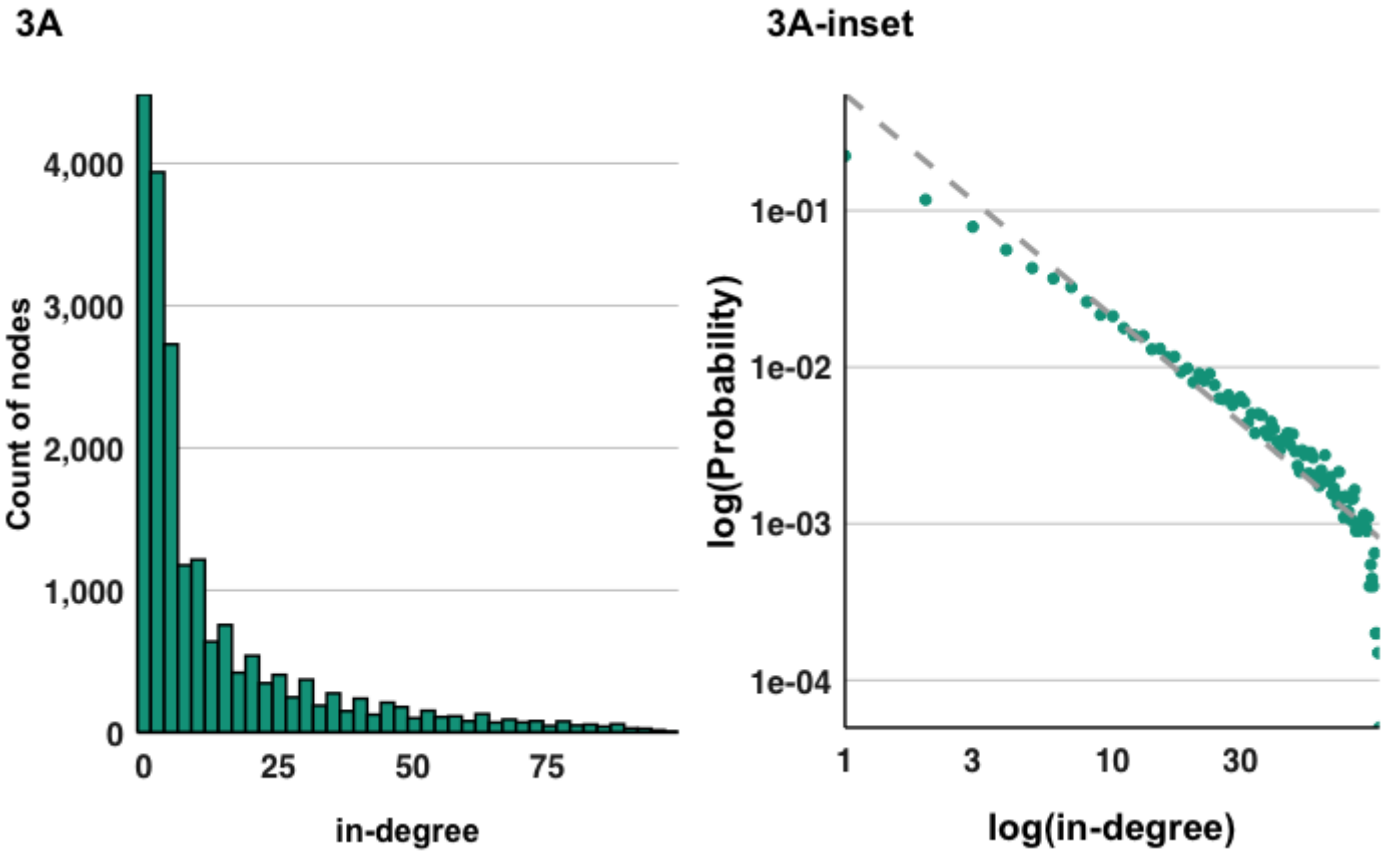
Regulatory Network Inference

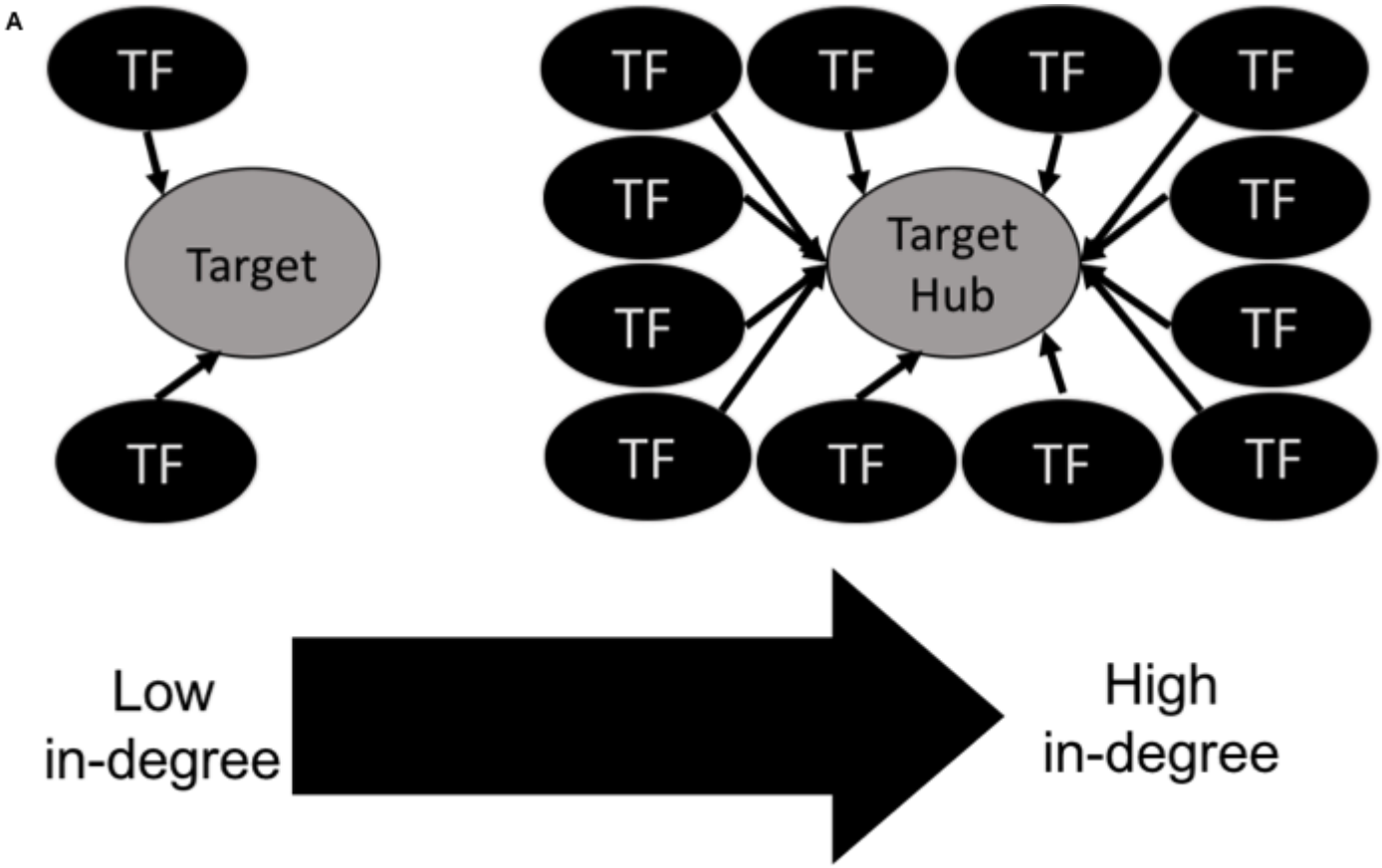
Code

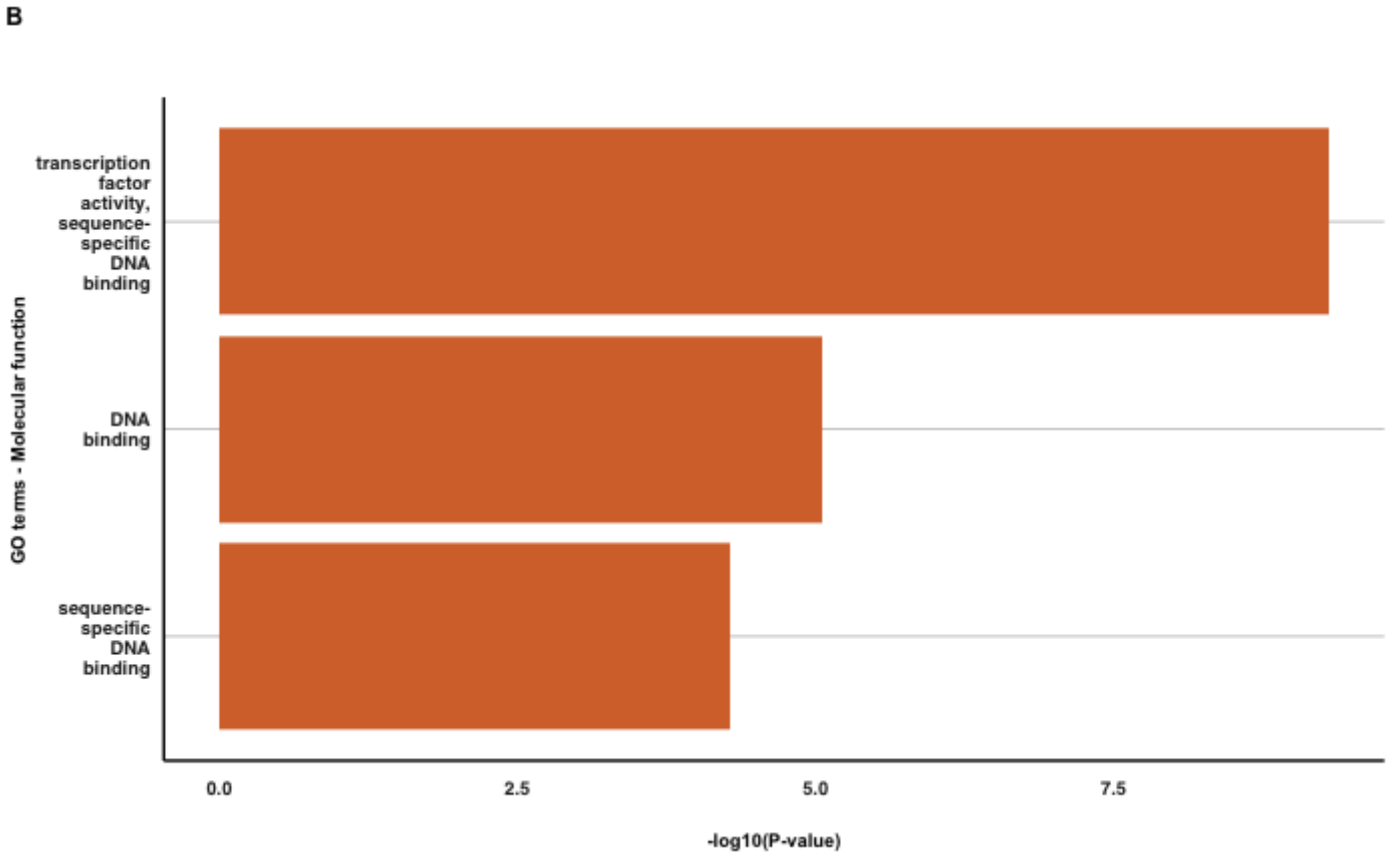
A common limiting factor for large-scale experiments is the signal to noise ratio. In the case of ChIP-seq, these often generate dubious sets of target genes, particularly when the interaction of the binding intensity is not considered. We hypothesize that a graph derived from true regulatory relationships should display topological features that set it apart from randomly connected ones. To test this, we reshaped the regulatory data into a graph to determine the feasibility of our data to pinpoint true regulatory relationships between TFs and target genes. For this we adopted the probabilistic framework used by the ENCODE project to identify high confidence proximal interactions (TIP, P-value < 0.05) 1 (http://doi.org/10.1093/bioinformatics/btr552). The resulting interactions render a graph with 20,179 nodes (including the 104 TFs) which encompass ~50% of the annotated genes in the maize genome.

The simplest statistic to describe the topology of a network is the distribution of the total number of connections for each node (i.e., the degree), which follows a Poisson distribution for random networks, and for real-world (e.g., biological) networks frequently approximates to a power-law 2 (http://doi.org/10.1126/science.286.5439.509). We evaluate the in-degree distribution (Fig 3A), or number of edges towards each node, and found it to be approximated to a power-law (R^2 = 0.882, P-value < 2e-16) (Fig 3A - inset), a landmark of scale-free networks 2 (http://doi.org/10.1126/science.286.5439.509).

