# Supporting Information

## Liu et al. 10.1073/pnas.1301009110

### SI Text

**Plant Growth Conditions and Sample Collection.** Seeds of *Zea mays* cv. White Crystal, a glutinous maize cultivar, were purchased from a local supplier and stored at room temperature. For germination, seeds were imbibed in distilled water at 6:00 PM by first shaking for 10 min at 200 rpm and then germinating on wet filter paper on Petri dishes in the greenhouse under natural sunlight conditions (approximately 6:00 AM to approximately 7:00 PM) during the summer. The first set of seeds was imbibed on 6/2/2011 and leaf tissue for RNA preparation was taken every 6 h for 2 d [up to the 48th hour (T48)]. A second set of seeds was imbibed at 6:00 PM on 7/24/2011 and leaf tissue for RNA preparation was taken at T54, T60, T66, and T72. The epicotyls were collected, and then coleoptiles were removed before the embryonic leaves were isolated. These isolated leaves were then stored in liquid nitrogen. The whole procedure was done in an hour. The maximum photosynthetic photon flux density around noon was ~1,600 $\mu mol \cdot m^{-2} \cdot s^{-1}$, and the day/night temperatures were 30–35 °C/25–28 °C, with a relative humidity of 60–70%. Samples of the embryonic leaves and the shoot apical meristem (SAM) were collected from dry seeds (T00) and germinating seeds every 6 h up to 72 h for leaf anatomic examination and for RNA isolation and transcriptomic analysis. For RNA isolation, ~200 plumules were dissected from the seeds and after removing the sheath tissue, the entire embryonic leaves and the SAM were stored in liquid $N_2$ until RNA extraction. Our tissue samples for transcriptomic analysis contained all of the four to five embryonic leaves in the maize seeds. As the first leaf is the largest, it should have the largest contribution to the transcriptomes. However, as new kranz structures (KSs) develop in these leaves, the transcriptomes are useful for hypothesizing the regulatory genes involved in KS development. The samples also contained SAM, but because SAM is small, its contribution to the transcriptomes may not be significant.

**Anatomical Studies.** For anatomical examination, samples of the embryonic leaves and SAM were fixed in 2.5% (wt/vol) glutaraldehyde and postfixed in 1% (wt/vol) OsO4, both in 0.1 M sodium phosphate buffer (1). After dehydration through an ethanol series, the samples were infiltrated and embedded in Spurr's resin. Semithin cross-sections (0.9 μm) were cut from the base of embryonic leaves with an Ultracut E Microtome (Reichert-Jung) and stained with 0.1% (wt/vol) toluidine blue and 0.1% (wt/vol) borax for observation with a light microscope (Olympus).

**RNA Extraction and Sequencing.** Total RNA was extracted using TRIzol reagent (Invitrogen), and 1 μL TURBO DNase (Ambion) per 10 μg RNA was added and incubated for 30 min at 37 °C to remove traces of contaminating DNA; and RNA was subsequently purified by the phenol:chloroform procedure. The RNAs were quantified using Qubit RNA Pico reagent (Invitrogen) and assessed for purity and quality using NanoDrop (Thermo Scientific) and BioAnalyzer (Agilent). RNA-seq libraries were prepared using the Illumina Standard mRNA-seq library preparation kit (Illumina) based on the manufacture's protocol with the following modifications. Briefly, 15 μg of total RNA was subjected to two rounds of oligo-dT bead purification, fragmented for 2.75 min, and primed for first strand cDNA using 0.8× amount of random primer. After double strand cDNA synthesis and adaptor ligation, each sample was fractionated on 2% (wt/vol) Low Range Ultra Agarose (Bio-Rad) gels, from which the libraries of two different size ranges (~300 bp and ~400 bp) were excised. The purified libraries were then independently amplified by 15 cycles of PCR and cleaned up using AMPure XP beads (Beckman Agencourt). RNA-seq libraries were subjected to Illumina sequencing by paired-end 2*101 nt (both ends with 101 nucleotides long) on a HiSeq2000 at the NGS Core Facility of Biodiversity Research Center in Academia Sinica, Taiwan. A total of 195–292 million pairs of raw reads was obtained from the two libraries (two lanes) for each time point sample; each pair of reads was counted as two reads. For each time point, ~390–584 million raw reads and 240–351 million mappable reads were obtained.

**Data Processing and Analysis.** Low-quality bases and reads were removed by three criteria: (*i*) the consecutive bases from the end of a read with a default low-quality score of 2 [phred score of 2 or Q2 (2)], (*ii*) the bases from the beginning of a read until all of the scores of the first 20 remaining bases were at least Q20 (the base call error rate of ~1%), and (*iii*) the trimmed reads with less than 60 remaining bases. (Phred score is a general metric for the accuracy of a sequencing platform (2). The Q2 indicator does not give a specific error rate, but rather indicates a specific portion of the read that should not be used in further analyses.) Each pair of reads was treated as two single reads. Then, all alignment and quantification processes followed the "alternative protocol B" in ref. 3). The processed reads were mapped to the maize genome (ZmB73_RefGen_v2) and the working gene set (ZmB73_5a_WGS, http://ftp.maizesequence.org/), using Tophat version 1.3.3 (http://tophat.cbcb.umd.edu/) and its embedded aligner Bowtie version 0.12.7 (http://bowtie-bio.sourceforge.net). Each read was aligned by the "-n" policy, and at most 10 hits were allowed. The expression level of each working gene set (WGS) gene was estimated in the reads per kilobase per million (RPKM) (4), using Cufflinks version 2.0.2 (http://cufflinks.cbcb.umd.edu/).

