Long title: Insights from the genomes of four diploid *Camelina* spp.

Sara L. Martin,[1] Beatriz Lujan Toro,[1] Tracey James[1], Connie A. Sauder[1], and Martin Laforest[2]

1 Ottawa Research and Development Centre, Agriculture and Agri-Food Canada, Ottawa, Ontario, Canada

2 Saint-Jean-sur-Richelieu Research and Development Centre, Agriculture and Agri-Food Canada Saint-Jean-sur-Richelieu, Quebec, J3B 3E6 Canada

*Corresponding author

E-mail: sara.martin@agr.gc.ca; slwmartin@gmail.com

Orchid ID: 0000-0003-2055-6498

All assembled genomes will be available on publication from the National Center for Biotechnology Information as part of Bioproject PRJNA750147.

© Her Majesty the Queen in Right of Canada, as represented by the Minister of Agriculture and Agri-Food Canada

Short title: Insights from diploid *Camelina* genomes

Keywords: allopolyploidy, chromosome evolution, genome evolution, phylogenomics, Camelina, Brassicaceae

Ottawa Research and Development Centre

1016, 960 Carling Avenue, Ottawa

Ontario, Canada K1A 0CA

Agriculture and Agri-Food Canada

Phone: 613-715-5406

Sara.Martin@agr.gc.ca

Abstract

Plant evolution has been a complex process involving hybridization and polyploidization. As a result, understanding the origin and evolution of a plant's genome is often challenging even once a published genome is available. The oilseed crop, *Camelina sativa* from the *Brassicaceae*, has a fully sequenced allohexaploid genome with potentially three unknown ancestors. To better understand which extant species best represent the ancestral genomes that contributed to *C. sativa*'s formation, we sequenced and assembled chromosome level draft genomes for four diploid members of *Camelina*: *C. neglecta C. hispida* var. *hispida*, *C. hispida* var. *grandiflora* and C. *laxa* using a combination of long and short read data scaffolded with proximity data. We then conducted phylogenetic analyses on regions of synteny and on genes described for *Arabidopsis thaliana*, from across each nuclear genome and the fully sequenced chloroplasts in order to examine the evolutionary relationships within *Camelina* and *Camelineae*. We conclude that the genome of *C. neglecta* is closely related to *C. sativa*'s sub-genome 1 and that *C. hispida* var. *hispida* and *C. hispida* var. *grandiflora* are most closely related to *C. sativa*'s sub-genome 3. Further, the abundance and density of transposable elements, specifically Helitrons, suggest that the progenitor genome that contributed *C. sativa*'s sub-genome 3 was more similar to the genome of *C. hispida* var. *hispida* than that of *C. hispida* var. *grandiflora*. These diploid genomes show few structural differences when compared to *C. sativa*'s genome indicating little change to chromosome structure following allopolyploidization. This work also indicates that *C. neglecta* and *C. hispida* are important resources for understanding the genetics of *C. sativa* and potential genetic resources for crop improvement.

Introduction

A key goal in evolutionary biology is to understand the evolution of genomes - how they

change in structure and content through time and how this affects the rate of evolution (Otto and Whitton 2002). Three closely linked processes play major roles in shaping genome structure and the evolution of plants: hybridization (Stebbins 1968; Rieseberg 1997; Soltis and Soltis 2009; Abbott et al. 2013), polyploidization (de Wet 1971; Levin 1983; Soltis and Soltis 1999; Husband et al. 2013) and chromosomal rearrangements (Rieseberg 2001; Rieseberg and Willis 2007). The publication of crop genomes such as maize, canola, soybean, sugarcane and wheat (Schmutz et al. 2010; Schnable et al. 2011; Chalhoub et al. 2014; The International Wheat Genome Sequencing Consortium 2014) have underscored that these processes occur frequently and have fundamentally shaped plant lineages. Understanding these processes are major areas of interest in plant whole genome-scale studies (Koenig and Weigel 2015). While there has been less effort focused on the wild diploid relatives of these crops (Michael and VanBuren 2015), sequencing the extant representative of potential diploid progenitors of these crop species can provide additional information on the origins of the crop and the processes that have shaped their genomes (Marcussen et al. 2014; Latta et al. 2019).

Camelina, in the Brassicaceae, has a number of strengths as a group for studying polyploidization. First, *Camelina* is part of the Camelineae Tribe within the Brassicaceae which includes three other genera with available sequenced genomes: *Arabidopsis*, *Capsella*, and *Neslia* (Al-Shehbaz 2012). Second, the genome of the emerging oil seed crop and allohexaploid *Camelina sativa* (L.) Crantz (2n = 40) has been sequenced and well described (Kagale et al. 2014). The diploid ancestors of the three parental genomes of *C. sativa* are estimated to have diverged from *Arabidopsis thaliana* (L.) Heynhold about 17 million years ago (Mya). These three parental genomes are thought to have diverged from each other recently between 2.5 Mya (Čalasan et al. 2019) and 5.4 Mya (Kagale et al. 2014). The dates of the hybridizations

contributing to *Camelina sativa*'s formation are unknown, but Kagale et al. (Kagale et al. 2014) suggest that they occurred between 5-10,000 years ago. The *Camelina* genus includes four taxa with individuals known to have diploid chromosome counts and relatively small genome sizes: *Camelina neglecta* (2n = 12, 1C = 265 Mb) J. Brock, Mandáková, Lysak & Al-Shehbaz, *Camelina laxa* C. A. Mey. (2n = 12, 1C = 275 Mb), *Camelina hispida* Boiss. var. *hispida* (2n = 14, 1C = 355 Mb) and *Camelina hispida* var. *grandiflora* (Boiss.) Hedge (2n = 14, 1C = 315 Mb) (Al-Shehbaz 2012; Martin et al. 2017; Brock et al. 2019). These species could be modern representatives of genomes involved in the evolution of *C. sativa* and could provide insights into the evolution of the crop, tribe and family.

The Brassicaceae have played an important role in developing our understanding of genome evolution (Lysak et al. 2016). In the 2000's, researchers defined 24 large conserved collinear regions or blocks (Ancestral Crucifer Karyotype or ACK) among crucifer genomes (Schranz et al. 2006; Murat et al. 2015; Lysak et al. 2016) and reconstructed an ancestral karyotype with 8 chromosomes similar to *Arabidopsis lyrata* (L.) O'Kane & Al-Shehbaz and *Capsella rubella* Reut. (Koch and Kiefer 2005). These blocks were later grouped into 16 ancestral Brassicaceae karyotype regions (ABK) corresponding to the arms of the ancestral chromosomes (Murat et al. 2015). Describing the genes in the conserved regions provided a resource for comparative studies for the evaluation of genome rearrangements and gene loss (Murat et al. 2015) and have facilitated the reconstruction of the sub-genomes of *Brassica rapa* L. (Cheng et al. 2013) and *C. sativa* (Kagale et al. 2014). Indeed, *C. sativa*'s genome shows a conserved ACK structure in each of the three sub-genomes, with only 21 in-block breaks that, most likely, primarily occurred in its diploid progenitors (Lysak et al. 2016). Here we extend our understanding of the evolution of *C. sativa* by assembling draft genomes and full chloroplast

sequences for *C. neglecta, C. laxa, C. hispida* var. *hispida* and *C. hispida* var. *grandiflora*. Using this data and the conserved ACK structure to facilitate the phylogenetic analysis we examine *C. sativa*'s sub-genomes, relationships within C*amelina,* and relationships within the Camelineae.

Materials and Methods

Plant Material and Nucleic Acid Isolation

Camelina neglecta, C. hispida var. hispida, C. hispida var. grandiflora, and *C. laxa* seed were obtained from the North Central Regional Plant Introduction Station (NCRPIS) (PI650135, PI650139, PI650133 and PI633185, respectively). Original seeds obtained from NCRPIS were stratified in petri dishes using filter paper that was moistened with 0.2% KNO_3, sealed with Parafilm (Pechiney Plastic Packaging Company, Illinois, USA), and placed at 4°C in the dark for 2 weeks. For seed germination, the plates were then placed at room temperature under growth lights with a 16 h/8 h day/night light cycle. Seedlings were then sown on soil (soil, peat, and sand; 1:2:1; Promix, Rivière-du-Loup, Québec, Canada) in deep trays, and placed in growth chambers with a photoperiod of 16 h 20°C days/8 h 18°C nights. For the largely self-incompatible species, *C. hispida* and *C. laxa* rosette leaves were then sampled for DNA extractions. For the self-compatible, *C. neglecta*, after 6 weeks, the plants were placed for another 6 weeks at 4°C with an 8 h/16 h day/night photoperiod for vernalization. Plants were then transplanted to 5 in pots with the same soil and allowed to self-pollinate and set seed in a 16 h/8 h day/night photoperiod with 20°C days and 18°C nights. Following approximately 3 months, the mature seeds were collected and this process was repeated to obtain a fifth-generation inbred line. Young leaves from the rosette stage of this inbred line were used for DNA extraction. Vouchers for each of the accessions used have been deposited in the DAO (Department of Agriculture Ottawa) herbarium (*C. neglecta* DAO 902176; *C. laxa*: DAO

902754; *C. hispida* var. *hispida*: DAO 902780; and *C. hispida* var. *grandiflora:* DAO 902768). In all cases immature leaves were harvested using dry ice throughout the growing period and stored at -80°C before extraction.

For Pacific Biosciences long read (PacBio; Pacific Biosciences, Menlo Park, CA, USA) and Illumina short reads (Illumina Inc., San Diego, California, U.S) sequencing, total DNA was extracted using a FastDNA Spin Kit (MP Biomedicals, Solon, OH), grinding was done in the FastPrep (MP Biomedicals) at 4.0 for 20 s, with the addition of one ceramic bead. Two DNA extractions were pooled for a total volume of 200 µl and precipitated with the addition of 20 µl 3 M NaOAc and 200 µl 100% ethanol. Following an overnight incubation at -20°C, the DNA was centrifuged at 13,000 rpm at 4°C for 30 min. The ethanol was decanted and the DNA pellet was washed with cold 70% ethanol, dried at 37°C for approximately 20 min and re-suspended in 100 µl 5 mM Tris-HCl (pH 8.5). The DNA concentration was determined by Qubit was 110 ng/µl. DNA quality was determined by running 1.0 µl on a 0.8% E-gel (Invitrogen by ThermoFisher) beside a 0.2 µg of 20kb ladder (GeneRuler 1kbPlus, ThermoFisher).

DNA was also extracted to generate sequence data using Oxford Nanopore Technologies (ONT; Oxford Nanopore Technologies, Oxford Science Park, UK). High molecular weight DNA extraction procedures were carried out as described in Workman et al. (Workman et al. 2018), which first isolates nuclei and uses the Nanobind Plant Nuclei Big DNA kit (Circulomics Inc., Baltimore, MD, U.S.) to obtain high molecular weight genomic DNA. DNA was quantified using Qubit and assessed for quality with DropSense (Trinean., Pleasanton, CA, USA). The Short Read Eliminator kit (Circulomics Inc., Baltimore, MD, U.S.) was used to reduce the amount of short pieces of DNA in the sample as per the directions and the sample size was assessed using TapeStation (Agilent., Santa Clara, CA, U.S.). molecular weight genomic DNA.

DNA was quantified using Qubit and assessed for quality with DropSense (Trinean). The Short Read Eliminator kit (Circulomics Inc) was used to reduce the amount of short pieces of DNA in the sample as per the directions.

Sequencing

Sequencing was conducted in five different locations – DNA was sent to four of these locations with additional sequencing completed within the Martin laboratory. McGill University Genome Quebec Innovation Centre completed PacBio sequencing using P6-C4 chemistry for *C. laxa, C. hispida* var. *hispida* and *C. neglecta*. Since DNA from *C. neglecta* was slightly degraded, the library was prepared without shearing. A total of 7 Single Molecule Real-Time (SMRT) cells were used for sequencing *C. neglecta* and 8 cells each were used for *C. laxa* and *C. hispida*. They also generated paired end Illumina data for *C. laxa*, *C. hispida* var. *hispida* and *C. hispida* var. *grandiflora* with runs of 2 x 150 bases. The sequencing facility at the Microbial Molecular Technologies Laboratory (MMTL) in Ottawa (Ottawa Research and Development Centre, Agriculture and Agri-Food Canada) was used for paired-end sequencing. Two libraries were prepared using the Ovation Ultralow Library Systems (NuGEN, San Carlos, CA, USA), with 500bp inserts, and sequenced using Illumina MiSeq v3 chemistry with runs of 2 x 300 bases. The Centre for Applied Genomics, The Hospital for Sick Children, Toronto, Canada prepared and sequenced four additional libraries: one paired-end (PE) library with 550 bp inserts using a Nano kit (Illumina, San Diego, CA, USA) and three mate-pair (MP) libraries: 3kb, 5kb and 10kb inserts using the Nextera mate-pair kit (Illumina). All four libraries were sequenced on an Illumina HiSeq-2000 using v4 chemistry and flow cells with runs of 2 × 126 bases. Finally, ONT library preparation of samples, including DNA repair/end-prep and adaptor ligation and clean up steps, was performed using the 1D Long Fragment enrichment Protocol using kit SQK-

LSK109 , the NEBNext Companion Module for Oxford Nanopore Technologies Ligation Sequencing (New England BioLabs), and Agencourt AMPure XP beads (Beckman Coulter). ONT sequencing data was then obtained using a MinION with one FLO-MIN106 flow cell run for each taxon examined here and run for 48h. Base pair calling for the ONT data generated was completed with Guppy v. 3.2.2+9fe0a78 (Wick et al. 2019). Finally, library preparation for chromosome conformation capture (Hi-C) analyses used the Proximo Hi-C 2.0 Kit from Phase Genomics Inc. (Seattle, WA, USA) and was completed in the Martin laboratory and sent to Phase Genomics for paired-end sequencing of the Hi-C libraries using Ilumina HiSeq 4000.

Quality Control and* de novo *Genome Assembly

Raw PacBio and ONT reads were self-corrected and assembled using Canu v1.8 (Koren et al. 2016) with the minimum read lengths set to 500 bp and the corrected error rate set to 10.5% for *C. neglecta* and *C. laxa*, which both had over 30X (Table 1) coverage with PacBio in addition to the ONT coverage. The corrected error rate was set to 14.4% for both *C. hispida* var. *grandiflora*, where we only had ONT data, and *C. hispida* var. *hispida,* where we had less than 30X coverage in PacBio. Quality control and trimming of the Illumina paired end and mate pair data was completed with trimmomatic v 0.33 using a sliding window of four base pairs requiring an average phred score greater than 15 (Bolger et al. 2014). Draft assemblies were polished using the Illumina data and up to three iterations of Pilon v 1.23 (Walker et al. 2014) using bowtie2 v 2.3.4.3 (Langmead and Salzberg 2012) to align paired end data and Burrows-Wheeler Aligner (bwa) (Li 2013) to align the mate pair data. For C. *hispida* var. *grandiflora* and *C. laxa* , the program Purge Haplotigs (Roach et al. 2018) was used to reduce redundancy, resulting from the heterozygosity of the genomes, in the draft assemblies.

