Horticulture Research

Horticulture Research, 2024, **11**: uhae247

https://doi.org/10.1093/hr/uhae247

Article

Exploring genetic diversity, population structure, and subgenome differences in the allopolyploid *Camelina sativa*: implications for future breeding and research studies

Jordan R. Brock[1,*], Kevin A. Bird[2], Adrian E. Platts[1], Fabio Gomez-Cano[3], Suresh Kumar Gupta[3], Kyle Palos[4], Caylyn E. Railey[4,5], Scott J. Teresi[1,6], Yun Sun Lee[3], Maria Magallanes-Lundback[1], Emily G. Pawlowski[3], Andrew D.L. Nelson[4], Erich Grotewold[3,*], and Patrick P. Edger[1,*]

[1]Department of Horticulture, Michigan State University, 1066 Bogue St, East Lansing, MI 48824, USA
[2]Department of Plant Sciences, University of California-Davis, 1 Shields Ave, Davis, CA 95616, USA
[3]Department of Biochemistry and Molecular Biology, Michigan State University, 603 Wilson Rd, East Lansing, MI 48824-6473, USA
[4]Boyce Thompson Institute, Cornell University, 533 Tower Rd, Ithaca, NY 14853, USA
[5]Plant Biology Graduate Field, Cornell University, 533 Tower Rd, Ithaca, NY 14853, USA
[6]Genetics and Genome Sciences Program, Michigan State University, 567 Wilson Rd Room 2165, East Lansing, MI 48824, USA
*Corresponding authors. E-mails: brockjor@msu.edu; grotewol@msu.edu; edgerpat@msu.edu

Abstract

Camelina (*Camelina sativa*), an allohexaploid species, is an emerging aviation biofuel crop that has been the focus of resurgent interest in recent decades. To guide future breeding and crop improvement efforts, the community requires a deeper comprehension of subgenome dominance, often noted in allopolyploid species, "alongside an understanding of the genetic diversity" and population structure of material present within breeding programs. We conducted population genetic analyses of a *C. sativa* diversity panel, leveraging a new genome, to estimate nucleotide diversity and population structure, and analyzed for patterns of subgenome expression dominance among different organs. Our analyses confirm that *C. sativa* has relatively low genetic diversity and show that the SG3 subgenome has substantially lower genetic diversity compared to the other two subgenomes. Despite the low genetic diversity, our analyses identified 13 distinct subpopulations including two distinct wild populations and others putatively representing founders in existing breeding populations. When analyzing for subgenome composition of long non-coding RNAs, which are known to play important roles in (a)biotic stress tolerance, we found that the SG3 subgenome contained significantly more lincRNAs compared to other subgenomes. Similarly, transcriptome analyses revealed that expression dominance of SG3 is not as strong as previously reported and may not be universal across all organ types. From a global analysis, SG3 "was only significant higher expressed" in flower, flower bud, and fruit organs, which is an important discovery given that the crop yield is associated with these organs. Collectively, these results will be valuable for guiding future breeding efforts in camelina.

Introduction

Camelina (*Camelina sativa*), also known as false-flax and gold-of-pleasure, is an ancient cruciferous oilseed crop consumed in Europe and Western Asia for over 6000 years for its calorie dense and oil-rich seeds [1]. As a food and feed crop, camelina benefits from high levels of omega-3 fatty acids and a favorable composition consisting largely of polyunsaturated fatty acids [2–4]. Generally, there are two types of camelina, the facultative annual spring-type and the obligate biennial winter-type, distinguished by the requirement of vernalization. These two types lend versatility to camelina as a crop, as the winter type can be grown as a cover crop over winter periods where fields may be fallow, whereas the spring type has a rapid generation time resulting in a faster harvest. In recent decades, camelina has emerged as a promising candidate for renewable energy, with aviation biofuels derived from camelina oil promising a 75% reduction in greenhouse gas emissions compared to petroleum-based fuels [5]. Cultivation of camelina can be achieved on otherwise non-arable land using minimal inputs of nitrogen fertilizer and has been described as a disease and drought tolerant relative to *Brassica napus* [6]. Additionally, camelina has been explored as a high-value chemical molecule factory capable of producing a variety of expensive industrial and pharmaceutical compounds through transgenesis [7–11]. Lastly, camelina has long been a model for evolution, specifically in regards to phenotypic plasticity and crop mimicry [12–14].

C. sativa is an allohexaploid ($2n = 6x = 40$), formed from recent hybrid origins involving the diploid species *C. neglecta* and *C. hispida*, and an unknown *C. neglecta*-like progenitor species [15–18]. Multiple naming schemes exist for the subgenomes of *C.*

Received: 1 May 2024; **Accepted:** 26 August 2024; **Published:** 9 September 2024; **Corrected and Typeset:** 1 November 2024

sativa including the *C. neglecta* subgenome (SG1/N6/Cs-G1), the *C. neglecta*-like subgenome (SG2/N7/Cs-G2) and the *C. hispida* subgenome (SG3/H7/Cs-G3) [16, 18–21]. A tetraploid intermediate species, *C. intermedia*, exists which contains SG1 and SG2 which later hybridized with *C. hispida* (SG3) forming the allohexaploid *C. sativa*. The polyploidization event resulting in *C. sativa* is dated at ~65 Kya [15]. Because relatives of progenitor species are extant, these species represent a valuable resource that may be used as a model system for studying polyploidy and uncovering the underlying genetics of important traits in camelina. Camelina is the most closely related crop species to Arabidopsis [22, 23] having diverged only ~8 Mya in contrast to *B. napus* which diverged from Arabidopsis ~23 Mya [24]. As such, camelina benefits from the genetic resources developed in Arabidopsis and situates camelina well as a model for translational research. The short-generation time of camelina, coupled with its ability to self-fertilize, also lends well to its use as a model system. Transformation of camelina is relatively easy and efficient with several protocols available [25, 26]. Lastly, many resources have already been developed in camelina, including databases for gene regulation [27, 28] and genomics (e.g., cruciferseq.ca and camregbase.org), as well as genomic resources for the diploid relatives *C. neglecta*, *C. laxa*, and *C. hispida* [20, 29, 30].

The genome of camelina [31] has enabled a swell of research in the system and made possible many new genetic and genomic discoveries. However, this genome was released nearly a decade ago using now outdated short-read sequencing technologies. Short-read genome assemblies are known to result in higher fragmented assemblies, particularly near or within repetitive regions of the genome [32, 33]. To further advance the field of camelina research, a new genome is required, one that includes high accuracy long-read (3rd generation) sequencing technology, ideally from a line that is already commonly used in research. The improved quality will enhance coverage of the gene space for RNAseq studies and increase the effectiveness of synthesized guide RNAs for genome editing.

As part of this study, we assembled a nearly complete genome of camelina variety "Suneson" which was sequenced with single molecule real-time sequencing from PacBio. The camelina variety "Suneson" is a spring-type advanced cultivar released by the Montana Agricultural Experiment Station in 2007 and named after the native Montanan and former UC-Davis breeder and USDA agronomist, Coit A. Suneson. It is routinely used as a model for molecular genetics and as a platform for transgenesis [7, 28, 34–37]. Our assembly represents a 211-fold increase in contig N50 relative to the short-read assembly, DH55. We recovered an additional ~19 Mb of the genome and 5326 additional protein-coding genes in our annotation. Annotations of transposable elements (TEs) and long noncoding RNAs (lncRNAs) were generated to supplement this genomic resource. It is worth noting that another version of the "Suneson" genome was published recently, which was sequenced using Oxford Nanopore technology [38].

We leveraged our new improved genome to obtain novel insights into the population genetics of camelina. For example, we identified genetically distinct cultivar and wild subpopulations which should be of valuable to public and private breeding programs. Subgenome expression was also explored to characterize the degree of subgenome dominance in various organ tissues. In summary, this high-quality reference for the lab model *C. sativa* "Suneson", alongside these discoveries, will enable future research opportunities and guide breeding efforts.

Results

Assembly of *C. sativa* Suneson genome

To generate a reference assembly of *C. sativa*, we employed long-read Pacific Biosciences (PacBio) (Menlo Park, CA) HiFi sequencing on the commonly used lab-model *C. sativa* "Suneson", with seeds provided by Yield10 Bioscience (Woburn, MA). We generated 32 Gb (~ 43× coverage) of PacBio HiFi reads with a read N50 of 12.9 kb. The final genome assembly consisted of 20 chromosomes (Fig. 1) and a total of 661 Mb with a scaffold N50 of 29.4 Mb representing an 11.5% increase in genome assembly size compared to the previous genome [31]. We assessed genome completeness of the assembly with Benchmarking Universal Single Copy Orthologs (BUSCO) revealing the assembly to be 99.7% complete (1.8% single copy, 97.9% duplicated, 0.1% fragmented, 0.2% missing). Hi-C sequencing was conducted to ensure proper assembly and orientation of the assembly (Supplemental Fig. S1).

