

Supplemental Information

Draft genome sequence of first monocot-halophytic species *Oryza coarctata* reveals stress-specific genes

Tapan Kumar Mondal*#, Hukam Chand Rawal#, Soni Chowrasia, Deepti Varshney, Alok Panda, Abhishek Majumdar, Harmeet Kaur, Kishor Gaikwad, Tilak Raj Sharma, Nagendra Kumar Singh.

ICAR-National Research Centre on Plant Biotechnology, Pusa, New Delhi-110012, India

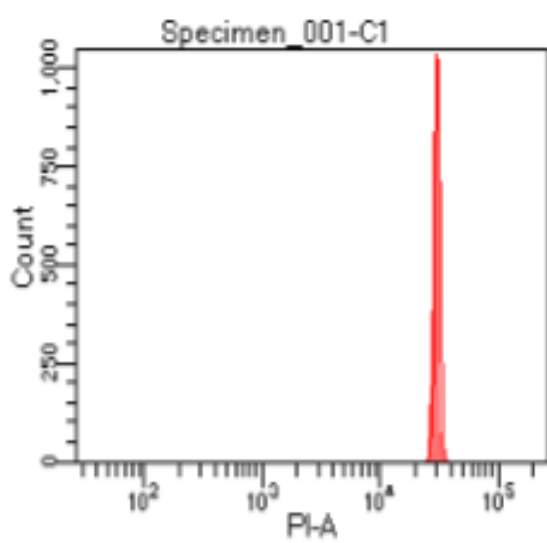
*corresponding author (mondaltk@rediffmail.com) ; #contributed equally.

Table of contents

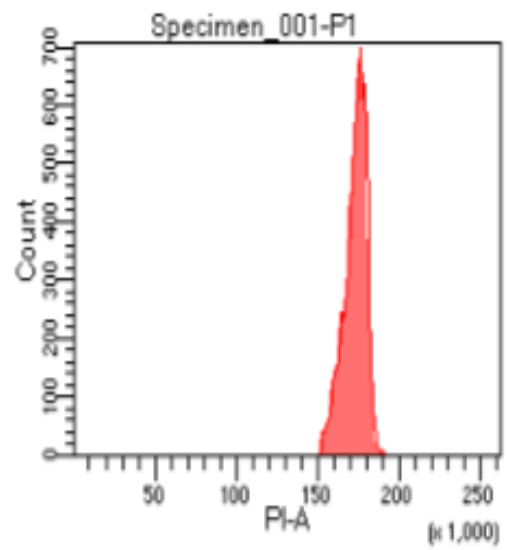
Supplementary figures: Figure S1 to S7	Page 1 to 7
Supplementary tables: Table S1 to S16	Page 8 to 19



a



P. sativum



O. coarctata

b

Figure S1a. Pot grown young *O. coarctata* plants, leaf of which was used for genome sequencing, **S1b.** Flow cytometry results revealed the estimated genome size for *O. coarctata* as to be of 665 Mb with *Pisium sativum* was used as standard. X axis and Y axis represents relative DNA content and relative fluorescence intensity

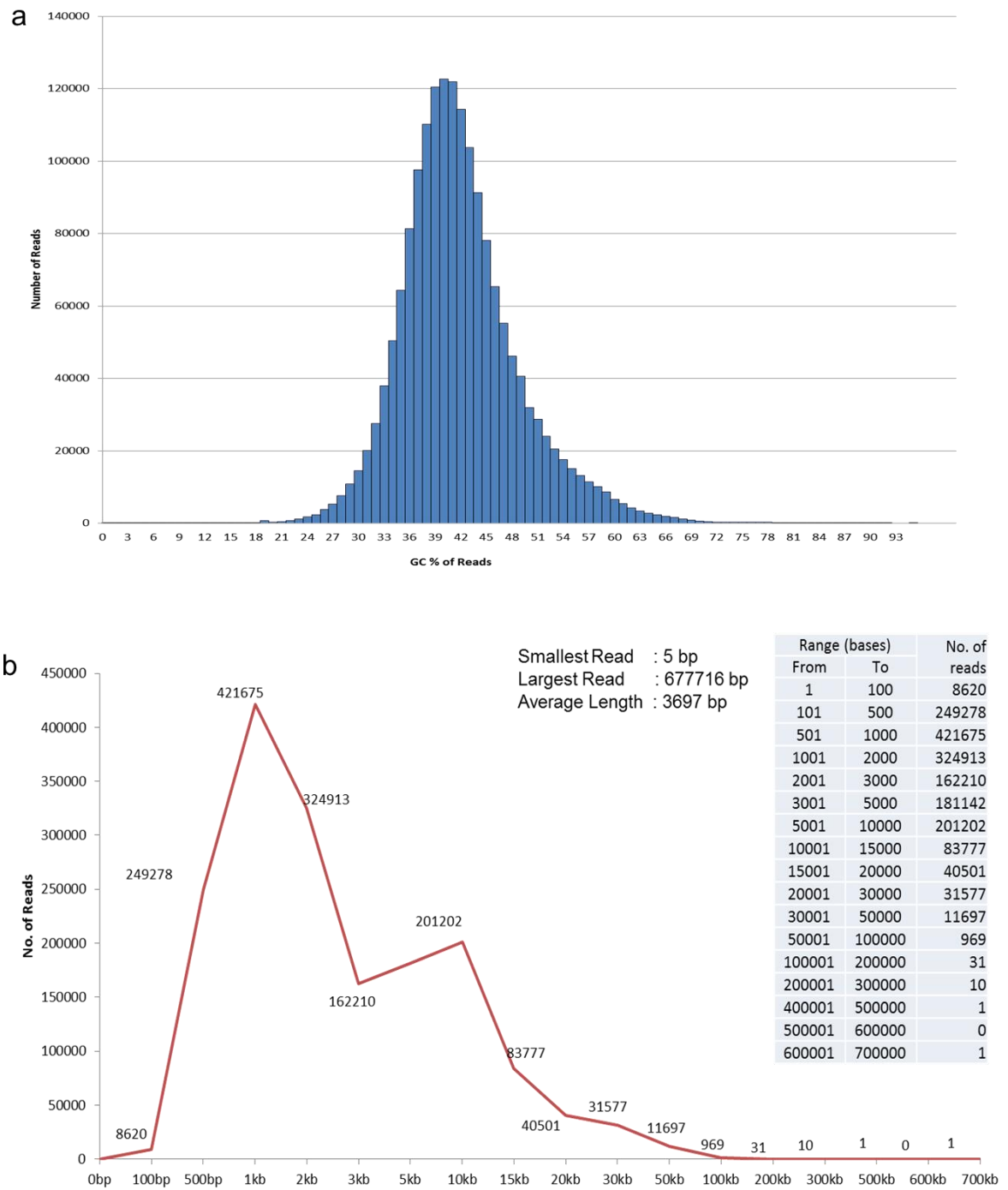


Figure S2a. GC content in Nanopore reads; **S2b.** Read length distribution of Nanopore reads

