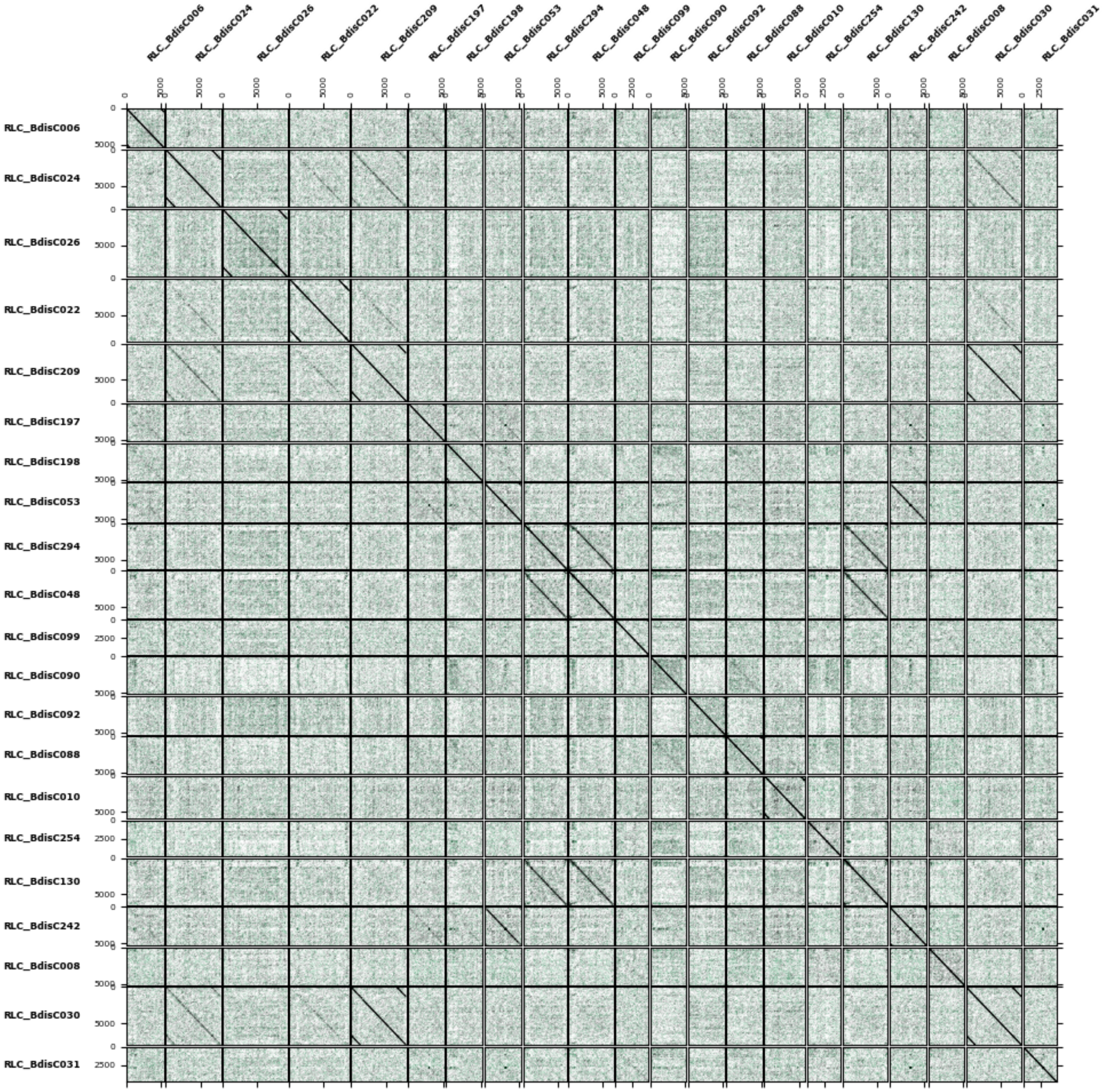
***New Phytologist* Supporting Information**

Article title: Diversity, dynamics and effects of LTR retrotransposons in the model grass *Brachypodium distachyon*
Authors: Stritt C, Wyler M, Gimmi EL, Pippel M, Roulin AC
Article acceptance date: 10th October 2019

The following Supporting Information is available for this article:

Fig. S1 Dotplots of the 40 LTR-RT families downloaded from TREP. a) *Copia* families. The following subfamilies were merged under the name of the most abundant subfamily (underscore): RLC_BdisC030, RLC_BdisC209; RLC_BdisC053, RLC_Bdis242; RLC_Bdis294, RLC_Bdis048, RLC_Bdis130. b) For *Gypsy* familes, the following subfamilies were merged: RLG_BdisC039, RLG_BdisC102, RLG_Bdis021; RLG_BdisC000, RLG_BdisC061; RLG_BdisC266, RLG_BdisC012.