Complementary to the new assembly, we developed a new annotation using both PacBio IsoSeq long-read RNA sequencing technology and short-read Illumina PE sequencing from 19 tissue-types and MeJa stress in *C. sativa* "Suneson". In total 117 688 gene models were predicted including 94 744 annotated protein-coding genes compared to 94 495 and 89 418, respectively, in the previous assembly of DH55 (Table 1). The assembly and annotation statistics are overall similar to the recently published Oxford Nanopore long-read assembly of *C. sativa* "Suneson" (Table 1) (Supplemental Fig. S2) [38]. The completeness of our annotation was assessed with BUSCO resulting in 99.6% completeness of our annotation (97.8% duplicated, 1.8% single copy, 0.0% fragmented, 0.4% missing). Our *C. sativa* "Suneson" assembly filled in the majority of gaps in anchored chromosomes that had existed in the DH55 genome (Supplemental Fig. S3). Our assembly contains 169 N-regions representing a total of 0.0028% gaps with no major difference in gap content between subgenomes, while the DH55 version contained 6.47% gaps with an unequal proportion of gaps among subgenomes (SG1 = 6.20%, SG2 = 5.90%, SG3 = 7.14%, see Supplemental Table S1). Canonical Arabidopsis-type TTTAGGG telomeric repeats (>10 consecutive repeats, within 10 kb of start/end of chromosome) were found on both ends for 7 chromosomes, one end for 10 chromosomes, and on neither end for 3 chromosomes (Supplemental Table S2). In contrast, telomeric repeats were not found on chromosome ends of the DH55 v2 genome.

Analysis of genome resequencing data

Following methods outlined in Li et al. 2021, we re-mapped the resequencing data of 222 accessions (Supplemental Table S3) of *C. sativa* to our "Suneson" assembly. Using the new assembly, we obtained ~25% more total unfiltered SNPs, ~5 million total SNPs, compared to ~4 million SNPs in the previous study [41]. However, we employed more strict filtering for our SNP dataset, including a minimum mapping quality of 20 and a minimum base quality of 30, yielding 3.98 million SNPs and 138 469 after final filtering (see methods). Genetic diversity was calculated for the 222 accessions of *C. sativa* revealing relatively low genetic diversity $\pi = 0.00086$. When assessing genetic diversity across only whole chromosomes, average nucleotide diversity was found to be $\pi = 0.00098$. The weighted average of nucleotide diversity for each subgenome was also calculated by separating chromosomes based on their subgenome of origin resulting in estimates for the *C. neglecta* subgenome (SG1) $\pi = 0.00100$, *C. neglecta*-like (SG2) subgenome $\pi = 0.00122$, and *C. hispida* subgenome (SG3) $\pi = 0.00079$ (Supplemental Table S4). Heterozygosity was calculated for all individuals both at the

Figure 1. Circos plot [39] of the hexaploid *C. sativa* "Suneson" genome assembly and synteny with the diploid *Capsella rubella* genome [40]. Camelina chromosomes are arranged by subgenome –SG1 (red) consists of chromosomes 4,7,8,11,14, and 19, while SG2 (blue) consists of chromosomes 1,3,6,10,13,16 and 18, and SG3 (green) consists of chromosomes 2,5,9,12,15,17, and 20. Scale indicates chromosome intervals with major ticks of 5mb and minor ticks of 1mb.

genome and subgenome level to address the overall amount of heterozygosity as well as to identify individuals and populations with potentially valuable heterozygosity for breeding programs (Supplemental Table S5). The individuals with the lowest heterozygosity were Suneson, Borow1, PRFGL76, Czestochowska, and Kirkizska ($H < 0.00035$). Those individuals with the highest heterozygosity were Borow2, Przybrodzka, and Auslese1 ($H > 0.20$). The average genome-wide heterozygosity for all individuals was 0.0337, with heterozygosity being highest in SG3 (SG1, $H = 0.0318$; SG2, $H = 0.0332$; SG3, $H = 0.0363$).

To address the degree of genetic diversity and understand groupings of individuals, we analyzed population structure for the resequencing lines. We ran ADMIXTURE on a range of K-values from 1 to 32 (Supplemental Fig. S4). We determined that the cross-validation error was lowest at $K = 13$ (CV = 0.64409, Supplemental Fig. S5). At $K = 13$, genetic clustering by population can be observed in principal component space such that the 13 genetic populations are generally occurring in distinct clusters (Fig. 2A). Aside from cases of highly admixed individuals at $K = 4$, we observe these four populations grouping together in a phylogenetic context (Fig. 3). Pairwise *Fst* values between the four subpopulations ranged from 0.055 (pop1/pop3) to 0.207 (pop2/pop3) suggesting low to moderate levels of genetic differentiation (Supplemental Table S6). Wild-collected and winter-type accessions were found to be interspersed throughout the tree with two individual clades in pop1 composed predominantly of wild-collected accessions (Fig. 3). Wild accessions group almost entirely with Western (pop1) and mostly Eastern (pop4) genetic populations. At $K = 13$, several interesting groups remain as distinct genetic populations. One such group includes the reference genome accession, Suneson, and may be largely composed of germplasm that was either used as breeding material for the generation of Suneson, or

Table 1. Genome assembly statistics of the new *C. sativa* variety Suneson genome compared to the previously published *C. sativa* DH55 genome. * = This study. ** = Oxford Nanopore assembly [38]

	Suneson*	**Suneson****	**DH55**
Assembly technology	PacBio	ONT	Illumina/Roche 454
Total coverage	43×	42×	123×
Assembled genome size	660.89 Mb	644.49 Mb	641.45 Mb
Anchored genome size	610.23 Mb	633.61 Mb	608.54 Mb
Scaffolds >200 bp	339	62	37 871
Gaps in anchored genome (%)	0.02 Mb (0.0028%)	0.01 Mb (0.0015%)	39.37 Mb (6.47%)
Scaffold N50	29.40 Mb	32.18 Mb	30.01 Mb
Contig N50	7.70 Mb	Not reported	32.17 Kb
Complete BUSCO genes (genome)	99.7%	99.5%	99.6%
Annotated protein coding genes	94 744	91 877	89 418
Gene models	117 688	133 355	94 495
Complete BUSCO genes (annotation)	99.6%	98.4%	99.7%
GC content %	37.02%	36.63%	33.99%

offspring that have since been renamed or interbred. Further, we identify two distinct wild populations at $K = 13$, one representing wild collected *C. sativa* from Czech Republic, Germany, Sweden, and Bulgaria and the other representing lines collected in the Republic of Georgia.

Subgenome dominance

Several allopolyploids have been shown to exhibit subgenome expression dominance, such that the dominant subgenome exhibits higher global expression of transcripts relative to the submissive subgenomes. This was observed in *C. sativa* with the previously released DH55 genome, where SG3 was observed to have more genes of significantly higher expression that are retained in 1:1:1 ratios across the three subgenomes [16, 31]. We reassessed the degree of subgenome dominance using the new *C. sativa* "Suneson" genome. First, we identified 11 269 syntelogs that were 1:1:1 in each subgenome and with a syntelogs in *Arabidopsis thaliana* [42]. Each subgenome can also be observed with comparisons to the genome of *Capsella rubella* [40] (Fig. 1). Six RNA-seq datasets were examined for subgenome dominance including leaf, leaf treated with methyl jasmonate (10 hr post-treatment), root, whole flower, flower bud, and young fruit (Supplemental Table S7). Median transcript per million (TPM) expression of SG3 was found to be marginally higher than SG1 and SG2 in most comparisons, and only significant in pairwise comparisons for flower, flower bud, and young fruit (Fig. 4, Supplemental Fig. S6). The number of biased genes (log2 fold expression difference greater than |2|) was only significantly higher in SG3 than SG1 in flower, flower bud, and young fruit, and higher than SG2 in flower, young fruit, methyl jasmonate (10 hr), and leaf (Supplemental Fig. S7). Pairwise measures of homoeolog expression bias in various tissues largely showed no, or very slight bias (Supplemental Figs. S8-S12). For instance in flower tissue, a slight bias toward SG3 genes was observed when compared to SG1 and SG2 (Fig. 5). Gene ontology (GO) enrichment uncovered that the biased genes in SG3 subgenome are enriched with functions associated with abiotic stress response, phytohormones (e.g., abscisic acid; ABA), among other related GO terms (Supplemental Table S8).

Annotation of long noncoding RNAs

Long noncoding RNAs (lncRNAs) are poorly understood when compared to protein-coding genes; however, because of their roles in regulating protein-coding gene expression under abiotic and biotic stress conditions [43, 44], they are a prime target for characterization for the improvement of crops. Illumina and Nanopore RNA sequencing was used to identify novel transcriptional units in the *C. sativa* "Suneson" genome, yielding a total of 1979 intergenic lncRNAs (lincRNAs). Six of these lincRNAs were novel to the new assembly and not annotated in the DH55 version of the genome with the same methods [45]. Further, 849 antisense lncRNAs (ASlncRNAs), which are transcribed from the antisense strand of a gene were identified (Supplemental Table S9) which are novel to this study. When analyzing the subgenome composition of lincRNAs, we found SG1 = 626, SG2 = 516, and SG3 = 837, whereas there was a more even distribution of ASlncRNAs with SG1 = 298, SG2 = 258, and SG3 = 293. When using a combined approach of identification with CPC2 and CPAT, 1633 lincRNAs, 830 ASlncRNAs, and 395 promoter associated lincRNAs (within 500 bp of promoter) were identified (Supplemental Table S10).

