Transcription factors mediating regulation of photosynthesis

Wiebke Halpape[#,1,2], Donat Wulf[#,1,2], Bart Verwaaijen[1,2,6], Anna Sophie Stasche[1], Sanja Zenker[1,2], Janik Sielemann[1,2], Sebastian Tschikin[1,2], Prisca Viehöver[2,3], Manuel Sommer[4], Andreas P. M. Weber[4], Carolin Delker[5], Marion Eisenhut[1,2], Andrea Bräutigam[*,1,2]

[1]Computational Biology, Faculty for Biology, Bielefeld University
[2]Center of Biotechnology, Bielefeld University
[3]Genetics and Genomics of Plants, Faculty for Biology, Bielefeld University
[4]Institute of Plant Biochemistry, Heinrich Heine University Düsseldorf, Cluster of Excellence on Plant Science (CEPLAS), Germany
[5]Institute of Agricultural and Nutritional Sciences at Martin-Luther-University Halle-Wittenberg
[6]Department of Genetics, Martin-Luther-University Halle-Wittenberg

[#]contributed equally to the work, [*]to whom correspondence should be addressed

Contributions: WH analyzed photosynthetic gene expression, calculated the networks, identified and analyzed BBX overexpression lines, analyzed FAIR data, and edited the manuscript, DW provided the scripts for network calculation, calculated the enrichments, analyzed promoters, produced and analyzed MYBS overexpression lines, and edited the manuscript, BV developed and conducted the walk-trap analyses, SZ analyzed DAP-seq data, JS analyzed DAP-seq data, ST identified and analyzed PIF8 overexpression lines, PV conducted sequencing, AS analyzed FAIR data of known photosynthetic regulators, MS and APMW provided GLK mutant data, CD suggested the overexpression strategy for analyses, ME edited the manuscript, AB conceived of the study, analyzed data and wrote the manuscript

Abstract

Photosynthesis by which plants convert carbon dioxide to sugars using the energy of light is fundamental to life as it forms the basis of nearly all food chains. Surprisingly, our knowledge about its transcriptional regulation remains incomplete. Effort for its agricultural optimization have mostly focused on post-translational regulatory processes[1-3] but photosynthesis is regulated at the post-transcriptional[4] and the transcriptional level[5]. Stacked transcription factor mutations remain photosynthetically active[5,6] and additional transcription factors have been difficult to identify possibly due to redundancy[6] or lethality. Using a random forest decision tree-based machine learning approach for gene regulatory network calculation[7] we determined ranked candidate transcription factors and validated five out of five tested transcription factors as controlling photosynthesis *in vivo*. The detailed analyses of previously published and newly identified transcription factors suggest that photosynthesis is transcriptionally regulated in a partitioned, non-hierarchical, interlooped network.

Main

Photosynthesis, the process which generates virtually all calories for human consumption, is a very productive and at the same time very dangerous pathway[8]. The photosynthetic electron transfer chain with its photosystems (PSs) and its light harvesting complexes (LHCs) harvests the energy of photons (**Fig. 1A**). This energy needs to be near immediately consumed by the Calvin Benson Bassham cycle (CBBc), photorespiration (**Fig. 1A**), and other reactions to avoid overoxidation and radical production[8]. If the balance between harvest and consumption was only regulated at the post-transcriptional level[4], large potential savings in protein investment in this high protein investment pathway (**Fig. 1B, SupData1**[9]) were impossible to realize. Therefore, we hypothesized that both pathways are at least partially independently controlled at the transcriptional level.

The light harvesting module of photosynthesis is at least partially controlled by transcription factors (TFs), which also control germination. Photosynthetic genes are transcriptionally activated by ELONGATED HYPOCOTYL 5 (HY5) [10] and repressed by PHYTOCHROME INTERACTING PROTEINS (PIFs) 1, 3, 4, and 5 as evidenced by the *pifq* mutant[11]. Neither mutant was initially identified for their effects on photosynthetic gene expression. Frequently depicted downstream of these regulators[5], the GOLDEN2-LIKE (GLK) 1 and 2[6,12] and the GATA-type TFs GATA, NITRATE-INDUCIBLE, CARBON METABOLISM-INVOLVED (GNC) [13] and GNC-like[6,14] of which only the original GOLDEN2 mutation in maize was characterized as a photosynthetic TF[15]. We hypothesized that the identification and characterization of additional photosynthetic regulators will reveal different modes of regulation of photosynthetic modules and will reveal if regulation is hierarchical or more complex.

