

Supplemental Figures and Tables for:

**KNOWN AND NOVEL POST-TRANSCRIPTIONAL REGULATORY SEQUENCES ARE
CONSERVED ACROSS PLANT FAMILIES.**

Justin N Vaughn, Sally R Ellingson, Flavio Mignone, and Albrecht von Arnim

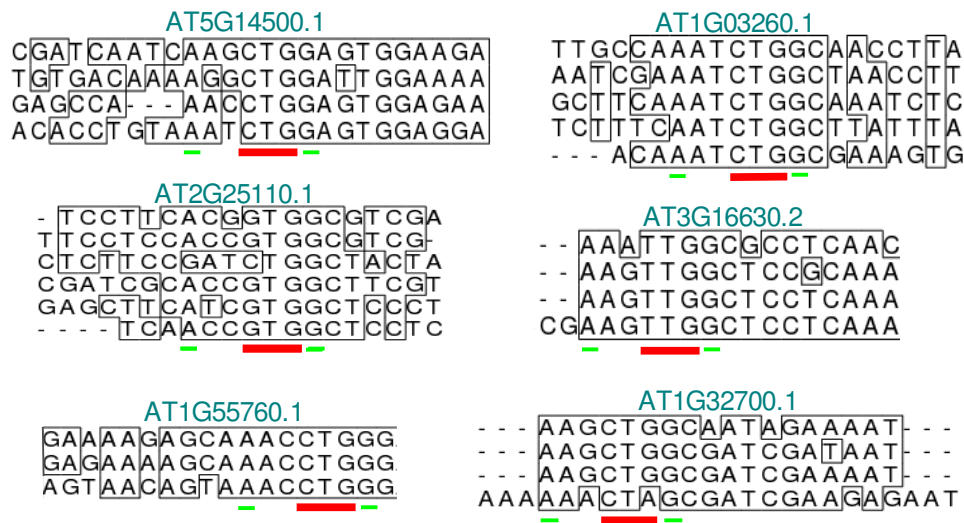
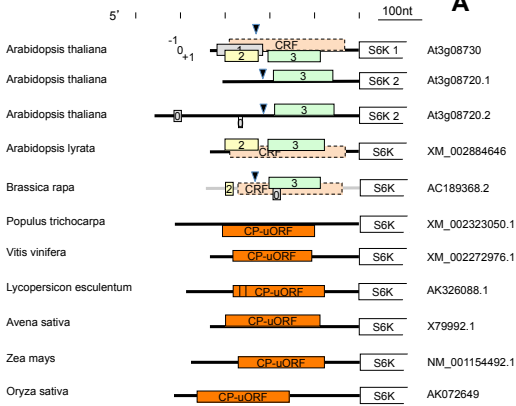
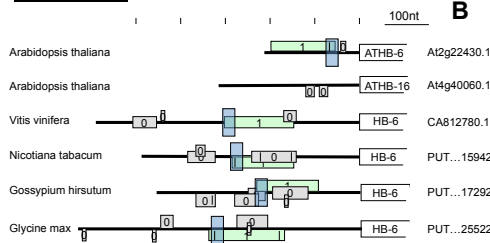


Figure S1. Non-canonical start sites have strong context and a variable +1 residue. Initially, continuous open reading frames that overlapped a conserved motif were aligned as protein. Alignments were then back-translated to create codon alignments. Each sub-panel shows the region of these alignments around each predicted start site. Conserved blocks are demarcated by black boxes. Red lines indicate the proposed start codon and green lines indicate the two most critical context sites. The *A. thaliana* accession from each group appears above the respective alignment.