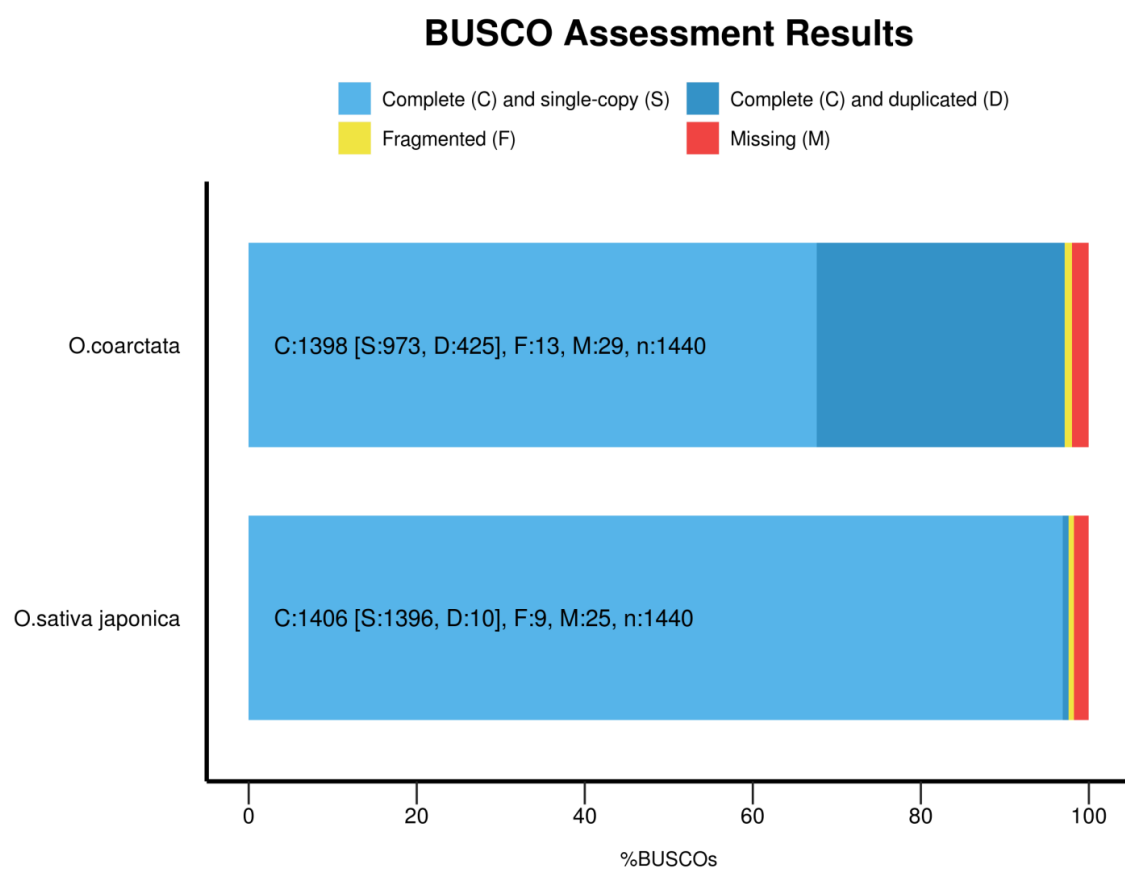


Figure S3. Completeness of *O. coarctata* and *O. sativa japonica* genome assemblies based on BUSCO analysis.

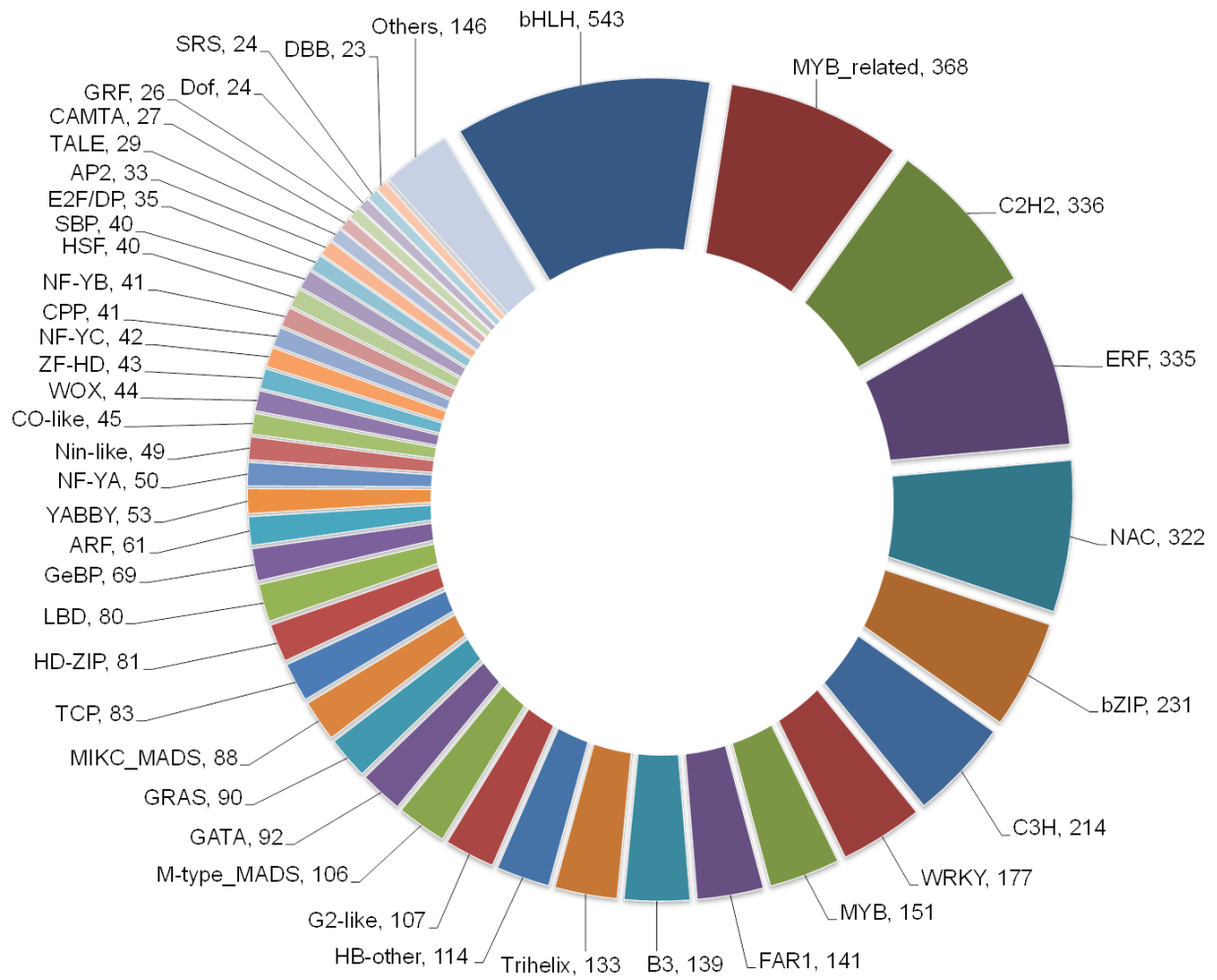


Figure S4. Transcription factor related genes in *O. coarctata* genome with high abundance of the stress-response related categories like MYB, MYB_related, NAC, bZIP, WRKY, HB, HSF and AP.

