Supplementary Material

GOLEM: A tool for visualizing the distribution of Gene regulatOry eLEMents within the plant promoters with a focus on male gametophyte

Lukáš Nevosád[1], Božena Klodová[2], Jiří Rudolf[1,3], Tomáš Raček[1,3], Tereza Přerovská[1,3], Alžbeta Kusová[1,3], Radka Svobodová[1,3], David Honys[2], and Petra Procházková Schrumpfová[1,3*]

[1]National Centre for Biomolecular Research, Faculty of Science, Masaryk University, Kotlářská 2, 611 37 Brno, Czech Republic

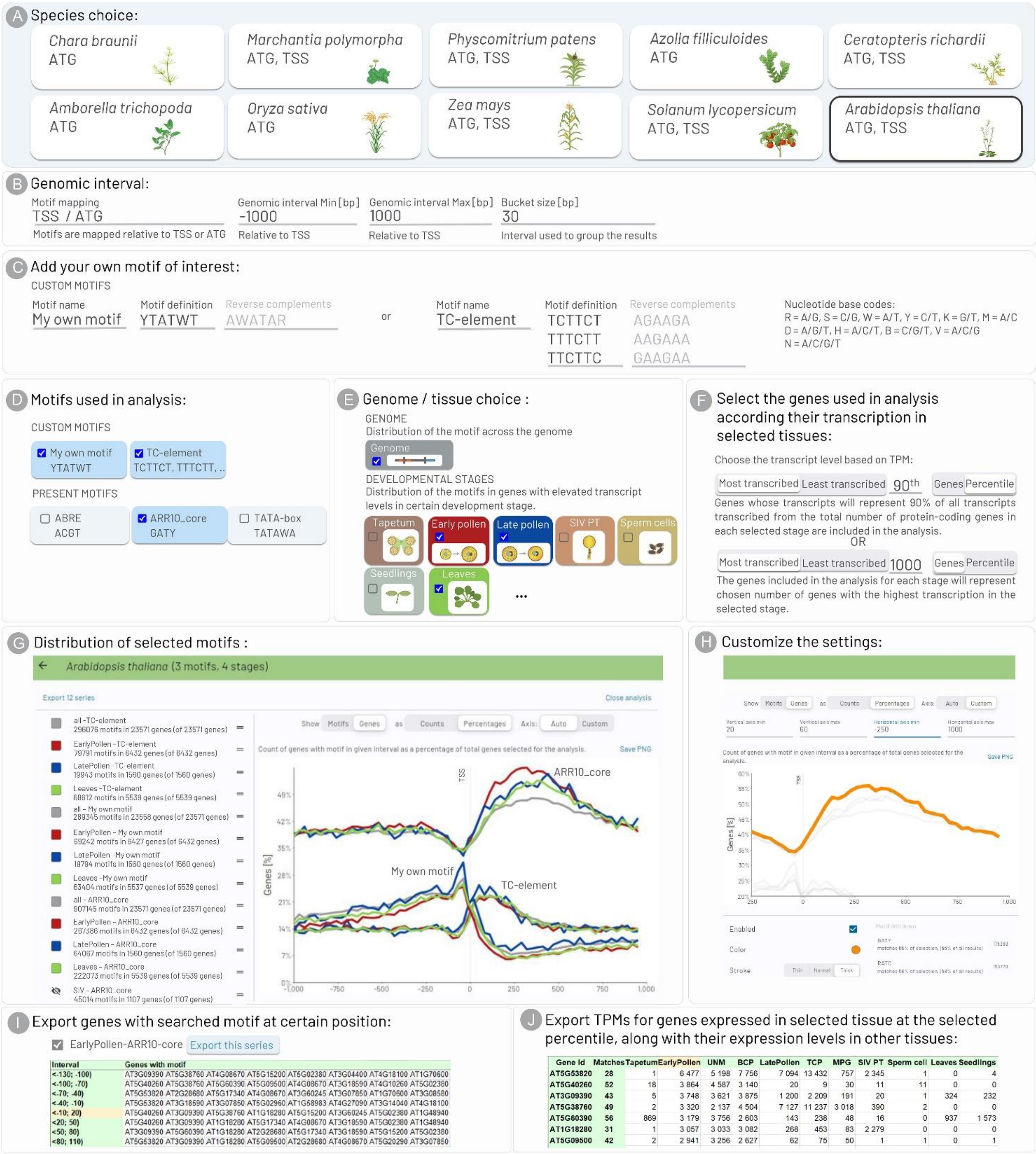
[2]Laboratory of Pollen Biology, Institute of Experimental Botany of the Czech Academy of Sciences, Rozvojová 263, 165 02 Prague, Czech Republic