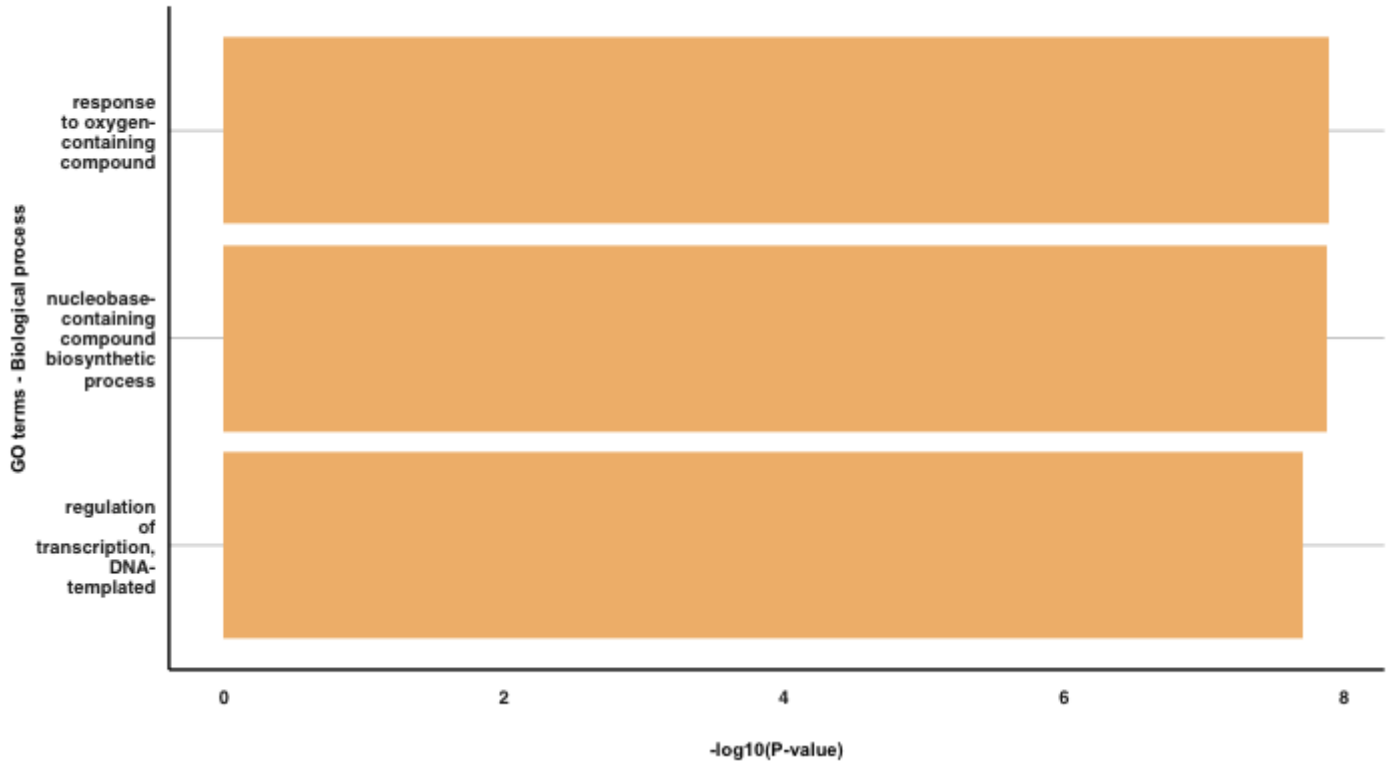


In scale-free topology, some nodes are considered critical for information flow (i.e., hubs), and appear more connected than others. We defined “hubs”, as nodes at the top percentile of the in-degree distribution (99th percentile), and obtained a set of 206 hubs-candidate genes (Supplementary figure 1A). A gene ontology (GO) analysis of the hubs-candidate genes showed statistical significant and strong enrichment for transcription regulation activity (molecular function, enrichment 4X; biological process, enrichment 4X) (Supplementary figure 1B-C), consistent with the role that hub nodes play in the transcriptional regulatory network.

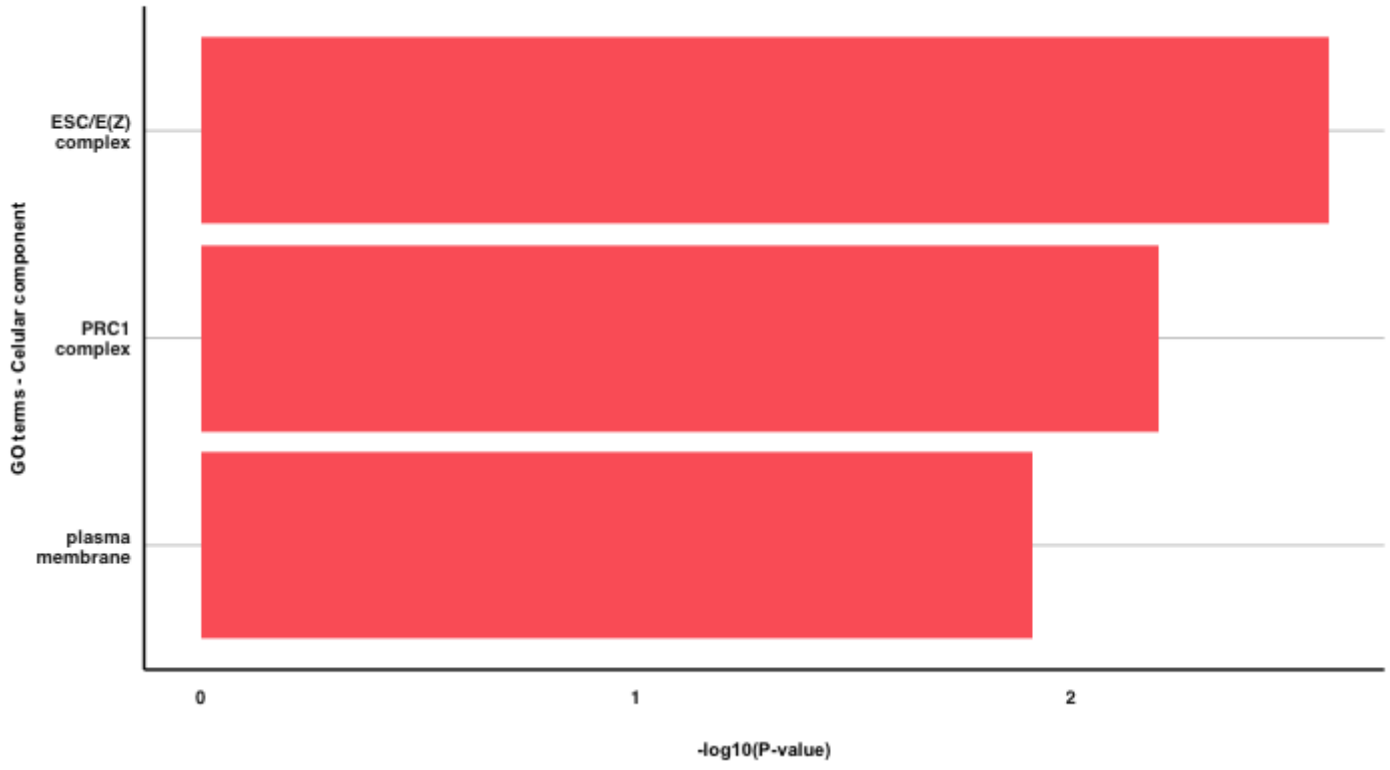

A
TF
Target
TF
TF
TF
TF
TF
TF
Target
Hub
TF
TF
TF
TF
TF
TF
TF
TF
Low
in-degree
High
in-degree