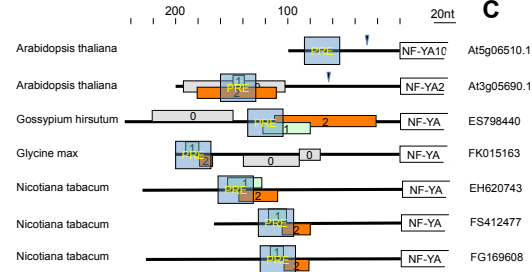
S6 Kinase



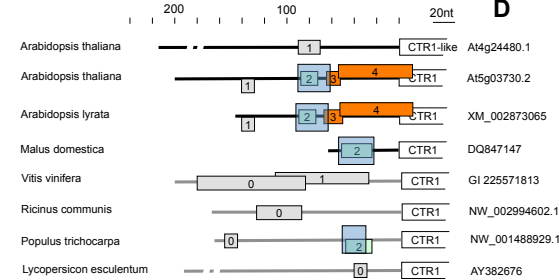
HD-ZIPATHB-6



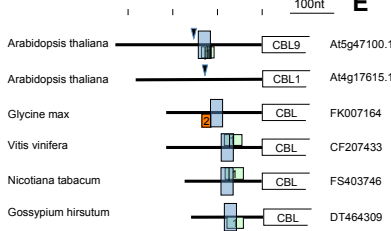
NF-YA subunit 10 and 2



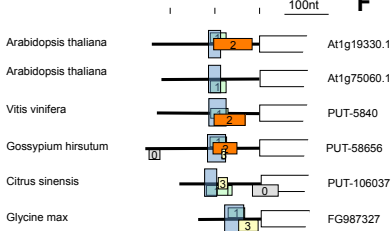
CTR1



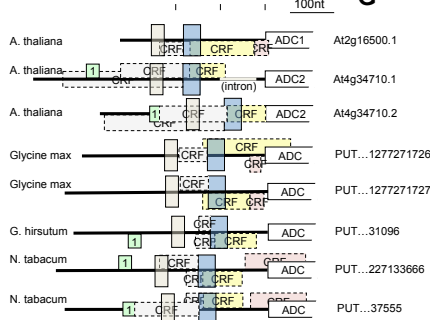
Calcineurin CBL9



Expressed protein AT1G19330

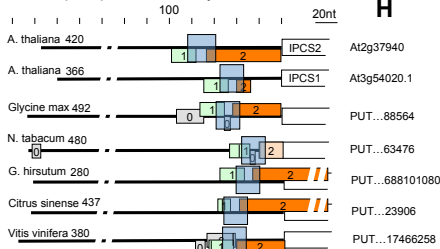


Arginine decarboxylase

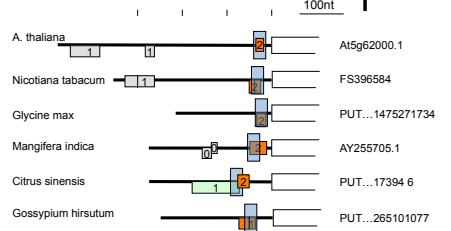


▼ Intron cDNA genomic DNA

Inositol phosphoceramide synthase



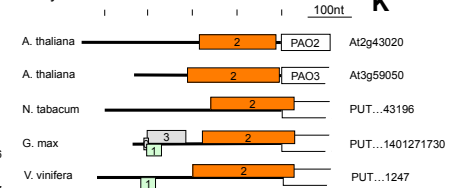
Auxin response factor ARF2



Polyamine oxidase PAO4



Polyamine oxidase PAO2/3



Supplemental Figure S2. Conservation and variation in the 5' UTR of mRNAs that harbor uAUG containing PREs and/or uORFs. Each subpanel shows the uORF pattern (colored rectangles) and, where present, the PRE (tall blue box) of one orthologous group; some orthologous groups were supplemented manually with members from additional species. The Gene Identifiers come from TAIR (At#g#####), PlantGDB (PUT...), or Genbank (all others). The beginning of the major ORF is symbolized by a white box. The 5' upstream region is indicated by a thick line, which is colored black in case the sequence is derived from spliced mRNA (5' UTR) and gray if derived from genomic DNA; the latter may contain introns. Vertical lines inside the uORF stand for internal in-frame AUG codons. Unless indicated, identical color or number does not imply sequence similarity. uORFs centered on the UTR are in the same frame as the major ORF, while those above the line and below the line are in the -1 and +1 frames, respectively. Selected exon-exon junctions are indicated by black arrowheads.

A) Ribosomal protein S6 kinase. This 5' UTR contains a single conserved-peptide uORF (orange), which was recognized in monocots (Tran et al., 2008) and also exists in many dicots. In the Brassicaceae (Brassica, Arabidopsis) the CP-uORF lacks an AUG start codon and is only present as a continuous reading frame (CRF, stippled). However, an alternative uORF (green) with partially conserved protein sequence is present, often together with an additional upstream uORF (yellow). Additional uORFs tend to be present in the Brassicaceae (gray). In Arabidopsis thaliana, S6 kinase is encoded by two tandemly repeated genes. S6K2 (At3g08720) is alternatively spliced and lacks the CRF, but contains the alternative uORF.