[3]Central European Institute of Technology, Masaryk University, Kamenice 5, 625 00 Brno, Czech Republic

*corresponding author

schpetra@sci.muni.cz

Figures



A Species choice:

Chara braunii ATG	*Marchantia polymorpha* ATG, TSS	*Physcomitrium patens* ATG, TSS	*Azolla filliculoides* ATG	*Ceratopteris richardii* ATG, TSS
Amborella trichopoda ATG	*Oryza sativa* ATG	*Zea mays* ATG, TSS	*Solanum lycopersicum* ATG, TSS	*Arabidopsis thaliana* ATG, TSS

B Genomic interval:

Motif mapping: TSS / ATG — Motifs are mapped relative to TSS or ATG

Genomic interval Min [bp]: -1000 — Relative to TSS

Genomic interval Max [bp]: 1000 — Relative to TSS

Bucket size [bp]: 30 — Interval used to group the results

C Add your own motif of interest:

CUSTOM MOTIFS

Motif name	Motif definition	Reverse complements
My own motif	YTATWT	AWATAR

or

Motif name	Motif definition	Reverse complements
TC-element	TCTTCT TTTCTT TTCTTC	AGAAGA AAGAAA GAAGAA

Nucleotide base codes:
R = A/G, S = C/G, W = A/T, Y = C/T, K = G/T, M = A/C
D = A/G/T, H = A/C/T, B = C/G/T, V = A/C/G
N = A/C/G/T

D Motifs used in analysis:

CUSTOM MOTIFS

☑ My own motif YTATWT
☑ TC-element TCTTCT, TTTCTT, ..

PRESENT MOTIFS

☐ ABRE ACGT
☑ ARR10_core GATY
☐ TATA-box TATAWA

E Genome / tissue choice :

GENOME
Distribution of the motif across the genome

☑ Genome

DEVELOPMENTAL STAGES
Distribution of the motifs in genes with elevated transcript levels in certain development stage.

☐ Tapetum
☑ Early pollen
☑ Late pollen
☐ SIV PT
☐ Sperm cells
☐ Seedlings
☑ Leaves
...

F Select the genes used in analysis according their transcription in selected tissues:

Choose the transcript level based on TPM:

Most transcribed | Least transcribed 90th Genes | Percentile

Genes whose transcripts will represent 90% of all transcripts transcribed from the total number of protein-coding genes in each selected stage are included in the analysis.

OR

Most transcribed | Least transcribed 1000 Genes | Percentile

The genes included in the analysis for each stage will represent chosen number of genes with the highest transcription in the selected stage.

G Distribution of selected motifs :

← *Arabidopsis thaliana* (3 motifs, 4 stages)

Export 12 series — Close analysis

🟫	all -TC-element 296076 motifs in 23571 genes (of 23571 genes)
🟥	EarlyPollen -TC-element 79791 motifs in 6432 genes (of 6432 genes)
🟦	LatePollen -TC-element 19943 motifs in 1560 genes (of 1560 genes)
🟩	Leaves -TC-element 68612 motifs in 5539 genes (of 5539 genes)
🟫	all - My own motif 289345 motifs in 23558 genes (of 23571 genes)
🟥	EarlyPollen - My own motif 69242 motifs in 6427 genes (of 6432 genes)
🟦	LatePollen - My own motif 19794 motifs in 1560 genes (of 1560 genes)
🟩	Leaves -My own motif 63404 motifs in 5537 genes (of 5539 genes)
🟫	all - ARR10_core 907145 motifs in 23571 genes (of 23571 genes)
🟥	EarlyPollen - ARR10_core 267386 motifs in 6432 genes (of 6432 genes)
🟦	LatePollen - ARR10_core 64067 motifs in 1560 genes (of 1560 genes)
🟩	Leaves - ARR10_core 222073 motifs in 5539 genes (of 5539 genes)
	SIV - ARR10_core 45014 motifs in 1107 genes (of 1107 genes)

Show Motifs | Genes as Counts | Percentages Axis: Auto | Custom

Count of genes with motif in given interval as a percentage of total genes selected for the analysis. Save PNG

