Supplementary Information for:

The genome of the extremophile crucifer *Thellungiella parvula*

Maheshi Dassanayake[1,9], Dong-Ha Oh[1,9], Jeffrey S. Haas[1,2], Alvaro Hernandez[3], Hyewon Hong[1,4], Shahjahan Ali[5], Dae-Jin Yun[4,6], Ray A. Bressan[4,6,7], Jian-Kang Zhu[6,7], Hans J. Bohnert[1,4,7,8] and John M. Cheeseman[1]

[1]Department of Plant Biology, University of Illinois at Urbana-Champaign, Urbana, IL 61801, USA. [2]Office of Networked Information Technology, School of Integrative Biology, University of Illinois at Urbana-Champaign, Urbana, IL 61801, USA. [3]Center for Comparative & Functional Genomics, University of Illinois at Urbana-Champaign, Urbana, IL 61801, USA. [4]Division of Applied Life Science, Gyeongsang National University, Jinju, Korea 660–701. [5]Bioscience Core Laboratory-Genomics, King Abdullah University of Science and Technology, Thuwal 21534, Saudi Arabia. [6]Department of Horticulture & Landscape Architecture, Purdue University , West Lafayette, Indiana 47907, USA. [7]Center for Plant Stress Genomics and Biotechnology, King Abdullah University of Science and Technology, Thuwal 21534, Saudi Arabia. [8]Department of Crop Sciences, University of Illinois at Urbana-Champaign, Urbana, IL 61801, USA. [9]These authors contributed equally to this work. Correspondence should be addressed to M.D. (maheshi.dassanayake@gmail.com), D.-H.O. (ohdongha@gmail.com) or D.-J.Y. (djyun@gnu.ac.kr)

Nature Genetics: doi:10.1038/ng.889

List of Supplementary Information:

Supplementary Table 1 Overview of the *Thellungiella parvula* assembly input sequences

Supplementary Table 2 Classification of sequences in the *T. parvula* draft genome

***Supplementary Table 3** Repetitive sequences in *T. parvula* draft genome

***Supplementary Table 4** List of *T. parvula* predicted ORFs and their annotations

Supplementary Table 5 Overview of the BLASTn results of the predicted *T. parvula* ORFs with NCBI nucleotide (nt) database

Supplementary Table 6 Overview of the largest 20 *T. parvula* contigs

***Supplementary Table 7** List of non-coding RNAs in *T. parvula* draft genome

***Supplementary Table 8** List and comparison of GO annotations of the *T. parvula* predicted ORFs and *A. thaliana* cDNAs

***Supplementary Table 9** Tandem local duplications in the *T. parvula* draft genome and the *A. thaliana* genome

***Supplementary Table 10** Assignments of the largest 40 *T. parvula* contigs in seven chromosomes

Supplementary Figure 1 Overview of the *T. parvula* genome sequencing strategy

Supplementary Figure 2 Colinearity between *T. parvula* contigs and *B. rapa* A3 chromosome

Supplementary Figure 3 Comparison of *T. parvula* predicted ORFs with *A. thaliana* cDNAs

Supplementary Figure 4 Comparison of *T. parvula* contigs c2 and c3 with *A. thaliana* chromosomes

*Supplementary Excel files.

Supplementary Table 1 Overview of the *Thellungiella parvula* assembly input sequences

Total number of reads sequenced	41.95 million
Total number of bases sequenced	7.82 Gb
Total number of reads produced by 454	26.86 million
Total number of bases produced by 454	6.61Gb
Average 454 read size	355 bp
Number of single-end 454 reads	17.46 million
Number of high quality 454 reads used in the Newbler assembly	17.74 million
Number of paired-end reads	9.4 million
Number with both pairs mapped with correct orientation	4.2 million
Fraction of bases in 454 contigs Q40 or greater	99.18%
Total number of reads produced by Illumina	15.09 million
Total number of bases produced by Illumina	1.21 Gb
Average Illumina read size	80 bp
Number of high quality Illumina reads used in the ABySS assemblies	14.37 million
Fraction of cleaned Illumina reads uniquely mapped to the genome	88%
Fraction of cleaned 454 reads uniquely mapped to the genome	53%