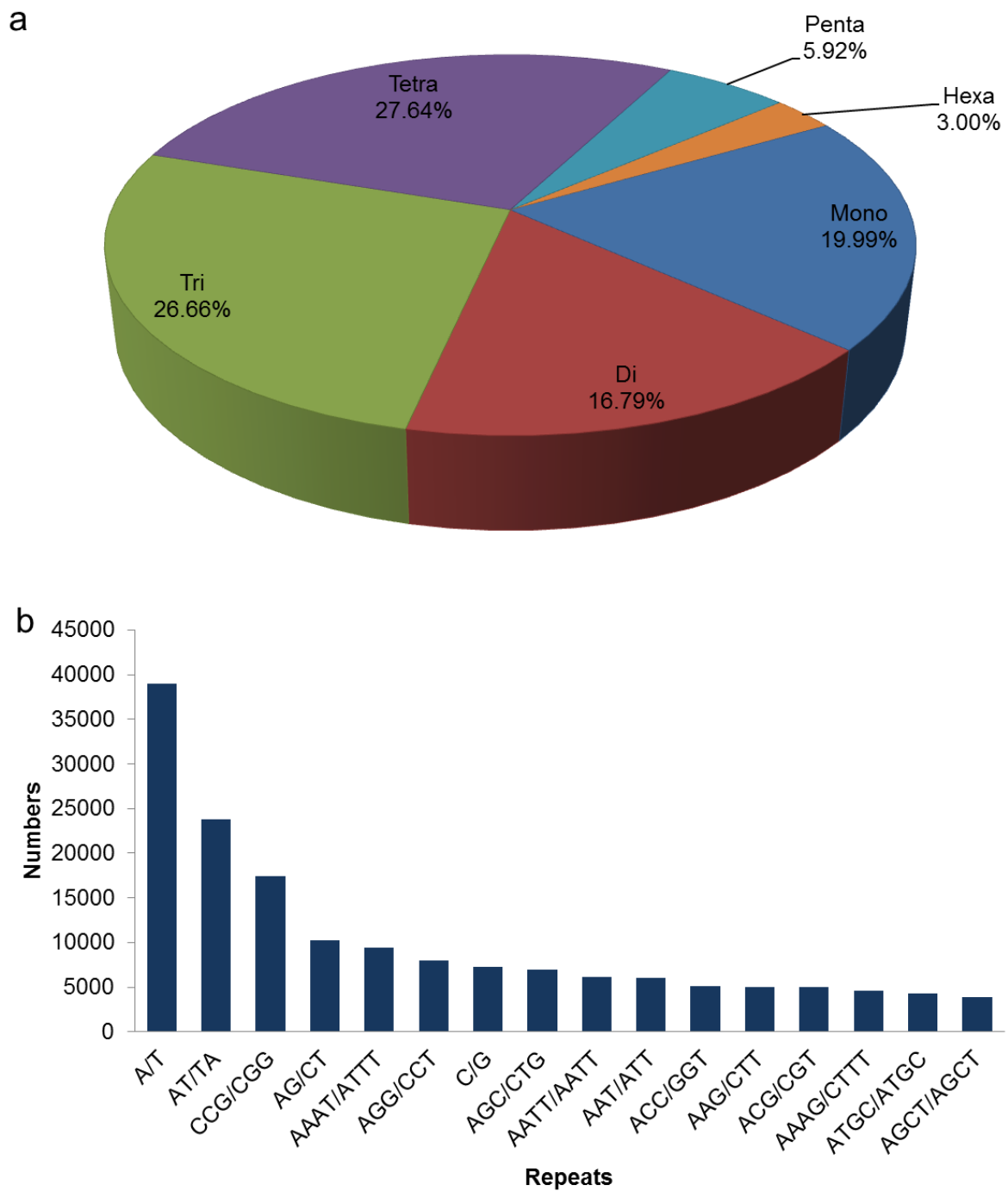


Figure S5a. Percentage distribution of identified SSRs into mono-, di-, tri-,tetra-, penta- and hexa-nucleotides; **S5b.** Distribution of mono-, di- and tri- types of SSRs

a

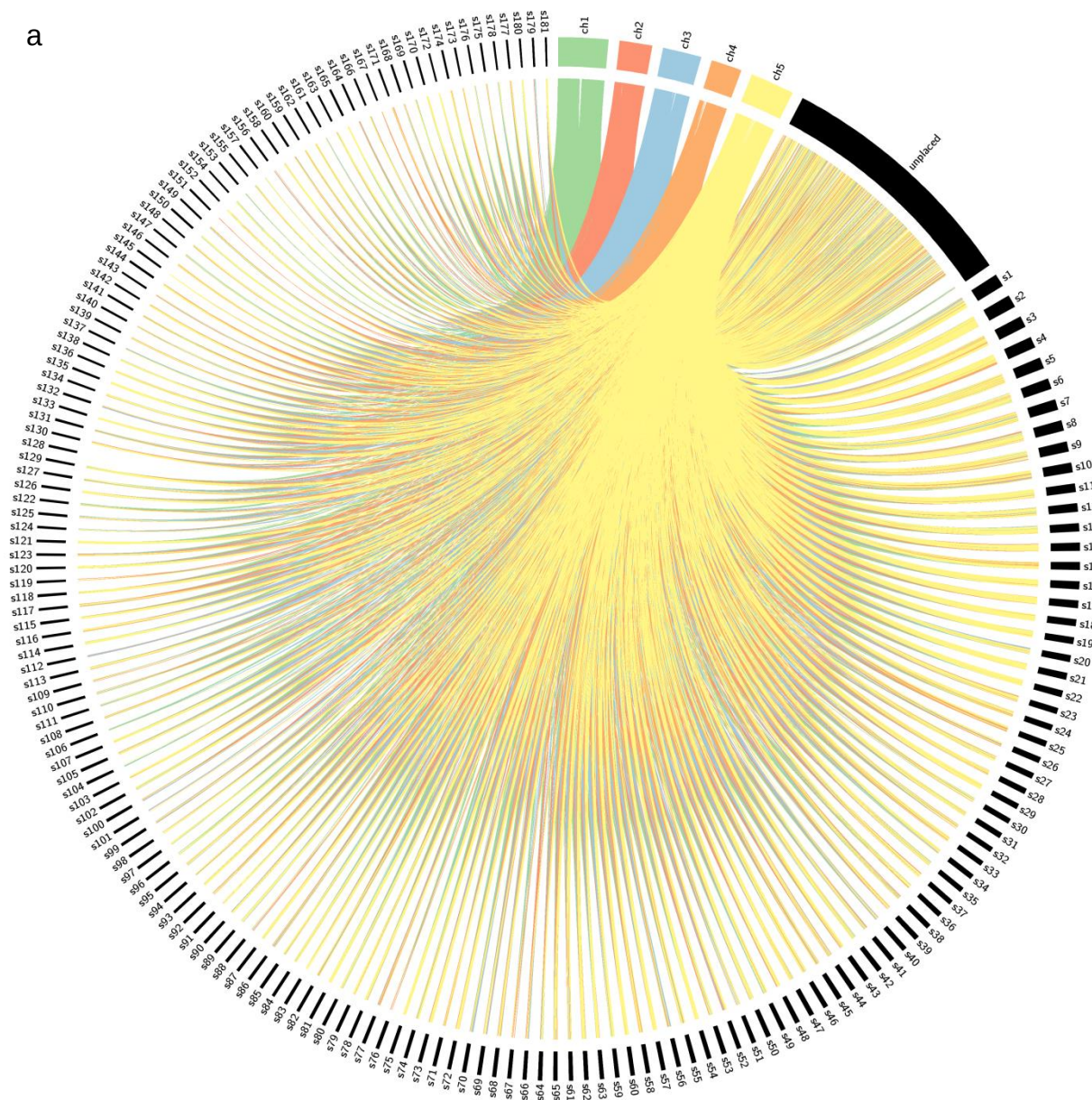


Figure S6a. Syntenic plots showing homologous blocks between *O. coarctata* and *A. thaliana*