Figure 1

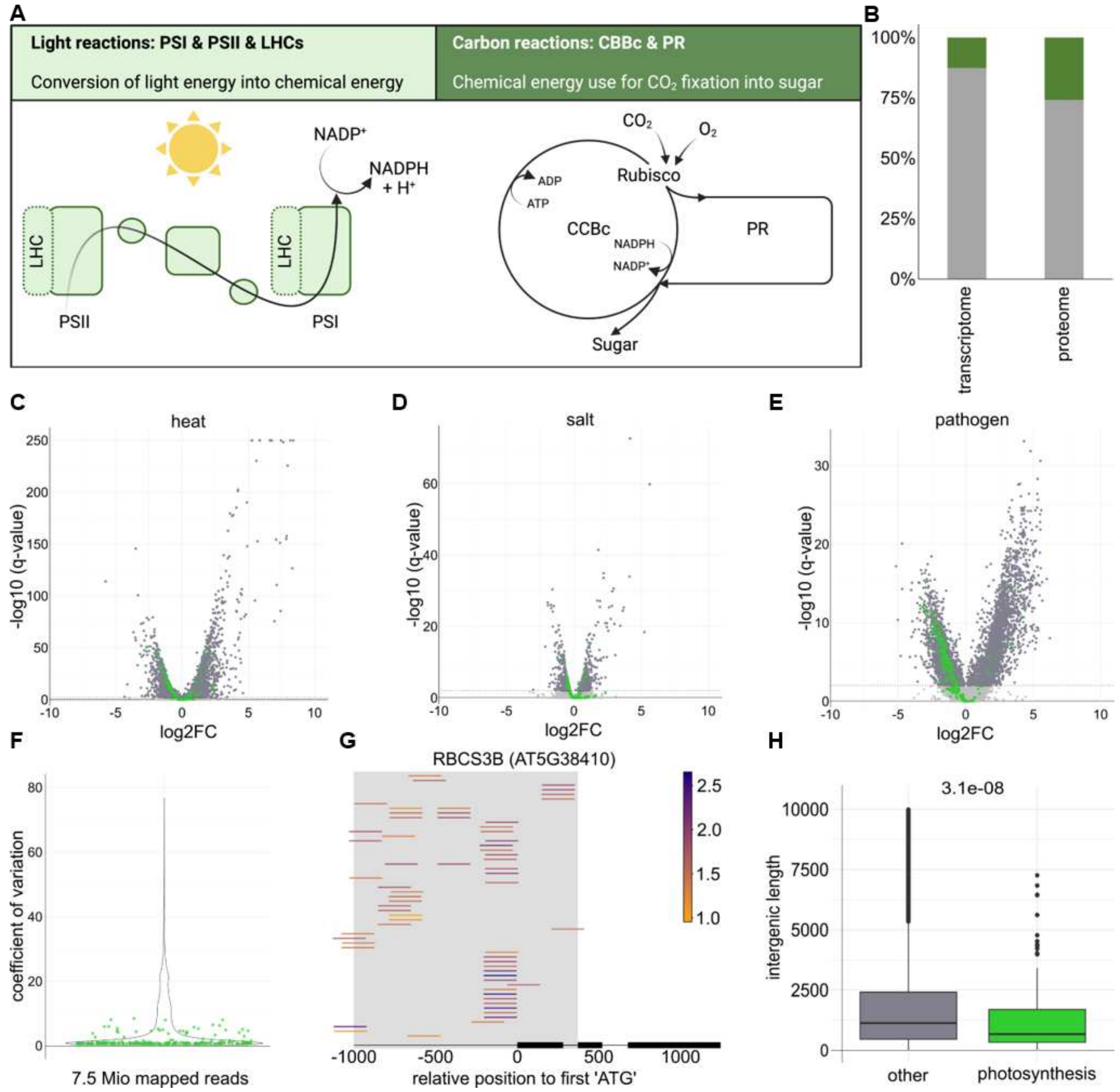


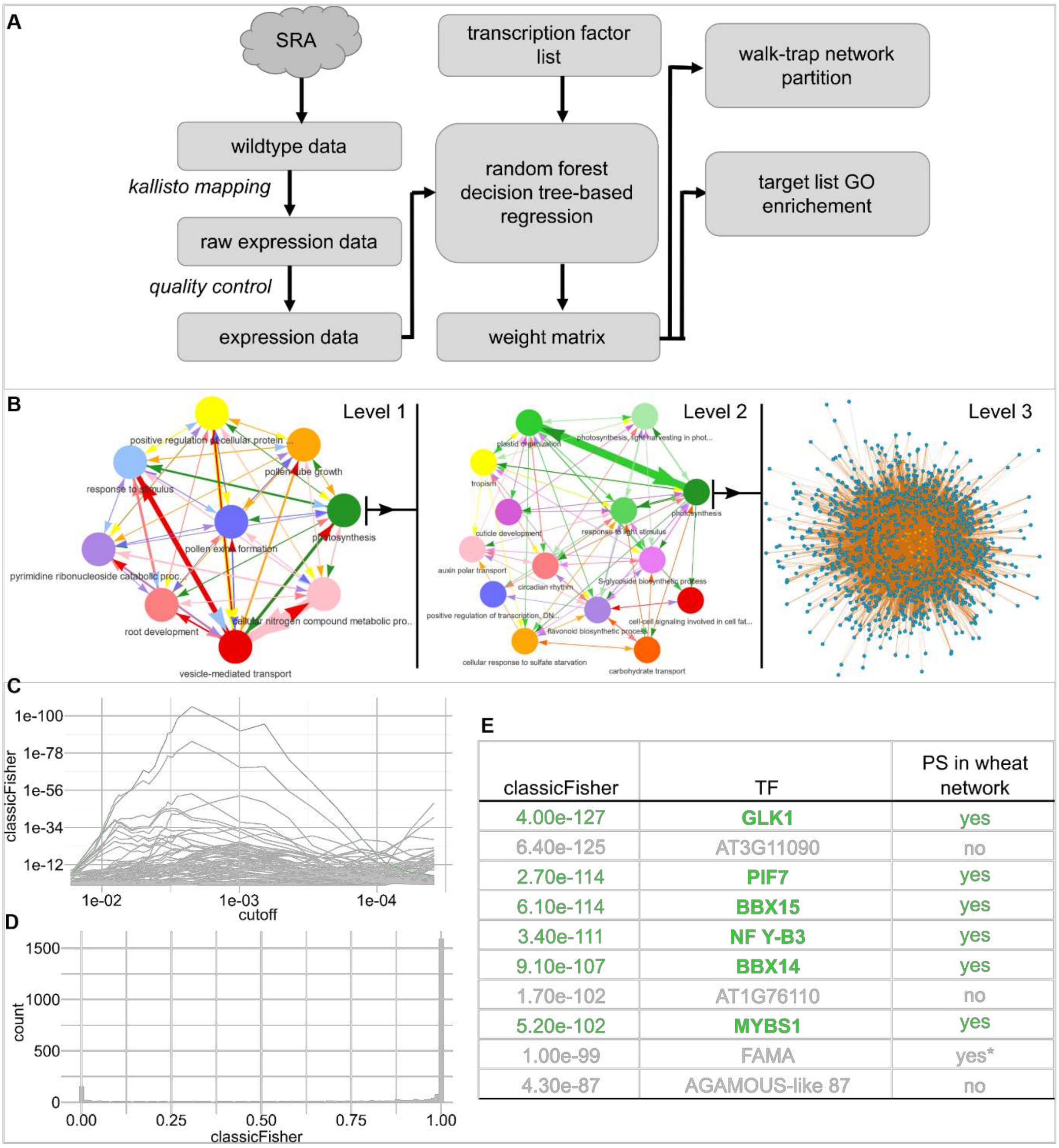
Figure 1: Photosynthetic gene expression is variable (A) schematic representation of photosynthesis; (B) proportion of *Arabidopsis thaliana* proteome and transcriptome invested in photosynthesis (green); (C-E) volcano plots of transcript abundance changes with photosynthetic genes in green; (F) coefficient of variance for all transcripts in 6,033 RNA-seq experiments, photosynthetic genes in green; (G) experimentally validated binding events on the RBCS3B promoter with relative DAP-seq peak height coded in color; (E) promoter lengths of all promoters vs. photosynthetic promoters (Wilcoxon Rank test, p<1e-7)

Photosynthetic transcript abundance is variable

To understand the regulatory processes related to photosynthesis, we started an in depth analysis of photosynthetic genes and their abundance variation. In mature leaves, photosynthetic transcripts significantly changed in abundance in response to stresses[16] with changes up to 10-fold (**Fig. 1C to E**) and were among those with the highest absolute changes (**SupFig.1**). Photosynthesis is indeed also controlled at the transcriptional level. To test overall variation, 7,468 RNA-seq experiments in wild type *Arabidopsis thaliana* plants were compiled (**SupData 2**) and 6,033 RNA-seq datasets with at least 7.5 million mapped reads per samples re-analyzed (**SupData 3**). The coefficient of variation for all transcripts was plotted with photosynthetic transcripts highlighted (**Fig. 1F**). Compared to all transcripts, photosynthetic transcripts varied close to the average of all transcripts and were not among those with a large coefficient of variation (**Fig. 1F**). To compile a list of candidate regulators, protein::DNA binding for 217 TFs was analyzed. From 0 to 118 (median = 57) regulators bind to photosynthetic promoters (**SupData 4**, analyzed from[17]). The RbcS promoter, for example (**Fig.1G**), is bound by 57 TFs with many binding to the same sites indicating highly complex regulation, likely with redundant contributions, and a necessity for identifying those TFs with a large contribution to regulation. A detailed analysis demonstrated that photosynthetic promoters were significantly shorter compared to all promoters ($p<10^{-7}$, Wilcoxon Sum Rank Test, **Fig. 1H**) with some as short as 250 bp.

