**Genome Report: Whole genome sequence of synthetically derived Brassica napus inbred strain, 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. The high degree of similarity between the subgenomes increases the difficulty of resolving the two subgenomes; it has also resulted in 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 a synthetic *Brassica napus* cultivar, Da-Ae. Through the use of long reads, linked-reads, and Hi-C proximity data, we assembled a new draft genome that provides the community with a more complete reference genome of *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 understanding 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 (“FAOSTAT 2018”). 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 2018”). 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 has been reported to occur more often in populations that have a newly synthesized *B. napus* as a parent (Udall *et al.* 2005). The exact mechanisms of this process are still unknown; however, the process is thought to be a driving factor in the large amount of diversity found within *B. napus*. Consequently, it has been challenging to generate a standard public consensus genome assembly for *B. napus*.

In 2014, a high-quality genomic reference assembly for *B. napus* was released to the public (Chalhoub *et al.* 2014). This assembly, hereby 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. 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 (Song *et al.* 2020). Concurrently we have generated a new genomic reference for a synthetic *B. napus* that includes a significant number of previously unscaffolded sequences. Additionally, this new assembly reveals shared and unique homoeologous exchange events in different *B. napus* lines.

**Methods and Materials**

*Creation of Synthetic Brassica napus (Da-Ae)*

The synthetic *B. napus* genotype Da-Ae (AACC, Korea patent number: 10-1432278-0000, 2014.08.13) was the focus of this study. Da-Ae 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 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 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 instead of the 10X Genomics’ suggested 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 single Da-Ae plant were grown to produce 4.5 g young leaf tissues that were 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 assembled. The Da-Ae 10X Davis and Da-Ae 10X Novogene reads were arbitrarily split in half creating four sets of reads with coverage ranging from 40–60X, following the guidelines from 10X Genomics for using 36X–56X coverage when using Supernova. The four sets of reads were then assembled independently. All four assemblies had similar assembly statistics with both N50 (~140 Kb; Table 1) and total assembly lengths (793–806 Mb; Table 1) being lower than expected. Assemblies were completed again upon the release of Supernova-2.0.0. The 10X *B. rapa,* 10X *B. oleracea*, and 10X Da-Ae Davisreads were used in this round of assembly. The 10X Da-Ae Novogene reads were excluded due to having near identical assembly performance when compared to the 10X Da-Ae Davis reads. For the assemblies generated with Supernova-2.0.0, the reads sets were not arbitrarily split. Instead, 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. napus, B. rapa,* and *B. oleracea* were 1.12 Gb, 530 Mb, and 630 Mb, respectively. These values were then input to Supernova-2.0.0 using the --maxreads parameter. Scaffolds from the 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 assembliy. 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 of 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 frequently low-quality sequence. The trimmed reads were then mapped to the Canu 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 assembly along with the Hi-C reads sequenced at the UCD Genome Center were sent to Dovetail Genomics for scaffolding. Both assemblies along with 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 was compared to the chromosomes of the publicly available Darmor-bzh genome (*B. napus* genome v4.1) hosted by the Brassica database (BRAD) (Cheng *et al.* 2011). The scaffolds from the HiRise generated assembly were independently aligned to the Darmor-bzh chromosomes using Nucmer with the parameters --maxmatch -l 100 -c 500. The alignments were filtered for quality and all scaffolds were plotted (Figures 1–3). 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 Assembly and the Public Reference Assembly*

The 21 largest scaffolds in the assembly were independently compared to their corresponding Darmor-bzh chromosomes. 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 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. If the region of discrepancy in the assembly had significant support from the mapped reads and scaffolds, the discrepancy was considered a true difference between our assembly and the Darmor-bzh assembly and retained. If there was no support, or the mapped reads and scaffolds disagreed with the Canu assembly, the region of discrepancy was considered a likely error and altered to match Darmor-bzh. All alterations performed were simple sequence flips to fix assembly inversions. All inversions, except one, were almost exactly encapsulated within the contig boundaries of a scaffold. After all identified discrepancies had been addressed, the assembly was considered final and annotation began (Figure 4, Supplementary Table 1).

*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) (Li *et al.* 2018). The mapped reads were then assembled by Cufflinks v2.2.1 (Trapnell *et al.* 2010) to transcripts with the help of reference annotations. The output GTF file generated by Cufflinks was fed to Cuffmerge and then Cuffcompare along with the annotations from the reference assembly. 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) with 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 1 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). Annotation using MAKER was run in two rounds. In order to speed up the annotation process, only the 19 named pseudomolecules were used and each pseudomolecule was annotated separately. In the first round of annotation, MAKER was run with the following parameters: The CDS transcripts from the Darmor-bzh assembly and the previously identified novel transcripts were used as expressed sequence tag (EST) evidence. The peptide sequences from *B. napus, B. oleracea,* and *B. rapa* downloaded from BRAD 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: Arabidopsis 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 stated above were left as the MAKER defaults. Due to an unresolved bioinformatic issue, 10 Kb of sequence of chrC01 starting at 47,446,387 had to be masked with N before MAKER would run to completion.