B
transcription factor activity, sequence-specific DNA binding
DNA binding
sequence-specific DNA binding
GO terms - Molecular function
0.0
2.5
5.0
7.5
-log10(P-value)

C

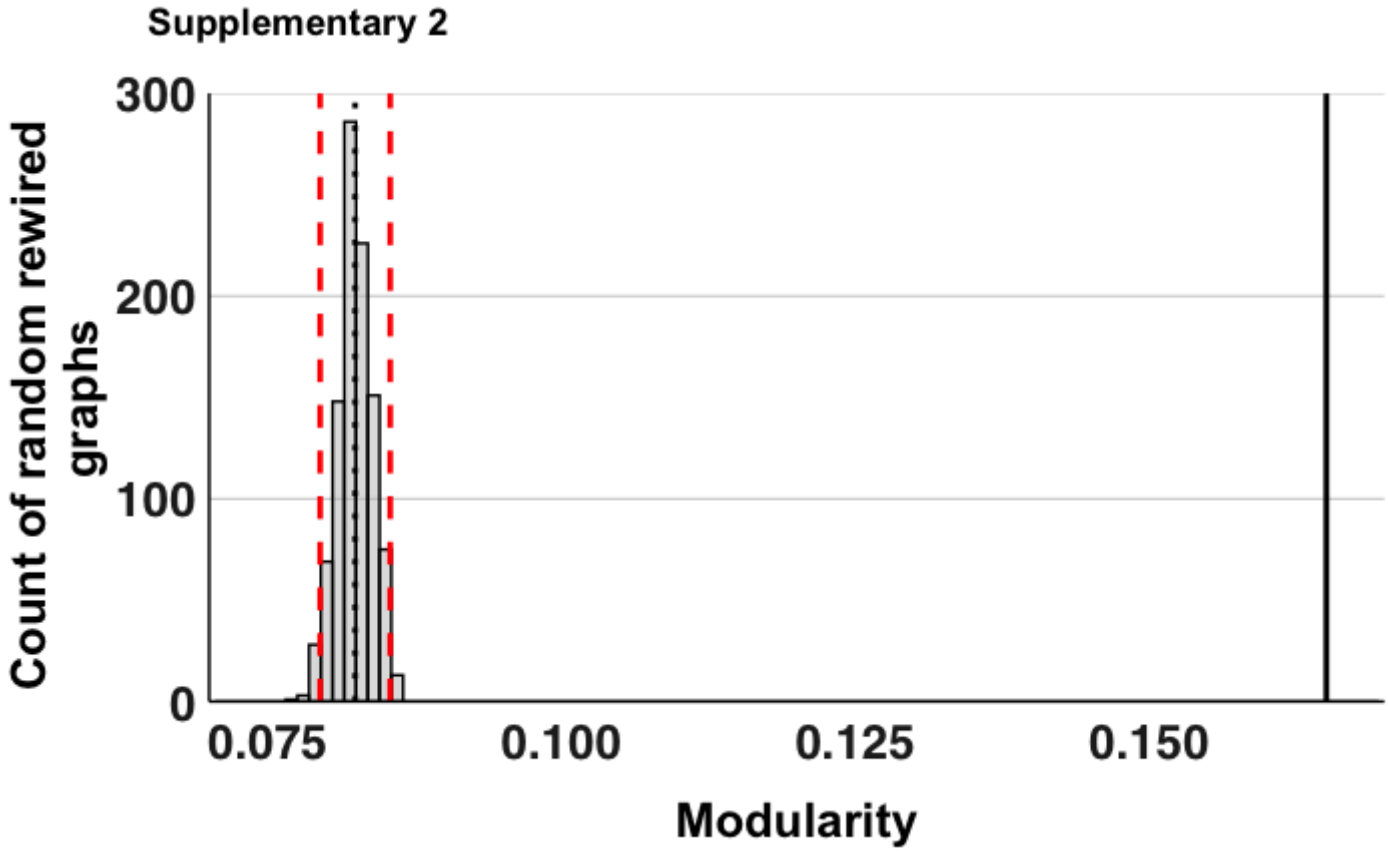


D

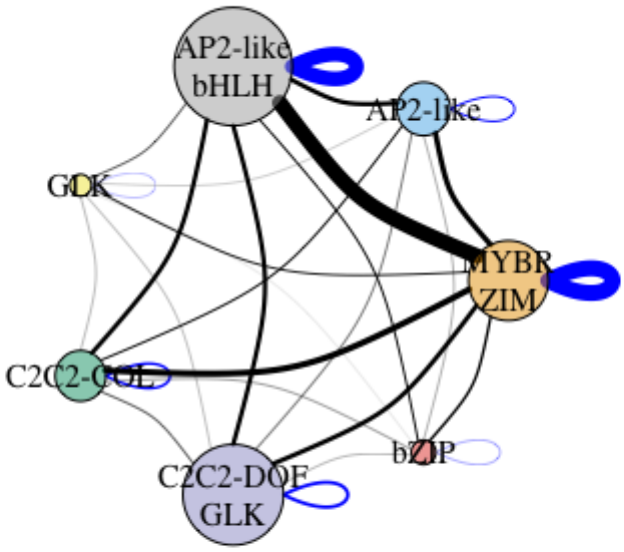


True biological networks, from food webs to protein-protein interaction networks, often exhibit topological and/or functional modularity *missing citations*.To determine topological modularity, we built a null distribution from an ensemble of random rewired graphs (H0: 1000 rewired graphs), while maintaining the number of nodes and number of edges per node, calculating for each a maximum modularity parameter 3

(http://doi.org/10.1103/PhysRevE.70.066111). This analysis shows statistically significant differences in modularity, which was large in in our graph (P-value < 0.05) versus randomly rewired ones (Supplementary figure 2).

