Plant Biotechnology Journal Open Access

Plant Biotechnology Journal (2023) **21**, pp. 521–535 doi: 10.1111/pbi.13968

Sequencing of *Camelina neglecta*, a diploid progenitor of the hexaploid oilseed *Camelina sativa*

Raju Chaudhary[1,2], Chu Shin Koh[2], Sampath Perumal[2], Lingling Jin[3], Erin E. Higgins[1], Sateesh Kagale[4], Mark A. Smith[1], Andrew G. Sharpe[2] and Isobel A. P. Parkin[1,*]

[1]*Agriculture and Agri-Food Canada, Saskatoon, SK, Canada*
[2]*Global Institute for Food Security, Saskatoon, SK, Canada*
[3]*Department of Computer Science, University of Saskatchewan, Saskatoon, SK, Canada*
[4]*National Research Council Canada, Saskatoon, SK, Canada*

Received 22 August 2022;
revised 26 October 2022;
accepted 9 November 2022.
**Correspondence* (Tel 1-306-385-9434; email isobel.parkin@agr.gc.ca)

Keywords: *Camelina*, progenitor, genome assembly, whole genome sequencing.

Summary

Camelina neglecta is a diploid species from the genus *Camelina*, which includes the versatile oilseed *Camelina sativa*. These species are closely related to *Arabidopsis thaliana* and the economically important *Brassica* crop species, making this genus a useful platform to dissect traits of agronomic importance while providing a tool to study the evolution of polyploids. A highly contiguous chromosome-level genome sequence of *C. neglecta* with an N50 size of 29.1 Mb was generated utilizing Pacific Biosciences (PacBio, Menlo Park, CA) long-read sequencing followed by chromosome conformation phasing. Comparison of the genome with that of *C. sativa* shows remarkable coincidence with subgenome 1 of the hexaploid, with only one major chromosomal rearrangement separating the two. Synonymous substitution rate analysis of the predicted 34 061 genes suggested subgenome 1 of *C. sativa* directly descended from *C. neglecta* around 1.2 mya. Higher functional divergence of genes in the hexaploid as evidenced by the greater number of unique orthogroups, and differential composition of resistant gene analogs, might suggest an immediate adaptation strategy after genome merger. The absence of genome bias in gene fractionation among the subgenomes of *C. sativa* in comparison with *C. neglecta*, and the complete lack of fractionation of meiosis-specific genes attests to the neopolyploid status of *C. sativa*. The assembled genome will provide a tool to further study genome evolution processes in the *Camelina* genus and potentially allow for the identification and exploitation of novel variation for *Camelina* crop improvement.

Introduction

Camelina is a genus of the Brassicaceae family, which contains a number of important vegetable and oilseed crops, and is closely related to the plant model *Arabidopsis thaliana* (Al-Shehbaz *et al*., 2006). *Camelina sativa* (L.) Crantz is an oilseed crop with a unique fatty acid composition compared with other vegetable oils that make it a desirable platform for multiple applications in the food, feed, and fuel markets (Zubr, 1997). Like most angiosperms, many Brassicaceae species have been identified as polyploids, ranging from paleopolyploids with remnants of older hybridization events to relatively young neopolyploids such as *Camelina sativa*, where the progenitor species have yet to be confirmed. The whole genome sequencing of *C. sativa* has confirmed the hexaploid nature of its genome (Kagale *et al*., 2014) and facilitated a number of studies deciphering the relationships among *Camelina* species (Brock *et al*., 2018); however, without knowledge of the progenitor genomes, our understanding of the formation and evolution of *C. sativa* is innately limited. There are several lower ploidy species of *Camelina* viz. *C. neglecta* (Brock *et al*., 2019), *C. hispida* (Boiss.) Hedge, *C. laxa* C.A. Mey., *C. microcarpa* Andrz. ex DC., and *C. rumelica* Velen., etc. (Martin *et al*., 2017). *Camelina neglecta* is a diploid six-chromosome species, recently re-classified from its original taxonomic identity as *C. microcarpa* (Brock *et al*., 2019), but notably recent work has shown that *C. neglecta*, has a higher affinity with the first subgenome (or SG1) of the hexaploid *C. sativa* (Chaudhary *et al*., 2020; Mandáková *et al*., 2019).

Allopolyploidy is an important source of variation in plant families, where the merger of two related genomes creates a novel genome structure with potentially greater fitness due to the inherent heterosis (Cheng *et al*., 2014). However, upon genome hybridization, genome fractionation and subgenome dominance have been reported in most young or neopolyploid species (Kagale *et al*., 2016; Schnable *et al*., 2011). These phenomena are thought to play a role in the adaptation of such species (Comai, 2005). Understanding the role and extent of these evolutionary processes is important to identify the changes that differentiate a fertile polyploid species from its progenitors. This knowledge can also facilitate the development of new synthetic lines and could be exploited to capture additional variation for traits of interest by diversification of individual subgenomes (Abel *et al*., 2005; Rosyara *et al*., 2019; Wei *et al*., 2016). All such analyses are predicated on the availability of the progenitor species, which allows the polyploid genome to be partitioned into subgenomes and facilitates the identification of fractionation and structural rearrangements that led to the adaptation of the new polyploid. Available *C. sativa* germplasm has very limited genetic diversity (Gehringer *et al*., 2006; Luo *et al*., 2019; Singh *et al*., 2015; Vollmann *et al*., 2005), which suggests a bottleneck created from a small number of hybridization events in its evolutionary trajectory, thus further knowledge of the progenitors for this crop may be exploited for genetic improvement.

Hybridization between *C. neglecta* and *C. sativa* has had limited success (Martin *et al*., 2019), which notwithstanding the chromosome imbalance indicated fundamental differences in the gene pool for these species; however, chromosome painting, syntenic analysis, and sequence alignment of short reads from *C. neglecta* with the reference *C. sativa* has shed some light on the relationship between these species (Chaudhary *et al*., 2020; Mandáková *et al*., 2019). Whole genome sequence-based comparative analysis remains the most conclusive and informative method to infer the homology between chromosomes of different species and to identify structural changes during the evolution of a polyploid. The development of single-molecule sequencing techniques such as Pacific Biosciences and scaffolding techniques such as chromosome conformation capture (Belton *et al*., 2012; Ghurye *et al*., 2017) have eased the assembly process for plant genomes, which tend to be highly recursive with large repetitive regions (Girollet *et al*., 2019; Song *et al*., 2020). Here, we report a high-quality genome of *C. neglecta* (Accession: PI650135), assembled to the chromosome level using long-read sequencing technology and chromosome conformation capture. This inferred progenitor of *C. sativa* was compared with the genome of its hexaploid relative, confirming the first subgenome of *C. sativa* has directly descended from *C. neglecta*, with minimal changes in the gene complement of the neopolyploid *C. sativa* post-polyploidization, whereby, a minor reshuffling of conserved genomic blocks was observed. Overall, the results will act as a repository for further genetic studies, tool development, and potential trait improvement in modern *Camelina sativa*.

Results

De novo assembly of the *Camelina neglecta* genome

Previous flow cytometry indicated the genome size of accession PI650135 was approximately one-third of that of the *C. sativa* reference genome (Martin *et al*., 2017), which was corroborated with k-mer-based genome size estimation (Marçais and Kingsford, 2011) that indicated a genome size of approximately 201.62 Mb (Figure S1). The *de novo* assembly was generated with ~2.4 M PacBio reads (72.9× coverage), which yielded 131 contigs with an N50 length of 7.9 Mb. The contigs were arranged into the expected six chromosomes with an N50 of 29.1 Mb using Chicago® in vitro proximity ligation followed by Dovetail™ HiC phasing technology (Cairns *et al*., 2016; Figure S2; Table 1). The assembled genome sequence of *C. neglecta* covered 94.58% of the estimated genome size. Assembly quality was assessed by mapping 99.33% of available Illumina reads (>146 × coverage) onto the final assembly using BWA version 0.7.17 (Li and Durbin, 2009). Qualimap v.2.2.1 (García-Alcalde *et al*., 2012) suggested a general error rate of 0.81% with the majority being homopolymer indels (66.33%; Figure 1, Table S1). The level of heterozygosity was also low (~0.2%), as shown by statistical analyses of the k-mer profile in GenomeScope v.2.0. Smudgeplot v0.2.3 (Ranallo-Benavidez *et al*., 2020) was further used to evaluate heterozygous k-mer pairs providing confirmation of the ploidy (Figure S3). Assessment of the expected gene content was carried out using the *brassicales_odb10* ($n = 4596$) Benchmarking Universal Single-Copy Orthogues (BUSCO) data set (Simão *et al*., 2015), giving a 97.6% score, with 95.1% complete and single-copy BUSCO genes, and 2.5% duplicated genes (Figure S4). In concert, these results suggested the high quality of the genome with regard to contiguity and content, which was comparable with that of the *C. sativa* genome (Kagale *et al*., 2014). Another assembly of *C. neglecta* was recently made available, along with additional *Camelina* diploid species, although no gene annotation was provided, nucmer comparison of the two assemblies indicated a strong correspondence, apart from one inversion on chromosome 2 of approximately 4 Mb (Martin *et al*., 2021; Figure S5). The HiC data and subsequent synteny analyses presented here showed no apparent anomalies, so no changes were made to the assembly.

Repeat analysis and gene model prediction

Repeat analysis identified 42.22% repeat elements (REs) in the whole genome assembly, with a total of 1162 full-length long terminal repeat (LTR) retrotransposon elements (Figure 1, Table S2–S4). The Gypsy and Copia retrotransposons and the Helitron DNA transposons were most prevalent, accounting for more than 29% of the genome (Table S2). A re-analysis of the hexaploid *C. sativa* genome identified a lower level of REs across the whole genome (40.41%; Table S2). In total, 1700 full-length LTR elements could be identified in *C. sativa* (hexaploid; Table S5), this is comparatively less than for *C. neglecta* (diploid), yet expected due to the short-read sequencing approach used for the *C. sativa* assembly. The higher percentage of full-length LTR sequences and the associated LTR assembly index (LAI) of *C. neglecta* (21.3%; LAI of 21.8) over *C. sativa* (8.5%; LAI of 6.98) support the better assembly of repeats in the long-read based *C. neglecta* assembly (Ou *et al*., 2018, Figure S6). In both *Camelina* species, three classes dominated the composition of transposons, the DNA transposon Helitron (12.08%–14.27%), and two retrotransposons, Gypsy (10.10%–12.26%), and Copia (4.57%–4.68%). The amount and composition of REs were similar when comparing *C. neglecta* with each of the three subgenomes (SG) of *C. sativa*; although SG2 contained a lower percentage of REs (35.28%), while SG3 contained higher

Table 1 Genome assembly statistics of the *C. neglecta* genome

Assembly	Value
Estimated genome size ($k = 21$) (Mb)	201.6
Assembly size (Mb)	192.5
Genome coverage (%)	95.5
No. of chromosomes	6
No. of sequences	131
Longest scaffold (Mb)	47.8
Scaffold N50 (Mb)	29.06
BUSCO (percentage complete)	97.6
No. of protein-coding genes	34 061
Mean gene length (bp)	2007
Mean cds length (bp)	1091
Mean exon length (bp)	265
Repetitive elements	
Helitron	44 789 (12.08%)
Copia	7752 (4.68%)
Gypsy	18 154 (12.26%)
Others	38 332 (13.20%)
Total	109 027 (42.22%)