Upon completion of the first round of MAKER, the GFF files from each chromosome were concatenated. The GFF annotations were then filtered using Genome Annotation Generator (GAG) (Hall 2014) to remove questionable features. Following filtering, the annotations were then used to train SNAP (Korf 2004) using default parameters to generate an HMM file. Upon generation of the HMM file, the second round of MAKER was executed.

The second round of MAKER included all the scaffolds and used the same repeat library along with the same protein and EST evidence. est2genome and protein2genome were both set to 0 and snaphmm used the previously generated HMM file. Unlike the first round of annotation, Da-Ae was used as the model species for Augustus, and the Da-Ae model species files were generated through BUSCO using the long parameter. 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” 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*. (Campbell *et al.* 2014a).

*Analysis of Homoeologous Exchange between Subgenomes*

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. Homoeologous exchange was explored using both gene and sequence level analyses. Gene-level pairwise alignments between diploid genomes of Da-Ae, Darmor-bzh, Tapidor, *B. rapa,* and *B. oleracea* were made using JCVI’s MCscan pipeline (“jcvi: JCVI utility libraries | Zenodo”). Complete conversions are events where both sister chromatids for a region in one subgenome are converted to the homoeologous version from the other subgenome but without a reciprocal exchange. As a result, the ratio of A:C or C:A at these homoeologous regions will become 4:0. By this criteria, homoeologous exchange was examined at both gene and sequence level contexts using genome and transcriptome information from Da-Ae, Darmor-bzh, *B. rapa, B. oleracea*, and an additional *B. napus* cultivar Tapidor (Bayer *et al.* 2017). Because our current assembly is unphased, attempting to identify potential 3:1 homoeologous ratios is inhibited by the assembler program creating a consensus sequence by either selecting one of the two homoeologous regions or creating a mashup of the two regions. In either case, the true underlying sequences are not being accurately represented by the assembly sequence. Thus, only complete conversions were explored.

To look for homoeologous exchange at the gene level, annotations of Da-Ae, Darmor-bzh, *B. rapa,* and *B. oleracea*, and an additional *B. napus* cultivar Tapidor, were used. CDS sequences for each gene were generated by gffread (Trapnell *et al.* 2010) using each assembly’s GFF and sequence files. GFF annotations were converted to BED format using JCVI's jcvi.formats.gff module (“jcvi: JCVI utility libraries | Zenodo”). These BED files along with their corresponding CDS sequences were used as input for the MCscan pipeline of JCVI. Prior to running the MCscan pipeline, the *B. napus* CDS and BED files were separated into their two subgenomes, A and C, creating a CDS and BED file for both the A and C subgenomes. For each *B. napus* genome, the MCscan pipeline was run five times corresponding to five different pairwise synteny searches*, B. rapa vs. B. napus A, B. rapa vs. B. napus C, B. oleracea vs. B. napus A, B. oleracea vs. B. napus C,* and *B. napus A vs. B. napus C (*Ar-An, Ar-Cn, Co-An, Co-Cn, and An-Cn ) with a cscore filter of ≥ 0.99 to identify the reciprocal best hit (RBH) of each gene. *B. rapa* and *B. oleracea* (Ar- Co) were also aligned to one another for a total of 16 alignments. Although using a cscore cutoff of 0.99 should return only RBHs, it is still possible for a tie to occur between multiple query and subject sequences. If a tie occurred, the alignments were filtered to contain the alignment that had the highest bit score. The resulting alignments were then analyzed in R (R Core Team 2020) to identify potential genes that may have been involved in homoeologous exchange. For simplicity of analysis, RBH gene alignments between *B. rapa* and *B. oleracea* were used to filter potential genes involved in homoeologous exchange between the A and C subgenomes of *B. napus*. A homoeologous gene pair was considered a possible site of homoeologous exchange if two requirements were met. First, one gene of the pair must align better to its homoeolog than it does to its ortholog. Second, the gene must also align better to its homoeolog's ortholog than it does to its own ortholog. For example, consider the case of a gene on the *B. napus* C subgenome being converted to the *B. napus* A subgenome form. The gene in the C subgenome will align better to its homoeolog in the A subgenome than to its ortholog in the *B. oleracea* genome. The gene in the C subgenome will also align better to its homoeolog’s ortholog in the *B. rapa* genome than to its ortholog in the *B. oleracea* genome (Figure 5). However, if an annotation is incomplete or erroneous, it can create both false positive and false negative results.