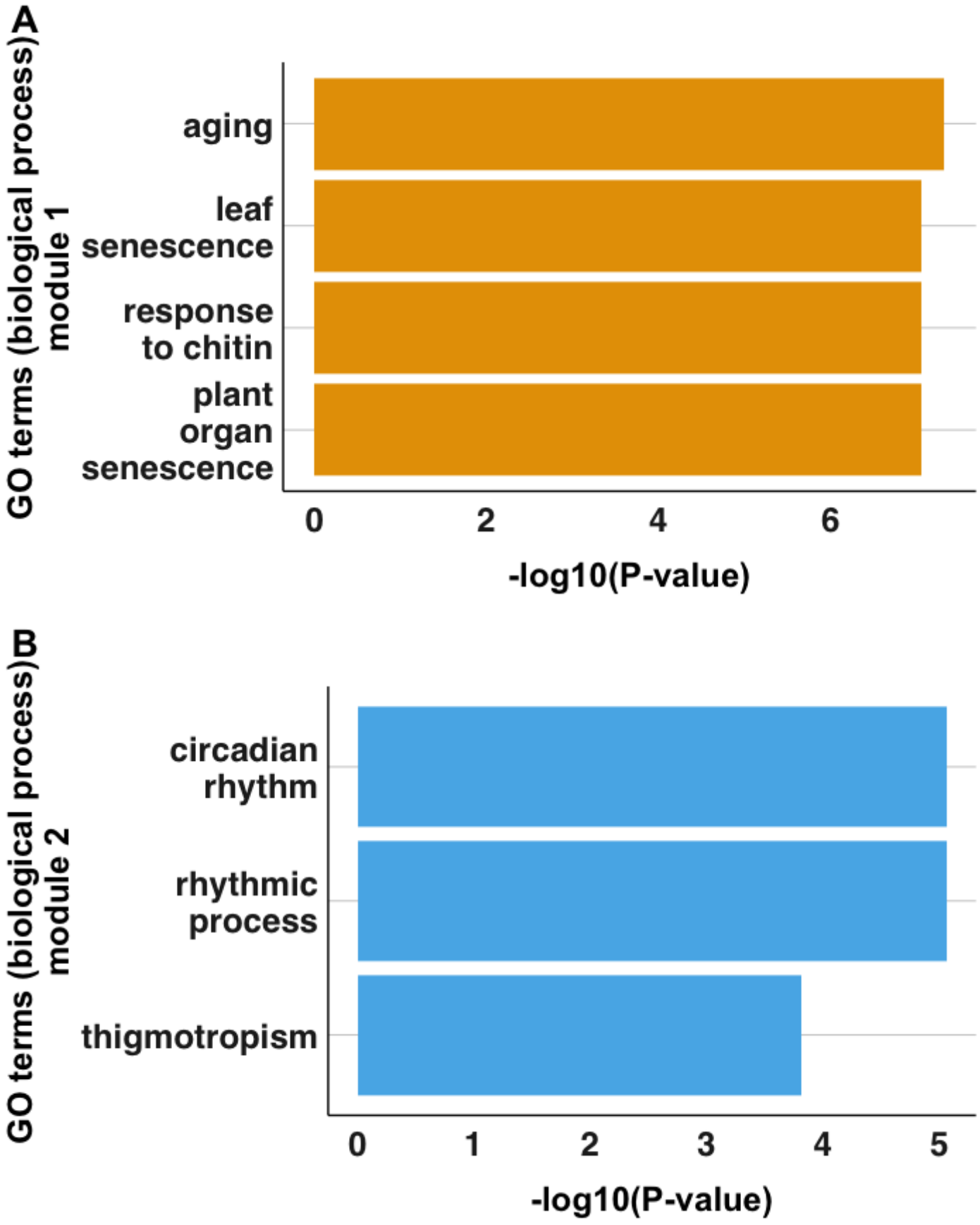


From a graph-theoretical point of view, we identified a total of seven modules, each containing from ~27% to ~5% of the total nodes (Fig 3B). Subgraphs containing only edges within the nodes from each module correspond only to ~40% of the edges, with ~60% of the edges among modules, which suggests a large information flow between modules (Fig 3B).