Supplementary Table 2 Classification of sequences in the *T. parvula* draft genome

	Length occupied (bp)	% occupied
Protein-coding regions	60,973,281	44.48%
Exons	36,171,204	26.39%
Introns	24,802,077	18.09%
Repetitive elements	10,227,489	7.46%
Retroelements	7,604,114	5.55%
SINEs	26,337	0.02%
LINEs	1,482,882	1.08%
Copia	2,634,223	1.92%
Gypsy	3,368,178	2.46%
DNA transposons	1,581,042	1.15%
Other repeats	1,042,333	0.76%
ncRNAs	86,596	0.06%
Total	137,086,937	

Supplementary Table 5 Overview of the BLASTn results of the predicted *T. parvula* ORFs with NCBI nucleotide (nt) database

Family	Genus	Species	Number of predicted ORFs
Brassicaceae	*Arabidopsis*	*Arabidopsis lyrata*	15438
		Arabidopsis thaliana	8340
		Arabidopsis arenosa	17
		Arabidopsis halleri	10
	Brassica	*Brassica rapa*	1447
		Brassica napus	327
		Brassica oleracea	121
		Brassica juncea	49
		Brassica carinata	7
	Thellungiella	*Thellungiella halophila*	73
	Other *Brassicaceae*	*Sisymbrium irio*	59
		Raphanus sativus	27
		Noccaea caerulescens	22
		Isatis tinctoria	17
		Sinapis alba	14
		Boechera divaricarpa	9
		Arabis hirsuta	6
		Thlaspi goesingense	5
		Other *Brassicaea* species	42
Other plants		*Zea mays*	242
		Vitis vinifera	39
		Populus trichocarpa	18
		Ricinus communis	18
		Solanum lycopersicum	9
		Glycine max	6
		Lotus japonicus	6
		Oryza sativa	2
Other			91
no hit			2440
Total			28901

Shown are the numbers of ORFs which have the best hits with a nucleotide sequence from each species.

