**Genome Report: Whole genome sequence of synthetically derived *Brassica napus* inbred cultivar Da-Ae**

John T. Davis\*, Ruijuan Li\*†, Seungmo Kim‡, Richard Michelmore§, Shinje Kim‡, Julin N. Maloof\*

\* Department of Plant Biology, University of California, Davis, Davis, CA, 95616

† Current Address: Inari Agriculture, Cambridge, MA

‡ FnP Co., Ltd., Jeungpyeong, South Korea

§ Genome Center and Department of Plant Sciences, University of California, Davis, Davis, CA, 95616

**Running Title:** Synthetic *B. napus* Genome Assembly

**Keywords:** Illumina, Dovetail, scaffolds, allotetraploid, subgenome

**Corresponding authors:**

Shinje Kim

Fungi and Plants, Corp.

60 Noam-ro, Doan-myeon

Jeungpyeong-gun, Chungbuk-do 27903

South Korea

[sjekim@fnpco.com](mailto:sjekim@fnpco.com)

+82-43-836-1751(tel)

Julin N Maloof

Department of Plant Biology

University of California, Davis

1 Shields Ave,

Davis, CA 95616

[jnmaloof@ucdavis.edu](mailto:jnmaloof@ucdavis.edu)

+1 (530) 752-8077

**Abstract**

*Brassica napus*, a globally important oilseed crop, is an allotetraploid hybrid species with two subgenomes originating from *B. rapa* and *B. oleracea.* The presence of two highly similar subgenomes has made the assembly of a complete draft genome challenging and has also resulted in natural homoeologous exchanges between the genomes, resulting in variations in gene copy number, which further complicates assigning sequences to correct chromosomes. Despite these challenges, high quality draft genomes of this species have been released. Using third generation sequencing and assembly technologies, we generated a new genome assembly for the synthetic *Brassica napus* cultivar Da-Ae. Through the use of long reads, linked-reads, and Hi-C proximity data, this draft genome that provides a high quality reference genome of a synthetic *Brassica napus*. In addition, we identified potential hotspots of homoeologous exchange between subgenomes within Da-Ae, based on their presence in other independently-derived lines. The occurrence of these hotspots may provide insight into the genetic rearrangements required for *B. napus* to be viable following the hybridization of *B. rapa* and *B. oleracea.*

**Introduction**

*Brassica napus*, commonly known as rapeseed, is the second most widely cultivated oilseed crop in the world (USDA). Historically, rapeseed oil was used primarily in the production of lubricants due to its high erucic acid content. In the late 1970s, new, edible, low erucic acid cultivars were created, enabling rapeseed oil to become a major component of most commercial vegetable oil products (Oplinger *et al.* 1989). The demand for rapeseed oil has caused global production to more than triple in the last few decades, with China and Canada being the world’s largest producers (“PSD Online 2020”). Numerous attempts are being made to understand the biology of *B. napus* with the goal of increasing production to keep up with demand.

The genetics of *B. napus* is challenging to untangle due to its genomic complexity. *B. napus* is an outcrossing species that originated from the hybridization of two different diploid parents, *B. rapa* and *B. oleracea* (Nagaharu 1935). Both *B. rapa* and *B. oleracea* are widely cultivated as human food crops such as cabbage, bok choy, and broccoli. It is believed that *B. napus* first appeared approximately 7,500 years ago when *B. rapa* hybridized with *B. oleracea* and underwent a chromosome doubling event, resulting in an allotetraploid (Chalhoub *et al.* 2014). *B. napus* (AACC) contains the diploid genomes of both *B. rapa* (AA) and *B. oleracea* (CC). While polyploidy has been hypothesized to provide plants with advantages, such as favorability in domestication (Bertioli *et al.* 2019), it also has genetic consequences that can cause several analytical challenges. In the case of *B. napus*, the A and C subgenomes are so similar that there can be homoeologous exchange of genetic information between the two subgenomes. Such exchanges range in size from a few base pairs (gene conversion) to larger chromosomal regions (Chalhoub *et al.* 2014). The rate and specifics of homoeologous exchange varies between *B. napus* populations and occurs more often in populations that have a newly synthesized *B. napus* as a parent (Udall *et al.* 2005; Higgins *et al.* 2018; Xiong *et al.* 2021) and loci affecting the rate have been identified (Higgins *et al.* 2021). Homoeologous exchange is thought to be a driving factor in the large amount of diversity found within *B. napus* (Gaeta *et al.* 2007; Stein *et al.* 2017; Lloyd *et al.* 2018; Hurgobin *et al.* 2018; Higgins *et al.* 2018; Raman *et al.* 2022). Consequently, it is important to have genome assemblies from multiple different *B. napus* varieties as an aid to building a pan-genome for this species.

In 2014, a genomic reference assembly for *B. napus* was released to the public (Chalhoub *et al.* 2014). This assembly, herein referred to as Darmor-bzh, was generated using short read sequencing data. Due to challenges associated with assembling and scaffolding short reads and the high similarity between the two subgenomes, a significant portion of the genome could not be confidently anchored in the assembly and was left unscaffolded (Table 1). Since the release of the Darmor-bzh assembly, new sequencing and assembly strategies, including long reads, linked-reads, and proximity data, have become available and fiscally feasible. Recently, new *B. napus* genomes using these technologies have been released to the public (Lee *et al.* 2020; Song *et al.* 2020; Rousseau-Gueutin *et al.* 2020). Concurrently, we generated a genomic assembly for a synthetic *B. napus* that includes multiple previously unscaffolded sequences relative to the Darmor-bzh v4.1 assembly. In addition, this new assembly reveals shared and unique homoeologous exchange events compared to different linesof *B. napus*.