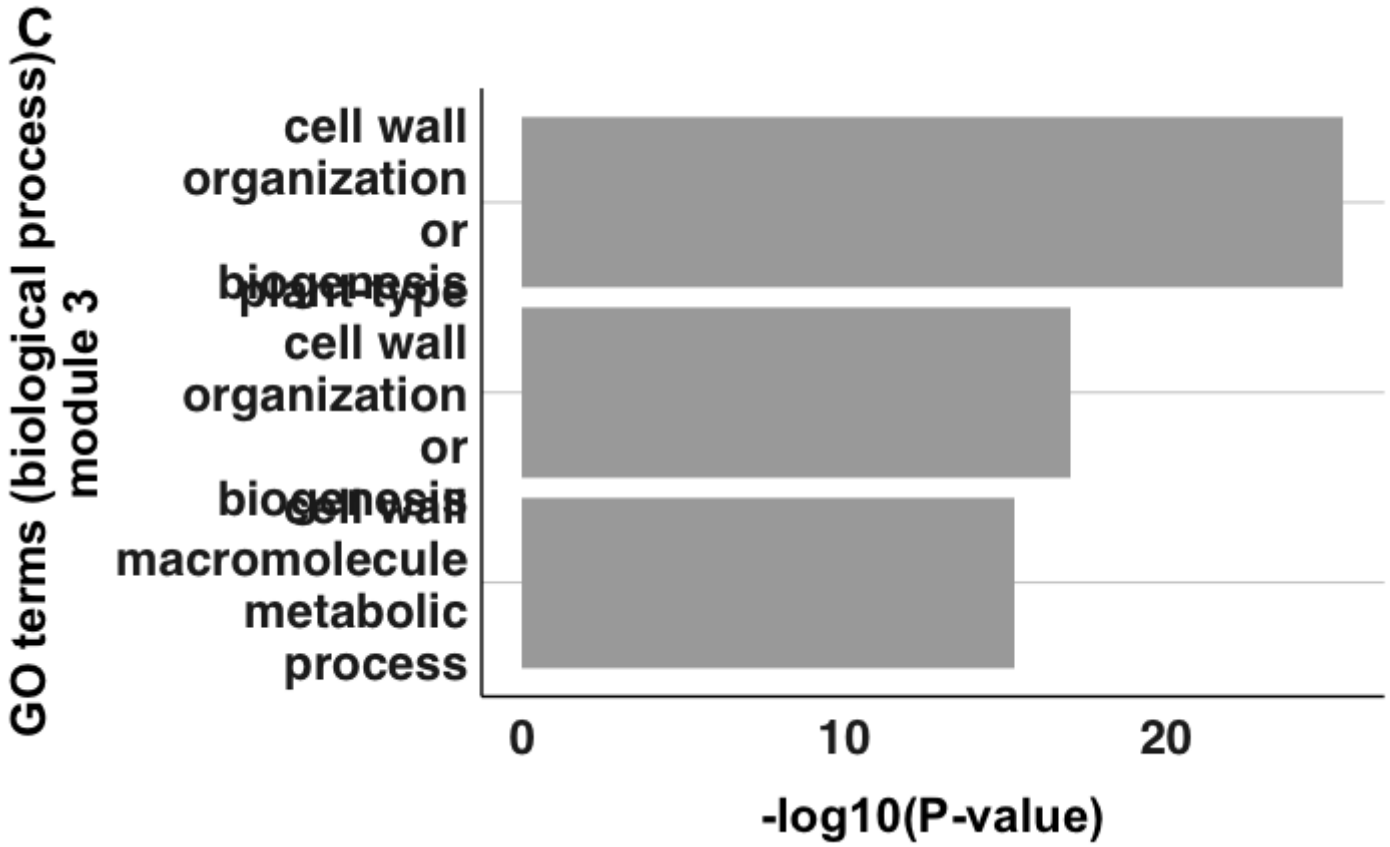


Next, we sought to determine if topological modularity could be related with functional modularity. For this, target genes present in each module were evaluated for enrichment of GO terms (biological processes). We found that all the modules were enriched for GO terms, with low overlap between them, suggesting functional modularity on top of the topological modularity (Supplementary figure 3).

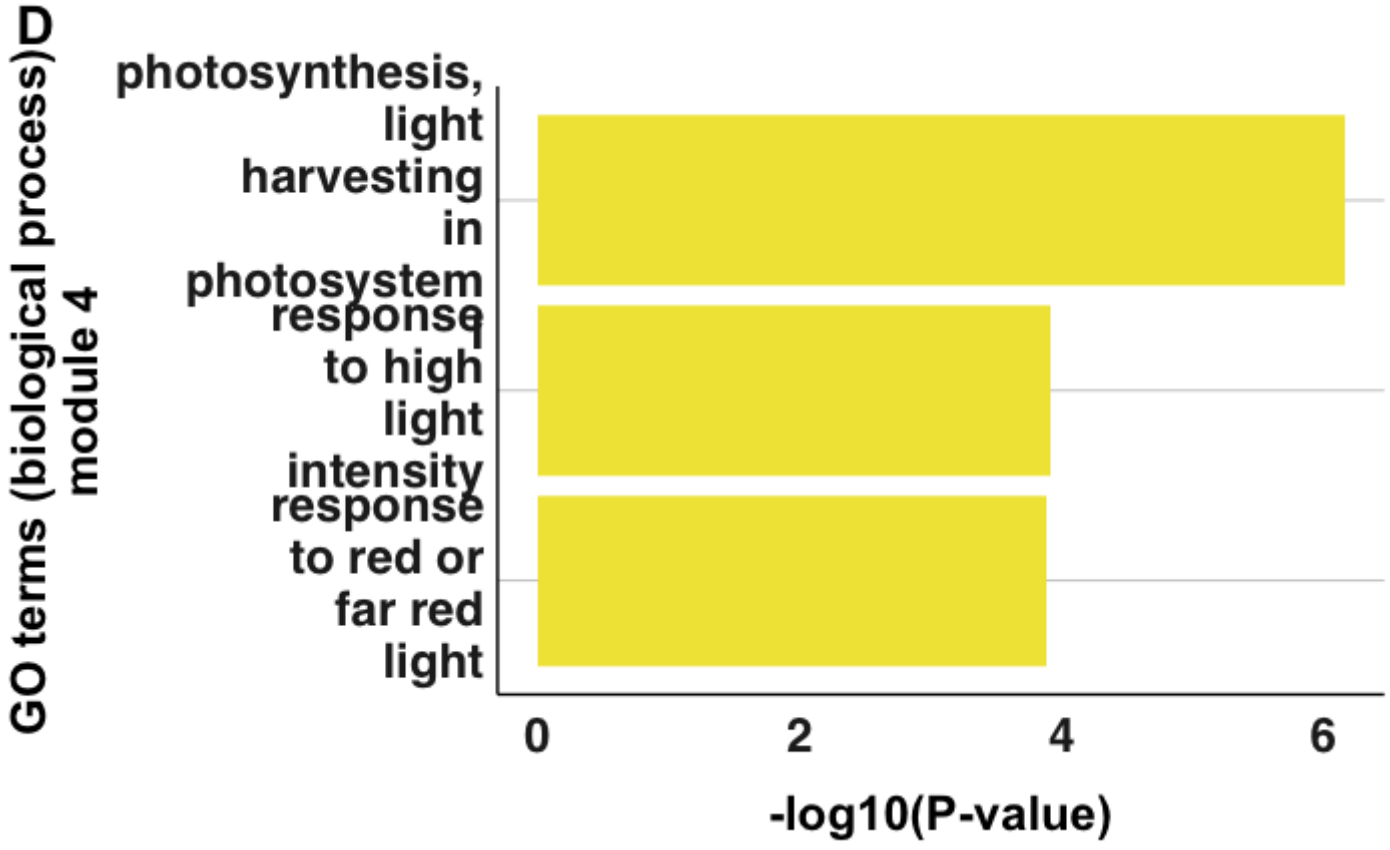
```
[1] 1
[1] 2
[1] 3
[1] 4
[1] 5
[1] 6
[1] 7
```