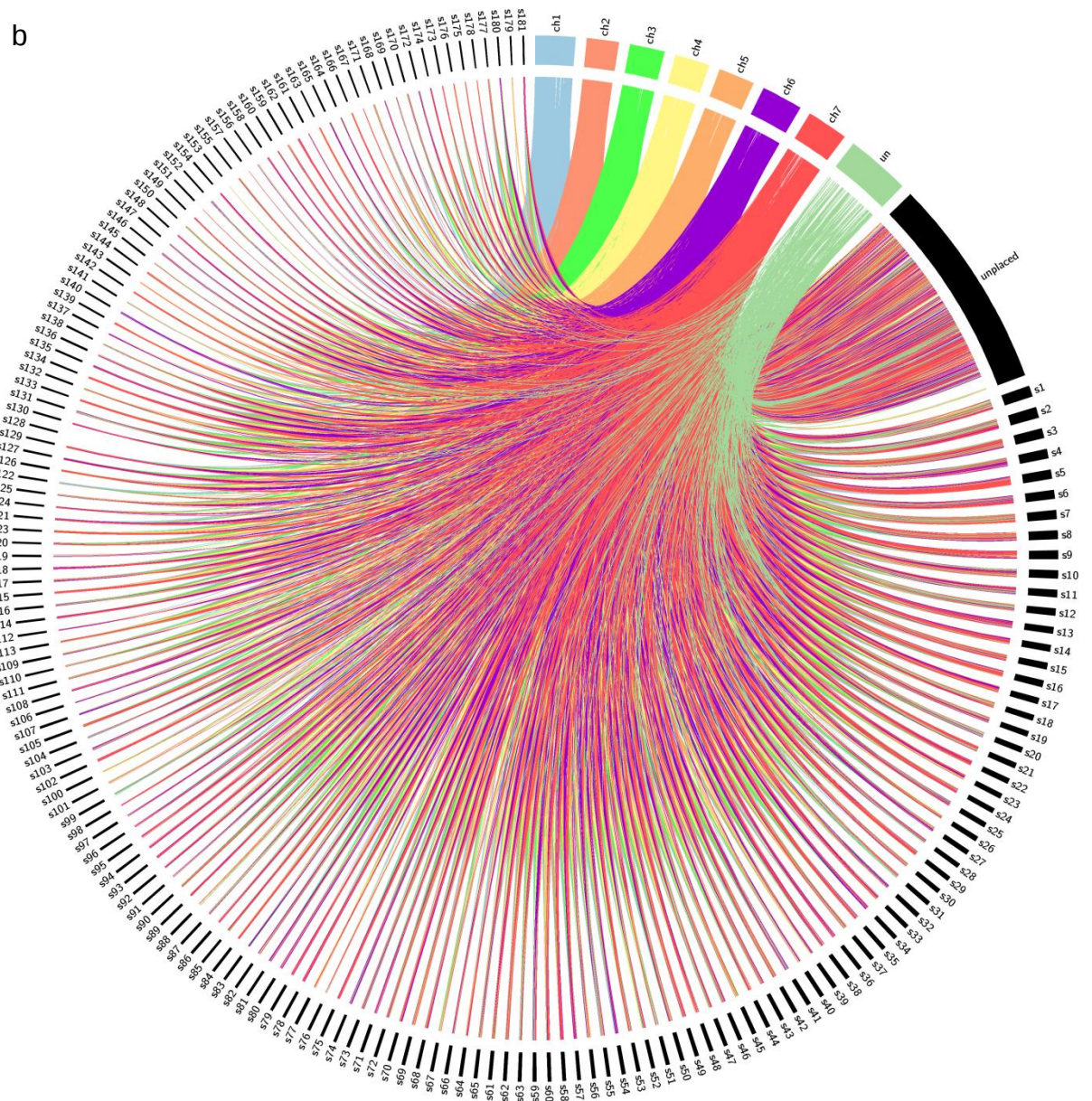


Figure S6b. Synteny plots showing homologous blocks between *O. coarctata* and *E. salsugineum*