**Methods and Materials**

*Creation of Synthetic Brassica napus, Da-Ae*

The synthetic *B. napus* cultivarDa-Ae (AACC, Korea patent number: 10-1432278-0000, 2014.08.13) used in this study was developed at FnPCo (South Korea) by crossing an inbred *B. rapa (*AA)Chinese cabbage (WC720) with an inbred *B. oleracea (*CC) red cabbage (BW716). After hybridization, the F1 underwent spontaneous chromosome doubling, producing a naturally occurring allotetraploid *B. napus (*AACC). The hybrid was self-fertilized, and seven seeds were obtained and planted. Only three of the seven plants germinated and flowered, with only one producing seeds. Progeny from this plant were then self-fertilized for six generations with the final generation being designated Da-Ae.

*Plant Materials, DNA Extraction, and Library Preparation*

Three plant lines were sequenced in this study: the highly inbred Da-Ae, the male parent *B. rapa* (AA, WC720), and the female parent *B. oleracea* (CC, BW716). For each line, 100 seeds from a single plant were germinated and grown for 8 to 10 days. The resulting seedlings were pooled separately for each line and high molecular weight genomic DNA was extracted by Amplicon Express (Amplicon Express Inc., Pullman, WA, US). The quality of the DNA collected from these three samples was assessed using a Bioanalyzer (Agilent Technologies, Inc. Santa Clara, CA, US). A 10X Genomics library was prepared by the University of California, Davis (UCD) Genome Center. The resulting libraries were sequenced on an Illumina HiSeq X10 by Novogene (Novogene Corporation Inc., Sacramento, CA, US) as 150 bp paired-end reads, producing ~451 million, ~380 million, and ~380 million reads for Da-Ae, the male parent, and the female parent, respectively. An additional 10X Genomics library for Da-Ae was constructed by the UCD Genome Center using a library prep involving sonication, in contrast to the 10X Genomics library prep without sonication. This library was then sequenced on a HiSeq 4000 at the UCD Genome Center, producing ~347 million 151 bp paired-end reads. For Pacific Biosciences (PacBio) sequencing, 32.9 µg high molecular weight DNA from Da-Ae was used for library construction and 19 SMRTcells were sequenced on a PacBio Sequel system (Pacific Biosciences, Menlo Park, CA, US) at the UCD Genome Center, producing ~6.6 million subreads with an average length of ~11.2 kb. An additional 100 seeds from the same Da-Ae plant were grown to produce 4.5 g young leaf tissue, which was sent to Dovetail Genomics (Dovetail Genomics, Scotts Valley, CA, US) for Hi-C library construction. The Hi-C library was then sequenced at the UCD Genome Center on an Illumina HiSeq 4000, producing ~374 million 150-bp paired-end reads.

*Generation of 10X Genomics Assemblies*

Initial assemblies of *B. napus* were generated using the default Supernova v1.1.5 pipeline (Weisenfeld *et al.* 2017) with an estimated genome size of 1.12 Gb. The 10X Genomics Da-Ae reads sequenced at the UCD Genome Center and Novogene (hereafter referred to as Da-Ae 10X Davis and Da-Ae 10X Novogene) were both independently assembled. The Da-Ae 10X Davis reads and the Da-Ae 10X Novogene reads resulted in near identical assemblies. As a result, only the Da-Ae 10X Davis reads were used in downstream Supernova assemblies. Upon the release of Supernova-2.0.0, the *B. rapa* 10X*, B. oleracea* 10X, and Da-Ae 10X Davisreads were each individually assembled using the newer software package. The number of reads required for 56X coverage was calculated using the formula genome size x 56 / read length. The expected genome sizes used for *B. rapa, B. oleracea,* and *B. napus* were 530 Mb, 630 Mb, and 1.12 Gb, respectively. These values were then input to Supernova-2.0.0 using the --maxreads parameter. Scaffolds from these three new Supernova assemblies were later used to assess mis-assemblies in Dovetail scaffolding-based assemblies.

*Generation of Pac-Bio Assemblies*

The PacBio reads were assembled using Canu version 1.6 (Koren *et al.* 2017) from Maryland Bioinformatics. Canu was configured for the 1.12 Gb genome size of *B. napus* and the reference suggestions for high coverage and polyploid organisms of corrected ErrorRate=0.040 and corOutCoverage=200. The Canu pipeline consisted of three separate steps: correction, trimming, and assembly.

*Polishing of Pac-Bio Assemblies*

Polishing was performed to improve the quality of the Canu Da-Ae assembly. Polishing was completed using the 10X Da-Ae Davis reads and the Broad Institute’s program Pilon v.1.22 (Walker *et al.* 2014). Following the guidelines from 10X Genomics, 23 bp at the start of read 1 and the first base pair of read 2 were removed using Trimmomatic v.0.33 (Bolger *et al.* 2014) in order to remove the 10X barcodes and the initial base of read 2 that is often low-quality. The trimmed reads were then mapped to the Canu Da-Ae assembly using bwa version 0.7.16a (Li and Durbin 2009). The assembly and the mapped read files were fed into Pilon. After polishing, the assembly had approximately the same size and N50 as its unpolished counterpart.

*Hi-C Scaffolding of Pac-Bio Assemblies*

The Canu Da-Ae assembly and the Hi-C reads sequenced at the UCD Genome Center were sent to Dovetail Genomics for scaffolding. The assembly and the Hi-C reads were run through Dovetail’s proprietary HiRise pipeline, where the individual contigs were scaffolded to create chromosome scale scaffolds.