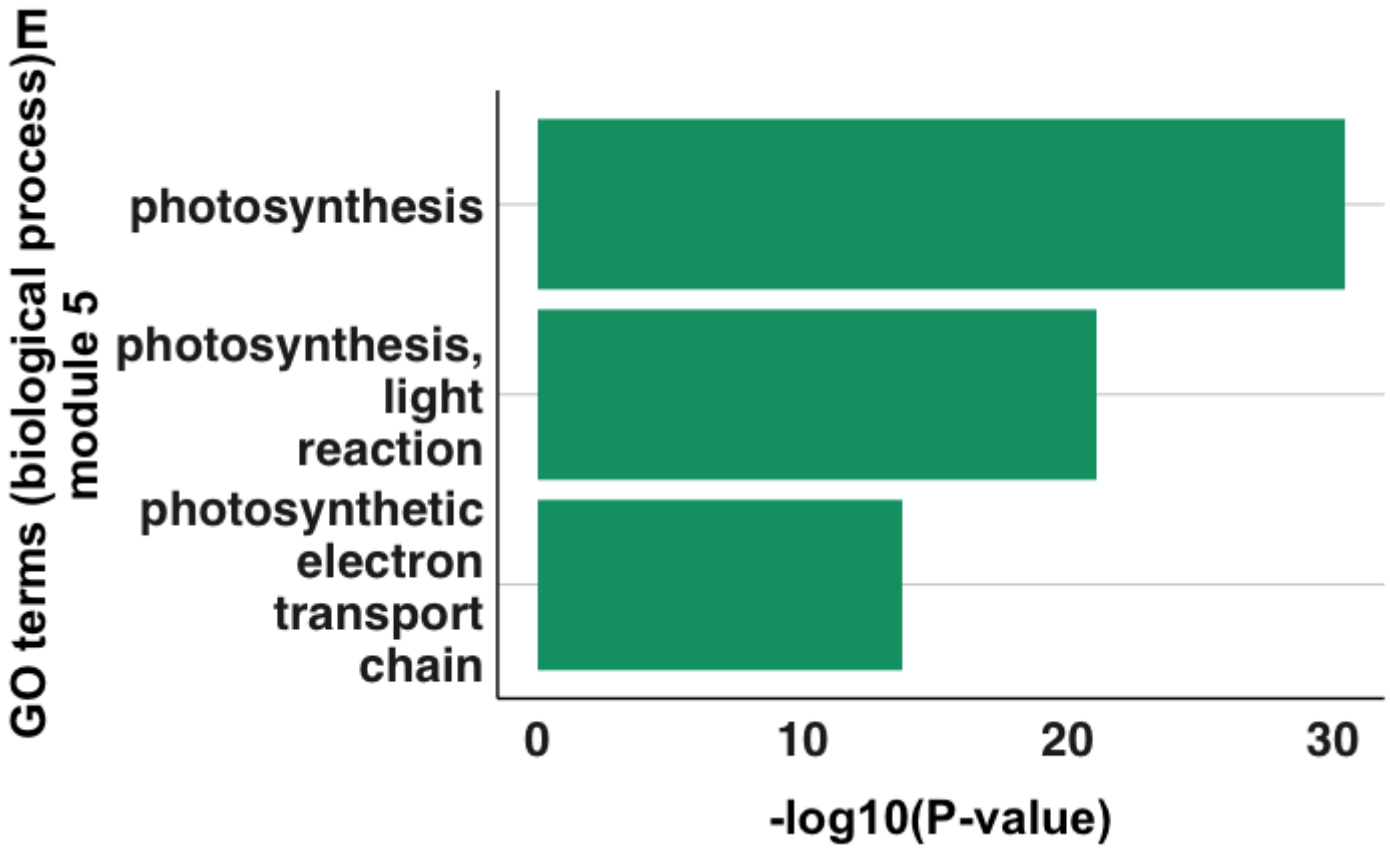

A
GO terms (biological process)
module 1
aging
leaf senescence
response to chitin
plant organ senescence
0
2
4
6
-log10(P-value)
B
GO terms (biological process)
module 2
circadian rhythm
rhythmic process
thigmotropism
0
1
2
3
4
5
-log10(P-value)