(a)



(b)

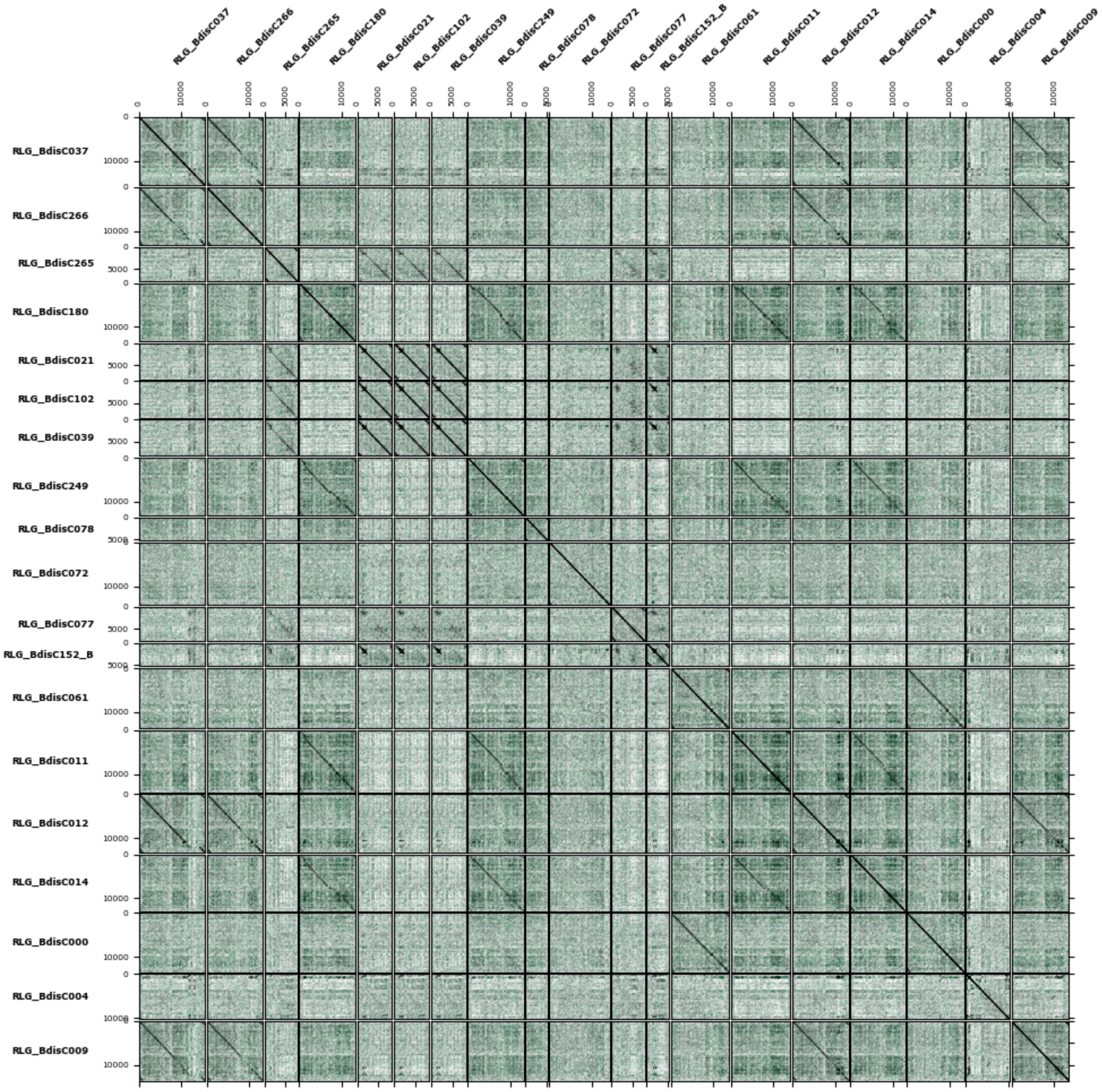

RLG_BdisC037
RLG_BdisC266
RLG_BdisC265
RLG_BdisC180
RLG_BdisC021
RLG_BdisC102
RLG_BdisC039
RLG_BdisC249
RLG_BdisC078
RLG_BdisC072
RLG_BdisC077
RLG_BdisC152_B
RLG_BdisC061
RLG_BdisC011
RLG_BdisC012
RLG_BdisC014
RLG_BdisC000
RLG_BdisC004
RLG_BdisC009
0
5000
10000

