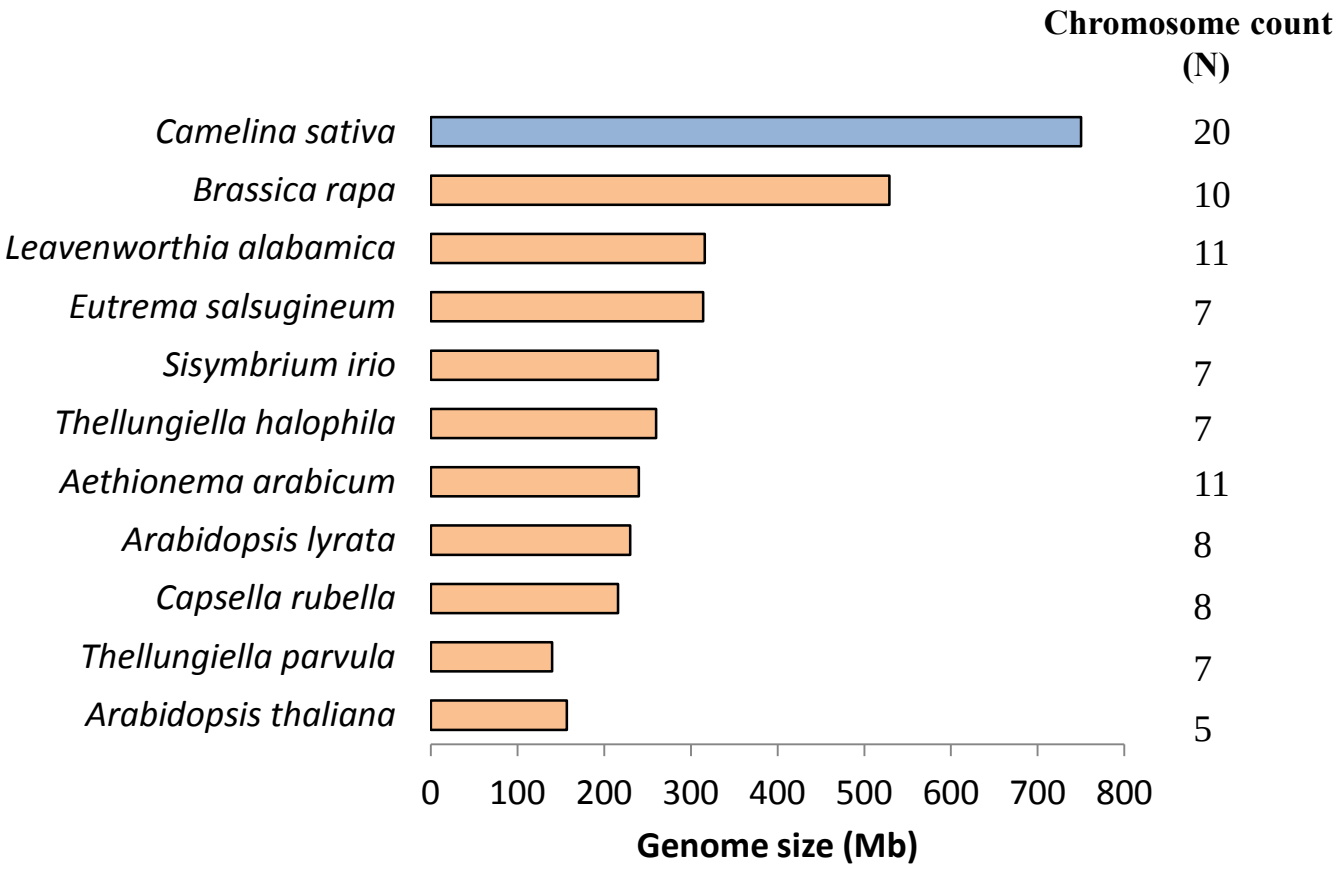
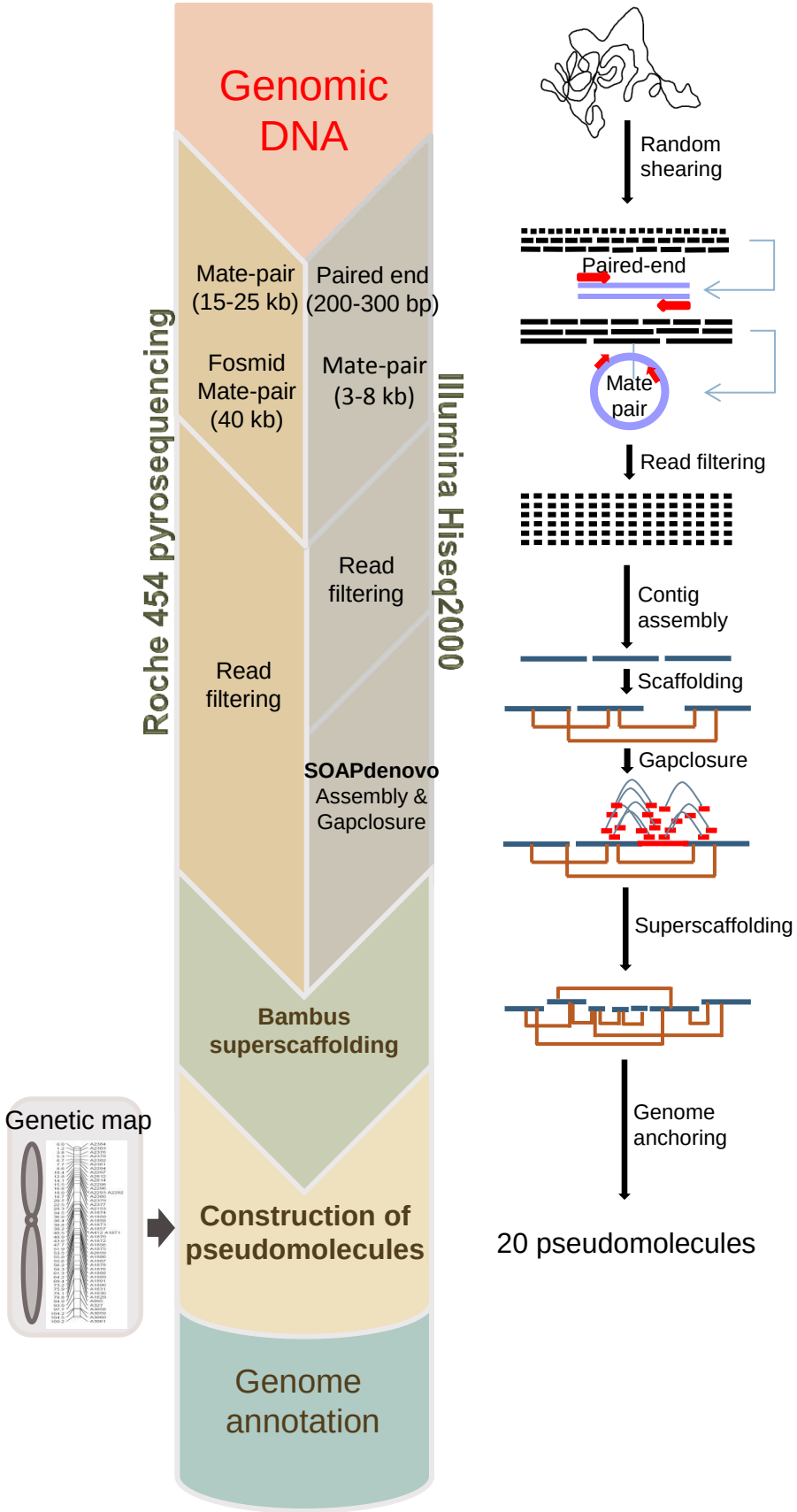
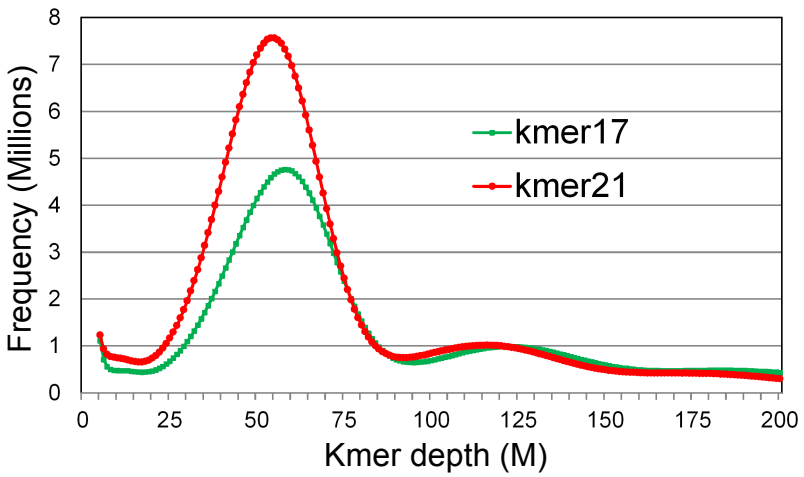


Supplementary Fig. 1. Genome size and chromosome number of Brassicaceae species. Genome size estimate for each Brassicaceae species was derived from published reports: *C. sativa*[7], *B. rapa*[29], *L. alabamica*[30], *E. salsugineum*[31], *S. irio*[29], *T. halophila*[32], *A. arabicum*[30], *A. lyrata*[29], *C. rubella*[31], *T. parvula*[33] *and A. thaliana*[29]. Base chromosome numbers were obtained from Brassibase Database[34].



Supplementary Fig. 2. *Camelina sativa* genome assembly workflow. The genome of *C. sativa* was sequenced using a hybrid Illumina and Roche 454 next generation sequencing approach. The draft genome sequence was assembled using a novel hierarchical assembly strategy employing SOAPdenovo[1] for contig assembly and scaffolding, and Bambus[2] for superscaffolding. A high density genetic linkage map was utilized for the construction of chromosome scale pseudomolecules.

