. 1967. The origin of variation in wild *Raphanus sativus* (Cruciferae) in California. *Genetica* 38:243–274.

Pellicer J, Leitch IJ. 2020. The Plant DNA C-values database (release 7.1): an updated online repository of plant genome size data for comparative studies. *New Phytol*. 226:301–305.

Quinlan AR, Hall IM. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*. 26:841–842.

Rask L, Andréasson E, Ekbom B, Eriksson S, Pontoppidan B, Meijer J. 2000. Myrosinase: gene family evolution and herbivore defense in Brassicaceae. *Plant Mol Biol*. 42:93–113.

Ray DK, Mueller ND, West PC, Foley JA. 2013. Yield trends are insufficient to double global crop production by 2050. *PLoS One*. 8:e66428.

Ridley CE, Ellstrand NC. 2008. Evolution of enhanced reproduction in the hybrid-derived invasive, California wild radish (*Raphanus sativus*). *Biol Invasions*. 11:2251.

Ridley CE, Ellstrand NC. 2010. Rapid evolution of morphology and adaptive life history in the invasive California wild radish (*Raphanus sativus*) and the implications for management. *Evol Appl*. 3:64–76.

Ridley CE, Kim SC, Ellstrand NC. 2008. Bidirectional history of hybridization in California wild radish, *Raphanus sativus* (Brassicaceae), as revealed by chloroplast DNA. *Am J Bot*. 95:1437–1442.

Seppey M, Manni M, Zdobnov EM. 2019. BUSCO: Assessing genome assembly and annotation completeness. *Methods Mol Biol*. 1962:227–245.

Shen D, Sun H, Huang M, Zheng Y, Li X, Fei Z. 2013. RadishBase: a database for genomics and genetics of radish. *Plant Cell Physiol*. 54:e3.

Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. 2015. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics*. 31:3210–3212.

Song JM, Guan Z, Hu J, Guo C, Yang Z, Wang S, Liu D, Wang B, Lu S, Zhou R, et al. 2020. Eight high-quality genomes reveal pangenome architecture and ecotype differentiation of Brassica napus. *Nat Plants*. 6:34–45.

Stamatakis A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*. 30:1312–1313.

Stanton ML. 1987. Reproductive biology of petal color variants in wild populations of *Raphanus sativus*: I. Pollinator response to color morphs. *Am J Bot*. 74:178–187.

Stanton ML, Snow AA, Handel SN, Bereczky J. 1989. The impact of a flower-color polymorphism on mating patterns in experimental populations of wild radish (*Raphanus raphanistrum* L.). *Evolution*. 43:335–346.

Strauss SY, Irwin RE, Lambrix VM. 2004. Optimal defence theory and flower petal colour predict variation in the secondary chemistry of wild radish. *J Ecol*. 92:132–141.

Swofford DL. 2003. *Phylogenetic analysis using parsimony (*and other methods)*. Sunderland (MA): Sinauer Associates.

Tarailo-Graovac M, Chen N. 2009. Using RepeatMasker to identify repetitive elements in genomic sequences. *Curr Protoc Bioinformatics*. 25:4.10.1–4.10.14.

Visualize gaps in the genome. 2022. *Bioinformatics Workbook*. Available from: https://bioinformaticsworkbook.org/dataWrangling/R/visualize-gaps-in-genomes.html.

Xu L, Dong Z, Fang L, Luo Y, Wei Z, Guo H, Zhang G, Gu YQ, Coleman-Derr D, Xia Q, et al. 2019. OrthoVenn2: a web server for whole-genome comparison and annotation of orthologous clusters across multiple species. *Nucleic Acids Res*. 47:W52–W58.

Xu Z, Wang H. 2007. LTR_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons. *Nucleic Acids Res*. 35:W265–W268.

Yu HJ, Baek S, Lee YJ, Cho A, Mun JH. 2019. The radish genome database (RadishGD): an integrated information resource for radish genomics. *Database*. 2019:1–10.

Zhang XH, Yue Z, Mei SY, Qiu Y, Yang XH. 2015. A de novo genome of a Chinese radish cultivar. *Horticult Plant*. 1:155–164.