H Customize the settings:

Show Motifs | Genes as Counts | Percentages Axis: Auto | Custom

Vertical axis min: 20
Vertical axis max: 60
Horizontal axis min: -250
Horizontal axis max: 1000

Count of genes with motif in given interval as a percentage of total genes selected for the analysis. Save PNG

Enabled ☑
Color ●
Stroke Thin | Normal | Thick

Motif drill down
GATY matches 66% of selection (66% of all results) 175288
RATC matches 58% of selection (58% of all results) 153770

I Export genes with searched motif at certain position:

☑ EarlyPollen-ARR10-core Export this series

Interval	Genes with motif							
<-130; -100)	AT3G09390	AT5G38760	AT4G08670	AT5G15200	AT5G02380	AT3G04400	AT4G18100	AT1G70600
<-100; -70)	AT5G40260	AT5G38760	AT5G60390	AT5G09500	AT4G08670	AT3G18590	AT4G10260	AT5G02380
<-70; -40)	AT5G53820	AT2G28680	AT5G17340	AT4G08670	AT3G60245	AT3G07850	AT1G70600	AT3G08580
<-40; -10)	AT5G53820	AT3G18590	AT3G07850	AT5G02960	AT1G58983	AT4G27090	AT3G14040	AT4G18100
<-10; 20)	AT5G40260	AT3G09390	AT5G38760	AT1G18280	AT5G15200	AT3G60245	AT5G02380	AT1G48940
<20; 50)	AT5G40260	AT3G09390	AT1G18280	AT5G17340	AT4G08670	AT3G18590	AT5G02380	AT1G48940
<50; 80)	AT3G09390	AT5G60390	AT1G18280	AT2G28680	AT5G17340	AT3G18590	AT5G15200	AT5G02380
<80; 110)	AT5G53820	AT3G09390	AT1G18280	AT5G09500	AT2G28680	AT4G08670	AT5G20290	AT3G07850

J Export TPMs for genes expressed in selected tissue at the selected percentile, along with their expression levels in other tissues:

Gene Id	Matches	Tapetum	EarlyPollen	UNM	BCP	LatePollen	TCP	MPG	SIV PT	Sperm cell	Leaves	Seedlings
AT5G53820	28	1	6 477	5 198	7 756	7 094	13 432	757	2 345	1	0	4
AT5G40260	52	18	3 864	4 587	3 140	20	9	30	11	11	0	0
AT3G09390	43	5	3 748	3 621	3 875	1 200	2 209	191	20	1	324	232
AT5G38760	49	2	3 320	2 137	4 504	7 127	11 237	3 018	390	2	0	0
AT5G60390	56	869	3 179	3 756	2 603	143	238	48	16	0	937	1 573
AT1G18280	31	1	3 057	3 033	3 082	268	453	83	2 279	0	0	0
AT5G09500	42	2	2 941	3 256	2 627	62	75	50	1	1	0	1

Figure S1. A detailed overview of the workflow of the GOLEM software.
(a) One plant species across the plant Tree of Life is chosen, and the data are downloaded on the web browser (*Chara, Marchantia*, *Physcomitrium*, *Amborella*, *Azolla, Ceratopteris*, *Oryza*, *Zea*, *Solanum*, and *Arabidopsis*). If available, the positions of both TSS and ATG are given.
(b) The defined region (genomic interval) in the vicinity of the TSS or ATG, within the selected bucket size (bp), is chosen.
(c) A single custom motif of interest can be defined by users. Additionally, multiple motifs as well as degenerate motifs can also be searched for by users.

(d) Optionally the motif can be chosen from several motifs present in the software.
(e) The promoters of genes showing expression in selected tissues and developmental stages (sporophyte, male gametophyte), along with an analysis of genome-wide distribution regardless of transcription, are chosen for the analysis.
(f) The selection of genes that are highly or minimally transcribed in tissues or developmental stages of interest can be determined by the user, based on a specified percentile (default is the 90th percentile) or a certain number of genes included in the analysis.
(g) The exemplified "My own motif, TC-element and ARR10_core" motifs show various distributions upstream/downstream of TSS. The motif TC-element shows higher prevalence in the promoters of genes transcribed during Late pollen development (blue) and the motif ARR10_core shows higher prevalence in the promoters of the genes transcribed during early pollen (red) stages, in comparison to the genome-wide distribution (all; grey). The symbol (=) is used to change the curve order. The individual stages can be made invisible.
(h) The customization options for the output graph include adjusting curve color/stroke, axes size, and displaying either percentages or counts of genes with the motif of interest. Additionally, the output graph can be saved in PNG format.
(i) The accession numbers of genes with certain motif at the selected interval may be exported in XLSX format tables.
(j) The normalized expression values of genes, represented as Transcript Per Million (TPM) in selected tissue at a specified percentile or for a chosen number of genes, along with their expression in other tissues, can be exported as a table in XLSX format. Some plant icons were created with BioeRender.com.