Pangenome annotation of TEs

The composition of TEs and their location in a genome plays key roles in genome structure, function, and evolution. Using a pangenome approach, we annotated TEs in *C. sativa* "Suneson" as well as its diploid relatives. We found that TEs account for 27%–38% of the genomes of *Camelina* diploid species, compared to 27.14% for the allohexaploid *C. sativa* (Supplemental Table S11). We also calculated the proportion of TEs within the three subgenomes of *C. sativa* and found considerable variation with SG1 = 26.91%, SG2 = 23.69%, SG3 = 33.36% (Supplemental Table S11). For most individual TE families (e.g., Helitron, Copia and Gypsy), SG3 had a higher relative percentage abundance compared to SG1 and SG2 subgenomes (Supplemental Table S11). However, the SG3 subgenome had lower amounts for hAT, CACTA, and Harbinger TEs. The total proportion of TEs, both by types and overall, annotated in *C. sativa* subgenomes SG1 and SG3 largely reflects that found in the diploid genomes *C. neglecta* and *C. hispida*, respectively.

Discussion

Here, we present new estimates for nucleotide diversity and relatedness among wild and cultivated camelina accessions. Our estimates of whole-genome nucleotide diversity, based on the same set of 222 accessions used in a previous resequencing study [41] resulted in a substantially lower estimate of $\pi = 0.00086$, compared to 0.0013. Another study using the DH55 genome assembly used genotype-by-sequencing to assess genetic diversity

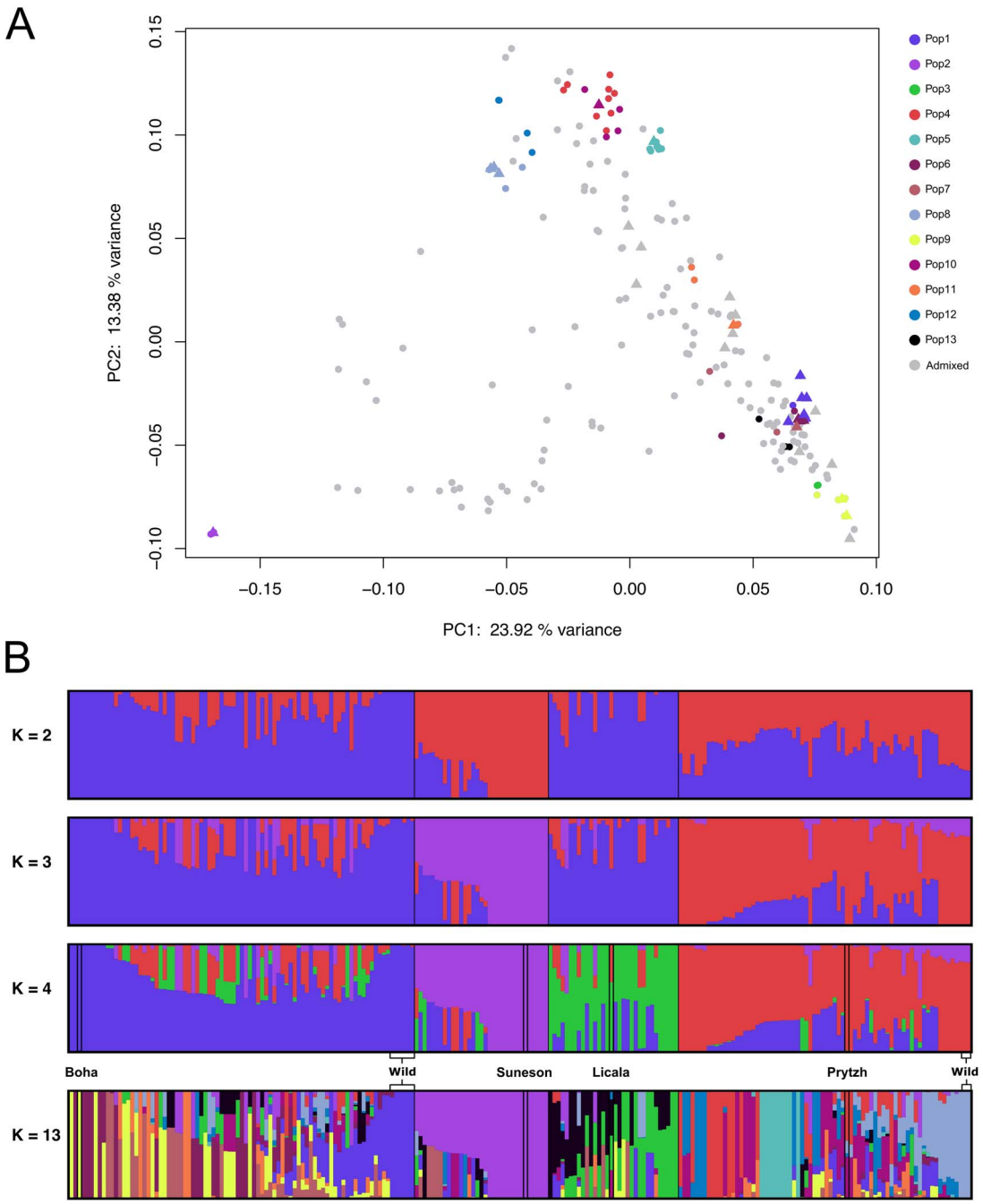

Figure 2. Analysis of genetic structure of the 222 resequenced accessions of *C. sativa*. (A) Clustering of individuals in PCA space with individuals colored according to genetic population identity at $K = 13$ with wild individuals plotted as triangles. (B) Population structure output from ADMIXTURE. Notable germplasm, including wild-collected individuals are highlighted at $K = 4$ and $K = 13$.

and also found an estimate of $\pi = 0.0013$ [1]. We predict that our estimate of nucleotide diversity may be lower due to two potential factors: (i) Genome quality and reduction in error rate of the new genome assembly resulting in fewer erroneous SNPs being called or (ii) more complete read mapping to previously unsequenced regions of the genome or areas where gaps were filled. We found that the *C. hispida* subgenome (SG3) of *C. sativa* has substantially lower genetic diversity ($\pi = 0.0007857$) compared to the other subgenomes but also the highest level of heterozygosity ($H = 0.0363$) (Supplemental Fig. S13). We suspect that this could be explained by differences in selective constraints between subgenomes such that the dominant subgenome was subjected to more large selective sweeps post-polyploidization. This is consistent with a previous study that showed that the *F. vesca* subgenome of *F.* × *ananassa* was the least genetically diverse and also the transcriptionally dominant subgenome across a panel of *F.* × *ananassa* accessions [46].

Despite being somewhat genetically depauperate, we found the lowest cross-validation (CV) error at the population cluster $K = 13$ (Supplemental Fig. S5). Previous studies in *C. sativa*, and its predomesticate, *C. microcarpa*, had found lower *K*-values to be optimal including $K = 2$ and $K = 4$ [47], $K = 4$ and $K = 8$ [41], $K = 3$ [1, 16]. It is plausible that the high optimal *K* value obtained here does not indicate the existence of distinct ancestral populations, but