,

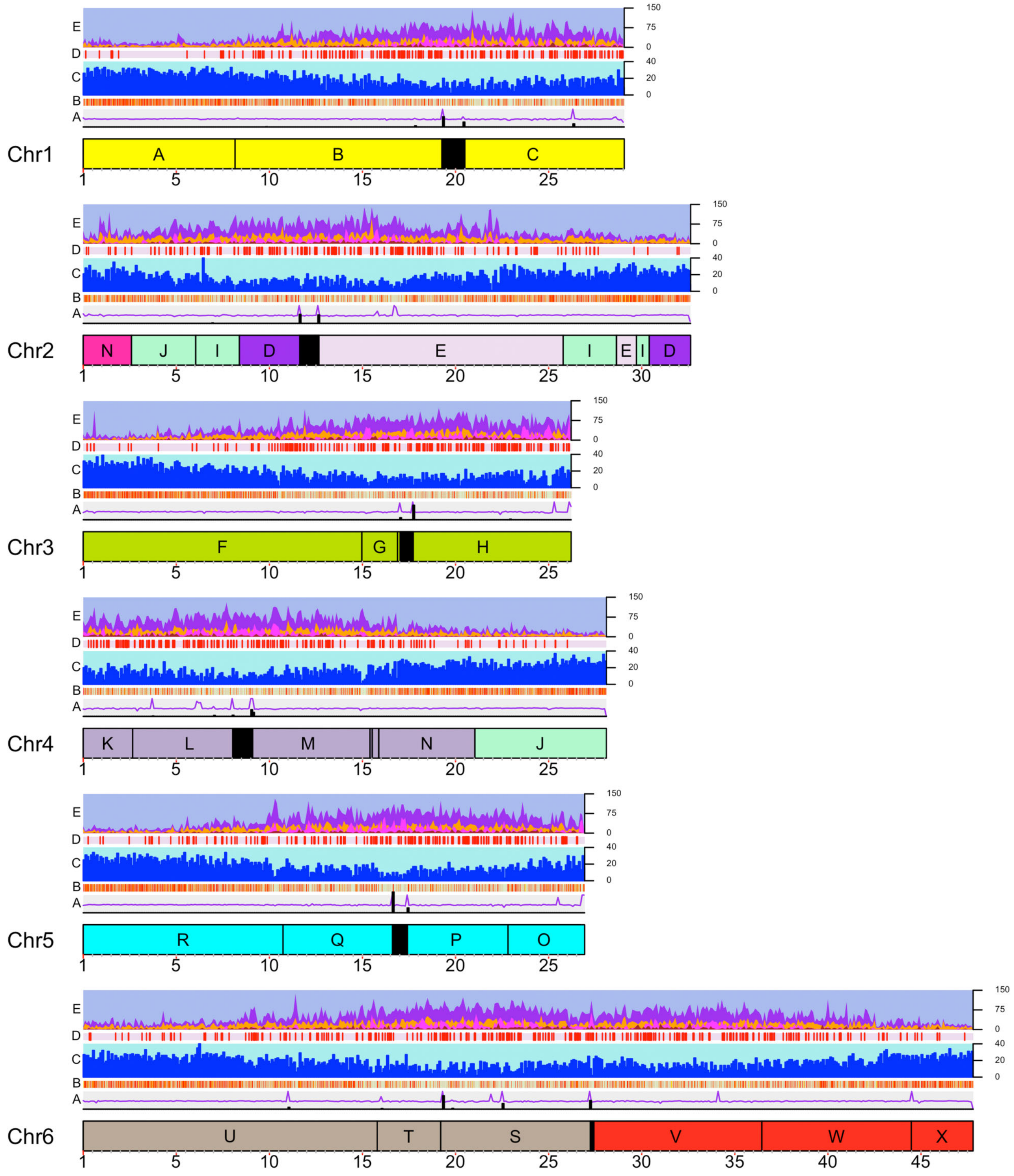


Figure 1 The genome of *Camelina neglecta*. Ideograms are shown for each chromosome with the ancestral karyotype genomic blocks (A–X) and probable centromeric locations (black block) indicated. Track A represents the density of normalized Illumina short reads mapped to the assembled genome (purple) and distribution of centromeric repeats (black). Track B is a heatmap showing the expression of genes in leaf tissue (log10[FPKM]). Track C shows the distribution of genes across the genome (blue). Track D shows the distribution of full-length long-terminal repeats across the assembled genome. Track E is the distribution of all repeat elements (purple), helitrons (orange), gypsy (magenta), and copia elements (brown) across the assembled genome.

amounts of REs (45.27%), largely contributed by the Gypsy (11.71%) and Helitron (16.41%) classes. The full-length LTR retrotransposons (FL-LTRs) were analysed to identify any clear expansion peaks, one major event was aged at approximately 1.08 million years ago (mya) in *C. neglecta*, with about 90% (1045 FL-LTRs) of the elements aged at <2 mya (Figure S7), these data coincide with the proliferation of LTR for other *Brassica* species (Song *et al*., 2021). In addition, family-level analysis of FL-LTRs suggests that Copia-Ale and Gypsy-Athila families were present in high copies in both *C. neglecta* and *C. sativa* genomes (Figure 2, Table S3). Among the three subgenomes of *C. sativa* SG3 possessed the highest number of FL-LTRs with 766 copies followed by 511 in SG1 and 384 in SG2. Although the analysis is limited by the different sequencing technologies adopted for *C. neglecta* (PacBio long-read) and *C. sativa* (Illumina short-read), this proliferation appears to be recent in comparison to the major proliferation event for full-length LTR in the first and second subgenomes of *C. sativa* (Figure 2, Table S5) and would align with the dominant expression pattern of *C. sativa* SG3 (Kagale *et al*., 2016).

Centromeric repeat analysis was performed using an LTR finder (Xu and Wang, 2007), TRF finder (Benson, 1999), and Repeat explorer (Novak *et al*., 2013), which identified a higher abundance of two types of centromeric tandem repeats in the *C. neglecta* genome, one specific to *C. neglecta* (CentCn) and one common with *C. sativa* (CentCs) (Figure S8). CentCs1 was found in both genomes, whereas a second sequence CentCs2 was unique to the *C. sativa* genome. A sequence homologous (~70% identity over 6156 bp of its length) to the centromeric associated retrotransposon in *Brassica* (CRB) was also identified and named Centromeric Retrotransposon of *Camelina* (CRC) (Lim *et al*., 2007; Figure S8B). The abundance of these centromeric tandem repeats enabled the position of centromeres in the *C. neglecta* genome to be provisionally located (Figure 1, Table S6). The centromere positions for most of the chromosomes were in accordance with the predictions of Mandáková *et al*. (2019), but in the case of chromosome 6, the result is more ambiguous and the centromere could lie between the ancestrally conserved genomic blocks S and V. Although it should be noted that the reduction in chromosome number of *C. neglecta* compared with the ancestral progenitor would necessarily result in remnants of centromeric regions being found in the genome (Figure S8C).

Based on a combination of *ab initio* gene prediction, experimental evidence from RNAseq data, and gene models from *A. thaliana* (Cheng *et al*., 2017), *A. lyrata* (Hu *et al*., 2011), *Thellungiella parvula* (Dassanayake *et al*., 2011), and *C. sativa* (Kagale *et al*., 2014), the assembled genome of *C. neglecta* was predicted to contain 34 061 protein-coding gene models. Among these, 33 700 were annotated to the six chromosomes, which represented a higher gene content and gene density (177 genes/Mb) in comparison to any of the subgenomes from the reference genome of *C. sativa* (141, 152, and 130 genes/Mb, respectively; Table 1, Table S7A). More than 98% of RNAseq data derived from seedling tissue could be mapped to the *C. neglecta* genome, with 94% of the reads mapping uniquely (Table S7B); 71% of the annotated genes showed expression of at least 0.1 FPKM (Figure 1, Table S7C). The subgenomes of *C. sativa* showed fractionation compared with *C. neglecta*, where the first, second, and third subgenomes retained 72.97%, 69.29%, and 70.47% of orthologues, respectively, with 7101 *C. neglecta* genes absent from the *C. sativa* genome (Figure S9). It was not unexpected to find syntenic genes present in *C. neglecta* yet absent from the first subgenome of *C. sativa* (Table S8), since orthologous copies were maintained on one of the other subgenomes. However, of these 1498 genes, 300 genes were not present in any of the subgenomes of *C. sativa*. There was no evidence of enrichment for any one biological function among the 1498 genes. Although, those 915 genes absent from *C. neglecta*, yet present in the first subgenome of *C. sativa* were found to be over-represented for genes related to defence responses (most significantly GO: 0031640: killing of cells of other organisms and GO: 0050832: defence response to fungus; Figure S10).

Synteny analysis and genome evolution trajectory

Collinearity of *C. neglecta* with *C. sativa* and *A. thaliana* was assessed with whole genome alignment using nucmer (Kurtz *et al*., 2004; Figure 3a,b). Triplication of the *C. sativa* genome compared with diploid *C. neglecta* is apparent, with *C. neglecta* sharing more similarity with the first subgenome of *C. sativa* (Figure 3a). In comparison with *A. thaliana*, although long stretches of chromosomes are conserved there is evidence of multiple large-scale rearrangements separating the two species (Figure 3b). However, synteny analysis with *A. thaliana* recovered all 24 ancestral karyotype Genomic Blocks (GB) of the Brassicaceae (Figure 3c) in the *C. neglecta* genome (Figure 3d; Table S8), and using these GBs a representative karyotype of *C. neglecta* was drawn to analyse rearrangements relative to the putative Brassicaceae progenitor genome. *Camelina neglecta* shared common chromosome structures with the ancestral karyotype for two chromosomes (AK1/CnChr1; AK3/CnChr3), while the remainder had undergone chromosome fusions and translocations to reduce the karyotype number (Figure 3c,d). The relationship between *C. neglecta* and SG1 of *C. sativa* genome was further visualized by SynVisio (Bandi and Gutwin, 2020; Figure 3e), which emphasized the conservation of syntenic genes between *C. neglecta* and the first subgenome of *C. sativa*. The chromosomal structure based on the GBs of the six *C. neglecta* chromosomes was markedly similarly to the first subgenome of *C. sativa* except for *C. neglecta* chromosome 5, which although aligning along much of its length with *C. sativa* chromosome 8, the two chromosomes were differentiated by a large pericentric inversion that led to GB R being split (Figure 3f, Figure S11). Assuming all three *C. sativa* subgenomes may have evolved from a species with a similar structure to *C. neglecta* a number of events, including inversions, translocations, and hybridization between chromosomes, could be inferred. However, a greater distance would be postulated between *C. neglecta* and the third subgenome resulting from the increased number of rearrangements invoked (Figure 3b, Figure S11).