Supplementary Table 6 Overview of the largest 20 *T. parvula* contigs

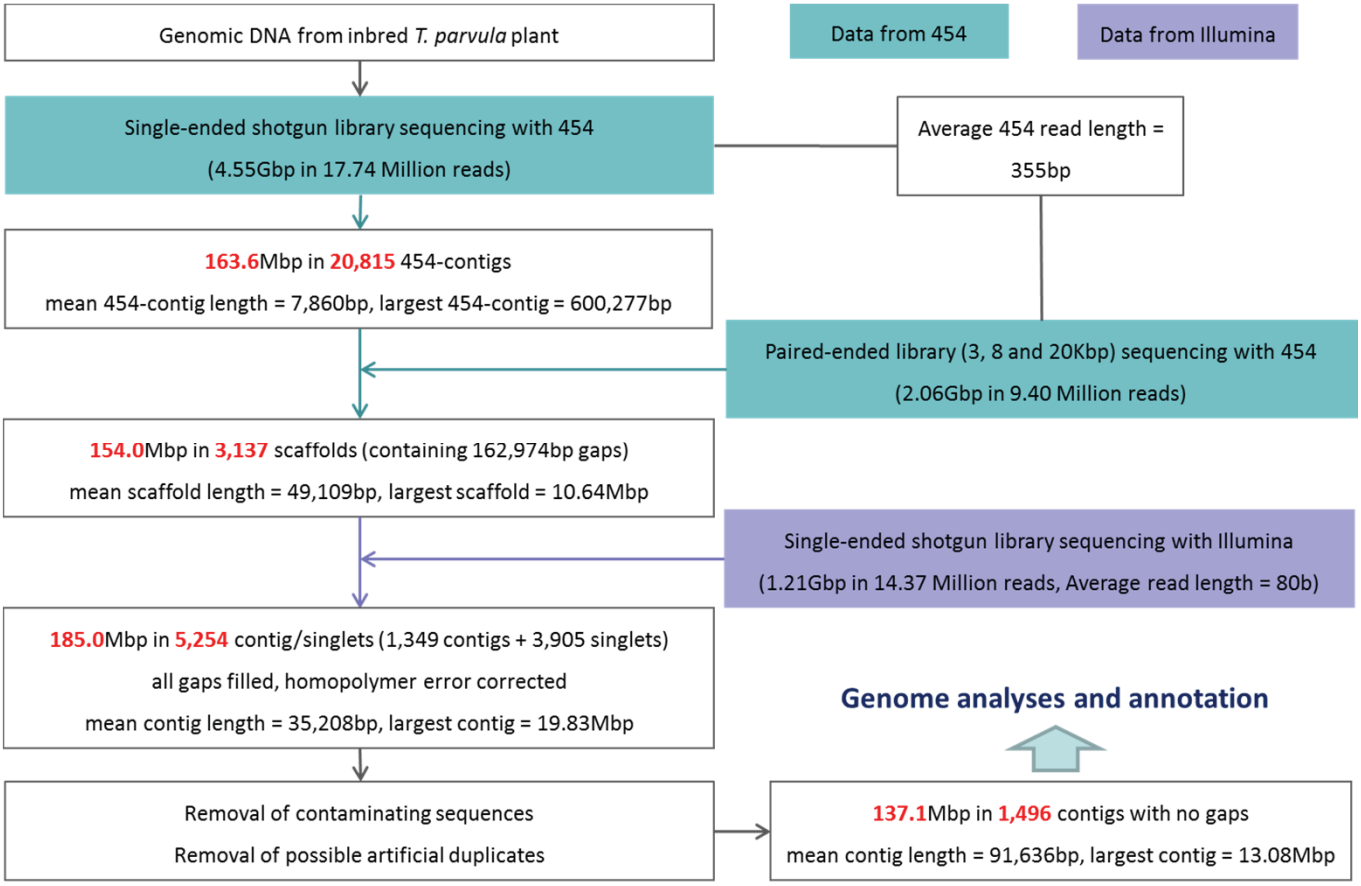
contig	length	# of ORFs	ORF frequency*	First hits by BLASTn found in**						% of repetitive sequences
				At Chr1	At Chr2	At Chr3	At Chr4	At Chr5	Not in At	
c1	13,079,760	2,839	4,607	2,561	34	31	34	41	138	3.91
c2	11,244,833	2,602	4,322	25	11	29	20	2,482	35	1.10
c3	9,613,411	2,170	4,430	41	216	334	26	1,487	66	2.40
c4	8,335,336	1,965	4,242	40	1,798	41	17	28	41	1.31
c5	7,893,823	2,093	3,772	32	17	1,974	17	29	24	0.73
c6	7,884,501	1,604	4,916	33	24	43	1,295	51	158	7.71
c7	6,749,000	1,507	4,478	14	34	1,377	11	38	33	2.41
c8	5,288,824	969	5,458	726	15	36	20	25	147	10.70
c9	3,484,480	504	6,914	29	257	20	15	13	170	19.98
c10	3,074,916	449	6,848	27	195	13	23	20	171	21.49
c11	3,007,254	641	4,692	22	8	7	560	16	28	3.48
c12	2,875,200	631	4,557	577	4	15	8	10	17	1.33
c13	2,863,558	747	3,833	711	10	6	5	7	8	0.71
c14	2,732,121	694	3,937	3	16	7	653	7	8	0.40
c15	2,574,192	437	5,891	322	13	12	6	17	67	10.58
c16	2,364,423	489	4,835	20	20	26	7	349	67	10.86
c17	1,885,384	489	3,856	6	10	1	466	3	3	0.32
c18	1,662,689	212	7,843	8	3	8	71	9	113	27.88
c19	1,604,931	245	6,551	11	4	5	17	92	116	25.64
c20	1,554,986	358	4,344	14	298	11	8	10	17	2.85
Total	99,773,622	21,645	5,016	5,222	2,987	3,996	3,279	4,734	1,427	8.00

* Numbers indicate the average length (bp) per ORF occurrence.