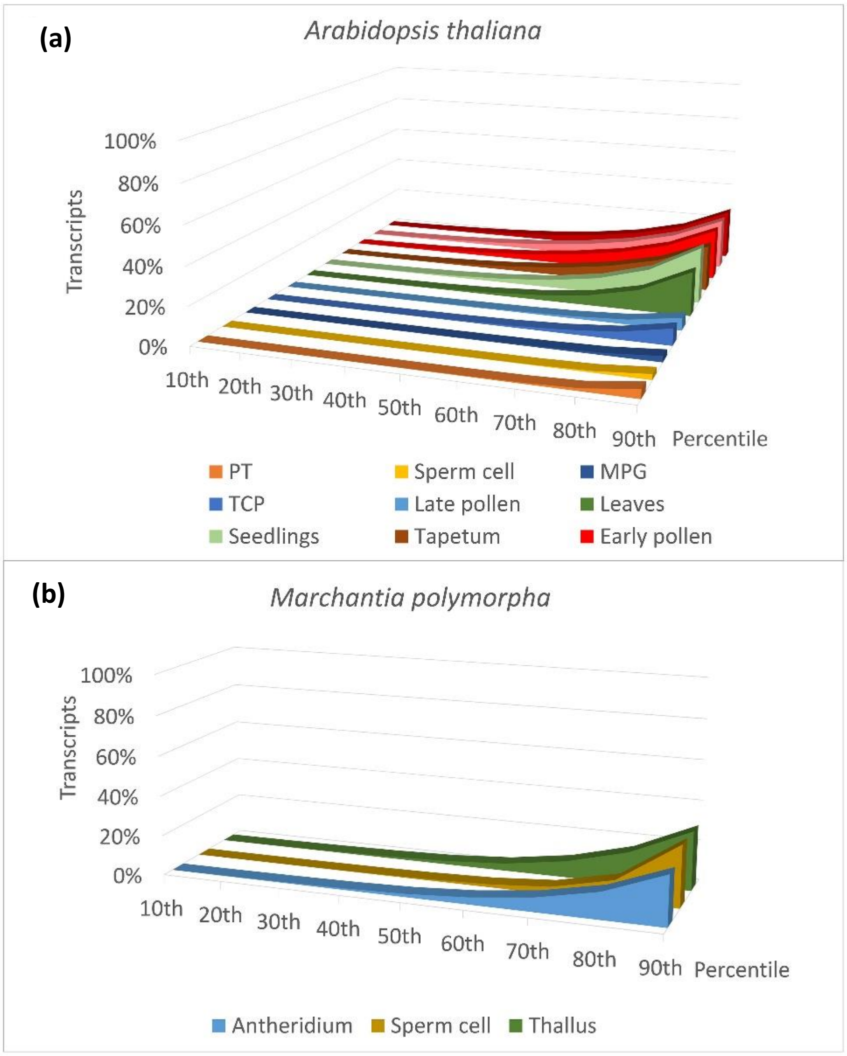


Figure S2. The number of genes contributing to expression programs varies between developmental stages or tissues.

(a) In early pollen stages, leaves and seedlings of *A. thaliana* the genes whose transcripts account for 90% of all transcripts transcribed from the total number of protein-coding genes (90th percentile) represent 27%, 24% and 30% of the total protein-coding genes, respectively. In late pollen stages, sperm cells and PT, those genes represent 7%, 3% and 5%, respectively.

(b) In bryophyte *M. polymorpha*, the genes whose transcripts comprise 90th percentile show more similar levels in antheridia, sperm cells and thallus, 25%, 30% and 29%, respectively. UNM, uninucleate microspore; BCP, bicellular pollen; early pollen, UNM + BCP; TCP, tricellular pollen; MPG, mature pollen grain; late pollen, TCP + MPG; PT, semi-in vivo grown pollen tube; Percentile, genes whose transcripts represent certain percent of all transcripts transcribed from the total number of protein-coding genes in each selected stage.