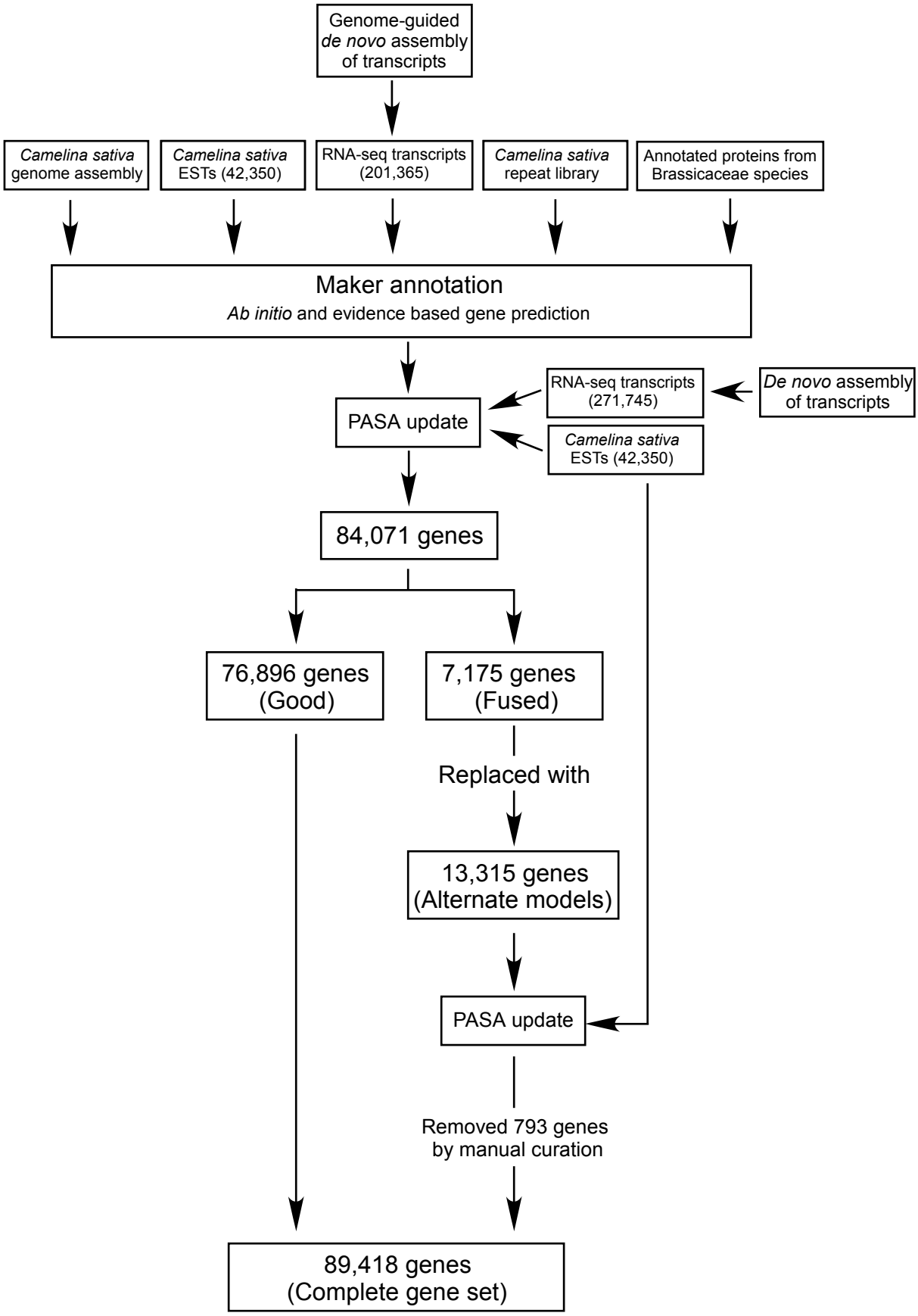
a



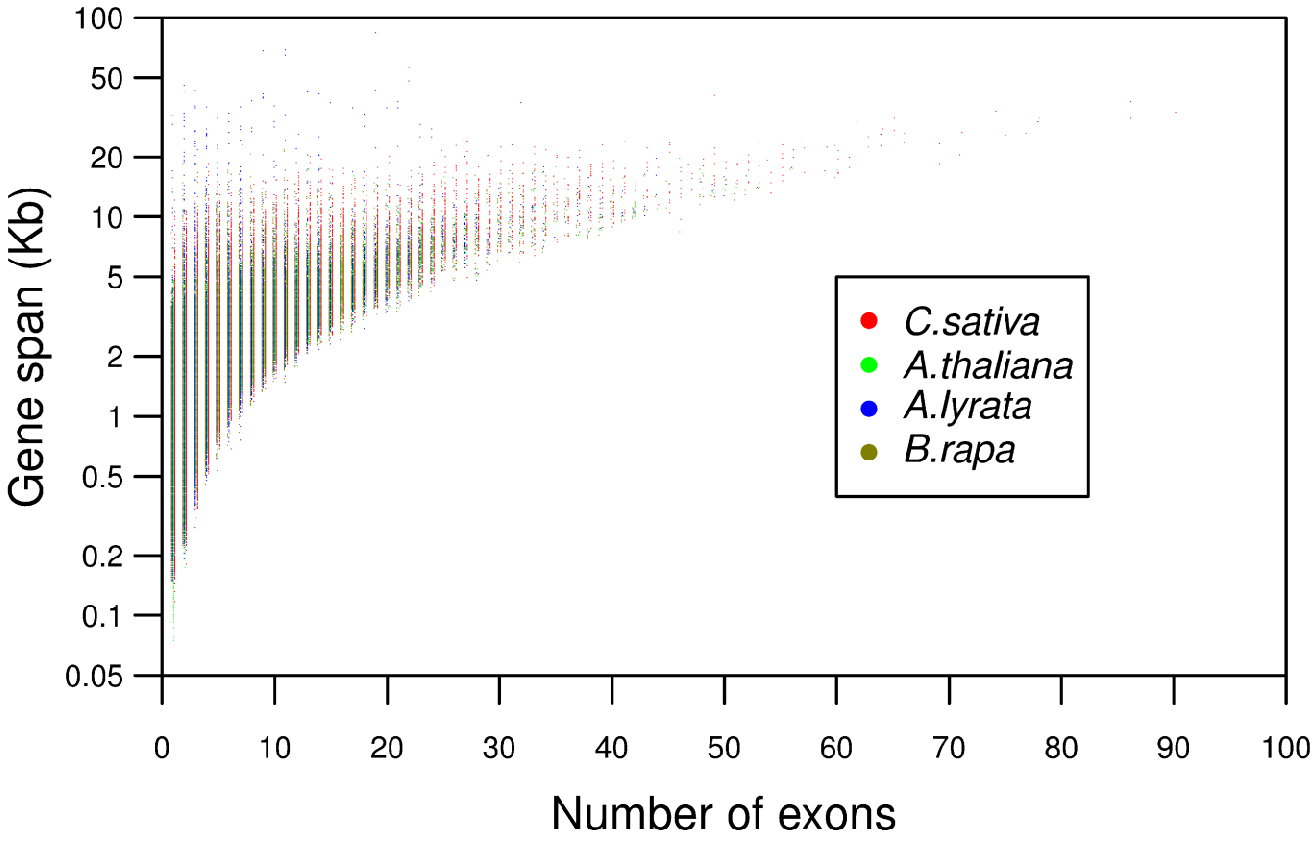
b

Kmer size (bp) (K)	Kmer Depth (M)	Read length (L)	Sequence Depth [N=(M*L)/(L-K+1)]	Total sequence length (bp) (SL)	Genome size (bp) (SL/N)
17	57.23	99.69	68.17	53,164,128,938	779,860,508
21	53.72	99.69	67.20	53,164,128,938	791,106,457
				Average	**785,483,483**

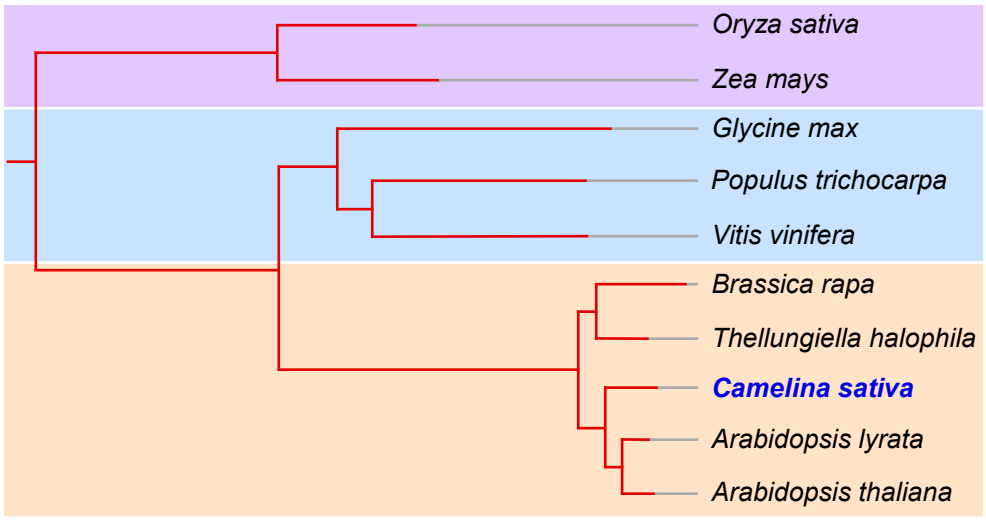
Supplementary Fig. 3. Estimation of *C. sativa* genome size based on Illumina sequencing coverage
(a) Distribution of 17- and 21mer frequency in filtered Illumina paired-end reads.
(b) Estimation of genome size based on kmer statistics. Sequencing depth (N) was estimated from the means of read length (L) and kmer frequency (M) determined by fitting a Gaussian mixture (+constant) model to the main peak. Genome size was then estimated by dividing the total sequence volume (SL) used to determine kmer frequency by sequencing depth (N).



Supplementary Fig. 4. An integrated gene annotation pipeline employed in this study.
For annotation of protein coding genes in *Camelina sativa*, two major genome annotation pipelines Maker[35] and PASA[36] were used. To improve the accuracy of gene prediction and determine intron-exon structures, external evidence such as RNAseq transcripts and protein databases from other sequenced Brassicaceae species was used. Gene models annotated using Maker/PASA were checked for accuracy and manually curated if necessary.

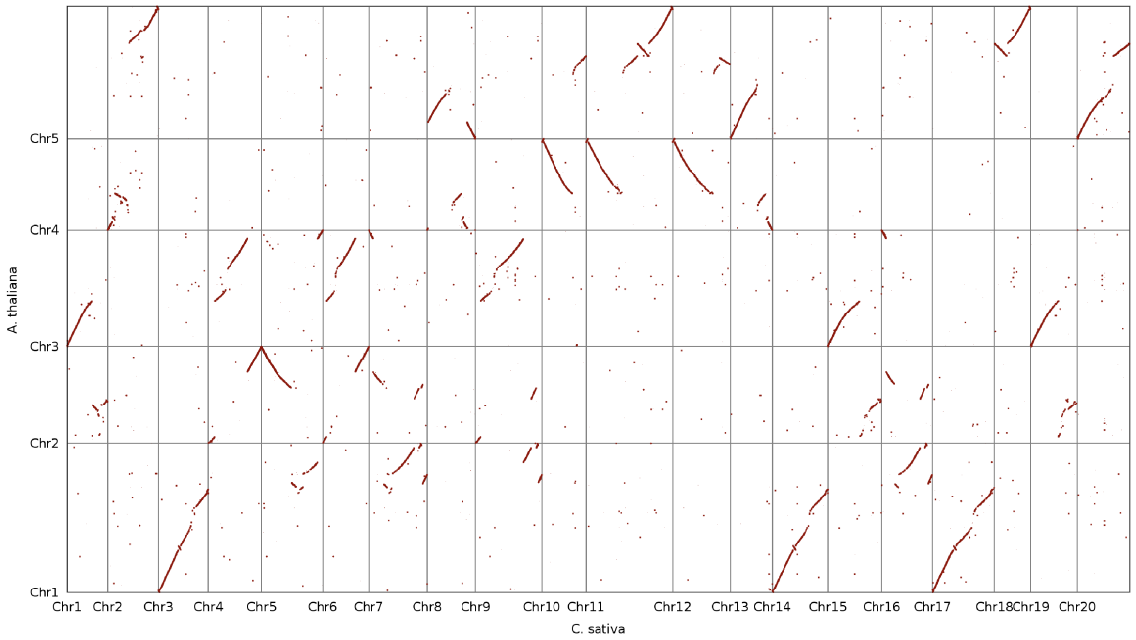


Supplementary Fig. 5. Comparison of gene model characteristics (length of coding sequences of genes plotted against their corresponding exon number) of *C. sativa* with a subset of related Brassicaceae species, including *A. thaliana*, *A. lyrata* and *B. rapa*.

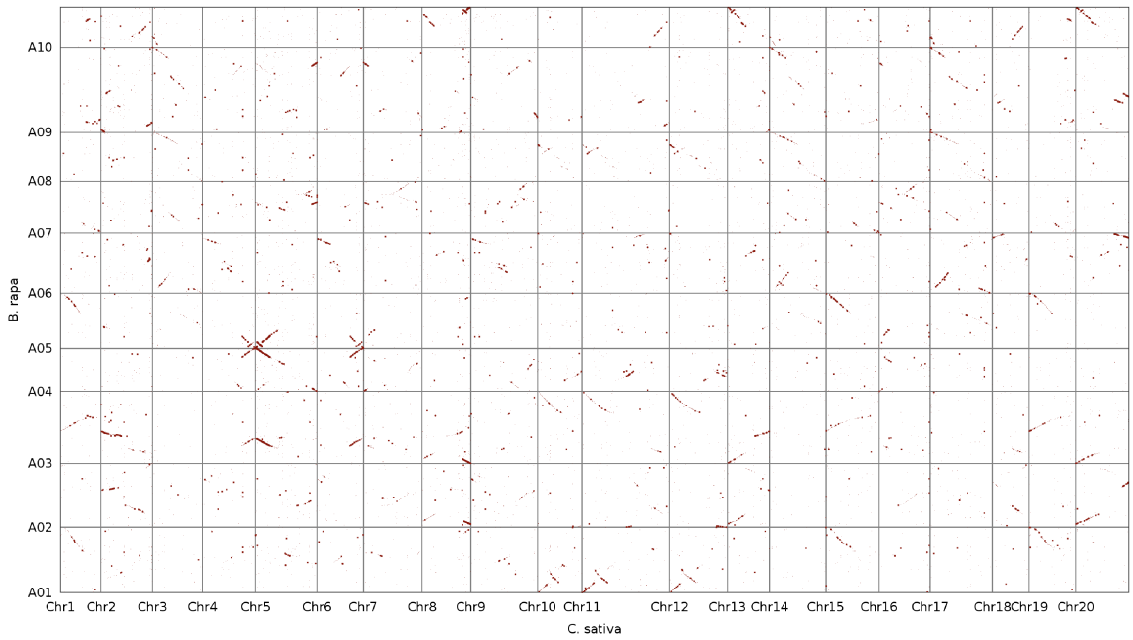


Supplementary Fig. 6. Phylogenetic relationship of *Camelina sativa* with Brassicaceae and other angiosperm species. A maximum likelihood tree produced from a supermatrix constructed based on 1,853 orthologous sequences in a concatenated alignment of 780,391 bp .