The quality of the polished assemblies was evaluated using QUAST 5.0.2 to obtain the

assembly metrics (Gurevich et al., 2013). The completion of the assembly's gene space was evaluated using BUSCO v3.0.2 (Simão et al. 2015) to identify conserved eukaryotic genes using the embryophyta_odb10 database.

The chloroplast genome for each species was extracted from the draft assemblies by aligning contigs with *C. sativa*'s chloroplast genome using nucmer v 4.0 from MUMmer v 4.0. Usually the chloroplast was represented by only a few contigs, which were ordered and orientated based on the position of match with *C. sativa*'s chloroplast, overlapped, and merged into a consensus assembly using a custom script written in R.

Genome Scaffolding

Phase genomics used their proprietary software, Proximo, to produce chromosome level scaffolds (Oddes et al. 2018) for *C. neglecta*, *C. hispida* var. *hispida* and *C. laxa* from the draft assemblies produced by Canu and polished by Pilon. A final round of polishing by Pilon was completed following scaffolding with the Hi-C data. The scaffolding tool, ntJoin 1.0.3-0 (Coombe et al. 2020), was used to scaffold the genome of *C. hispida* var. *grandiflora* using the Hi-C scaffolded assembly of *C. hispida* var. *hispida.* The primary contigs from the scaffolded assembly were then polished with Pilon and evaluated with QUAST and BUSCO.

All genomes are available from as part of Bioproject PRJNA750147.

Genome Annotation

Ab initio gene prediction was completed using AUGUSTUS 3.3.2 (Stanke et al. 2006) and transposable elements (TEs) were located using the Extensive *de novo* TE Annotator 1.8.3 (EDTA; Ou et al., 2010) and EAHelitron 1.5.1 (Hu et al. 2019). We determined whether Helitrons detected in *C. sativa*'s sub-genome 3 were also detected in the syntenic regions of the genomes of *C. hispida var. hispida* and *C. hispida* var. *grandiflora* using a script written in R

and output from EAHelitron and nucmer. Specifically, we examined all regions of all three genomes over 2000 bp in length that showed synteny, checked each genome region to determine if a Helitron was detected for the species, and then determined how many of the cases were shared.

Synteny between **Camelina** *diploids and* **C. sativa** *and* **A. lyrata**

To evaluate collinearity between the chromosome level draft genomes we used nucmer with both *C. sativa* and *A. lyrata* as references for comparison and scripts written in R employing the package circlize 0.4.11 (Gu et al. 2014) to order and visualize the alignments.

Phylogenetic analysis of **C. sativa** *sub-genomes and Camelineae diploids*

We determined the phylogenetic relationships among genomes for 1) diploid members of the Camelineae: *Arabidopsis lyrata* (Ensembl Genomes version 1.0), *Capsella rubella* (NCBI v. ANNY00000000.1), *Neslia paniculata* (L.) Desvaux (S. Wright personal communication) 2) the three sub-genomes of *C. sativa* (NCBI version JFZQ0000000.1) and 3) the four diploid *Camelina* species sequenced here. We used three methods to extract regions of the genomes to construct phylogenetic trees. Our first method identified random fragments within homologous regions of the genomes, our second method used a reciprocal best hit method to genes from *Arabidopsis thaliana* (TAIR version 10) represented in all assemblies, and third we used Orthofinder 2.3.11 (Emms and Kelly 2017; Emms and Kelly 2019) to identify orthologous gene groups and estimate species trees based on genes predicted *in silico* by AUGUSTUS.

First we determined homologous regions between the genomes by finding fragments within each ACK block as described in the *A. lyrata* genome that aligned well for all genomes of interest (Schranz et al. 2006) (Table S1). Each ACK region was cut into 1000 bp fragments, aligned to *A. lyrata*'s genome using bowtie2 (Langmead and Salzberg 2012) with the –very-

sensitive parameter, and filtered to exclude fragments that aligned to the genome more than once. Pairwise collinearity between this set of filtered fragments and each diploid Camelineae and *C. sativa* was determined using nucmer from MUMmer (v4.0). Then a custom R script was then used to find the collinear fragments that overlapped between all genomes within each of the 24 blocks. For a fragment to be included in the analysis it had to be at least 1000bp long (before trimming) and had to be found in three copies within a consistent set of *C. sativa*'s chromosomes within an AKB (Table S2). Each ACK block was aligned with the msa function in the R package msa 1.18.0 (Bodenhofer et al. 2015), which uses the "ClustalW" methods as a default.

Phylogenetic analysis was then conducted on each set of fragments. We first used stepping stone models in MrBayes v3.2.1 (Ronquist et al. 2012) to run for 6,000,000 generations to determine the appropriate model, ran the preferred model for each sequence for 5,000,000 generations, checked convergence of each run with the r package rwty (Warren et al. 2017), and then calculated seven metrics to allow exclusion of biased sequences or sources of misleading phylogenetic signal using TreeCmp 2.0 (Bogdanowicz et al. 2012) and TreSpEx 1.1 (Struck 2014). Specifically, following Nikolov et al. (2019) we calculated the number of matching splits and the Robinson-Foulds tree distances using TreeCmp; the upper quartile and standard deviation of the long-branch scores, average patristic differences, and R^2 of the saturation score and slope with TreSpEx. Fragments or genes were then excluded from further consideration if they failed the convergence checks or were outliers for one or more of the seven phylogenetic metrics at the 99th percentile. For each set of trees belonging to ABKs identified as located on the same set of *C. sativa* chromosomes, we estimated coalescent species trees with ASTRAL-III (5.1.1) (Zhang et al. 2018). This information was used to understand the sub-genome structure of *C. sativa* and assign specific chromosomes to a particular sub-genome based on phylogenetic

relationship.

Our second method used this information clarifying the sub-genome structure of *C. sativa*, to identify reciprocal best hits (Chen et al. 2017) for *A. thaliana* genes to be used in phylogenetic analysis. Specifically, information on the locations of genes identified for *A. thaliana* in the TAIR 10 release were extracted from the genomic feature file (available from www.arabidopsis.org) and used to extract the sequences for each genome from *A. thaliana*'s genome assembly (TAIR 10 Release). BLAST 2.2.31 (Altschul et al. 1990) was then used to find these genes in each of the genomes included in the study. Following recommendations by Chen et al. (2017) these hits were screened to include hits with an e-value of less or equal to 0.0001, that were at represented 70% or more of the original query length, had 70% or greater identity with the query, and had a bit score ratio between the first and second BLAST hit of 1.2 or greater. The sequences of these putative matches were then extracted from each genome and BLAST was used to determine if these genes mapped back to the expected location in *A. thaliana*'s genome when screened with the same criteria as above. The intersection of these sets from all genomes of interest were then divided in the ACK and ABK groups based on their position in *A. lyrata*'s genome and included in the phylogenetic analysis.

The phylogenetic analysis of these best hits was then completed as above using MrBayes and ASTRAL-III to estimate coalescent species trees by ABK. However, we then extended the phylogenetic analysis for both the selected fragments and the reciprocal best hit genes. One randomly selected sequence from each ACK, resulting in 24 sequences sampled across the genome 25 times, to represent 25 (x2) sets of unlinked data was used for analysis using StarBEAST2 2.6.3 (Ogilvie et al. 2017; Suchard et al. 2018; Bouckaert et al. 2019), PhyloNet 3.8.2 (Than et al. 2008; Wen et al. 2018; Cao et al. 2019) and to generate a consensus tree by

ASTRAL-III.

StarBEAST2 co-estimates gene and species trees and takes into consideration gene tree heterogeneity resulting from incomplete lineage sorting (Ogilvie et al. 2017). StartBeast2 was run for between 20 and 100 million generations as required to produce estimated sample sizes (ESS) values above 200 with the GTR site model using a configuration file created by BEAUti 2.6.3 (Bouckaert et al. 2019), convergence was examined with Tracer (1.7.1) (Rambaut et al. 2018), summarized with TreeAnnotator 2.6.3 (Bouckaert et al. 2019) and plotted with the densiTree function in the R package phangorn 2.5.5 (Schliep 2011).

Similarly, PhyloNet accounts for incomplete lineage sorting, but allows Bayesian inference of phylogenetic networks where hybridization has resulted in a reticulate evolutionary history (Wen et al. 2018). PhyloNet's MCMC_Seq command was run with chain lengths of between 20 and 80 million as required to produce ESS values above 200 and the maximum number of reticulations set to default number of 4. Networks and trees produced by PhyloNet were visualized using plotTree function from the package phytools 0.7.20 (Revell 2012).

The third method was the most straight forward. AUGUSTUS was run on the final version of the assembled and reference genomes to predict genes *in silico* using *Arabidopsis* as the reference species. Amino acid sequences were then provided to OrthoFinder, which grouped them into orthogroups, genes descended from a single gene in the last common ancestor of the species included in the analysis. OrthoFinder then produced a rooted species tree inferred from all the orthogroups with support levels using the default species tree inference from all genes (STAG) method and by the multiple sequence alignment method which concatenates single-copy orthogroups (Emms and Kelly 2018). By considering all of the orthologs detected this method can overcome situations where similarity based methods such as the reciprocal best hit methods

when orthologs are not identified or are misidentified (Emms and Kelly 2019).

Entire chloroplast genomes for *Arabis alpina*, *Arabidopsis thaliana*, *A. lyrata*, *Capsella bursa-pastoris*, *C. grandiflora*, *C. rubella* and *Camelina sativa* were downloaded from the Chloroplast Genome Database (https://rocaplab.ocean.washington.edu/old_website/tools/cpbase) and aligned with the four *Camelina* chloroplasts using the msa function in R (see above). A consensus tree was then estimated with MrBayes as above. Annotation of the chloroplast genomes was completed with GeSeq 1.77 (Tillich et al. 2017), which is available as one of the tools on the CHLOROBOX website (https://chlorobox.mpimp-golm.mpg.de/index.html), using *Camelina sativa*, *A. lyrata,* and *Capsella rubella*'s chloroplasts as reference sequences. The chloroplast genomes with their annotations were then visualized with OrganellarGenomeDRAW (OGDRAW) 1.3.1 (Greiner et al. 2019) also available on the CHLOROBOX website.

Additional software used

The version of R used here was 3.6.3 (2020-02-29) -- "Holding the Windsock." Sequence handling, tree plotting and graphical display were facilitated by numerous R packages in addition to those mentioned above including: ape 5.0 (Paradis and Schliep 2019), apex 1.0.4 (Schliep et al. 2020), Biostrings 2.56.0 (Pagès et al. 2020), pals 1.7 (Wright 2021), pBrackets 1.0.1 (Schulz 2021), plotrix 3.7.8 (Lemon 2006), plyr 1.8.6 (Wickham 2011), ips 0.0.11 (Heibl 2008), IRanges 2.22.2 (Lawrence et al. 2013), outliers 0.14 (Komsta 2011), Rsamtools 2.4.0 (Morgan et al. 2020), seqinr 3.6.1 (Bastolla et al. 2007), stringr 1.4.0 (Wickham 2019), and treeio 1.10.0 (Wang et al. 2020).

The program FigTree 1.4.4 (Rambaut 2018) was used to convert trees including their support values to a format easily readable in R.

Unless otherwise specified all tools were run using default settings.

Results

Genome Assemblies

Camelina neglecta plants were inbred through self-pollination for five generations to reduce heterozygosity, while the three outcrossing species were sequenced without inbreeding. The variation in sequencing effort and in the technology used resulted in assemblies with varied initial levels of fragmentation and completeness. Coverage of the genome by each sequence type differed for each species with *C. neglecta* receiving the majority of our sequencing efforts with a total a total of 286x raw coverage (Table 1). Following assembly with Canu and polishing with Pilon, *C. neglecta* had the most contiguous assembly with 204 contigs and an NG50 of 11,493,634, while *C. hispida* var. *hispida* had the most fragmented assembly with 2,779 contigs and an NG50 of 566,799 (Table 2). Scaffolding using Hi-C data and Phase Genomics' Proximo resulted in chromosome level assemblies for *C. neglecta* (n = 6), *C. hispida* var. *hispida* (n = 7) and *C. laxa* (n =6) with approximately 70% or more of their expected lengths and NG50s of 29,279,412. 39,460,631, and 31,147,072 respectively. Following scaffolding by ntJoin, 70% of the expected genome length for *C. hispida* var. *grandiflora* was incorporated into a chromosome level assembly (n = 7). The completeness of gene space, as evaluated by estimating the proportion of core conserved eukaryotic genes using BUSCO, indicated that all assemblies had at least 90% of the expected genes (Table 2) with the Hi-C scaffolded genomes: *C. neglecta* having 98% of these genes, *C. hispida* var. *hispida* 96% and *C. laxa* 96%.

Phylogenetic relationships among *C. sativa* and Camelineae diploids

In total 1,444 fragments within the ACK blocks described for *A. lyrata* genome were found to be shared among the eight core Camelineae taxa with three copies distributed across *C. sativa*'s genome. Fragments from each ancestral chromosome showed the same pattern of localization in *C. sativa*'s genome in the majority of cases (Table 3). For example ACK blocks A, B, and C from ancestral chromosome 1 all showed the highest number of hits on *C. sativa*'s chromosomes 3, 14 and 17. Where differences occurred, there was consistency corresponding to ancestral chromosome arms and, therefore, ABK group. For example, ABK blocks AK11 (OP) and AK12 (QR) from ancestral chromosome 6 showed slightly different patterns of distributions with the greatest number of hits on *C. sativa*'s chromosomes 8, 13, and 12 and 8, 13, and 20 respectively. The pattern of hits across the genomes resulted in eleven groupings used in further phylogenetic analyses that corresponded to either ancestral chromosome arms or entire chromosomes (Table 4).

Coalescent species trees produced by ASTRAL-III using trees produced for each fragment within these eleven summary groups indicated the phylogenetic relationship between each set of *C. sativa*'s chromosomes and the other taxa. Summary trees produced by ABK did not differ from these relationships and the trees in each group were largely congruent as indicated by the high normalized quartet scores (Table 4). Here we defined sub-genome 1 as the genome with closest relationship to *C. neglecta*, while sub-genome 3 was defined as that with the closest relationship to the *C. hispida* varieties (Table 4; Fig. 1). This sub-genome definition differs from that originally published by Kagale et al. (Kagale et al. 2014) and uses a different nomenclature than our previous work (Lujan Toro 2017), but is concordant with the recently published revised genome structure for *C. sativa* (Chaudhary et al. 2020).