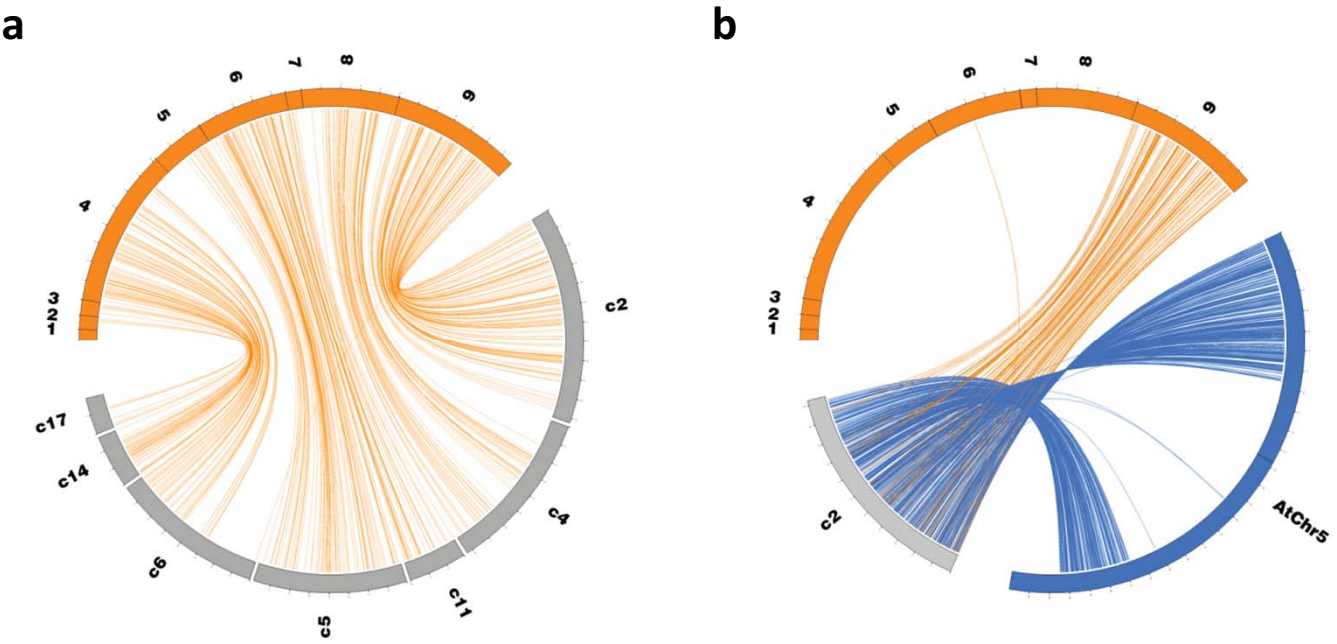
**BLASTn was performed against the TAIR9 cDNA database for *A. thaliana.*

Categories of ORFs constituting the major part of each contig are shaded.