B) Homeodomain-zipper proteins ATHB-6 and 16. The PRE initiates a single long uORF of about 150 bases (green). Additional uORFs are abundant, and about half of them are exceptionally short. In Arabidopsis the PRE is located unusually close to the main AUG and the uORF has shifted farther upstream. A second Arabidopsis paralog lacks the PRE and the long uORF.

C) CCAAT enhancer binding protein, Nuclear factor YA subunits 10 and 2. The conserved PRE usually initiates two uORFs that are in a consistent frame arrangement (green, orange). Additional uORFs may be present (gray). Arabidopsis has an unusual uORF pattern.

D) CONSTITUTIVE TRIPLE RESPONSE1 kinase (CTR1). CTR1 is included here to represent uORF-containing 5'UTRs that lack a recognizable PRE. Arabidopsis CTR1 has four uORFs, one of which overlaps the mORF. uORF 2 can also be recognized in Malus (apple) and Populus (poplar) by virtue of weak conservation around its start codon (boxed, not a PRE). However, the uORF pattern is highly variable. Note the scale bar.

E) Calcineurin CBL9/CBL1. The PRE initiates a short uORF that usually contains additional start codons (green). In soybean, the typical uORF is replaced by a novel one slightly upstream (orange). Between the two Arabidopsis paralogs, one lacks the PRE and uORF, although the 5' UTRs are related in sequence (not shown).

F) Expressed protein At1g19330. The highly conserved PRE initiates up to three overlapping uORFs. uORF1 (green) is always present. Two Arabidopsis paralogs differ in the presence of uORF2 (orange).

G) Arginine decarboxylase (ADC). Aside from the known uORF1 (green, Chang et al., 2000), there is a conserved PRE (blue box), and a second conserved motif (brown box; TYCTTCTGAAN₁-₄TNRNNGTGGRGA). Several continuous reading frames (AUG-less, stippled) that end or begin in the PRE are indicated. One CRF (yellow) consistently harbors CGG arginine codons. A second CRF with consistent arginine codons lies upstream of the start codon of ADC (pink).

H) Inositol phosphoceramide synthase, also annotated as 'phosphatidic acid phosphatase-related'. The PRE contains the stop of an upstream uORF (#1) followed by the start of a longer downstream uORF (#2), which in some cases overlaps the mORF. uORFs #1 have limited sequence similarity, as do uORFs #2. In N. tabacum, the uAUG #2 starts an N-terminal extension of the mORF.

I) Auxin response factor 2. The conserved uORF is consistently short in length and does not overlap the mORF.

J) Polyamine oxidase 4. In this example, the frame arrangement between the two uORFs is variable. uORFs #2 are related in amino acid sequence to uORF #2 of PAO2 (panel K), except for that of Arabidopsis which has diverged. uORFs #1 are unrelated to each other.

K) Polyamine oxidase 2 and 3. uORF#2 is a CPuORF (Hayden and Jorgensen, 2007).

Table S1: A numerical account of the clustering of dicot mRNA sequence data into orthologous groups.

Species	Putative transcripts ^b	Number of sequences after clustering ^b	Number of orthologous groups with the species present ^c	Number of orthologous groups containing the given species and an <i>A. thaliana</i> sequence ^c
<i>Total</i>	698,941	158,613	31,909	10,122
<i>Arabidopsis thaliana</i> ^a	39,640	29,823	11,887	10,122
<i>Gossypium hirsutum</i>	98,420	28,385	14,541	8,016
<i>Citrus sinensis</i>	105,294	34,063	15,168	7,187
<i>Glycine max</i>	258,849	26,192	14,124	8,216
<i>Vitis vinifera</i>	64,796	14,091	9,763	6,380
<i>Nicotiana tabacum</i>	131,942	26,059	13,995	8,060

^aFrom TAIR9 release, as opposed to the informant species from PlantGDB. ^bA majority of putative transcripts represented fragmentary and unspliced mRNAs and were eliminated by orthologous clustering. ^cThough there are only 10,122 groups with an *A. thaliana* and an informant species representative, there are 31,909 groups total. Some of these groups represent lineage-specific genes, but many groups represent sequence fragments that were too short to cluster with their appropriate full-length transcripts (Figure S1).