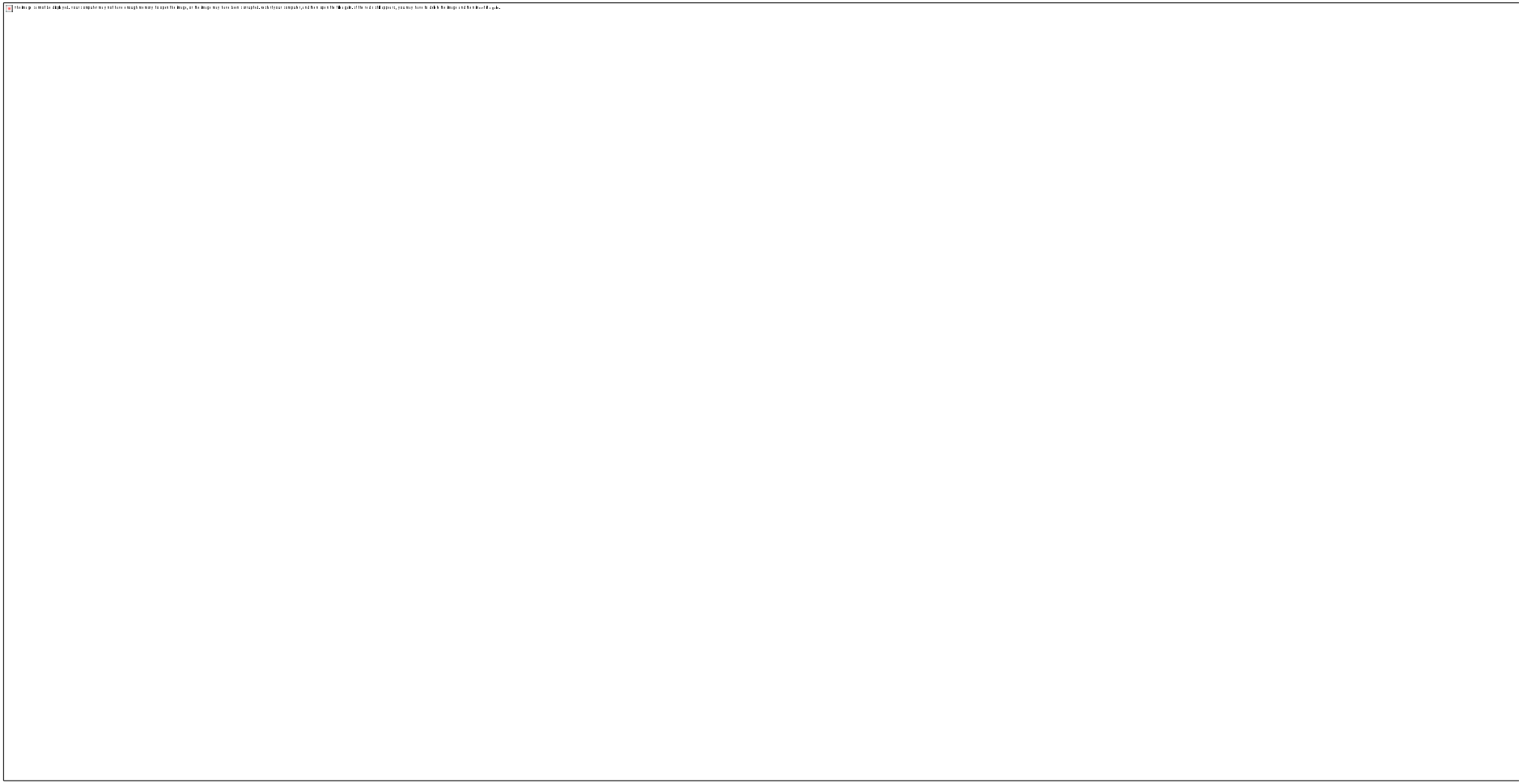


Fig 1. Coalescent Species Trees. These coalescent species trees were produced by ASTRAL-III using trees produced for each fragment identified by bowtie 2's alignment of fragments from *A. lyrata* by MrBayes and grouped within the eleven ABK groups, corresponding to either whole ancestral chromosomes or chromosome arms. The ACK(s) included in each tree are indicated in the box to the left of each tree and the ancestral chromosome structure with division into ACK and ABK is shown in the lower right.

The consensus species trees estimated by StarBEAST2 for either set of sequence data indicated strong consistency in the close relationship of *C. sativa*'s sub-genome 1 and *Camelina neglecta*, with *C. sativa*'s sub-genome 2 falling into a clade with these two genomes, and *C. sativa*'s sub-genome 3 close relationship with the varieties of *C. hispida* (Fig. 2 and 3). Also consistent was the placement of *Arabidopsis* as the most distant genus and *Neslia* as the closest genus of those studied from within the Camelinaeae. As with the consensus trees from ASTRAL-III, the primary uncertainty in the topology of the tree was the position of *C. laxa* - whether it is basal to all the *Camelina* spp. analyzed here or whether it is in a clade with *C. sativa*'s sub-genome 3 and the varieties of *C. hispida*. Specifically, of the randomly chosen 25 sets of 24 sequences from across the genome, ASTRAL-III consensus trees produced trees with *C. laxa* basal to the *Camelina* group just over half the time - 14 sets out of 25 for the fragments and 15 sets out of 25 for the reciprocal best hit genes. The consensus trees produced by StarBEAST2 showed greater disparity in how many fragments indicated *C. laxa* is basal between the two sets of sequence data with 18 of 25 suggesting *C. laxa* is basal, and with two trees not corresponding to either of the two common topologies. For the reciprocal best hit genes, only 6 sets out of 25 suggested *C. laxa* is basal with a thin majority -14 out of 25 suggesting *C. laxa* in a clade with *C. sativa*'s sub-genome 3 and four trees not corresponding to either of the two common topologies. Density tree plots of trees generated by StarBEAST2 indicated that the species trees occasionally showed mixed signals for *C. laxa*'s position with both sequence fragments (Fig. 2A) and genes (Fig. 2B). These plots also indicated that some trees differed in the timing of divergence in genes among *C. laxa*, *C. sativa*'s sub-genome 3, and the varieties of *C. hispida* (Fig. 2C). Both ASTRAL-III and StartBeast2 account for incomplete lineage sorting; however, given the propensity of Brassicaceae taxa to hybridize, reticulate evolution was further

investigated with PhyloNet. For many of the sets of 24 unlinked fragments or genes, PhyloNet indicated trees without reticulations were the most credible representations of the evolutionary history of the species', both for the fragments (10 of 25) or the reciprocal best hit genes (17 of 25). However, with other sets, one reticulation was included in the credible networks. The placement of these reticulations was not consistent, but often (5 of 8) included a link between the node before or the terminal branch of *C. laxa* linking to the base of the *Camelina* clade or to the clade with *C. sativa*'s sub-genome 3 in the consensus (Fig. 3 A) and most credible networks (Fig. 3 B, C, D) produced by the set of fragments or genes.

Fig. 2. Density tree plots. Density tree plots of 1,000 randomly chosen tree estimated by StarBEAST2 for a set of A) fragment sequences, and B) genes, C) two sets of reciprocal gene sequences.

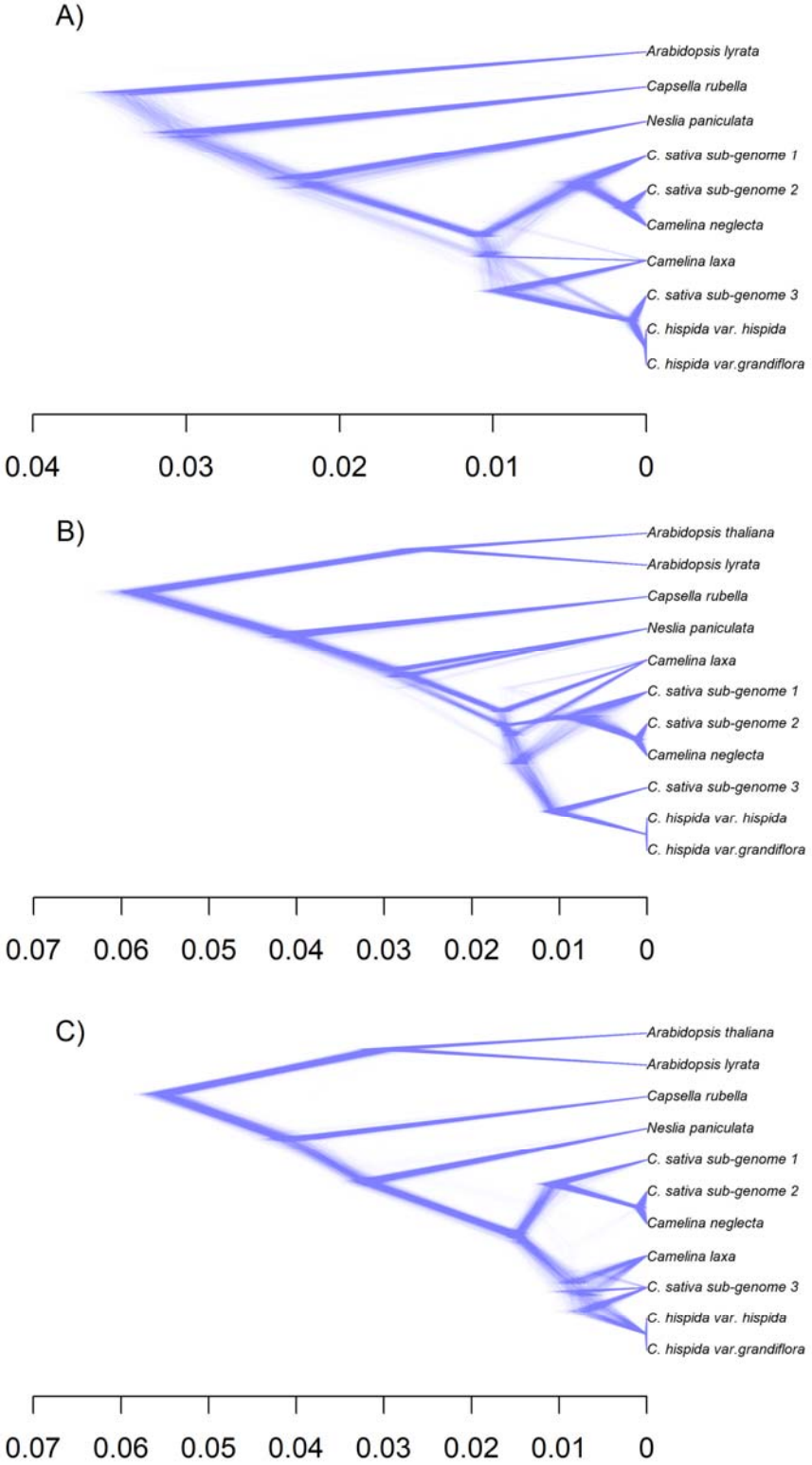
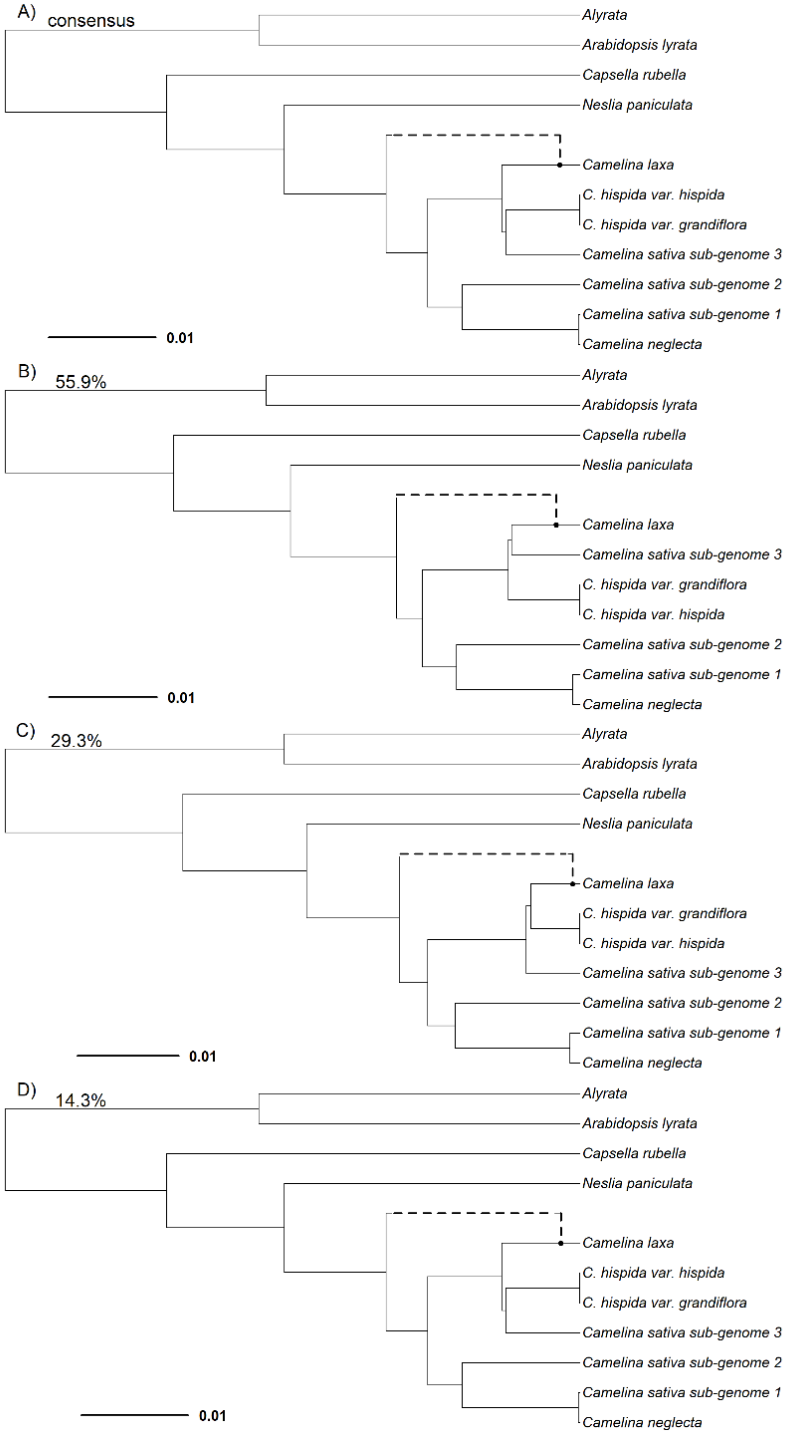