For sequence level analysis of homoeologous exchange, the barcode removed 10X Da-Ae Davis reads, genomic reads from Darmor-bzh, and genomic reads from Tapidor were used. All reads were trimmed for quality using Trimmomatic and the adapter sequences with the parameters ILLUMINACLIP:adapters.fa:2:30:10 LEADING:3 TRAILING:3 SLIDINGWINDOW:4:15 MINLEN:36 before being mapped with BWA to an *in silico* *B. napus* genome constructed by combining both the *B. rapa* (Cheng *et al.* 2013)and the *B. oleracea* (Liu *et al.* 2014)chromosomes. To find possible sites of homoeologous exchange, we first filtered 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 files were 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 from these alignments were then mapped to their source genomes and filtered for alignments with a MAPQ of five or greater. bedcov from Samtools (Li *et al.* 2009) was then used to calculate the coverage across the genomes and the coverage of the individual potential genes previously identified. To calculate coverage across the genomes, a window size of 100 Kb with a step size of 20 Kb was used. The calculated coverages were standardized based on the chromosome 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. These regions were considered potential sites of homoeologous exchange. The coverage of the potential homoeologous genes was also compared to the genome coverage to look for agreement between the two methods.

**Results**

To develop a high-quality, more complete assembly of *B. napus* we took advantage of contemporary technologies by using a combination of 10X Genomics, Pacific Biosciences, and Dovetail / Hi-C methods (Figure 6). The application of each is described in turn below, followed by the results of the annotation and homoeologous exchange analysis.

*Supernova assemblies*

The first assembly attempts were made using 10X Da-Ae Davis and 10X Da-Ae Novogene reads along with the default Supernova-1.1.5 pipeline and an estimated genome size of 1.12 Gb. A total of four pseudohap assembly files, two from 10X Da-Ae Davis and two from 10X Da-Ae Novogene, were created (see methods). The assembly lengths ranged from 793–806 Mb with an average size of 801 Mb and the N50s ranged from 140–150 Kb with an average size of 143 Kb. All assemblies had approximately 80,000 scaffolds (Table 1) and poor BUSCO scores for the number of complete BUSCOs and proportion of single to duplicate BUSCOs when compared to the public reference (Table 1). Given that *B. napus* is a recent allotetraploid, one would expect to see a higher number of duplicate BUSCO genes due to a copy being present in both subgenomes. It is not standard procedure when generating 10X libraries to perform fragmentation using sonication, but given the highly similar assembly results, it suggests that the method of fragmentation did not alter assembly performance. Following the release of Supernova v2.0.0, the 10X Da-Ae Davis reads were reassembled. The new assembly had a length of 918 Mb and an N50 of 1.5 Mb. Additionally, the number of scaffolds was halved to ~36,000 (Table 1). Notably, the BUSCO scores of this new assembly greatly improved, approaching the scores of the Darmor-bzh (Table 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 (Table 1). 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 encouraged the use of the assembly scaffolds in the manual curation of future assemblies.

*PacBio assemblies*

The long read assembler Canu was used in this project. The Canu assembly was polished using Pilon and the 10X Da-Ae reads. After polishing with Pilon, the Canu assembly had a larger N50, smaller assembly size, and more complete BUSCOs compared to its unpolished counterpart.

*Dovetail Scaffolding*

The Pilon-polished Canu assembly was scaffolded using the HiRise pipeline by Dovetail Genomics. After HiRise scaffolding, the Canu assembly showed a large increase in N50 from 1.59 Mb to 42.79. The Canu assembly was now composed of 3,190 scaffolds. The Canu assembly also had 23 scaffolds greater than 1 Mb, with the largest being 74.2 Mb. Regarding BUSCO scores, the scaffolding caused the single to duplicate ratio to decrease in the Canu assembly, resulting in an increase in the number of complete BUSCOs in the Canu assembly (Table 1).

*Assigning Scaffolds to Chromosomes*

To assign the scaffolds to the established chromosomes, the assembly was aligned to the Darmor-bzh assembly using Nucmer. The 19 Darmor-bzh 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 in pairs to span ChrC06 and ChrC07 (Figure 2). Names were then assigned to the scaffolds based on which Darmor-bzh chromosome they aligned to.