C

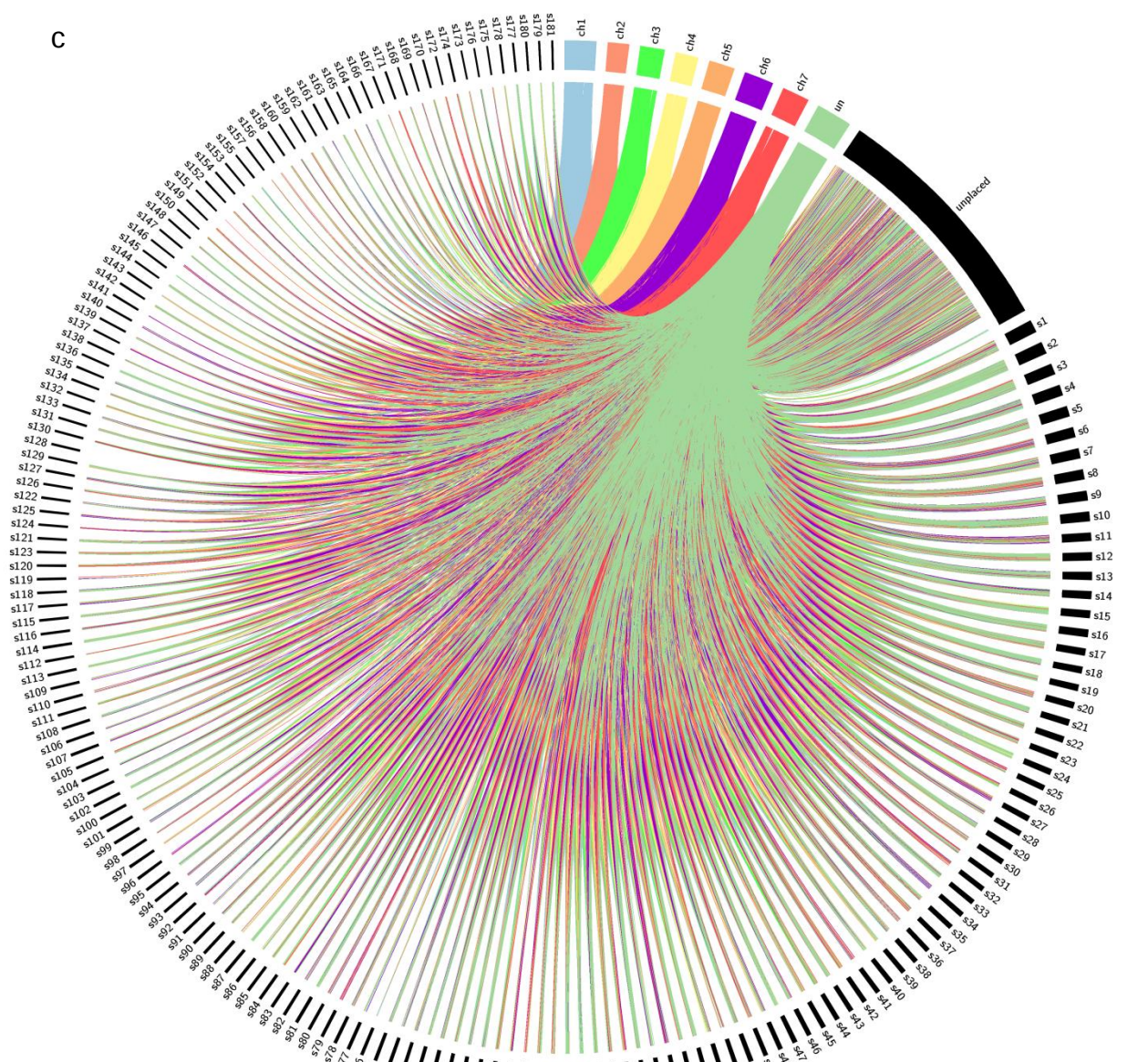


Figure S6c. Synteny plots showing homologous blocks between *O. coarctata* and *S. parvula*

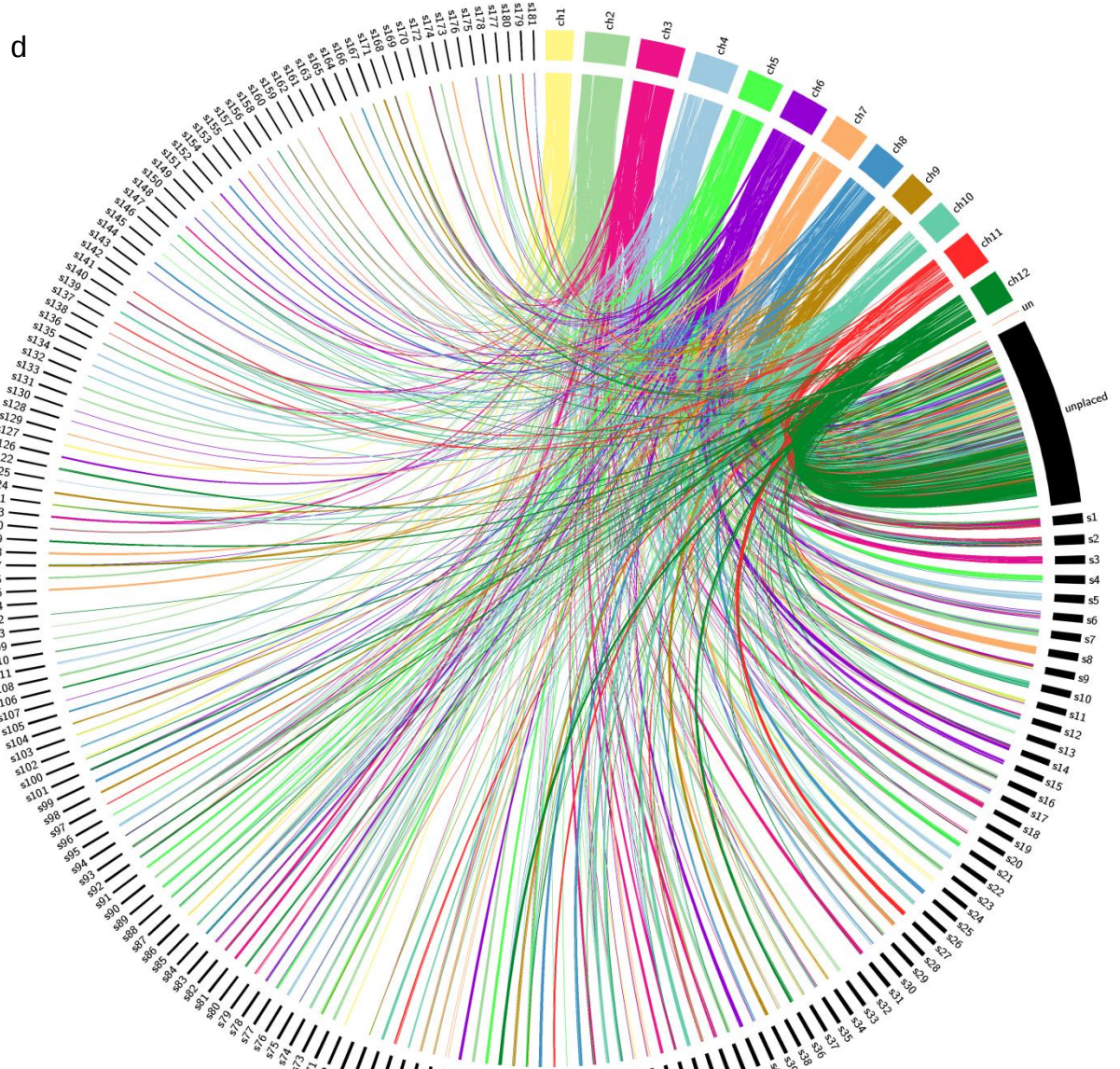


Figure S6. Synteny plots showing homologous blocks between *O. coarctata* and (a) *A. thaliana*, (b) *E. salsugineum*, (c) *S. parvula* and (d) *O. sativa japonica*. Each line link one homologous gene pair while the line colour representing each chromosome set of reference genome. The 181 linkage groups of *O. coarctata* are prefixed with “S” followed by the scaffold numbers, and the reference genomes chromosomes are designated with “ch” followed by their chromosome numbers.

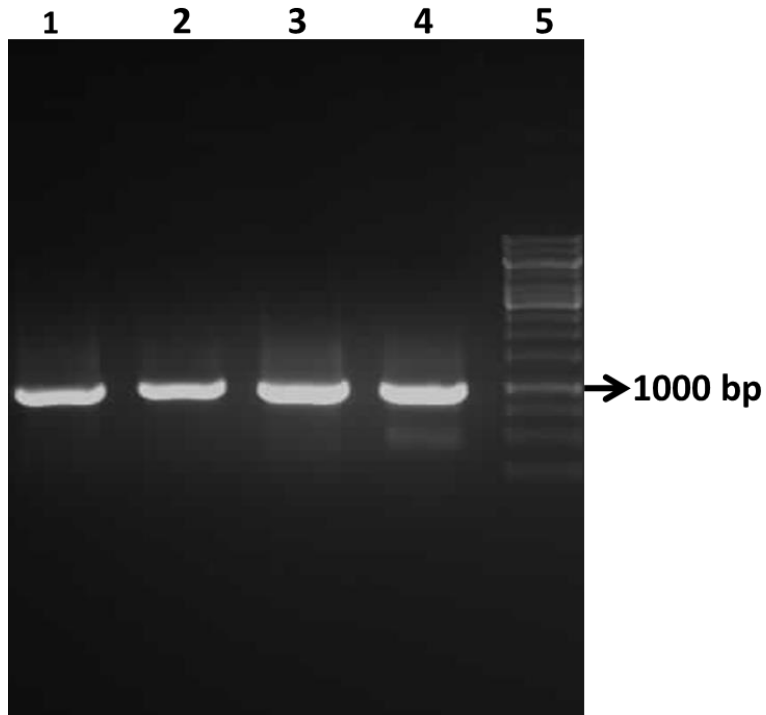


Figure S7. PCR validation of the chloroplast genome of *O. coarctata*. The agarose gel electrophoresis (1.5%) of the PCR amplified products of the four chloroplast genome junctions. The lane 1 represents the Junction of LSC/IRa; lane 2 represents Junction of IRa/SSC; lane 3 represents Junction of SSC/IRb ;lane 4 represents Junction of IRb/SSC and lane 5 represents 1 kbp (kilobase pair) ladder.