Age of divergence of the subgenomes of *Camelina sativa* compared with *Camelina neglecta*

Using *A. thaliana* as a basal genome to define orthologues within and between the *Camelina* species, the distributions of synonymous substitutions per site rate (*K*s) were calculated for all possible duplicated syntenic gene pairs (Table S8) to identify and age any whole genome duplication events (Kagale *et al*., 2014). *K*s analysis among the genes with duplicate copies within the *C. neglecta* genome reflected the remnants of the alpha, beta, and gamma ancient whole genome duplication events found in *A. thaliana* (Figure 4a) and all angiosperms (Bowers *et al*., 2003; Table S9). Independently studying orthologues from the three subgenomes of *C. sativa* against *C. neglecta* further corroborated the similarity of the *C. neglecta* genome with that of the first subgenome of *C. sativa*. The small peak observed at 0.02 *K*s suggested the first subgenome of *C. sativa* diverged around 1.2 mya from *C. neglecta*, while peaks at 0.071 and 0.086 *K*s, for

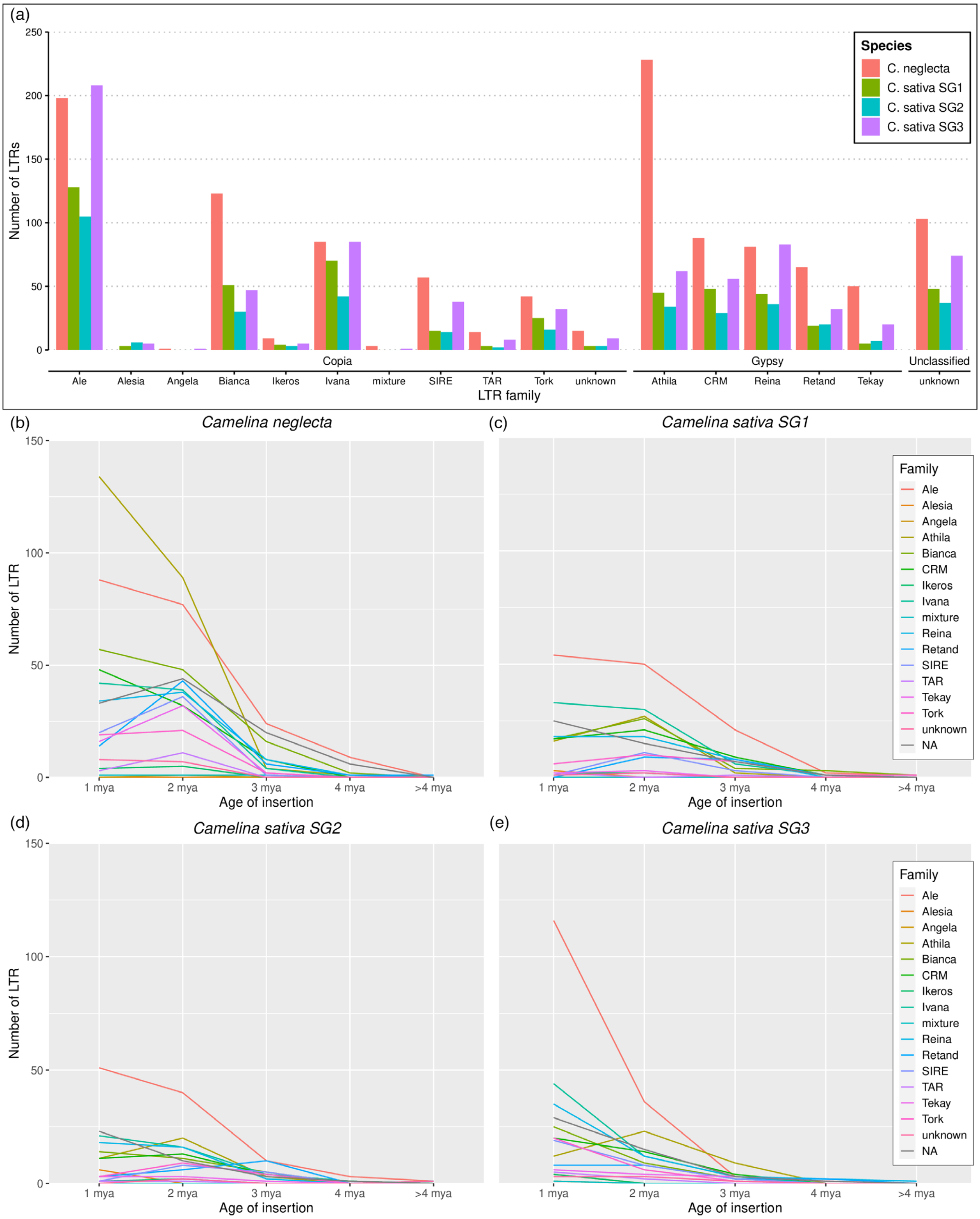


Figure 2 Full-length LTRs (FL-LTRs) of *C. neglecta* and *C. sativa* genome. Copy number of FL-LTRs families present in the *C. neglecta* and *C. sativa* subgenomes (a); age distribution of FL-LTRs families present in *C. neglecta* (b), subgenome 1 (c), subgenome 2 (d), and subgenome 3 of *C. sativa* (e).

the second and third subgenomes, respectively, corresponded to divergence dates of 4.3 and 5.2 mya. Comparing orthologues between *A. thaliana* and *C. neglecta* dated their divergence from a common ancestor to 17.8 mya, which is comparable to that estimated for the three *C. sativa* subgenomes (Kagale *et al.*, 2014; Figure 4b).

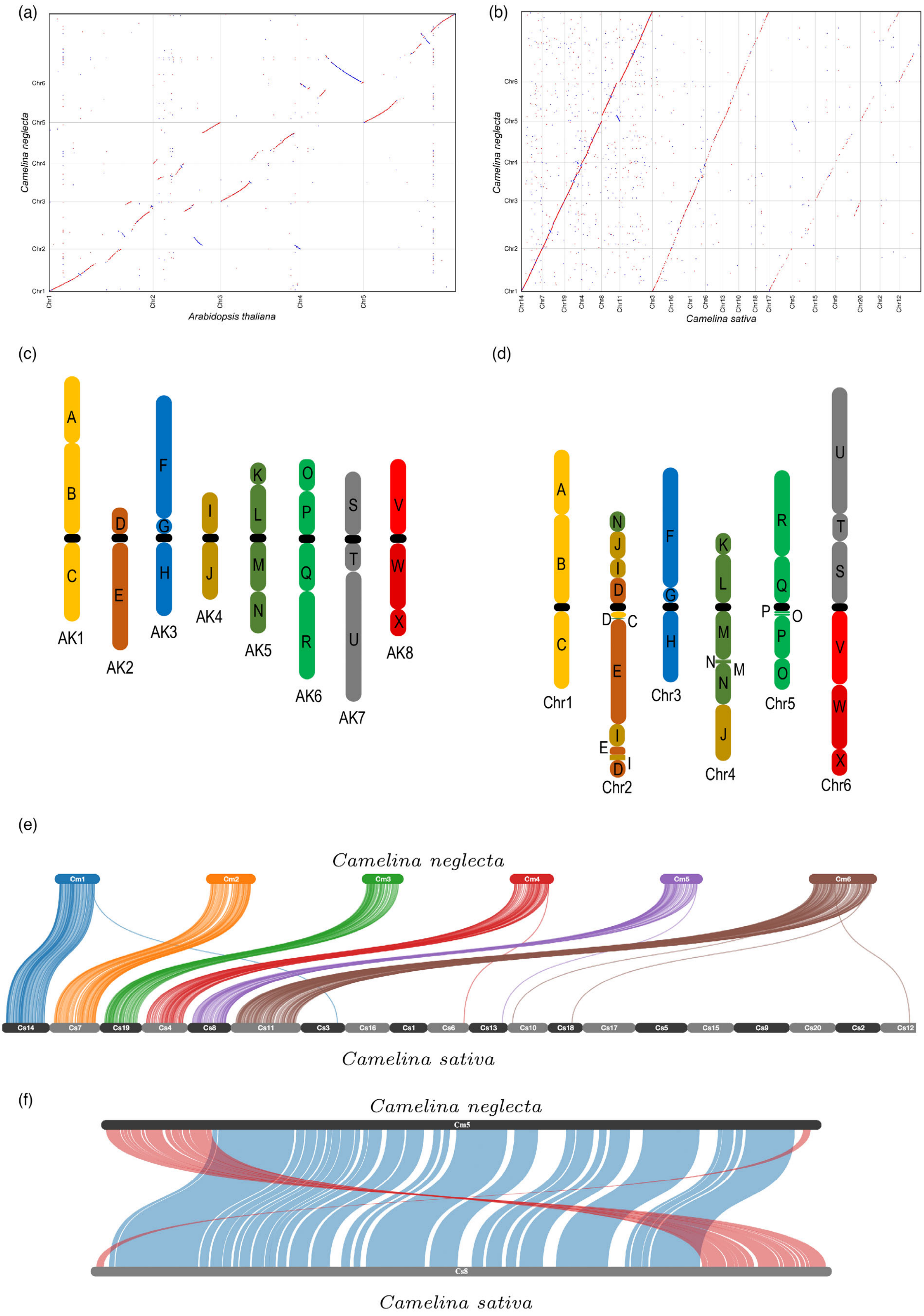

(a)
Camelina neglecta
Chr1
Chr2
Chr3
Chr4
Chr5
Chr6
Arabidopsis thaliana
Chr1
Chr2
Chr3
Chr4
Chr5
(b)
Camelina neglecta
Chr1
Chr2
Chr3
Chr4
Chr5
Chr6
Camelina sativa
Chr14
Chr7
Chr19
Chr4
Chr8
Chr11
Chr3
Chr16
Chr1
Chr6
Chr13
Chr10
Chr18
Chr17
Chr5
Chr15
Chr9
Chr20
Chr2
Chr12
(c)
A
B
C
AK1
D
E
AK2
F
G
H
AK3
I
J
AK4
K
L
M
N
AK5
O
P
Q
R
AK6
S
T
U
AK7
V
W
X
AK8
(d)
A
B
C
Chr1
N
J
I
D
D
C
E
I
E
I
D
Chr2
F
G
H
Chr3
K
L
M
N
M
N
J
Chr4
R
Q
P
O
P
O
Chr5
U
T
S
V
W
X
Chr6
(e)
Camelina neglecta
Cm1
Cm2
Cm3
Cm4
Cm5
Cm6
Cs14
Cs7
Cs19
Cs4
Cs8
Cs11
Cs3
Cs16
Cs1
Cs6
Cs13
Cs10
Cs18
Cs17
Cs5
Cs15
Cs9
Cs20
Cs2
Cs12
Camelina sativa
(f)
Camelina neglecta
Cm5
Cs8
Camelina sativa

Figure 3 Synteny analysis of *Camelina neglecta* genome. (a) Nucmer plot showing the relationship between *C. neglecta* (vertical) and *Arabidopsis thaliana* genomes (horizontal); (b) Nucmer plot showing the relationship between *C. neglecta* genome (vertical) and three *C. sativa* subgenomes (horizontal); (c) representation of genomic block organization in ancestral crucifer karyotype (ACK); (d) arrangement of 24 ancestral karyotype genomic blocks in the *C. neglecta* genome; (e) syntenic analysis of *C. neglecta* with *C. sativa* genome; and (f) comparison of chromosome 5 of *C. neglecta* with chromosome 8 of *C. sativa* showing terminally inverted region (red) in the chromosomes.

Assessing the role of recursive genome duplication in the hexaploid *C. sativa*

Duplicate genes in *C. neglecta* were analysed using *DupGen_finder* (Qiao *et al*., 2019) which identified 1653 tandem duplicates, 1203 proximal duplicates, and 3549 transposed duplicated genes (Table S10). The tandem duplicated genes were mainly enriched with the gene ontology categories: defence response (GO:0006952); secondary metabolite biosynthesis process (GO:0044550); and toxin catabolic process (GO:0009407) (Figure S12). The *Ks* analysis among the gene pairs representing tandem duplication, proximal duplication, and transposed duplication within the genome of *C. neglecta* showed some level of differentiation among paralogues (Figure S13). The same analysis of the subgenomes of *C. sativa* identified a lower number of tandem and proximal duplicate genes (Table S10), suggesting either gene duplication occurred in *C. neglecta* after the common progenitor contributed to the formation of hexaploid *C. sativa*, or additional gene copies were deleted post-hybridisation from the hexaploid.

OrthoFinder version 2.5.2 (Emms and Kelly, 2019) was used to identify species-specific orthologues in *C. neglecta* and *C. sativa*.