We hypothesized that the variation in photosynthetic transcript abundances (**Fig. 1C-F**) opens the door to RNA-based gene regulatory networks (GRNs) of photosynthetic gene expression. This approach allows to test all TFs for their contribution. Co-expression and linear models have previously been pursued for network construction[18]. We hypothesized that unlike linear approaches, random forest regression decision trees (RF) [19] will construct a GRN that includes non-linear and combinatorial interactions between TFs and targets genes and captures direct and indirect interactions. 2,399 TFs were obtained from published sources[20], curated, and used as regulators for a wrapped GENIE3[19] algorithm**.** For validation, biology-dependent evaluation workflows were developed for RF based GRNs (**Fig. 2A**).

Figure 2



classicFisher	TF	PS in wheat network
4.00e-127	GLK1	yes
6.40e-125	AT3G11090	no
2.70e-114	PIF7	yes
6.10e-114	BBX15	yes
3.40e-111	NF Y-B3	yes
9.10e-107	BBX14	yes
1.70e-102	AT1G76110	no
5.20e-102	MYBS1	yes
1.00e-99	FAMA	yes*
4.30e-87	AGAMOUS-like 87	no

Figure 2: Gene regulatory network based identification of candidate TFs for photosynthesis (A) scheme of the workflow; (B) iterative walk-trap based partitioning of the RF output and hairball plot of the photosynthetic sub-subcommunity; (C) GO term enrichment for all biological process GO terms for the transcription factor GLK1 at different weight cut-offs in the network; (D) p-value histogram for enrichment of 2,399 TFs in GO:0015979 "photosynthesis"; (E) overlap of ranked list with wheat GRN[21]

Gene regulatory networks suggest candidates for photosynthetic regulation

The algorithm outputs dimensionless numbers, called weights. These represent the contribution of each TF to abundance values of each target transcript. At a weight cut-off of 0.005, 2,398 of 2,399 TFs were assigned at least one target gene and 33,105 of 33,599 target genes were assigned at least one candidate regulating TF (**Sup.Fig. 2).** The initial publication of GENIE3[19] and subsequent users[22] identified a high false positive rate (up to a 80%[19]) at full recall when single edges connecting targets to TFs were considered. This indicates a substantial amount of noise in the data. To test if the communities defined by the GENIE3 output contained biological information, a series of random walks[23] was conducted on the data (**Fig. 2A, B**). The initial walk-trap detected nine communities of which each enriched for particular gene ontology (GO) terms[24,25] (**Fig. 2B**). The photosynthesis community was extracted and re-partitioned into 14 subcommunities, which included photosynthesis *per se*, plastid organization, and response to light stimulus enriching subcommunities (**Fig. 2B**). Photosynthesis and plastid organization partitioned into different sub-communities indicating that both processes are at least partially independently regulated, and that the RF based GRN separates both processes (**Fig. 2B**). Plotting the complete network of the photosynthesis subcommunity resulted in 55 candidate TFs (**Fig. 2B**). We tested the community for known photosynthetic regulators and identified GLK1[6,12], GLK2[6,12], GNC[6,14], and GNC-like[6,14] (**SuppData 5**). The GO term enrichments demonstrate that the weight matrix constructed by GENIE3 contained biological information (**Fig. 2B**). The identification of known photosynthetic regulators indicated that regulators of enriched pathways co-localize in communities with their targets. To facilitate analysis of pathways other than photosynthesis, the complete community information is provided as **SuppData 6**. We tested the target genes of each TF for enrichment with a sliding weight cut-off (**Fig. 2A, C**) to obtain a ranked list of photosynthesis regulation candidates to prioritize further validation. We hypothesized that by testing for enrichment of genes in a particular GO term, noise is filtered, and the TF assigned a putative function based on the GO term (**Fig. 2A**). This analysis overcomes the expected high noise level[19] as the importance of single edges is diminished in favor of significant biological relevance based on many edges. The known photosynthetic transcription factor GLK1[12,26] was used to test the method (**Fig. 2C**). At different weight cut-offs all GO terms in biological processes were tested for enrichment. At a cut-off of 0.005, the GO:0015979 term "photosynthesis" (**SuppData 7)** enriched with a minimal p-value <1e-100. Other GO terms semantically related to photosynthesis also enriched significantly (**Fig. 2B**). A second peak of enrichment at a lower weight cut-off was observed for many TFs, which contained mostly general GO terms with large membership (**SuppData 8**). All TFs in the dataset were tested for enrichment in photosynthesis at a weight cut-off of 0.005. Enrichment p-values ranged from 7.4E-138 to 1 in a typical significance pattern (**Fig. 2D**) resulting in 95 candidates for regulation of photosynthesis. To select the candidates to enter analysis for nuclear photosynthetic regulation, the list was manually curated from the top. Plastid targeted gene products were excluded from validation analysis as were ZP1 known to function in root hairs[27], DOT3 known to affect venation patterning[28], FAMA known to only be expressed in the epidermis[29], and NCH1 known to interact with phototropins[30] (**SuppData 10**). The results of a wheat GRN[21] were overlayed with the remaining top candidates in the Arabidopsis GRN (**Fig.2E**). Six candidates identified as photosynthetic in both wheat and Arabidopsis GRNs were pursued further. The top candidate is again GLK1 (**Fig. 2E**), a known photosynthetic regulator reported to control light harvesting complexes[12,26]. PIF7 is a member of the phytochrome interacting factors[11], which include the photosynthetic repressors PIF1, PIF3, PIF4 and PIF5[11]. However, PIF7 is not degraded upon light exposure[11]. The B-BOX transcription factors BBX14 and BBX15 are closely related proteins of the BBX family[31], which also include CONSTANS[31] and BBX proteins known to interact with HY5[32]. Nuclear factor Y-B3 has been reported as being involved in heat tolerance[33] and MYBS1 was reported to mediate sugar signaling[34].