Fig. S2 Dotplots of the consensus sequences for the five centromeric CRM families and RLG_BdisC152, which is also centromere-specific. As can be seen, the latter shares most of its LTR and part the 3' flanking sequence with the CRM elements.

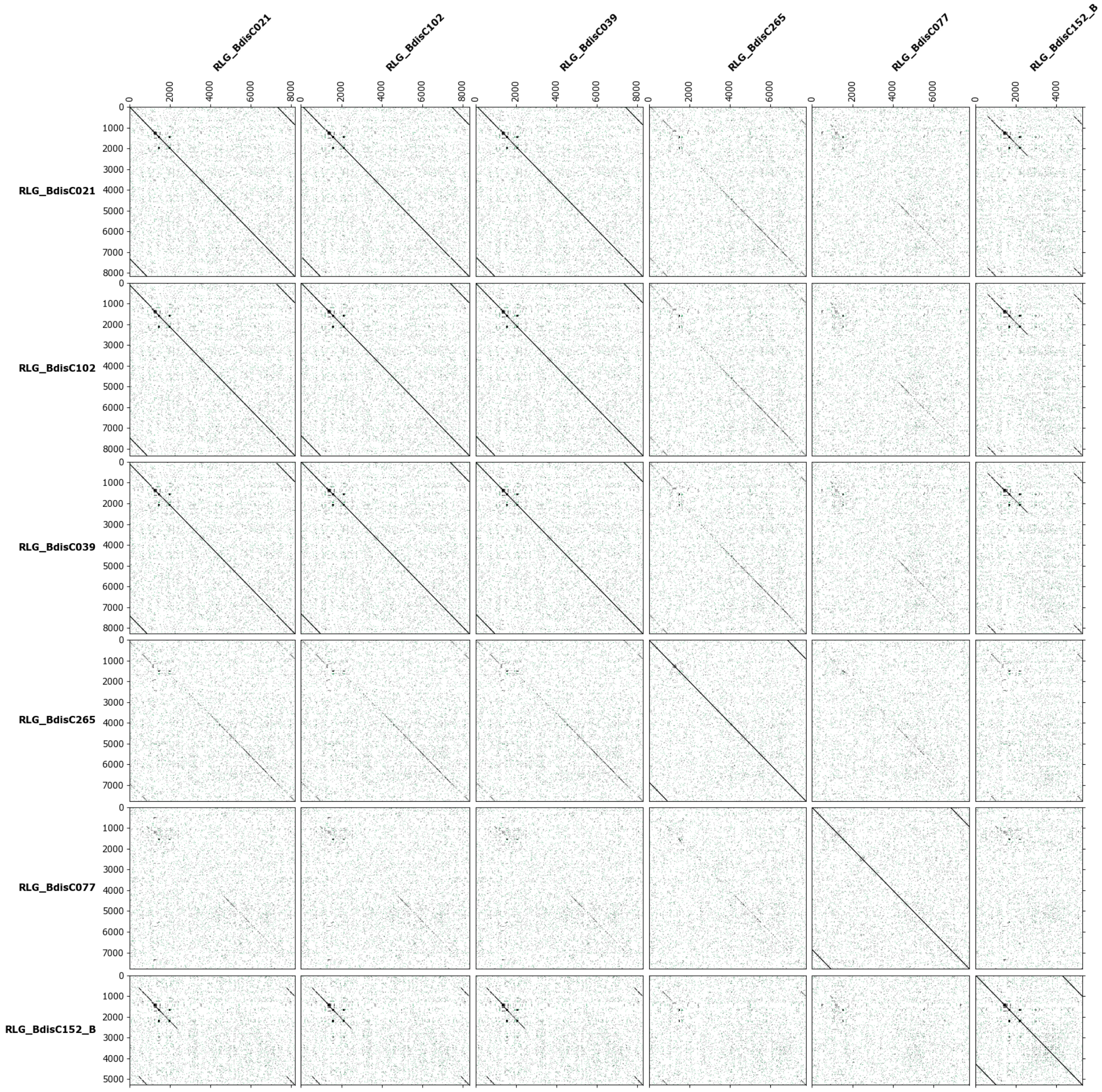
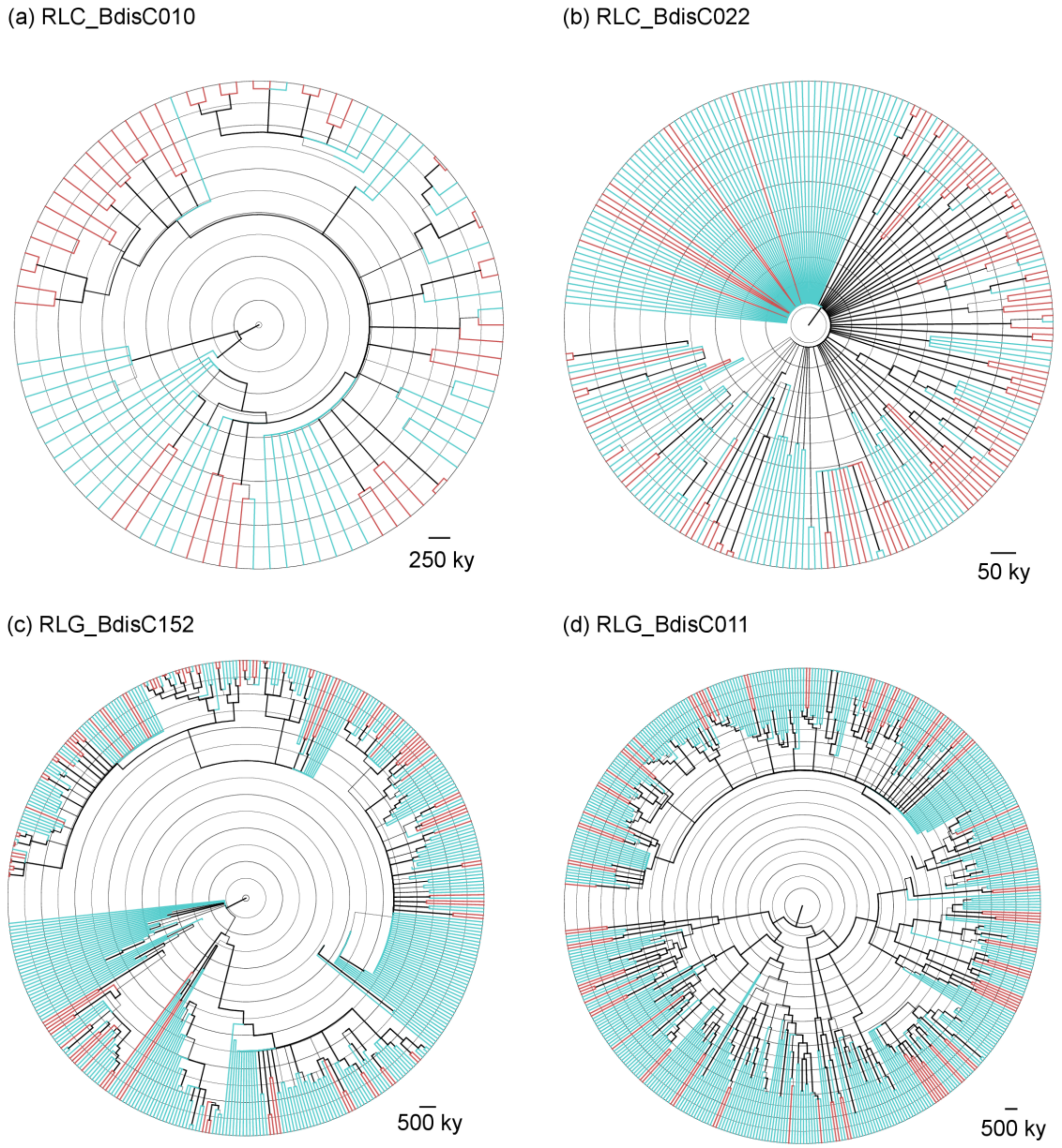


Fig. S3 Four exemplary LTR genealogies (RLC_BdisC10, RLC_BdisC022, RLG_BdisC152, RLG_BdisC011) estimated with MrBayes, showing single (blue) and paired (red) LTRs. Note the star-like phylogeny and different timescale for RLC_BdisC022, a high-turnover family. The terminal branch lengths of these trees were used as estimates for the age of the copies.