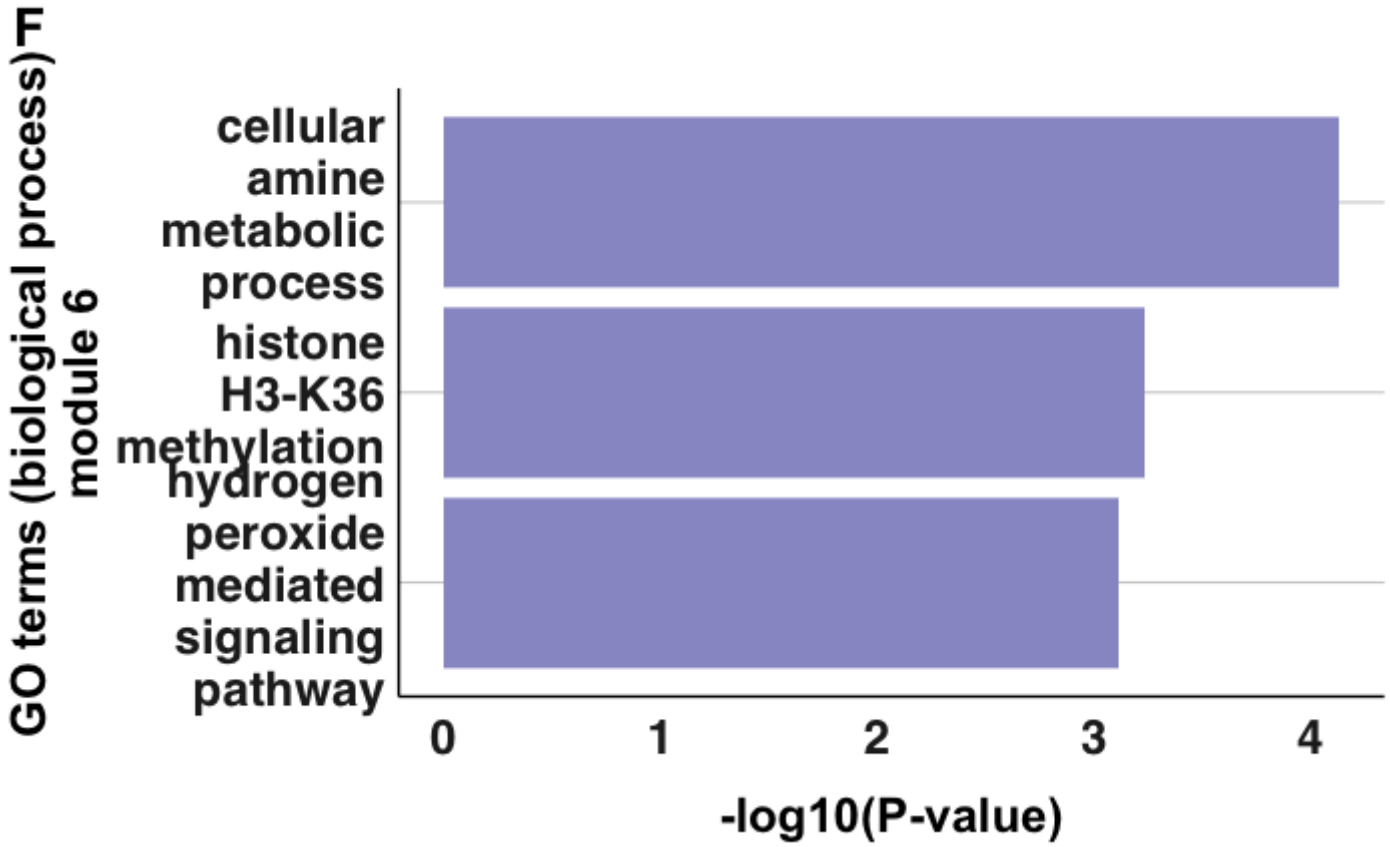
C

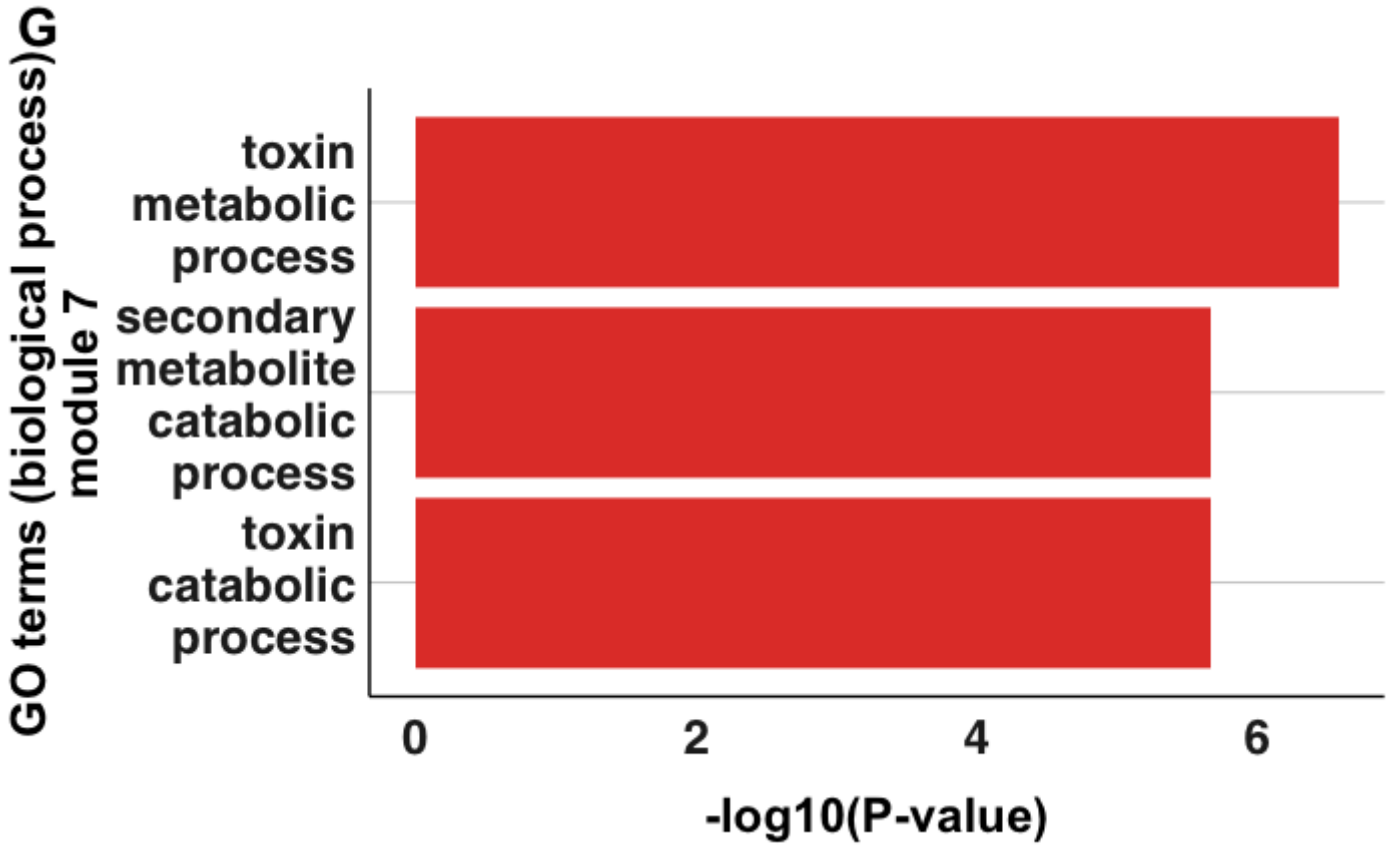

GO terms (biological process) module 3
cell wall organization or biogenesis
plant-type cell wall organization or biogenesis
cell wall macromolecule metabolic process
0
10
20
-log10(P-value)

D


GO terms (biological process) module 4
photosynthesis, light harvesting in photosystem I
response to high light intensity
response to red or far red light
0
2
4
6
-log10(P-value)


GO terms (biological process) module 5
photosynthesis
photosynthesis, light reaction
photosynthetic electron transport chain
0
10
20
30
-log10(P-value)


GO terms (biological process) module 6
cellular amine metabolic process
histone H3-K36 methylation
hydrogen peroxide mediated signaling pathway
0
1
2
3
4
-log10(P-value)



We found two photosynthesis-related modules. Module 4 mainly corresponds to genes regulated by two GLK transcription factors 4 (http://doi.org/10.1105/tpc.108.065250), while module 5 mainly correspond to target genes for CONSTANS(CO)-like TFs (Fig 3B). GLK and CO are known regulator of photosynthesis and related plant developmental processes (*missing citations*). Both, module 4 and 5 are enriched for “response to high light intensity”, and “response to low light intensity”, with module 4 be highly specialized and module 5 more generice (several GO terms appear enriched in addition to PS), which might suggest a split in the detailed mode of action of different TFs (Supplementary figure 3). Our evaluation of the topological features of the maize leaf regulatory graph not only served as a validation of the known TF to target gene interactions, but also provide new clues to understand those uncharacterized genes and TFs.

Hide

```
#References
#1. Cheng, C., Min, R., & Gerstein, M. (2011). TIP: A probabilistic method for identifyi
ng transcription factor target genes from ChIP-seq binding profiles. Bioinformatics (Oxf
ord, England), 27(23), 3221–3227.
#2. Barabási, A.-L., & Albert, R. (1999). Emergence of Scaling in Random Networks. Scien
ce, 286(5439), 509–512.
#3. Clauset, A., Newman, M. E. J., & Moore, C. (2004). Finding community structure in ve
ry large networks. Physical Review. E, Statistical, Nonlinear, and Soft Matter Physics,
 70(6 Pt 2), 066111.
#4. Waters, M. T., Wang, P., Korkaric, M., Capper, R. G., Saunders, N. J., & Langdale,
 J. A. (2009). GLK transcription factors coordinate expression of the photosynthetic app
aratus in Arabidopsis. The Plant Cell, 21(4), 1109–1128.
```