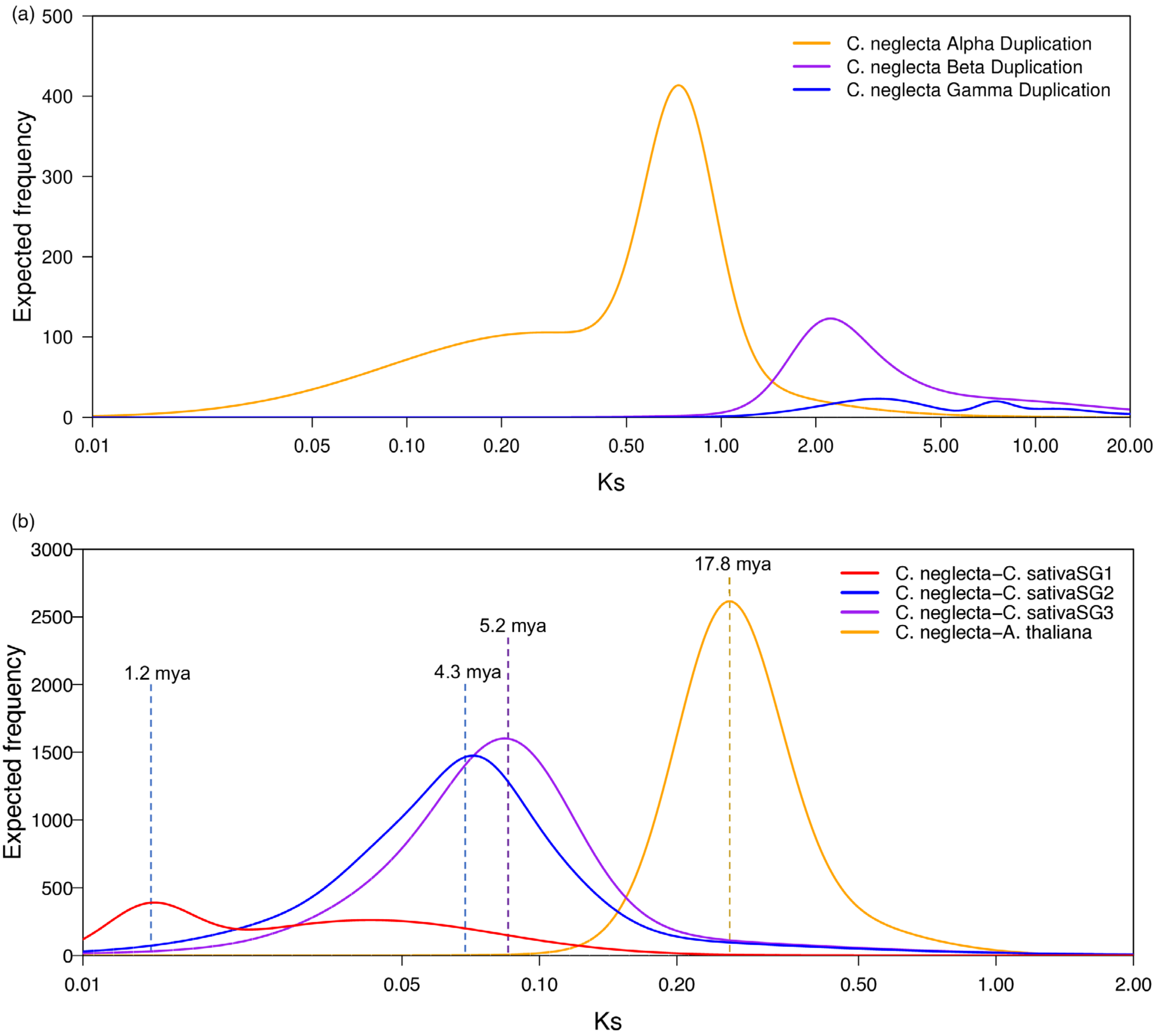


Figure 4 Evolutionary relationship between diploid *C. neglecta*, *A. thaliana*, and hexaploid *C. sativa* genome. (a) Distribution of synonymous substitution per synonymous site rate (*Ks*) among ancestral paralogs in *C. neglecta*; and (b) distribution of *Ks* values, the peaks show the age of divergence of *C. sativa* subgenomes (red, blue and purple, respectively) and the *A. thaliana* genome (orange) relative to *C. neglecta*.

Around 93.8% of genes from *C. neglecta* were assigned to gene families or orthogroups, whereas only 64.7% genes from *C. sativa* were similarly assigned. The number of orthogroups specific to *C. neglecta* (544) was lower in comparison to *C. sativa* (6769) (Table S11). The unique orthogroups from *C. neglecta* were not enriched for biological function; however, each of the subgenomes *C. sativa* had multiple unique orthogroups or genes enriched for a variety of biological processes (Tables S12–S16).

In an attempt to understand the impact of genome duplication on functional genes, the copy number of genes known to be responsible for a range of physiological processes, and often studied in the context of genome adaptation, was assessed. Disease resistance gene analogs (RGAs) are defined primarily based on their homology to known *R*-genes, and the conserved domains and motifs of *R*-genes that confer roles in resistance to specific pathogens can be used to identify all such genes in any one genome using the pipeline RGAugury (Li *et al*., 2016). RGAugury identified a total of 935 and 2967 RGAs in *C. neglecta* and *C. sativa*, respectively (Table S17). Although the level of gene duplication within *C. sativa* (3.17 fold) reflected the hexaploid nature of the genome, the pattern of expansion in each RGA category varied. The transmembrane leucine-rich repeat (TM-LRR) RGA families were more extensively replicated (3.3–5 fold) while the nucleotide-binding site leucine-rich repeat RGA families were replicated on average 2.2 times in the *C. sativa* genome. Among the TM-LRRs the greatest expansion was seen for membrane-associated receptor-like proteins (RLPs), which increased fivefold (Table S17). Only 409 RGAs identified in *C. neglecta* were found to be conserved across all three subgenomes of *C. sativa*, and the fractionation level of the remaining RGAs was equivalent (28%) in each subgenome (Table S17D).

Likewise, the assembled *C. neglecta* genome possessed 376 orthologues of 405 flowering-related genes that have been identified in *A. thaliana* (Sasaki *et al*., 2015) (Table S18). Notably, *C. neglecta* does not appear to have an orthologue of *FRIGIDA (FRI)*, an important gene responsible for the vernalization requirement in *A. thaliana* (Johanson *et al*., 2000). Similarly in hexaploid *C. sativa* although three genes share homology with *FRI* they are not found in a syntenic position, suggesting they are also not orthologous (Kagale *et al*., 2014). However, all the additional components of the *FRI* complex (FRL1, SUF4, FES1, and FLX) are found as expected in both the *C. neglecta* and *C. sativa* genomes. *Flowering Locus C* (Michaels and Amasino, 1999), which in *Arabidopsis* is upregulated by *FRI*, is present in three copies in *C. sativa* and could be a determinant of vernalization requirement in winter *Camelina* species (Anderson *et al*., 2018), including *C. neglecta*.

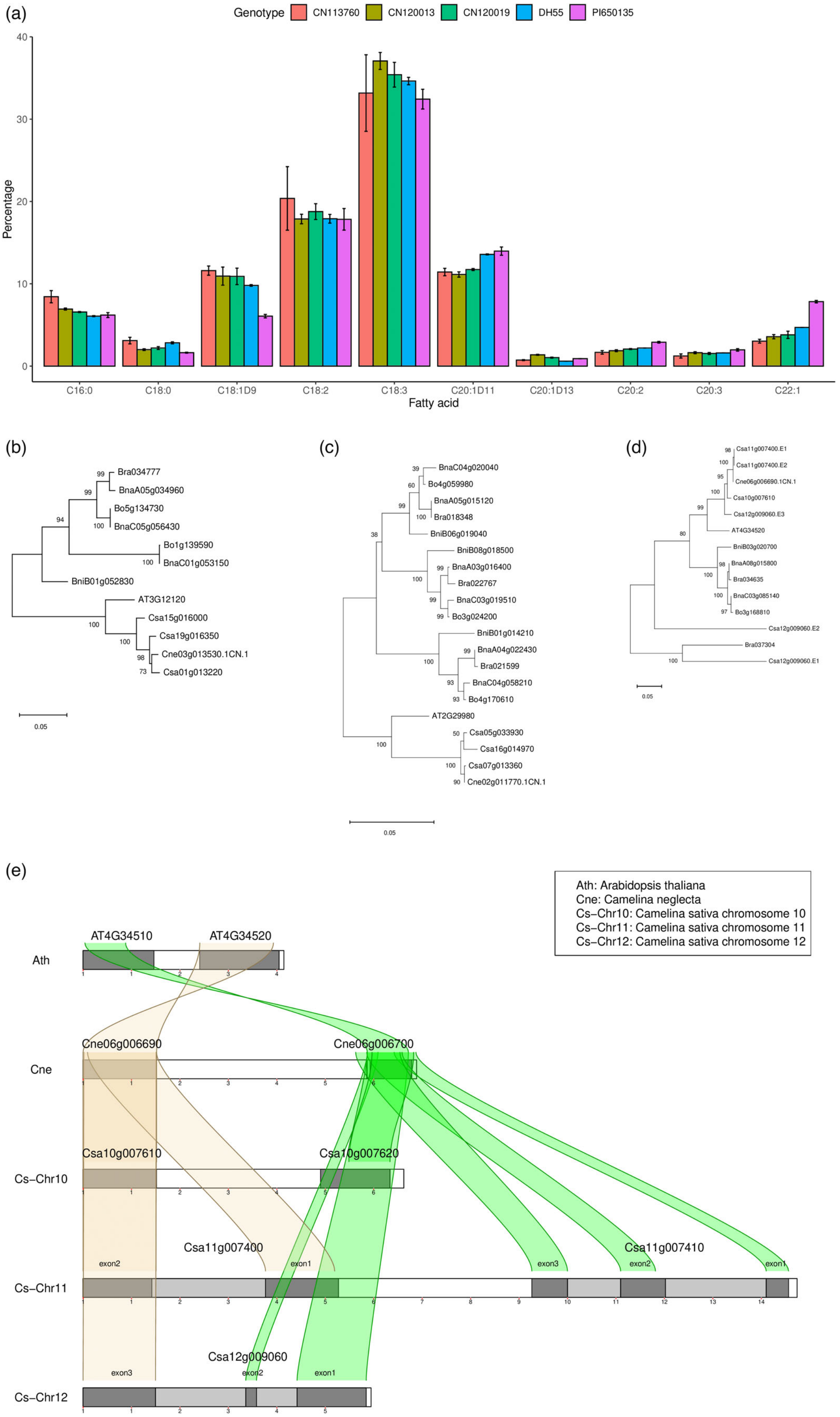
The genome of *C. neglecta* also contained a full complement of meiotic genes, which have previously been identified to have a role in faithful recombination in *A. thaliana*. In *C. sativa* the majority of these genes have been retained in triplicate (94.3%), indicating the recent formation of this allopolyploid, since such genes have been shown to fractionate more rapidly in established polyploids (Table S19; Lloyd *et al*., 2014).