Figure 3

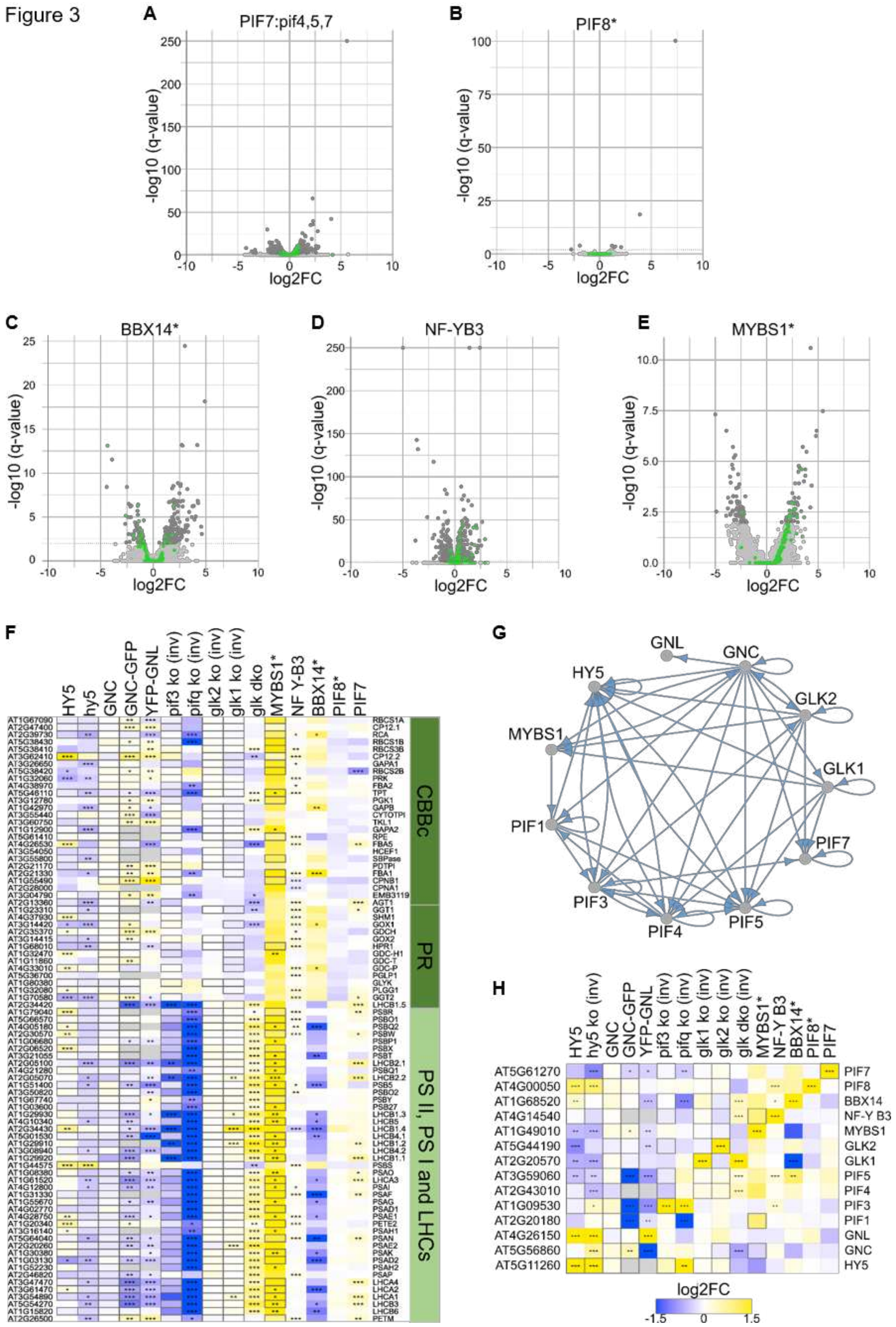


Figure 3: Characterization of photosynthetic TF candidates (A-E) volcano plots of differential transcript abundance with photosynthetic genes in green for candidate TFs; (F) heat map of transcript abundance changes in photosynthetic genes with significant changes denoted by stars and experimentally validated binding denoted by boxes; starred gene names indicate induced overexpression, mutant data was inverted to homogenize the visualization; (G) diagram of promoter binding data for known photosynthetic TFs; (H) as in F, transcription factors

New photosynthetic transcription factors show partitioned, interlooped regulation

We hypothesized that photosynthetic gene expression is deregulated in plants with changes in expression for their controlling TF. For all candidate photosynthesis regulators, expression data of complementants, knock-out mutants, or overexpression lines were obtained through FAIR data (**Fig. 3A and D**) or RNA-seq data was produced for newly established inducible overexpression lines (**Fig. 3B, C, and E**). Volcano plots showed that photosynthetic genes were significantly deregulated in all lines tested except PIF8 indicating that these TFs indeed control photosynthetic gene expression (**Fig. 3A-E**). Photosynthetic genes were binned to the functions of photosystems and light harvesting complexes (LHCs), Calvin Benson Bassham cycle (CBBc), and photorespiration. Photosynthetic transcript abundance was plotted for new candidates (**Fig. 2E**) and for known photosynthetic regulators to identify shared and different patterns (**Fig. 3F**). Re-analysis of the PIF7 complementant in a *pif4,5,7* knock-out mutant background[35] showed PIF7 to be a transcriptional activator of the photosystems and LHCs (**Fig. 3F**) in stark contrast to the other PIFs, which strongly downregulate them[11] (**Fig. 3F**). This observation is in line with PIF7s observed role as an activator[36]. PIF8 showed similar patterns in photosynthetic transcript abundance without any significantly changed photosynthetic genes after 72 hours of induction (**Fig. 3F**). BBX14 decoupled the photosystems and LHCs from photorespiration and the CBBc with repression of the photosystems and LHCs, and induction of the CBBc and photorespiration (**Fig. 3F**) after 72 hours of induction. Permanent induction of either BBX14 or the closely related BBX15 resulted in near white plantlets (**Supp.Fig. 3**) as expected for plants in which the photosystems and LHCs are reduced (**Supp.Data 11**). BBX14 enables the plants to balance light harvesting and processing in the photosystems with consumption of the harvested energy in the CBBc and photorespiration (**Fig. 3F**). Both NF-Y B3 and MYBS1 function as general activators of photosynthesis as indicated by transcriptional effects observed in the constitutive overexpressor of NF-Y B3[33] and inducible overexpression of MYBS1 (**Fig. 3F**). The effect of permanent NF-Y B3 overexpression was rather small, likely since it requires interactors, such as a NF-Y C and a protein with a CCT domain for its action[37]. Permanent overexpression of NF-Y B3 in wheat supports a photosynthetic role[38]. MYBS1 binds the promoters of most of the significantly increased photosynthetic genes (**Fig. 3G**) pointing toward direct upregulation. The analyses clearly demonstrated that the list of photosynthetic TFs[5], which currently contains GLK1 and GLK2[6,12,26], GNC and GNL[6,14], PIF1,3,4 and 5[11], and HY5[10] needs to be amended with PIF7, BBX14 and 15, NF-Y B3 and MYBS1. To see the underlying regulatory network, we also reanalyzed DNA-binding data for HY5, MYBS1[17], PIF1, PIF3, PIF4, PIF5, PIF5, GLK1, GLK2, GNC and GNL (**Supp. Data 14**). The results demonstrated that the TFs are extensively cross-regulated (**Fig. 3G**). One major integrator of photosynthetic gene expression is the G-box (CACGTG), which is bound by all PIFs and HY5 (**Fig. 3G**). These TFs likely bind with different affinities depending on the particular DNA-shape of the G-box[39], and with different effects on photosynthetic transcript abundance (**Fig. 3F**). MYBS1-binding is enriched in photosynthetic promoters (Fishers Exact Test, p<1e-21, Fig. 3G) at GATAA[17]. It binds more than 50 photosynthetic genes (**Fig. 3F**) and three photosynthetic regulators (**Fig. 3H**). GLKs bind at GGATT[6,40] and GNC binds at GATC[6] and both binding sites are present and occupied in several other TF promoters (**Fig. 3F**) and in photosynthetic promoters (**Fig. 3G, SuppData 12**). Regulation of photosynthetic gene expression is controlled by a complex, non-hierarchical, interlooped network of TFs that frequently bind not only photosynthetic target gene promoters but also each other's promoters.