(a) RLC_BdisC010

(b) RLC_BdisC022

(c) RLG_BdisC152

(d) RLG_BdisC011

Fig. S4 Number of annotated LTRs in Bd21 and BdTR7a. The dashed line shows the slope of 1, family names of high copy-number outliers are indicated.

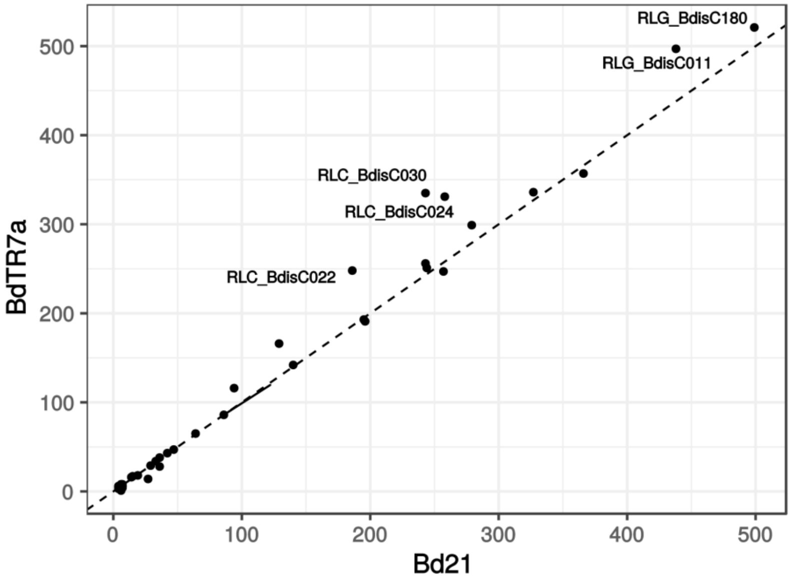


Fig. S5 Genomic distribution of LTR retrotransposons along the five chromosomes of *B. distachyon.* Colored dashed lines show the distribution of single families with lineages, while the solid black line is the average across families.