Table S1. Different Data used in this study:

Data Type	Database	Website Link/Accession no.
For comparative study between Oryza genus		
Protein sequences of 11 Oryza species	Gramene	http://www.gramene.org/
For Synteny analysis- Chromosome level assembly of 4 species		
<i>Arabidopsis thaliana</i>	Assembly Database at NCBI (https://www.ncbi.nlm.nih.gov/assembly/)	GCA_000001735.1
<i>Schrenkiella parvula</i>		GCA_000218505.1
<i>Eutrema salsugineum</i>		GCA_000325905.2
<i>Oryza sativa japonica</i>		GCA_001433935.1
Transcriptomic data used as expression evidence for gene prediction		
Transcriptome data (5 Samples) of <i>Oryza (Porteresia) coarctata</i>	SRA database at NCBI https://www.ncbi.nlm.nih.gov/sra	SRP019245
Chloroplast genomes of 11 <i>Oryza</i> species used as references for chloroplast assembly of <i>O. coarctata</i>		
<i>Oryza sativa japonica</i>	Nucleotide database at NCBI https://www.ncbi.nlm.nih.gov/nuccore/	X15901.1
<i>Oryza rufipogon</i>		KF359902.1
<i>Oryza sativa indica</i>		NC_008155.1
<i>Oryza nivara</i>		NC_005973.1
<i>Oryza meridionalis</i>		NC_016927.1
<i>Oryza longistaminata</i>		KM088024.1
<i>Oryza punctata</i>		KF359908.1
<i>Oryza glumaepatula</i>		KF359905.1
<i>Oryza brachyantha</i>		KF359917.1
<i>Oryza barthii</i>		KM088023.1
<i>Oryza glaberrima</i>		KF359903.1
Mitochondrial genomes used as references and in phylogeny for mitochondrial assembly of <i>O. coarctata</i>		
<i>Oryza sativa japonica</i>	Nucleotide database at NCBI https://www.ncbi.nlm.nih.gov/nuccore/	NC_011033.1
<i>Oryza sativa indica</i>		NC_007886.1
<i>Oryza rufipogon</i>		AP011076.1
<i>Orya minuta</i>		KU176938.1
<i>Triticum aestivum</i>		NC_007579.1
<i>Sorghum bicolor</i>		NC_008360.1
<i>Glycine max</i>		NC_020455.1
<i>Arabidopsis thaliana</i>		NC_001284.2

Table S2. List of pairs of primers used for PCR validation

O. coarctata plantlets were grown in glass house under 16 h light (25 °C)/ 8 h dark (16 °C) conditions. Fresh leaves from the plantlet were collected and total genomic DNA was isolated from 100 mg of leaves by using CTAB method. The residual RNA was removed by RNase I treatment (Thermo Fisher Scientific, USA) and integrity of DNA was checked in 0.8% agarose. *In silico* gene map of chloroplast genome of *O. coarctata* was designed. In order to validate the gene map, the primers were prepared from the four junctions between LSC/IRs and SSC/IRs of the chloroplast gene map. All primers were designed by using Vector NTI software (Lu et al., 2004). The primer sequences along with its T_m (°C) and product length are provided in Table 1.

The four junctions of chloroplast genome were amplified using the polymerase of GoTaq® DNA Polymerase (Promega). The reaction mixture of 25 µ L was prepared for PCR amplification. The amplification of the LSC/IRs and SSC/IRs regions was performed in Agilent PCR: 5 min at 97 °C for denaturation; 35 cycles of 1 min at 95 °C (denaturation), 1 min at 59- 62°C (annealing), and 2 min at 72 °C (extension); and a final step of 10 min at 72 °C for extension. The PCR products were checked on 1% agarose and also send for sequences for confirmation.

Junctions	Sequence	T _m (°C)	Product size(base pair)
Junction of LSC/IRa	F:GGTGATTTTGGGTGTACCAAGTCTGAAAC R: CGAGATATTGGACATCGAATGCCACTC	62	966
Junction of IRa/SSC	F:CAAGTATTCAGTTTCACCAATAAGATACGGAGAC R: TTCCAAGAAGTTCTAATTCGTCCATAAATTC	59	990
Junction of SSC/IRb	F: CCCAATTTACAGCTTCTTCTCCGCTAA R: GTTTCACCAATAAGATACGGAGACTTGCTTC	61	972
Junction of IRb/SSC	F: GGATTCACAATCCACTGCCTTGATCC R: AGGAAACCACTGAAATGAATCTGCTAATGAG	62	948

Reference

Lu, G., Moriyama, E.N. (2004). Vector NTI, a balanced all-in-one sequence analysis suite. *Brief. Bioinform*5, 378-388. doi: 10.1093/bib/5.4.378

Table S3. Genome coverage and sequencing reads statistics

S. No.	Reads Type	Genome coverage (x) using Estimated Genome size (~665 Mb)	High-Quality Processed Data	
			Reads	Bases
1.	PE	186.14	412,607,853	123,782,355,900
2.	2 Kb	16.73	37,093,166	11,127,949,800
3.	4 Kb	4.61	10,218,333	3,065,499,900
4.	6 Kb	16.89	37,439,915	11,231,974,500
5.	8 Kb	16.74	37,105,001	11,131,500,300
6.	Nanopore	9.55	1,717,607	6,350,953,369
	Total	250.66	536,181,875	166,690,233,769
PE – Illumina paired-end reads; 2 Kb, 4 Kb, 6 Kb and 8 Kb – Mate-pair reads with insert size of 2 Kb, 4 Kb, 6 Kb and 8 Kb, respectively.				