Table S2. CPuORF monocot homologs. 'Implicit' indicates that the monocot homolog's start codon is not covered by the current 5'UTR annotation. Species are listed in the order that they appear in the alignment.

CPuORF accession	Protein sequence	
24n - implicit <i>Zea mays</i> , <i>Oryza sativa</i> , <i>Arabidopsis thaliana</i>	NM_001143483.1_1 NM_001068020.2_1 AT2G11890.1_1 NM_001143483.1_1 NM_001068020.2_1 AT2G11890.1_1	--TSNPTRIELTSSDRDELEDHLRAAAAATTTKDPSGYTTPSPVLGPQTSNPLLQFLHPK -----IELRSSDRDELEDHLRAAAAAAS-TPTASSTPTTTPPPSNSNPLLHLLHPP MLRRKPTKIQLKIEDREELEQSRKSPSTTTTAPSSSSAAS-----SLHHLIDPK *: * *: *: *: : : : : : : *: . : : : . * : : : . * PGAVPSKSQRIGIGLSTPPAPAPNPRPPHPPHGG PGAAPSKSHRIGL----PTNPNPNPKP----- HKNPSSKSDRIGLS----- . * * . * * :
25n - explicit <i>Saccharum officinarum</i> , <i>Zea mays</i> , <i>Sorghum bicolor</i> , <i>Arabidopsis thaliana</i>	AY644463.1_1 NM_001157169.1_1 XM_002445603.1_1 AT2G22500.1_1	MSQRASVPHSSCIGCALHSHLLVSSEMCPSSYWQQ MSQRASVPHSSCIAFALHSHLLVSSEMCPSSYWQQ MSQRASVPHSSCIAFALHSHLLVSSEMCPSSSWQQ MSQRSLIPHSSSIAFGLHSHLLISSEISSNSNWSL ****: : ****. * . *****: ****: . . . * *
26n - explicit <i>Arabidopsis thaliana</i> , <i>Oryza sativa</i>	AT1G67480.1_2 NM_001053449.1_3 AT1G67480.1_2 NM_001053449.1_3	MTFIDTEMCMRRNNINLTVIDSNEAIGMEHELDSARHQYSS-----VLTAIPFFSATL MAIDNCAMCVGGKGFYLSKETS DPSRKNH SKVSQYRMAFDAPRIKTETSKLKNLISAS * : : * : : : * : * : : . : : . * : : : : : : : FIPLSL----- FKPLSLTIPIGDGFHELFPVGHCHL * * * * *

Table S3. Mono- and di- nucleotide repeats in the 5' UTR.

Group Acc	Arabidopsis Acc	Annotation	repeat	type
ORTHOMCL311.27	AT1G01360.1	Encodes RCAR1 (regulatory components of ABA receptor). Interacts with and regulates the type 2C protein phosphatases (PP2Cs) ABI1 and ABI2. Functions as abscisic acid sensor.	AG	di
ORTHOMCL503.13	AT1G24260.2	Member of the MADs box transcription factor family. SEP3 is redundant with SEP1 and 2. Flowers of SEP1/2/3 triple mutants show a conversion of petals and stamens to sepals. SEP3 forms heterotetrameric complexes with other MADS box family members and binds to the CArG box motif.	AG	di
ORTHOMCL603.12	AT1G17590.2	"nuclear factor Y, subunit A8" (NF-YA8)	AG	di
ORTHOMCL681.24	AT3G15480.1	Protein of unknown function (DUF1218)	AG	di
ORTHOMCL760.25	AT3G09980.1	Family of unknown function (DUF662)	AG	di
ORTHOMCL842.18	AT3G16780.1	Ribosomal protein L19e family protein	AG	di
ORTHOMCL1033.8	AT5G28770.1	bZIP protein BZO2H3 mRNA, partial cds	AG	di
ORTHOMCL1107.28	AT2G45850.1	AT hook motif DNA-binding family protein	AG	di
ORTHOMCL1135.11	AT2G19520.1	Controls flowering.	AG	di
ORTHOMCL1141.24	AT1G79750.1	The malic enzyme (EC 1.1.1.40) encoded by AtNADP-ME4 is localized to chloroplasts. The gene is expressed throughout the whole plant and during embryogenesis and germination. A possible involvement in the fatty acid biosynthesis has been proposed.	AG	di
ORTHOMCL1558.26	AT5G56590.1	O-Glycosyl hydrolases family 17 protein	AG	di
ORTHOMCL1932.5	AT5G10580.2	Protein of unknown function, DUF599	AG	di
ORTHOMCL2102.23	AT2G34690.1	Gene product transports the glycolipid precursor sphingosine between membranes in vitro. Mutant constitutively expresses defense-related genes that accompany the hypersensitive response normally triggered by avirulent pathogens.	AG	di
ORTHOMCL2175.3	AT1G54460.1	TPX2 (targeting protein for Xklp2) protein family	AG	di
ORTHOMCL2204.14	AT2G35480.1	unknown protein	AG	di
ORTHOMCL2364.20	AT5G57330.1	Galactose mutarotase-like superfamily protein	AG	di
ORTHOMCL2603.5	AT3G62790.1	NADH-ubiquinone oxidoreductase-related	AG	di
ORTHOMCL2615.15	AT2G45400.1	involved in the regulation of brassinosteroid metabolic pathway	AG	di
ORTHOMCL3307.7	AT5G15790.1	RING/U-box superfamily protein	AG	di
ORTHOMCL3414.6	AT2G03350.1	Protein of unknown function, DUF538	AG	di
ORTHOMCL3825.1	AT5G54310.1	A member of ARF GAP domain (AGD), A thaliana has 15 members, grouped into four classes. Regulates	AG	di