Supplementary Figure 1 Overview of the *T. parvula* genome sequencing strategy



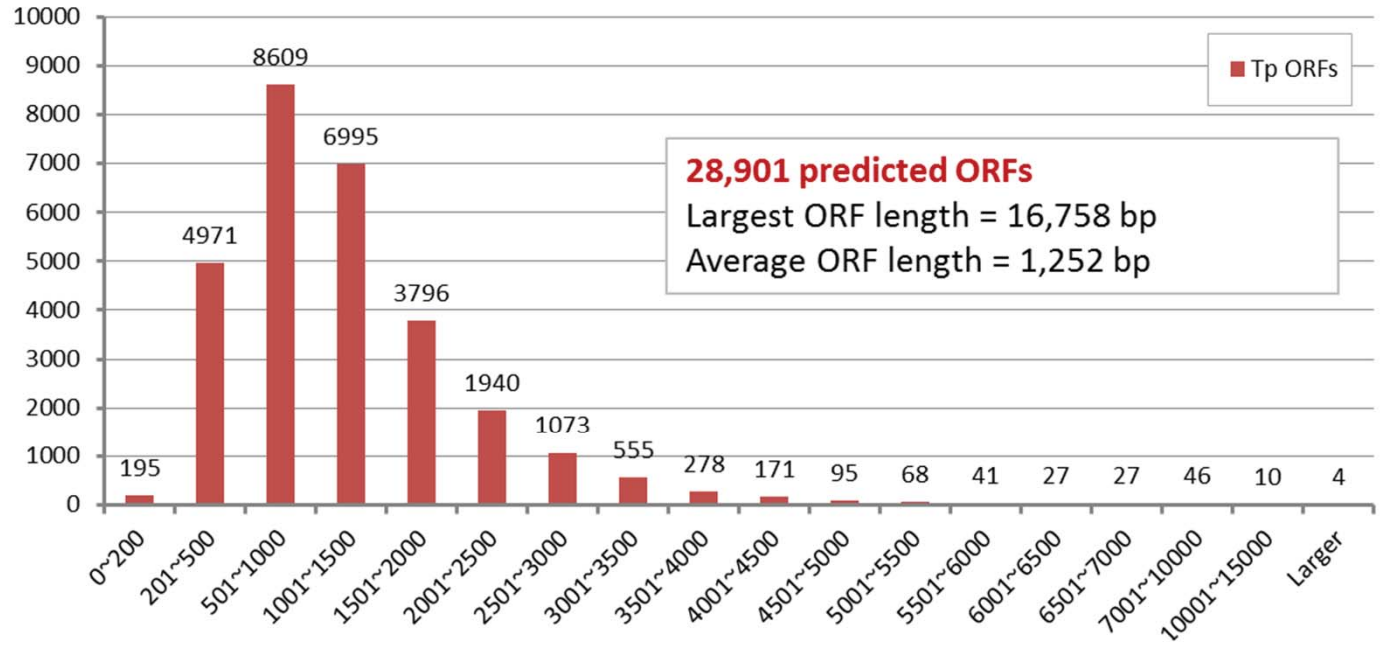
Supplementary Figure 2 Colinearity between *T. parvula* contigs, *A.thaliana* chromosome 5 and *B. rapa* A3 chromosome