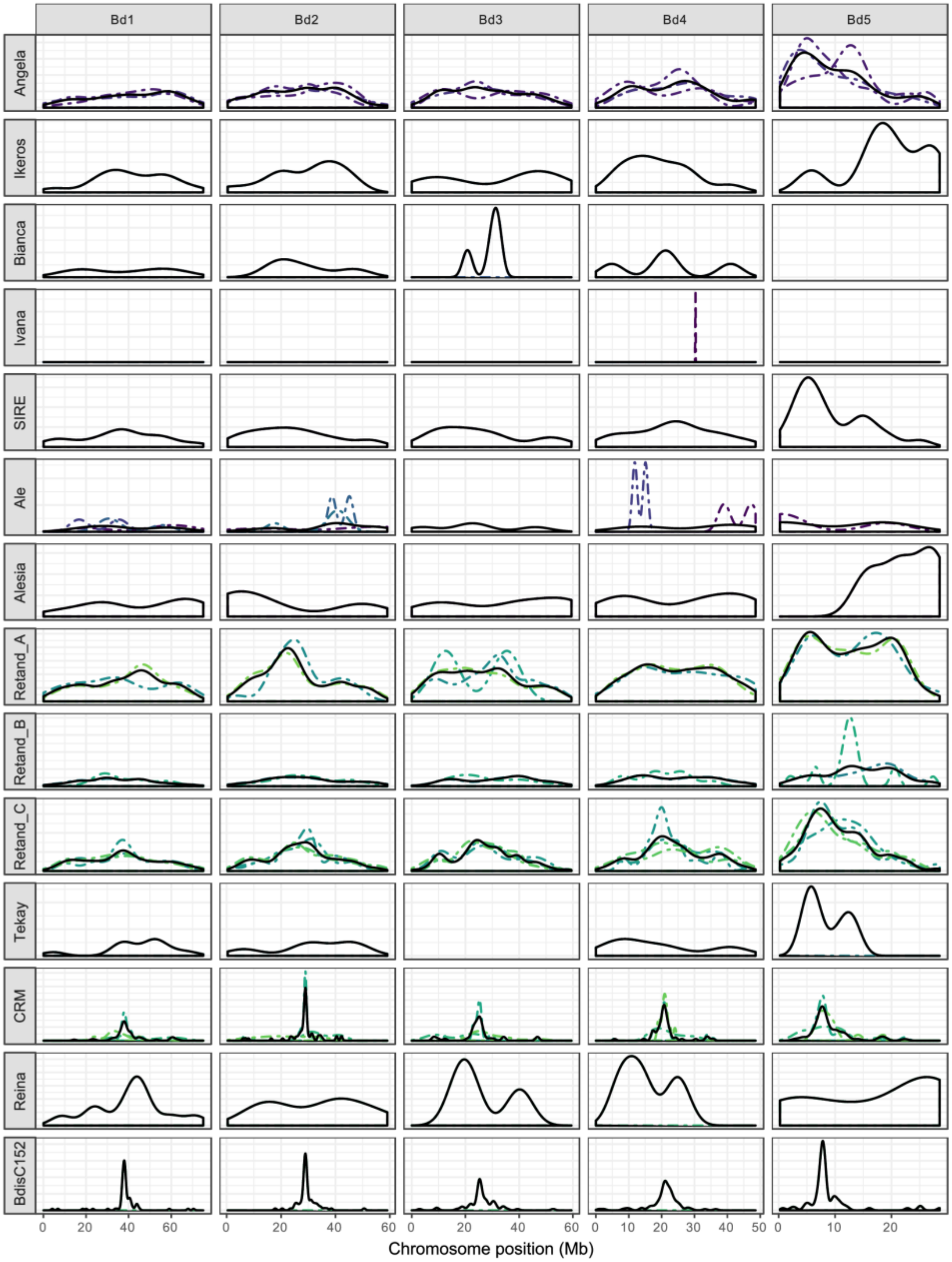
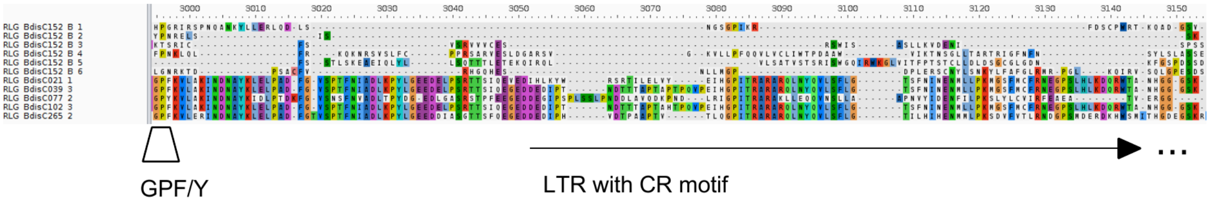


Fig. S6 Putative chromatin targeting domain (PTD) of the CRM families of *B. distachyon.* Also shown in the alignment are the six-frame translated sequences for the centromere-specific non-autonomous family RLG_BdisC152, where the PTD is not present. See Figure 3 of Neumann et al. 2011.



Literature cited

Neumann P, Navrátilová A, Koblížková A, Kejnovsk E, Hřibová E, Hobza R, Widmer A, Doležel J, MacAs J. 2011. Plant centromeric retrotransposons: A structural and cytogenetic perspective. *Mobile DNA* **2**: 1–16.

Table S3 Random forest confusion matrix, showing how well the TE lineages could be distinguished from each other based on features of their genomic niche.

	Angela	Ikeros	SIRE	Alesia	Retand_A	Retand_B	Retand_C	CRM	RLG_BdisC152	class.error
Angela	407	5	6	10	11	110	30	43	1	0.35
Ikeros	8	3	4	0	21	11	16	1	0	0.95
SIRE	5	0	161	2	9	5	1	13	46	0.33
Alesia	11	1	4	8	23	10	3	0	0	0.87
Retand_A	13	3	11	14	356	45	9	27	2	0.26
Retand_B	177	15	15	11	40	198	39	22	1	0.62
Retand_C	64	10	2	1	18	110	920	3	1	0.19
CRM	25	0	9	2	14	12	1	109	16	0.42
RLG_BdisC152	3	0	81	0	10	4	0	31	148	0.47

