Title: Decoding the regulatory architecture of the maize leaf

Authors: Xiaoyu Tu[1,2,*], María Katherine Mejía-Guerra[3,*,†], Jose A Valdes Franco[4], David Tzeng[2], Po-Yu Chu[2], Xiuru Dai[1], Pinghua Li[1,†], Edward S Buckler[3,4,5] Silin Zhong[2,†].

Affiliations:

[1]College of Agronomy, Shandong Agricultural University, China.

[2]State Key Laboratory of Agrobiotechnology, School of Life Sciences, The Chinese University of Hong Kong, Hong Kong, China

[3]Institute for Genomic Diversity, Cornell University, Ithaca, NY, USA

[4]School of Integrative Plant Sciences, Section of Plant Breeding and Genetics, Cornell University, Ithaca, NY, USA

[5]Agricultural Research Service, United States Department of Agriculture, Ithaca, NY, USA

†Corresponding author. Email: mm2842@cornell.edu (M.K.M.); pinghuali@sdau.edu.cn (P.L.); silin.zhong@cuhk.edu.hk (S.Z.)

*These authors contributed equally to this work.

Abstract:

Characterize transcription factors (TF) binding to cis-regulatory elements is key to understand the effect of genetic variability on phenotypic differences. Here, we used 104 TF ChIP-seq to annotate the regulatory landscape of the maize leaf. The ~2 millions of identified TF binding sites co-localized with open chromatin, are evolutionarily conserved, and show enrichment for GWAS-hits and cis-expression QTLs. The resulted TF-target regulatory network covers half of the annotated genes and shows characteristics of real-word networks such as scale-free topology and modularity (structural and functional). Machine-learning analyses reveal that sequence preferences are alike within TF families, and that co-localization is key for binding specificity. Our comprehensive approach provides a starting point to decipher the gene regulatory architecture in plants.

One Sentence Summary: Integration of 104 ChIP-seq experiments highlights TF binding redundancy and modularity as key properties of the maize regulatory network.

Main Text:

Genome non-coding regions are major contributors to observable phenotypic differences and adaptation, between and within species (*1*), likely by regulating gene expression. However, assessing their true functions are an ongoing challenge. For example, 70% of maize GWAS hits are located in the non-coding regions without functional annotation (*2*). Recently, it was shown that sequence variation in the open chromatin within non-coding regions could explain up to ~40% of the phenotypic variation of key agronomic traits (*3*). Several large-scale projects (e.g., ENCODE and modENCODE) have been launched, aiming to systematically characterize all functional elements in the genome (*4*); but similar efforts are seldom feasible for individual laboratories, and have yet to be fully explored outside metazoan (*5*).

Maize is an important crop plant and model for biology research. Its leaves are among the most photosynthetically efficient of all plants, yet the information flow modulated by transcription factors (TFs) that allows this design is largely unknown. In order to characterize its regulatory architecture, we developed a modified ChIP-seq assay that eliminates the need for specific antibodies and reduces the amount of material required for library preparation, two major limitations for quality and scalable assays (*6*). We successfully profiled the binding for 104 TFs expressed in the maize leaves (Fig. 1 and table S1). We applied the ENCODE uniform pipeline for quality-control and processing of the ChIP-seq data (*6*) (fig. S1), and identified a total of 2,147,346 reproducible peaks, that cluster in 144,890 non-overlapping TF binding loci (~2% of the genome, fig. S2, table S2). In addition, we measured chromatin accessibility in the same tissue using ATAC-seq (*7*). In total, we identified 38,713 biologically reproducible open chromatin regions (table S3).