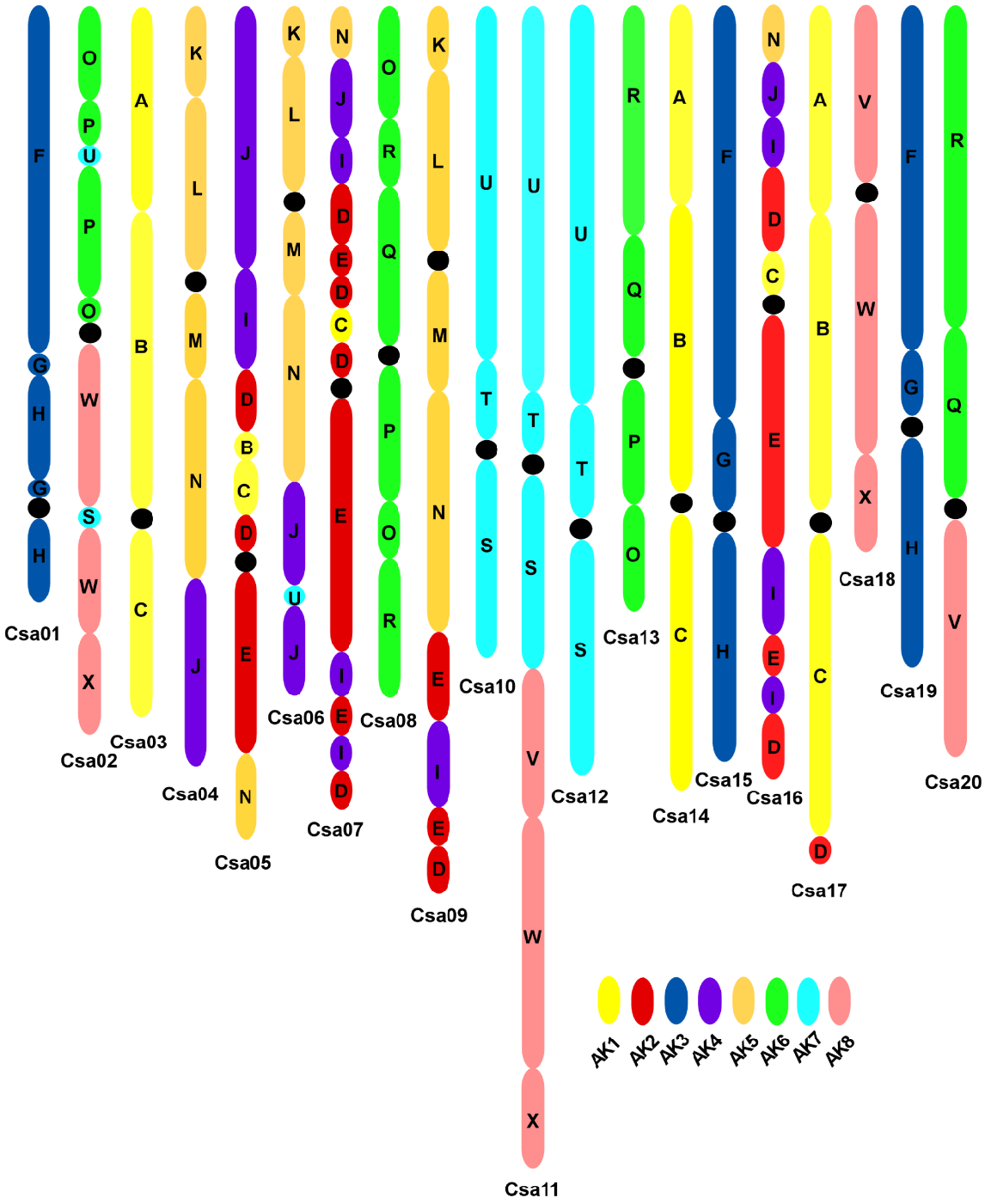
a



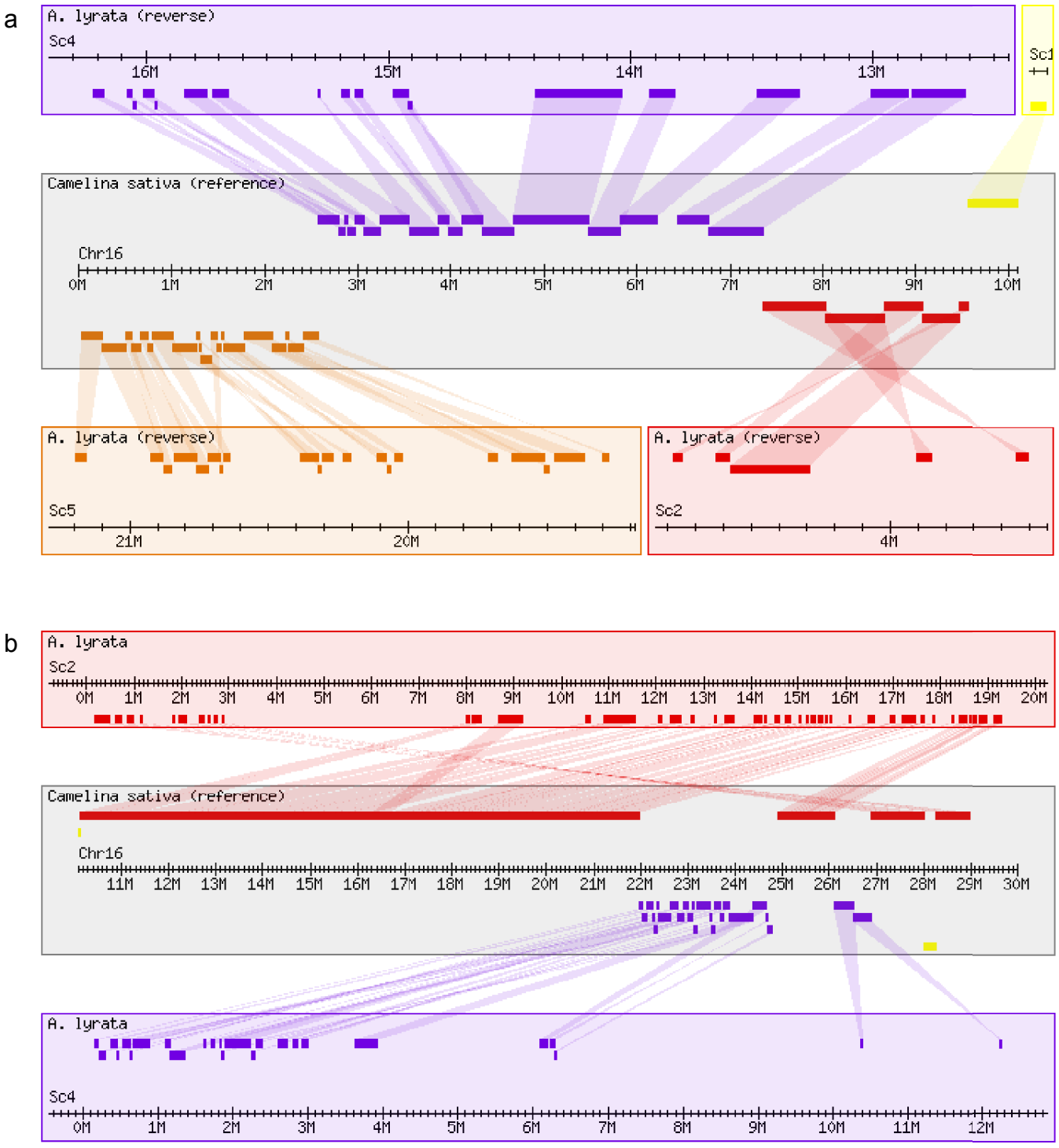
b



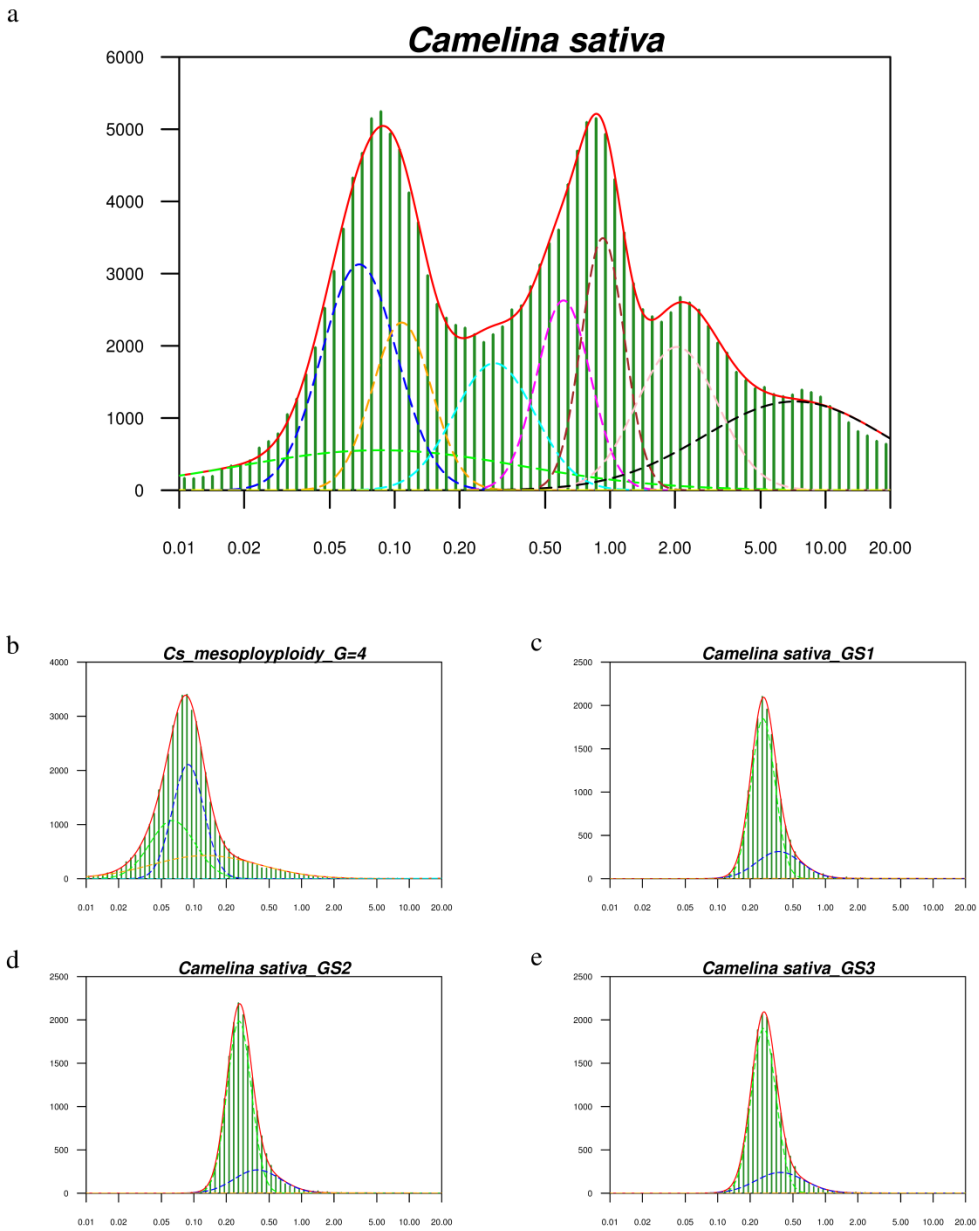
Supplementary Fig. 7. Collinearity of *Camelina sativa* genome with *Arabidopsis thaliana* and *Brassica rapa*. The draft genome sequence of *C. sativa* was compared to the genome sequences of related crucifer species, including *A. thaliana* (a) and *B. rapa* (b) by NUCmer[7].



Supplementary Fig. 8. Distribution of genomic blocks (A-X) among *Camelina sativa* chromosomes.



Supplementary Fig. 9. Gbrowse-syn view of the comparison between *C. sativa* short (a) or long (b) arm of the chromosome 16 and *Arabidopsis lyrata* genome



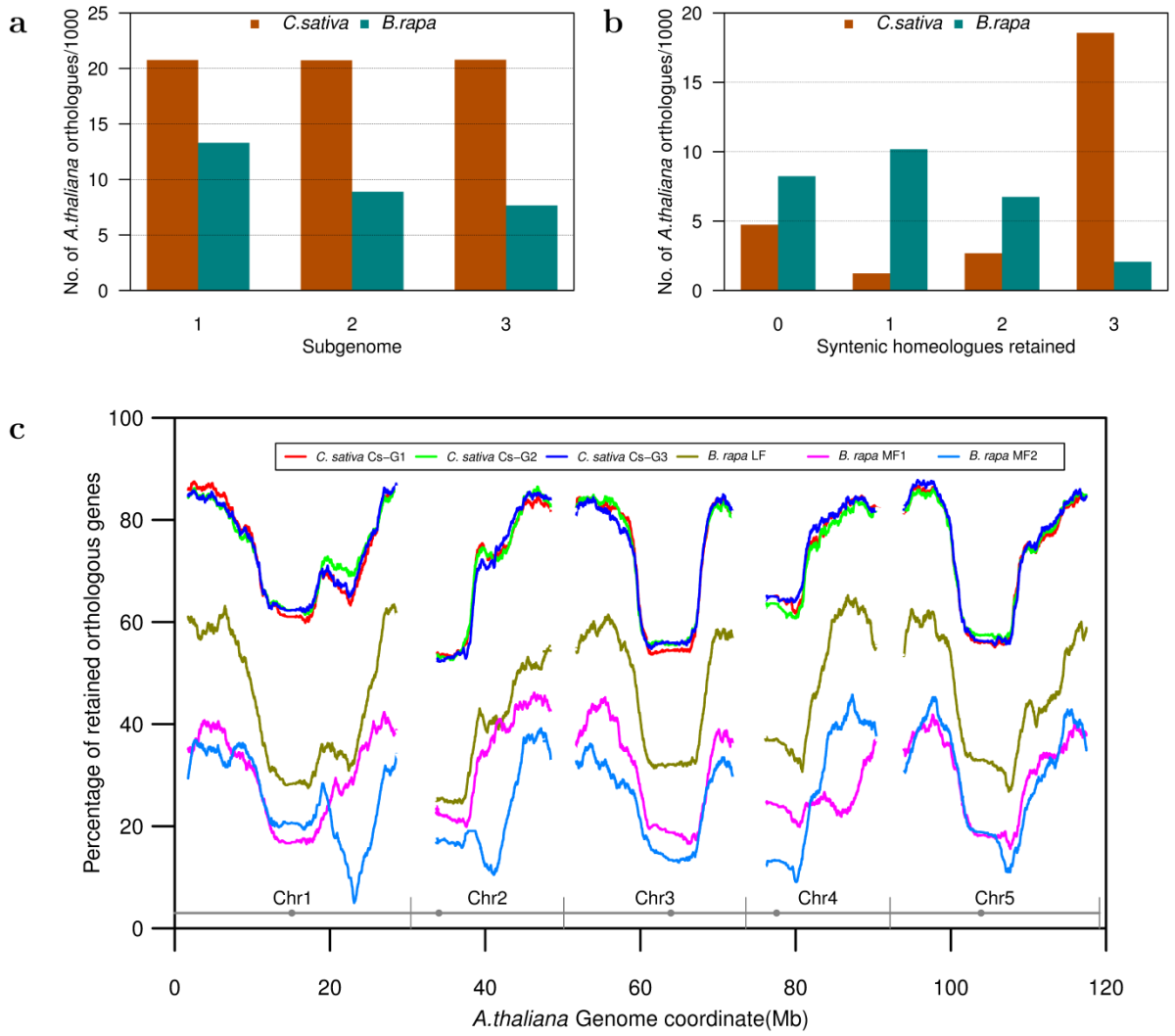
Supplementary Fig. 10. Age distributions of paralogous and orthologous genes in *Camelina sativa* indicated by histograms showing the frequency distributions of *Ks* values

(a) comparing pairs of all paralogues in *C. sativa*

(b) comparing pairs of all syntenic homologues in *C. sativa*

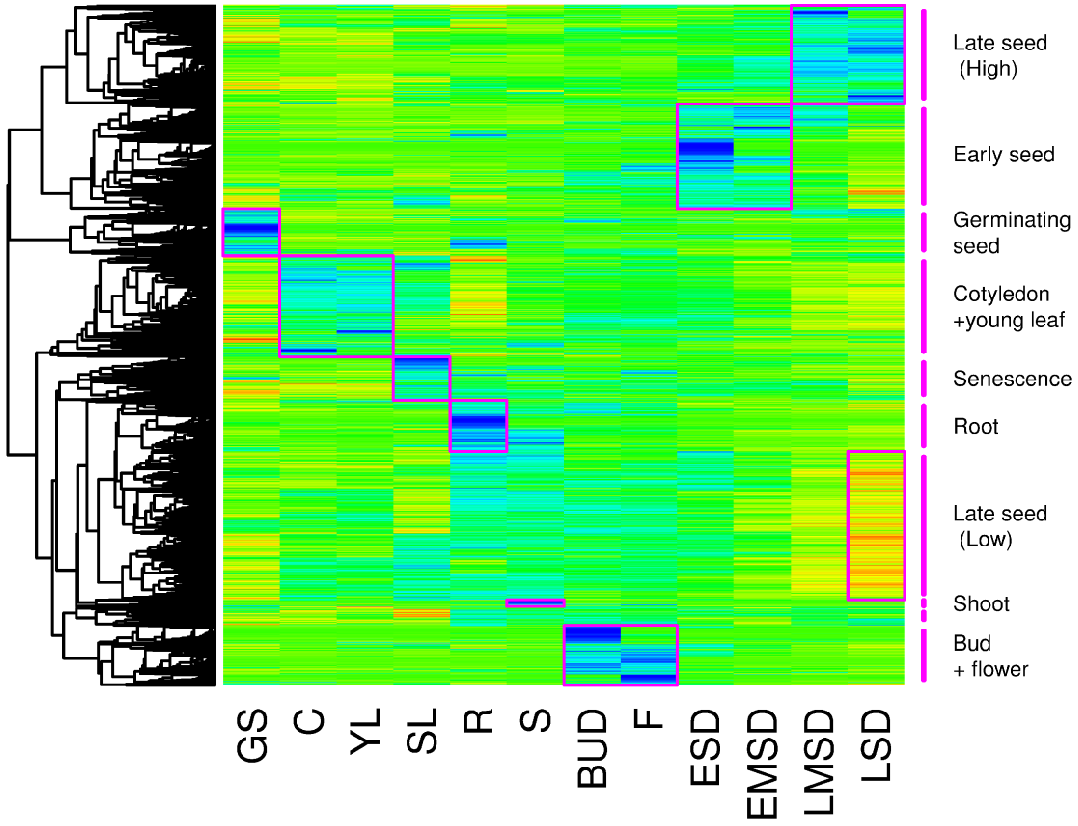
(c, d and e) comparing pairs of all orthologues between *C. sativa* subgenomes and *A. thaliana*.