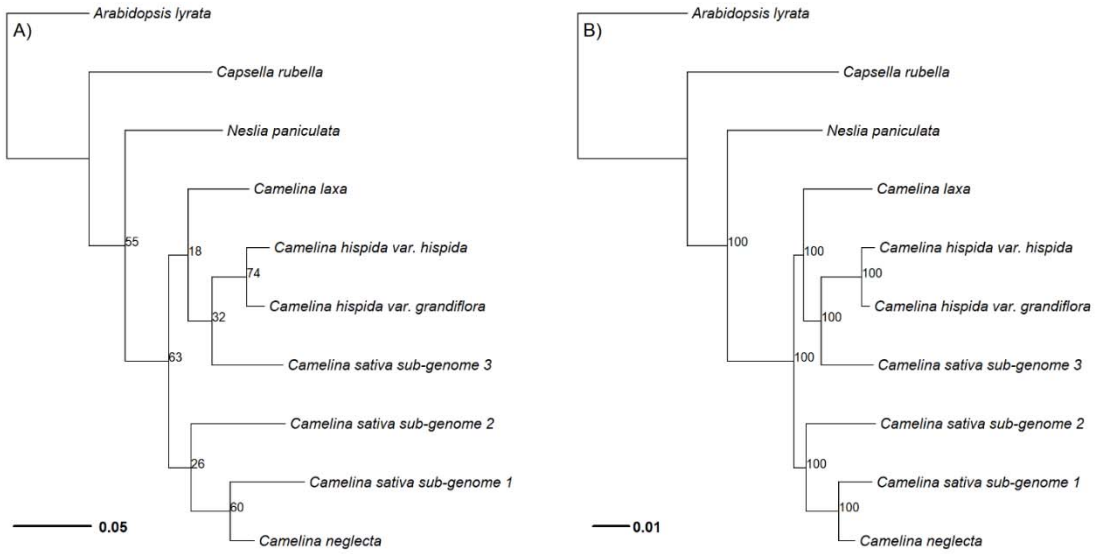
Fig. 3. **Consensus networks.** The consensus networks generated with PhyloNet using one set of reciprocal gene sequences (A) and the three most credible networks contributing to this consensus (B-D) with the percentage of credible networks represented.



Results from OrthoFinder

Given the ten sub-genomes and genomes and the *in silico* predictions of amino acids produced by AUGUSTUS, OrthoFinder was able to assign 93.4% of the 419,400 predicted genes to orthogroups with 14,513 genes common to all ten genomes. The final species tree predicted by the default STAG method (Emms and Kelly 2018), a consensus tree based 14,513 genes trees with an inferred root using STRIDE method (Emms and Kelly 2017), indicated that *A. lyrata* was the best root for the tree, that *C. neglecta* was most closely related to *C. sativa*'s sub-genome 1 and then to sub-genome 2, and that these taxa were sister to a clade containing *C. laxa*, both *C. hispida* varieties and *C. sativa's* sub-genome 3 (Fig. 4). However, the support values for each bipartition based on the number of individual species trees that contained that bipartition, indicated that this topology was often represented by a small proportion of the gene trees (Fig. 4a). The tree produced by OrthoFinder using the option to create a tree based on the concatenated multiple sequence alignment of 7,401 shared, single-copy genes resulted in a tree with the same topology, but with higher levels of support as is typical in analyses of concatenated data (Fig. 4b).

Fig 4. Trees from OrthoFinder. Phylogenetic trees generated by OrthoFinder based on A) a consensus of 14,513 gene trees with an inferred root and B) a concatenated multiple sequence alignment of 7,401 shared, single-copy genes. Node labels indicate support values for each bipartition.

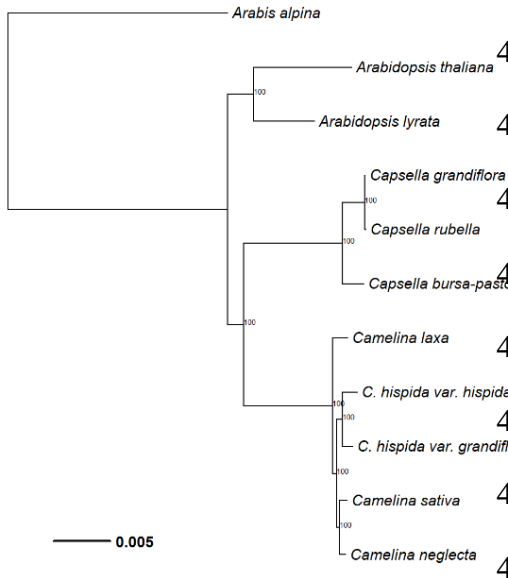


Results for Chloroplasts

The chloroplast (cp) assemblies produced for the four diploid *Camelina* species sequenced here, based on the scaffolding of contigs using *C. sativa*'s cp, resulted in contigs between 152,239 bp (*C. hispida var. grandiflora*) and 153,366 bp (*C. neglecta*) in comparison to the 153,044 bp for *C. sativa's* cp assembly. Annotation by GeSeq indicated expected features such as the large inverted repeat and that ARAGORN 1.2.38, invoked by GeSeq, found the same 39 genes in all four of the *Camelina* cp's assembled here and that of *C. sativa*. The cp assemblies varied in the degree of fragmentation of the genes within the assembly with *C. hispida var. grandiflora* showing the greatest number of fragmented features (89) and *C. neglecta* the fewest (20) (S1 Fig). The resulting tree from entire chloroplast sequences of all eleven taxa included in the analysis, showed the chloroplast sequence from *C. sativa* was most closely related to that of *C. neglecta* and that these taxa were sister to a clade containing the varieties of *C. hispida* with *C. laxa* basal to the *Camelina* species included in the analysis. This suggests the maternal lineage of *C. sativa* is most closely related to *C. neglecta*. As with the analysis of nuclear genes, the

chloroplast shows the clade of *Capsella* species as more closely related to *Camelina* than the *Arabidopsis* clade (Fig. 5).

Fig. 5. **Chloroplast Based Phylogeny**. Phylogeny constructed from whole chloroplast alignment using MrBayes. Node labels indicate posterior probabilities.



Synteny between Camelina diploids and A. lyrata and C. sativa

As expected from previous work on species from the Brassicaceae and, more specifically, within the Camelinaeae, the four diploid species sequenced here showed extensive synteny with both *A. lyrata* and *C. sativa* genomes with conservation of the ACK and ABK blocks (Fig. 6). The analysis indicates a high level of synteny between *A. lyrata* and *C. neglecta*, and to a greater degree, between *C. neglecta* and *C. sativa* sub-genome 1 (Fig. 6, Fig. S1). Similarly, there is strong synteny between *A. lyrata* and *C. hispida* and, to a greater extent, between *C. hispida* and *C. sativa* sub-genome 3 (Fig. 6, Fig. S2). These results echo the phylogenetic analysis suggesting less phylogenetic distance between *C. sativa* sub-genome 1 and *C. neglecta*, and, *C. sativa* sub-genome 3 and *C. hispida*. In contrast the genome of *C. laxa* has more extensive rearrangements and breaks within ACK and ABK blocks (Fig. 7, Fig. S3) including fragmentation of the E block, and separation of parts of the F, J, U, and W blocks on to separate chromosomes. Some of

this fragmentation maybe an artifact of the fragmentation of the genome prior to scaffolding with HiC data that would be reduced with an improved assembly. However, it appears that, in contrast to *C. neglecta*, where the reduction in chromosome number was the result of the fusion of two chromosomes (Al07 and Al08), portions of several of the ancestral chromosomes has been incorporated into several chromosomes (Fig. 7). For example, AK05 (H) is part of C. laxa's chromosome 6 while the upper arm of the chromosome AK06 (FG) has been split and is found on *C. laxa*'s chromosomes 3 and 5. As a result, the reduction of chromosome number from 7 to 6 in *C. laxa* is the result of more extensive genomic rearrangement than the reduction in *C. neglecta*.

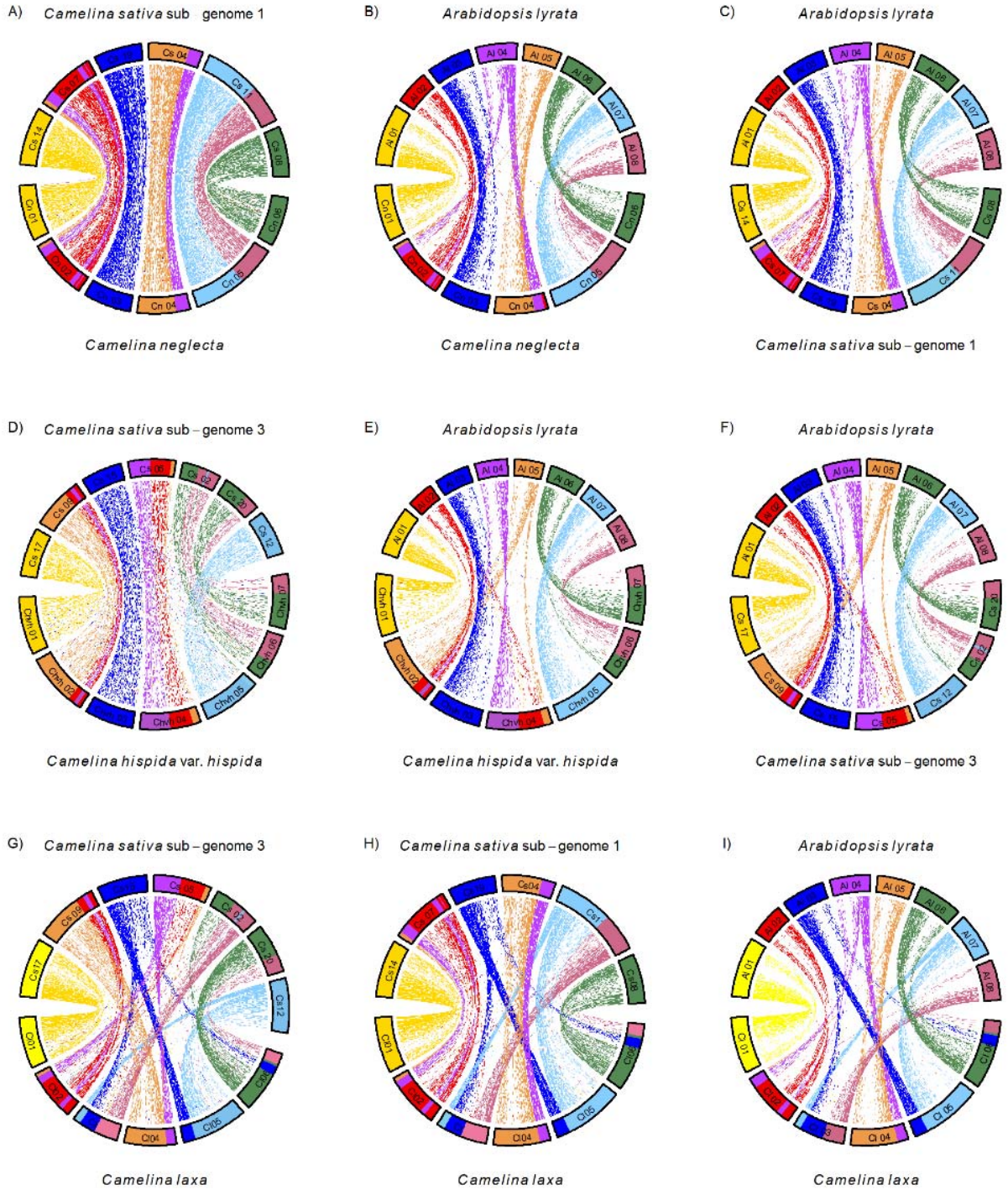


Fig. 6. Synteny Plots Among Genomes. Plots indicating regions of synteny as indicated by nucmer among the diploid *Camelina* species *C. neglecta*, *C. hispida* var. *hispida* and *C. laxa* with *C. sativa*'s sub-genomes (A, D, G, H) and with *A. lyrata* (B, E , I). Synteny between *C. sativa* sub-genomes and *A. lyrata* are shown for comparison (C, F). All plots are coloured by alignment with *A. lyrata's* chromosomes, which are analogs for the ancestral chromosomes.

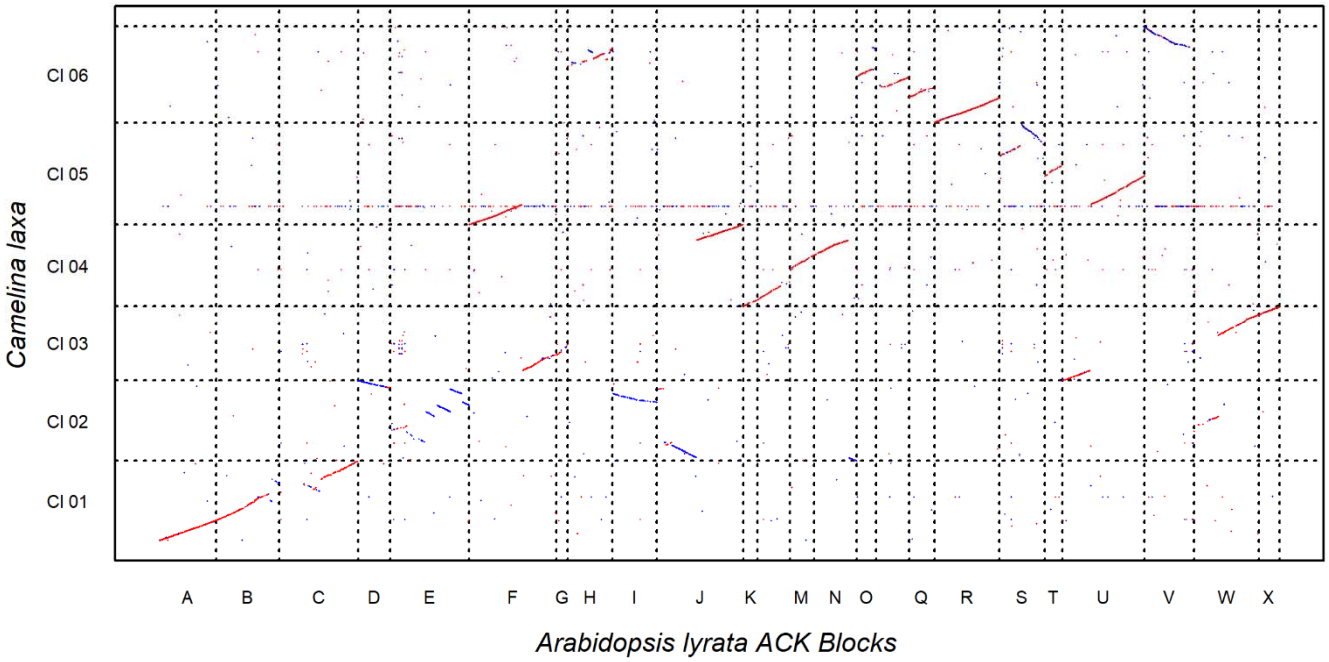


Fig. 7. Synteny Plot between AKB Regions and *C. laxa*. Plot of synteny between ACK regions of *A. lyrata*'s genome and *C. laxa*'s genome indicating breaks and inversions with portions of the blocks F, J, U and W distributed across several chromosomes.