From the analysis of large-scale TF ChIP-seq of ENCODE and modENCODE, we expect plant TF peaks would also frequently locate within open chromatin regions (*4*, *8*). We found that the maize TF binding loci and open chromatin regions show similar genome-wide distributions, frequently proximal to gene bodies (+/- 2.5kb), with preferences for the 5' end (Fig. 1B). The distance between TF binding to genes after excluding regions that overlap with gene bodies is bimodal (Fig. 1C). Despite the larger space available for distal regulation (grey curve, Fig. 1C), we observed that regions between 10-100kb constitute ~15% of the TF binding loci, and ~17% of the open chromatin regions. Layering ATAC-seq and ChIP-seq data show that TF binding loci and

open chromatins overlap (fig. S3, P-value < 10-5). On average, ~74% of the peaks for a given TF (IQR25-75 64%-87%) intersect with open chromatin regions (fig. S4) confirming the relevance of the identified TF binding sites within the chromatin context. Collectively, 98% of the open chromatin regions overlap with TF peaks (Fig. 1D) suggesting a large number of possible TF combinations that co-regulate transcription.

TF binding sites are key determinants of transcriptional regulation, and if purifying selection is effective in these regions, they should exhibit low sequence diversity. We examined the conservation of the TF binding sites by assessing the overall nucleotide diversity represented in the maize HapMap (*9*) while controlling for overall SNP density in function to the distance of TF's peak summit (Fig. 2A). The result confirmed that sequence variation is, in fact, reduced.

While most TFs binding is constrained, TF and their binding sites are key to local adaptation or domestication (e.g., ZmTB1 a known TF important for maize domestication) (*12*). Hence, we predict that TF binding loci will be enriched for common SNP variations controlling gene expression and downstream traits. This was first tested in a panel of 282 inbred breeding lines for their effect on mRNA expression using common and likely adaptive variants (*13*). We found twofold enrichment of TF binding loci (95% credible interval 2.26-2.46), similar to the enrichment around 5'UTRs (95% credible interval 1.80-2.11) and 3'UTRs (95% credible interval 2.37-2.73) (Fig. 2B). Distal and proximal TF binding loci are both enriched (fig. S5). The enrichment pattern is ubiquitous across the 104 TFs that we examined (fig S6 and table S4). Overall, these are strong and consistent supports for common variation in mRNA expression being controlled by TF binding site variation.

Next, we tested the overlap with functional variations associated to traits others than gene expression. To this end, we calculated the enrichment in GWAS hits for seven traits related to metabolites (*14*), leaf architecture (*15*), and photoperiodicity (*16*) measured in the US NAM population. Overall, TF binding loci are enriched for four of the traits (Fig. 2C). Across TFs, simpler traits, such as metabolites, show few TFs enriched for GWAS hits (e.g. malate and nitrate, fig. 7A and table S5). In contrast, many TF bindings overlap with GWAS hits for complex traits, which are known to be polygenic and influenced by a large number of genetic variants (e.g., days

after silking and days after anthesis, Fig. 2D, fig. S7B-C). The general trend for GWAS enrichment supports our hypothesis that non-coding region variations related to traits are often mediated by TF. Furthermore, our finding shows the potential of our data to connect sequence variation in *cis* to *trans*-regulators and highlights the regulatory networks implicated in complex phenotypes (Fig. 2D).

Real-world networks frequently exhibit a scale-free topology (*17*). To determine the feasibility of our data to pinpoint true regulatory relationships, we tested if a network derived from our data could fit such topology. We reshaped the regulatory data into a graph, adopting the ENCODE probabilistic framework to identify high confidence proximal interactions (P-value < 0.05) (*4*, *6*, *18*). This approach renders a graph with 20,179 nodes (~45% of the annotated genes and ~77% of the leaf expressed genes, table S7). We evaluate the in-degree distribution (Fig. 3A), or number of edges towards each node, and found it to follow a linear trend in the log-scale (R2 = 0.882, P-value < 2e-16), as expected for power-law distribution (goodness of fit P-value = 0.67) (Fig. 3A). This distribution is a landmark of scale-free networks characterized for few nodes critical for information flow, having high in-degree (*17*). This suggests that the graph of TFs-targets derived from our data resembles a real-world network.