We selected only genes with RPKM ≥1 in two or more time points for further analysis. To compare the expression levels of the selected genes across all time points, we applied the upper quartile normalization procedure (5), which reduces the bias in RPKM values due to highly expressed genes. We used the transcriptome at T42 as the reference because its RPKMs were most evenly distributed among the 13 transcriptomes. The RPKM values at a time point were scaled by a constant factor to make the 75th percentile of the RPKM values for all expressed genes at that time point equal to the one at T42.

**Expression Profile Correlation and Clustering.** The clustering tool gCLUTO (Graphical Clustering Toolkit) was downloaded from http://glaros.dtc.umn.edu/gkhome/cluto/gcluto/overview. It includes several hierarchical clustering methods and we selected the repeated bisection method, which is a top-down hierarchical clustering algorithm that separates samples with the largest distance. We calculated the Pearson correlation coefficient (PCC) between each pair of expression profiles of the genes under study and used the PCC values to cluster the expression profiles. A scalar quantity, called the figure of merit (FOM) (6), was used to assess the quality of the clustering algorithm and to determine the optimal number of clusters.

To determine the functional enrichment for each cluster, we used the MapMan functional annotation (7). The annotated entries with the functional category name of "not assigned.unknown" were excluded in this study. We then used Fisher's exact test to examine whether a function was significantly over- or un-

derrepresented in a selected cluster of genes against the set of all expressed genes with MapMan annotations.

**Identification of Differentially Expressed Genes.** We used the nonparametric method of Tarazona et al. (8) to identify differentially expressed genes (DEGs) between two samples. Here, we set the $q$ value (differentially expression probability) in the method to be 0.7, because in our dataset genes with $q > 0.7$ have at least a 2.3-fold change in RPKM between the two samples. The test is stringent. For example, the number of DEGs between T06 and T12 was only 39, and because 30,255 genes (tests) were used, the probability of passing a test was only ~39/30,000 = 0.0013, which is much smaller than 0.01 (the 1% significance level). Thus, the test is conservative and can be used to identify DEGs between two time points.

The above analysis was between two time points. To find all genes differentially expressed during leaf development, we have developed the following method to make all pairwise comparisons between time points, which takes care of the problem of multiple testing. Suppose the probability of "success" in a single test is $p$ and a total of $N$ tests are made. We use the binomial expansion to calculate the probability of at least $X$ successes in $N$ tests. In the above, we have seen that $p$ is only ~0.0013. To be conservative, let us take $p = 0.005$. Among the 13 time points there are 78 possible pairwise comparisons, that is $n = 78$. Then the probability to have at least $X = 4$ "successes" is only ~0.0001. Under this criterion, we find that 13,907 (46%) of the 30,255 expressed genes and 590 (47%) of the 1,238 expressed TF genes are differentially expressed during the 72-h time course.


1. Sheue CR, et al. (2007) Bizonoplast, a unique chloroplast in the epidermal cells of microphylls in the shade plant Selaginella erythropus (Selaginellaceae). *Am J Bot* 94(12):1922–1929.
2. Ewing B, Hillier L, Wendl MC, Green P (1998) Base-calling of automated sequencer traces using phred. I. Accuracy assessment. *Genome Res* 8(3):175–185.
3. Trapnell C, et al. (2012) Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat Protoc* 7(3):562–578.
4. Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B (2008) Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat Methods* 5(7):621–628.
5. Bullard JH, Purdom E, Hansen KD, Dudoit S (2010) Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments. *BMC Bioinformatics* 11:94.
6. Yeung KY, Haynor DR, Ruzzo WL (2001) Validating clustering for gene expression data. *Bioinformatics* 17(4):309–318.
7. Thimm O, et al. (2004) MAPMAN: A user-driven tool to display genomics data sets onto diagrams of metabolic pathways and other biological processes. *Plant J* 37(6):914–939.
8. Tarazona S, García-Alcalde F, Dopazo J, Ferrer A, Conesa A (2011) Differential expression in RNA-seq: A matter of depth. *Genome Res* 21(12):2213–2223.


| | T00 | T06 | T12 | T18 | T24 | T30 | T36 | T42 | T48 | T54 | T60 | T66 | T72 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| base | 0.21 | 0.19 | 0.19 | 0.22 | 0.33 | 0.39 | 0.51 | 0.44 | 0.46 | 0.63 | 0.64 | 0.49 | 0.52 |
| -1 cm | 0.04 | 0.03 | 0.04 | 0.03 | 0.06 | 0.06 | 0.08 | 0.07 | 0.08 | 0.10 | 0.11 | 0.10 | 0.10 |
| +4 cm | 0.00 | 0.00 | 0.01 | 0.00 | 0.01 | 0.01 | 0.01 | 0.01 | 0.02 | 0.01 | 0.01 | 0.02 | 0.02 |
| tip | 0.02 | 0.02 | 0.02 | 0.02 | 0.03 | 0.03 | 0.04 | 0.04 | 0.04 | 0.03 | 0.03 | 0.04 | 0.04 |

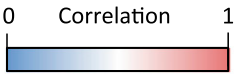


**Fig. S1.** Comparison between mature leaf and embryonic leaf time course transcriptomes. Pearson's correlation coefficients (PCCs) between the four mature leaf transcriptomes of Li et al. (1), as indicated in the first column, and our time course transcriptomes, as indicated in the first row.