Figure 4

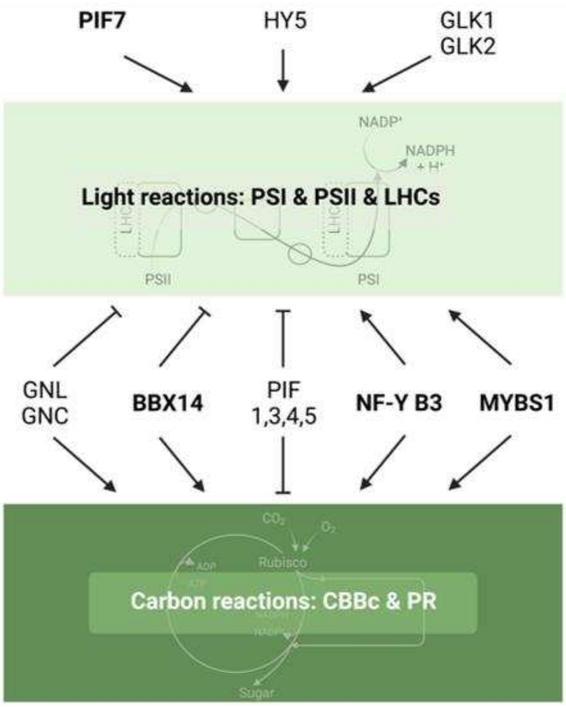


Figure 4: Scheme of photosynthetic gene regulation by a non-hierarchical, interloped network of TFs. The model bases on transcript abundance patterns in mutants and overexpression lines. Dashed lines indicate mixed regulation patterns as depicted in Fig 3G. Newly identified TFs are highlighted in bold.

Despite the highly interconnected regulators, the transcript abundance patterns of photosynthetic genes (**Fig. 3G**) allow classification of the roles for the previously known and newly identified TFs. The *glk1glk2* double mutant identifies GLKs as activators for abundance of photosystem and LHC transcripts (**Fig.3 G, Fig.4**). HY5 also mostly acts as an activator of those although HY5 data is reportedly highly variable between experiments[41]. PIF7 also activates photosystems and LHC genes. In contrast, both NF-Y B3 and MYBS1 activate photosynthetic gene expression throughout while the *pifq* mutant indicates that PIF1, 3, 4, and 5 act as repressors throughout. GNC and GNL partially decouple photosystem and LHC transcript abundance from CBBs and photorespiration transcript abundances while BBX14 fully decouples (**Fig.3 G, Fig.4**). The newly described TFs have sufficient regulatory capacity to up-regulate photosynthetic transcript abundance in general and to decouple the light from the carbon reactions (**Fig. 4**). In concert with the known TFs, general upregulation, general downregulation, upregulation of only light harvesting, and upregulation of only carbon fixation can be accomplished (**Fig. 4**).

GRNs calculated using RF, followed by GO term based post-processing, and multi-species comparison are an excellent tool for determining candidate TFs for a specific process (**Fig. 2B and C**, data for all processes available in **SuppTable4, SuppTable5, SuppData2**). The GRNs provided a ranked list of candidate TFs for the process of photosynthesis, all of which except PIF8 could be experimentally validated (**Fig. 3G**). The data demonstrates that photosynthesis is not regulated en bloc, but rather regulated in groups of genes belonging to photosystems and their LHCs (light reaction) and belonging to the CBBc and photorespiration (carbon reactions), respectively (**Fig. 3G, Fig. 4**). The data also demonstrates photosynthesis is regulated through a complex, non-hierarchical, interlooped network of TFs. This more complete list of TFs regulating the different sub-pathways of photosynthesis opens the door to targeted engineering and to a better understanding of the complex outcomes of genetic reconstructions[42].

1 Leister, D. Enhancing the light reactions of photosynthesis: Strategies, controversies, and perspectives. *Mol. Plant.* **16**, 4-22 (2023). https://doi.org:https://doi.org/10.1016/j.molp.2022.08.005

2 Long, S. P., Zhu, X. G., Naidu, S. L. & Ort, D. R. Can improvement in photosynthesis increase crop yields? *Plant Cell Environ.* **29**, 315-330 (2006). https://doi.org:10.1111/j.1365-3040.2005.01493.x

3 Zhu, X. G., Long, S. P. & Ort, D. R. in *Annu. Rev. Plant Biol.* Vol. 61 *Annual Review of Plant Biology* (eds S. Merchant, W. R. Briggs, & D. Ort) 235-261 (Annual Reviews, 2010).