*Analysis of Hi-C Results*

The N50, assembly size, and BUSCO scores of the HiRise scaffolded assembly was measured. Next, all scaffolds from the HiRise generated assembly were compared to the chromosomes of the publicly available Darmor-bzh v4.1 genome hosted by the Brassica database (BRAD) (Cheng *et al.* 2011). The scaffolds from the HiRise generated assembly were independently aligned to the chromosomes of Darmor-bzh v4.1 using Nucmer with the parameters --maxmatch -l 100 -c 500. The alignments were filtered for quality and all scaffolds 1 Mbp or greater were plotted (Figures 1). If a scaffold aligned best to one reference chromosome, it was assigned a name based on its alignment. All remaining scaffolds in the assembly were not renamed and retained their HiRise designated sequence IDs

*Assessing Discrepancies Between the Canu Da-Ae Assembly and the Public Reference Assembly*

The 21 largest scaffolds in the assembly were independently compared to their corresponding Darmor-bzh v4.1 chromosomes (Darmor-bzh v10 was not available at this time). Regions of discrepancy between the assembly and the reference assembly were identified. The validity of each discrepancy was then tested by aligning PacBio reads and 10X ancestral parent scaffolds to the Canu Da-Ae assembly. The PacBio reads were aligned using BLASR (Chaisson and Tesler 2012) with a minimum subread length of 10 kb. The 10X ancestral parent scaffolds were aligned using nucmer from the MUMmer software suite (Marçais *et al.* 2018). If the region of discrepancy in the assembly had substantial support from the mapped reads and scaffolds, the discrepancy was considered a true difference between our assembly and the Darmor-bzh v4.1 assembly; thus, it was retained. If there was no support, or the mapped reads and scaffolds disagreed with the Canu Da-Ae assembly, the region of discrepancy was considered a likely error and altered to match Darmor-bzh v4.1. All alterations performed were simple sequence flips to fix assembly inversions. All inversions, except one, were almost exactly encapsulated by the contig boundaries of a scaffold. After all identified discrepancies had been addressed, the assembly was considered final and annotation began (Figure 2; Supplementary Table 1). After Darmor-bzh v10 was available we compared v4 and v10 and found them to be essentially co-linear. The one exception was an inversion on C07, a region where Da-Ae also showed a supported inversion relative to Darmor-bzh v4 (and which we did not change). We use juicer to make a Hi-C contact plant (figure X)

*Transcriptome Assembly and Structural Annotation of Novel Transcripts*

RNA-seq reads from thirteen RNA sequencing libraries generated from five tissues (young leaf, flower, bolting tissue, 1 cm silique, and 5 cm silique) of Da-Ae (Li *et al.* 2018) were used for transcriptome assembly and annotation. The raw sequencing data were preprocessed and mapped to the published genome sequence of Darmor-bzh (*Brassica napus* genome v4.1) as described in Li et al., (2018). The mapped reads were then assembled to transcripts using Cufflinks v2.2.1 (Trapnell *et al.* 2010) with the help of reference annotations. The output GTF files generated by Cufflinks were fed to Cuffmerge and then compared to the annotations from the reference assembly using Cuffcompare. From the output file, transcripts with code “u” were considered novel. Redundant isoforms among these novel transcripts were removed using CAP3 (Huang and Madan 1999), and only transcripts with open reading frames detected using TransDecoder (Haas *et al.* 2013) were retained for the next step. For *de novo* assembly, post-processed high-quality reads were pooled together and assembled using Trinity (Grabherr *et al.* 2011) set to default parameters. The abundance of transcripts was estimated using the Kallisto (Bray *et al.* 2016) method implemented in the Trinity pipeline, and those with less than one transcript per kilobase million were removed. Transcripts with detected open reading frames were aligned to the Darmor-bzh coding sequences (CDS) using BLASTN (AltschuP *et al.*) with an E-value cutoff of 1e-6, and those with high identity (≥ 95%) to Darmor-bzh CDS were filtered. An additional BLASTX search was conducted against NCBI non-redundant protein database using E-value 1e-6 to remove transcripts with no homology to known plant genes. The resulting assembly from reference-based and *de novo* methods were combined for structural annotation using DAMMIT (Scott 2016) with default parameters to generate the final GFF3 file. BUSCO scores for the final assembly were calculated to assess transcriptome completeness (Cantarel *et al.* 2008; Campbell *et al.* 2014).

*Annotation Using MAKER*

Annotation was performed using MAKER v.3.01.02-beta (Cantarel *et al.* 2008; Campbell *et al.* 2014a). Prior to running the MAKER pipeline, a custom repeat library was constructed using the MAKER-P Repeat Library Construction-Advanced (Campbell *et al.* 2014b). MAKER was run with the following parameters: the CDS transcripts from the Darmor-bzh v4.1 assembly (Chalhoub *et al.* 2014), Darmor-bzh v10 assembly (Rousseau-Gueutin *et al.* 2020), and the eight *B. napus* assemblies(ZS11, Westar, No2127, Zheyou7, Gangan, Shengli, Tapidor, and Quinta) from Song et. al 2020; the previously identified novel transcripts were used as expressed sequence tag (EST) evidence. The peptide sequences from each *B. napus* assembly mentioned above as well as *B. oleracea* HDEM*, B. rapa* Z1 v2downloaded from genoscope.cns.fr, and the *A. thaliana* Araport11 peptides downloaded from the TAIR Project (Berardini *et al.* 2015) were used as evidence for protein homology. MAKER parameters that were modified included the following: A custom Augustus gene prediction species model of Da-Ae created using BUSCO v3.0.2 with the long parameter was used as the model species for Augustus; repeat library was set to the custom repeat library we constructed using the MAKER-P Repeat Library Construction-Advanced protocol; est2genome was set to 1; protein2genome was set to 1. All other parameters not listed above were left as the MAKER defaults. Due to an unresolved bioinformatics issue, 10 kb of chrC01 sequence starting at 47,446,387 had to be masked with N before MAKER would run to completion.

Once annotation of each chromosome was completed, the MAKER proteins were compared to the Uniref90 protein set using BLASTP. Protein domains were then identified using InterProScan on the MAKER predicted proteins. Using accessory scripts provided with MAKER, the MAKER genes were then renamed with the prefix “Bna,” the suffix “Da-Ae,” and the BLASTP and InterProScan results were integrated into the GFF annotation files. Finally, the annotations were filtered to remove any annotation that contained an Annotation Edit Distance (AED) score greater than 0.5. The cutoff of 0.5 was selected based on the recommendation listed in Campbell *et al*. (2014a).