Table S4 Variable importance of the random forest model

	Angela	Ikeros	SIRE	Alesia	Retand_A	Retand_B	Retand_C	CRM	RLG_BdisC152	Mean Decrease Accuracy
chromosome	0.24	1.77	-0.02	0.42	0.09	1.00	1.80	5.06	5.44	4.26
cM_per_Mb	2.72	2.35	1.78	6.93	6.64	6.01	6.84	11.00	15.82	18.33
dist_to_gene	11.91	-0.52	1.76	7.14	15.45	5.58	17.87	0.96	5.87	24.80
CHG	31.26	2.82	21.70	1.49	67.79	21.73	37.30	29.35	32.74	57.28
CpG	11.11	0.18	64.41	3.23	22.58	9.83	21.84	13.22	45.24	41.04
CHH	48.36	3.86	13.14	2.94	20.36	17.48	41.38	41.56	9.86	51.52
S	4.72	0.43	2.59	0.55	4.27	3.62	6.67	3.48	5.87	11.97
π	6.46	1.54	3.10	1.13	3.43	2.84	6.85	3.81	5.63	12.38
Tajima's D	2.41	1.06	1.41	-0.74	0.85	1.09	4.06	5.94	6.58	7.86
Z_{ns}	0.48	-1.97	1.23	-1.58	4.60	4.56	2.47	3.45	8.43	8.77

Methods S1 BdTR7a genome assembly.

The BdTR7a assembly was created by combining PacBio sequencing with Bionano optical mapping. DNA was isolated with the Bionano Prep High Polyphenols Plant Tissue DNA Isolation Protocol (bionanogenomics.com/technology/platform-technology; August 28, 2018, date last accessed). The Bionano Sapyhr System at the Functional Genomic Center Zurich was used to produce an optical map of the genome. To do so, the extracted DNA was labelled using the PrepTM Direct Label and Stain (DLS) protocol and subsequently loaded onto a Saphyr Chip. In parallel, PacBio (Pacific Biosciences) seqencing was performed, also at the Functional Genomic Center Zurich.

Raw PacBio reads were assembled using the MARVEL assembler (Grohme *et al.*, 2018; Nowoshilow *et al.*, 2018). MARVEL consists of three major steps, namely the setup phase, patch phase and the assembly phase. In the setup phase, reads were filtered by choosing only the best read of each ZMW and requiring subsequently a minimum read length of 4 kb. The resulting 1.1 million reads (63-fold coverage) were stored in an internal database. The patch phase detects and corrects read artifacts including missed adaptors, polymerase strand jumps, chimeric reads and long low-quality read segments that are the primary impediments to long contiguous assemblies. The patched reads (55-fold coverage) were then used for the final assembly phase, which stitches short alignment artifacts resulting from bad sequencing segments within overlapping read pairs. This step was followed by repeat annotation and the generation of the overlap graph. To this end, we used the tool LAq with a quality cutoff of 27 to calculate a quality and a trim annotation track. In addition, alignments were forced through low quality regions (<200 bp) that remained in the patched reads. LArepeat in coverage auto-detection mode was used to create a repeat annotation track based on overlap coverage anomalies. The final assembled contigs were generated by touring the overlap graph. To correct base errors, we first used the correction module of MARVEL, which makes use of the final overlap graph and corrects only the reads that were used to build the contigs. Corrected contigs were further polished using PacBio's Arrow tool (github.com/PacificBiosciences/GenomicConsensus).

Using the Bionano optical map, the 257 contigs resulting from the PacBio assembly were combined into 7 supercontigs. Finally, the these supercontigs were aligned to the Bd21 reference genome with Mummer's nucmer algorithm (Marçais *et al.*, 2018)□. This step revealed that chromosomes Bd2 and Bd4 were split in two at the centromeres in the BdTR7a hybrid assembly; the final set of five chromosomes was therefore obtained by concatenating the respective supercontigs to full chromosomes.

Literature cited

Grohme MA, Schloissnig S, Rozanski A, Pippel M, Young GR, Winkler S, Brandl H, Henry I, Dahl A, Powell S, *et al.* 2018. The genome of *Schmidtea mediterranea* and the evolution of core cellular mechanisms. *Nature* **554**: 56–61.

Marçais G, Delcher AL, Phillippy AM, Coston R, Salzberg SL, Zimin A. **2018**. MUMmer4: A fast and versatile genome alignment system. *PLoS Computational Biology* **14**: 1–14.