4 Rochaix, J.-D. Regulation of photosynthetic electron transport. *Biochimica et Biophysica Acta (BBA) - Bioenergetics* **1807**, 375-383 (2011). https://doi.org:https://doi.org/10.1016/j.bbabio.2010.11.010

5 Cackett, L., Luginbuehl, L. H., Schreier, T. B., Lopez-Juez, E. & Hibberd, J. M. Chloroplast development in green plant tissues: the interplay between light, hormone, and transcriptional regulation. *New Phytol.* **233**, 2000-2016 (2022). https://doi.org:https://doi.org/10.1111/nph.17839

6 Zubo, Y. O. *et al.* Coordination of Chloroplast Development through the Action of the GNC and GLK Transcription Factor Families. *Plant Physiol.* **178**, 130-147 (2018). https://doi.org:10.1104/pp.18.00414

7 Breiman, L. Random Forests. *Machine Learning* **45**, 5-32 (2001). https://doi.org:10.1023/a:1010933404324

8 Oelze, M. L., Kandlbinder, A. & Dietz, K. J. Redox regulation and overreduction control in the photosynthesizing cell: complexity in redox regulatory networks. *Biochim Biophys Acta* **1780**, 1261-1272 (2008). https://doi.org:10.1016/j.bbagen.2008.03.015

9 Wang, M. *et al.* PaxDb, a database of protein abundance averages across all three domains of life. *Mol Cell Proteomics* **11**, 492-500 (2012). https://doi.org:10.1074/mcp.O111.014704

10 Toledo-Ortiz, G. *et al.* The HY5-PIF Regulatory Module Coordinates Light and Temperature Control of Photosynthetic Gene Transcription. *PLoS Genet.* **10**, e1004416 (2014). https://doi.org:10.1371/journal.pgen.1004416

11 Shin, J. *et al.* Phytochromes promote seedling light responses by inhibiting four negatively-acting phytochrome-interacting factors. *Proceedings of the National Academy of Sciences* **106**, 7660-7665 (2009). https://doi.org:doi:10.1073/pnas.0812219106

12 Waters, M. T. *et al.* GLK Transcription Factors Coordinate Expression of the Photosynthetic Apparatus in Arabidopsis. *Plant Cell* **21**, 1109-1128 (2009).

13 Bi, Y.-M. *et al.* Genetic analysis of Arabidopsis GATA transcription factor gene family reveals a nitrate-inducible member important for chlorophyll synthesis and glucose sensitivity. *The Plant Journal* **44**, 680-692 (2005). https://doi.org:https://doi.org/10.1111/j.1365-313X.2005.02568.x

14 Bastakis, E., Hedtke, B., Klermund, C., Grimm, B. & Schwechheimer, C. LLM-Domain B-GATA Transcription Factors Play Multifaceted Roles in Controlling Greening in Arabidopsis. *Plant Cell* **30**, 582-599 (2018). https://doi.org:10.1105/tpc.17.00947

15 Langdale, J. A. & Kidner, C. A. Bundle-Sheath Defective, A Mutation That Disrupts Cellular-Differentiation In Maize Leaves. *Development* **120**, 673-681 (1994).

16 Sewelam, N. *et al.* Molecular plant responses to combined abiotic stresses put a spotlight on unknown and abundant genes. *J. Exp. Bot.* **71**, 5098-5112 (2020). https://doi.org:10.1093/jxb/eraa250

17 O'Malley, R. C. *et al.* Cistrome and Epicistrome Features Shape the Regulatory DNA Landscape. *Cell* **165**, 1280-1292 (2016). https://doi.org:10.1016/j.cell.2016.04.038

18 Langfelder, P. & Horvath, S. WGCNA: an R package for weighted correlation network analysis. *Bmc Bioinformatics* **9** (2008). https://doi.org:559
10.1186/1471-2105-9-559

19 Huynh-Thu, V. A., Irrthum, A., Wehenkel, L. & Geurts, P. Inferring Regulatory Networks from Expression Data Using Tree-Based Methods. *PLoS ONE* **5** (2010). https://doi.org:e12776
10.1371/journal.pone.0012776

20 Perez-Rodriguez, P. *et al.* PlnTFDB: updated content and new features of the plant transcription factor database. *Nucl. Acids. Res.* **38**, D822-D827 (2010). https://doi.org:10.1093/nar/gkp805

21 Ramírez-González, R. H. *et al.* The transcriptional landscape of polyploid wheat. *Science* **361**, eaar6089 (2018). https://doi.org:10.1126/science.aar6089

22 Walley, J. W. *et al.* Integration of omic networks in a developmental atlas of maize. *Science* **353**, 814-818 (2016). https://doi.org:10.1126/science.aag1125

23 Xin, Y., Xie, Z.-Q. & Yang, J. An adaptive random walk sampling method on dynamic community detection. *Expert Systems with Applications* **58**, 10-19 (2016).

24 Swarbreck, D. *et al.* The Arabidopsis Information Resource (TAIR): gene structure and function annotation. *Nucl. Acids. Res.* **36**, D1009-D1014 (2008).

25 Alexa, A. & Rahnenfuhrer, J. topGO: topGO: Enrichment analysis for Gene Ontology. *R package version 2.22.0* (2010).

26 Waters, M. T., Moylan, E. C. & Langdale, J. A. GLK transcription factors regulate chloroplast development in a cell-autonomous manner. *Plant J.* **56**, 432-444 (2008).

27 Han, G. *et al.* Arabidopsis ZINC FINGER PROTEIN1 Acts Downstream of GL2 to Repress Root Hair Initiation and Elongation by Directly Suppressing bHLH Genes[OPEN]. *The Plant Cell* **32**, 206-225 (2019). https://doi.org:10.1105/tpc.19.00226

28 Petricka, J. J., Clay, N. K. & Nelson, T. M. Vein patterning screens and the defectively organized tributaries mutants in Arabidopsis thaliana. *The Plant Journal* **56**, 251-263 (2008). https://doi.org:https://doi.org/10.1111/j.1365-313X.2008.03595.x

29 Ohashi-Ito, K. & Bergmann, D. C. Arabidopsis FAMA Controls the Final Proliferation/Differentiation Switch during Stomatal Development. *The Plant Cell* **18**, 2493-2505 (2006). https://doi.org:10.1105/tpc.106.046136

30 Suetsugu, N. *et al.* RPT2/NCH1 subfamily of NPH3-like proteins is essential for the chloroplast accumulation response in land plants. *Proceedings of the National Academy of Sciences* **113**, 10424-10429 (2016). https://doi.org:doi:10.1073/pnas.1602151113

31 Gangappa, S. N. & Botto, J. F. The BBX family of plant transcription factors. *Trends Plant Sci.* **19**, 460-470 (2014). https://doi.org:https://doi.org/10.1016/j.tplants.2014.01.010

32 Bursch, K. *et al.* Identification of BBX proteins as rate-limiting cofactors of HY5. *Nature Plants* **6**, 921-928 (2020). https://doi.org:10.1038/s41477-020-0725-0

33 Sato, H., Suzuki, T., Takahashi, F., Shinozaki, K. & Yamaguchi-Shinozaki, K. NF-YB2 and NF-YB3 Have Functionally Diverged and Differentially Induce Drought and Heat Stress-Specific Genes. *Plant Physiol* **180**, 1677-1690 (2019). https://doi.org:10.1104/pp.19.00391