Camelina neglecta has a unique fatty acid profile as compared to multiple accessions of *C. sativa*; although different *C. sativa* lines showed some variation in the shorter chain fatty acids, *C. sativa* consistently has a higher proportion of oleic acid (C18:1) and a significantly lower proportion of erucic acid (C22:1) compared with *C. neglecta* (Figure 5a; Table S20). These two long-chain fatty acids have economic importance for the food, feed, and industrial feedstock oil industries, and the manipulation of genes controlling fatty acid profiles could play an important role in these sectors. More than 83% of the genes identified as playing a role in acyl-lipid biosynthesis pathways in *A. thaliana* (http://aralip.plantbiology.msu.edu) have been retained in *C. neglecta* (Table S21). Studying the fractionation pattern of these genes in *C. sativa* did not identify an obvious pattern of gene loss or gain in the hexaploid; moreover, although some orthologues showed evidence of positive selection in *C. sativa* (*Ka*/*Ks* >1), no specific pathway appeared to be under selection (Table S21). Among these genes, the activity of *Fatty Acid Desaturase 2* (*FAD2*), *Fatty Acid Desaturase 3* (*FAD3*), and *Fatty Acid Elongase 1* (*FAE1*) have been associated with variations in oleic and erucic acid levels (Okuley *et al*., 1994; Yang *et al*., 2012) and one ortholog of each were found in *C. neglecta*. The phylogenetic relationship among the orthologs of each of these genes in *C. neglecta*'s closely related species: *C. sativa* (Kagale *et al*., 2014); *A. thaliana* (Kaul *et al*., 2000); *Brassica rapa* (Wang *et al*., 2011); *Brassica oleracea* (Parkin *et al*., 2014); *Brassica napus* (Chalhoub *et al*., 2014); and *Brassica nigra* (Perumal *et al*., 2020), suggested a fairly predictable pattern for *FAD2* and *FAD3*, with a closer relationship among the species from Brassica lineage I (Figure 5b,c). However, orthologs of *FAE1* have higher differentiation in the subgenomes of *C. sativa* with tandem duplication of *FAE1* in subgenome 1 of *C. sativa*. The analyses were complicated by miss-annotation of the *FAE* genes in the *C. sativa* genome, with tandemly duplicated copies of *FAE* on subgenome 3 annotated as a single gene (Csa12g009060) and *FAE* showing apparent tandem duplication on subgenome 1, yet lying at the site of a scaffold boundary, which confounded confirming this gene organization (Csa11g007400; Figure 5d,e, Figure S14). The low level of erucic acid in *C. sativa* might suggest selection after whole genome duplication from the progenitor *C. neglecta*, but there is no evidence that gene fractionation plays a role. Additional tissue-specific gene expression analysis may help to elucidate the role of duplicated gene segments in causing variation in oil profile in hexaploid *C. sativa* in comparison to the diploid *C. neglecta*.

Discussion

Whole genome duplication is widespread in plant evolution and is often associated with increased fitness of the newly derived

Figure 5 A comparison of the seed fatty acid (FA) profile and pertinent genes in diploid *Camelina neglecta* and hexaploid *Camelina sativa*. A comparison of per cent content of different length carbon chain FAs in hexaploid *C. sativa* (CN113760, CN120013 and CN120019, and DH55) and diploid *C. neglecta* (PI650135) seed, error bars show standard deviation (a). Maximum likelihood phylogenetic relationship of *FAD2* (b), *FAD3* (c), and *FAE1* (d) orthologs from different *Brassica* species. Genes for each species are identified as AT: *A. thaliana*, Bna: *Brassica napus*, Bni: *Brassica nigra*, Bo: *Brassica oleracea*, Bra: *Brassica rapa*, Cne: *C. neglecta*, and Csa: *C. sativa*. The support value (1000 replications) is shown at each branch, and a scale bar indicating the branch length is provided for each tree. A schematic showing alignment of *FAE*1-related genes in *C. sativa* compared with *C. neglecta* (e), the links represents gene identity of more than 90%, and kb distance is indicated on the chromosomes.


(a)
Genotype
CN113760
CN120013
CN120019
DH55
PI650135
Percentage
0
10
20
30
40
C16:0
C18:0
C18:1D9
C18:2
C18:3
C20:1D11
C20:1D13
C20:2
C20:3
C22:1
Fatty acid
(b)
Bra034777
BnaA05g034960
Bo5g134730
BnaC05g056430
Bo1g139590
BnaC01g053150
BniB01g052830
AT3G12120
Csa15g016000
Csa19g016350
Cne03g013530.1CN.1
Csa01g013220
99
99
94
100
100
100
100
98
73
0.05
(c)
BnaC04g020040
Bo4g059980
BnaA05g015120
Bra018348
BniB06g019040
BniB08g018500
BnaA03g016400
Bra022767
BnaC03g019510
Bo3g024200
BniB01g014210
BnaA04g022430
Bra021599
BnaC04g058210
Bo4g170610
AT2G29980
Csa05g033930
Csa16g014970
Csa07g013360
Cne02g011770.1CN.1
39
60
99
100
38
100
99
99
99
100
99
93
93
100
50
100
90
0.05
(d)
Csa11g007400.E1
Csa11g007400.E2
Cne06g006690.1CN.1
Csa10g007610
Csa12g009060.E3
AT4G34520
BniB03g020700
BnaA08g015800
Bra034635
BnaC03g085140
Bo3g168810
Csa12g009060.E2
Bra037304
Csa12g009060.E1
98
100
95
100
99
80
98
100
100
97
100
0.05
(e)
Ath: Arabidopsis thaliana
Cne: Camelina neglecta
Cs–Chr10: Camelina sativa chromosome 10
Cs–Chr11: Camelina sativa chromosome 11
Cs–Chr12: Camelina sativa chromosome 12
AT4G34510
AT4G34520
Ath
Cne06g006690
Cne06g006700
Cne
Csa10g007610
Csa10g007620
Cs–Chr10
Csa11g007400
Csa11g007410
exon2
exon1
exon3
exon2
exon1
Cs–Chr11
Csa12g009060
exon3
exon2
exon1
Cs–Chr12

polyploid plant; defining the constituent genomes of any polyploid provides not only insights into plant evolution but can form a repository for novel variation. *Camelina neglecta* is a relatively recent discovery among relatives of the crop *C. sativa* and has been suggested to be closely related to the progenitor of subgenome 1 of the hexaploid *C. sativa* (Chaudhary *et al*., 2020). The high-quality sequence of *C. neglecta* enforces this relationship, with only a single major chromosomal rearrangement separating *C. neglecta* from subgenome 1 of *C. sativa*. Sequencing of the *C. neglecta* genome has provided a valuable genomic resource with the potential for *Camelina* breeding, and also an opportunity to characterize and compare genomic features across the different ploidy levels found among related *Camelina* species. The *C. neglecta* genome along with those released for related *Camelina* species (Martin *et al*., 2021) should assist with efforts to elucidate all progenitor genomes of the hexaploid crop. A recent study has suggested tetraploid *Camelina microcarpa* (or *C. intermedia*), which based on chromosome painting and comparative analyses is an allopolyploid formed from the fusion of two *C. neglecta*-like genomes, as the maternal progenitor of *C. sativa* (Mandáková and Lysak, 2022). Together these studies lay the groundwork for the artificial resynthesis of *C. sativa*, which could aid in expanding the limited gene pool of the hexaploid species.

Comparative analysis of *C. neglecta* genome with *C. sativa* and *A. thaliana* genomes suggested that rearrangements of genomic blocks in *C. sativa* subgenome 1 likely occurred after genome duplication/hybridisation. The landscape of the genome largely matches that identified by chromosomal painting (Mandáková *et al*., 2019), although the presence of centromeric specific elements could suggest the centromere of chromosome six lies between ancestral GB S and V rather than S and T. The position between S and V also represents a potential site of chromosomal condensation compared with the ancestral karyotype, which means chromosome 6 would be expected to show evidence of an ancient centromere. Analyses of full-length LTR REs show a continuous expansion of the diploid genome in comparison to subgenomes of the hexaploid (Figure 2). The level of full-length repeat proliferation appears older in the submissive subgenome (subgenome 1) compared with the dominant (subgenome 3) and the progenitor *C. neglecta*, suggesting genome dominance impacts more than gene expression and fractionation levels. The amount of REs is similar in *C. neglecta* and subgenome 1 of *C. sativa*, but there was a higher abundance of helitrons elements in all subgenomes of *C. sativa*, indicating a recent expansion of these elements potentially triggered by the polyploidization event (s) (Table S2). Perhaps as an antithesis to the hexaploid state of *C. sativa*, the diploid showed a higher prevalence of tandemly duplicated genes, that could provide material for gene functional differentiation.

Although the third subgenome of *C. sativa* is dominant with regard to gene expression, there was no evidence of gene fractionation bias compared with the diploid, further emphasizing the neopolyploid nature of the hexaploid. Interestingly *C. sativa* retained over 900 genes in the first subgenome, which were presumably deleted from *C. neglecta* subsequent to the polyploidisation event, yet these genes appeared to be enriched for defence responses. The availability of a fully annotated genome allowed the study of genes involved in well-studied physiological processes in the two species, identifying a number of interesting features. Of particular note is the absence of an orthologue of the flowering time gene *FRIGIDA (FRI)*. *Camelina neglecta* has a winter habit compared with *C. sativa*, and vernalisation is essential to initiate the transition to flowering. Although *FRI* has an established role in controlling the vernalisation response in *A. thaliana*, such that a functional copy is required to maintain *FLC* expression, some ecotypes have been identified, which are late flowering despite carrying a mutation in *FRI* (Werner *et al*., 2005). It is possible that the *Camelina* species utilize an alternative pathway to control *FLC* expressions, such as through the interaction of *ART1/HUA2* and *FLC* (Doyle *et al*., 2005), or they have evolved an independent mechanism. The availability of a progenitor genome for *C. sativa* allows for the potential of resynthesis to generate novel variation in the crop, a method that has proved effective in the breeding of *Brassica napus* (Gaeta *et al*., 2007). It is apparent that there are novel phenotypes available in *C. neglecta*, both flowering time and modified oil profiles have been indicated here, and the divergent pattern of RGA gene expansion could suggest novel sources of resistance. However, *C. neglecta* has only one accession in current germplasm collections though it has been suggested that the species could be common in southern France, possibly other regions of Europe, and even across the Eurasian steppe (Brock *et al*., 2022). As such, the collection of new material from these regions would be warranted to identify additional diversity within the diploid, and perhaps the derived tetraploid that represents subgenomes 1 and 2 of *C. sativa*. Notwithstanding this limitation, the genome of *C. neglecta* will prove a foundational resource for further evolutionary studies among the genus.

Methods

Plant material

Camelina neglecta accession PI650135 was selected for this study. Leaf samples were harvested from three 1-month-old individual plants and flash-frozen samples representing the same accession (10 g) were stored at −80 °C prior to sending to Dovetail Genomics (Scotts Valley, CA). High molecular weight DNA isolation, genome sequencing, and assembly were carried out by Dovetail Genomics.

PacBio library preparation and sequencing

For the whole genome sequencing of *C. neglecta*, PacBio SMRTbell libraries (~20 kb) for PacBio Sequel were constructed using SMRTbell Template Prep Kit 1.0 (PacBio, Menlo Park, CA) using standard manufacturer protocol. The pooled library was bound to polymerase using the Sequel Binding Kit 2.0 (PacBio) and loaded onto PacBio Sequel using the MagBead Kit V2 (PacBio). Sequencing was performed on 2 PacBio Sequel SMRT cells (Pacific BioSciences).

Genome assembly, polishing, scaffolding, and evaluation

Initial assemblies were performed using the FALCON 1.8.8 pipeline from Pacific Bioscience, where three stages of the FALCON pipeline using whole genome, single-molecule, real-time sequencing (SMRT) data resulted in a *C. neglecta* genome comprising 131 primary contigs containing 192.4 Mbp with an N50 contig length of 7.9 Mbp. Finally, the assembly was polished through PacBio's Arrow algorithm from SMRT Link 5.0.1, using the original raw reads.