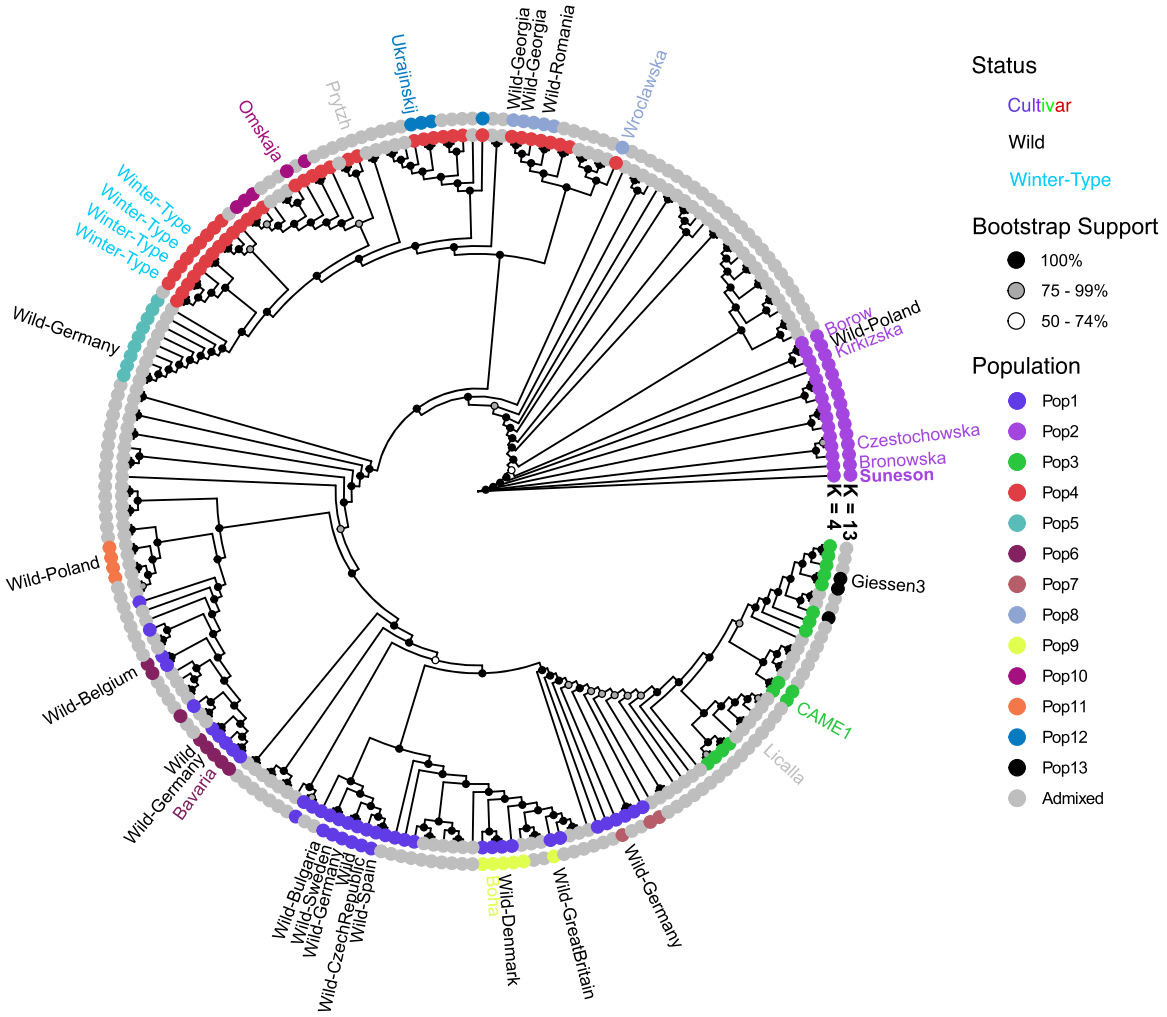

Figure 3. Phylogenetic tree of the 222 resequenced *C. sativa* lines colored by population identity when >75% identity to a single population, otherwise individuals were considered admixed and colored gray. Tip labels indicate select cultivars (colored according to population ID), wild-collected individuals (black), and winter-type cultivars (light-blue).

instead diverse founder individuals or groups that were instrumental in breeding programs. Much of the germplasm in the set of 222 *C. sativa* is derived from breeding material and is expected to be either admixed or highly inbred with a smaller set of genetically distinct cultivars. For instance, at $K = 13$, many of the clusters include well-known cultivars with complete genetic identity to their respective clusters including varieties such as Ukrajinskij, Zarjasocialisma1, CAME1, Boha, Sortandinskij, Voronezh349, as well as two clusters consisting of wild populations, one Western and one Eastern (Fig. 2B). Such cultivars likely represent founders, as we have no evidence of distinct geographical or domestication histories for these lines which would explain the high number of genetic populations. The genetic population containing Suneson includes varieties such as Borow1, Bronowska, Kirkizska, Czestochowska, and breeding material from Germany, of which also contain the lowest levels of heterozygosity (Supplemental Table S5). Because Suneson is a more recent advanced cultivar, it is likely that it shares ancestry with these cultivars and breeding lines. We argue that a lower population cluster value between $K = 3$ (CV = 0.77969) or $K = 4$ (CV = 0.72540), is more representative of the number of distinct genetic populations of *C. sativa*. This would be reflective of the recent neopolyploidization ~65 Kya [15] and domestication history 6–8 Kya [1], previously reported lack of genetic diversity, and absence of distinct clustering in principal component analysis (PCA) space (Fig. 2A). However, we caution against making conclusions on the evolutionary or domestication history of *C. sativa* using these data, as much of the passport data associated with these 222 resequencing lines is either missing or potentially erroneous, especially for country of origin. Nevertheless, cultivars showing minimal to no observed admixture at $K = 13$ likely represent distinct cultivars which represent valuable targets for future breeding programs and potential assignment into heterotic groups. The relative dearth of winter-type lines present in the resequencing panel studied here, as well as others [1, 16, 47], points to the need for the collection of new wild or cultivated winter-type lines which could be used to inject additional diversity into breeding programs. Together, the insights we provide for the resequencing lines may be valuable for future assignment of heterotic groups to facilitate breeding progress in this crop.

Previous analyses of subgenome-specific expression patterns showed the *C. hispida* (SG3) subgenome of *C. sativa* to be dominantly expressed across all tissue types examined [16, 31]. These measures of expression dominance were determined as those 1:1:1 orthologs which were expressed significantly higher in one subgenome than the other two. However, our results suggest that expression dominance of SG3 is not as strong as previously reported, and may not be universal across all tissue-types. When employing methods for determining homoeolog expression

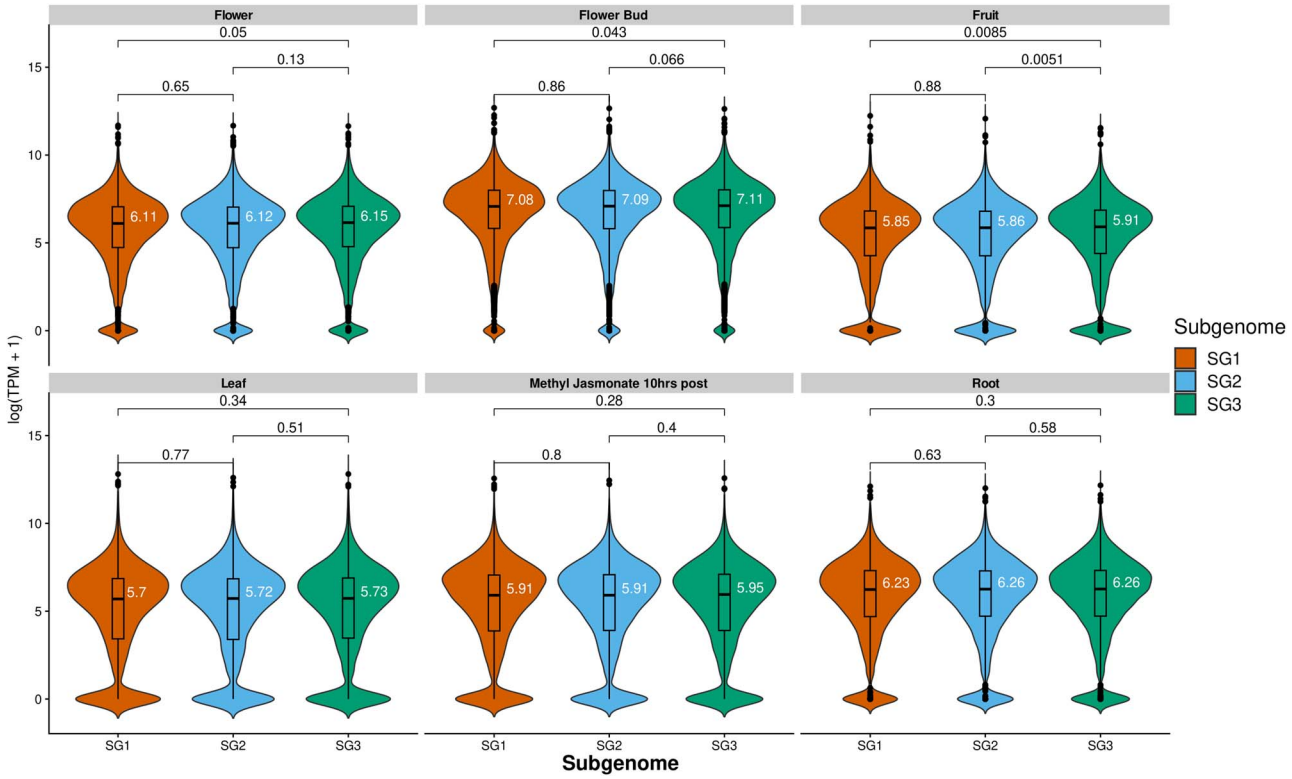

Figure 4. Expression of camelina syntelogs maintained in a 1:1:1 ratio to Arabidopsis for six experiments. Median log(TPM + 1) expression (white numbers) is shown for each subgenome/tissue-type. Syntenic homoeologs for each subgenome were tested for significant differences with pair-wise wilcoxon tests, *P* values displayed above each comparison.

bias that are commonly used in other systems [48–50], we found that only in flower, flower bud, and young fruit, were SG3 genes expressed significantly higher than one or both of the other subgenomes (Fig. 4). We found no significant differences in the number of biased homoeologs between SG1 and SG2 across all tissues (Fig. 4). The tissue-dependent subgenome expression dominance described here did not result in a complete shift of subgenome expression dominance from the dominant subgenome to a submissive subgenome, but this phenomena has been observed in wheat [51] and blueberry [52] and may exist in other tissue/cell/stress types in *C. sativa*. Interestingly, where we found a bias, it was only marginally toward the SG3 subgenome (Figs 3, 4; Supplemental Figs S6, S7). These findings suggest that although subgenome dominance was found across tissues using the older versions of *C. sativa* genomes [16, 31], it might be tissue-dependent, line dependent, or some combination thereof. It is also possible that by generating a new, nearly gap-free assembly, we were able to more precisely map reads to their correct subgenomes, thus reducing false signal.