The red line in each histogram represents the fitted mixture model with normal distribution (Gaussian) components. The dotted lines represent the individual Gaussian components. The Complete list of mixture model Gaussian components is provided in Supplementary Table 23.



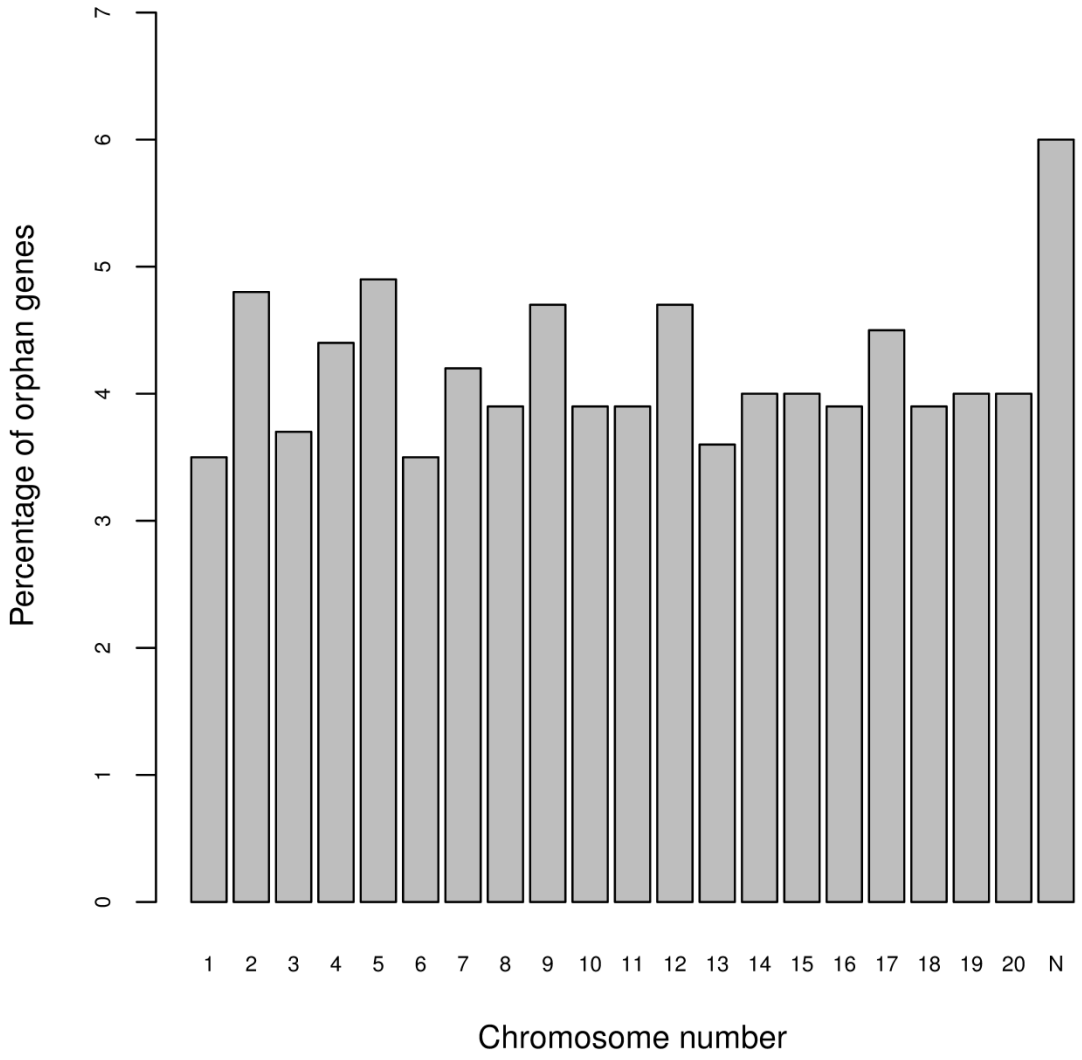
Supplementary Fig. 11. Fractionation of genes in *C. sativa* following the whole genome triplication

(a) Comparison of the number of syntelogs of *A. thaliana* genes retained in the three sub-genomes of *C. sativa* (Cs-G1, 2 and 3) and *B. rapa* (LF, MF1 and MF2).
(b) Comparison of the copy number of syntelogs of *A. thaliana* genes retained in the *C. sativa* and *B. rapa* genomes.
(c) The density of syntelogs of *A. thaliana* retained in the three sub-genomes of *C. sativa* and *B. rapa* are shown. The X- and Y-axis represent the physical positions of the *A. thaliana* chromosomes and the percentages of retained syntelogs around each *A. thaliana* gene (average of retention of 500 genes upstream and downstream), respectively.



Supplementary Fig.12. Clustering and heatmap of fully-retained genes belonging to the interaction group.

The clustering of 12,112 genes belonging to the triplets that showed statistically significant magnitude of interaction [$p < 0.05$ (ANOVA test for interaction); STDEV (GxT random effects)>0.25] was carried out using 1 minus the Pearson sample correlation as the distance measure. The average linkage method was used for clustering. The heatmap was drawn using the statistical package R where the colours were 256 levels of rainbow, starting with red representing low values and ending with blue representing high values. G, germinating seed; C, cotyledon; YL, young leaf; SL, senescing leaf; R, root; S, stem; F, flower; ESD, early seed development; EMSD, early-mid seed development; LMSD, late-mid seed development; LSD, late seed development.



Supplementary Fig. 13. Percentage of protein coding orphan genes on 20 chromosomes of *Camelina sativa*

Supplementary Table 1. Sequencing libraries used for genome assembly of *Camelina sativa*

Illumina sequencing

Library insert size (bp)	Average read length (bp)	Total raw data (Gb)	Total filtered data (Gb)	Sequence depth (X)*
Paired-end libraries				
225	2x100	7.08	5.57	7.10
	2x100	12.97	10.92	13.91
325	2x100	9.04	7.00	8.92
	2x100	15.58	12.86	16.38
Short span mate-pair libraries				
3000	2x100	41.46	26.24	33.43
5,000	2x100	39.66	19.68	25.07
8,000	2x100	39.63	14.26	18.17
Total		**165.42**	**96.53**	**122.97**

Roche 454 sequencing

Library insert size (bp)	Average read length (bp)	Total raw data (Mb)	Total reads	Usable mate pairs
Medium span mate-pair libraries				
15Kb	282	94.04	333,025	53,216
	341	264.96	769,457	147,941
	344	247.79	724,019	135,533
20Kb	254	124.88	490,496	42,643
	343	250.62	730,832	102,406
	351	172.22	489,738	67,463
25Kb	344	247.73	720,175	141,349
	344	220.45	639,989	167,086
Long span mate-pair libraries				
40 Kb	272	254.22	720,321	283,634
	271	267.67	661,989	302,417
Total		**2,144.58**	**6,280,041**	**1,443,688**

*Estimated genome size=785 Mb

Supplementary Table 2. Summary of *Camelina sativa* draft genome assembly statistics

Assembly Statistics	Hierarchical genome assembly stages		
	SOAPdenovo + Gapcloser	Bambus superscaffolding	Pseudomolecule construction[1]
Number of scaffolds (> 200 bp)	39,514	37,871	37,418
Total span (Mb)	641.39	641.45	641.45
Coverage (785 Mb genome)	81.65%	81.66%	81.66%
Longest (Mb)	3.91	16.18	49.7
Mean length (Kb)	16.23	16.93	17.14
Scaffold N50	603.04 Kb	2.16 Mb	30.09 Mb
Scaffold L50	289	78	10
Contig N50	33.42 Kb	32.17 Kb	33.41 Kb
Contig L50	5,380	5618	5384
GC content (%)	33.99	33.99	33.99
Gap length (Mb)	44.99	45.09	45.05
Gap size (%)	7.02	7.03	7.02

[1]Chromosome scale pseudomolecules were constructed by anchoring and orienting assembled scaffolds using a high resolution SNP/Indel-based genetic linkage map.

Supplementary Table 3. Anchoring of the *Camelina sativa* genome sequence to the high density SNP/Indel-based genetic linkage map
Genetic linkage map of *C. sativa* representing polymorphic SNP/Indel markers was constructed by restriction site associated DNA (RAD) and Illumina GoldenGate genotyping assay approches.

Linkage group/ Pseudomolecule	Number of SNP/Indel/SSR markers			No. of anchored scaffolds	Pseudomolecules size (bp)	Genetic distance (cM)
	GoldenGate assay	RAD	Total			
CsLG01/Csa01	27	97	124	11	23,241,285	94.16
CsLG02/Csa02	13	123	136	39	29,139,806	150.88
CsLG03/Csa03	36	182	218	18	28,500,605	222.11
CsLG04/Csa04	34	138	172	26	30,110,606	177.55
CsLG05/Csa05	25	137	162	52	34,971,402	206.21
CsLG06/Csa06	20	68	88	14	26,391,736	113.98
CsLG07/Csa07	31	195	226	20	33,617,777	225.36
CsLG08/Csa08	35	183	218	25	27,720,346	212.55
CsLG09/Csa09	40	151	191	50	38,094,978	216.35
CsLG10/Csa10	26	176	202	14	25,316,904	167.20
CsLG11/Csa11	58	310	368	36	49,703,607	307.07
CsLG12/Csa12	16	85	101	48	33,044,737	118.15
CsLG13/Csa13	38	110	148	20	24,101,694	150.93
CsLG14/Csa14	30	195	225	23	31,758,599	219.05
CsLG15/Csa15	20	120	140	47	30,521,352	172.99
CsLG16/Csa16	24	147	171	14	29,110,212	197.18
CsLG17/Csa17	29	143	172	50	35,492,130	236.75
CsLG18/Csa18	28	101	129	13	20,869,551	142.84
CsLG19/Csa19	25	158	183	22	26,736,940	186.02
CsLG20/Csa20	24	177	201	46	30,099,736	188.36
Total	579	2996	**3575**	**588**	**608,544,003**	**3705.69**
Unanchored scaffolds				**37398**	**32,908,259**	

Supplementary Table 4. Summary of the *Camelina sativa* genome features

Features	
Genome assembly	
Predicted genome size	785.50 Mb
Assembled genome span	641.45 Mb
Genome coverage	81.66%
Number of scaffolds (>200 bp)	37,418
Number of scaffolds (>2 kb)	1,561
Number of contigs	71,834
Scaffold N50	30.09 Mb
Contig N50	33.41 Kb
Gap length	45.04 Mb
Gap size	7.02%
GC content	33.99%
Genetic map and scaffold anchoring	
Number of polymorphic markers	3,575
No. of pseudomolecules/linkage groups	20
Anchored genome size	608.54 Mb/94.87%
Unachored genome size	32.90 Mb/5.13%
Longest pseudomolecule (CsChr11)	49.70 Mb
Shortest pseudomolecule (CsChr18)	20.86 Mb
Genome annotation	
Gene span	213.8 Mb
Number of predicted genes	89,418
Number of gene models	94,495
Gene encoding alternatively spliced variants	4,753 (5.3%)
Genes on chromosomes	85,274 (95.36%)
Genes on unanchored scaffolds	4,144 (4.63%)
Mean gene length	2391 bp
Mean predicted ORF length	1223 bp
Repeat annotation	
Total transposable elements	180.1 Mb (28%)
Retrotransposons	123.5 Mb (19%)
DNA transposons	21.0 Mb (3%)
Unclassified	32.6 Mb (5%)

Supplementary Table 5. Assessment of the quality of the *Camelina sativa* genome assembly by aligning it with Baterial artificial chromosomes (BAC) sequences

Slightly lower identity percentages were due to the gaps in both *C. sativa* assembly and BAC scaffold sequences