Group Acc	Arabidopsis Acc	Annotation	repeat	type
		membrane trafficking and organ separation.		
ORTHOMCL4176.1	AT3G16150.1	N-terminal nucleophile aminohydrolases (Ntn hydrolases) superfamily protein	AG	di
ORTHOMCL4445.5	AT1G52240.2	Encodes a member of KPP-like gene family, homolog of KPP (kinase partner protein) gene in tomato. Also a member of the RopGEF (guanine nucleotide exchange factor) family, containing the novel PRONE domain (plant-specific Rop nucleotide exchanger), which is exclusively active towards members of the Rop subfamily .	AG	di
ORTHOMCL4570.6	AT2G02820.2	Encodes a putative transcription factor (MYB88), involved in stomata development, double loss of MYB88 and FLP (MYB124) activity results in a failure of guard mother cells (GMCs) to adopt the guard cell fate, thus they continue to divide resulting in abnormal stomata consisting of clusters of numerous guard cell-like cells. This phenotype is enhanced in double mutants over the single mutant flp phenotype.	AG	di
ORTHOMCL5703.1	AT1G31070.2	Encodes a protein that functions as an N-acetylglucosamine-1-phosphate uridylyltransferase that catalyzes the formation of UDP-N-acetylglucosamine (UDP-GlcNAc). This is an essential precursor for glycolipid and glycoprotein synthesis and is also used for regulatory protein modification in signaling pathways. The enzyme can also catalyze the reverse reaction using both UDP-GlcNAc and the less common UDP-N-acetylgalactosamine as substrates.	AG	di
ORTHOMCL8163.2	AT4G36960.2	RNA-binding (RRM/RBD/RNP motifs) family protein	AG	di
ORTHOMCL8225.1	AT4G22360.1	SWIB complex BAF60b domain-containing protein	AG	di
ORTHOMCL9007.2	AT1G10030.1	homolog of yeast ergosterol28 (ERG28)	AG	di
ORTHOMCL653.1	AT3G12630.1	A20/AN1-like zinc finger family protein	CT	di
ORTHOMCL885.15	AT3G46460.1	ubiquitin-conjugating enzyme 13 (UBC13)	CT	di
ORTHOMCL921.30	AT2G28950.1	Encodes an expansin. Naming convention from the Expansin Working Group (Kende et al, 2004. Plant Mol Bio). Involved in the formation of nematode-induced syncytia in roots of Arabidopsis thaliana.	CT	di
ORTHOMCL1149.19	AT1G68140.1	Protein of unknown function (DUF1644)	CT	di
ORTHOMCL1626.21	AT5G19090.1	Heavy metal transport/detoxification superfamily protein	CT	di
ORTHOMCL1776.10	AT1G20330.1	Encodes a sterol-C24-methyltransferases involved in sterol biosynthesis. Mutants display altered sterol composition, serrated petals and sepals and altered cotyledon vascular patterning as well as ectopic	CT	di