*Analysis of Homoeologous Exchange Between Subgenomes*

We examined homoeologous exchange using two methods: synteny analysis and read coverage. For synteny analysis we first identified “trusted” syntenic regions of the A and C subgenomes by performing a nucmer (Marçais *et al.* 2018) alignment of *B. rapa* (A chromosomes) and *B. oleracea* (C chromosomes) assemblies; hits were filtered to require > 85% identity for > 1000bp. We next used nucmer to align each *B. napus* assembly in our comparison (Da-Ae, Darmor-BZH\_V10, GanganF73, No2127, QuintaA, Shengli3, Tapidor, Westar, Zheyou73, and ZS11) to an *in silico* *B. napus* genome constructed by combining both the *B. rapa* (Istace *et al.* 2021)and the *B. oleracea* (Belser *et al.* 2018) chromosomes (hereafter referred to as “ancestral”). Candidate homoeologous exchange regions were defined as those where the sequence was from an “A” subgenome in the *B. napus* assembly but had its highest hit to a “C” region in the “ancestral” assembly or *vice versa*. Candidate homoeologous exchange regions were filtered to retain only those with greater than 90% identity and a length of greater than 100bp and that overlapped with “trusted” syntenic regions in the ancestral assembly (to eliminate false positives that could be caused by assembly gaps in *B. rapa* or *B. oleracea).* All nucmer runs used version 4.0.0 with default parameters; all nucmer results werefiltered to retrain the one best alignment using the delta-filter program with option “-1”.

When sequence reads from one genome are mapped to an assembly from a different genome, differences in homoeologous exchange between the two genomes will lead to decreased or increased read coverage. To utilize this kind of information, we performed a coverage analysis using the 10X Da-Ae Davis reads. All reads were trimmed for quality using Trimmomatic and the adapter sequences were removed with the parameters ILLUMINACLIP:adapters.fa:2:30:10 LEADING:3 TRAILING:3 SLIDINGWINDOW:4:15 MINLEN:36 before being mapped with BWA to the *in silico* “ancestral” *B. napus* genome described above. To find possible sites of homoeologous exchange, we first filtered the 10X Da-Ae Davis reads to retain those that could reliably be described as coming from either the A or C subgenome (i.e., those with unique and trustworthy mapping locations). To do so, the alignment file was filtered to only contain alignments that had a MAPQ of five or greater, were properly paired, had no supplementary alignments, and were primary alignments. Reads that passed these filters were then mapped to nine *B. napus* genomes. The coverageBed function from bedtools2 v2.29.2 (Quinlan and Hall 2010) was then used to calculate the coverage across the genomes and the coverage of the individual potential genes previously identified. The alternate mapping sites were also captured using the “XA” tag from the bwa output. Using edit distance as a filtering parameter, alternate mapping sites that had an edit distance equal to or less than the primary alignment’s edit distance were added to the coverage calculation. To calculate coverage across the genomes, median coverage in a window size of 100 kb with a step size of 20 kb was used. The calculated coverages were standardized based on the genome-wide average using R (R Core Team 2020). Prior to standardization, regions that contained ≥10X mean coverage of their chromosome were removed from further analysis. The coverages were then plotted to identify regions across the genome with higher or lower than average coverage. Coverages were plotted using ggplot2 (Wickham 2009) in R. Plots combining the coverage and synteny analyses were plotted using a modified version of the plotsr program (Goel and Schneeberger 2022).

Annotation of genes in shared homoeologous exchange regions was done by using BLASTP to query an Arabidopsis ARAPORT11 protein database with *Brassica rapa* or *Brassica oleracea* protein sequences at phytozome (Goodstein *et al.* 2012), keeping the one best hit, and downloading ARAPORT11 annotations from phytozome.

**Results**

To develop a high-quality assembly of this new, synthetic *B. napus* we took advantage of contemporary technologies by using a combination of 10X Genomics, Pacific Biosciences, and Dovetail / Hi-C methods (Figure 4). The application of each is described in turn below, followed by the results of the annotation and homoeologous exchange analysis.

*Supernova Assemblies*

The Da-Ae 10X Davis reads were assembled with Supernova v2.0.0. The assembly had a length of 918 Mb and an N50 of 1.5 Mb. Notably, the BUSCO scores of this new assembly approached the scores of the Darmor-bzh v4.1. The 10X reads for both *B. rapa* and *B. oleracea* assembled using Supernova v2.0.0 also showed promising results. Both assemblies had N50 values over 2 Mb and consisted of less than 20,000 scaffolds. Although all assemblies were smaller than the expected genome sizes, they were all on par with the sizes of the public references. The assembly metrics and BUSCO scores supported the use of the assembly scaffolds in the manual curation of the subsequent *B. napus* Da-Ae assembly.

*Dovetail Scaffolding*

The Pilon-polished Canu Da-Ae assembly was scaffolded using the HiRise pipeline by Dovetail Genomics. After HiRise scaffolding, the Canu Da-Ae assembly showed a large increase in N50 from 1.59 Mb to 42.79 Mb, and had 3,190 scaffolds. Twenty-three of the scaffolds were greater than 1 Mb, with the largest being 74.2 Mb. Regarding BUSCO scores, the scaffolding caused the single to duplicate ratio to increase in the Canu Da-Ae assembly while the percentage of complete BUSCOs, 98.6%, did not change in the Canu Da-Ae assembly (Supplemental Table 3).