BAC Scaffold number	*C. sativa* Chromosome aligned	% identity	BAC Scaffold number	*C. sativa* Chromosome aligned	% identity	BAC Scaffold number	*C. sativa* Chromosome aligned	% identity
1	Chr7	94.7594	28	Chr15	98.3786	55	Chr20	98.04
2	Chr11	91.5403	29	Chr13	90.5067	56	Chr20	94.3869
3	Chr14	97.2046	30	Chr5	92.1088	57	Chr12	97.869
4	Chr14	96.2752	31	Chr6	96.8471	58	Chr15	94.1187
5	Chr2	91.0852	32	Chr2	89.9192	59	Chr16	99.8078
6	Chr14	90.629	33	Chr5	89.8238	60	Chr5	92.71
7	Chr19	95.9205	34	Chr6	98.0294	61	Chr2	90.1238
8	Chr6	97.6735	35	Chr5	95.598	62	Chr12	97.1042
9	Chr19	98.8106	36	Chr5	91.5759	63	Chr7	99.1329
10	Chr15	97.3878	37	Chr10	95.04	64	Chr4	98.49
11	Chr14	93.2389	38	Chr5	90.8525	65	Chr13	97.8684
12	Chr13	95.9456	39	Chr3	91.2544	66	Chr12	94.4231
13	Chr1	92.4452	40	Chr18	98.3567	67	Chr12	97.7118
14	Chr5	89.1346	41	Chr12	93.8675	68	Chr4	95.3103
15	Chr11	95.0806	42	Chr7	96.375	69	Chr19	90.6737
16	Chr11	91.36	43	Chr20	89.9744	70	Chr14	97.0086
17	Chr13	97.8254	44	Chr3	91.0748	71	Chr2	90.9054
18	Chr10	95.9774	45	Chr12	96.9729	72	Chr7	98.0056
19	Chr8	90.3522	46	Chr9	91.6987	73	Chr20	95.4435
20	Chr20	94.5386	47	Chr12	97.5814	74	Chr14	94.1021
21	Chr2	90.1989	48	Chr19	92.0881	75	Chr12	97.2943
22	Chr7	95.5662	49	Chr9	96.4484	76	Chr16	96.9023
23	Chr9	89.1141	50	Chr6	95.916	77	Chr2	93.8778
24	Chr13	99.0869	51	Chr20	98.7038	78	Chr9	97.0446
25	Chr2	97.5267	52	Chr4	97.9479	79	Chr3	95.4374
26	Chr6	98.1808	53	Chr5	90.4718	80	Chr12	95.115
27	Chr9	94.4005	54	Chr9	99.874	81	Chr18	97.0116

Supplementary Table 6. Alignment of Illumina short read sequences onto the *Camelina sativa* genome assembly.

Library insert size (bp)		Number of filtered reads	Assembled reads		Unassembled reads	
			Number	%	Number	%
Paired end reads						
225	Read1	82,582,236	75,675,263	92%	6,906,973	8%
	Read2	82,582,236	75,754,992	92%	6,827,244	8%
325	Read1	99,709,678	91,406,657	92%	8,303,021	8%
	Read2	99,709,678	91,506,659	92%	8,203,019	8%
Mate pair reads						
3,000	Read1	134,056,177	103,881,069	77%	30,175,108	23%
	Read2	134,056,177	104,470,910	78%	29,585,267	22%
5,000	Read1	100,853,282	79,926,033	79%	20,927,249	21%
	Read2	100,853,282	81,222,281	81%	19,631,001	19%
8,000	Read1	73,866,314	58,796,501	80%	15,069,813	20%
	Read2	73,866,314	59,308,871	80%	14,557,443	20%
Total/Average		982,135,374	821,949,236	**84%**	160,186,138	**16%**

Supplementary Table 7. Genomic mapping of *Camelina sativa* EST and 454 read sequences

Sequences			Alignment statistics		
Type	**Number**	**Total sequence data (bp)**	**Number**	**Average sequence coverage**	**Average sequence identity**
Sanger ESTs	10,254	9,272,068	10,167	94.53%	99.21%
454 ESTs	32,096	27,431,012	32,013	91.45%	98.43%

Supplementary Table 8. Assessment of completeness, continuity and coverage of gene space in *Camelina sativa* assembly by CEGMA analysis

Core eukaryotic gene (CEG) set includes 248 gene models which are divided into four subsets based on their degree of protein sequence conservation. Group 1 represents the least conserved of all 248 CEGs, with the degree of conservation increasing in subsequent groups through to group 4.

CEG classification	Number of CEGs	Full-length CEGs mapped in the genome	
		Number	%
Group 1	66	66	100
Group 2	56	55	98.21
Group 3	61	61	100
Group 4	65	65	100
Total	248	247	99.55

Supplementary Table 9. Classification and composition of transposable elements in the *Camelina sativa* genome

RepeatClass	MaskedBase	Copy number	%assembly
Retrotransposons			
LTR-Retrotransposons			
LTR/Gypsy	65,416,672	90,542	10.20%
LTR/Copia	22,792,697	36,390	3.55%
RC/Helitron	13,003,747	42,924	2.03%
LTR/ERV1	1,374,909	3,126	0.21%
LTR	1,951,951	4,993	0.30%
LTR/ERVK	207,933	556	0.03%
LTR/Caulimovirus	162,201	122	0.03%
Non-LTR-Retrotransposons			
LINE/L1	18,373,883	26,010	2.86%
LINE/I	32,947	156	0.01%
SINE/tRNA	152,338	1,107	0.02%
SINE	65,241	166	0.01%
DNA transposons			
DNA/MULE-MuDR	3,838,798	13,799	0.60%
DNA/MuDR	3,719,444	7,127	0.58%
DNA/En-Spm	3,107,100	5,886	0.48%
DNA/hAT-Ac	2,822,522	10,735	0.44%
DNA	1,707,516	7,632	0.27%
DNA/TcMar-Pogo	1,403,636	5,416	0.22%
DNA/Harbinger	1,262,516	3,869	0.20%
DNA/CMC-EnSpm	1,187,831	5,239	0.19%
DNA/hAT-Tag1	630,879	3,217	0.10%
DNA/PIF-Harbinger	450,943	1,120	0.07%
DNA/hAT	382,624	1,891	0.06%
DNA/TcMar-Stowaway	310,341	1,599	0.05%
DNA/TcMar-Mariner	128,706	739	0.02%
DNA/TcMar	55,952	262	0.01%
Satellites	2,253,299	2,856	0.35%
Simple_repeats	324,100	1,294	0.05%
Other/Composites	317,323	1,751	0.05%
Unclassified elements	32,686,651	143,788	5.10%
Total	180,124,700	424,312	28.08%

Supplementary Table 10. RNAseq libraries used for transcriptome characterization of the *Camelina sativa*

Growth stages	Replicate Number	Average read length (bp)	Total raw data (bp)	Total filtered data (bp)
Vegetative				
Germinating seed (GS)	1	2x100	4,827,561,200	4,412,077,343
	2	2x100	9,678,451,800	8,887,406,459
	3	2x100	470,856,800	428,774,166
Cotyledon (C)	1	2x100	1,575,872,200	1,458,855,242
	2	2x100	2,373,721,600	2,195,340,661
	3	2x100	6,796,837,400	6,263,802,383
Young leaf (YL; two week old seedling)	1	2x100	1,390,013,400	1,327,940,893
	2	2x100	1,388,316,800	1,330,537,403
	3	2x100	3,085,135,200	2,946,287,373
Senescing leaf (OL; eight week old seedling)	1	2x100	898,662,800	861,002,057
	2	2x100	1,901,790,400	1,819,475,986
	3	2x100	1,556,933,000	1,489,005,380
Root (R; four week old seedling)	1	2x100	3,006,788,600	2,873,976,582
	2	2x100	2,717,533,800	2,598,433,059
	3	2x100	2,778,536,800	2,659,203,079
Stem (S; four week old seedling)	1	2x100	2,560,075,000	2,454,066,142
	2	2x100	1,864,700,800	1,786,277,633
	3	2x100	2,220,999,600	2,123,131,915
Reproductive				
Bud (B)	1	2x100	863,765,800	805,934,907
	2	2x100	689,403,400	644,811,841
	3	2x100	1,946,654,200	1,817,691,599
Flower (F)	1	2x100	641,596,200	600,591,560
	2	2x100	1,403,225,000	1,311,005,059
	3	2x100	2,626,887,000	2,467,539,986
Early seed development (ESD; 4-12 DPA*)	1	2x100	1,740,685,600	1,636,306,370
	2	2x100	1,538,385,000	1,454,376,957
	3	2x100	865,246,600	796,885,652
Early-mid seed development (EMSD; 18 & 22 DPA)	1	2x100	1,047,840,600	950,936,443
	2	2x100	2,279,660,600	2,147,754,756
	3	2x100	1,658,205,600	1,571,110,448
Late-mid seed development (LMSD; 26 & 30 DPA)	1	2x100	1,648,937,800	1,570,446,299
	2	2x100	2,445,500,800	2,319,358,756
	3	2x100	1,429,953,400	1,351,795,713
Late seed development (LSD; 32 & 40 DPA)	1	2x100	1,556,199,800	1,479,035,186
	2	2x100	1,522,260,400	1,419,733,552
	3	2x100	1,513,994,600	1,433,853,177
Total			**78,511,189,600**	**73,694,762,017**

*days post anthesis

Supplementary Table 11. A comparison of genome composition of Brassicaceae species

Genome composition	***Camelina sativa***	***Arabidopsis lyrata***	***Arabidopsis thaliana***	***Brassica rapa***
Genome size (Mb; assembled)	641.45	207	120	283.8
Genic (%)	31.33	28.7	42.3	29
Intergenic (%)	38.58	41.6	34	31.4
Transposable elements (%)	28.08	29.7	23.7	39.5
No. of genes	89,418	32,670	27,206	41,174
Gene density (genes/Mb)	139	158	227	145
Average gene length (bp)	2,391	2,080	2,242	2,015

Supplementary Table 12. Genome size and gene number of completely sequenced plant species

Plant species	Abbreviation	Genome size	Gene number
Manihot esculenta	Mes	533	30,666
Ricinus communis	Rco	400	31,221
Linum usitatissimum	Lus	318.3	43,471
Populus trichocarpa	Ptr	422.9	41,335
Medicago truncatula	Mtr	257.6	44,135
Phaseolus vulgaris	Pvu	521.1	27,197
Glycine max	Gma	975	54,175
Cucumis sativis	Csa	203	21,491
Prunus persica	Ppe	227.3	27,864
Malus domestica	Mdo	881.3	63,538
Fragaria vesca	Fve	240	32,831
Arabidopsis thaliana	Ath	135	27,416
Arabidopsis lyrata	Aly	207	32,670
Camelina sativa	**Csa (*Camelina sativa*)**	**641.4**	**89,421**
Capsella rubella	Cru	134.8	26,521
Brassica rapa	Bra	283.8	41,174
Thellungiella halophila	Tha	243.1	26,351
Carica papaya	Cpa	135	27,332
Gossypium raimondii	Gra	761.4	37,505
Theobroma cacao	Tca	346	29,452
Citrus sinensis	Csi	319	25,346
Eucalyptus grandis	Egr	691	36,376
Vitis vinifera	Vvi	487	26,346
Solanum tuberosum	Stu	800	35,119
Solanum lycopersicum	Sly	900	34,727
Mimulus guttatus	Mgu	321.7	26,718
Aquilegia coerulea	Aco	302	24,823
Sorghum bicolor	Sbi	697.5	34,496
Zea mays	Zma	2066	39,656
Setaria italica	Sit	405.7	35,471
Panicum virgatum	Pvi	1358	65,878
Oryza sativa	Osa	372	39,049
Brachypodium distachyon	Bdi	272	26,552
Selaginella moellendorfii	Smo	212.5	22,285
Physcomitrella patens	Ppa	480	32,272

Supplementary Table 13. Comparison of sequence and structural features of orphan and non-orphan genes in the *Camelina sativa* genome

	Mean		Median	
	Orphans	**Non-Orphans**	**Orphans**	**Non-Orphans**
CDS length	324.54±3.389	1236.05±3.351	261.0±3.389	1020.0±3.351
Percent GC	42.81±0.089	44.97±0.011	42.60±0.089	44.61±0.011
Percent GC1	46.00±0.141	50.25±0.016	45.83±0.141	50.28±0.016
Percent GC2	40.53±0.141	40.59±0.019	40.21±0.141	40.22±0.019
Percent GC3	41.92±0.136	44.07±0.023	41.73±0.136	43.24±0.023
Number of introns	1.7±0.032	4.514±0.019	1.0±0.032	3.0±0.019

Supplementary Table 14. Evolutionary origins of orphan genes in *Camelina sativa* genome

Mode of evolution	Number of orphans (% of orphans)
Duplication	1737 (46.18%)
Out-of-frame CDS hits in Brassicaceae(Camelina specific)	210 (12.68%)
Non-coding region hits in Brassicaceae (Camelina specific)	144 (8.69%)
Overlapping gene models	53 (1.41%)
Out-of-frame CDS hits in non-Brassicaceae	16 (0.43%)
Non-coding region hits in non-Brassicaceae	9 (0.24%)

Supplementary Table 15. Genomic composition of the three subgenomes within *Camelina sativa*

	Subgenome I	Subgenome II	Subgenome III
Number of chromosomes	6	7	7
Chromosomes	17, 16, 15, 04, 13, 11	14, 07, 19, 06, 08, 10, 18	03, 05, 01, 09, 20, 02, 12
Genome size (bp)	199,039,601	192,411,853	217,092,549
Genic (bp)	69,407,192	69,058,569	71,249,998
Intergenic (bp)	75,467,826	73,080,545	81,816,353
Transposable elements (bp)	54,164,583	50,272,739	64,026,198
Number of genes	28,254	27,813	29,207
Syntenic genes	20,754	20,605	20,479
Syntenic genes retained (%)	76%	76%	75%
Non-syntenic genes	7,500	7,208	8,728
Repeat content (bp)	54,164,583	50,272,739	64,026,198
Repeat content (%)	27.21%	26.13%	29.49%
Gene density (genes/Mb)	142	145	135

Supplementary Table 16. Comparison of GB associations of *Camelina sativa* karyotype with ancestral karyotypes in the Brassicaceae