Further, a Chicago library (Putnam *et al*., 2016) and a Dovetail HiC library (Lieberman-Aiden *et al*., 2009) were prepared where chromatin was fixed in place with formaldehyde in the nucleus and then extracted fixed chromatin was digested with *Dpn*II; and the DNA was then sheared to ~350 bp mean fragment size, and sequencing libraries were generated using NEBNext Ultra enzymes and Illumina-compatible adapters. The libraries were sequenced on an Illumina HiSeqX, generating 251 M, and 174 M PE150 reads, respectively, for the Chicago library and Dovetail HiC library.

For scaffolding, the *de novo* assembly, Chicago library reads, and Dovetail HiC library reads were used as input data for HiRise Scaffolder (Putnam *et al*., 2016). First, Chicago library sequences were aligned to the draft assembly using a modified SNAP read mapper (http://snap.cs.berkeley.edu). The separations of Chicago read pairs mapped within draft scaffolds were analysed by HiRise to produce a likelihood model for the genomic distance between read pairs, and the model was used to identify and break putative misjoins, score prospective joins, and make joins above a threshold. After aligning and scaffolding the Chicago data, Dovetail HiC library sequence was aligned and scaffolded following the same method. The HiC contact map was visualized using the software Juicebox (Durand *et al*., 2016).

Illumina paired-end 125 bp reads were generated on an Illumina HiSeq 2500 platform for the same accession PI650135. The reads obtained from Illumina were trimmed for poor quality reads, short reads (<55 bp), and adapter contamination by Trimmomatic v.0.39 (Bolger *et al*., 2014). A total of 228 million clean raw reads were mapped to the assembled genome using BWA version 0.7.17 (Li and Durbin, 2009) with default parameters, and the assembly was evaluated for with Qualimap (García-Alcalde *et al*., 2012). Further, the completeness of genes was assessed utilizing BUSCO v4.1.4 (Simão *et al*., 2015) with *brassicales_odb10* (*n* = 4596) data set.

Genome size estimation

Jellyfish v.2.2.6 (Marçais and Kingsford, 2011) was used to estimate the genome size where 17-mer, 21-mer, 25-mer, and 31-mer were used to estimate genome size (Table S22**)**. The histogram obtained from jellyfish was used to estimate the genome size using an online platform of GenomeScope (http://qb.cshl.edu/genomescope/) (Vurture *et al*., 2017; Figure S1). Smudgeplot (Ranallo-Benavidez *et al*., 2020) was used to predict genome structure using heterozygous k-mers (*hetkmers* function with default parameters; Figure S3).

Gene annotation

BRAKER1 (Hoff *et al*., 2019) was used to annotate the genes in the assembled genome of *C. neglecta* where GeneMark-ES (Lomsadze *et al*., 2005) and AUGUSTUS version 3.4 (Stanke and Morgenstern, 2005) were used for prediction of gene models. Evidence from RNAseq data generated from the same accession at the early growth stage, protein homology of *C. sativa* genome (Kagale *et al*., 2014), and gene models from *A. thaliana*, *A. lyrata*, *Thellungiella parvula*, and *C. sativa* were utilized for the prediction of gene models. Initially, BRAKER annotated 32 830 gene models, which was refined using PASA (Haas *et al*., 2008) to yield a total of 34 061 gene models.

Repeat annotation

Repetitive elements (REs), including transposable elements (TEs) and tandem repeats (TRs), were characterized from *C. neglecta* genome using a combination of structural and homology-based approaches. EDTA tool v.1.9.9 (Ou *et al*., 2019) was used to predict the whole genome repeat content in the two *Camelina* genomes. LTR assembly quality in the *Camelina* genomes was estimated using the LTR assembly index (LAI) tool (Ou *et al*., 2018). In addition, structure-based characterization including the age of full-length long terminal repeat retrotransposon (FL-LTR-RTs) was done using the LTR retriever program (Ou and Jiang, 2018), which used the inputs from LTR_finder (Xu and Wang, 2007) and LTRHarvest (Ellinghaus *et al*., 2008). RepeatMasker was implemented to identify and classify homologous repeat elements in the genome using Camelina-lib against the *C. neglecta* genome assembly. These repeat elements were plotted along the genome of *C. neglecta* using KaryotypeR (Gel and Serra, 2017) from R statistical software (Team, 2021).

Synteny analysis

Syntenic analysis of *C. neglecta* was performed as described by Kagale *et al*. (2014) where the protein models of *C. neglecta* and *A. thaliana* were compared using reciprocal BLASTP. Further, DAGChainer (Haas *et al*., 2004) with default parameters was used to identify *C. neglecta–A. thaliana* gene pairs and construct a syntelog table. MCscanX (Wang *et al*., 2012) with parameters of Match_score 50, Match_size 10, Gap_penalty -1, Overlap_window 10, E-value 1e-05, and Max gaps 5 was used to identify syntenic genes, which were visualized with SynVisio (Bandi and Gutwin, 2020).

From the synteny table, gene pairs between *C. neglecta* and subgenomes of *C. sativa* as well as *A. thaliana* were identified and used to calculate the rates of synonymous substitutions per synonymous site (*K*s) using GenoDup pipeline (Mao, 2019). This pipeline utilized protein and nucleotide sequences of a gene pair to calculate (*K*s). MAFFT (Katoh *et al*., 2002) was used for the alignment of protein sequences of each gene family. TranslatorX (Abascal *et al*., 2010) was used for the alignment of coding sequences based on the alignment of protein sequences, and codeml package in PAML (Yang, 2007) was used to calculate *K*s value. The value of *K*s was plotted using the Gaussian mixture model using mclust version 5.4.7 (Scrucca *et al*., 2016) in R statistical software (Team, 2021). Further, a comparison between the assembled genome of PI650135 and recently assembled genome of *C. neglecta* (Martin *et al*., 2021) was performed using MUMer v.3.23 (Kurtz *et al*., 2004).

Prediction of duplicate genes and orthogroups

From the annotated genes of *C. neglecta*, duplicated genes as a result of tandem duplication, proximate duplication, and transposed duplication were identified using pipeline *DupGen_finder.pl* (Qiao *et al*., 2019). In the case of *C. neglecta*, 'all-versus-all' BLASTP alignment (*E*-value <1e-10), as well as BLASTP alignment (*E*-value <1e-10) with *A. thaliana* as an outgroup, was performed to identify the duplication events; whereas in the case of *C. sativa* (Kagale *et al*., 2014), *C. neglecta* was considered as an outgroup to identify duplicated genes with the same parameters as used for *C. neglecta* duplicate gene prediction. For these analyses, the MCScanX (Wang *et al*., 2012) parameters were set as Match_score 50, Match_size 5, Gap_penalty -1, Overlap_window 5, E_value 1e-10, and Max gaps 25. Tandem duplicated genes, proximal duplicate genes, and transposed genes were visualized using circos v. 0.69-6 (Krzywinski *et al*., 2009). The protein sequences of annotated genes from *C. neglecta* and *C. sativa* genomes were implemented in OrthoFinder version 2.5.2 (Emms and Kelly, 2019) to identify orthogroups specific to each genome.

Identification of disease resistance genes analogs

Resistance gene analogs (RGAs) were identified in the *C. neglecta* genome using *RGAugury* pipeline (Li *et al*., 2016). The protein sequences of annotated genes were used to identify disease resistance genes where four classes of RGAs were analysed such as NBS-encoding protein, receptor-like protein kinases, receptor-like proteins, and transmembrane-coiled coil proteins.

RNA sequencing and gene expression analysis

Five biological replicates of *C. neglecta* (PI650135) were taken for the transcriptome study, among these three samples were collected from 1-week-old seedling (leaf sample) grown in a CYG seed germination pouch (Mega International, Newport, MN), whereas two samples were collected from the plant (leaf sample) kept at vernalization conditions of 4 °C for 30 days. Total RNA was extracted using a standard RNeasy Plant Qiagen kit as described by the manufacturer with on-column DNA digestion. RNA was quantified using a Qubit (Invitrogen, Waltham, MA, USA), and the quality was determined using an RNA Nano lab chip on a Bioanalyzer (Agilent Technologies, Mississauga, ON, Canada). Paired-end RNAseq libraries were constructed using the TruSeq RNA preparation kit (Illumina, San Diego, CA, USA), with 100 ng of RNA used for cDNA synthesis followed by RNA library preparation. The final library quality was checked using a Bioanalyzer (Agilent Technologies). Sequencing was conducted on an Illumina HiSeq2000 platform at National Research Council Canada in Saskatoon (2 × 125 bp).

Sequence data were filtered for low-quality reads (<40), short reads (<55 bp), and adapter contamination using Trimmomatic v.0.33 (Bolger *et al*., 2014). High-quality reads were aligned with the annotated genome of PI650135 using STAR v. 2.7.6 (Dobin *et al*., 2013). Utilization of *GeneCounts* features from STAR provided the number of transcripts per annotated gene. A maximum of four mismatches was allowed while mapping transcripts to the reference genome. Normalization of transcripts was done using Fragment Per Kilobase of transcripts per Million mapped reads (FPKM) method.

Gene ontology enrichment analysis

The corresponding orthologue in *A. thaliana* was identified based on the syntenic table for all the differentially expressed genes or the unique genes present in *C. neglecta* compared with the first subgenome of the *C. sativa* or vice versa. The significantly enriched Gene Ontology (GO) terms (false discovery rate <0.05) were identified for *C. neglecta* and *C. sativa* using agriGO v.2.0 online program (http://systemsbiology.cau.edu.cn/agriGOv2/; Tian *et al*., 2017).

Fatty acid analysis and phylogenetic analysis of *FAD2*, *FAD3*, and *FAE1*

Fatty acid extraction was performed using around 30 mg seeds per sample digested with 1.5 mL of sulphuric acid in methanol (1.5% v/v) in a Pyrex screw-top methylation tube. Further, 0.4 mL of Toluene was added along with 50 μL of internal standard (10 mg/mL triheptadecanoin) and incubated at 90 °C for 2 h for complete digestion. The digested solution was allowed to cool, and 1 mL of 0.9% NaCl and 1 mL of hexane were added, the solution was centrifuged at 1000 rpm at room temperature in the Thermo Legend XTR centrifuge for 1 min to separate the phases. 200 μL of the top hexane phase was collected in a GC vial containing a glass insert. The prepared sample was analysed for the fatty acid profile in a GC autosampler with the method MSFAMES1.

Identification of *FAD2*, *FAD3*, and *FAE1* orthologs in Brassica species was performed with nucleotide blast (Altschul *et al*., 1990) search against the protein-coding genes from the genome of *C. sativa* (Kagale *et al*., 2014), *C. neglecta*, *Brassica rapa* (Wang *et al*., 2011), *Brassica oleracea* (Parkin *et al*., 2014), *Brassica napus* (Chalhoub *et al*., 2014), and *Brassica nigra* (Perumal *et al*., 2020). All the identified orthologs have been listed in Table S23. Further, the sequences of these genes were aligned with MUSCLE (Edgar, 2004) in MEGA X (Stecher *et al*., 2020) with default parameters. The maximum likelihood method and Tamura-Nei model (Tamura and Nei, 1993) were used to construct the phylogenetic trees with a bootstrap of 1000 replications in MEGA X.

Acknowledgements

The research was funded in part by Agriculture and Agri-Food Canada's Crop Genomics Initiative with additional support from the Global Institute for Food Security, Saskatoon.

Conflicts of interest

The authors declare no conflicts of interest.

Author contributions

R.C. and E.E.H. performed the experiments, and R.C. wrote the draft manuscript. I.A.P.P. and A.G.S. conceived the project, secured the funding and finalized the manuscript. R.C., C.K., L.J., and S.P. performed the analysis, S.K. and M.S. provided expert advice and reviewed the manuscript.