*Assigning Scaffolds to Chromosomes*

To assign the scaffolds to the established chromosomes, the assembly was aligned to the Darmor-bzh v4.1 assembly using Nucmer. The 19 Darmor-bzh v4.1 chromosomes were covered by the 21 largest Canu scaffolds; 17 spanned the full length of their sister Darmor-bzh scaffold, while the remaining four scaffolds had to be concatenated into pairs to span ChrC06 and ChrC07 (Figure 1). Names were then assigned to the scaffolds based on which Darmor-bzh chromosome they aligned to.

*Assembly Discrepancies*

Comparison of the Canu Da-Ae assembly to the Darmor-bzh v4.1 assembly revealed 24 assembly discrepancies (Supplemental Table 1). These discrepancies included inversions, lack of contiguity, and introduction of new sequence. To assess the validity of these discrepancies, both the parental 10X scaffolds and the PacBio reads were mapped to the Canu Da-Ae assembly. In 15 of the 24 discrepancies, the Canu Da-Ae assembly was supported by either read mapping or scaffold evidence. In ChrC06 and ChrC07, two scaffolds spanned the whole reference chromosome but failed to be scaffolded together. These scaffolds were joined with 100 Ns to signify a scaffolding gap and were then able to span the entire Darmor-bzh v4.1 chromosome as one scaffold. In six cases, the Canu Da-Ae assembly had unsupported inversions with four of the inversions spanning from one scaffold gap to another scaffold gap. For each case, the sequence was inverted to match the Darmor-bzh v4.1 assembly. The most prominent discrepancy occurred on ChrA05. Alignment to Darmor-bzh v4.1 suggested that both chromosome arms were inverted at their junction with the centromere. As there was no read or scaffolding evidence to support this, both chromosome arms were inverted to match Darmor-bzh. Although our ChrA05 now agrees with the Darmor-bzh v4.1 assembly, the orientation and centromeric region remains questionable. After all discrepancies were addressed, the assembly was deemed final and annotation began. Darmor-bzh v10 was released after our assembly was finalized. Darmor-bzh v4.1 and v10 are nearly co-linear. In the one one place where they are not, C07, Da-Ae matches v10, so there was no need to update our assembly.

*Annotation*

MAKER analysis of the Da-Ae assembly predicted 125,439 protein coding genes after filtering, compared to the 101,400 and 108,190 genes annotated in the Darmor-bzh v4.1 and v10 assemblies. To explore these differences, we determined the location of the predicted genes in their respective assemblies. Da-Ae contains more gene models than Darmor-bzh v4.1 and v10, with 123,488 of the Da-Ae gene models being present on its 19 pseudomolecules compared to Darmor-bzh v4.1 and v10 which contain 80,927 and 106,885 gene models on their 19 pseudomolecules, respectively. These discrepancies could be due to the differences in length of time since polyploidization. Since Da-Ae is a new synthetic it has had much less time for gene loss after the polyploidization event.

*Final Assembly Comparison*

The final Da-Ae assembly improves upon the Darmor-bzh v4.1 assembly by a number of criteria (Tables 1 and 2). Comparing the full assemblies and the pseudomolecule assemblies, respectively, the N50 is 24% to 32% longer, and there are 36% to 47% more unambiguous bases incorporated into the Da-Ae assembly (Table 1). When compared to the Darmor-bzh v10 assembly, the full Da-Ae assembly and the pseudomolecule assembly each have 4% shorter N50s. However, the full Da-Ae assembly has 12% more unambiguous bases than Darmor-bzh v10, while the pseudomolecule Da-Ae assembly has 4% fewer unambiguous bases than Darmor-bzh v10 (Table 1). When comparing BUSCO scores using the brassicales\_odb10 dataset, both the Da-Ae assembly and the Darmor-bzh v10 assemblies had BUSCO complete scores of 98.5%, while Darmor-bzh v4.1 had a slightly lower score of 98.2%. Both Darmor-bzh assemblies had a higher percentage of complete single copy-BUSCOs, whereas the Da-Ae assembly had a higher percentage of duplicated BUSCOs.

*Genome Completeness Analysis*

Genome completeness of Da-Ae and Darmor-bzh v10 was analyzed using the public unigene set of 133,127 *Brassica* sequences. Of the 133,127 sequences, 116,897 (87.81%) were present in the pseudomolecules of both genomes. Overall Darmor-bzh v10 contained the most unigene sequences, 118,199, with Da-Ae a close second with 118,193 unigene sequences. A total of 13,632 (10.24%) were missing from both genomes. To determine if there were classes of genes that were deleted/missing in these genomes, we looked for enriched GO terms among the set of genes missing from the two genomes. Among the enriched categories, enrichment for genes involved in responses to biotic and abiotic stressors was particularly noticeable, as seen in the pink box in the left of Figure 7 (Supek *et al.* 2011) . We also looked for unigenes present in Da-Ae but not in Darmor-bzh v10 and vice versa. Here, among the enriched categories, we noted an enrichment for genes involved in very long chain fatty acid metabolism, perhaps reflecting different breeding selection targets for these oil-seed crops, as seen in the teal box in the bottom middle of the treemap (Figure 8).

*Homoeologous Exchange*

Homoeologous exchange is the exchange of genetic material from one subgenome to the other. This could result in the conversion of an A subgenome gene to a C subgenome gene or vice versa. Because *B. napus* is an allotetraploid containing two diploid subgenomes, A and C, homoeologous exchange can result in homoeolog ratios of 2:2, 3:1, or 4:0, corresponding to reciprocal, partial, or complete conversions, respectively. We used two criteria to identify and characterize candidate homoeologous exchange regions: 1) a synteny analysis in which the *B. napus* genome assemblies were aligned to concatenated *B. rapa* and *B. oleracea* genomes, as proxies for the ancestral A and C subgenomes; 2) a coverage analysis in which we looked for regions of decreased or increased coverage that would result when two *B. napus* genomes had different homoeologous exchange (see material and methods).