Chr	GB Association	No.	Chr	GB Association	No.	Chr	GB Association	No.	Chr	GB Association	No.	*C. sativa* Chromosomes
ACK (n=8)			PCK (n=7)			tPCK (n=7)			dACK (n=7)*			
AK1	A/B	3	AK1	A/B	3	AK1	A/B	3	AK1	A/B	3	Csa17, Csa14, Csa03
	B/C	4		B/C	4		B/C	4		B/C	4	Csa17, Csa14, Csa04, Csa05
AK2	D/E	5	AK2	D/E	5	AK2/5	N/M	5	AK2/4	I/J	3	Csa16, Csa07, Csa05
AK3	F/G	3	AK3	F/G	3		M/E	0		D/I	5	Csa16, Csa16, Csa07, Csa07, Csa05
	G/H	5		G/H	5	AK3	F/G	3		D/E	5	Csa07, Csa07, Csa07, Csa05, Csa09
AK4	I/J	3	AK4	I/J	3		G/H	5		I/E	8	Csa16, Csa16, Csa16, Csa07, Csa07,Csa07, Csa09, Csa09
AK5	K/L	3	AK7	S/T	3	AK4	I/J	3	AK3	F/G	3	Csa15, Csa19, Csa01
	L/M	3		T/U	3	AK7	S/T	3		G/H	5	Csa15, Csa19, Csa01, Csa01, Csa01
	M/N	3	AK6/8	O/P	4		T/U	3	AK5	K/L	3	Csa04, Csa06, Csa09
AK6	O/P	4		P/W	0	AK6/8	O/P	4		L/M	3	Csa04, Csa06, Csa09
	P/Q	2		W/R	0		P/W	0		M/N	3	Csa04, Csa06, Csa09
	Q/R	3	AK5/6/8	N/M	3		W/R	0	AK6	O/P	4	Csa13, Csa08, Csa02, Csa02
AK7	S/T	3		M/V	0	AK2/5/6	N/M	3		P/Q	2	Csa13, Csa08
	T/U	3		V/K	0		D/V	0		Q/R	3	Csa13, Csa08, Csa20
AK8	V/W	2		K/L	3		V/K	0	AK7	S/T	3	Csa11, Csa10, Csa12
	W/X	3		L/Q	0		K/L	3		T/U	3	Csa11, Csa10, Csa12
	Q/X	0		L/Q	0	AK8	V/W	2				Csa11, Csa18
				Q/X	0		W/X	3				Csa11, Csa18, Csa02

*Inferred from the present study

Supplementary Table 17. *Camelina sativa* genes retained in ancestral genomic blocks

Block	Cs-G1	Cs-G2	Cs-G3
A	1689	1653	1661
B	1156	1158	1158
C	750	767	752
D	495	515	503
E	1357	1366	1372
F	2192	2183	2141
G	99	94	93
H	482	491	456
I	678	668	694
J	1612	1626	1638
K	222	219	212
L	330	334	342
M	218	235	223
N	1426	1414	1433
O	406	399	415
P	310	300	299
Q	391	400	409
R	1819	1818	1829
S	446	445	438
T	211	212	236
U	2235	2204	2238
V	486	482	479
W	1121	1130	1138
X	624	628	625
Total	**20755**	**20741**	**20784**

Supplementary Table 18. Rate of gene loss in the sub-genomes of *Camelina sativa* and *Brassica rapa*

Sub-genomes	Triplication age (Mya)	Fraction retained	Rate of gene loss (k)*
Camelina sativa			
Cs-G1	5.41	0.759306	0.050897
Cs-G2	5.41	0.759280	0.050903
Cs-G3	5.41	0.760969	0.050492
Brassica rapa			
LF	22.50	0.482921	0.032351
MF1	22.50	0.321917	0.050376
MF2	22.50	0.283525	0.056020

*Rate of gene loss was calculated assuming that they are lost at a fixed rate

Exponential decay equation f=exp(-Kt), where t= time in million years, f=fraction retained, and k is the rate constant

Supplementary Table 19. ANOVA test to determine significant differences in exon count between homeologues compared at genomic block and subgenome level

Sample size (N)=Total df+1

Subgenomewise comparison

		sos	df	ms	F-test	P-value	significance
Subgenomes	Between	5.925	2	2.963	0.091	9.13E-01	ns
	Within	1762130	54414	32.384			
	Total	1762136	54422				

sos, sum of squared deviations; df, degrees of freedom; ms, mean squares

Genomic blockwise comparison

Block		sos	df	ms	F-test	P-value	significance
Block A	Between	4.178	2	2.089	0.057	9.45E-01	ns
	Within	168963	4602	36.715			
	Total	168967	4610				
Block B	Between	5.174	2	2.587	0.085	9.18E-01	ns
	Within	91812.6	3033	30.271			
	Total	91817.8	3041				
Block C	Between	15.534	2	7.767	0.181	8.35E-01	ns
	Within	84752.7	1971	43			
	Total	84768.2	1979				
Block D	Between	13.354	2	6.677	0.204	8.15E-01	ns
	Within	41198.6	1260	32.697			
	Total	41211.9	1268				
Block E	Between	16.335	2	8.168	0.235	7.90E-01	ns
	Within	129054	3717	34.72			
	Total	129071	3725				
Block F	Between	7.329	2	3.664	0.108	8.98E-01	ns
	Within	199852	5889	33.937			
	Total	199860	5897				
Block G	Between	7.056	2	3.528	0.121	8.87E-01	ns
	Within	7113.46	243	29.274			
	Total	7120.52	251				
Block H	Between	32.826	2	16.413	0.473	6.23E-01	ns
	Within	42389.1	1221	34.717			
	Total	42422	1229				
Block I	Between	0.041	2	0.021	0.001	9.99E-01	ns
	Within	58089.8	1776	32.708			
	Total	58089.9	1784				
Block J	Between	0.222	2	0.111	0.003	9.97E-01	ns
	Within	141805	4401	32.221			
	Total	141806	4409				

Block K	Between	5.016	2	2.508	0.097	9.07E-01	ns
	Within	14680.6	570	25.755			
	Total	14685.6	578				
Block L	Between	5.343	2	2.672	0.115	8.92E-01	ns
	Within	18905.5	813	23.254			
	Total	18910.8	821				
Block M	Between	28.772	2	14.386	0.443	6.43E-01	ns
	Within	17652.3	543	32.509			
	Total	17681	551				
Block N	Between	0.466	2	0.233	0.007	9.93E-01	ns
	Within	121437	3870	31.379			
	Total	121437	3878				
Block O	Between	0.511	2	0.256	0.009	9.91E-01	ns
	Within	31768.8	1098	28.933			
	Total	31769.3	1106				
Block P	Between	1.027	2	0.514	0.02	9.81E-01	ns
	Within	20020.2	768	26.068			
	Total	20021.3	776				
Block Q	Between	22.425	2	11.213	0.257	7.74E-01	ns
	Within	44188	1011	43.707			
	Total	44210.5	1019				
Block R	Between	1.847	2	0.923	0.029	9.72E-01	ns
	Within	160224	4965	32.271			
	Total	160226	4973				
Block S	Between	2.629	2	1.315	0.061	9.41E-01	ns
	Within	24046.8	1107	21.722			
	Total	24049.4	1115				
Block T	Between	4.922	2	2.461	0.084	9.19E-01	ns
	Within	15764.3	540	29.193			
	Total	15769.2	548				
Block U	Between	10.259	2	5.13	0.17	8.44E-01	ns
	Within	172025	5691	30.228			
	Total	172035	5699				
Block V	Between	8.662	2	4.331	0.128	8.80E-01	ns
	Within	14531.3	429	33.872			
	Total	14540	437				
Block W	Between	1.463	2	0.731	0.026	9.74E-01	ns
	Within	83693.8	3015	27.759			
	Total	83695.2	3023				
Block X	Between	2.604	2	1.302	0.04	9.61E-01	ns
	Within	54817.1	1674	32.746			
	Total	54819.7	1682				

Supplementary Table 20. Homeologue silencing in *Camelina sativa*

Number of homeologous triplets where one or two homeologues were silenced (FPKM=0 across all tissue types) but remaining homeologues were expressed are presented

Tissue type	Subgenome to which silenced homeologue(s) belong						
	Cs-G1	Cs-G2	Cs-G3	Cs-G1 and G2	Cs-G1 and G3	Cs-G2 and G3	Cs-G1, G2 and G3
Cotyledon (C)	390	355	405	368	348	309	2169
Early-mid seed development (EMSD; 18 & 22 DPA)	465	440	420	344	345	285	1680
Early seed development (ESD; 4-12 DPA)	461	442	418	319	278	281	1267
Flower (F)	398	411	369	279	261	249	887
Germinating seed (GS)	413	429	430	336	292	306	1411
Bud (B)	398	388	400	275	245	254	851
Late seed development (LSD; 32 & 40 DPA)	529	500	525	499	490	481	3342
Late-mid seed development (LMSD; 26 & 30 DPA)	460	417	453	412	379	336	2059
Mature leaf (OL)	456	456	449	416	402	410	2432
Root (R)	426	383	424	332	304	323	1502
Stem (S)	387	383	410	340	327	330	1698
Young leaf (YL)	426	410	416	362	335	326	2404

Supplementary Table 21. Retention pattern of acyl-lipid metabolism genes in *Camelina sativa*

Pathway	Frequency								Proportion (%)			
				Camelina sativa					*Camelina sativa*			
				Syntenic					Syntenic			
	Arabidopsis thaliana	*Camelina sativa* (Total)	Expansion of gene families in *C. sativa* in comparison to *A. thaliana*	Fully-retained	2 copy	1 copy	0 copy	Non-syntenic	Fully-retained	2 copy	1 copy	0 copy
Fatty Acid Synthesis	34	93	173.5%	26	2	0	6	11	76.5%	5.9%	0.0%	17.6%
Fatty Acid Synthesis; Prokaryotic Galactolipid, Sulfolipid, & Phospholipid Synthesis 1	15	45	200.0%	11	2	0	2	8	73.3%	13.3%	0.0%	13.3%
Fatty Acid Elongation & Wax Biosynthesis	198	651	228.8%	149	26	11	12	141	75.3%	13.1%	5.6%	6.1%
Fatty Acid Elongation, Desaturation & Export From Plastid	14	42	200.0%	11	1	0	2	7	78.6%	7.1%	0.0%	14.3%
Triacylglycerol Biosynthesis	70	192	174.3%	48	12	8	2	16	68.6%	17.1%	11.4%	2.9%
Triacylglycerol & Fatty Acid Degradation	64	204	218.8%	55	5	4	0	25	85.9%	7.8%	6.3%	0.0%
Eukaryotic Galactolipid & Sulfolipid Synthesis	65	201	209.2%	59	3	2	1	16	90.8%	4.6%	3.1%	1.5%
Prokaryotic Galactolipid, Sulfolipid, & Phospholipid Synthesis 1	7	22	214.3%	7	0	0	0	1	100.0%	0.0%	0.0%	0.0%
Prokaryotic Galactolipid, Sulfolipid, & Phospholipid Synthesis 2	4	12	200.0%	4	0	0	0	0	100.0%	0.0%	0.0%	0.0%
Prokaryotic Galactolipid, Sulfolipid, & Phospholipid Synthesis 2; Eukaryotic Galactolipid & Sulfolipid Synthesis	10	34	240.0%	9	0	0	1	7	90.0%	0.0%	0.0%	10.0%
Prokaryotic Galactolipid, Sulfolipid, & Phospholipid Synthesis 1; Eukaryotic Galactolipid & Sulfolipid Synthesis;Phospholipid Signaling	9	29	222.2%	8	1	0	0	3	88.9%	11.1%	0.0%	0.0%
Cutin Synthesis & Transport 1	26	81	211.5%	22	1	1	2	12	84.6%	3.8%	3.8%	7.7%
Suberin Synthesis & Transport 1	47	151	221.3%	42	0	2	3	23	89.4%	0.0%	4.3%	6.4%
Mitochondrial Fatty Acid & Lipoic Acid Synthesis	10	32	220.0%	8	1	0	1	6	80.0%	10.0%	0.0%	10.0%
Mitochondrial Lipopolysaccharide Synthesis	12	23	91.7%	6	1	1	4	2	50.0%	8.3%	8.3%	33.3%
Mitochondrial Phospholipid Synthesis	14	43	207.1%	14	0	0	0	1	100.0%	0.0%	0.0%	0.0%
Oxylipin Metabolism 1	17	50	194.1%	10	1	1	5	17	58.8%	5.9%	5.9%	29.4%
Oxylipin Metabolism 2	3	13	333.3%	3	0	0	0	4	100.0%	0.0%	0.0%	0.0%
Oxylipin Metabolism 1; Oxylipin Metabolism 2	43	159	269.8%	42	1	0	0	31	97.7%	2.3%	0.0%	0.0%
Oxylipin Metabolism 1; Oxylipin Metabolism 2;Phospholipid Signaling	6	20	233.3%	5	1	0	0	3	83.3%	16.7%	0.0%	0.0%
Phospholipid Signaling	91	296	225.3%	78	7	3	3	45	85.7%	7.7%	3.3%	3.3%
Sphingolipid Biosynthesis 1	17	55	223.5%	15	0	2	0	8	88.2%	0.0%	11.8%	0.0%
Sphingolipid Biosynthesis 1;Sphingolipid Biosynthesis 2	5	14	180.0%	4	0	0	1	2	80.0%	0.0%	0.0%	20.0%
Sphingolipid Biosynthesis 2	7	24	242.9%	6	1	0	0	4	85.7%	14.3%	0.0%	0.0%
Lipid Trafficking	10	36	260.0%	9	0	0	1	9	90.0%	0.0%	0.0%	10.0%
Pathway, function or subcellular location uncertain	58	202	248.3%	40	12	5	1	53	69.0%	20.7%	8.6%	1.7%
Total	**856**	**2724**	**217.0%**	**691**	**78**	**40**	**47**	**455**	**83.5%**	**6.5%**	**2.8%**	**7.2%**

Supplementary Table 22. Roche 454 pyrosequencing, read filtering and transcriptome assembly statistics for lower chromosome number *Camelina* species

Species	Total raw reads	Total data (bp)	Clean data (bp)	Number of isogroups	Number of isotigs
Camelina hispida	730,040	289,831,958	253,990,405	23,391	30,639
Camelina laxa	727,870	285,946,338	240,013,877	20,963	22,477
Camelina rumelica ssp. Iran	736,560	320,322,283	263,285,196	23,906	32,239
Camelina rumelica ssp. transcaspica	749,447	303,539,215	256,396,754	21,260	27,891
Camelina rumelica ssp. USSR	803,764	340,536,335	290,533,930	25,088	38,567