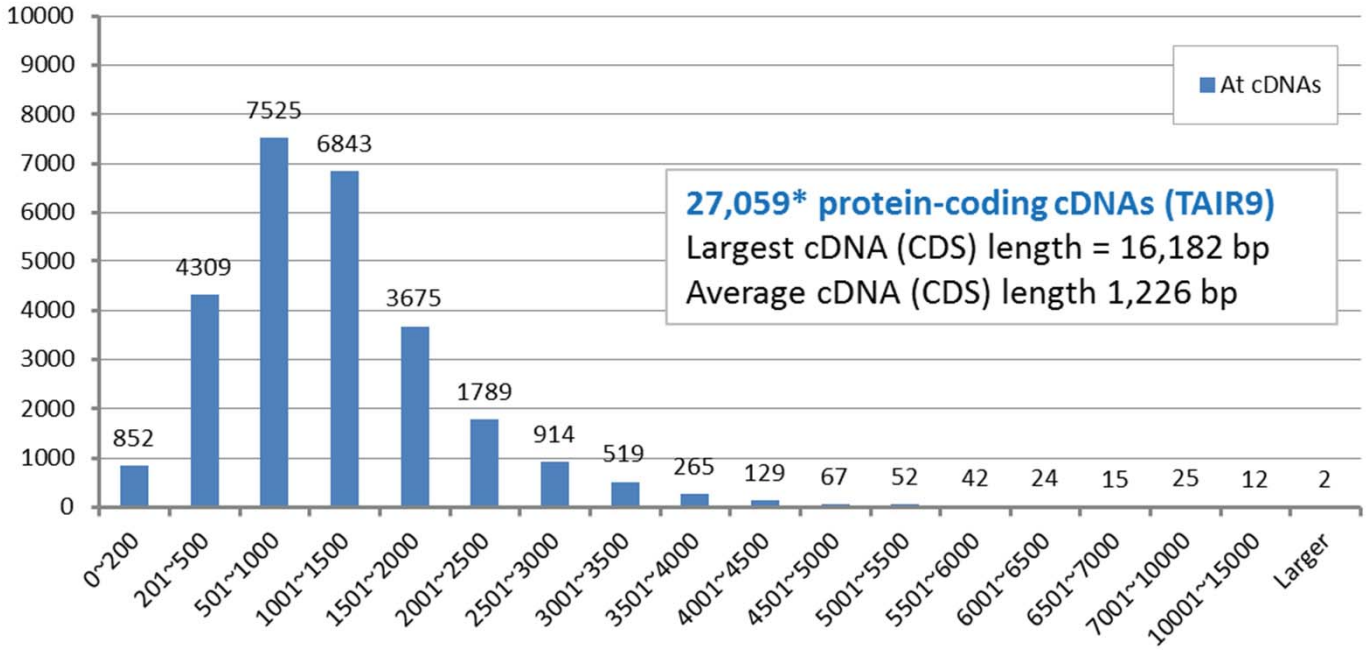


Alignment of the *B. rapa* chromosome A3 and (**a**) *T. parvula* contigs or (**b)** the *T. parvula* contig c2 and *A. thaliana* chromosome 5. The links are identified as connecting genomic regions showing 75% similarity over a minimum of 2,000 bp with maximum gap allowance of 1,000 bp. The *B. rapa* chromosome A3, *A. thaliana* chromosome 5 and *T. parvula* contigs are indicated as orange, blue and gray blocks, respectively.

The *B. rapa* chromosome A3 can be covered with *T. parvula* contigs showing co-linearity over the entire chromosome (**a**). *T. parvula* contig c2 showed synteny with *A. thaliana* chromosome 5 and *B. rapa* chromosome A3. However, macro-scale rearrangements were found only with *A. thaliana* chromosome 5.

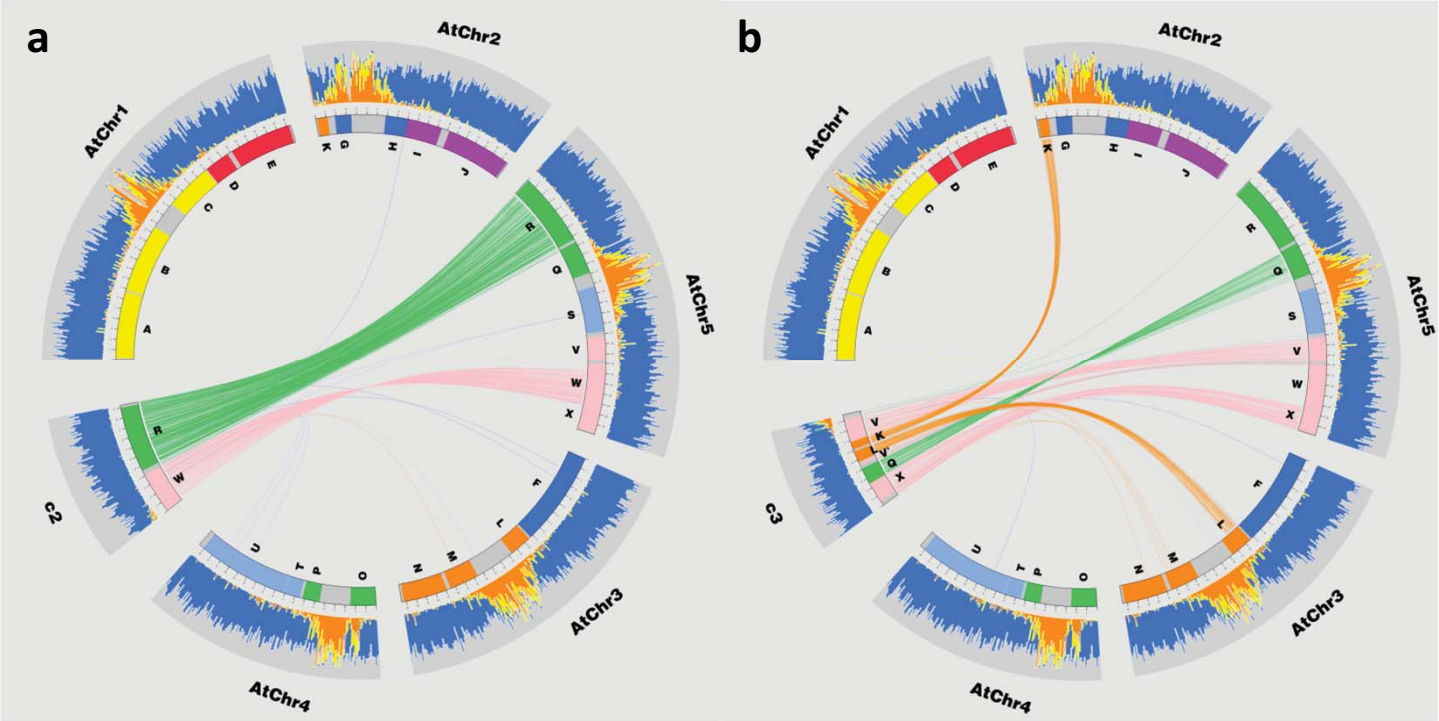
Supplementary Figure 3 Comparison of *T. parvula* predicted ORFs with *A. thaliana* cDNAs