For the coverage analysis we examined read coverage of Da-Ae when mapped to itself, a pseudo “ancestral” genome of concatenated *B. rapa* and *B. oleracea,* and nine exisiting *B. napus* assemblies (Supplemental Figures SX to SY). Note that in these plots, average coverage is normalized to “1”; a partial conversion (one but not both homologs) will result in readings of ~0.5 and ~1.5, whereas a complete conversion will give coverages of ~0 and ~2 on this scale. Each chromosome shows a region of coverage elevated to 4X or higher, likely representing centromeric regions where the repeats are collapsed in the assembly. In addition, we see numerous regions with coverage in the 0, 0.5, 1.5, or 2X range.

To determine if any of regions with increased or decreased coverage might result from homoeologous exchange versus aneuploidy (citation), we plotted the coverage and synteny analysis together (Figures SXX to SYY). In many cases the change in coverage is due to homoeologous exchange. For example, examining ZS11 and Da-Ae reveals that ZS11 had a reciprocal exchange between the right-hand sides of A01 and C01, whereas in Da-Ae this region of C01 was converted to A01 (Figure 5A). As a consequence, there is low Da-Ae coverage at the end of ZS11 A01 and high coverage at the end of ZS11 (since that region is corresponds to ancestral A01 and Da-Ae has two homoeologs matching A01 in this region). Comparing Zheyou73 and Da-Ae A02 and C02 reveals that both ends of Da-Ae C02 have been converted to A02 and a region in the middle of Zheyou73 A02 has been converted to C02, with read coverage changing as expected (Figure 5B). Other regions with increased or decreased coverage but no evidence of homoeologous exchange could result from insertion/deletion differences between the genomes, aneuploidy, or incomplete genome assemblies.

The synteny plots (SXX-SYY) reveal that there are numerous regions where homoeologous exchange has occurred in the same place in different genomes. Since Da-Ae and No2127 are independent synthetic *B. napus* lines, this suggests that there are hotspots of homoeologous exchange. Figure SZ shows the similarity in homoeologous exchange regions across the varieties; as expected, the two synthetic varieties, Da-Ae and No2127 are the most dissimilar from the other varieties. We next asked if there were any homoeologous exchange regions shared among all varieties. We found 31 homoeologous exchange regions encompassing a total of 39kb that were common across all varieties. This is a much overlap than predicted by chance; based on the proportion of each genome involved in homeologous exchange we would expect zero bases to be common across all varieties. There are a total of 16 genes in the conserved exchange regions, 14 of which had strong homologs in the Arabidopsis genome (Table SX). Of these fourteen, one, *BolC8t52214H*, is a nucleotide binding site leucine-rich repeat protein whose closest Arabidopsis homolog is *AT1G12210* or *RPS5-LIKE 1*, a close paralog of the defense R gene *RPS5*. Two other genes have leucine-rich repeats although their relationship to plant immunity is less clear.

**Discussion**

Since the release of the first reference genome (Chalhoub *et al.* 2014), multiple research groups have released genome assemblies of different *B. napus* cultivars, analyzed homoeologous exchange, and identified quantitative trait loci (QTLs) related to key agricultural traits (Wang *et al.* 2015; Bayer *et al.* 2017; Samans *et al.* 2017; Stein *et al.* 2017; Song *et al.* 2020; Rousseau-Gueutin *et al.* 2020; Boideau *et al.* 2022). These efforts all contribute to untangling the genome biology of *B. napus* that will one day be combined to create a species-wide pangenome.

The first *B. napus* reference was assembled and released during a time when sequencing technologies from PacBio, 10X Genomics, and Dovetail Genomics were in their infancy and/or not fiscally feasible for most research groups. As a result, the first release of the *B. napus* genome was not able to benefit from the analytical power of these technologies. This is reflected in the assembly size of the Darmor-bzh V4.1 genome (Chalhoub *et al.* 2014). Although the expected size of the *B. napus* genome is over 1 Gb, the Darmor-bzh V4.1 genome assembly is only approximately 850 Mb, of which 650 Mb is contained in 19 chromosome-scale pseudomolecule scaffolds. By using a recently created synthetic *B. napus*, *Da-Ae*, along with long-read, linked-read, and proximity ligation technologies, we were able to generate a new synthetic *B. napus* genome that exceeded the first high-quality reference genome by several metrics and is on par with more contemporary assemblies. Our assembly of Da-Ae is over 1 Gb, with more than 800 Mb contained within 19 chromosome-scale pseudomolecule scaffolds. While our assembly is larger compared to both the Darmor-bzh V4.1 and v10 assemblies, it still maintains a high level of sequence collinearity with the two Darmor-bzh assemblies. On a gene level, the Darmor-bzh v4.1 and v10 references have fewer annotated genes than our assembly. The differences in the high-quality assemblies may reflect differences in genome content due to the synthetic Da-Ae having had fewer generations in which to “purge” extra material (resulting in, for example, larger number of bases and more duplicated BUSCOs) or could reflect differences resulting from the assembly process. It is not possible to distinguish between these causes with our available data. The improved assembly enabled by third generation sequencing technologies will serve as an excellent resource for *B. napus* geneticists and scientists aiming to identify genes underlying agronomic traits.

Homoeologous exchange is a biological process observed in allopolyploids, like *B. napus*, where highly similar yet different regions of the two diploid subgenomes exchange genetic material with one another. The result is new chromosome structures that, while being primarily composed of one ancestral genome, now also contain regions belonging to a different ancestral genome. To investigate the occurrence of homoeologous exchange in Da-Ae, we investigated both genome coverage and synteny across the genomes of *B. napus* Da-Ae, and nine other cultivars. Our results indicate that homoeologous exchange has occurred in both small and large regions throughout the whole genome. Each cultivar of *B. napus* had many unique homoeologous exchange events. More surprising was that there are multiple regions of homoeologous exchange that are shared among the *B. napus* cultivars. It is possible that these homoeologous exchange regions are shared among multiple varieties because particular combinations of homoeologous genes affect plant fitness or agroeconomic traits. Alternatively, these sites could be shared because sequence homology and chromosome topology favors recombination at these sites.t. These findings further build upon the previous work done to identify hotspot regions (Higgins *et al.* 2018).