34 Chen, Y.-S. *et al.* Two MYB-related transcription factors play opposite roles in sugar signaling in Arabidopsis. *Plant Mol. Biol.* **93**, 299-311 (2017). https://doi.org:10.1007/s11103-016-0562-8

35 Chung, B. Y. W. *et al.* An RNA thermoswitch regulates daytime growth in Arabidopsis. *Nat Plants* **6**, 522-532 (2020). https://doi.org:10.1038/s41477-020-0633-3

36 Leivar, P. & Monte, E. PIFs: systems integrators in plant development. *The Plant cell* **26**, 56-78 (2014). https://doi.org:10.1105/tpc.113.120857

37 Lv, X. *et al.* Structural insights into the multivalent binding of the Arabidopsis FLOWERING LOCUS T promoter by the CO–NF–Y master transcription factor complex. *The Plant Cell* **33**, 1182-1195 (2021). https://doi.org:10.1093/plcell/koab016

38 Stephenson, T. J., McIntyre, C. L., Collet, C. & Xue, G.-P. TaNF-YB3 is involved in the regulation of photosynthesis genes in Triticum aestivum. *Functional & Integrative Genomics* **11**, 327-340 (2011). https://doi.org:10.1007/s10142-011-0212-9

39 Sielemann, J., Wulf, D., Schmidt, R. & Bräutigam, A. Local DNA shape is a general principle of transcription factor binding specificity in Arabidopsis thaliana. *Nat. Commun.* **12**, 6549 (2021). https://doi.org:10.1038/s41467-021-26819-2

40 Tu, X. *et al.* Limited conservation in cross-species comparison of GLK transcription factor binding suggested wide-spread cistrome divergence. *Nat. Commun.* **13**, 7632 (2022). https://doi.org:10.1038/s41467-022-35438-4

41 Burko, Y. *et al.* Chimeric Activators and Repressors Define HY5 Activity and Reveal a Light-Regulated Feedback Mechanism. *The Plant Cell* **32**, 967-983 (2020). https://doi.org:10.1105/tpc.19.00772

42 Wang, P. *et al.* Re-creation of a Key Step in the Evolutionary Switch from C-3 to C-4 Leaf Anatomy. *Curr. Biol.* **27**, 3278-+ (2017). https://doi.org:10.1016/j.cub.2017.09.040

Methods

Variation analysis

RNA-seq data of [1] was downloaded, mapped with kallisto[2] 0.44 onto the TAIR10 transcriptome and tested for differential expression using edgeR[3]. P-values were Benjamini-Hochberg[4] corrected and labeled significantly different at $p<0.01$. 7,048 wildtype RNA-seq datasets were curated from the sequence read archive (SRA) information and downloaded. Each dataset was mapped with kallisto[2] 0.44.0 in single end (length 200 bp, standard deviation 20) or in paired end mode as specified in SRA. Experiments were filtered for at least 7.5 million mapped reads. The coefficient of variation was plotted for all transcripts as a violin plot with photosynthetic transcripts overlayed in green.

Promoter analysis

DAP-seq data DNA binding data was downloaded from[5] and visualized on promoters defined as 1 kb upstream plus sequences including the first intron using a color scale to denote DAP-seq peak height. The intergenic length was determined from the TAIR10[6] annotation. To test the difference between the intergenic length of the genes with the gene ontology term photosynthesis using a Wilcoxon rank sum test.

GRN construction and analysis

To infer a gene regulatory network of *A. thaliana* 6033 wildtype RNA-seq datasets (SuppData 13) were downloaded from the sequence read archive (SRA) [7]. Each dataset was mapped with kallisto[2] 0.44.0 in single end (length 200 bp, standard deviation 20) or in paired end mode as specified in SRA. Experiments were filtered for at least 7.5 million mapped reads. The tpm values were used as the expression matrix for GENIE3. A manually curated list of DNA-binding transcription factors was derived from TAP-scan[8]. Orthogroups were created using Orthofinder2[9] using *C. rheinhardtii* v5.5, *V. carteri* v2.0, *M. polymorpha* v3.1, *V. vinifera* v2.1, *S. lycopersicum* iTAG2.4, *A. thaliana* TAIR10, *B. napus* v5, *Z. mays* RefGen v4, *S. bicolor* v3.1.1, *O. sativa* v7.0, *B. distachyon* v3.1, *T. aestivum* IWGSC v1.0, and *H. vulgare* v1 proteins. Proteins with Interpro domains “DNA-binding” and “transcription factor” were added to the curation list. Each entry was manually inspected for functional annotation and those containing the following key words were deselected: cell cycle DNA replication / repair, nucleases, helicases, RNA-binding / splicing, telomere binding, IQ domain and FYVE/PHD-type zinc finger proteins, lipid binding STAR, pentatricopeptide repeats, pleckstrin homolog, protein kinase C-like, protein phosphatase 2C, DUF833. WD40 proteins were kept if they contained CTLH or LisH domains as those were shown to be present in transcription factors. If an orthogroup contained at least one gene annotated as a transcription factor as described above, all orthogroup members were labeled TFs. 2,399 TFs were used as regulators for GENIE3. COP-1 and paralogues included in the list of 2399 transcription factors. For GENIE3 a R C-wrapper was used with the random forest method 1000 trees and the square root of the number of regulators was used for the tree construction. Edges with a feature importance greater than 0.005 were defined as target genes and used for subsequent analysis. For community detection GENIE3 weights were filtered with a 0.005 minimum cutoff and converted to directional edges. Optimized numbers of rounds of iterative community detection were calculated with the walktrap algorithm from the igraph R package (https://igraph.org) with a walk length range 1:20. For each subsequent iteration the walk length resulting in the highest modularity value was picked. Visualizations were created with the visNetwork R package (http://datastorm-open.github.io/visNetwork). The GO-term enrichment for biological process for the target genes of each transcription factor was calculated with the topGO[10] (V2.5, nodesize=10, classicFisher mode) Rpackage in R 4.2 at different weight cut-offs for the matrix. The GenTable function was modified to report p-values with as small as 1e-1500.