Nowoshilow S, Habermann B, Knapp D, Fei J-F, Dahl A, Pang AWC, Cao H, Roscito JG, Hiller M, Winkler S, *et al.* 2018. The axolotl genome and the evolution of key tissue formation regulators. *Nature* **554**: 50–55.

Methods S2 Whole-genome bisulfite sequencing.

For whole-genome bisulfite sequencing, three replicates of the *B. distachyon* accession Bd21 where planted in small pots (50 ml) on a soil-sand mixture. The plants were grown for 24 days in a growth chamber with following conditions: 16 h with 200 µMol lightvolume, constant 20°C and 60 – 70% of relative air humidity.

The second and third leaf from the top were sampled, flash-frozen and disrupted using a Bead Ruptor 24 (OMNI) with following settings: T = 1:00, S = 2.50, C = 01, D = 0:00. Subsequent DNA extraction was performed with the DNeasy Plant Mini Kit (Qiagen). Around 500 ng of DNA where physically shared for 13 min with 20% of amplitude on a Q800R2 Sonicator (Qsonica). End repair and A tailing of the samples were performed with the Kapa Hyper Prep Kit (Kapa Biosystems). Methylated NimbleGen SeqCap (Roche) adapters were ligated with the above-mentioned kit and the post-ligation cleanup was performed with 0.8x AMPure XP Beads (Agencourt) and two 80% ethanol washes. Bisulfite conversion was performed with EZ DNA Methylation-Gold (ZYMO Research). The converted libraries where eluted with 20 µL for five cycles of amplification using the Kapa HiFi HotStart Uracil+ ReadyMix (Kapa Biosystems) with the Pre-LM-PCR Oligos 1 & 2 from the NimbleGen SeqCap (Roche) kit. Post-amplification cleanup was performed similarly to the above mentioned post-ligation cleanup, but with 1x of AMPure XP Beads (Agencourt). Libraries were size-selected with 0.7x of AMPure XP Beads (Agencourt) and eluted with 20 µL of 10mM Tris-HCl, pH 8.0. For library quality assessment the High Sensitivity D5000 kit on a TapeStation 2200 (Agilent Technologies) was used. For library quantification the Kapa Library Quantification Kit for Illumina Platforms (Kapa Biosystems) was used on a 7500 Fast Real-Time PCR System (Applied Biosystems). Sequencing of 151 bp paired-end reads was performed on a Illumina HiSeq 2500 (Illumina).

Reads were trimmed with trim_galore (0.4.5, bioinformatics.babraham.ac.uk/ projects/trim_galore) and subsequently mapped against the reference genome of *B. distachyon* Bd21 (v.3) with Bismark (0.19.0, Krueger & Andrews, 2011)□ using Bowtie2 (2.3.2, Langmead & Salzberg, 2012)□. An average of 0.55% of the reads where identified as duplicates resulting from an overamplification of the library and removed with the deduplicate_bismark script from the Bismark. To asses the quality of the converted reads, Mbias plots (Hansen *et al.*, 2012)□were produced with the bismark_methylation_extractor script using following parameters: --comprehensive –mbias_only. Visual assessment allowed to identify inconsistent sequencing of the first 2 bp and 1 bp of the 5' end of read 1 respectively read 2. Consequently methylation levels of each covered cytosines in the specific CpG, CHG or CHH context were assessed with bismark_methylation_extractor specifying the following parameters: --comprehensive --bedGraph --CX --ignore 2 --ignore_r2 1. From reads mapping to the chloroplast sequence, we assessed an average conversion efficiency of 99%. The weighted methylation level for each annotated TE was calculated for all three context following the method described in Schultz, Schmitz & Ecker (Schultz *et al.*, 2012)□.

Literature cited

Hansen KD, Langmead B, Irizarry RA. **2012**. BSmooth: from whole genome bisulfite sequencing reads to differentially methylated regions. *Genome Biology* **13**: R83.

Krueger F, Andrews SR. **2011**. Bismark: A flexible aligner and methylation caller for Bisulfite-Seq applications. *Bioinformatics* **27**: 1571–1572.

Langmead B, Salzberg SL. **2012**. Fast gapped-read alignment with Bowtie 2. *Nature Methods* **9**: 357–359.

Schultz MD, Schmitz RJ, Ecker JR. **2012**. 'Leveling' the playing field for analyses of single-base resolution DNA methylomes. *Trends in Genetics* **28**: 583–585.