Table S4. Completeness of the genome assembly by CEGMA based analysis

Pathotypes	<i>O. coarctata</i>		Reference	
	Complete	Partial	Complete	Partial
No. of proteins	229	241	232	245
% Completeness	92.34	97.18	93.55	98.79
Total	611	727	426	492
Average	2.67	3.02	1.84	2.01
% Ortho	83.41	91.29	51.72	57.96
Normalized % Completeness	98.70	98.36	100	100

*Reference= *Oryza sativa* spp. *japonica* Nipponbare -IRGSP-1.0

Keys:

- # Prots = Number of 248 ultra-conserved CEGs present in genome
- # %Completeness = Percentage of 248 ultra-conserved CEGs present
- # Total = Total number of CEGs present including putative orthologs
- # Average = Average number of orthologs per CEG
- # %Ortho = Percentage of detected CEGs that have more than 1 ortholog
- # Normalized % = Completeness with respect to reference

Table S5. Number of genes present in 11 *Oryza* species genomes used for comparative study

S.No.	Species	Total	Range (aa)	Begins with M	Begins with M and >50 aa
1	<i>O. coarctata</i>	42284	4569-23	34851	33627
2	<i>O. sativa japonica</i>	42132	5342-14	32384	32218
3	<i>O. sativa indica</i>	40745	5024-32	40745	40701
4.	<i>O. rufipogon</i>	47441	6304-1	47235	47042
5.	<i>O. punctata</i>	41060	5284-1	40946	40803
6.	<i>O. nivara</i>	48360	5485-1	48163	47982
7.	<i>O. meridionalis</i>	43455	5440-1	43251	43059
8.	<i>O. longistaminata</i>	31686	5154-2	28237	27660
9.	<i>O. glumaepatula</i>	46893	5488-1	46723	46535
10.	<i>O. glaberrima</i>	33164	4738-23	28958	28869
11.	<i>O. brachyantha</i>	32037	5111-28	30420	30312
12.	<i>O. barthii</i>	41595	5436-1	41431	41260

Table S6. Statistics of genes per species in the plant gene database used for annotation (top 50 are shown)

S.No.	Species	Proteins
1	<i>Brassica napus</i>	113056
2	<i>Camelina sativa</i>	107566
3	<i>Gossypium hirsutum</i>	91320
4	<i>Nicotiana tabacum</i>	84981
5	<i>Helianthus annuus</i>	73959
6	<i>Glycine max</i>	71209
7	<i>Arabidopsis thaliana</i>	63768
8	<i>Chenopodium quinoa</i>	63172
9	<i>Raphanus sativus</i>	61234
10	<i>Malus domestica</i>	60654
11	<i>Zea mays</i>	59499
12	<i>Gossypium raimondii</i>	59094
13	<i>Hevea brasiliensis</i>	58083
14	<i>Medicago truncatula</i>	57753
15	<i>Arachis ipaensis</i>	57604
16	<i>Brassica oleracea</i> var. <i>oleracea</i>	56622
17	<i>Juglans regia</i>	55711
18	<i>Aegilops tauschii</i> subsp. <i>tauschii</i>	55630
19	<i>Lupinus angustifolius</i>	52833
20	<i>Arachis duranensis</i>	52811
21	<i>Brassica rapa</i>	52555
22	<i>Eucalyptus grandis</i>	52553
23	<i>Ipomoea nil</i>	51068
24	<i>Populus euphratica</i>	49783
25	<i>Nicotiana tomentosiformis</i>	49041
26	<i>Nicotiana sylvestris</i>	48312
27	<i>Musa acuminata</i> subsp. <i>malaccensis</i>	47720
28	<i>Gossypium arboreum</i>	47619
29	<i>Pyrus x bretschneideri</i>	47064
30	<i>Populus trichocarpa</i>	46066
31	<i>Capsicum annuum</i>	45495
32	<i>Oryza sativa</i> Japonica Group	44930
33	<i>Daucus carota</i> subsp. <i>sativus</i>	44587
34	<i>Nicotiana attenuata</i>	44494
35	<i>Manihot esculenta</i>	43333
36	<i>Elaeis guineensis</i>	41875
37	<i>Vitis vinifera</i>	41380
38	<i>Tarenaya hassleriana</i>	41179
39	<i>Sorghum bicolor</i>	39414
40	<i>Arabidopsis lyrata</i> subsp. <i>lyrata</i>	39157
41	<i>Cajanus cajan</i>	38966
42	<i>Phoenix dactylifera</i>	38567
43	<i>Solanum tuberosum</i>	38362

44	<i>Nelumbo nucifera</i>	38232
45	<i>Vigna angularis</i>	37747
46	<i>Ziziphus jujuba</i>	37646
47	<i>Asparagus officinalis</i>	36857
48	<i>Solanum lycopersicum</i>	36515
49	<i>Physcomitrella patens</i>	36066
50	<i>Ananas comosus</i>	35784

Table S7. Rice (IRGSP1.0) vs *O. coarctata* annotation analysis showing differences (of atleast 5) for number of genes with stress related GO terms.

GO-IDs	Gene counts		GO category (Function)
	<i>O. coarctata</i>	Rice	
GO:0009733	91	1	response to auxin
GO:0006950	94	42	response to stress
GO:0042744	11	0	hydrogen peroxide catabolic process
GO:0050832	10	0	defense response to fungus
GO:0009269	9	0	response to desiccation
GO:0006654	7	0	phosphatidic acid biosynthetic process
GO:0009607	7	0	response to biotic stimulus
GO:0048017	7	0	inositol lipid-mediated signaling
GO:0000724	6	0	double-strand break repair via homologous recombination
GO:0009737	6	0	response to abscisic acid
GO:0010112	6	0	regulation of systemic acquired resistance
GO:2000028	6	0	regulation of photoperiodism, flowering
GO:0009651	5	0	response to salt stress
GO:0009742	5	0	brassinosteroid mediated signaling pathway
GO:0009958	5	0	positive gravitropism

Table S8. Statistics of BLAST and annotation with Blast2GO

Data type	Genes
Predicted genes	33627
Annotated with Blast against Plant Genes DB	26569
Genes with assigned GO Terms	16357
Genes with InterProScan Match	30886
Genes with Enzyme Commission Classification	2,844
Genes with No Blast hit	7058
Genes with No GO terms	17270
Genes with No BLAST hits but GO terms	120
Genes with No BLAST hits but IPS match	4842
Genes with neither BLAST hits nor GO terms nor IPS match	2216

Table S9. Number of TF containing genes of *O. coarctata* shared across other species of plants

Organism Name	No. of genes shared
<i>Oryza punctata</i>	560
<i>Oryza brachyantha</i>	261
<i>Malus domestica</i>	209
<i>Oryza meridionalis</i>	186
<i>Oryza sativa subsp. japonica</i>	181
<i>Fragaria vesca</i>	163
<i>Oryza longistaminata</i>	150
<i>Actinidia chinensis</i>	148
<i>Oryza glumaepatula</i>	147
<i>Oryza sativa subsp. indica</i>	141
<i>Leersia perrieri</i>	136
<i>Oryza nivara</i>	111
<i>Sisymbrium irio</i>	110
<i>Zoysia matrella</i>	88
<i>Oryza glaberrima</i>	88
<i>Fragaria ananassa</i>	87
<i>Dorcoceras hygrometricum</i>	84
<i>Phyllostachys heterocycla</i>	83
<i>Oryza rufipogon</i>	82
<i>Ipomoea trifida</i>	81
<i>Brassica napus</i>	80
<i>Oryza barthii</i>	74
<i>Musa acuminata</i>	70
<i>Daucus carota</i>	70
<i>Zoysia japonica</i>	64
<i>Gossypium hirsutum</i>	63
<i>Dichanthelium oligosanthes</i>	60
<i>Linum usitatissimum</i>	59
<i>Oropetium thomaeum</i>	55
<i>Raphanus sativus</i>	51
<i>Solanum melongena</i>	50
Others	1124
Total	4916