Transposable element annotation and comparison

Annotation of the transposable elements (TEs) in the genomes assembled here by the Extensive *de novo* TE Annotator (EDTA) indicated that TEs generally made up 34-35% of the diploid *Camelina* genomes, with the exception of *C. hispida* var. *hispida* where TEs accounted for 50% of the genome (Fig. 8, Table S2). The largest groups of TEs identified belonged to the Helitron (DHH) and Gypsy

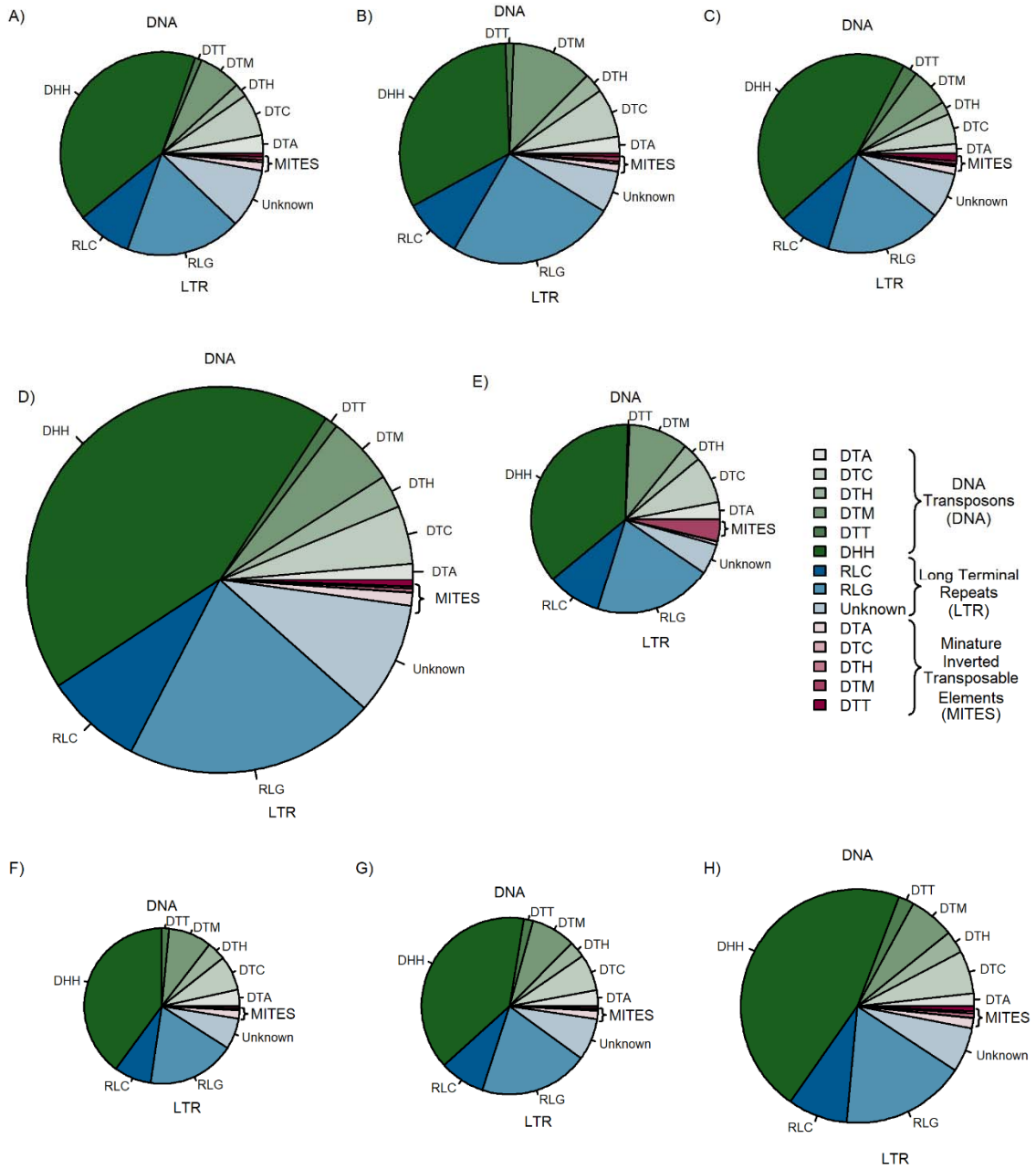


Fig. 8. **Transposable Element Types and Abundance.** Types of transposable elements (TE) in A) *Camelina neglecta*, B) *C. laxa*, C) *C. hispida* var. *grandiflora*, D) *C. hispida* var. *hispida*, E) *Arabidopsis lyrata*, F) *C. sativa* sub-genome 1, G) *C. sativa* sub-genome 2, and H) *C. sativa* sub-genome 3, as annotated by Extensive De novo Transposable Element (EDTA). Pie graphs are scaled by the percentage of the genome attributed to TEs which at highest is 49.6% for *C. hispida* var. *hispida* (D). Classification of transposable element type follows the unified classification system for eukaryotic transposable elements that uses a three letter code indicating class, order and superfamily (34). Here these are divided into three groups: DNA transposons (DNA), which includes Helitrons (DHH); long terminal repeats (LTR), which includes retrotransposons RLC (Copia) and RLG (Gypsy); and miniature inverted transposable elements (MITES).

superfamilies (RLC). EAHelitron identified fewer Helitrons than EDTA (Table 5). However, the difference in the number of Helitrons in *C. neglecta* and *C. hispida* var. *hispida* was even more divergent with Helitron densities of 9.6 and 20.4 per 1,000,000 bp respectively. The percentage of *Camelina hispida* var. *grandiflora*'s genome (34%) and the Helitron density (5.7) was much lower than that of *C. hispida* var. *hispida*. The three sub-genomes of *C. sativa* also showed differences in the number of TEs and, specifically Helitrons, with sub-genome 1 and 2 showing similar percentages of TEs at 30.2% and 26.6% and similar Helitron densities at 6.1 and 5.8, but sub-genome 3 showing higher values at 40% and 14.7. In syntenic regions of *C. sativa*'s sub-genome 3, *C. hispida* var. *grandiflora*, and *C. hispida* var. *hispida*, sites with Helitrons were found to be shared almost twice as often between *C. sativa*'s sub-genome 3 and *C. hispida* var. *hispida* (88 shared sites) compared to *C. sativa*'s sub-genome 3 and *C. hispida* var. *grandiflora* (46 shared sites) (Table 6).

Discussion

Whole genome sequencing is allowing a deeper understanding of the processes that have shaped plant evolution including polyploidization, hybridization and chromosomal rearrangements. Allopolyploid crops have become excellent systems for understanding these genomic changes because they have often received substantial sequencing effort (e.g. *Triticum aestivum* L. (Appels et al. 2018), *Solanum lycopersicum* L. (Sato et al. 2012), *Zea mays* ssp. *mays* L. (Schnable et al. 2009; Jiao et al. 2017), and *Arachis hypogaea* L. (Bertioli et al. 2019)). Moreover, as work on these species has shown, a greater understanding of the genomic changes associated with allopolyploidization or autopolyploidization, the genomic consequences of domestication, and the potential breeding resources can be achieved when related, extant diploid species are also sequenced (e.g. 19,35,39,40).

The genome of the allohexaploid *Camelina sativa* (L.) Crantz (camelina; 2n = 40) has been sequenced (Kagale et al. 2014), but the diploid ancestors of the three parental genomes have not been identified. Four diploid taxa are known within the genus: *Camelina neglecta*, *Camelina laxa*, *Camelina hispida* var. *hispida* and *Camelina hispida* var. *grandiflora*. Here we sequenced the genomes of these four diploids with the goal of examining their phylogenetic relationship with *C. sativa* and of comparing the structure of these genomes to the *C. sativa*'s sub-genomes. Specifically, we produced chromosome level drafts for all four genomes. Three of these, *C. neglecta*, *C. hispida* var. *hispida*, and *C. laxa*, were scaffolded based on Hi-C data while the *C. hispida* var. *grandiflora* was scaffolded using the chromosome level draft of *C. hispida* var. *hispida*. Evaluation of the gene space by BUSCO suggests that more than 90% of the core genes are represented in these assemblies containing 70% or more of the expected sequence length (Table 2). Each genome showed strong synteny with *A. lyrata* with conservation of the ACK and ABK blocks, with each region mapping once, confirming a diploid structure (Fig. 6, 7). This conservation of ACK and ABK blocks allowed for further dissection of the genomes.

Coalescent species trees produced by ASTRAL-III using trees produced from sequence data from each ABK or ACK indicated the phylogenetic relationships between these regions from *C. sativa*'s chromosomes and the other taxa included here (Fig. 1). The trees in each group were largely congruent as indicated by high normalized quartet scores (Table 4). These phylogenies suggest that the assignment of *C. sativa*'s chromosomes to sub-genomes requires reassessment compared to that suggested based on a visual evaluation of the level of synteny with *A. lyrata* by Kagale et. al (Kagale et al. 2014). However, genome rearrangement and fractionation are not predominant mechanisms in *C. sativa*, with *C. sativa's* chromosomes

showing very little reshuffling providing limited visual information for this division (Lysak et al. 2016). Here, we define sub-genome 1 as composed of the chromosomes with the closest phylogenetic relationship with *C. neglecta*, sub-genome 2 as those chromosomes sister to the clade with *C. neglecta* and sub-genome 1, and sub-genome 3 as chromosomes with the closest phylogenetic relationship to the varieties of *C. hispida* (Table 4; Fig. 1). This results in composition, ordered by ACK regions, of sub-genome 1: *Camelina sativa* chromosome (Cs) 14, Cs07, Cs19, Cs04, Cs08, and Cs11; sub-genome 2: Cs03, Cs16, Cs01, Cs06, Cs13, Cs10, and Cs18; and sub-genome 3: Cs17, Cs09, Cs15, Cs05, Cs02, Cs20, and Cs12. This contrasts with the composition sub-genome 1: Cs17, Cs16, 01Cs15, Cs04, Cs13 and Cs11; sub-genome 2: Cs14, Cs07, Cs19, Cs06, Cs08, Cs10, and Cs18; and sub-genome 3: Cs03, Cs05, Cs01, Cs09, Cs20, and Cs02 suggested by Kagale et. al (2014). This sub-genome definition uses a different nomenclature than our previous work which defined sub-genome 2 as most closely related to *C. neglecta* (Lujan Toro 2017). However, it is concordant with the revised sub-genome definition recently published for *C. sativa* (Chaudhary et al. 2020). In that analysis, the authors characterized 193 accessions of *Camelina*, including *C. neglecta*, *C. laxa*, *C. hispida*, *C. rumelica* Velen., tetraploid and hexaploid *C. microcarpa* Andrz. ex DC., and *C. sativa*, using a genotyping by sequencing (GBS) approach that mapped SNPs in sequences for these accessions to *C. sativa*'s genome. They determined that sequences from *C. neglecta* aligned to six of *Camelina sativa*'s chromosomes and *C. hispida* showed bias toward alignment with sub-genome 3. Additionally, tetraploid *C. microcarpa* aligned with 13 of the chromosomes. However, they determined that alignment reads did not correspond to the sub-genome structure published in Kagale et. al (Kagale et al. 2014) and refined the composition of the sub-genomes accordingly with *C. neglecta* sequences aligning to sub-genome 1, the tetraploid *C. microcarpa* sequences

aligning to sub-genomes 1 and 2, and *C. hispida* aligning to sub-genome 3 (Chaudhary et al. 2020). This indicates that very different approaches, an analysis based on SNPs and a phylogenetic analysis of whole genome sequences, has resulted in similar conclusions about refinements to the sub-genome structure of *C. sativa*.

With these alterations to the sub-genome structure of *C. sativa*, there is strong synteny between the draft *C. neglecta* and *C. sativa*'s sub-genome 1 and 2 (Fig. 7 A-C). Specifically, two larger areas with inversions are apparent when *C. neglecta*'s genome and *C. sativa*'s sub-genome 1 are compared: one at the start of *C. neglecta'*s Chromosome 6 (Cn06) and end of Cs08; and one within Cn02 and Cs07. Interestingly, the difference between Cn06 and Cs08 shows that Cn06 is more similar to *A. lyrata*'s genome structure suggesting this change occurred in *C. sativa*'s genome 1 following allopolyploidization (Fig. 7 A-C). Similarly, when *C. neglecta*'s genome and *C. sativa*'s sub-genome 2 are compared the largest difference is the lack of fusion in Cs10 and Cs18 compared to Cn05, and several areas of inversions within Cs16 and Cn02, and within Cs01 and Cn03. This raises the question of the origin of the n = 7 for sub-genome 2 and whether it came from an ancestor or sister species of *C. neglecta* where these chromosomes had not fused or represents a re-breakage of the fused chromosome. There is also strong synteny between *C. hispida* var. *hispida* and sub-genome 3, with two inversions, one the start of Cs09 and one at the end of Cs12 represented in the Hi-C scaffolded draft (Fig. 6 D, G). In all three cases, many of the changes in the genome structure compared to *A. lyrata* are shared between the diploid genomes and the sub-genomes of *C. sativa* (Fig. 7 E-G). This conservation of structure in these closely related genomes is in line with Lysak's expectations that the majority of the changes in chromosome structure seen in *C. sativa* compared to the ancestral chromosomes were likely present in the ancestral diploid genomes (Lysak et al. 2016). However, as these drafts are

based on sequencing data, verification of these inversions should be completed using genome *in situ* hybridization (GISH) using bacterial artificial chromosomes (BACs) constructed for that purpose (Lysak et al. 2016).