True biological networks often exhibit topological and/or functional modularity (*19*, *20*). We tested for this predicted property by contrasting the maximum modularity (*21*) in our network to a null distribution from an ensemble of random rewired graphs (H0: 1000 rewired graphs). The result confirmed that our network exhibits a significant increase in modularity (P-value < 0.05, fig. S8). We identified seven modules (Gephi resolution 1.0, table S8), each containing from ~27% to ~5% of the total nodes, and ~40% of the total edges occurring within the modules, suggestive of large information flow between them (Fig. 3B).

We predicted that topological modularity should be related to function in known biological processes. GO enrichment analyses show that all seven modules were indeed enriched for specific GO terms (Fig. 3B and fig. S9). We identified two photosynthesis-related modules enriched in targets of GLK and CONSTANT(CO)-like TFs, known regulators of photosynthesis, circadian and light signalling (*22*, *23*) (table S9). This suggests that the network structural and functional

split could contribute to fine tuning the regulation of different aspects of photosynthesis (Fig. 3C-D). On the other hand, the presence of multiple TFs across modules to regulate genes in the same pathway suggests intricate mode of actions to coordinate transcription, in contrast to the classic view of a singular or few master regulators. Altogether, the topological and functional features of the regulatory graph provide validation of known TF to target gene interactions, and additional clues to understand a vast number of uncharacterized genes and TFs.

With a genome larger than two billion bp, any given TF (avg. 6.8 bp) could have affinity for roughly a third of a million locations across the genome (*24-25*). Under this scenario, achieving the observed TF binding specificity would likely require extra cues. We predict that co-binding and combinatorial binding must be key for specificity. To test this, we created machine-learning models based on either sequence information or co-localization information (Fig. 1A). To model TF binding from sequence we applied a "bag-of-k-mers" (*24*) model to discriminate TF binding regions from other regions in the genome, which resulted in reliable models for all the TFs (5-fold cross-validation, average accuracy for each TF > 70%) (table S9, fig. S10). Using average k-mer weights from the models (table S10), we derived a distance matrix among TFs, and summarized TFs relationships (Fig. 4A). After removal of singleton families, we observed that for 85% of the TFs families, most of their members (>= 50%) cluster into the same group in a dendrogram (fig. S11).

This observation prompted us to evaluate whether TF sequence preference has persisted across angiosperm evolution, as many TF protein families are well conserved. Using the top 1% of the predictive k-mers for each TF, we examined their similarity to a large collection of *in vitro Arabidopsis* TF binding position weight matrices (PWMs) (*25*). After removal of families that did not have a counterpart (or were poorly represented), 50 out of 81 (61%) of the evaluated TFs preferentially match PWMs to their corresponding family in *Arabidopsis* (P-value < 0.001, fig S12). This sequence preference conservation suggests strong constraint over more than 150 million years, which also agrees with a reduced SNP variation at the peaks (Fig. 2A).

To model TF binding from co-localization, we adopted a machine-learning approach used in the ENCODE project to learn non-linear dependencies among TFs (*4*). To fit a model for each TF (i.e.,

"focus TF" or context) we built a co-localization matrix, by overlapping peaks for the focus-TF with peaks of all remaining TFs (i.e., partners) (Fig. 1A). The co-localization model was aimed to discriminate between the true co-localization matrix and a randomized version of the same (*26*). The output of each model is a set of combinatorial rules that can predict TF binding. For each TF, the average of 10 models with independent randomized matrices have area under the receiver operating curve > 0.9 (table S12). The high performance of the model agrees with the hypothesis that co-localization has vast information content to determine binding specificity.

Using the rules derived from the co-localization models, we scored the relative importance (RI) of each partner TF for the joint distribution of the set of peaks for a focus TF. TFs with large RI score are predictive of the binding of specific focus TF (fig. S13-14). For instance, the co-localization model for GLK2 suggests that MYBR26 is a top partner (table S12). The relationship between these two TFs was also suggested in the network view of the regulation of chlorophyll biosynthetic genes, and with the bag-of-k-mer models of the promoter of a photosynthetic gene (Fig 3B, 4B and fig. S15). Overall, the results of the co-localization model indicate a large number of possible combinations of TF bindings, that appear key to binding specificity.