1. Li P, et al. (2010) The developmental dynamics of the maize leaf transcriptome. *Nat Genet* 42(12):1060–1067.

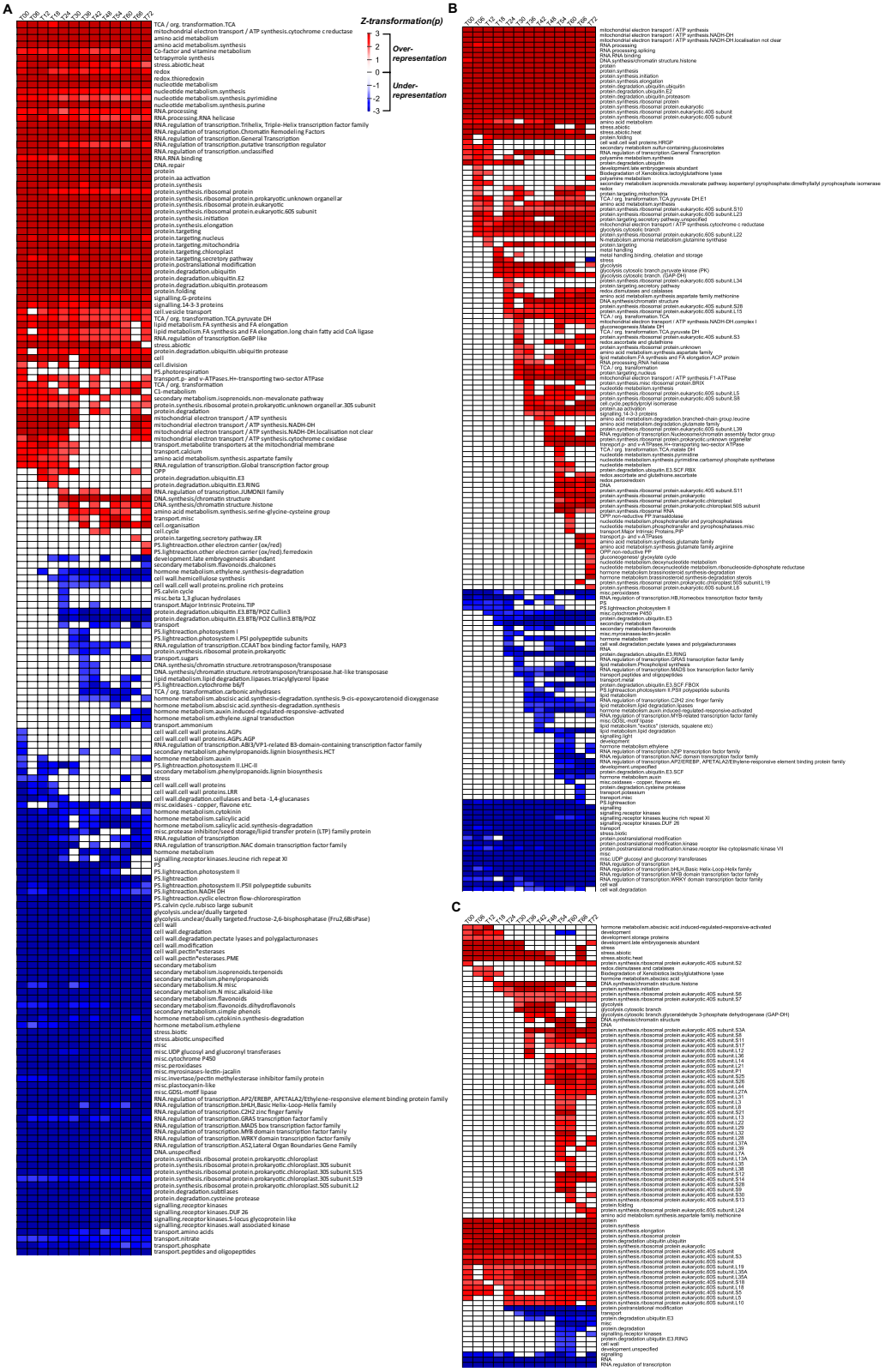


**Fig. S2.** MapMan functional categories enriched in expressed genes at each time point. Enrichment tests conducted using (*A*) all expressed genes and (*B*) genes with top 10% or (*C*) top 1% RPKM values in each time point. The heatmap is color-coded based on the Z transformation of adjusted *P* values (1). Blue and red represent under- and overrepresented categories, respectively.

1. Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J R Stat Soc, B* 57(1):289–300.