Supplementary Table 23. Complete mixture model estimates of *Ks* distributions

	No. of duplicate pairs	G (# Gaussian components)	Chisquare	p-value	mean ln (*Ks*)	variance ln (*Ks*)	fraction of data	mean *Ks*	Merged peak*	Associated polyploidy event
Camelina sativa	194806	9	631.50177	6E-89	-2.68059	0.16064	0.1615667	0.068522	0.082	Recent Mesopolyploidy
					-2.4552	2.28313	0.1080536	0.085846		
					-2.225	0.10002	0.09422479	0.108068		
					-1.23511	0.18329	0.09686181	0.290803		
					-0.49729	0.08261	0.09721083	0.60818	0.755	α Duplication
					-0.07711	0.05289	0.1032682	0.925786		
					0.715967	0.18678	0.1103641	2.046165		β duplication
					1.976215	0.95901	0.1548896	7.215381		γ duplication
					4.21513	0.0044	0.07356034	67.70297		
Cs-G1--*A. thaliana*	15647	3	65.561246	1E-05	-1.33366	0.06465	0.7519223	0.263511	0.286	
					-0.9975	0.22271	0.2369449	0.368801		
					0.708826	3.00966	0.01113282	2.031604		
Cs-G2--*A. thaliana*	16421	3	99.272658	3E-10	-1.33388	0.06729	0.7858025	0.263454	0.285	
					-0.94712	0.24577	0.2035503	0.387858		
					0.730498	2.827	0.01064716	2.076113		
Cs-G3--*A. thaliana*	15675	3	67.105546	6E-06	-1.32159	0.06784	0.7918447	0.26671	0.287	
					-0.96337	0.2744	0.2007897	0.381605		
					1.115325	3.49089	0.00736565	3.050559		
C. sativa -- Mesopolyploidy	43350	4	109.83305	0.0002	-2.75942	0.23293	0.2996091	0.063328	0.089	
					-2.42505	0.1138	0.4112391	0.088474		
					-2.0529	1.31612	0.2841752	0.128362		
					3.468451	0.98736	0.00497664	32.087		

*Components shaded in tan colour were merged as described in **Supplementary Note 9**.

Supplementary Notes

1. Genome sequencing and assembly strategy

The presence of repetitive sequences in eukaryotic genomes can have a direct bearing on the completeness and contiguity of draft genomes assembled via whole genome shotgun sequencing approach. Paired-end deep sequencing with a combination of short, medium and long insert sizes is required to achieve maximal coverage of the genome and generate megabase sized contiguous scaffolds. To ensure genome-wide deep coverage and generate contiguous chromosome scale super-scaffolds, we used a hybrid sequencing and hierarchical assembly approach (**Supplementary Fig. 2**), utilizing a combination of Illumina and Roche 454 next generation sequencing technologies. Illumina sequencing of short span paired-end and mate-paired reads enables ultra-deep coverage at moderate cost, whereas, the longer reads along with medium to long span mate-paired reads generated by 454 pyrosequencing aid in resolving repeats and bridging low complexity regions.

The implementation of a hierarchical assembly strategy, based on (i) production of larger contigs with fewer gaps and greater accuracy of the final consensus sequence by assembling the larger pool of short span Illumina reads using SOAPdenovo[1], (ii) construction of superscaffolds by joining contigs with the help of medium to long span 454 mate-pair read information using Bambus superscaffolder[2], and (iii) construction of chromosome-scale pseudomolecules by ordering and orienting scaffolds via anchoring to comprehensive genetic linkage map(s) developed by high density sequence tag mapping in segregating population, has been proven in our laboratories to be a highly efficient and cost effective strategy for assembling high quality draft genomes of various plant species, including *C. sativa* (this study) and a few other Brassicaceae species.

2. Estimation of genome size

2.1. Flow cytometric estimation

Nuclei were extracted from young leaf tissue of *C. sativa* following the protocol described by Galbraith et al.[3] with minor alterations. Approximately 2 cm^2 of tissue was chopped in 0.5 ml of lysis buffer (0.1M Tris-HCl pH7; 2 mM MgCl2; 0.1M NaCl; 0.05% Triton) at room temperature.

The volume of buffer was adjusted to 1 ml and filtered through a 30 μm pore prior to analysis. DNA content of the nuclei from each species was estimated using DAPI staining and fluorescence measurements made using the CyFlow Ploidy analyser (Partec NJ, USA). A standard curve was generated by linear regression using the fluorescence data and genome size estimates for the genomes of *A. thaliana*, *B. carinata*, *B. juncea*, *B. napus*, *B. nigra*, *B. oleracea* and *B. rapa*. The genome size for *C. sativa* was estimated using their fluorescence measurement and the standard curve. This analysis revealed the genome size of *C. sativa* to be 825 Mb (1C content of 0.84 pg).

2.2. k-mer frequency analysis based genome size estimation

Analysis of k-mer frequency distribution in short read data provides a true estimate of the sequencing depth that can be utilized to determine the genome size[4]. The real sequencing depth (*N*) is correlated with the peak of the k-mer frequency (*M*) in sequenced reads, read length (*L*) and k-mer length (*K*), and can be derived using the formula $N=M * L/(L-K+1)$[4]. To determine the genome size of *C. sativa*, the occurrences of 17 and 21 bp k-mers in ~53 Gb of filtered Illumina PE data were counted using Jellyfish[5]. Genome size was estimated by dividing the total length of sequencing reads used in the analysis by sequencing depth (**Supplementary Fig. 3**). This analysis suggested a genome of 785 Mb, which is in agreement with the average of flow cytometry based estimates of 750 Mb[6] and 825 Mb.

3. Genome assembly validation

3.1 Alignment of BAC sequences to the genome assembly

The quality of the genome assembly was assessed by comparing it with a set of BAC scaffolds independently assembled from Illumina sequence generated from 768 indexed and pooled BACs. The alignment of all BAC scaffolds >50Kb to the *C. sativa* genome assembly using NUCmer (version 3.07)[7] showed no misassemblies (**Supplementary Table 5**).

3.2 Evaluation of assembled and unassembled short reads

To assess overall genome coverage and structural completeness of the *de novo* assembly, all filtered short Illumina reads (982 million) were mapped back to the assembled genome using Bowtie (version 2.1.0)[8], with parameters requiring end to end mapping and allowing only 2 mismatches per read. Overall, 84% of the Illumina reads could be aligned to the assembly (**Supplementary Table 6**). The proportion of unmapped reads (16%) is largely consistent with the estimated unassembled portion (18%) of the genome, likely enriched with repetitive sequences, indicating good coverage of the euchromatic genome space.

3.3 Genomic mapping of EST sequences

To assess coverage of gene space in the genome assembly, *C. sativa* ESTs derived from 454 or Sanger cDNA reads were aligned back to the genome sequence using GMAP (release 2012-01-11)[9]. About 99% of *C. sativa* ESTs aligned to the genome sequence with sequence coverage of at least 90% and identity greater than 99% (**Supplementary Table 7**), again confirming high coverage of the gene complement in the assembly and gene annotations.

3.4 CEGMA analysis

CEGMA pipeline[10], which provides a rapid and accurate method for annotation of a set of 248 highly conserved core eukaryotic genes (CEGs) that are expected to be present in all eukaryotes, was used to assess completeness of the gene catalogue in the assembled genome. This analysis revealed the *C. sativa* genome assembly and annotated gene set harbour 99.6% (247/248) of complete CEG models (**Supplementary Table 8**), thus demonstrating the completeness and very high coverage of gene complement in *the draft* assembly.

3.5 Collinearity analysis

The draft genome sequence of *C. sativa* was compared to other sequenced crucifer genomes, including A. *thaliana, A. lyrata* and *B. rapa,* by NUCmer[7]. A striking degree of synteny conservation was observed between the draft genome sequence of *C. sativa* and *Arabidopsis* species (**Fig. 3; Supplementary Fig. 7**). This comparative analysis also identified a few misassembled scaffolds, which were split and repositioned to correct false joins or insertions by utilizing additional evidence from the genetic map.

4. Orphan genes in *Camelina sativa*

Orphan genes have been identified in almost all biological domains of life yet their origins and biological functions remain poorly understood[11]. Orphan genes arising in specific lineages can potentially acquire lineage-specific functions, including in plant species such as rice[12], pigeonpea[13] and *A. thaliana*[14] where lineage-specific orphan genes have been identified.

To identify orphan genes restricted to the *C. sativa* genome, we used a BLAST-based filtering approach to firstly identify all of the Brassicaceae-specific orphans. A total of 3,761 Brassicaceae-restricted orphan genes were identified from the 89,418 *C. sativa* gene models, which represents 4.2% of the protein-coding genome. The BLAST-filtering approach was then further extended to identify the subset of Brassicaceae-restricted orphan genes that are found only in the *C. sativa* genome, resulting in the identification of 1,656 protein-coding orphans (1.85% of the *C. sativa* genome). This indicates that there are 1,656 protein-coding orphan genes in *C. sativa* that could be candidate genes for lineage-specific adaptations in this species.

The characterization of features associated with the orphan genes in *C. sativa* revealed that they have shorter sequence length, lower GC content and lower numbers of introns (than non-orphan genes) (**Supplementary Table 13**). The percentage of orphan genes on each chromosome in *C. sativa* exhibited no major bias towards any specific chromosome (**Supplementary Fig. 13**). Such characteristics are consistent with findings of other studies[13, 14, 15].

Using a BLAST-based sequence similarity approach, we identified the evolutionary origins of 2,169 of the 3,761 Brassicaceae-restricted orphan genes found within the *C. sativa* genome (**Supplementary Table 14**). A total of 1,737 of these orphan genes have significant BLASTP hits with *C. sativa* non-orphan genes which indicates that nearly half of *C. sativa* orphan genes have arisen by gene duplication of non-orphan genes. An additional 367 orphans were found to have a significant BLASTN hit at CDS level. A total of 354 (9.4%) orphans have significant BLASTN hits with other Brassicaceae species at either CDS or non-coding level. A small number of orphans (0.7%) have displayed significant similarity to non-coding or out-of-frame CDS in non-Brassicaceae species. A total of 53 orphan genes were found to have arisen by overprinting (i.e. generating a new ORF from an existing ORF) of other non-orphan genes. For the orphan genes whose mode of origin we could decipher, our analysis indicates that gene

duplication from non-orphan genes was the predominant (46.18%) mechanism by which orphan genes have arisen in the *C. sativa* genome.

5. Phylogenetic analysis of *Camelina* species

5.1. RNA extraction and mRNA enrichment

Total cellular RNA was extracted using the ToTALLY RNA kit (Ambion) according to the manufacturer's instructions. The integrity and quantity of total RNA was assessed using RNA 6000 Nano labchip on a BioAnalyzer (Agilent). Poly(A)$^+$ RNA was purified from DNase I-treated total RNA samples by two sequential rounds of oligo(dT) enrichment using the Poly(A) Purist mRNA isolation kit (Ambion). The concentration of mRNA was checked using Ribogreen (Molecular Probes) and the quality was verified using RNA 6000 Pico labchip on a BioAnalyzer (Agilent).

5.2. cDNA library construction and 454 sequencing

cDNA library construction and sequencing was performed according to standard Roche 454 FLX and Titanium protocols. Briefly, mRNA was fragmented with a zinc chloride based RNA fragmentation solution. Double stranded (ds)-cDNA was synthesized from purified fragmented mRNA using Roche's cDNA synthesis system and random hexamer primers. Subsequent steps, including end-polishing of ds-cDNA, adapter ligation, removal of smaller fragments, and quality and quantity assessment of the final library, were performed as per standard instructions provided in cDNA rapid library preparation guide (Roche). For sequencing, adapter ligated DNA fragments were denatured to generate a single-stranded library, which was then amplified by emulsion PCR. Individual titrated single strand DNA libraries were sequenced in one half-plate run on the 454 GS-FLX platform using Titanium chemistry. For each species, on average about half a million reads were generated (**Supplementary Table 22**).

5.3. De novo sequence assembly

A rigorous read pre-processing pipeline was adapted, utilizing in turn (i) the in-built quality, adapter and length trimming tool in the CLCbio genomics workbench (http://www.clcbio.com), (ii) CD-HIT-454 read clustering program[16] and (iii) the SILVA comprehensive ribosomal RNA database[17]. Standard flowgram files (SFF) generated by 454 sequencing were initially imported

into the CLCbio genomics workbench 4.7 and reads were trimmed, by removing adapter sequences, ambiguous nucleotides and low quality sequence, using default settings. Filtered reads were clustered at 100% identity using CD-HIT-454, and exactly identical PCR duplicates were removed. Any reads shorter than 100 bp were also removed. Subsequently, ribosomal RNA reads were identified based on significant nucleotide similarity [BLASTN[18] with E-value cutoff of 1E-40, identity >85% and match length >50 bp] to the SILVA rRNA database and discarded.

The clean, unique and non-ribosomal reads were assembled using the 454 GS *De novo* Assembler software (Version 2.6). An incremental transcriptome (-cDNA flag) assembly approach was implemented through a command line interface with parameters, including isogroup threshold (-ig) of 10000, isotig threshold (-it) of 10000, isotig contig count threshold (-icc) of 200, minimum overlap length (-ml) of 125 and minimum identity (-mi) of 98%. Additionally, the use read tips (-urt) option that helps in obtaining contigs from rare or low coverage transcripts was activated.

For each assembly, the 454 GS *De novo* Assembler generated isogroup (analogous to genes) and corresponding isotig (analogous to individual transcripts) sequences. Different isotigs within an isogroup are considered to represent alternative splice variants of the same gene. In almost all transcript assemblies we noticed substantial redundancy among isotigs belonging to a subset of isogroups, some of which were nearly identical. The unusually inflated number of isotigs per isogroup is likely an artifact of the greedy assembly algorithm. To avoid bias in downstream analyses due to redundancy, only one isotig (longest) per isogroup was selected for subsequent phylogenetic analysis.