The position of *C. laxa* in the phylogenetic tree was the least stable element across analyses with two main alternatives, either as basal to all the other *Camelina* species sampled here or as basal to the clade containing the varieties of *C. hispida* and *C. sativa*'s sub-genome 3. More rarely a third topology with *C. laxa* basal to the clade with *C. neglecta* and *C. sativa*'s sub-genomes 1 and 2 was also produced. The consensus from previous work is that *C. laxa* should be considered basal to the other species of *Camelina*. Brock et. al (Brock et al. 2018) created a phylogeny using data from ddRADseq for 48 specimens from gene bank material and field collections from Turkey, Jordan, and Armenia for *C. sativa*, *C. microcarpa* hexaploids, *C. rumelica*, *C. laxa* and *C. hispida*. In the maximum likelihood consensus tree generated from this data, *C. laxa* was basal to the other species. A neighbor joining tree produced using GBS data indicated *C. laxa* was basal to the rest of the clade except the tetraploid *C. rumelica* (Chaudhary et al. 2020). In a more comprehensive analysis of the Camelineae that included *C. alyssum* (Mill.) Thell. and *C. anomala* Boiss. & Hausskn. ex Boiss. as well as representatives from many other genera included in the tribe *Nelisa*, *Capsella*, *Arabidopsis*, *Catolobus*, *Pseudoarabidopsis*, *and Chryochamela*, Čalasan et al. (Čalasan et al. 2019) constructed trees using both maximum likelihood and Bayesian inference with sequence data from the EST region of the ribosome. In this tree *C. laxa* was also placed as basal to the other species of *Camelina*. Inter-genome recombination has been observed in some allopolyploids, including resynthesized *Brassica napus* (Pires et al. 2004), and if this has occurred in *C. sativa*'s genome it could result in conflicting signals in some loci, as could historical gene flow and introgression between *C. laxa*

and other *Camelina* species. Some genes appear to show a signal of hybridizations among *C. laxa*, *C. sativa*'s sub-genome 3 and the varieties of *C. hispida* (Fig 2., Fig 3.), and further evaluations of this possibility could be completed with broader population level sampling of the taxa and evaluation of the evidence of introgression (Martin et al. 2015; Martin and Jiggins 2017; Crowl et al. 2020) . The results presented by Čalasan et al. (Čalasan et al. 2019) also suggest that *C. anomala*, a taxon with unknown chromosome number, is sister to *C. hispida* raising the possibility that this species could also be a close relative of *C. sativa*'s sub-genome 3.

The sub-genomes of the tetraploid and hexaploid members of the genus could also be more closely related to *C. sativa*'s sub-genomes than the taxa represented here as suggested by the affinity observed between *C. sativa*'s sub-genome 2 and that of the tetraploid *C. microcarpa* (Chaudhary et al. 2020). Future work to evaluate the phylogenetic position of the sub-genomes of the tetraploid and hexaploid species or a phylogenetic analysis of markers that explicitly accounts for polyploidy would be of value in increasing our understanding of the genus and of the evolutionary history of *C. sativa*.

While the phylogenies indicate that both varieties of *C. hispida* share a common ancestor with *C. sativa*'s sub-genome 3, the frequency and distribution of TEs and, in particular, the Helitrons detected here suggest *C. hispida* var. *hispida*'s genome may most closely resemble the genome that contributed sub-genome 3 to *C. sativa*. The frequency of TEs varies across *C. sativa*'s sub-genome's with sub-genome 3 having the highest number of these elements, at 40% of the sub-genome, compared to 30% in the other two (Fig. 8., Table 5, SI Table 2). Among the diploids sequenced here this percentage is closest to the percentage observed for *C. hispida* var. *hispida* at almost 50%. The largest portion of these TEs belong to the Helitron family and make up 15% of *C. sativa*'s sub-genome 3 and 16% of *C. hispida* var. *hispida's* genome. Helitrons

were first discovered in the genomes of *Arabidopsis thaliana*, *Oryza sativa* L. and *Caenorhabditis elegans* Maupas (Kapitonov and Jurka 2001). Helitrons are DNA transposons with lengths ranging from 0.5-3kb that transpose by rolling-circle replication without an RNA intermediate (Hu et al. 2019). Hu et al. (Hu et al. 2019) have suggested that, because Helitron density is highly evolutionarily labile, highly variable within genera, they could be used as characteristic for species identification. In this case, they may also provide clues to the characteristics of the genomes contributing to the formation of allopolyploids. Here we used the software developed by Hu et al. (Hu et al. 2019), EAHelitron, to examine whether Helitrons detected in *C. sativa*'s sub-genome 3 were more often shared in regions of synteny with *C. hispida* var. *hispida* or *C. hispida* var. *grandiflora*. This analysis indicated that almost twice the number of Helitron insertion sites found in *C. sativa*'s sub-genome 3 were also found in *C. hispida* var. *hispida's* genome compared to *C. hispida* var. *grandiflora's* genome (Table 6). This suggests that the genome that contributed *C. sativa*'s sub-genome 3 was more similar to the genome of *C. hispida* var. *hispida* with its high TE content.

Hu et al. (Hu et al. 2019) noted in their analysis of 53 Brassicaceae genomes that are predominately outcrossing species tend to have a higher Helitron density than predominantly self-pollinating species. For example, the Helitron density of the outcrossing *A. lyrata* (16.7) is higher than that of the self-pollinating *A. thaliana* (5.6) (Table 5). While this difference in reproductive system may explain the strong difference in Helitron density between *C. hispida* var. *hispida* and *C. hispida* var. *grandiflora* as out-crossing rates have not been estimated for these species, in our greenhouse work with the plants both varieties appear to be largely self-incompatible (S. L. Martin personal observation). We note that a similar difference in Helitron density was observed in genomes *Capsella rubella* (14.1) and *Capsella grandiflora* Boiss. (2.1)

(Table 5), but *C. rubella* is self-compatible while *C. grandiflora* is self-incompatible, but went through an extreme bottleneck on formation (Guo et al. 2009). This suggests future work is needed to clarify how reproductive biology and population size alter TE and Helitron abundance.

Here we see reduced TEs abundance and Helitron density in sub-genome 3 from levels observed in *C. hispida* var. *hispida* and in sub-genome 1 compared to *C. neglecta*. While one expectation is that genomic shock and relaxed selection pressure following allopolyploidization could allow increased in TE abundance, the consequences of allopolyploidization for TEs appears to be complex and potentially specific to the both the genomes combined and the type of TEs (Parisod and Senerchia 2012; Vicient and Casacuberta 2017). For example, proliferation of TEs, likely following allopolyploidization, has been observed in allotetraploid *Coffea arabica* L.. Specifically, copia elements have increased the size of sub-genome C compared to that observed in the diploid representative, *C. canephora* L., the progenitor genome (Yu et al. 2011). Similarly, the transposon Sunfish, was observed to be released from repression in the early generations of synthetic autopolyploids of *Arabidopsis thaliana* and *A. arenosa* (L.) Lawalrée as well as their allopolyploid derivative *A. suecica* (Fries.) Norrlin, but the same was not true for other groups of TEs (Madlung et al. 2005). Hu et al. (Hu et al. 2019) observed that Helitron density was lower in the sub-genomes of the allotetraploid *Brassica napus* L. compared to the extant representatives of the parental genomes *B. rapa* L. and *B. oleraceae* L.. This is in line with Sarilar el al.'s (Sarilar et al. 2013) conclusion that early generations of synthetic *Brassica napus*, did not show increased activity in three different groups of TEs, *Athila*-like retrotransposons, the MITE *BraSto*, or CACTA transposons Bot1. As a result, if sub-genomes 1 and 3 have lost TEs since allopolyploidization, these elements may either have been eliminated early in *C. sativa*'s history or lost gradually as a result of greater levels of autogamy. However, further comparisons of the

TE content of the genomes of allopolyploids and representatives of their diploid progenitors from within the Brassicaceae and beyond are needed to understand whether there are any patterns to changes in TEs following allopolyploidization or whether they are esoteric. Alternatively, early changes of TE abundance and distribution could be studied in resynthesized lines of *C. sativa* to determine whether the apparent changes are repeatable.

As domestication has led to decreased genetic variability in *C. sativa* accessions (Manca et al. 2013) other species of the genus might be of value for the improvement of *C. sativa*. Here our results indicate that *C. neglecta* and *C. hispida* var. *hispida* could be useful species to investigate for variation in traits such as seed size, disease resistance and oil profiles. However, the collections available for these species are limited and international efforts to collect and preserve a broader selection of germplasm for the genus should be considered as we look to these species as sources of desirable traits for the improvement of the crop species (Ford-Lloyd et al. 2011).

Data Availability

All assembled genomes are available from the National Center for Biotechnology Information as part of Bioproject PRJNA750147.

Acknowledgements

We would like to thank the ORDC growth facility team for support rearing the plant material, the ORDC Molecular Technology Laboratory for support with sequencing, and the U.S. National Plant Germplasm System and GRIN-Global for conserving and providing the germplasm needed for this research. We thank Dr. S. Wright for sharing the draft genome sequence of *Neslia paniculata* and Julia Mata for help processing convergence data. Funding was

provided by Agriculture and Agri-Food Canada as part of the project "Gene flow, diversity and relationships within the Brassicaceae: Focus on *Camelina*" (J-001029).

Table 1. Genome sizes (1C) as estimated from flow cytometry and the fold coverage (x) for the four genomes sequenced here provided by long read data (PacBio and ONT) and short read data including Illumina paired-end reads (PE) and Illumina mate pairs (MP).

SPECIES	GENOME SIZE (MB)	PACBIO	ONT	ILLUMINA		RAW COVERAGE
				PE	MP	
Camelina neglecta	265	38	60	92	96	286
Camelina hispida var. *hispida*	355	38	53	52	-	143
Camelina hispida var. *grandiflora*	315	-	32	64	-	96
Camelina laxa	275	57	74	47	-	177

Table 2. Statistics on genome assemblies for *C. neglecta*, *C. hispida* var. *hispida*, *C. hispida* var. *grandiflora,* and *C. laxa* generated by QUAST and BUSCO based on the embryophyta_odb10 database before and after scaffolding. The initial values for *C. hispida* var. *grandiflora* and *C. laxa* are for assemblies following assembly by Canu and reduction in the number of contigs by Purge Haplotigs. In all cases, assemblies were polished by Pilon.

	Camelina neglecta		*Camelina hispida* var. *hispida*		*Camelina hispida* var. *grandiflora*		*Camelina laxa*	
ASSEMBLY STRATEGY	CANU	HI-C	CANU	HI-C	CANU & PURGE HAPLOTIGS	NTJOIN	CANU & PURGE HAPLOTIGS	HI-C
contigs (>= 0 bp)	204	6	2,779	7	1,072	7	900	6
contigs (>= 1000 bp)	202	6	2,747	7	1,018	7	870	6
contigs (>= 5000 bp)	180	6	2,591	7	925	7	797	6
contigs (>= 10000	157	6	2,539	7	896	7	785	6
contigs (>= 25000	152	6	2,157	7	839	7	633	6
contigs (>= 50000	139	6	1,600	7	748	7	530	6
Total length (>= 0	210,252,877	194,776,448	446,327,369	283,171,403	247,512,955	247,668,533	223,800,267	199,991,310
Total length (>=	210,251,602	194,776,448	446,302,478	283,171,403	247,473,876	247,668,533	223,776,735	199,991,310
Total length (>=	210,014,359	194,776,448	445,578,173	283,171,403	247,070,138	247,668,533	223,546,568	199,991,310
Total length (>=	209,462,969	194,776,448	417,775,663	283,171,403	242,642,354	247,668,533	217,296,729	199,991,310
Largest contig	19,486,679	48,079,882	6,140,829	45,593,769	3,518,166	41,100,942	4,931,297	39,604,730
Total length	210,252,877	194,726,046	446,327,369	283,171,403	247,512,955	247,668,533	223,800,267	199,991,310
Estimated length	265,000,000	265,000,000	360,000,000	360,000,000	320,000,000	320,000,000	275,000,000	275,000,000
Percent of Expected	79%	74%	124%	79%	77%	77%	81%	73%
N50	13,701,387	30,499,177	566,799	41,586,011	477,640	37,976,760	651,797	31,824,321
NG50	11,493,634	29,279,412	851,053	39,460,631	346,627	30,680,265	477,199	31,147,072
N's per 100 kbp	-	1	-	16	-	32,233	-	20
BUSCO	2,121	2,121	2,121	2,121	2,121	2,121	2,121	2,121
Complete (%)	2,082 (98)	2,084 (98)	2,055 (97)	2,025 (96)	2,063 (97)	1,901 (90)	2,079 (98)	2032 (96)
Complete Single Copy (%)	2,051 (97)	2,055 (97)	1,938 (91)	1,971 (93)	1,863 (88)	1,853 (88)	1,939 (91)	1,978 (93)
Complete Duplicate (%)	31 (1)	29 (1)	117 (6)	54 (3)	200 (9)	48 (2)	140 (7)	54 (3)
Fragmented (%)	14 (<1)	14 (<1)	28 (1)	43 (2)	18 (<1)	65 (3)	14 (<1)	18 (<1)
Missing (%)	25 (1)	23 (1)	38 (2)	53 (3)	40 (2)	155 (7)	28 (1)	71 (3)

Table 3. Distribution of the top three chromosomes with successful alignments of fragments from each of *Arabidopsis lyrata*'s ACKs across the genome of *Camelina sativa* as determined by bowtie2 ordered by number of hits.

ACK	ABK	*Camelina sativa* TOP CHROMOSOME	*Camelina sativa* 2nd CHROMOSOME	*Camelina sativa* 3rd CHROMOSOME
A	AK2	3	14	17
B	AK2	14	3	17
C	AK1	3	14	17
D	AK3	7	16	9
E	AK4	7	16	9
F	AK6	1	15	19
G	AK6	1	15	19
H	AK5	15	1	19
I	AK7	7	16	9
J	AK8	5	6	4
K	AK9	4	6	9
L	AK9	4	6	9
M	AK10	9	6	4
N	AK10	4	6	9
O	AK11	2	13	8
P	AK11	13	2	8
Q	AK12	8	13	20
R	AK12	8	13	20
S	AK13	10	11	12
T	AK14	10	11	12
U	AK14	11	12	10
V	AK15	11	20	18
W	AK16	18	11	2
X	AK16	11	2	18

Table 4. Summary of ACK and ABK groups across *Camelina sativa*'s chromosomes based on bowtie hits showing preservation of ancestral genome structure across ABK regions, but movement of some regions across chromosomes. Chromosomes were assigned to sub-genomes based on summary trees constructed by MrBayes from the ACK fragments grouped by ABK using ASTRAL III. The normalized quartet score, indicating the similarity of trees from individual fragments (out of one), produced by ASTRAL-III is presented. Sub-genome 1 was defined as the genome with closest relationship to *C. neglecta* while sub-genome 3 was defined as that with the closest relationship to the *C. hispida* varieties.