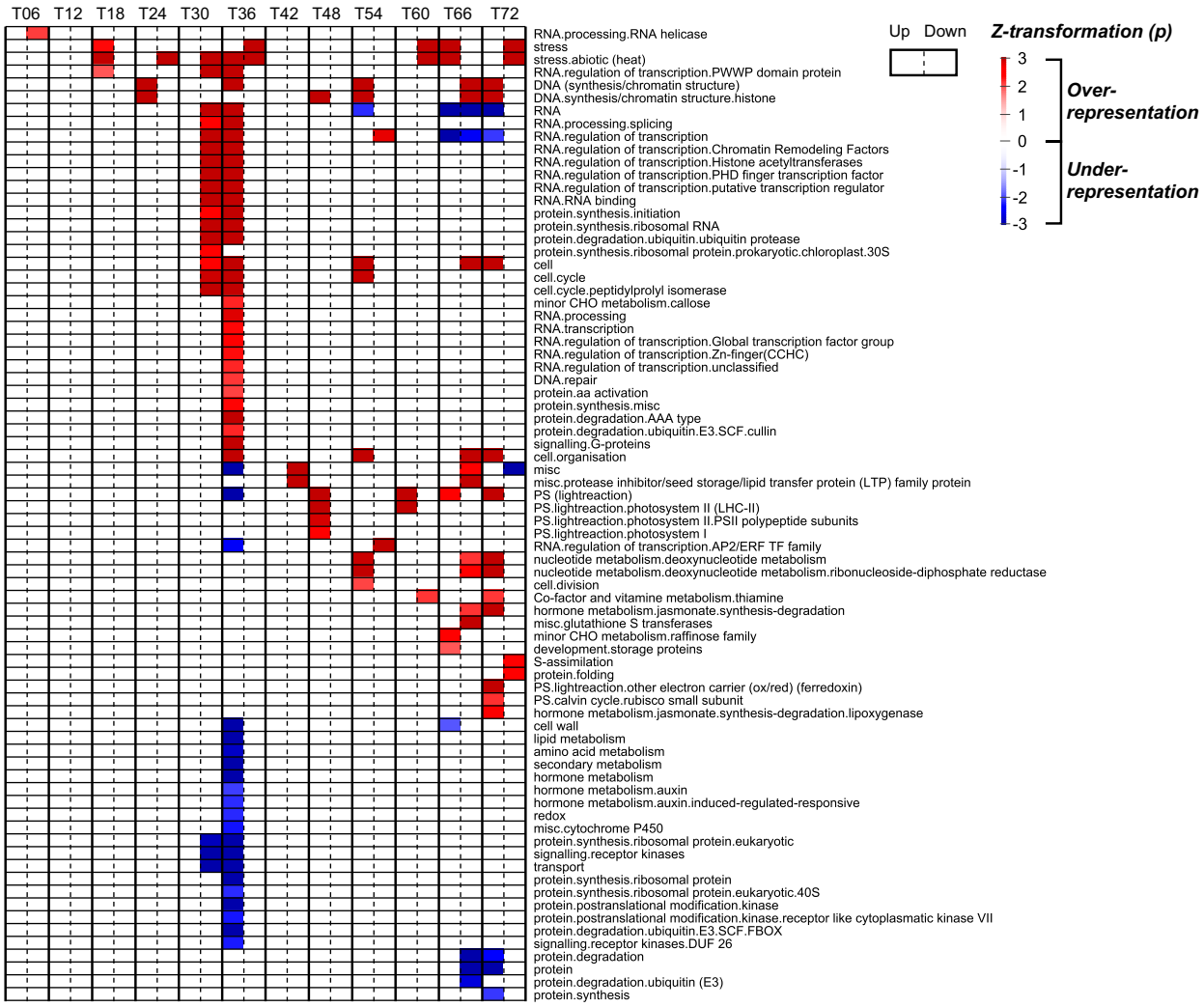


**Fig. S3.** (Continued)

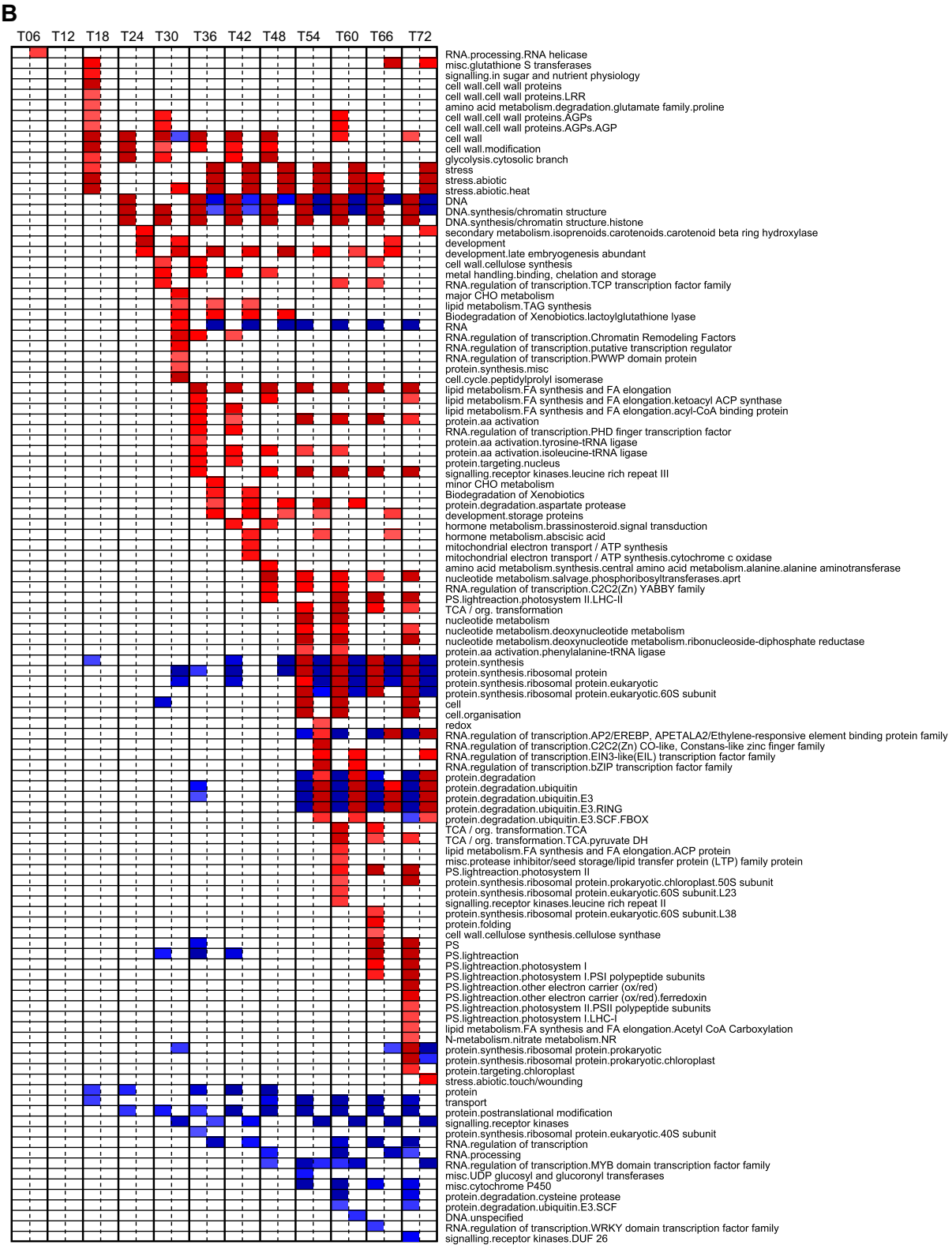


**Fig. S3.** (Continued)

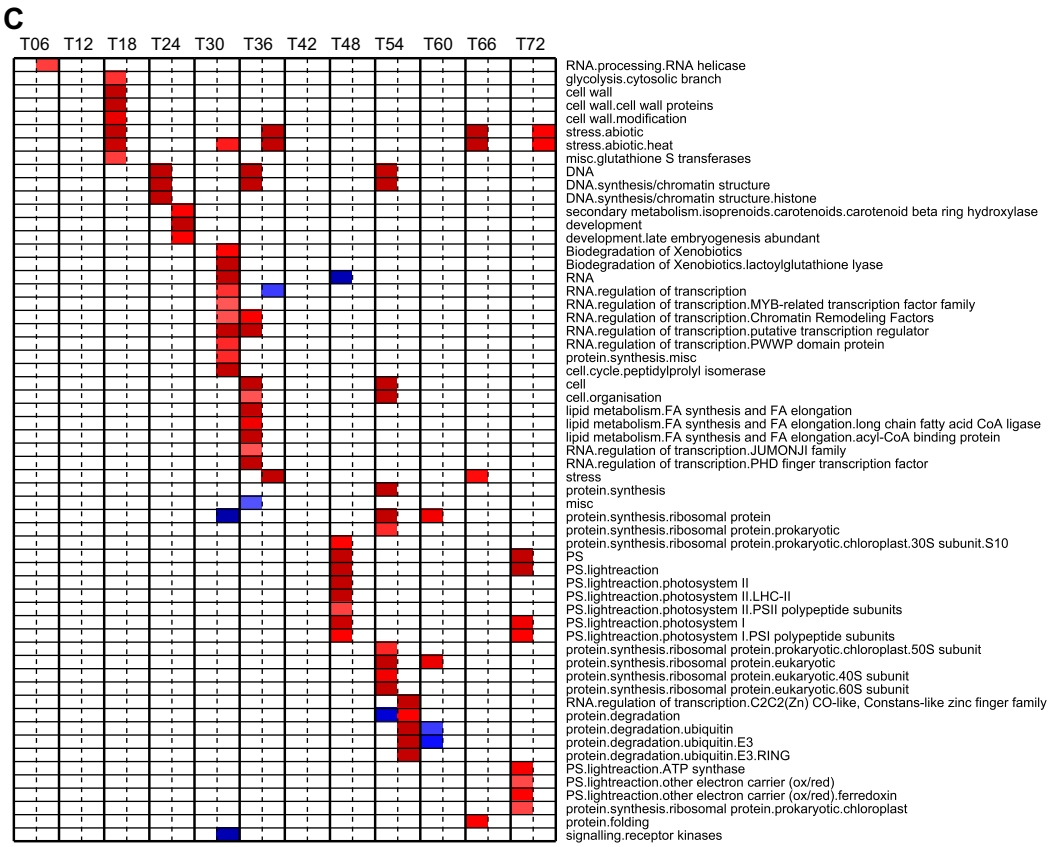


**Fig. S3.** (Continued)


**Fig. S3.** MapMan functional categories enriched in differentially expressed genes (DEGs). Enrichment tests with (*A*) type 1 DEGs, (*B*) type 2 DEGs, and (*C*) type 3 DEGs. (*D*) Clusters of all three types of DEGs. (*E*) Enrichment tests on each DEG cluster shown in *D*.