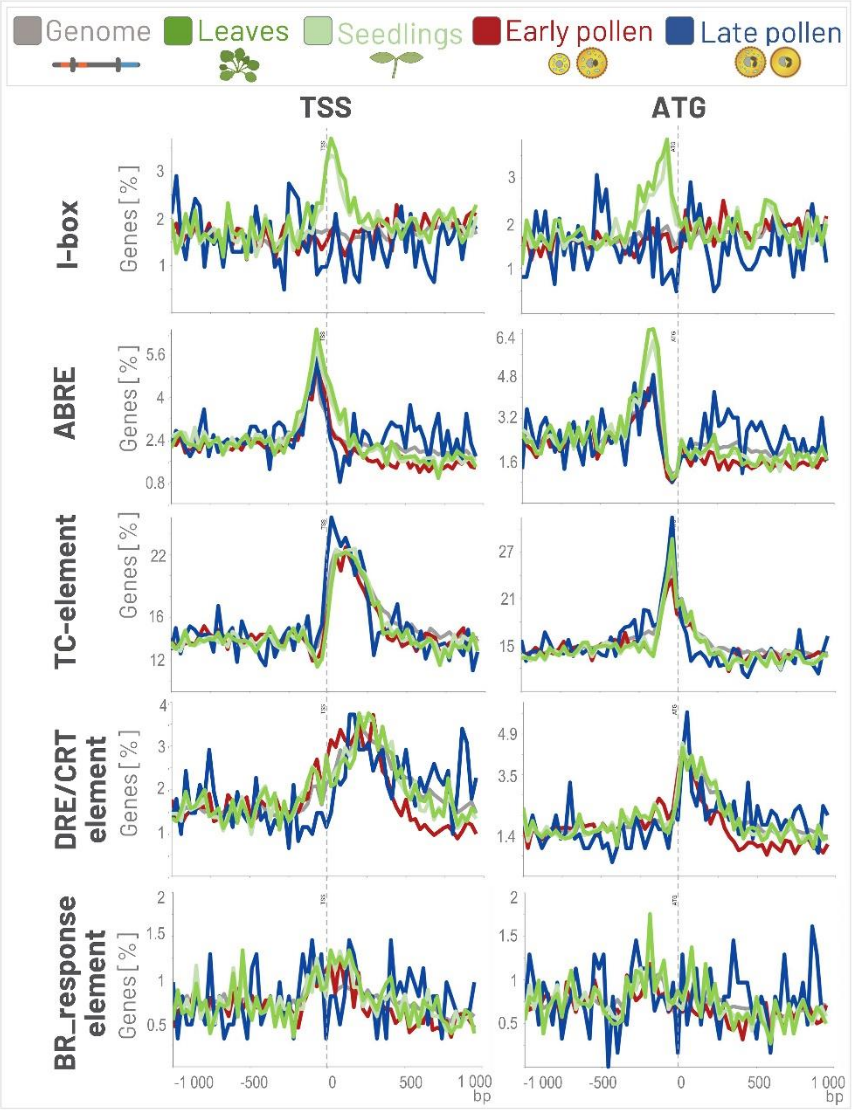


Figure S3. Example of the distribution of various motifs in the vicinity of TSS and ATG in *A. thaliana* with a focus on plant leaves and seedlings. Colored lines represent different datasets and indicate the percentage of genes containing selected motifs at specific positions in the promoters of the genes whose transcripts represent 80% of all transcripts transcribed from the total number of protein-coding genes in each selected stage: early pollen, late pollen, leaves, seedling and regardless of the transcription level (genome). The motifs were searched in the interval <-1000, 1000> bp, within the bucket size 30 bp, and the axis size was adjusted. I-motif (GATAAG); ABRE (ACGTG); TC_element (TCTTCT, TTTCTT, TTCTTC); DRE/CRT_element (CANNTG); BR_response element (CGTGYG); TSS, transcription start site; ATG, translation start site; early pollen, UNM + BCP; late pollen, TCP + MPG; bp, base pair.