Group Acc	Arabidopsis Acc	Annotation	repeat	type
		endoreduplication. This suggests that suppression of endoreduplication is important for petal morphogenesis and that normal sterol composition is required for this suppression.		
ORTHOMCL1920.7	AT4G39850.1	Encodes a peroxisomal protein of the ATP binding cassette (ABC) transporter class (PMP subfamily) with significant identity to the human X-linked adrenoleukodystrophy protein (ALDP). The gene product promotes germination and represses embryo dormancy. ABI3, ABA1, FUS3 and LEC1 are epistatic to this gene. Mutants accumulate fatty acyl CoA suggesting a defect in uptake of fatty acyl CoA into the peroxisome.	CT	di
ORTHOMCL1954.7	AT5G47200.1	RAB GTPase homolog 1A (RAB1A)	CT	di
ORTHOMCL2359.10	AT5G60980.2	Nuclear transport factor 2 (NTF2) family protein with RNA binding (RRM-RBD-RNP motifs) domain	CT	di
ORTHOMCL2449.7	AT4G28660.1	Similar to PsbW subunit of photosystem II.	CT	di
ORTHOMCL3527.11	AT1G28280.2	VQ motif-containing protein	CT	di
ORTHOMCL5168.3	AT4G09650.1	Encodes the chloroplast ATPase delta-subunit.	CT	di
ORTHOMCL66.25	AT4G38920.1	vacuolar-type H(+)-ATPase C3 (VHA-C3)	GA	di
ORTHOMCL103.25	AT1G26880.1	Ribosomal protein L34e superfamily protein	GA	di
ORTHOMCL163.19	AT4G15000.1	Ribosomal L27e protein family	GA	di
ORTHOMCL222.9	AT1G22400.1	UGT85A1	GA	di
ORTHOMCL822.2	AT1G34760.2	Encodes a 14-3-3 protein. Binds H+-ATPase in response to blue light.	GA	di
ORTHOMCL876.10	AT3G57870.1	Encodes a SUMO liage that directs the attachment of the small protein SUMO to target proteins via an isopeptide bond. This enzyme is localized to the nucleus and plants with reduced levels of this protein show higher sensitivity to ABA in root growth inhibition assays. It has high similarity to the yeast UBC9 SUMO ligase and is sometimes referred to by that name.	GA	di
ORTHOMCL898.17	AT3G12120.1	Major enzyme responsible for the synthesis of 18:2 fatty acids in the endoplasmic reticulum. Contains His-rich motifs, which contribute to the interaction with the electron donor cytochrome b5. Mutations in this gene suppress the low temperature-induced phenotype of Arabidopsis tocopherol-deficient mutant vte2.	GA	di
ORTHOMCL1071.13	AT5G65790.1	Encodes a putative MYB transcription factor.	GA	di
ORTHOMCL1557.29	AT4G31290.1	ChaC-like family protein	GA	di
ORTHOMCL1798.27	AT1G62880.1	Cornichon family protein	GA	di
ORTHOMCL2004.6	AT5G41060.1	DHHC-type zinc finger family protein	GA	di