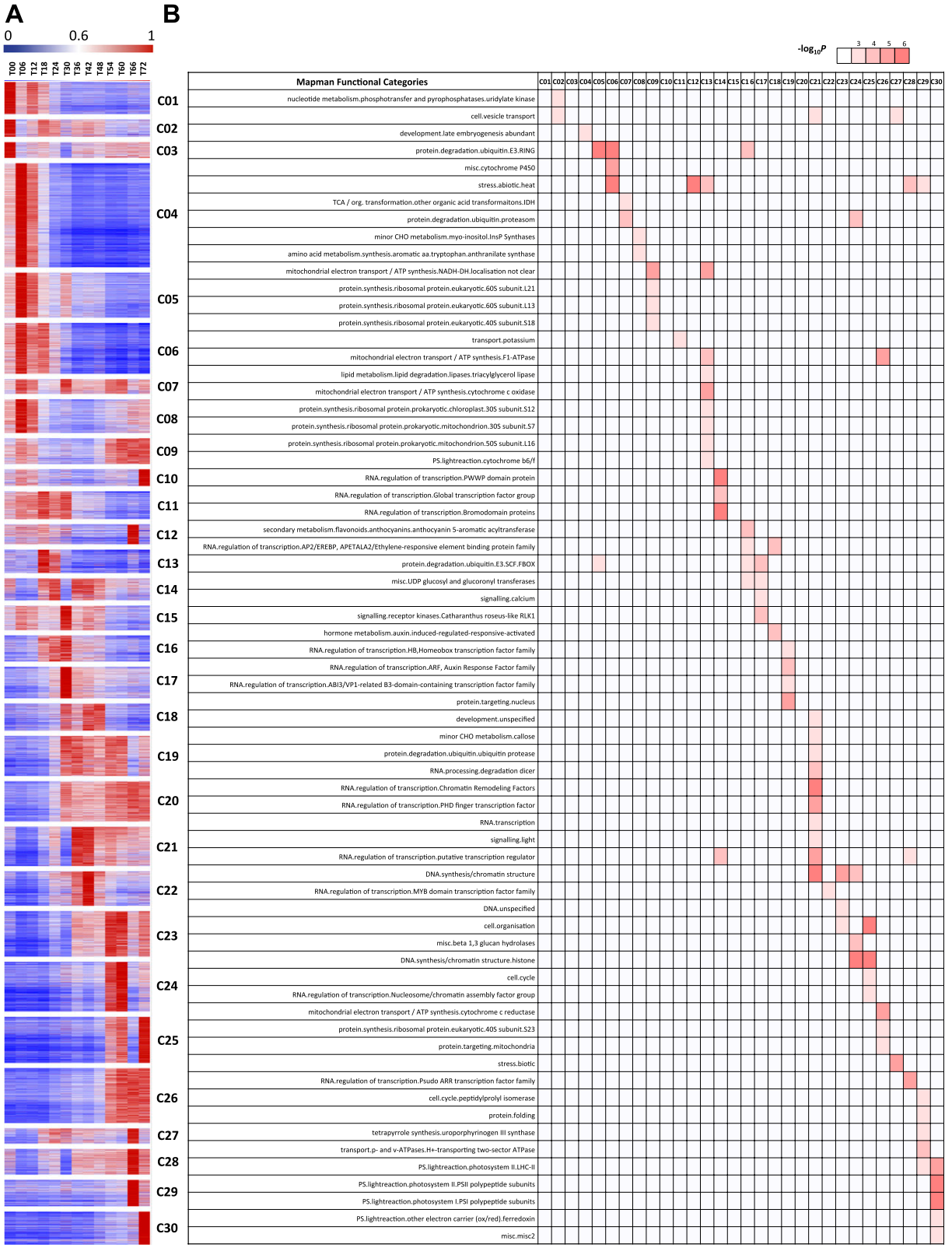


**Fig. S4.** (Continued)

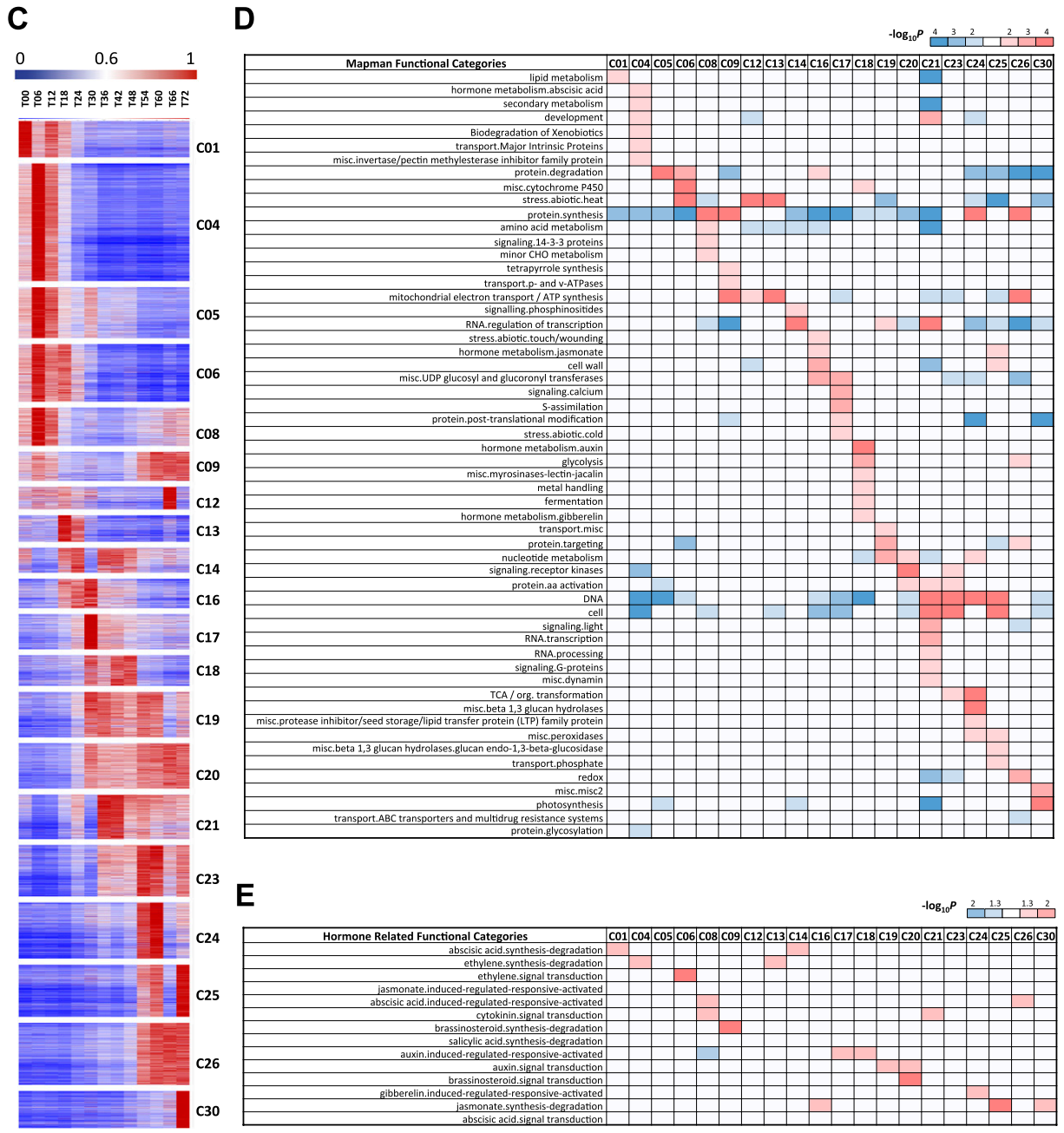


**Fig. S4.** MapMan functional categories enriched in genes of various coexpression gene clusters. (*A*) Expression profiles of genes sorted according to cluster orders. (*B*) Enrichment tests of detailed overrepresented functional categories on each cluster as shown in *A*. (*C*) Expression profiles of genes in selected clusters. (*D*) Enrichment tests of both over- and underrepresented high-level categories on each cluster as shown in *C*. (*E*) Enrichment tests of both over- and underrepresented hormone-related genes on each cluster as shown in *C*.