Table S10. Repeat masking statistics of *O. coarctata* genome

TE Class	Family	Sub-family	No. of Elements	Length Occupied (bp)	% in the genome
Class I (Retroelements)	Non-LTR elements	SINEs:	1450	217932	0.04
		LINEs:	3826	2456111	0.43
	LTR elements	Ty1/Copia	47330	46943177	8.24
		Gypsy/DIRS1	73040	34417003	6.04
		Total LTR elements	124458	82529172	14.48
Total Retroelements ^a			129734	85203215	14.95
Class II (DNA transposons)	DNA transposons	hobo-Activator	4857	1305561	0.23
	DNA transposons	Tc1-IS630-Pogo	15111	2448176	0.43
	DNA transposons	Tourist/Harbinger	16361	2564908	0.45
Total DNA Transposons ^b			79468	20121450	3.53
Unclassified ^c			3092	892369	0.16
Total interspersed repeats ^(a+b+c)				106217034	18.63
Small RNA			1605	284272	0.05
Satellites			141	24922	0
Simple repeats			136349	5835615	1.02
Low complexity			23175	1153463	0.2
Total bases masked				113352647	19.89
Full length LTR retrotransposons			218	1308198	0.23

Table S11. Statistics of SSRs

Stats	Count
Total number of sequences examined	58362
Total size of examined sequences (bp)	569994164
Total number of identified SSRs	230968
Number of SSR containing sequences	12926
Number of sequences containing more than 1 SSR	5171
Number of compound SSRs	26494
Mono	46172
Di	38786
Tri	61568
Tetra	63849
Penta	13668
Hexa	6925

Table S12. Annotation of conserved non-coding RNA genes in *O. coarctata* genome

ncRNA Type	#Loci	Average length (bp)	Total length (bp)	Percentage of genome (%)
miRNA	200	135	27,018	0.00
tRNA	900	74	66,822	0.01
rRNA (LSU)	226	263	59,482	0.01
rRNA(SSU)	142	255	36,158	0.01
rRNA(5_8S_rRNA)	8	130	1,037	0.00
rRNA (5S_rRNA)	2,734	108	294,806	0.05
SnoRNA	1,184	119	140,685	0.02
snRNA	118	141	16618	0.00

Table S13. Features of the chloroplast genome of *O. coarctata*

Feature	Length (bp)	Number count
Genome	134750	1
LSC	80816	1
SSC	12334	1
Inverted Repeat (IR)Regions	20800	2 (IRa, IRb)
Protein Coding genes	59086	82
tRNA	2419	33
rRNA	9182	8

Table S14. List of genes present in the chloroplast genome of *O. coarctata*

Category	Gene Name
Photosystem I	<i>psaA,B,C,I</i>
Photosystem II	<i>psbA,B,C,D,E,F,H,I,J,K,L,M,N,T,Z</i>
Cytochrome b6/f	<i>petA,B,D,G,L,N</i>
ATP Synthase	<i>atpA,B,E,F,H,I</i>
Rubisco	<i>rbcL</i>
NADH-dehydrogenase	<i>ndhA,B[#],C,D,E,F,G,H,I,J,K</i>
Large subunit ribosomal proteins	<i>rpl2[#],14,16,20,22,23[#],32,33,36</i>
Small subunit ribosomal proteins	<i>rps11, rps12[#], rps14, rps15[#], rps16, rps18, rps19[#], rps2, rps3, rps4, rps7[#], rps8</i>
DNA dependent RNA polymerase	<i>rpoA,B,C1,C2</i>
Other Proteins	<i>ccsA, cemA, clpP, infA, matK</i>
Proteins of unknown Function	<i>ycf3, ycf4</i>
Ribosomal RNAs	<i>rrn16S[#], rrn23S[#], rrn4.5S[#], rrn5S[#]</i>
Transfer RNAs	<i>trnA-TGC[#], trnC-GCA, trnD-GTC, trnE-TTC, trnF-GAA, trnI[#]M-CAT, trnG-GCC, trnH-GTG[#], trnI-CAT[#], trnL-CAA[#], trnL-TAG, trnM-CAT, trnN-GTT[#], trnP-TGG, trnQ-TTG, trnR-ACG[#], trnR-TCT, trnS-GCT, trnS-GGA, trnS-TGA, trnT-GGT, trnT-TGT, trnV-GAC[#], trnW-CCA, trnY-GTA</i>

[#]Two gene copies in the IRs

Table S15. Features of the mitochondrial genome of *O. coarctata*

Features	Statistics
Genome size (bp)	491065
GC contents (%)	43.07
Total Genes	91
Protein Coding genes	46 (40206 bp)
tRNA	37 (2807 bp)
rRNA	4 (5671 bp)
Total Gene Length (bp)	48960 (9.97% of genome)

Table S16. Genetic contents in the mitochondria genome of *O. coarctata*

Functional Group	Genes
Complex I (NADH dehydrogenase subunits)	<i>nad1, nad2, nad3, nad4, nad4l, nad5, nad6, nad7, nad9</i>
Complex III (Apocytochrome b)	<i>cob</i>
Complex IV (Cytochrome c oxidase subunits)	<i>cox1, cox2, cox3</i>
Complex V (ATP synthase subunits)	<i>atp1, atp4, atp6_0, atp6_1, atp6_2, atp8, atp9_0, atp9_1</i>
Cytochrome C	<i>ccmB, ccmC, ccmFc, ccmFn</i>
Ribosomal large subunit	<i>rpl2, rpl5, rpl16</i>
Ribosomal small subunit	<i>rps1, rps2, rps3, rps4, rps7, rps12, rps13, rps19,</i>
Ribosomal RNAs (rRNA)	<i>rrnL, rrnS_0-a, rrnS_0-b, rrnS_1</i>
Transfer RNAs (tRNA)	<i>trnC_0(gca), trnC_1(gca), trnC_2(gca), trnD(gtc), trnF_0(gaa), trnF_1(gaa), trnH_0(gtg), trnH_1(gtg), trnH_2(gtg), trnK(ttt), trnL1(---), trnL2(caa), trnM_0(cat), trnM_1(cat), trnM_2(cat), trnM_3(cat), trnM_4(cat), trnM_5(cat), trnM_6(cat), trnM_7(cat), trnM_8(cat), trnN(gtt), trnP_0(tgg), trnP_1(tgg), trnP_2(tgg), trnQ(ttg), trnR(tct), trnS1_0(gct), trnS1_1(gct), trnS2_0(tga), trnS2_1(tga), trnS2_2(gga), trnV(gac), trnW_0(cca), trnW_1(cca), trnY_0(gta), trnY_1(gta)</i>
Origins of replication	<i>OH-a, OH-b, OH-c, OH-d</i>
Intron maturase	<i>mat-r</i>

Table S17. Summary of the sequenced halophyte genomes

Organism/species		<i>T. parvula</i> (<i>E. parvulum</i> or <i>S. parvula</i>)	<i>T. salsuginea</i> (<i>T. halophila</i>)	<i>E. salsugineum</i> (<i>T. halophila</i>)	<i>O.</i> <i>coarctata</i>
Dicot or monocot		Dicot	Dicot	Dicot	Monocot
Genome size	Expected	140	260	260	665
	Assembled	137.09	233.7	243.1	569.99
Coverage (X)		~50	~134	~8	250.66
Repetitive elements (%)		7.46%	52%	51.40%	19.89%
Protein-coding genes		30419	28457	26,531	33,627
Reference		Dassanayake et al., (2011)	Wu et al., (2012)	Yang et al., (2013)	In the present study