*Assembly Discrepancies*

Comparison of the Canu assembly to the Darmor-bzh 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 assembly. In 15 of the 24 discrepancies, the Canu 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 chromosome as one scaffold. In six cases, the Canu 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 assembly. The most prominent discrepancy occurred on ChrA05. Alignment to Darmor-bzh 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 assembly, the orientation and centromeric region remains questionable. After all discrepancies were addressed, the assembly was deemed final and annotation began (Figure 4).

*Annotation*

MAKER analysis of the Da-Ae assembly predicted 96,442 protein coding genes after filtering, compared to the 101,400 genes annotated in the reference assembly. To explore these differences, we determined the location of the predicted genes in their respective assemblies. While Da-Ae contains fewer gene models than Darmor-bzh, 88,605 of the Da-Ae gene models are present on its 19 pseudomolecules compared to Darmor-bzh, which contains 80,927 gene models on its 19 pseudomolecules (Table 1). This indicates the improved assembly of pseudomolecules in the Da-Ae assembly.

To further explore the discrepancy in annotated gene number, we determined how much of the discrepancy was due to differences in annotation versus differences in assembly. Of the 101,040 predicted Darmor-bzh genes, 100,575 are present in the Da-Ae genome assembly and 91,949 are present in the Da-Ae predicted gene set (8,626 Darmor-bzh predicted genes are present in the Da-Ae assembly but not annotated as genes). Similarly, of the 96,442 predicted Da-Ae genes, 95,991 are present in the Darmor-bzh genome assembly and 88,303 are present in the Darmor-bzh predicted gene set (7,688 Da-Ae predicted genes are present in the Darmor-bzh assembly but not annotated as genes). Thus, almost all of the genes predicted from one genome are present in the other genome, but 8-8.5% of the predicted genes from one genome were not annotated in the other genome. One possible explanation for genes that are only present in one of the two annotations is that they are not true genes. Indeed, while the average length of predicted Darmor-bzh genes that have a match among Da-Ae predicted gene is 1,048 bases, those that are present in the Da-Ae genome but missing from the Da-Ae annotation average only 536 bases in length. Thus, much of the discrepancy in annotation is due to small predicted gene products that may not be true genes or are difficult to reliably annotate.

*Final Assembly Comparison*

The final Da-Ae assembly improves upon the Darmor-bzh assembly by a number of criteria (Table 3). Comparing the full assemblies and the pseudomolecule assemblies, respectively, the N50 is 24% to 32% longer; there are 36% to 47% more unambiguous bases incorporated into the Da-Ae assembly; and there are 1% to 4% more complete BUSCOs in the Da-Ae assembly. As for gene models, Da-Ae had 5% less than Darmor-bzh in the full assembly, but 9% more gene models incorporated into pseudomolecules.

*Genome Completeness Analysis*

Genome completeness of Da-Ae and Darmor-bzh was analyzed using the public Unigene set of 133,127 Brassica sequences. Of the 133,127 sequences, 117,447 (88.22%) were present in both genomes, 1,300 (0.98%) were present in only Da-Ae, 1,198 (0.90%) were present in only Darmor-bzh, and 13,182 (9.90%) were missing from both genomes. To determine there were particular classes of genes that were deleted in these genomes, we looked for enriched GO terms among the set of genes that were either present in Da-Ae and missing in Darmor-bzh or present in Darmor-bzh but missing in Da-Ae. We found an enrichment for genes involved in very long chain fatty acid metabolism, perhaps reflecting different breeding selection targets for these oil-seed crops (Figure 7). We also found enrichment for genes involved in several hormone pathways and in cuticle development, potentially representing adaptations to different environmental stressors (Figure 7).

*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. *B. napus* is an allotetraploid containing two diploid subgenomes A and C, meaning homoeologous exchange can result in homoeolog ratios of 2:2, 3:1, or 4:0, corresponding to reciprocal, partial, or complete conversions, respectively. For ease of detection given our unphased assembly we focused on complete conversions for our homoeologous exchange analysis.