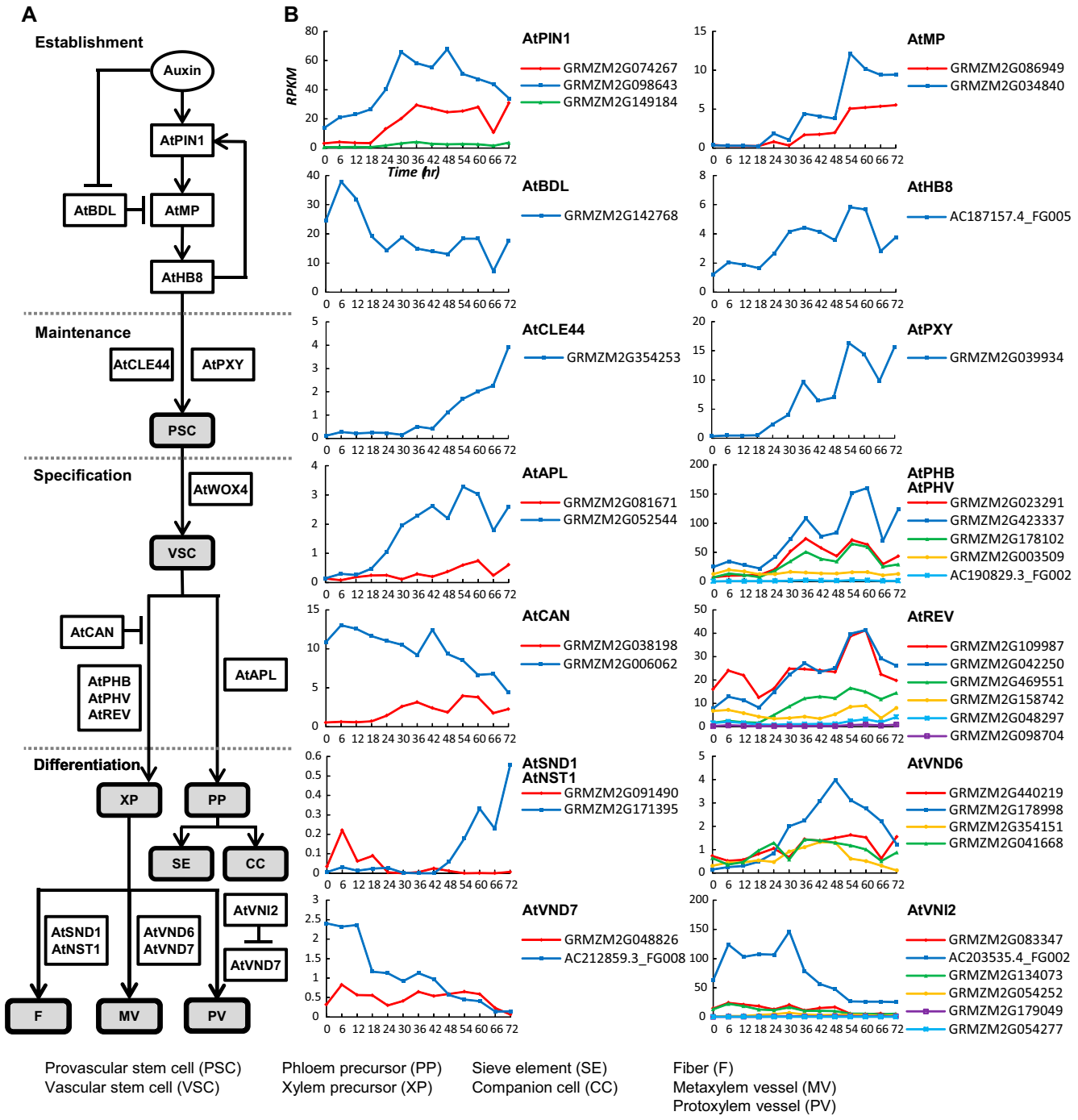


**Fig. S5.** Maize transcription factors likely regulating vascular development. (*A*) The current known components and relationships of vascular development genes in *Arabidopsis thaliana*. (*B*) Expression patterns of putative maize orthologs of *A. thaliana* vascular development genes.

**Table S1. Read count statistics of the deep sequencing data for the 13 time points studied**

| | | Quality filtering | | Mappable read | |
|---|---|---|---|---|---|
| Sample | Library length, bp | Total reads | Rate, % | Total | Rate, % |
| T00_ST-DA01 | 300 | 218,477,277 | 75.9 | 277,488,186 | 76.6 |
| | 400 | 143,975,388 | 86.8 | | |
| T06_ST-DA02 | 300 | 213,618,797 | 74.8 | 255,413,163 | 71.7 |
| | 400 | 142,589,424 | 85.7 | | |
| T12_ST-DA03 | 300 | 206,947,451 | 74.6 | 275,874,436 | 75.9 |
| | 400 | 156,344,245 | 82.1 | | |
| T18_ST-DA04 | 300 | 210,270,688 | 76.5 | 279,889,765 | 76.4 |
| | 400 | 155,975,570 | 84.0 | | |
| T24_ST-DA05 | 300 | 190,425,691 | 78.2 | 294,733,426 | 78.8 |
| | 400 | 183,789,389 | 79.9 | | |
| T30_ST-DA06 | 300 | 195,697,940 | 80.2 | 271,280,993 | 77.9 |
| | 400 | 152,421,099 | 85.6 | | |
| T36_ST-DA07 | 300 | 215,565,717 | 79.9 | 308,395,082 | 80.0 |
| | 400 | 169,797,300 | 84.6 | | |
| T42_ST-DA08 | 300 | 153,005,724 | 85.2 | 242,223,816 | 72.2 |
| | 400 | 182,552,562 | 81.7 | | |
| T48_ST-DA09 | 300 | 156,382,336 | 83.7 | 240,264,229 | 73.1 |
| | 400 | 172,108,685 | 84.8 | | |
| T54_ST-DB05 | 300 | 222,228,743 | 76.8 | 340,039,005 | 81.2 |
| | 400 | 196,448,375 | 82.1 | | |
| T60_ST-DB06 | 300 | 236,437,704 | 76.3 | 351,089,370 | 80.6 |
| | 400 | 199,181,671 | 82.3 | | |
| T66_ST-DB07 | 300 | 237,623,498 | 74.6 | 339,580,019 | 75.6 |
| | 400 | 211,538,295 | 79.9 | | |
| T72_ST-DB08 | 300 | 208,071,129 | 80.7 | 320,303,598 | 75.3 |
| | 400 | 217,502,039 | 76.5 | | |

**Table S2. Numbers of annotated genes and expressed features**

| | | No. of expressed features | | | |
|---|---|---|---|---|---|
| Gene set | A/E | CDS | TF | Pseudo | TE |
| WGS | A | 63,363 | 2,070 | 17,344 | 29,024 |
| | E | 30,255 | 1,238 | 3,040 | 1,986 |
| FGS | A | 39,441 | 2068 | 95 | 53 |
| | E | 23,900 | 1,237 | 95 | 53 |

A, all; CDS, protein coding gene; E, expressed (RPKM ≥1 in more than one time point); FGS, filtered gene set; pseudo, pseudogene; TE, transposable element; TF, transcription factor; WGS, working gene set.

## Other Supporting Information Files

Dataset S1 (XLSX)
Dataset S2 (XLSX)