Data availability statement

All associated data have been submitted to NCBI under the Bioproject ID: PRJNA869687.

References

Abascal, F., Zardoya, R. and Telford, M.J. (2010) TranslatorX: multiple alignment of nucleotide sequences guided by amino acid translations. *Nucleic Acids Res.* **38**, W7–W13.

Abel, S., Möllers, C. and Becker, H. (2005) Development of synthetic *Brassica napus* lines for the analysis of "fixed heterosis" in allopolyploid plants. *Euphytica*, **146**, 157–163.

Al-Shehbaz, I.A., Beilstein, M.A. and Kellogg, E.A. (2006) Systematics and phylogeny of the Brassicaceae (Cruciferae): an overview. *Plant Syst. Evol.* **259**, 89–120.

Altschul, S.F., Gish, W., Miller, W., Myers, E.W. and Lipman, D.J. (1990) Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410.

Anderson, J.V., Horvath, D.P., Doğramaci, M., Dorn, K.M., Chao, W.S., Watkin, E.E., Hernandez, A.G. *et al*. (2018) Expression of FLOWERING LOCUS C and a frameshift mutation of this gene on chromosome 20 differentiate a summer and winter annual biotype of *Camelina sativa*. *Plant Direct*, **2**, e00060.

Bandi, V. and Gutwin, C. (2020) *Interactive exploration of genomic conservation*. URL https://graphicsinterface.org/wp-content/uploads/gi2020-9.pdf

Belton, J.-M., McCord, R.P., Gibcus, J.H., Naumova, N., Zhan, Y. and Dekker, J. (2012) Hi–C: a comprehensive technique to capture the conformation of genomes. *Methods*, **58**, 268–276.

Benson, G. (1999) Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res.* **27**, 573–580.

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

Bowers, J.E., Chapman, B.A., Rong, J. and Paterson, A.H. (2003) Unravelling angiosperm genome evolution by phylogenetic analysis of chromosomal duplication events. *Nature*, **422**, 433–438.

Brock, J.R., Donmez, A.A., Beilstein, M.A. and Olsen, K.M. (2018) Phylogenetics of *Camelina* Crantz. (Brassicaceae) and insights on the origin of gold-of-pleasure (*Camelina sativa*). *Mol. Phylogenet. Evol.* **127**, 834–842.

Brock, J.R., Mandakova, T., Lysak, M.A. and Al-Shehbaz, I.A. (2019) *Camelina neglecta* (Brassicaceae, Camelineae), a new diploid species from Europe. *PhytoKeys*, **115**, 51–57.

Brock, J.R., Mandáková, T., McKain, M., Lysak, M.A. and Olsen, K.M. (2022) Chloroplast phylogenomics in *Camelina* (Brassicaceae) reveals multiple origins of polyploid species and the maternal lineage of *C. sativa*. *Hortic. Res.* **9**, uhab050.

Cairns, J., Freire-Pritchett, P., Wingett, S.W., Várnai, C., Dimond, A., Plagnol, V., Zerbino, D. *et al*. (2016) CHiCAGO: robust detection of DNA looping interactions in Capture Hi-C data. *Genome Biol.* **17**, 1–17.

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

Chaudhary, R., Koh, C.S., Kagale, S., Tang, L., Wu, S.W., Lv, Z., Mason, A.S. *et al*. (2020) Assessing diversity in the *Camelina* genus provides insights into the genome structure of *Camelina sativa*. *G3 (Bethesda)*, **10**, 1297–1308.

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

Cheng, C.Y., Krishnakumar, V., Chan, A.P., Thibaud-Nissen, F., Schobel, S. and Town, C.D. (2017) Araport11: a complete reannotation of the *Arabidopsis thaliana* reference genome. *Plant J.* **89**, 789–804.

Comai, L. (2005) The advantages and disadvantages of being polyploid. *Nat. Rev. Genet.* **6**, 836–846.

Dassanayake, M., Oh, D.-H., Haas, J.S., Hernandez, A., Hong, H., Ali, S., Yun, D.-J. *et al*. (2011) The genome of the extremophile crucifer *Thellungiella parvula*. *Nat. Genet.* **43**, 913–918.

Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P. *et al*. (2013) STAR: Ultrafast universal RNA-seq aligner. *Bioinformatics*, **29**, 15–21.

Doyle, M.R., Bizzell, C.M., Keller, M.R., Michaels, S.D., Song, J., Noh, Y.S. and Amasino, R.M. (2005) HUA2 is required for the expression of floral repressors in *Arabidopsis thaliana*. *Plant J.* **41**, 376–385.

Durand, N.C., Robinson, J.T., Shamim, M.S., Machol, I., Mesirov, J.P., Lander, E.S. and Aiden, E.L. (2016) Juicebox provides a visualization system for Hi-C contact maps with unlimited zoom. *Cell Syst.* **3**, 99–101.

Edgar, R.C. (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* **32**, 1792–1797.

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

Emms, D.M. and Kelly, S. (2019) OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biol.* **20**, 1–14.

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

García-Alcalde, F., Okonechnikov, K., Carbonell, J., Cruz, L.M., Götz, S., Tarazona, S., Dopazo, J. *et al*. (2012) Qualimap: evaluating next-generation sequencing alignment data. *Bioinformatics*, **28**, 2678–2679.

Gehringer, A., Friedt, W., Lühs, W. and Snowdon, R.J. (2006) Genetic mapping of agronomic traits in false flax (*Camelina sativa* subsp. *sativa*). *Genome*, **49**, 1555–1563.

Gel, B. and Serra, E. (2017) karyoploteR: an R/Bioconductor package to plot customizable genomes displaying arbitrary data. *Bioinformatics*, **33**, 3088–3090.

Ghurye, J., Pop, M., Koren, S., Bickhart, D. and Chin, C.-S. (2017) Scaffolding of long read assemblies using long range contact information. *BMC Genomics*, **18**, 1–11.

Girollet, N., Rubio, B., Lopez-Roques, C., Valière, S., Ollat, N. and Bert, P.-F. (2019) De novo phased assembly of the *Vitis riparia* grape genome. *Sci. Data*, **6**, 1–8.

Haas, B.J., Delcher, A.L., Wortman, J.R. and Salzberg, S.L. (2004) DAGchainer: a tool for mining segmental genome duplications and synteny. *Bioinformatics*, **20**, 3643–3646.

Haas, B.J., Salzberg, S.L., Zhu, W., Pertea, M., Allen, J.E., Orvis, J., White, O. *et al*. (2008) Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments. *Genome Biol.* **9**, 1–22.

Hoff, K.J., Lomsadze, A., Borodovsky, M. and Stanke, M. (2019) Whole-genome annotation with BRAKER. In *Gene Prediction*(Kollmar, M., ed), pp. 65–95. New York, NY: Springer.

Hu, T.T., Pattyn, P., Bakker, E.G., Cao, J., Cheng, J.-F., Clark, R.M., Fahlgren, N. *et al*. (2011) The *Arabidopsis lyrata* genome sequence and the basis of rapid genome size change. *Nat. Genet.* **43**, 476–481.

Johanson, U., West, J., Lister, C., Michaels, S., Amasino, R. and Dean, C. (2000) Molecular analysis of FRIGIDA, a major determinant of natural variation in *Arabidopsis* flowering time. *Science*, **290**, 344–347.

Kagale, S., Koh, C., Nixon, J., Bollina, V., Clarke, W.E., Tuteja, R., Spillane, C. *et al*. (2014) The emerging biofuel crop *Camelina sativa* retains a highly undifferentiated hexaploid genome structure. *Nat. Commun.* **5**, 1–11.

Kagale, S., Nixon, J., Khedikar, Y., Pasha, A., Provart, N.J., Clarke, W.E., Bollina, V. *et al*. (2016) The developmental transcriptome atlas of the biofuel crop *Camelina sativa*. *Plant J.* **88**, 879–894.

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

Kaul, S., Koo, H.L., Jenkins, J., Rizzo, M., Rooney, T., Tallon, L.J., Feldblyum, T. *et al*. (2000) Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature*, **408**, 796–815.

Krzywinski, M., Schein, J., Birol, I., Connors, J., Gascoyne, R., Horsman, D., Jones, S.J. *et al*. (2009) Circos: an information aesthetic for comparative genomics. *Genome Res.* **19**, 1639–1645.

Kurtz, S., Phillippy, A., Delcher, A.L., Smoot, M., Shumway, M., Antonescu, C. and Salzberg, S.L. (2004) Versatile and open software for comparing large genomes. *Genome Biol.* **5**, R12.

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

Li, P., Quan, X., Jia, G., Xiao, J., Cloutier, S. and You, F.M. (2016) RGAugury: a pipeline for genome-wide prediction of resistance gene analogs (RGAs) in plants. *BMC Genomics*, **17**, 1–10.

Lieberman-Aiden, E., Van Berkum, N.L., Williams, L., Imakaev, M., Ragoczy, T., Telling, A., Amit, I. *et al*. (2009) Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science*, **326**, 289–293.

Lim, K.B., Yang, T.J., Hwang, Y.J., Kim, J.S., Park, J.Y., Kwon, S.J., Kim, J. *et al*. (2007) Characterization of the centromere and peri-centromere retrotransposons in *Brassica rapa* and their distribution in related Brassica species. *Plant J.* **49**, 173–183.

Lloyd, A.H., Ranoux, M., Vautrin, S., Glover, N., Fourment, J., Charif, D., Choulet, F. *et al*. (2014) Meiotic gene evolution: can you teach a new dog new tricks? *Mol. Biol. Evol.* **31**, 1724–1727.

Lomsadze, A., Ter-Hovhannisyan, V., Chernoff, Y.O. and Borodovsky, M. (2005) Gene identification in novel eukaryotic genomes by self-training algorithm. *Nucleic Acids Res.* **33**, 6494–6506.

Luo, Z., Brock, J., Dyer, J.M., Kutchan, T.M., Augustin, M., Schachtman, D.P., Ge, Y. *et al*. (2019) Genetic diversity and population structure of a *Camelina sativa* spring panel. *Front. Plant Sci.* **10**, 184.

Mandáková, T. and Lysak, M.A. (2022) The identification of the missing maternal genome of the allohexaploid camelina (*Camelina sativa*). *Plant J.* **112**, 622–629.

Mandáková, T., Pouch, M., Brock, J.R., Al-Shehbaz, I.A. and Lysak, M.A. (2019) Origin and evolution of diploid and allopolyploid Camelina genomes were accompanied by chromosome shattering. *Plant Cell*, **31**, 2596–2612.

Mao, Y. (2019) GenoDup Pipeline: a tool to detect genome duplication using the dS-based method. *PeerJ*, **7**, e6303.

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

Martin, S.L., Smith, T.W., James, T., Shalabi, F., Kron, P. and Sauder, C.A. (2017) An update to the Canadian range, abundance, and ploidy of *Camelina* spp.(Brassicaceae) east of the Rocky Mountains. *Botany*, **95**, 405–417.

Martin, S.L., Lujan-Toro, B.E., Sauder, C.A., James, T., Ohadi, S. and Hall, L.M. (2019) Hybridization rate and hybrid fitness for *Camelina microcarpa* Andrz. ex DC (♀) and *Camelina sativa* (L.) Crantz(Brassicaceae) (♂). *Evol. Appl.* **12**, 443–455.

Martin, S.L., Toro, B.L., James, T., Sauder, C.A. and Laforest, M. (2021) *Insights from the genomes of four diploid Camelina* spp. *bioRxiv*, https://doi.org/10.1101/2021.08.23.455123

Michaels, S.D. and Amasino, R.M. (1999) FLOWERING LOCUS C encodes a novel MADS domain protein that acts as a repressor of flowering. *Plant Cell*, **11**, 949–956.