At the gene level, there were 2,189, 1,848, and 823 potential gene pairs in Da-Ae, Darmor-bzh, and Tapidor where the C subgenome gene was a copy of the A subgenome gene. Conversely, there were 1,815, 1,666, and 666 potential gene pairs where the A subgenome gene was a copy of the C subgenome gene. To further validate these candidates, homoeologous exchange candidate gene pairs were next filtered based on their genomic sequencing coverage. If a C to A conversion has taken place, the expected average coverage ratio between orthologs should be 3:1 or greater when mapping reads to an *in silico* combined *B. rapa* + *B. oleracea* reference genome and should be 1:1 between homoeologs in the *B. napus* genome. Thus, a candidate exchange gene pair was retained if the ratio of coverage between the *B. rapa* and *B. oleracea* orthologs was at least 2.5 and the ratio of coverage between the two *B. napus* homeologs was between 0.5 and 1.5. After filtering, 234, 137, and 80 gene pairs remained in the C converted to A case, and 123, 150, and 31 in the A converted to C case for Da-Ae, Darmor-bzh, and Tapidor, respectively. Between the three *B. napus* genomes, only six C to A and one A to C gene conversions were shared (Figure 8). Interestingly, there was only one *B. rapa* to *B. oleracea* gene pair that showed opposite conversions between the Da-Ae and Darmor-bzh genome, where an A to C conversion took place in Da-Ae and a C to A conversion took place in Darmor-bzh.

At the sequence level, homoeologous exchange was examined by looking at the coverage across the genome using the previously described alignments. In regions where homoeologous exchange has occurred, we would expect an increase in the coverage of reads mapped to the donor region and a decrease in the coverage of reads mapped to the recipient region in the *in silico* *B .rapa* + *B. oleraceae* combined genome. This is due to the *in silico* recipient region being replaced with the donor region in the *B. napus* genome. In the *B.* napus genome, there would be an equal increase in coverage for reads mapped to both homoeologous exchange regions since both regions will be identical, allowing reads to map to both regions equally well. We observed sites of possible homoeologous exchange on every chromosome in the *B. napus* genome in regions ranging from 100 Kb to greater than 1 Mb. There are several large regions that appear to have undergone homoeologous exchange in two or more *B. napus* genomes (Figure 9). At the same time, each *B. napus* genome appears to contain numerous smaller sites of homoeologous exchange that are unique to their genome (Figure 10).

**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, 2016; Bayer *et al.* 2017; Samans *et al.* 2017; Stein *et al.* 2017; Song *et al.* 2020). 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 original *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 *B. napus* genome reference that exceeded the previous high-quality reference genome by several metrics. 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 the Darmor-bzh V4.1 assembly, it still maintains a high level of sequence collinearity with much of the increase in length being due to sequences in the Darmor-bzh assembly that were not anchored in the 19 chromosome pseudomolecules being included in the Da-Ae assembly. On a gene level, the Darmor-bzh reference does have slightly more annotated genes than our assembly, but the great majority of these are very small in length and most likely do not reflect true genes. While Darmor-bzh has more annotated genes, our Da-Ae assembly has a higher number of gene models located on the 19 pseudomolecules. 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 gene content across the genomes of three assemblies of *B. napus*, Da-Ae, Darmor-bzh, and Tapidor. 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 large regions of homoeologous exchange that are shared among the three *B. napus* cultivars. These shared regions may be homoeologous exchange hotspots for chromosomal rearrangements, which are required for viable *B. napus* cultivars to exist and further build upon the previous work done to identify hotspot regions (Higgins *et al.* 2018). Further investigation is needed to see how prevalent these shared homoeologous exchange regions are in the *B. napus* species.

In conclusion, using several recent sequencing technologies, we created a genome assembly that improves upon previous assemblies. We were able to include sequences that were previously unassigned, thus increasing the completeness of the *B. napus* genome 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 apangenome for *B. napus*.

**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

**Citations**

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.

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.

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.

Cheng, F., T. Mandáková, J. Wu, Q. Xie, M. A. Lysak *et al.*, 2013 Deciphering the Diploid Ancestral Genome of the Mesohexaploid *Brassica rapa*. Plant Cell 25: 1541–1554.

FAOSTAT.

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:.

Hall, B., 2014 *GAG: the Genome Annotation Generator*.

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

jcvi: JCVI utility libraries | Zenodo.

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.

Korf, I., 2004 Gene finding in novel genomes. BMC Bioinformatics 9.

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

Li, H., B. Handsaker, A. Wysoker, T. Fennell, J. Ruan *et al.*, 2009 The Sequence Alignment/Map format and SAMtools. Bioinformatics 25: 2078–2079.

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:.

Liu, S., Y. Liu, X. Yang, C. Tong, D. Edwards *et al.*, 2014 The Brassica oleracea genome reveals the asymmetrical evolution of polyploid genomes. Nat. Commun. 5: 3930.

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.

R Core Team, 2013 *R: A language and environment for statistical computing.* R Foundation for Statistical Computing, Vienna, Austria.

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.

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.

Wang, H., H. Cheng, W. Wang, J. Liu, M. Hao *et al.*, 2016 Identification of BnaYUCCA6 as a candidate gene for branch angle in Brassica napus by QTL-seq. Sci. Rep. 6: 38493.

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.