TEs and lncRNAs have the potential to elicit regulatory changes in protein-coding genes [53]. To enhance the utility of our reference genome as a resource, we annotated TEs and lncRNAs. We found only six additional novel lincRNAs when compared to a previous lncRNA annotation using the DH55 genome [45]. Additionally, we annotated 849 ASlncRNAs which are novel to this study. We observed a mostly even distribution of ASlncRNAs across the three subgenomes; however, there was a significant difference for lincRNAs (χ^2, $p < 2.2e{-}16$), SG3 contained 837 annotated lincRNAs, versus 626 and 516 for SG2 and SG1, respectively. The TE annotation revealed that SG3 also contained substantially more TEs relative to the other two subgenomes, although these values were in line with the values annotated for TEs from the diploid progenitor species (Supplemental Table S11), and those found in a previous study [20].

C. sativa, a rising biofuel crop, is gaining renewed attention. Understanding its subgenome dominance and genetic diversity is crucial for breeding advancements. Through population genetics and transcriptome analyses, we found low genetic diversity, with the SG3 subgenome notably less diverse. Despite this, we identified thirteen distinct subpopulations, including two distinct wild population, in the surveyed diversity population. Additionally, while SG3 was previously thought to be dominantly expressed, our findings suggest a more nuanced picture, with its dominance being largely restricted to floral and fruit organs, offering valuable insights for future breeding strategies in camelina.

Methods

Sequencing, assembly, and annotation

Seeds of *C. sativa* "Suneson" were provided by Yield10 Bioscience. High molecular weight DNA was isolated from *C. sativa* "Suneson" fresh young leaves at University of Delaware. A total of 32 GBase of Pacbio HiFi genomic long read data (~43× depth) was assembled using Hifiasm [54] version 0.17.3 with options—hg-size 720 m -k 61—n-hap 6. The likely genome size (—hg-size 720 m) was based on an initial estimate, but changing this variable did result in a notable change in assembly size. The resulting primary assembly had a size of 661 Mb with a scaffold/contig N50 of 14.2 Mb. Scaffolds were organized according to their synteny with the layout generated by Kagale et al. (GCF_000633955) using Ragtag [55] version 2.1 in correct and scaffold modes with -f 100 000 and remove-small flags. BUSCO [56] version 5.3.2 analysis with the brassicales_odb10 database estimated the assembly to be 99.7% complete. HiC data was used to validate that there were

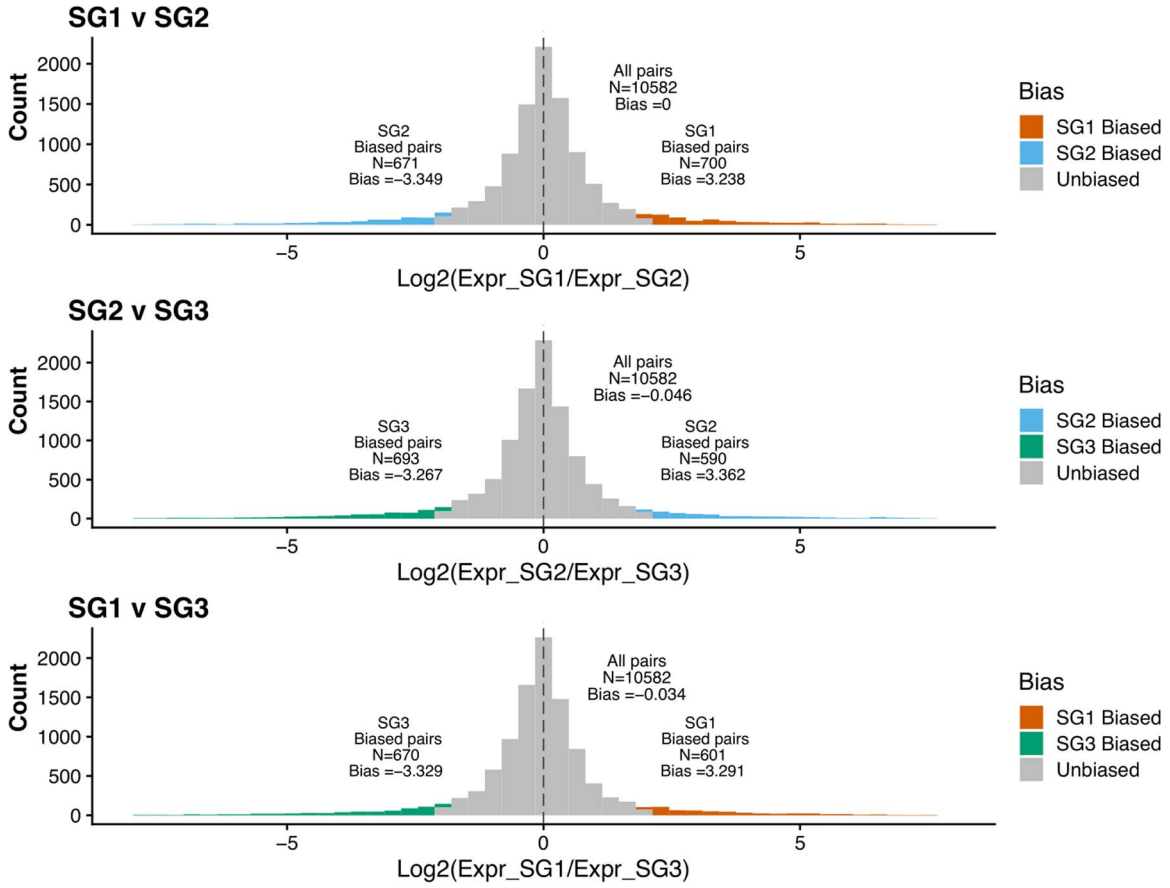

Figure 5. Homoeolog expression bias of 1:1:1 syntelogs of *C. sativa* measured pair-wise between camelina subgenomes from RNAseq data generated from flower tissue. Biased homoeologs with log2 expression fold difference greater than |2| are colored: Orange (SG1, *C. neglecta* subgenome), blue (SG2, *C. neglecta*-like subgenome), green (SG3, *C. hispida* subgenome).

no structural incompatibilities between the genomic layout of the DH55 line employed by Kagale *et al.* and our assembly of Suneson by aligning reads with the Burrow–Wheeler Aligner and filtering the resulting sam file for uniquely matching reads.

To annotate the assembly, total RNA was generated from mixed-tissues including seed (early, mid, mid-late, germinating), leaf, old-leaf, root, flower, fruit (young and old), seedling, flower-bud, and stem using the PureLink RNA Mini Kit (Invitrogen, Waltham, MA). Individual libraries were prepared as half-volume reactions for each tissue and sequenced deeply (1.4 billion reads) on the short-read illumina Novaseq 6000 platform using the Illumina Stranded mRNA Library Preparation, Ligation Kit (Illumina, San Diego, CA) with IDT for Illumina Unique Dual Index adapters (IDT, Coralville, IA) (see Supplemental Table S7). Additionally, RNA was pooled from all tissue types and sequenced using the PacBio long-read IsoSeq platform (9.4 million reads). IsoSeq data was aligned using minimap2 with options "-ax splice:hq -uf", converting and sorting the bam file with SAMtools version 1.3, while illumina data was aligned with STAR [57] version 2.7.9a and initially assembled using stringTie [58] version 2.2.1 in mixed long and short read mode. Maker [59] version 2.31.10 was then used to combine the stringtie models with EST and protein evidence from GCF_000633955, a custom repeat database (repeatmodeler, https://www.repeatmasker.org/RepeatModeler/) and SNAP, Augustus and Genemark HMM models. SNAP and Augustus were trained on DH55 gene models (https://vcru.wisc.edu/simonlab/bioinformatics/programs/augustus/docs/tutorial2015/training.html) and Genemark used the generic Eukaryotic HMM. BUSCO completeness in transcriptome mode relative to brassicales_odb10 was 99.6%. The resulting annotation was further processed through the defusion pipeline using default settings (https://wjidea.github.io/defusion/Introduction.html) to split tandem fused tandem duplicates and fused chimeric genes. This output was cleaned using the agat utility agat_convert_sp_gxf2gxf.pl (https://www.doi.org/10.5281/zenodo.3552717) and renamed. Analysis of BUSCO completeness was also conducted for the assembly and annotation of the DH55 v2 genome (Table 1).