Distribution of the lengths of the predicted *T. parvula* ORFs (Tp ORFs) and *A. thaliana* protein-coding cDNAs (At cDNAs). Numbers for *A. thaliana* cDNAs are based on the annotation in TAIR9, counting protein-coding genes excluding mitochondrial/chloroplast genes and overlapping ORFs (ftp://ftp.arabidopsis.org/home/tair/Sequences/blast_datasets/TAIR9_blastsets/TAIR9_cdna_20090619). The much lower number of ORFs at size 0-200 nucleotides in *T. parvula* is based on the conservative cut-off (i.e. minimum of predicted ORF length = 150bp) that was chosen by the ORF prediction programs. Actual numbers may approach the numbers seen in *Arabidopsis*.

Supplementary Figure 4 Comparison of *T. parvula* contigs c2 and c3 with *A. thaliana* chromosomes



Compared are *T. parvula* contigs **(a)** c2 and **(b)** c3 with the ancestral karyotype (AK) segments [1,2] in *A. thaliana* chromosomes.

The AK segments in both *A. thaliana* chromosomes and *T. parvula* contigs were presented with colors as defined in Figure 4A in the main article and by Lysak et al. [1,2]. Regions containing more than 75% similarity over a minimum of 2,000bp with maximum gap allowance of 1,000bp, were connected with lines of colors matching those used for coloring the AK segments. Ticks in each chromosome/contig block indicate lengths in 1Mb. The distributions of protein coding regions and repetitive sequences are shown in the outer circles, with the percentage of protein coding genes, DNA transposons and retrotransposons shown in blue, yellow and orange, respectively, with a window size of 0.1Mb. In the *T. parvula* contigs, predicted protein coding genes without BLASTn hits (e value <0.0001) against the *A. thaliana* cDNA database are shown in green color.

1. Lysak, M.A. & Koch, M.A. Phylogeny, Genome, and Karyotype Evolution of Crucifers (*Brassicaceae*). *Genetics and Genomics of the Brassicaceae* pp. 1–31 (Springer, New York, 2011)

2. Mandáková, T. & Lysak, M.A. Chromosomal phylogeny and karyotype evolution in x=7 crucifer species (*Brassicaceae*). *Plant Cell* **20**, 2559-2570 (2008).

Nature Genetics: doi:10.1038/ng.889