Single gene analyses

Inducible overexpression lines[11] were ordered from NASC where available and tested for overexpression in induced plants vs controls using reverse transcription followed by semi-quantitative PCR (data not shown). Overexpression was confirmed using RNA-seq. For BBX14 and PIF8, plantlets were incubated on plates, induced daily with beta-estradiol and harvested after 72 hours of induction. For MYBS1, the coding sequence of MYBS1 with a stop codon was cloned (CGGGCTCAGGCCTGGATGGAGAGTGTGGTGGCAACATG, CCGGGAGCGGTACCCTCAGTGCATTGTCGACGGAGCT) with Gibson assembly between the AscI and XhoI restriction site of the UBQ10:sXVE:HAC:Bar vector[12]. The vector was transformed into *Arabidopsis thaliana* Col-0 using *Agrobacterium tumefaciens* using the floral-dip method[13]. Primary transformed seeds were selected on ½ MS with 25 mg/L glufosinate ammonium. After one week seedlings were picked and placed on soil and were grown in short day conditions for 6 weeks. For each plant two leaves were harvested and vacuum infiltrated 3 times for 20 seconds with 0.2 % ethanol water solution with or without 100 µM β-estradiol. The submerged leaves were put into long day conditions for 12 h. RNA was extracted with the QIAGEN RNeasy Plant Mini Kit with on the column DNAse digest using the QIAGEN RNase-Free DNase. After reverse transcription semi-quantitative PCR with a gene specific and a vector specific primer confirmed the induction of the β-estradiol treated leaves (ATCTGGAACATCGTATGGATACCCGGGAG, ACTGTGAACAATCAAGCTCCTGCGG) and validated via RNA-seq. RNA was isolated using the Qiagen plant RNeasy kit. A sequencing library was constructed with the TruSeq Illumina kit and sequenced on a NextSeq 550. RNA-seq data was further processed as described above.

ChIP and DAP seq data (SuppData 14) was processed using trimmomatic-0.39[14], aligned to the TAIR10 Arabidopsis genome using bowtie2 version 2.4.1[15], and filtered for nuclear genome hits, sorted, and converted to bam format with samtools-1.3[16]. For DAP-seq data, peaks were called with GEM version 3.1.4(-d Read_Distribution_default.txt -a 6 -icr 1 -print_bound_seqs -k_min 6 -k_max 20 -k_seqs 600) [17]. For ChIP-seq data, peaks were called with MACS version 2.2.7.1 (callpeak -gsize 1.118e8 -nomodel -nolambda -keep-dup auto -q 0.05 -call_summits) [18]. If replicates were available, overlapping peaks were determined[19] and only peaks present in all replicates kept. A promoter is called bound, if at least one experiment detects TF::promoter interaction at -750 to 0 relative to the transcriptional start site of the target gene. For MYBS1, binding data was downloaded from[5], processed as above, and enrichment was tested using Fishers Exact Test

Motif occurrence in photosynthetic promotors

The sequence 750 bp upstream of the transcriptional start site of photosynthesis genes was searched for motif occurrence for the motif of the GLK (GGATT), MYBS1 (GATAA), GNC (GATC) and the G-box (CACGTG) with FIMO[20]. The number of motif occurrences in the promotor was counted per gene.

1 Sewelam, N. *et al.* Molecular plant responses to combined abiotic stresses put a spotlight on unknown and abundant genes. *J. Exp. Bot.* **71**, 5098-5112 (2020). https://doi.org:10.1093/jxb/eraa250

2 Bray, N. L., Pimentel, H., Melsted, P. & Pachter, L. Near-optimal probabilistic RNA-seq quantification. *Nat Biotech* **34**, 525-527 (2016). https://doi.org:10.1038/nbt.3519 http://www.nature.com/nbt/journal/v34/n5/abs/nbt.3519.html#supplementary-information

3 Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**, 139-140 (2010). https://doi.org:10.1093/bioinformatics/btp616

4 Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate - a practical and powerful approach to multiple testing. *J. R. Stat. Soc. Ser. B-Methodol.* **57**, 289-300 (1995).

5 O'Malley, R. C. *et al.* Cistrome and Epicistrome Features Shape the Regulatory DNA Landscape. *Cell* **165**, 1280-1292 (2016). https://doi.org:10.1016/j.cell.2016.04.038

6 Swarbreck, D. *et al.* The Arabidopsis Information Resource (TAIR): gene structure and function annotation. *Nucl. Acids. Res.* **36**, D1009-D1014 (2008).

7 Leinonen, R., Sugawara, H., Shumway, M. & Collaboration, o. b. o. t. I. N. S. D. The Sequence Read Archive. *Nucl. Acids. Res.* **39**, D19-D21 (2010). https://doi.org:10.1093/nar/gkq1019

8 Wilhelmsson, P. K. I., Muehlich, C., Ullrich, K. K. & Rensing, S. A. Comprehensive Genome-Wide Classification Reveals That Many Plant-Specific Transcription Factors Evolved in Streptophyte Algae. *Genome Biol. Evol.* **9**, 3384-3397 (2017). https://doi.org:10.1093/gbe/evx258

9 Emms, D. M. & Kelly, S. OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biol.* **20**, 238 (2019). https://doi.org:10.1186/s13059-019-1832-y

10 Alexa, A. & Rahnenfuhrer, J. topGO: topGO: Enrichment analysis for Gene Ontology. *R package version 2.22.0* (2010).

11 Coego, A. *et al.* The TRANSPLANTA collection of Arabidopsis lines: a resource for functional analysis of transcription factors based on their conditional overexpression. *The Plant Journal* **77**, 944-953 (2014). https://doi.org:https://doi.org/10.1111/tpj.12443

12 Schlücking, K. *et al.* A new β-estradiol-inducible vector set that facilitates easy construction and efficient expression of transgenes reveals CBL3-dependent cytoplasm to tonoplast translocation of CIPK5. *Mol Plant* **6**, 1814-1829 (2013). https://doi.org:10.1093/mp/sst065

13 Bent, A. F. Arabidopsis in planta transformation. Uses, mechanisms, and prospects for transformation of other species. *Plant Physiol.* **124**, 1540-1547. (2000).

14 Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114-2120 (2014). https://doi.org:10.1093/bioinformatics/btu170

15 Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357-U354 (2012). https://doi.org:10.1038/nmeth.1923

16 Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078-2079 (2009).

17 Guo, Y., Mahony, S. & Gifford, D. K. High Resolution Genome Wide Binding Event Finding and Motif Discovery Reveals Transcription Factor Spatial Binding Constraints. *PLoS Comput. Biol.* **8**, e1002638 (2012). https://doi.org:10.1371/journal.pcbi.1002638

18 Zhang, Y. *et al.* Model-based Analysis of ChIP-Seq (MACS). *Genome Biol.* **9**, R137 (2008). https://doi.org:10.1186/gb-2008-9-9-r137

19 Zhu, L. J. *et al.* ChIPpeakAnno: a Bioconductor package to annotate ChIP-seq and ChIP-chip data. *BMC Bioinformatics* **11**, 237 (2010). https://doi.org:10.1186/1471-2105-11-237

20 Bailey, T. L. *et al.* MEME SUITE: tools for motif discovery and searching. *Nucl. Acids. Res.* **37**, W202-W208 (2009). https://doi.org:10.1093/nar/gkp335