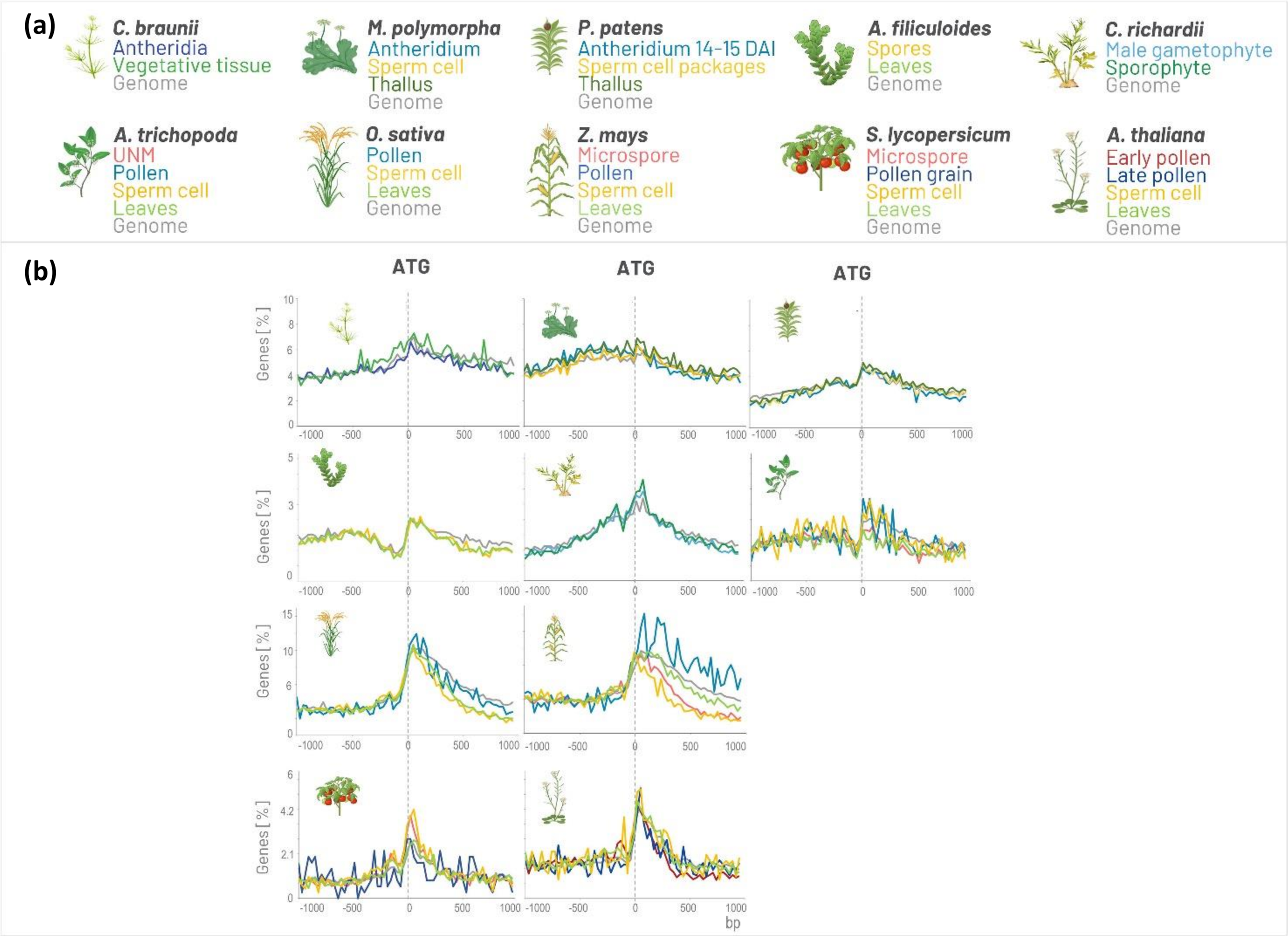


Figure S4. Example of the distribution of DRE/CRT element.

(a) The genomes analyzed across plant evolution include one streptophyte alga (*Chara*), two mosses (*Marchantia* and *Physcomitrium*), two ferns (*Azolla* and *Ceratopteris*), two monocots (*Oryza* and *Zea*), and two dicots (*Solanum* and *Arabidopsis*), as well as selected tissues and developmental stages.

(b) The distribution of DRE/CRT (CCGAC) element shows a gravitating around the start codon with higher occurrence in gene body than in 5' UTR region. This element seems to be over-represented in monocot genes. The genes expressed in the 90th percentile are shown within the range <-1000, 1000> bp, with a bucket size 30 bp; bp, base pair. The axis size was adjusted in each row.