ACK	ABK	SUB-GENOME 1	SUB-GENOME 2	SUB-GENOME 3	FRAGMENTS IN GROUP	NORMALIZED QUARTET SCORE
ABC	AK1/2	14	3	17	262	0.78
DE	AK3/4	7	16	9	98	0.80
FGH	AK5/6	19	1	15	269	0.79
I	AK7	7	16	9	28	0.82
J	AK8	4	6	5	104	0.77
KLMN	AK9/10	4	6	9	115	0.79
OP	AK11	8	13	2	37	0.80
QR	AK12	8	13	20	180	0.79
STU	AK13/14	11	10	12	202	0.78
V	AK15	11	18	20	19	0.80
WX	AK16	11	18	2	130	0.78

Table 5. Results of Helitron (DHH) annotation by EAHelitron. Helitron density is calculated as the number of Helitrons detected in the genome for each 1,000,000 bp.

SPECIES	HELITRONS	HELITRON DENSITY
Camelina neglecta	1877	9.6
Camelina hispida var. *hispida*	5776	20.4
Camelina hispida var. *grandiflora*	1439	5.7
Camelina laxa	937	4.6
Camelina sativa sub-genome 1	1220	6.1
Camelina sativa sub-genome 2	1024	5.8
Camelina sativa sub-genome 3	3395	14.7
Arabidopsis lyrata	3242	16.7
Arabidopsis thaliana	665	5.6
Capsella rubella	1826	14.1
Capsella grandiflora	226	2.1
Capsella bursa-pastoris	2576	9.6
Neslia paniculata	1003	8.9

753 **Table 6.** The number of Helitron insertion sites detected by EAHelitron in *C. sativa*'s sub-genome 3 within regions
754 of synteny among *C. sativa*'s sub-genome 3, *C. hispida* var. *grandiflora,* and *C. hispida* var. *hispida* by chromosome
755 and the number of these sites where Helitron insertion sites were also detected in these regions in EAHelitron's
756 analysis of the genomes of *C. hispida* var. *grandiflora* and *C. hispida* var. *hispida's*.
757

C. sativa CHROMOSOME	*C. sativa* HELITRON COUNT	HELITRONS SHARED WITH		COMMON HELITRONS
		C. hispida var. *hispida*	*C. hispida* var. *grandiflora*	
Cs17	21	10	7	5
Cs09	34	16	15	8
Cs15	22	11	5	2
Cs05	34	23	9	9
Cs12	24	12	5	4
Cs02	16	10	2	1
Cs20	15	6	3	3
Total	166	88	46	32

References

Abbott R, Albach D, Ansell S, Arntzen JW, Baird SJE, Bierne N, Boughman J, Brelsford A, Buerkle CA, Buggs R, et al. 2013. Hybridization and speciation. J Evol Biol. 26(2):229–246. doi:10.1111/j.1420-9101.2012.02599.x.

Al-Shehbaz IA. 2012. A generic and tribal synopsis of the Brassicaceae (Cruciferae). Taxon. 61(5):931–954.

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic Local Alignment Search Tool. J Mol Biol. 215:403–410.

Appels R, Eversole K, Feuillet C, Keller B, Rogers J, Stein N, Pozniak CJ, Choulet F, Distelfeld A, Poland J, et al. 2018. Shifting the limits in wheat research and breeding using a fully annotated reference genome. Science (80-). 361(6403). doi:10.1126/science.aar7191.

Bastolla U, Porto M, Roman HE, Vendruscolo M. 2007. Seqin{R} 1.0-2: a contributed package to the {R} project for statistical computing devoted to biological sequences retrieval and analysis. In: Structural approaches to sequence evolution: Molecules, networks, populations. New York: Springer Verlag. p. 207–232.

Bertioli DJ, Cannon SB, Froenicke L, Huang G, Farmer AD, Cannon EKS, Liu X, Gao D, Clevenger J, Dash S, et al. 2016. The genome sequences of *Arachis duranensis* and *Arachis ipaensis*, the diploid ancestors of cultivated peanut. Nat Genet. 48(4):438–446. doi:10.1038/ng.3517.

Bertioli DJ, Jenkins J, Clevenger J, Dudchenko O, Gao D, Seijo G, Leal-Bertioli SCM, Ren L, Farmer AD, Pandey MK, et al. 2019. The genome sequence of segmental allotetraploid peanut *Arachis hypogaea*. Nat Genet. 51(5):877–884. doi:10.1038/s41588-019-0405-z. http://dx.doi.org/10.1038/s41588-019-0405-z.

Bodenhofer U, Bonatesta E, Horejs-Kainrath C, Hochreiter S. 2015. msa: an R package for multiple sequence alignment. Bioinformatics. 31(24):3997–3999. doi:10.1093/bioinformatics/btv176.

Bogdanowicz D, Giaro K, Wróbel B. 2012. TreeCmp: Comparison of trees in polynomial time. Evol Bioinforma. 2012(8):475–487. doi:10.4137/EBO.S9657.

Bolger AM, Lohse M, Usadel B. 2014. Trimmomatic: A flexible trimmer for Illumina sequence data. Bioinformatics. 30(15):2114–2120. doi:10.1093/bioinformatics/btu170.

Bouckaert R, Vaughan TG, Barido-Sottani J, Duchêne S, Fourment M, Gavryushkina A, Heled J, Jones G, Kühnert D, De Maio N, et al. 2019. BEAST 2.5: An advanced software platform for Bayesian evolutionary analysis. PLoS Comput Biol. 15(4):1–28. doi:10.1371/journal.pcbi.1006650.

Brock JR, Dönmez AA, Beilstein MA, Olsen KM. 2018. Phylogenetics of Camelina Crantz. (Brassicaceae) and insights on the origin of gold-of-pleasure (Camelina sativa). Mol Phylogenet Evol. 127(April):834–842. doi:10.1016/j.ympev.2018.06.031. https://doi.org/10.1016/j.ympev.2018.06.031.

Brock JR, Mandáková T, Lysak MA, Al-Shehbaz IA. 2019. *Camelina neglecta* (Brassicaceae, Camelineae), a new diploid species from Europe. PhytoKeys. 115:51–57.

doi:10.3897/phytokeys.115.31704.

Čalasan AŽ, Seregin AP, Hurka H, Hofford NP, Neuffer B. 2019. The Eurasian steppe belt in time and space: Phylogeny and historical biogeography of the false flax (*Camelina* Crantz, Camelineae, Brassicaceae). Flora Morphol Distrib Funct Ecol Plants. 260(October):151477. doi:10.1016/j.flora.2019.151477. https://doi.org/10.1016/j.flora.2019.151477.

Cao Z, Liu X, Ogilvie H, Yan Z, Nakhleh L. 2019. Practical Aspects of Phylogenetic Network Analysis Using PhyloNet. doi:10.1101/746362.

Chalhoub B, Denoeud F, Liu S, Parkin IAP, Tang H, Wang X, Chiquet J, Belcram H, Tong C, Samans B, et al. 2014. Early allopolyploid evolution in the post-Neolithic *Brassica napus* oilseed genome. Science (80-). 345(6199):950–953. doi:10.1126/science.1253435. http://www.sciencemag.org/cgi/doi/10.1126/science.1253435.

Chaudhary R, Koh CS, Kagale S, Tang L, Wu SW, Lv Z, Mason AS, Sharpe AG, Diederichsen A, Parkin IAP. 2020. Assessing diversity in the camelina genus provides insights into the genome structure of *Camelina sativa*. G3 Genes, Genomes, Genet. 10(4):1297–1308. doi:10.1534/g3.119.400957.

Chen MY, Liang D, Zhang P. 2017. Phylogenomic resolution of the phylogeny of laurasiatherian mammals: Exploring phylogenetic signals within coding and noncoding sequences. Genome Biol Evol. 9(8):1998–2012. doi:10.1093/gbe/evx147.

Cheng F, Mandáková T, Wu J, Xie Q, Lysak M a., Wang X, Mandakova T. 2013. Deciphering the diploid ancestral genome of the mesohexaploid *Brassica rapa*. Plant Cell. 25(5):1541–1554. doi:10.1105/tpc.113.110486. http://www.plantcell.org/cgi/doi/10.1105/tpc.113.110486%5Cnhttp://www.ncbi.nlm.nih.gov/pubmed/23653472.

Coombe L, Nikolić V, Chu J, Birol I, Warren RL. 2020. ntJoin: Fast and lightweight assembly-guided scaffolding using minimizer graphs. Bioinformatics. 36(12):3885–3887. doi:10.1093/bioinformatics/btaa253.

Crowl AA, Manos PS, McVay JD, Lemmon AR, Lemmon EM, Hipp AL. 2020. Uncovering the genomic signature of ancient introgression between white oak lineages (*Quercus*). New Phytol. 226(4):1158–1170. doi:10.1111/nph.15842.

Emms DM, Kelly S. 2017. STRIDE: Species tree root inference from gene duplication events. Mol Biol Evol. 34(12):3267–3278. doi:10.1093/molbev/msx259.

Emms DM, Kelly S. 2018. STAG: Species Tree Inference from All Genes. bioRxiv.:267914. doi:10.1101/267914. https://www.biorxiv.org/content/10.1101/267914v1.full.

Emms DM, Kelly S. 2019. OrthoFinder: Phylogenetic orthology inference for comparative genomics. Genome Biol. 20(1):1–14. doi:10.1186/s13059-019-1832-y.

Ford-Lloyd B V, Schmidt M, Armstrong SJ, Barazani O, Engels J, Hadas R, Hammer K, Kell SP, Kang D, Khoshbakht K, et al. 2011. Crop wild relatives - Undervalued, underutilized and under threat? Bioscience. 61(7):559–565. doi:10.1525/bio.2011.61.7.10.

Greiner S, Lehwark P, Bock R. 2019. OrganellarGenomeDRAW (OGDRAW) version 1.3.1: Expanded toolkit for the graphical visualization of organellar genomes. Nucleic Acids Res. 47(W1):W59–W64. doi:10.1093/nar/gkz238.

Gu Z, Gu L, Eils R, Schlesner M, Brors B. 2014. Circlize implements and enhances circular visualization in R. Bioinformatics. 30(19):2811–2812. doi:10.1093/bioinformatics/btu393.

Guo YL, Bechsgaard JS, Slotte T, Neuffer B, Lascoux M, Weigel D, Schierup MH. 2009. Recent speciation of *Capsella rubella* from *Capsella grandiflora*, associated with loss of self-incompatibility and an extreme bottleneck. Proc Natl Acad Sci U S A. 106(13):5246–5251. doi:10.1073/pnas.0808012106.

Heibl C. 2008. PHYLOCH: R language tree plotting tools and interfaces to diverse phylogenetic software packages. http://www.christophheibl.de/Rpackages.html.

Hu K, Xu K, Wen J, Yi B, Shen J, Ma C, Fu T, Ouyang Y, Tu J. 2019. Helitron distribution in Brassicaceae and whole Genome Helitron density as a character for distinguishing plant species. BMC Bioinformatics. 20(1):1–20. doi:10.1186/s12859-019-2945-8.

Husband BC, Baldwin SJ, Suda J. 2013. The incidence of polyploid in natural plant populations: major patterns and evolutionary processses. In: Leitch IJ, Al. E, editors. Plant Genome Diversity. Springer-Verlag. p. 255–276. http://link.springer.com/10.1007/978-3-7091-1160-4.

Jiao Y, Peluso P, Shi J, Liang T, Stitzer MC, Wang B, Campbell MS, Stein JC, Wei X, Chin CS, et al. 2017. Improved maize reference genome with single-molecule technologies. Nature. 546(7659):524–527. doi:10.1038/nature22971. http://dx.doi.org/10.1038/nature22971.

Kagale S, Koh C, Nixon J, Bollina V, Clarke WE, Tuteja R, Spillane C, Robinson SJ, Links MG, Clarke C, et al. 2014. The emerging biofuel crop *Camelina sativa* retains a highly undifferentiated hexaploid genome structure. Nat Commun. 5:3706–3717. doi:10.1038/ncomms4706. [accessed 2014 Aug 13]. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=4015329&tool=pmcentrez&rendertype=abstract.

Kapitonov V V., Jurka J. 2001. Rolling-circle transposons in eukaryotes. Proc Natl Acad Sci U S A. 98(15):8714–8719. doi:10.1073/pnas.151269298.

Koch MA, Kiefer M. 2005. Genome evolution among cruciferous plants: a lecture from the comparison of the genetic maps to three diploid species - Capsella rubella, Arabidopsis lyrata subsp. petraea, and A. thaliana. Am J Bot. 92(4):761–767.

Koenig D, Weigel D. 2015. Beyond the thale: comparative genomics and genetics of Arabidopsis relatives. Nat Rev Genet. 16(5):285–298. doi:10.1038/nrg3883. http://www.nature.com/doifinder/10.1038/nrg3883.

Komsta L. 2011. outliers: Tests for outliers.

Koren S, Walenz BP, Berlin K, Miller JR, Phillippy AM. 2016. Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. bioRxiv. doi:10.1101/071282.

Langmead B, Salzberg SL. 2012. Fast gapped-read alignment with Bowtie 2. Nat Methods. 9(4):357–359. doi:10.1038/nmeth.1923.Fast.

Latta RG, Bekele WA, Wight CP, Tinker NA. 2019. Comparative linkage mapping of diploid, tetraploid, and hexaploid *Avena* species suggests extensive chromosome rearrangement in ancestral diploids. Sci Rep. 9(1):1–12. doi:10.1038/s41598-019-48639-7. http://dx.doi.org/10.1038/s41598-019-48639-7.

Lawrence M, Huber W, Pagès H, Aboyoun P, Carlson M, Gentleman R, Morgan M, Carey V. 2013. Software for computing and annotating genomic ranges. PLoS Comput Biol. 9. doi:10.1371/journal.pcbi.1003118.

Lemon J. 2006. Plotrix: a package in the red light district of R. R-News. 6(4):8–12.

Levin DA. 1983. Polyploidy and novelty in flowering plants. Am Nat. 122(1):1–25.

Li H. 2013. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. 00(00):1–3. http://arxiv.org/abs/1303.3997.

Lujan Toro BE. 2017. Genome assembly of Camelina microcarpa Andrz. Ex DC, A step towards understanding genome evolution in Camelina. Carleton University.

Lysak MA, Mandakova T, Schranz ME. 2016. Comparative paleogenomics of crucifers: Ancestral genomic blocks revisited. Curr Opin Plant Biol. 30:108–115. doi:10.1016/j.pbi.2016.02.001.