Analysis of genome resequencing data

Genome resequencing data previously published by Li *et al.* 2021 was downloaded from NCBI. This study had identified and removed several identical accessions from their analyses, and thus we followed their selection of 222 nonredundant accessions for our analyses of resequencing data. Briefly, a total of 222 samples were previously generated with 150 bp paired-end reads with an average coverage depth of ~35 × (Supplemental Table S3). For mapping and filtering of the resequencing data, we largely followed the methods provided by the previous study [41], with a few modifications. Reads were trimmed using Trimmomatic [60] v. 0.38 keeping only reads >50 bp and bases of quality $q > 20$. The Burrows–Wheeler alignment tool [61] was then used to align reads to a draft *C. sativa* "Suneson" genome using default parameters. Some of the scaffolds represent the chloroplast genome; however, downstream analyses of genetic diversity focused on only nuclear chromosomes. The alignment files were

then filtered with SAMtools v1.9 to remove any regions with a coverage depth of more than one logarithmic scale higher than the average depth so as to remove any regions that include simple sequence repeats or abnormal mapping rates. Two mpileup files for all samples were then generated with SAMtools, one without quality filtering and the other excluding secondary alignments and filtering to include only sites with a minimum mapping quality (q) of 20 and a minimum base quality (Q) of 30. We then called SNPs from the mpileup file using the VarScan [62] mpileup2snp function. Without quality filtering, 5.00 million SNPs were recovered while the quality filtered dataset contained 3.98 million SNPs. We focused on the quality filtered dataset to conduct all further analyses. VCFtools [63] was used to filter the resulting SNPs (3.98 million) to include only biallelic SNPs which have <0.5 heterozygosity and < 10% missing data, leaving 3.89 million SNPs. We applied a linkage disequilibrium filter of r2 < 0.4 in PLINK [64] v1.9 after which 936 131 SNPs remained. Finally, PLINK was used to filter out all sites with minor allele frequencies over 0.1 resulting in the final filtered dataset (138 469 SNPs). Nucleotide diversity and fixation index (Fst) metrics were calculated using VCFtools.

Population genetic structure was assessed with ADMIXTURE v1.3.0 [65] on the final SNP set using *K* values 1–32. Cross-validation (CV) scores obtained by ADMIXTURE were used to identify the optimal number of subpopulations based on the lowest CV score (Supplemental Fig. S5). Population structure results obtained from ADMIXTURE were plotted using the visualization program pong [66]. Filtered SNPs were also used to visualize genetic clustering in a PCA using R. Measures of subpopulation genetic differentiation (*Fst*) were calculated with VCFtools. A tree was generated with IQ-TREE v. 2.1.3 [67] with the flag-m MFP to run extended model selection, which determined TVMe+R10 to be the best model, and -B 1000 for bootstrap replication.

Subgenome dominance analyses

We identified syntelogs between *A. thaliana* and *C. sativa* using SynMap [68] and QUOTA-ALIGN [69] on the CoGe platform (genomevolution.org) and filtered to retain only 1:1:1 syntelogs with Arabidopsis, resulting in 11 525 syntenic triplets. These syntelogs groups were further filtered to ensure each subgenome was represented once for each set, reducing the final number of 1:1:1 syntelogs 11 269. We removed Illumina Truseq 3 adapter sequences from the raw RNAseq reads with Trimmomatic v 0.39 (ILLUMINACLIP:TruSeq3-PE-2.fa:2:30:10:8:True) [55]. To ensure high confidence mapping to homoeologous genes, we aligned the RNAseq to the "Suneson" reference genome with STAR v 2.6.0 [57] with stringent filtering which excluded alignments which had more than one mismatch across the entire read length and removed reads which mapped to more than one position. For expression quantification, all analyses were performed on the subgenomes separately, defined by physical chromosomes, to account for size differences between subgenomes. Transcript abundance was quantified with StringTie v 2.1.2 [70] and abundance was converted to transcripts per million (TPM) with tximport [71].

To identify homoeolog expression bias, we analyzed expression of our 1:1:1 syntelogs across six experiments: leaf, flowers, flower buds, young fruits, roots, and leaves 10 hours post methyl jasmonate treatment. Visualization and identification of homoeolog expression bias was performed in R v 4.3.0 (R Core team 2023). Homoeolog expression bias was defined as a log2 fold change expression difference greater than |2| in pairwise comparisons of homoeologs from the three subgenomes. Significant differences in median TPM was determined with pair-wise wilcoxon tests between subgenomes with a *p*-value cutoff of 0.05 and using the R package ggpubr [72]. Significant biases in the number of homoeologs exhibiting biased expression was determined using a chi-squared test of observed proportions of biased homoeologs against the expectation of equal proportion of biased homoeologs between subgenomes. Final visualizations were done with the ggplot2 [73] and cowplot [74] packages in R. Gene ontology enrichment analyses on biased genes were done using the R package TopGO [75] for biological processes (algorithm = "elim" & statistic = "fisher").

Annotating putative protein-coding genes and long non-coding RNAs

Publicly available Illumina RNA-sequencing data for *C. sativa* (PRJEB49403, PRJNA397728, PRJNA231618) and Nanopore RNA-sequencing (PRJNA765684) were downloaded from the NCBI SRA. Raw Illumina reads were mapped to the camelina genome as in [45] using Hisat2 v 2.2.1 [76] with the following arguments: "–max-intronlen 10 000", "–dta-cufflinks", and "–rna-strandness" with the appropriate strand parameter. A Hisat2 index was built before aligning reads with exon and splice site coordinates from annotated protein coding transcripts along with the genome sequence file. Nanopore reads were mapped to the camelina genome using Minimap2 v 2.17-r941 [77] with the following parameters: "-ax splice", and "-G 10000".

Transcript assembly for Illumina sequencing was performed using Stringtie v 2.2.1 [70] with the following parameters: -f 0.05, –j 5, –c 5, –s 10, along with the appropriate strand parameter. Transcript assembly for Nanopore sequencing was also performed with Stringtie using the following parameters: –fr, –L, –f 0.05, –j 5, –c 5, and -s 10. All Stringtie outputs from Illumina and Nanopore sequencing were merged using Stringtie "merge" with the following parameters: -m 200, –c 5, and -f 0.05. Transcript assembly for both Illumina and Nanopore sequencing was performed with a reference annotation of protein-coding genes using the -G option.

Newly assembled transcripts were classified relative to annotated protein-coding genes using Gffcompare v 0.12.2 [78]. Transcripts antisense to protein-coding genes were identified using the Gffcompare classification code "x" while intergenic transcripts were identified using the classification code "u". Single-exon intergenic transcripts that could not be assigned to the forward or reverse strand were discarded. Antisense and intergenic transcripts were processed through the Coding Potential Calculator 2 (CPC2) webserver at http://cpc2.gao-lab.org/index.php. Transcripts were separated on the basis of the "noncoding" or "coding" classification label assigned by CPC2, with transcripts scored at <0.5 retained for further analysis. Non-coding intergenic transcripts (putative long intergenic non-coding RNAs – lincRNAs) were further filtered for other "housekeeping" RNAs (e.g., ribosomal RNA, small nuclear/nucleolar RNAs, etc.) using the RNA families (Rfam) [79] webserver. All transcripts with hits to the Rfam database were removed. Finally, antisense and intergenic transcripts that were classified as "coding" by CPC2 were scanned for putative protein domains and families. All open-reading frames in the forward frame were identified and translated using the EMBOSS getorf tool [80]. All translated proteins were searched for protein domains and families using PfamScan (http://www.ebi.ac.uk/Tools/pfa/pfamscan) along with HMMER v 3.3.2 [81].

To compare the annotation of putative protein-coding genes and long non-coding RNAs between the DH55 genome assembly (Ensembl - Camelina_sativa.Cs.54.gff3) and our assembly, the

aforementioned publicly available datasets were mapped using the Hisat2 parameters with the currently available camelina genome on Ensembl [82]. Transcript assembly and classification was performed with the same Stringtie and Gffcompare parameters, respectively. Antisense and intergenic transcripts were processed through the command-line version of CPC2 (v1.0.1) and separated according to "noncoding" or "coding" classification label. Congruently, non-coding intergenic transcripts (or putative lincRNAs) were further filtered for other "housekeeping" RNAs using the Rfam webserver. All transcripts with hits to the Rfam database were removed. In addition to using CPC2, we also used the Coding-Potential Assessment Tool (CPAT) [83] to determine the reliability of identified lncRNAs and to reduce false positives.

Final putative protein-coding genes and long non-coding RNA annotations were generated in command-line and submitted to Liftoff—using default parameters v1.6.3 [84]. Liftoff outputs were used to determine differences in annotation of putative protein-coding genes and long noncoding RNAs between the DH55 genome assembly and our assembly. R code available at https://rpubs.com/cer246/1049254.