In conclusion, using several sequencing technologies, we created a genome assembly similar in quality to other recently published assemblies that used third generation sequencing, allowing for an improvement upon the original Darmor-bzh v4.1 published assembly. We also identified potential hotspots of homoeologous exchange along with single-copy BUSCOs that are shared among different cultivars of *B. napus*. Our assembly and analysis of Da-Ae is another step forward toward the realization of apan-genome for *B. napus*.

**Data Availability**

All raw reads and the nuclear genome assembly and annotation are available at NCBI under BioProject PRJNA627442.

**Acknowledgements**

We would like to thank the members of the Michelmore Lab (UC Davis), especially Kyle Fletcher, Will Palmer, and Sebastian Reyes Chin Wo, for countless hours of advice and support throughout this project.

**Conflict of Interest**

The authors report no conflict of interest

**Funder information**

Need to pull from online

**References**

AltschuP, S. F., W. Gish, W. Miller, E. W. Myers, and D. J. Lipman Basic Local Alignment Search Tool. 8.

Bayer, P. E., B. Hurgobin, A. A. Golicz, C.-K. K. Chan, Y. Yuan *et al.*, 2017 Assembly and comparison of two closely related Brassica napus genomes. Plant Biotechnol. J. 15: 1602–1610.

Belser, C., B. Istace, E. Denis, M. Dubarry, F.-C. Baurens *et al.*, 2018 Chromosome-scale assemblies of plant genomes using nanopore long reads and optical maps. Nat. Plants 4: 879–887.

Berardini, T. Z., L. Reiser, D. Li, Y. Mezheritsky, R. Muller *et al.*, 2015 The arabidopsis information resource: Making and mining the “gold standard” annotated reference plant genome. genesis 53: 474–485.

Bertioli, D. J., J. Jenkins, J. Clevenger, O. Dudchenko, D. Gao *et al.*, 2019 The genome sequence of segmental allotetraploid peanut Arachis hypogaea. Nat. Genet. 51: 877–884.

Boideau, F., G. Richard, O. Coriton, V. Huteau, C. Belser *et al.*, 2022 Epigenomic and structural events preclude recombination in Brassica napus. New Phytol. 234: 545–559.

Bolger, A. M., M. Lohse, and B. Usadel, 2014 Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 30: 2114–2120.

Bray, N. L., H. Pimentel, P. Melsted, and L. Pachter, 2016 Near-optimal probabilistic RNA-seq quantification. Nat. Biotechnol. 34: 525–527.

Campbell, M. S., C. Holt, B. Moore, and M. Yandell, 2014a Genome Annotation and Curation Using MAKER and MAKER-P: Genome Annotation and Curation Using MAKER and MAKER-P, pp. 4.11.1-4.11.39 in *Current Protocols in Bioinformatics*, edited by A. Bateman, W. R. Pearson, L. D. Stein, G. D. Stormo, and J. R. Yates. John Wiley & Sons, Inc., Hoboken, NJ, USA.

Campbell, M. S., M. Law, C. Holt, J. C. Stein, G. D. Moghe *et al.*, 2014b MAKER-P: A Tool Kit for the Rapid Creation, Management, and Quality Control of Plant Genome Annotations. Plant Physiol. 164: 513–524.

Cantarel, B. L., I. Korf, S. M. C. Robb, G. Parra, E. Ross *et al.*, 2008 MAKER: An easy-to-use annotation pipeline designed for emerging model organism genomes. Genome Res. 18: 188–196.

Chaisson, M. J., and G. Tesler, 2012 Mapping single molecule sequencing reads using basic local alignment with successive refinement (BLASR): application and theory. BMC Bioinformatics 13: 238.

Chalhoub, B., F. Denoeud, S. Liu, I. A. P. Parkin, H. Tang *et al.*, 2014 Early allopolyploid evolution in the post-Neolithic Brassica napus oilseed genome. Science 345: 950–953.

Cheng, F., S. Liu, J. Wu, L. Fang, S. Sun *et al.*, 2011 BRAD, the genetics and genomics database for Brassica plants. BMC Plant Biol. 11: 136.

Gaeta, R. T., J. C. Pires, F. Iniguez-Luy, E. Leon, and T. C. Osborn, 2007 Genomic changes in resynthesized Brassica napus and their effect on gene expression and phenotype. Plant Cell 19: 3403–3417.

Goel, M., and K. Schneeberger, 2022 plotsr: visualizing structural similarities and rearrangements between multiple genomes. Bioinformatics 38: 2922–2926.

Goodstein, D. M., S. Shu, R. Howson, R. Neupane, R. D. Hayes *et al.*, 2012 Phytozome: a comparative platform for green plant genomics. Nucleic Acids Res. 40: D1178–D1186.

Grabherr, M. G., B. J. Haas, M. Yassour, J. Z. Levin, D. A. Thompson *et al.*, 2011 Trinity: reconstructing a full-length transcriptome without a genome from RNA-Seq data. Nat. Biotechnol. 29: 644–652.

Haas, B. J., A. Papanicolaou, M. Yassour, M. Grabherr, P. D. Blood *et al.*, 2013 De novo transcript sequence reconstruction from RNA-Seq: reference generation and analysis with Trinity. Nat. Protoc. 8:.

Higgins, E. E., W. E. Clarke, E. C. Howell, S. J. Armstrong, and I. A. P. Parkin, 2018 Detecting de Novo Homoeologous Recombination Events in Cultivated Brassica napus Using a Genome-Wide SNP Array. G3 Bethesda Md 8: 2673–2683.