The 454 GS *De novo* assembler trims the polyA tails prior to assembling, so the true orientation of reads in the assembly cannot be determined and an isotig may be output as the reverse complement of the true biological transcript. The orientation of isotigs from each species was determined by aligning (BLASTX with E-value cutoff of 1E-20) them against the *A. thaliana* proteome database and improperly oriented isotigs were reverse complemented.

5.4. Phylogenetic analysis

Individual isotig sequences from each *Camelina* species and coding DNA sequences (CDS) of the three sub-genomes (Cs-G1/2/3) of *C. sativa,* were mapped (BLASTN with E-value cut off of

1E-06) onto full-length CDS of *A. thaliana*. The reciprocal best BLAST hit method was used to determine putative orthologues between *A. thaliana* and each of the *Camelina* species. Based on this analysis, a phylogenomic data matrix consisting of 4,867 unique orthologous gene sets was constructed. Sequences from individual orthologous gene sets were locally aligned using ClustalW[19]. Gaps and missing data from each alignment were removed using an automated alignment trimming tool trimAL[20] with a gap threshold (-gt) value set to 1. Trimmed alignments were concatenated using the Phyutility program[21] to produce the final data matrix comprising a total alignment length of 3,484,241 bps. Phylogenetic analysis was performed using the optimality criteria of Maximum Likelihood (ML) implemented in RAxML[22]. ML tree was calculated assuming GTR+GAMMA model of sequence evolution. Robustness of phylogenetic inference was assessed by running 1000 bootstrap replicates using GTR+CAT approximation. Final tree was visualized using the interactive Tree of Life [iTOL[23]] web server.

6. Construction of a 40 kb mate-pair library

Approximately 40 µg of high quality genomic DNA was sheared using Hydroshear device (Genomic solutions; shearing assembly size: large, volume: 140 µl, speed code: 25 and cycle number: 20). About 10 ug of sheared DNA was end-repaired to create blunt ends with 5' phosphate groups for ligation into the blunt, dephosphorylated fosmid vector. End-repaired DNA fragments were resolved by field inversion gel electrophoresis (FIGE; program: 7 and runtime: 20 hrs) and 30-75 kb fragments were recovered from the gel using Elutrap electroelution system (Whatman) and concentrated using Microcon-YM-50 columns. Eluted fragments were then ligated to the pNGS-FOS vector (Lucigen) and packaged *in vitro* using Gigapack III Gold (Agilent) bacteriophage lambda packaging extract. Packaged fosmids were transfected into the replicator fos strain (Lucingen) and plated on agar plates containing chloramphenicol and 1X arabinose induction medium. A pool of 300,000-500,000 fosmid colonies was removed en masse and recombinant fosmid DNA was purified using the Qiagen large construct kit. Next, purified fosmid library was digested with *Bfa*I restriction enzyme. The linearized vector and remnant anchored paired-ends from completely digested fosmid inserts were purified, and re-ligated to reconstitute the restriction site and create a mate-pair junction. Mate-pair insert DNA molecules were enriched for 454 sequencing by PCR amplification using a custom set of primers (Titan-A:

CCA TCT CAT CCC TGC GTG TCT CCG ACT CAG and Titan-B: CCT ATC CCC TGT GTG CCT TGG CAG TCT CAG).

7. Read filtering and correction

All Illumina reads were filtered using Trimmomatic (version 0.17) (http://www.usadellab.org/cms/index.php?page=trimmomatic) by (i) removing adapter sequences, (ii) trimming leading and trailing low quality sequences (below 15), (iii) scanning the read with a 4-base wide sliding window and cutting when the average quality per base dropped below 15, and (iv) keeping only those pairs where both reads were longer than 55 bp. Additionally, PCR duplicates and reads with ambiguous residues (Ns) were removed using custom Perl scripts.

454 pyrosequencing reads generated for 15-25 kb mate-pair libraries were pre-processed using SFFtoCA module (generated by 454 life sciences; http://sourceforge.net/apps/mediawiki/wgs-assembler/index.php?title=SffToCA), which removes perfect prefix duplicates and splits each read into two mate-paired reads based on the presence of the titanium linker sequence. Sequencing datasets generated for 40 kb mate-pair library by 454 sequencing were filtered by (i) removing vector and adapter sequences, (ii) splitting each read into mate-paired reads based on the presence of *Bfa*I restriction site, and (iii) keeping only those mate-pairs where both reads were longer than 50 bp.

8. Statistical test for single outlier

The statistical test for a single outlier was carried out after log transforming the x (Genome size) and y (Gene number) data and fitting the linear regression to it. The test statistic was the maximum of the absolute values of the residual (y – the predicted value of y) divided by its estimated standard deviation. The formula for the estimated standard deviation of the residuals and the critical values of the test statistic under the assumption that the residuals are independent can be found in Barnett and Lewis[24] and Tietjen Moore and Beckman[25]. The critical values (for which there are slight discrepancies) for the test reported here were obtained from Barnett and Lewis[24]. The curves for the confidence intervals in Figure 2 are the sets of points for which if *C. sativa* was placed at that point, the most outlying point (*C. sativa*) is just significant at the 5% and 1% levels.

9. Mixture model analysis of *Ks* distributions

Estimation of the level of synonymous substitution (*Ks*) between paralogous and orthologous sequences was carried out using the Bioperl[26] script bp_pairwise_kaks.pl, which takes a pair of sequences as an input, translates them into their corresponding protein sequences, aligns the protein sequences using ClustalW[19], and then using the protein alignments together with the cDNA sequences calculates the *Ka*, *Ks*, and *Ka/Ks* ratio by implementing the codeml method of the PAML distribution[27]. Synonymous substitution rate of 8.22×10^{-9} substitutions/synonymous site/year for Brassicaceae species was extrapolated based on a fossil established age of 22.5 Mya for the whole genome triplication event in diploid brassicas[28]. Mixture model analysis of *Ks* distributions was performed to account for multiple duplication events within each lineage.

Each *Ks* dataset was trimmed by removing values of 0.001 or less because they are affected by rounding errors and result in spurious frequency peaks. Histograms were generated using log transformed *Ks* values with a bin width of 0.1 (**Supplementary Fig. 10**). To identify significant peaks in the ln (*Ks*) distribution, Gaussian mixture models were fitted to the ln (*Ks*) values using the R package Mclust, and the number of Gaussian components (*G*), the mean of each component and corresponding variance, standard deviation and fractions of data were calculated. The Bayesian Information Criterion (BIC) was used to determine the best fitting model to the data. The number of Gaussian components was increased to improve the fit, if necessary. All *Ks* data were included to develop the best fitting model and this allowed the detection of weak Gaussian components with means up to about 20. *Ks* values beyond 20 are difficult to meaningfully interpret and we restricted the display of our results to this boundary.

The fit of the determined models were confirmed by χ^2 tests. The upper limit of ln (*Ks*) = 3 in the χ^2 calculation was used to cut-off the *Ks* values beyond ~20. The number of degrees of freedom for the model was estimated as $3(G-1)$ for the single species datasets (that had a spurious peak around *Ks*=68) and $3G$ for the interspecific comparisons accounting for the number of parameters that significantly affect the model for *Ks* <20.

In the comparison involving all paralogous and orthologous gene pairs, skewed peaks were observed in the distribution of *Ks* values (**Supplementary Fig. 10**), this deviation resulted from

the combined effects of multiple overlapping Gaussian components. These complications were resolved by combining overlapping Gaussian components and representing them with the mean and standard deviation of their Gaussian mixtures (**Supplementary Table 23**). Thus for a Gaussian mixture distribution $x \sim \sum_i \pi_i N(\mu_i, \sigma_i^2)$ with proportions π_i that sum to 1, with means μ_i, and standard deviations σ_i, we determine the combined mean as $\bar{x} = \sum_i \pi_i \mu_i$ and the combined standard deviation as $\sqrt{V(x)} = \sqrt{\sum_i \pi_i(\sigma_i^2 + \mu_i^2) - (\sum_i \pi_i \mu_i)^2}$ enabling these peaks to be plotted as single points (**Supplementary Fig. 10**). The number of Gaussian components, Chi-square statistics, the ln-transformed mean of each component and corresponding variance and standard deviation are provided in **Supplementary Table 23**.

Supplementary References

1. Luo, R, *et al.* SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. *Gigascience* **1**, 18 (2012).
2. Pop, M., Kosack, D.S. & Salzberg, S.L. Hierarchical scaffolding with Bambus. *Genome Res.* **14**, 149-159 (2004).
3. Galbraith, D.W., Harkins, K.R., Maddox, J.M., Ayres, N.M., Sharma, D.P. & Firoozabady, E. Rapid flow cytometric analysis of the cell cycle in intact plant tissues. *Science* **220**, 1049-1051 (1983).
4. Li, R., *et al.* The sequence and de novo assembly of the giant panda genome. *Nature* **463**, 311-317 (2010).
5. Marcais, G. & Kingsford, C. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics* **27**, 764-770 (2011).
6. Hutcheon, C., *et al.* Polyploid genome of *Camelina sativa* revealed by isolation of fatty acid synthesis genes. *BMC Plant Biol.* **10**, 233 (2010).
7. Kurtz, S., *et al.* Versatile and open software for comparing large genomes. *Genome biol.* **5**, R12 (2004).
8. Langmead, B. & Salzberg, S.L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357-359 (2012).
9. Wu, T.D. & Watanabe, C.K. GMAP: a genomic mapping and alignment program for mRNA and EST sequences. *Bioinformatics* **21**, 1859-1875 (2005).
10. Parra, G., Bradnam, K. & Korf, I. CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. *Bioinformatics* **23**, 1061-1067 (2007).
11. Tautz, D. & Domazet-Loso, T. The evolutionary origin of orphan genes. *Nat. Rev. Genet.* **12**, 692-702 (2011).
12. Guo, W.J., Li, P., Ling, J. & Ye, S.P. Significant comparative characteristics between orphan and nonorphan genes in the rice (*Oryza sativa* L.) genome. *Comp. Func. Genomics* **2007**, 21676 (2007).
13. Varshney, R.K., *et al.* Draft genome sequence of pigeonpea (*Cajanus cajan*), an orphan legume crop of resource-poor farmers. *Nat. Biotechnol.* **30**, 83-89 (2012).
14. Donoghue, M.T., Keshavaiah, C., Swamidatta, S.H. & Spillane, C. Evolutionary origins of Brassicaceae specific genes in *Arabidopsis thaliana*. *BMC Evol. biol.* **11**, 47 (2011).
15. Wilson, G.A., Feil, E.J., Lilley, A.K. & Field, D. Large-scale comparative genomic ranking of taxonomically restricted genes (TRGs) in bacterial and archaeal genomes. *PLoS ONE* **2**, e324 (2007).
16. Li, W. & Godzik, A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* **22**, 1658-1659 (2006).
17. Quast, C, *et al.* The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Res.* **41**, D590-596 (2013).
18. Altschul, S.F., Gish, W., Miller, W., Myers, E.W. & Lipman, D.J. Basic local alignment search tool. *J. Mol. Biol.* **215**, 403-410 (1990).
19. Larkin, M.A., *et al.* Clustal W and Clustal X version 2.0. *Bioinformatics* **23**, 2947-2948 (2007).
20. Capella-Gutierrez, S., Silla-Martinez, J.M. & Gabaldon, T. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* **25**, 1972-1973 (2009).

21. Smith, S.A. & Dunn, C.W. Phyutility: a phyloinformatics tool for trees, alignments and molecular data. *Bioinformatics* **24**, 715-716 (2008).
22. Stamatakis, A. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics* **22**, 2688-2690 (2006).
23. Letunic, I. & Bork, P. Interactive Tree Of Life v2: online annotation and display of phylogenetic trees made easy. *Nucleic Acids Res.* **39**, W475-478 (2011).
24. Barnett, V. & Lewis, T. Outliers in Statistical Data" (2nd Edition), Wiley series in Probability and Mathematical Statistics. (1978).
25. Tietjen, G.L., Moore, R.H. & Beckman, R.J. Testing for a single outlier in simple linear regression. *Technometrics* **15**, 717-721 (1973).
26. Stajich, J.E., *et al.* The Bioperl toolkit: Perl modules for the life sciences. *Genome Res.* **12**, 1611-1618 (2002).
27. Yang, Z. PAML 4: phylogenetic analysis by maximum likelihood. *Mol. biol. evol.* **24**, 1586-1591 (2007).
28. Beilstein, M.A., Nagalingum, N.S., Clements, M.D., Manchester, S.R. & Mathews, S. Dated molecular phylogenies indicate a Miocene origin for *Arabidopsis thaliana*. *Proc. Natl. Acad. Sci. USA* **107**, 18724-18728 (2010).
29. Johnston, J.S. *et al.* Evolution of genome size in Brassicaceae. *Ann. Bot.* **95**, 229-235 (2005).
30. Haudry, A., *et al.* An atlas of over 90,000 conserved noncoding sequences provides insight into crucifer regulatory regions. *Nat Genet*. 45,891–898 (2013).
31. Lysak, M.A., Koch, M.A., Beaulieu, J.M., Meister, A. & Leitch, I.J. The dynamic ups and downs of genome size evolution in Brassicaceae. *Mol. Biol.Evol.* **26**, 85-98 (2009).
32. Wu, H.J. *et al.* Insights into salt tolerance from the genome of Thellungiella salsuginea. *Proc. Natl. Acad. Sci. USA* **109**, 12219-12224 (2012).
33. Dassanayake, M. *et al.* The genome of the extremophile crucifer Thellungiella parvula. *Nat. Genet.* **43**, 913-918 (2011).
34. Koch, M.A. *et al.* BrassiBase: Tools and biological resources to study characters and traits in the Brassicaceae-version 1.1. *Taxon* **61**, 1001-1009 (2012).
35. Cantarel, B.L. *et al.* MAKER: an easy-to-use annotation pipeline designed for emerging model organism genomes. *Genome Res.* **18**, 188-196 (2008).
36. Haas, B.J. *et al.* Improving the Arabidopsis genome annotation using maximal transcript alignment assemblies. *Nucleic Acids Res.* **31**, 5654-5666 (2003).