Pangenome annotation of TEs

TEs were annotated *de novo* via a pangenome approach. The genomes of *C. hispida* (GCA_023864115 [20]), *C. laxa* (GCA_024034495 [20]), *C. neglecta* (GCA_029034625 [30]), and *A. thaliana* (Araport11) were downloaded and along with the *C. sativa* Suneson genome, panEDTA v2.1.2 [85] was run with default parameters. This approach generates individual genome annotations with EDTA and then creates a common pangenome repeat library to finally re-annotate each genome. The scripts associated with this analysis can be found at https://github.com/sjteresi/Camelina_TE_Annotation.

Acknowledgements

This work was supported by the Department of Energy Office of Biological and Environmental Research (Grant no. DE-SC0022987 to E.G. and P.P.E.), National Science Foundation (NSF) Postdoctoral Research Fellowship in Biology (PRFB-2109178 to J.R.B.; PRFB-2208944 to K.A.B), National Science Foundation Plant Genome (PGRP-2029959 to P.P.E.), National Science Foundation (IOS-2023310 and NSF DBI-2243562 to A.D.L.N).

Author contributions

P.P.E, E.G., and J.R.B. conceived the project. J.R.B., S.K.G., and E.G.P. collected samples and S.K.P. and M.M.L. extracted DNA and RNA for sequencing. J.R.B analyzed camelina resequencing data, K.A.B. analyzed camelina homoeolog-expression bias, A.E.P. assembled and annotated the genome, S.J.T. annotated TEs, K.P. and C.E.R. annotated lncRNAs. Y.S.L, J.R.B., F.G.C and S.K.G. manually curated target genes. J.R.B. P.P.E., A.D.L.N., and E.G. wrote the manuscript. All authors gave feedback and comments on the final manuscript version.

Data availability

The reference genome and transcriptome of *C. sativa* variety "Suneson" will be publicly available after publication at CamRegBase (https://camregbase.org/) and CoGe (https://genomevolution.org/coge/). Raw RNAseq reads and HiFi genomic and transcriptomic reads have been submitted to NCBI SRA under accession numbers SRX25825928-SRX25825946, SRX25826254, SRX25827247, and SRX25827248.

Conflict of interest

The authors declare no conflict of interest.

Supplementary Data

Supplementary data is available at *Horticulture Research* online.

References

1. Brock JR, Ritchey MM, Olsen KM. Molecular and archaeological evidence on the geographical origin of domestication for *Camelina sativa*. *Am J Bot*. 2022;**109**:1177–90
2. Pilgeram AL, Sands DC, Boss D. *et al. Camelina sativa, a Montana Omega-3 and Fuel Crop*. 2007.
3. Berhow MA, Polat U, Glinski JA. *et al.* Optimized analysis and quantification of glucosinolates from *Camelina sativa* seeds by reverse-phase liquid chromatography. *Ind Crop Prod*. 2013;**43**:119–25
4. Brock JR, Scott T, Lee AY. *et al.* Interactions between genetics and environment shape Camelina seed oil composition. *BMC Plant Biol*. 2020;**20**:423
5. Shonnard DR, Williams L, Kalnes TN. Camelina-derived jet fuel and diesel: sustainable advanced biofuels. *Environ Prog Sustain Energy*. 2010;**29**:382–92
6. Séguin-Swartz G, Eynck C, Gugel RK. *et al.* Diseases of *Camelina sativa* (false flax). *Can J Plant Pathol*. 2009;**31**:375–86
7. Augustin JM, Brock JR, Augustin MM. *et al.* Field performance of terpene-producing *Camelina sativa*. *Ind Crop Prod*. 2019;**136**:50–8
8. Augustin JM, Higashi Y, Feng X. *et al.* Production of mono- and sesquiterpenes in *Camelina sativa* oilseed. *Planta*. 2015;**242**:693–708
9. Augustin MM, Shukla AK, Starks CM. *et al.* Biosynthesis of *Veratrum californicum* specialty chemicals in *Camelina sativa* seed. *Plant Biotechnol Rep*. 2017;**11**:29–41
10. Iven T, Hornung E, Heilmann M. *et al.* Synthesis of oleyl oleate wax esters in *Arabidopsis thaliana* and *Camelina sativa* seed oil. *Plant Biotechnol J*. 2016;**14**:252–9
11. Xia Y-H, Wang H-L, Ding B-J. *et al.* Green chemistry production of Codlemone, the sex pheromone of the codling moth (*Cydia pomonella*), by metabolic engineering of the oilseed crop Camelina (*Camelina sativa*). *J Chem Ecol*. 2021;**47**:950–67
12. Barrett SH. Crop mimicry in weeds. *Econ Bot*. 1983;**37**:255–82
13. Tedin O. Vererbung, variation und Systematik in der Gattung Camelina. *Hereditas*. 1925;**6**:275–386
14. Zinger HB. On the species of Camelina and Spergularia occurring as weeds in sowings of flax and their origin. *Trudy Bot Muz Imp Akad Nauk*. 1909;**6**:1–303
15. Brock JR, Mandáková T, McKain M. *et al.* Chloroplast phylogenomics in Camelina (Brassicaceae) reveals multiple origins of polyploid species and the maternal lineage of *C. sativa*. *Hortic Res*. 2022;**9**:9
16. Chaudhary R, Koh CS, Kagale S. *et al.* Assessing diversity in the Camelina Genus provides insights into the genome structure of *Camelina sativa*. *G3 (Bethesda)*. 2020;**10**:1297–308
17. Mandáková T, Pouch M, Brock JR. *et al.* Origin and evolution of diploid and allopolyploid Camelina genomes were accompanied by chromosome shattering. *Plant Cell*. 2019;**31**:2596–612
18. Mandáková T, Lysak MA. The identification of the missing maternal genome of the allohexaploid camelina (*Camelina sativa*). *Plant J*. 2022;**112**:622–9
19. Zhang Z, Meng F, Sun P. *et al.* An updated explanation of ancestral karyotype changes and reconstruction of evolutionary