Madlung A, Tyagi AP, Watson B, Jiang H, Kagochi T, Doerge RW, Martienssen R, Comai L. 2005. Genomic changes in synthetic *Arabidopsis* polyploids. Plant J. 41(2):221–230. doi:10.1111/j.1365-313X.2004.02297.x.

Manca A, Pecchia P, Mapelli S, Masella P, Galasso I. 2013. Evaluation of genetic diversity in a *Camelina sativa* (L.) Crantz collection using microsatellite markers and biochemical traits. Genet Resour Crop Evol. 60(4):1223–1236. doi:10.1007/s10722-012-9913-8.

Marcussen T, Sandve SR, Heier L, Pfeifer M, Kugler KG, Zhan B, Spannagl M, Pfeifer M, Jakobsen KS, Wulff BBH, et al. 2014. Ancient hybridizations among the ancestral genomes of bread wheat. Science. 345(6194):1250092. doi:10.1126/science.1251788. http://www.sciencemag.org/content/345/6194/1250092.abstract.

Martin SH, Davey JW, Jiggins CD. 2015. Evaluating the use of ABBA-BABA statistics to locate introgressed loci. Mol Biol Evol. 32(1):244–257. doi:10.1093/molbev/msu269.

Martin SH, Jiggins CD. 2017. Interpreting the genomic landscape of introgression. Curr Opin Genet Dev. 47:69–74. doi:10.1016/j.gde.2017.08.007.

Martin SL, Smith T, James T, Shalabi F, Kron P, Sauder CA. 2017. An update to the Canadian range and abundance of *Camelina* spp. (Brassicaceae) east of the Rocky Mountains. Botany. 95:405–417.

Michael TP, VanBuren R. 2015. Progress, challenges and the future of crop genomes. Curr Opin Plant Biol. 24:71–81. doi:10.1016/j.pbi.2015.02.002. http://dx.doi.org/10.1016/j.pbi.2015.02.002.

Morgan M, Pagès H, Obenchain V, Hayden N. 2020. Rsamtools: Binary alignment (BAM), FASTA, variant call (BCF), and tabix file import. http://bioconductor.org/packages/Rsamtools.

Murat F, Louis A, Maumus F, Armero A, Cooke R, Quesneville H, Crollius HR, Salse J. 2015. Understanding Brassicaceae evolution through ancestral genome reconstruction. Genome Biol. 16(1):262. doi:10.1186/s13059-015-0814-y. http://genomebiology.com/2015/16/1/262.

Nikolov LA, Shushkov P, Nevado B, Gan X, Al-Shehbaz IA, Filatov D, Bailey CD, Tsiantis M. 2019. Resolving the backbone of the Brassicaceae phylogeny for investigating trait diversity. New Phytol. 222(3):1638–1651. doi:10.1111/nph.15732.

Oddes S, Zelig A, Kaplan N. 2018. Three invariant Hi-C interaction patterns: Applications to

genome assembly. bioRxiv. 142:89–99. doi:10.1016/j.ymeth.2018.04.013.

Ogilvie HA, Bouckaert RR, Drummond AJ. 2017. StarBEAST2 brings faster species tree inference and accurate estimates of substitution rates. Mol Biol Evol. 34(8):2101–2114. doi:10.1093/molbev/msx126.

Otto SP, Whitton J. 2002. Polyploid Incidence and Evolution. Annu Rev Genet. 34(1):401–437. doi:10.1146/annurev.genet.34.1.401.

Ou S, Su W, Liao Y, Chougule K, Agda JRA, Hellinga AJ, Lugo CSB, Elliott TA, Ware D, Peterson T, et al. 2019. Benchmarking transposable element annotation methods for creation of a streamlined, comprehensive pipeline. Genome Biol. 20(1):1–18. doi:10.1186/s13059-019-1905-y.

Pagès H, Aboyoun P, Gentleman R, DebRoy S. 2020. Biostrings: Efficient manipulation of biological. https://bioconductor.org/packages/release/bioc/html/Biostrings.html.

Paradis E, Schliep K. 2019. ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. Bioinformatics. 35:526–528.

Parisod C, Senerchia N. 2012. Responses of Transposable Elements to Polyploidy. In: Grandbastien M-A, Casacuberta JM, editors. Plant Transposable Elements: Impact on Genome Structure and Function - Topics in Current Genetics. Vol. 24. Heidelberg: Springer. p. 147–168.

Pires CJ, Zhao J, Schranz ME, Leon EJ, Quijada PA, Lukens LN, Osborn TC. 2004. Flowering time divergence and genomic rearrangements in resynthesized *Brassica* polyploids (Brassicaceae). Biol J Linn Soc. 82(4):675–688. [accessed 2011 Apr 29]. http://www3.botany.ubc.ca/rieseberglab/plantevol/Piresetal2004.pdf.

Rambaut A. 2018. FigTree. tree.bio.ed.ac.uk/software/figtree/.

Rambaut A, Drummond AJ, Xie D, Baele G, Suchard MA. 2018. Posterior summarization in Bayesian phylogenetics using Tracer 1.7. Syst Biol. 67(5):901–904. doi:10.1093/sysbio/syy032.

Ramos-Madrigal J, Smith BD, Moreno-Mayar JV, Gopalakrishnan S, Ross-Ibarra J, Gilbert MTP, Wales N. 2016. Genome Sequence of a 5,310-Year-Old Maize Cob Provides Insights into the Early Stages of Maize Domestication. Curr Biol. 26(23):3195–3201. doi:10.1016/j.cub.2016.09.036.

Revell LJ. 2012. phytools: Phylogenetic tools for comparative biology (and other things). Methods Ecol Evol. 3:217–223. doi:10.1111/j.2041- 210X.2011.00169.x. https://cran.r-project.org/web/packages/phytools/index.html.

Rieseberg LH. 1997. Hybrid origins of plant species. Annu Rev Ecol Syst. 28(1):359–389. doi:10.1146/annurev.ecolsys.28.1.359. http://www.annualreviews.org/doi/10.1146/annurev.ecolsys.28.1.359.

Rieseberg LH. 2001. Chromosomal rearrangements and speciation. Trends Ecol Evol. 16(7):351–358. http://www.ncbi.nlm.nih.gov/pubmed/11403867.

Rieseberg LH, Willis JH. 2007. Plant speciation. Science. 317(5840):910–4. doi:10.1126/science.1137729. http://science.sciencemag.org/content/317/5840/910.abstract.

Roach MJ, Schmidt SA, Borneman AR. 2018. Purge Haplotigs: Allelic contig reassignment for third-gen diploid genome assemblies. BMC Bioinformatics. 19(1):1–10. doi:10.1186/s12859-018-2485-7.

Ronquist F, Teslenko M, Van Der Mark P, Ayres DL, Darling A, Höhna S, Larget B, Liu L, Suchard MA, Huelsenbeck JP. 2012. Mrbayes 3.2: Efficient bayesian phylogenetic inference and model choice across a large model space. Syst Biol. 61(3):539–542. doi:10.1093/sysbio/sys029.

Sarilar V, Palacios PM, Rousselet A, Ridel C, Falque M, Eber F, Chèvre AM, Joets J, Brabant P, Alix K. 2013. Allopolyploidy has a moderate impact on restructuring at three contrasting transposable element insertion sites in resynthesized Brassica napus allotetraploids. New Phytol. 198(2):593–604. doi:10.1111/nph.12156.

Sato S, Tabata S, Hirakawa H, Asamizu E, Shirasawa K, Isobe S, Kaneko T, Nakamura Y, Shibata D, Aoki K, et al. 2012. The tomato genome sequence provides insights into fleshy fruit evolution. Nature. 485(7400):635–641. doi:10.1038/nature11119.

Schliep K, Jombart T, Kamvar ZN, Archer E, Harris R. 2020. apex: Phylogenetic Methods for Multiple Gene Data. https://cran.r-project.org/package=apex%7D.

Schliep KP. 2011. phangorn□: phylogenetic analysis in R. Bioinformatics. 27(4):592–593. doi:10.1093/bioinformatics/btq706.

Schmutz J, Cannon SB, Schlueter J, Ma J, Mitros T, Nelson W, Hyten DL, Song Q, Thelen JJ, Cheng J, et al. 2010. Genome sequence of the palaeopolyploid soybean. Nature. 463(7278):178–183. doi:10.1038/nature08670.

Schnable JC, Springer NM, Freeling M. 2011. Differentiation of the maize subgenomes by genome dominance and both ancient and ongoing gene loss. Proc Natl Acad Sci. 108(10):4069–4074. doi:10.1073/pnas.1101368108. http://www.pnas.org/cgi/doi/10.1073/pnas.1101368108.

Schnable PS, Pasternak S, Liang C, Zhang J, Fulton L, Graves TA, Minx P, Reily AD, Courtney L, Kruchowski SS, et al. 2009. The B73 Maize Genome: Complexity, Diversity, and Dynamics. Science (80-). 326(June):1112–1115.

Schranz ME, Lysak M a., Mitchell-Olds T. 2006. The ABC's of comparative genomics in the Brassicaceae: building blocks of crucifer genomes. Trends Plant Sci. 11(11):535–542. doi:10.1016/j.tplants.2006.09.002.

Schulz A. 2021. pBrackets: Plot Brackets. https://cran.r-project.org/package=pBrackets.

Simão FA, Waterhouse RM, Ioannidis P, Kriventseva E V, Zdobnov EM. 2015. BUSCO: Assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics. 31(19):3210–3212. doi:10.1093/bioinformatics/btv351.

Soltis D, Soltis P. 1999. Polyploidy: recurrent formation and genome evolution. Trends Ecol Evol. 14(9):348–352. http://www.ncbi.nlm.nih.gov/pubmed/10441308.

Soltis PS, Soltis DE. 2009. The role of hybridization in plant speciation. Annu Rev Plant Biol. 60:561–88. doi:10.1146/annurev.arplant.043008.092039. http://www.ncbi.nlm.nih.gov/pubmed/19575590.

Stanke M, Schöffmann O, Morgenstern B, Waack S. 2006. Gene prediction in eukaryotes with a generalized hidden Markov model that uses hints from external sources. BMC Bioinformatics. 7:1–11. doi:10.1186/1471-2105-7-62.

Stebbins GL. 1968. The significance of hybridization for plant taxonomy and evolution. Taxon. 18(1):26–35.

Struck TH. 2014. Trespex-detection of misleading signal in phylogenetic reconstructions based

on tree information. Evol Bioinforma. 10:51–67. doi:10.4137/EBo.s14239.

Suchard MA, Lemey P, Baele G, Ayres DL, Drummond AJ, Rambaut A. 2018. Bayesian phylogenetic and phylodynamic data integration using BEAST 1.10. Virus Evol. 4(1):1–5. doi:10.1093/ve/vey016.

Than C, Ruths D, Nakhleh L. 2008. PhyloNet: A software package for analyzing and reconstructing reticulate evolutionary relationships. BMC Bioinformatics. 9:1–16. doi:10.1186/1471-2105-9-322.

The International Wheat Genome Sequencing Consortium. 2014. A chromosome-based draft sequence of the hexaploid bread wheat (*Triticum aestivum*) genome. Science. 345(6194):1251788. doi:10.1126/science.1251788. http://www.sciencemag.org/content/345/6194/1250092.abstract.

Tillich M, Lehwark P, Pellizzer T, Ulbricht-Jones ES, Fischer A, Bock R, Greiner S. 2017. GeSeq - Versatile and accurate annotation of organelle genomes. Nucleic Acids Res. 45(W1):W6–W11. doi:10.1093/nar/gkx391.

Vicient CM, Casacuberta JM. 2017. Impact of transposable elements on polyploid plant genomes. Ann Bot. 120(2):195–207. doi:10.1093/aob/mcx078.

Walker BJ, Abeel T, Shea T, Priest M, Abouelliel A, Sakthikumar S, Cuomo CA, Zeng Q, Wortman J, Young SK, et al. 2014. Pilon: An integrated tool for comprehensive microbial variant detection and genome assembly improvement. PLoS One. 9(11):e112963. doi:10.1371/journal.pone.0112963.

Wang L-G, Lam TT-Y, Xu S, Dai Z, Zhou L, Feng T, Guo P, Dunn CW, Jones BR, Bradley T, et al. 2020. treeio: an R package for phylogenetic tree input and output with richly annotated and associated data. Mol Biol Evol. 37(2):599–603. doi:10.1093/molbev/msz240.

Warren DL, Geneva AJ, Lanfear R, Rosenberg M. 2017. RWTY (R We There Yet): An R package for examining convergence of Bayesian phylogenetic analyses. Mol Biol Evol. 34(4):1016–1020. doi:10.1093/molbev/msw279.

Wen D, Yu Y, Zhu J, Nakhleh L. 2018. Inferring phylogenetic networks using PhyloNet. Syst Biol. 67(4):735–740. doi:10.1093/sysbio/syy015.

de Wet JMJ. 1971. Polyploidy and evolution in plants. Taxon. 20(1):29–35.

Wick RR, Judd LM, Holt KE. 2019. Performance of neural network basecalling tools for Oxford Nanopore sequencing. Genome Biol. 20(1):1–10. doi:10.1186/s13059-019-1727-y.

Wickham H. 2011. The split-apply-combine strategy for data analysis. J Stat Softw. 20(1):1–29. http://www.jstatsoft.org/v40/i01/.

Wickham H. 2019. stringr: simple, consistent wrappers for common string operations. https://cran.r-project.org/package=stringr.

Workman R, Fedak R, Kilburn D, Hao S, Liu K, Timp W. 2018. High molecular weight DNA extraction from recalcitrant plant species for third generation sequencing. Protoc Exch. version 1:1–15. doi:10.1038/protex.2018.059.

Wright K. 2021. pals: Color Palettes, Colormaps, and Tools to Evaluate Them.

Yu Q, Guyot R, De Kochko A, Byers A, Navajas-Pérez R, Langston BJ, Dubreuil-Tranchant C, Paterson AH, Poncet V, Nagai C, et al. 2011. Micro-collinearity and genome evolution in the

vicinity of an ethylene receptor gene of cultivated diploid and allotetraploid coffee species (*Coffea*). Plant J. 67(2):305–317. doi:10.1111/j.1365-313X.2011.04590.x.

Zhang C, Rabiee M, Sayyari E, Mirarab S. 2018. ASTRAL-III: Polynomial time species tree reconstruction from partially resolved gene trees. BMC Bioinformatics. 19(Suppl 6):15–30. doi:10.1186/s12859-018-2129-y. http://dx.doi.org/10.1186/s12859-018-2129-y.