Higgins, E. E., E. C. Howell, S. J. Armstrong, and I. A. P. Parkin, 2021 A major quantitative trait locus on chromosome A9, BnaPh1, controls homoeologous recombination in Brassica napus. New Phytol. 229: 3281–3293.

Huang, X., and A. Madan, 1999 CAP3: A DNA Sequence Assembly Program. Genome Res. 9: 868–877.

Hurgobin, B., A. A. Golicz, P. E. Bayer, C.-K. K. Chan, S. Tirnaz *et al.*, 2018 Homoeologous exchange is a major cause of gene presence/absence variation in the amphidiploid Brassica napus. Plant Biotechnol. J. 16: 1265–1274.

Istace, B., C. Belser, C. Falentin, K. Labadie, F. Boideau *et al.*, 2021 Sequencing and Chromosome-Scale Assembly of Plant Genomes, Brassica rapa as a Use Case. Biology 10: 732.

Koren, S., B. P. Walenz, K. Berlin, J. R. Miller, N. H. Bergman *et al.*, 2017 Canu: scalable and accurate long-read assembly via adaptive *k* -mer weighting and repeat separation. Genome Res. 27: 722–736.

Lee, H., H. S. Chawla, C. Obermeier, F. Dreyer, A. Abbadi *et al.*, 2020 Chromosome-Scale Assembly of Winter Oilseed Rape Brassica napus. Front. Plant Sci. 11:.

Li, H., and R. Durbin, 2009 Fast and accurate short read alignment with Burrows–Wheeler transform. Bioinformatics 25: 1754–1760.

Li, R., K. Jeong, J. T. Davis, S. Kim, S. Lee *et al.*, 2018 Integrated QTL and eQTL Mapping Provides Insights and Candidate Genes for Fatty Acid Composition, Flowering Time, and Growth Traits in a F2 Population of a Novel Synthetic Allopolyploid Brassica napus. Front. Plant Sci. 9:.

Lloyd, A., A. Blary, D. Charif, C. Charpentier, J. Tran *et al.*, 2018 Homoeologous exchanges cause extensive dosage-dependent gene expression changes in an allopolyploid crop. New Phytol. 217: 367–377.

Marçais, G., A. L. Delcher, A. M. Phillippy, R. Coston, S. L. Salzberg *et al.*, 2018 MUMmer4: A fast and versatile genome alignment system. PLOS Comput. Biol. 14: e1005944.

Nagaharu, U., 1935 Genome Analysis in Brassica with Special Reference to the Experimental Formation of B. Napus and Peculiar Mode of Fertilization. Jpn. J. Bot. 389–452.

Oplinger, E. S., L. L. Hardman, E. T. Gritton, J. D. Doll, and K. Kelling, 1989 Canola (Rapeseed): Alternative Field Crops Manual. Collect. Altern. Field Crops Man.

PSD Online.

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

R Core Team, 2013 R: A language and environment for statistical computing.

Raman, H., R. Raman, R. Pirathiban, B. McVittie, N. Sharma *et al.*, 2022 Multienvironment QTL analysis delineates a major locus associated with homoeologous exchanges for water-use efficiency and seed yield in canola. Plant Cell Environ. 45: 2019–2036.

Rousseau-Gueutin, M., C. Belser, C. Da Silva, G. Richard, B. Istace *et al.*, 2020 Long-read assembly of the Brassica napus reference genome Darmor-bzh. GigaScience 9: giaa137.

Samans, B., B. Chalhoub, and R. J. Snowdon, 2017 Surviving a Genome Collision: Genomic Signatures of Allopolyploidization in the Recent Crop Species Brassica napus. Plant Genome 10: plantgenome2017.02.0013.

Scott, C., 2016 dammit: an open and accessible de novo transcriptome annotator. Prep.

Song, J.-M., Z. Guan, J. Hu, C. Guo, Z. Yang *et al.*, 2020 Eight high-quality genomes reveal pan-genome architecture and ecotype differentiation of Brassica napus. Nat. Plants 6: 34–45.

Stein, A., O. Coriton, M. Rousseau‐Gueutin, B. Samans, S. V. Schiessl *et al.*, 2017 Mapping of homoeologous chromosome exchanges influencing quantitative trait variation in Brassica napus. Plant Biotechnol. J. 15: 1478–1489.

Supek, F., M. Bošnjak, N. Škunca, and T. Šmuc, 2011 REVIGO Summarizes and Visualizes Long Lists of Gene Ontology Terms. PLOS ONE 6: e21800.

Trapnell, C., B. A. Williams, G. Pertea, A. Mortazavi, G. Kwan *et al.*, 2010 Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat. Biotechnol. 28: 511–515.

Udall, J. A., P. A. Quijada, and T. C. Osborn, 2005 Detection of chromosomal rearrangements derived from homologous recombination in four mapping populations of Brassica napus L. Genetics 169: 967–979.

USDA, F. A. S. Publication | Oilseeds: World Markets and Trade | ID: tx31qh68h | USDA Economics, Statistics and Market Information System.

Wang, X., K. Yu, H. Li, Q. Peng, F. Chen *et al.*, 2015 High-Density SNP Map Construction and QTL Identification for the Apetalous Character in Brassica napus L. Front. Plant Sci. 6:.

Weisenfeld, N. I., V. Kumar, P. Shah, D. M. Church, and D. B. Jaffe, 2017 Direct determination of diploid genome sequences. Genome Res. 27: 757–767.

Wickham, H., 2009 *ggplot2: elegant graphics for data analysis*. Springer New York.

Xiong, Z., R. T. Gaeta, P. P. Edger, Y. Cao, K. Zhao *et al.*, 2021 Chromosome inheritance and meiotic stability in allopolyploid Brassica napus. G3 Bethesda Md 11: jkaa011.