trajectories to form *Camelina sativa* chromosomes. *BMC Genomics*. 2020;**21**:705
20. Martin SL, Lujan Toro B, James T. *et al.* Insights from the genomes of 4 diploid Camelina spp. *G3 (Bethesda, Md)*. 2022;**12**:jkac182
21. Blume RY, Kalendar R, Guo L. *et al.* Overcoming genetic paucity of *Camelina sativa*: possibilities for interspecific hybridization conditioned by the genus evolution pathway. *Front Plant Sci*. 2023;**14**:1259431
22. Beilstein MA, Al-Shehbaz IA, Kellogg EA. Brassicaceae phylogeny and trichome evolution. *Am J Bot*. 2006;**93**:607–19
23. Beilstein MA, Al-Shehbaz IA, Mathews S. *et al.* Brassicaceae phylogeny inferred from phytochrome A and ndhF sequence data: tribes and trichomes revisited. *Am J Bot*. 2008;**95**:1307–27
24. Hohmann N, Wolf EM, Lysak MA. *et al.* A time-calibrated road map of Brassicaceae species radiation and evolutionary history. *Plant Cell*. 2015;**27**:2770–84
25. Liu X, Brost J, Hutcheon C. *et al.* Transformation of the oilseed crop *Camelina sativa* by Agrobacterium-mediated floral dip and simple large-scale screening of transformants. *In Vitro Cell Dev Biol Plant*. 2012;**48**:462–8
26. Lu C, Kang J. Generation of transgenic plants of a potential oilseed crop *Camelina sativa* by agrobacterium-mediated transformation. *Plant Cell Rep*. 2008;**27**:273–8
27. Gomez-Cano F, Carey L, Lucas K. *et al.* CamRegBase: a gene regulation database for the biofuel crop, *Camelina sativa*. *Database*. 2020;**2020**:baaa075
28. Gomez-Cano F, Chu Y-H, Cruz-Gomez M. *et al.* Exploring *Camelina sativa* lipid metabolism regulation by combining gene co-expression and DNA affinity purification analyses. *Plant J*. 2022;**110**:589–606
29. Wang S, Blume RY, Zhou Z-W. *et al.* Chromosome-level assembly and analysis of *Camelina neglecta*: a novel diploid model for Camelina biotechnology research. *Biotechnol Biofuels Bioprod*. 2024;**17**:17
30. Chaudhary R, Koh CS, Perumal S. *et al.* Sequencing of *Camelina neglecta*, a diploid progenitor of the hexaploid oilseed *Camelina sativa*. *Plant Biotechnol J*. 2023;**21**:521–35
31. Kagale S, Koh C, Nixon J. *et al.* The emerging biofuel crop *Camelina sativa* retains a highly undifferentiated hexaploid genome structure. *Nat Commun*. 2014;**5**:3706
32. Michael TP, VanBuren R. Building near-complete plant genomes. *Curr Opin Plant Biol*. 2020;**54**:26–33
33. Edger PP. The power of chromosome-scale, haplotype-resolved genomes. *Mol Plant*. 2022;**15**:393–5
34. Ozseyhan ME, Kang J, Mu X. *et al.* Mutagenesis of the FAE1 genes significantly changes fatty acid composition in seeds of *Camelina sativa*. *Plant Physiol Biochem*. 2018;**123**:1–7
35. Na G, Mu X, Grabowski P. *et al.* Enhancing microRNA167A expression in seed decreases the α-linolenic acid content and increases seed size in *Camelina sativa*. *Plant J*. 2019;**98**:346–58
36. Li J, Su Y, Shapiro CA. *et al.* Phosphate deficiency modifies lipid composition and seed oil production in Camelina. *Plant Sci*. 2023;**330**:111636
37. Bengtsson JD, Wallis JG, Bai S. *et al.* The coexpression of two desaturases provides an optimized reduction of saturates in Camelina oil. *Plant Biotechnol J*. 2023;**21**:497–505
38. Fang C, Hamilton JP, Vaillancourt B. *et al.* Cold stress induces differential gene expression of retained homeologs in *Camelina sativa* cv Suneson. *Front Plant Sci*. 2023;**14**:1271625
39. Krzywinski M, Schein J, Birol I. *et al.* Circos: an information aesthetic for comparative genomics. *Genome Res*. 2009;**19**:1639–45
40. Slotte T, Hazzouri KM, Ågren JA. *et al.* The Capsella rubella genome and the genomic consequences of rapid mating system evolution. *Nat Genet*. 2013;**45**:831–5
41. Li H, Hu X, Lovell JT. *et al.* Genetic dissection of natural variation in oilseed traits of camelina by whole-genome resequencing and QTL mapping. *Plant Genome*. 2021;**14**:e20110
42. Cheng C-Y, Krishnakumar V, Chan AP. *et al.* Araport11: a complete reannotation of the *Arabidopsis thaliana* reference genome. *Plant J*. 2017;**89**:789–804
43. Gullotta G, Korte A, Marquardt S. Functional variation in the non-coding genome: molecular implications for food security. *J Exp Bot*. 2023;**74**:2338–51
44. Palos K, Yu L, Railey CE. *et al.* Linking discoveries, mechanisms, and technologies to develop a clearer perspective on plant long noncoding RNAs. *Plant Cell*. 2023;**35**:1762–86
45. Palos K, Nelson Dittrich AC, Yu L. *et al.* Identification and functional annotation of long intergenic non-coding RNAs in Brassicaceae. *Plant Cell*. 2022;**34**:3233–60
46. Hardigan MA, Lorant A, Pincot DDA. *et al.* Unraveling the complex hybrid ancestry and domestication history of cultivated strawberry. *Mol Biol Evol*. 2021;**38**:2285–305
47. Luo Z, Brock J, Dyer JM. *et al.* Genetic diversity and population structure of a *Camelina sativa* spring panel. *Front Plant Sci*. 2019;**10**:184
48. Woodhouse MR, Cheng F, Pires JC. *et al.* Origin, inheritance, and gene regulatory consequences of genome dominance in polyploids. *Proc Natl Acad Sci USA*. 2014;**111**:5283–8
49. Edger PP, Smith R, McKain MR. *et al.* Subgenome dominance in an interspecific hybrid, synthetic allopolyploid, and a 140-year-old naturally established neo-allopolyploid monkeyflower. *Plant Cell*. 2017;**29**:2150–67
50. Bird KA, Niederhuth CE, Ou S. *et al.* Replaying the evolutionary tape to investigate subgenome dominance in allopolyploid *Brassica napus*. *New Phytol*. 2021;**230**:354–71
51. Pfeifer M, Kugler KG, Sandve SR. *et al.* Genome interplay in the grain transcriptome of hexaploid bread wheat. *Science*. 2014;**345**:1250091
52. Colle M, Leisner CP, Wai CM. *et al.* Haplotype-phased genome and evolution of phytonutrient pathways of tetraploid blueberry. *Gigascience*. 2019;**8**:giz012
53. Schmitz RJ, Grotewold E, Stam M. Cis-regulatory sequences in plants: their importance, discovery, and future challenges. *Plant Cell*. 2022;**34**:718–41
54. Cheng H, Concepcion GT, Feng X. *et al.* Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. *Nat Methods*. 2021;**18**:170–5
55. Alonge M, Soyk S, Ramakrishnan S. *et al.* RaGOO: fast and accurate reference-guided scaffolding of draft genomes. *Genome Biol*. 2019;**20**:224
56. Manni M, Berkeley MR, Seppey M. *et al.* BUSCO: assessing genomic data quality and beyond. *Curr Protoc*. 2021;**1**:e323
57. Dobin A, Davis CA, Schlesinger F. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*. 2013;**29**:15–21
58. Shumate A, Wong B, Pertea G. *et al.* Improved transcriptome assembly using a hybrid of long and short reads with StringTie. *PLoS Comput Biol*. 2022;**18**:e1009730
59. Campbell MS, Holt C, Moore B. *et al.* Genome annotation and curation using MAKER and MAKER-P. *Curr Protoc Bioinform*. 2014;**48**:4.11.1–39
60. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*. 2014;**30**:2114–20
61. Li H, Durbin R. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics*. 2009;**25**:1754–60

62. Koboldt DC, Zhang Q, Larson DE. *et al.* VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Res*. 2012;**22**:568–76
63. Danecek P, Auton A, Abecasis G. *et al.* The variant call format and VCFtools. *Bioinformatics*. 2011;**27**:2156–8
64. Purcell S, Neale B, Todd-Brown K. *et al.* PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet*. 2007;**81**:559–75
65. Alexander DH, Novembre J, Lange K. Fast model-based estimation of ancestry in unrelated individuals. *Genome Res*. 2009;**19**:1655–64
66. Behr AA, Liu KZ, Liu-Fang G. *et al.* Pong: fast analysis and visualization of latent clusters in population genetic data. *Bioinformatics*. 2016;**32**:2817–23
67. Minh BQ, Schmidt HA, Chernomor O. *et al.* IQ-TREE 2: new models and efficient methods for phylogenetic inference in the genomic era. *Mol Biol Evol*. 2020;**37**:1530–4
68. Lyons E, Pedersen B, Kane J. *et al.* The value of nonmodel genomes and an example using SynMap within CoGe to dissect the hexaploidy that predates the rosids. *Trop Plant Biol*. 2008;**1**:181–90
69. Tang H, Lyons E, Pedersen B. *et al.* Screening synteny blocks in pairwise genome comparisons through integer programming. *BMC Bioinformatics*. 2011;**12**:102
70. Pertea M, Pertea GM, Antonescu CM. *et al.* StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat Biotechnol*. 2015;**33**:290–5
71. Soneson C, Love MI, Robinson MD. Differential analyses for RNA-seq: transcript-level estimates improve gene-level inferences. *F1000Res*. 2015;**4**:1521
72. Kassambara A, Kassambara MA. 2020 Package 'ggpubr'. *R package version 01*. https://cran.microsoft.com/snapshot/2017-02-26/web/packages/ggpubr/ggpubr.pdf.
73. Wickham H. Data analysis. In: Wickham H, ed. *ggplot2: elegant graphics for data analysis*. Cham: Springer International Publishing, 2016,189–201
74. Wilke CO. cowplot: streamlined plot theme and plot annotations for 'ggplot2'; 2020. *R package version*.
75. Alexa A, Rahnenfuhrer J 2010 topGO: enrichment analysis for gene ontology. *R package version* 2010.
76. Kim D, Paggi JM, Park C. *et al.* Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nat Biotechnol*. 2019;**37**:907–15
77. Li H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics*. 2018;**34**:3094–100
78. Pertea G, Pertea M. GFF utilities: GffRead and GffCompare. *F1000Res*. 2020;**9**:304
79. Griffiths-Jones S, Bateman A, Marshall M. *et al.* Rfam: an RNA family database. *Nucleic Acids Res*. 2003;**31**:439–41
80. Rice P, Longden I, Bleasby A. EMBOSS: the European molecular biology open software suite. *Trends Genet*. 2000;**16**:276–7
81. Potter SC, Luciani A, Eddy SR. *et al.* HMMER web server: 2018 update. *Nucleic Acids Res*. 2018;**46**:W200–4
82. Howe KL, Achuthan P, Allen J. *et al.* Ensembl 2021. *Nucleic Acids Res*. 2021;**49**:D884–91
83. Wang L, Park HJ, Dasari S. *et al.* CPAT: coding-potential assessment tool using an alignment-free logistic regression model. *Nucleic Acids Res*. 2013;**41**:e74
84. Shumate A, Salzberg SL. Liftoff: accurate mapping of gene annotations. *Bioinformatics*. 2021;**37**:1639–43
85. Ou S, Collins T, Qiu Y. *et al.* Differences in activity and stability drive transposable element variation in tropical and temperate maize bioRxiv. 2022