Group Acc	Arabidopsis Acc	Annotation	repeat	type
ORTHOMCL2076.17	AT2G42590.2	14-3-3 gene. Binds calcium and displays induced structural changes.	GA	di
ORTHOMCL2205.9	AT1G32210.1	Encodes protein involved in suppression of apoptosis. Complements a mammalian apoptosis suppressor mutation.	GA	di
ORTHOMCL3077.6	AT5G65640.2	beta HLH protein 93 (bHLH093)	GA	di
ORTHOMCL3293.11	AT3G05000.1	Transport protein particle (TRAPP) component	GA	di
ORTHOMCL3922.2	AT5G11790.1	N-MYC downregulated-like 2 (NDL2)	GA	di
ORTHOMCL4350.1	AT2G14910.1	unknown protein	GA	di
ORTHOMCL4611.5	AT3G46830.1	RAB GTPase homolog A2C (RABA2c)	GA	di
ORTHOMCL4647.1	AT1G01225.1	NC domain-containing protein-related	GA	di
ORTHOMCL5564.3	AT1G79430.2	Encodes gene product that is required for several aspects of phloem development in the root: (1) the specific divisions organizing the phloem pole, (2) sieve element differentiation and (3) the expression of a companion-specific gene. Mutant has a defect in the organization of phloem poles in the root. apl seedlings have a short, determinate root with only occasional lateral branches.	GA	di
ORTHOMCL7228.1	AT1G77610.1	EamA-like transporter family protein	GA	di
ORTHOMCL7299.2	AT1G09815.1	polymerase delta 4 (POLD4)	GA	di
ORTHOMCL10767.2	AT1G67630.1	DNA polymerase alpha 2 (POLA2)	GA	di
ORTHOMCL250.26	AT5G61130.1	Encodes a plasmodesmatal localized protein with callose binding activity. The function of PDCB1 is unknown but overexpression leads to the hyperaccumulation of callose.	TC	di
ORTHOMCL434.20	AT1G52140.1	unknown protein	TC	di
ORTHOMCL1178.20	AT1G23260.1	MMZ1/UEV1A encodes a protein that may play a role in DNA damage responses and error-free post-replicative DNA repair by participating in lysine-63-based polyubiquitination reactions. UEV1A can form diubiquitin and triubiquitin chains in combination with UBC13A/UBC35 in vitro. It can also functionally complement an mms2 mutation in budding yeast, both by increasing mms2 mutant viability in the presence of the DNA damaging agent MMS, and by reducing the rate of spontaneous DNA mutation. However, a combination of MMZ1/UEV1A and UBC13A do not do a good job of rescuing an mms2 ubc13 double mutant in yeast. MMZ1/UEV1A transcripts are found at low levels in most plant organs, but cannot be detected in the pollen. Transcript levels do not appear to be stress-inducible. The uev1a-1 mutant shows normal sensitivity to MMS in germination assays suggesting that UEV1A is not	TC	di

Group Acc	Arabidopsis Acc	Annotation	repeat	type
		required for DNA damage tolerance during this developmental stage.		
ORTHOMCL1196.30	AT1G14685.1	Arabidopsis GBP Basic Penta Cysteine 1	TC	di
ORTHOMCL2008.9	AT3G24160.1	Encodes a putative Type 1 membrane protein (PMP).	TC	di
ORTHOMCL4026.4	AT4G16780.1	Encodes a homeodomain-leucine zipper protein that is rapidly and strongly induced by changes in the ratio of red to far-red light. It is also involved in cell expansion and cell proliferation and in the response to auxin.	TC	di
ORTHOMCL4214.5	AT3G06740.1	Encodes a member of the GATA factor family of zinc finger transcription factors.	TC	di
ORTHOMCL1802.11	AT1G08650.1	Encodes a phosphoenolpyruvate carboxylase kinase that is expressed at highest levels in leaves. Expression is induced by light.	A	mono
ORTHOMCL2157.30	AT5G38650.1	Proteasome maturation factor UMP1	A	mono
ORTHOMCL2478.5	AT4G13040.1	Integrase-type DNA-binding superfamily protein	A	mono
ORTHOMCL3106.9	AT4G38690.1	PLC-like phosphodiesterases superfamily protein	A	mono
ORTHOMCL4276.4	AT2G40840.1	Encodes a cytosolic protein with transglucosidase and amylomaltase activity. It is an essential component of the pathway from starch to sucrose and cellular metabolism in leaves at night. The protein binds to heteroglycans and utilizes glucose, mannose and xylose as acceptors. Fucose and galactose can also act as acceptors but less efficiently than the previous three. It was also was also recently reported to act on maltodextrins. On the other hand, arabinose and fructose were not efficiently used. Its role probably includes metabolizing maltose exported from the chloroplast. Studies using maltose extracted from the double mutant be2-1 be3-2 showed that this enzyme is preferentially active of β	A	mono
ORTHOMCL4314.6	AT2G27020.1	Encodes 20S proteasome alpha 7 subunit PAG1.	A	mono
ORTHOMCL4567.3	AT2G02180.1	Necessary for the efficient multiplication of tobamoviruses.	A	mono
ORTHOMCL1526.27	AT5G52840.1	NADH-ubiquinone oxidoreductase-related	G	mono
ORTHOMCL3556.6	AT1G17370.2	oligouridylate binding protein 1B (UBP1B)	T	mono