Novak, P., Neumann, P., Pech, J., Steinhaisl, J. and Macas, J. (2013) RepeatExplorer: a Galaxy-based web server for genome-wide characterization of eukaryotic repetitive elements from next-generation sequence reads. *Bioinformatics*, **29**, 792–793.

Okuley, J., Lightner, J., Feldmann, K., Yadav, N. and Lark, E. (1994) *Arabidopsis* FAD2 gene encodes the enzyme that is essential for polyunsaturated lipid synthesis. *Plant Cell*, **6**, 147–158.

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

Ou, S., Chen, J. and Jiang, N. (2018) Assessing genome assembly quality using the LTR Assembly Index (LAI). *Nucleic Acids Res.* **46**, e126.

Ou, S., Su, W., Liao, Y., Chougule, K., Agda, J.R., Hellinga, A.J., Lugo, C.S.B. *et al*. (2019) Benchmarking transposable element annotation methods for creation of a streamlined, comprehensive pipeline. *Genome Biol.* **20**, 1–18.

Parkin, I.A., Koh, C., Tang, H., Robinson, S.J., Kagale, S., Clarke, W.E., Town, C.D. *et al*. (2014) Transcriptome and methylome profiling reveals relics of genome dominance in the mesopolyploid *Brassica oleracea*. *Genome Biol.* **15**, 1–18.

Perumal, S., Koh, C.S., Jin, L., Buchwaldt, M., Higgins, E.E., Zheng, C., Sankoff, D. *et al*. (2020) A high-contiguity *Brassica nigra* genome localizes active centromeres and defines the ancestral *Brassica* genome. *Nat. Plants*, **6**, 929–941.

Putnam, N.H., O'Connell, B.L., Stites, J.C., Rice, B.J., Blanchette, M., Calef, R., Troll, C.J. *et al*. (2016) Chromosome-scale shotgun assembly using an ,ro method for long-range linkage. *Genome Res.* **26**, 342–350.

Qiao, X., Li, Q., Yin, H., Qi, K., Li, L., Wang, R., Zhang, S. *et al*. (2019) Gene duplication and evolution in recurring polyploidization–diploidization cycles in plants. *Genome Biol.* **20**, 1–23.

Ranallo-Benavidez, T.R., Jaron, K.S. and Schatz, M.C. (2020) GenomeScope 2.0 and Smudgeplot for reference-free profiling of polyploid genomes. *Nat. Commun.* **11**, 1–10.

Rosyara, U., Kishii, M., Payne, T., Sansaloni, C.P., Singh, R.P., Braun, H.-J. and Dreisigacker, S. (2019) Genetic contribution of synthetic hexaploid wheat to CIMMYT's spring bread wheat breeding germplasm. *Sci. Rep.* **9**, 1–11.

Sasaki, E., Zhang, P., Atwell, S., Meng, D. and Nordborg, M. (2015) "Missing" G x E variation controls flowering time in *Arabidopsis thaliana*. *PLoS Genet.* **11**, e1005597.

Schnable, J.C., Springer, N.M. and Freeling, M. (2011) Differentiation of the maize subgenomes by genome dominance and both ancient and ongoing gene loss. *Proc. Natl Acad. Sci. USA*, **108**, 4069–4074.

Scrucca, L., Fop, M., Murphy, T.B. and Raftery, A.E. (2016) mclust 5: clustering, classification and density estimation using Gaussian finite mixture models. *R J.* **8**, 289.

Simão, F.A., Waterhouse, R.M., Ioannidis, P., Kriventseva, E.V. and Zdobnov, E.M. (2015) BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics*, **31**, 3210–3212.

Singh, R., Bollina, V., Higgins, E.E., Clarke, W.E., Eynck, C., Sidebottom, C., Gugel, R. *et al*. (2015) Single-nucleotide polymorphism identification and genotyping in *Camelina sativa*. *Mol. Breed.* **35**, 35.

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

Song, X., Wei, Y., Xiao, D., Gong, K., Sun, P., Ren, Y., Yuan, J. *et al*. (2021) *Brassica carinata* genome characterization clarifies U's triangle model of evolution and polyploidy in *Brassica*. *Plant Physiol.* **186**, 388–406.

Stanke, M. and Morgenstern, B. (2005) AUGUSTUS: a web server for gene prediction in eukaryotes that allows user-defined constraints. *Nucleic Acids Res.* **33**, W465–W467.

Stecher, G., Tamura, K. and Kumar, S. (2020) Molecular evolutionary genetics analysis (MEGA) for macOS. *Mol. Biol. Evol.* **37**, 1237–1239.

Tamura, K. and Nei, M. (1993) Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. *Mol. Biol. Evol.* **10**, 512–526.

Team, R.C. (2021) *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. URL https://www.R-project.org

Tian, T., Liu, Y., Yan, H., You, Q., Yi, X., Du, Z., Xu, W. *et al*. (2017) agriGO v2. 0: a GO analysis toolkit for the agricultural community, 2017 update. *Nucleic Acids Res.* **45**, W122–W129.

Vollmann, J., Grausgruber, H., Stift, G., Dryzhyruk, V. and Lelley, T. (2005) Genetic diversity in camelina germplasm as revealed by seed quality characteristics and RAPD polymorphism. *Plant Breed.* **124**, 446–453.

Vurture, G.W., Sedlazeck, F.J., Nattestad, M., Underwood, C.J., Fang, H., Gurtowski, J. and Schatz, M.C. (2017) GenomeScope: fast reference-free genome profiling from short reads. *Bioinformatics*, **33**, 2202–2204.

Wang, X., Wang, H., Wang, J., Sun, R., Wu, J., Liu, S., Bai, Y. *et al*. (2011) The genome of the mesopolyploid crop species *Brassica rapa*. *Nat. Genet.* **43**, 1035–1039.

Wang, Y., Tang, H., DeBarry, J.D., Tan, X., Li, J., Wang, X., Lee, T.-H. *et al*. (2012) MCScanX: a toolkit for detection and evolutionary analysis of gene synteny and collinearity. *Nucleic Acids Res.* **40**, e49.

Wei, Z., Wang, M., Chang, S., Wu, C., Liu, P., Meng, J. and Zou, J. (2016) Introgressing subgenome components from *Brassica rapa* and *B. carinata* to *B. juncea* for broadening its genetic base and exploring intersubgenomic heterosis. *Front. Plant Sci.* **7**, 1677.

Werner, J.D., Borevitz, J.O., Uhlenhaut, N.H., Ecker, J.R., Chory, J. and Weigel, D. (2005) FRIGIDA-independent variation in flowering time of natural *Arabidopsis thaliana* accessions. *Genetics*, **170**, 1197–1207.

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

Yang, Z. (2007) PAML 4: phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol.* **24**, 1586–1591.

Yang, Q., Fan, C., Guo, Z., Qin, J., Wu, J., Li, Q., Fu, T. *et al*. (2012) Identification of FAD2 and FAD3 genes in *Brassica napus* genome and development of allele-specific markers for high oleic and low linolenic acid contents. *Theor. Appl. Genet.* **125**, 715–729.

Zubr, J. (1997) Oil-seed crop: *Camelina sativa*. *Ind. Crops Prod.* **6**, 113–119.

Supporting information

Additional supporting information may be found online in the Supporting Information section at the end of the article.

Figure S1 Genome size estimation of *C. neglecta* genome accession PI650135 using k-mer analysis with different k-mer lengths.
Figure S2 High-throughput chromosome confirmation capture (HiC) contact map of *C. neglecta* genome.
Figure S3 (A) Smudgeplot showing highly diploidized nature of *C. neglecta* genome. (B) Smudgeplot showing triplicated nature of *C. sativa* genome.
Figure S4 BUSCO assessment of the gene content of the *C. neglecta* genome.
Figure S5 Alignment of *C. neglecta* assembly (horizontal access) to one recently submitted (vertical access) (Martin *et al*., 2021).
Figure S6 Boxplot showing distribution of long-terminal repeat assembly index (LAI) in *C. neglecta* and *C. sativa* genomes based on a window of 3 Mb and a sliding step of 300 Kb.
Figure S7 Age of proliferation of full length LTR in *C. neglecta* and subgenomes of *C. sativa*.

Figure S8 Centromere associated retrotransposons in *C. neglecta*.

Figure S9 Number of *C. neglecta* genes sharing homology (E-value = 1^{e-10}) with genes in the subgenomes of *C. sativa*.

Figure S10 Gene Ontology analysis of genes absent in *C. neglecta* genome.

Figure S11 Possible rearrangements in the *C. neglecta* like genome during the formation of chromosomes of *C. sativa*.

Figure S12 Gene Ontology analysis of tandem duplicate genes present in *C. neglecta* genome.

Figure S13 The distribution of Ks value among the whole genome duplication (A), tandem (B), proximal (C) and transposed (D) duplicate genepairs in the *C. neglecta* genome.

Figure S14 Dot plot of protein sequences of FAE1 genes from *C. neglecta* (Cne06g006690.1CN.1), *C. sativa* (Csa11g007400.1, Csa10g007610.1 and Csa12g009060.1) and *A. thaliana* (AT4G34520.1).

Table S1 Summary statistics of mapping Illumina data to the assembled genome.

Table S2 Distribution of repeat elements in *Camelina neglecta* and *Camelina sativa* genomes.

Table S3 Summary of family-level distribution of full-length long terminal repeats retrotransposons in *C. neglecta* and *C. sativa* genomes.

Table S4 List of full-length long terminal repeats retrotransposons in *C. neglecta* genome.

Table S5 List of full-length long terminal repeats retrotransposons in *C. sativa* genome.

Table S6 Distribution of centromeric repeats in *C. neglecta* genome.

Table S7 (A) Total number of genes in *Camelina neglecta* and *C. sativa* genomes with average length of the gene. (B) Mapping of RNASeq reads to the genome of *C. neglecta* using STAR. (C) Gene expression (FPKM) in the leaf tissue of *C. neglecta*.

Table S8 Syntelog table for *Arabidopsis thaliana*, *Camelina neglecta* and subgenomes of *Camelina sativa* (Cs-SG).

Table S9 (A–C) Genepairs identified to participate in the alpha, beta and gamma genome duplication.

Table S10 Number of duplicate genes across *C. neglecta* and *C. sativa* genomes.

Table S11 Clustering analysis of gene families across *C. neglecta* and *C. sativa* genomes.

Table S12 List of unique orthogroups in *C. neglecta* identified through orthofinder.

Table S13 List of unique orthogroups in *C. sativa* identified through orthofinder.

Table S14 Gene Ontology of *Camelina sativa* subgenome 1 gene having unique orthogroups compared with *Camelina neglecta*.

Table S15 Gene Ontology of *Camelina sativa* subgenome 2 genes having unique orthogroups compared with *Camelina neglecta*.

Table S16 Gene Ontology of *Camelina sativa* subgenome 3 genes having unique orthogroups compared with *Camelina neglecta*.

Table S17 (A–D) Number and list of RGAs found in *C. neglecta* and *C. sativa* genomes.

Table S18 List of flowering-related genes present in the *C. neglecta* genome.

Table S19 List of meiosis-related genes present in the *Camelina neglecta* and *Camelina sativa* subgenomes.

Table S20 Seed oil content and fatty acid composition of different *Camelina* species.

Table S21 List of *Arabidopsis thaliana* genes responsible for acyl-lipid biosynthesis with orthologs in *Camelina neglecta* and *C. sativa* genomes.

Table S22 Genome size estimation using k-mer analysis.

Table S23 List of Brassica genes sharing similarity with FAD2, FAD3, and FAE1 identified through blastn.