In this study, we present a bottom-up approach to systematically identify emergent properties of transcriptional regulation in plant leaves. To do so, we resolved binding profiles for 104 TFs in the maize leaf and describe how TF binding sites massively overlap with open chromatin regions, are conserved, and enriched in eQTLs and GWAS hits, suggesting that binding loci are under selection and functional. Our TF-target dataset captures the backbone of a complex network that is scale-free and shows a modular organization at topological and functional level. In the future, we foresee the incorporation of distal TF binding sites (~ 30% of the total) will enrich the topology of the proposed network.

Finally, we generated quantitative and highly interpretable models of TF binding, and TF co-localization. On a practical level, we expect these models to be integrated into pipelines to predict effects of non-coding variants, both common and rare, on TF binding, to pinpoint causal sites. The possibility to predict and generate novel variation not seen in nature could fundamentally change future plant breeding.

References and Notes:

1. M. C. King, A. C. Wilson, Evolution at two levels in humans and chimpanzees. Science. **188**, 107–116 (1975).
2. J. G. Wallace et al., Association Mapping across Numerous Traits Reveals Patterns of Functional Variation in Maize. PLoS Genet. **10**, e1004845 (2014).
3. E. Rodgers-Melnick, D. L. Vera, H. W. Bass, E. S. Buckler, Open chromatin reveals the functional maize genome. Proc. Natl. Acad. Sci. U.S.A. **113**, E3177–84 (2016).
4. M. B. Gerstein et al., Architecture of the human regulatory network derived from ENCODE data. Nature. **489**, 91–100 (2012).
5. Lü et al., Genome encode analyses reveal the basis of convergent evolution of fleshy fruit ripening. Nature Plants. **4**, 784–791 (2018).
6. S. G. Landt et al., ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. Genome Res. **22**, 1813–1831 (2012).
7. J. D. Buenrostro, P. G. Giresi, L. C. Zaba, H. Y. Chang, W. J. Greenleaf, Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. Nat. Methods. **10**, 1213–1218 (2013).
8. S. Neph et al., Circuitry and Dynamics of Human Transcription Factor Regulatory Networks. Cell. **150**, 1274–1286 (2012).
9. R. Bukowski et al., Construction of the third-generation Zea mays haplotype map. Gigascience. **7**, 1–12 (2018).
10. Y.-X. Li et al., Identification of genetic variants associated with maize flowering time using an extremely large multi-genetic background population. Plant J. **86**, 391–402 (2016).
11. P. Alter et al., Flowering Time-Regulated Genes in Maize Include the Transcription Factor ZmMADS1. Plant Physiol. **172**, 389–404 (2016).
12. J. Doebley, A. Stec, L. Hubbard, The evolution of apical dominance in maize. Nature. **386**, 485–488 (1997).
13. K. A. G. Kremling et al., Dysregulation of expression correlates with rare-allele burden and fitness loss in maize. Nature. **555**, 520–523 (2018).
14. N. Zhang et al., Genome-wide association of carbon and nitrogen metabolism in the maize nested association mapping population. Plant Physiol. **168**, 575–583 (2015).
15. F. Tian et al., Genome-wide association study of leaf architecture in the maize nested association mapping population. Nature Genetics. **43**, 159–162 (2011).
16. E. S. Buckler et al., The Genetic Architecture of Maize Flowering Time. Science. **325**, 714–718 (2009).
17. A.-L. Barabási, R. Albert, Emergence of Scaling in Random Networks. Science. **286**, 509–512 (1999).
18. C. Cheng, R. Min, M. Gerstein, TIP: A probabilistic method for identifying transcription factor target genes from ChIP-seq binding profiles. Bioinformatics. **27**, 3221–3227 (2011).

19. J. M. Olesen, J. Bascompte, Y. L. Dupont, P. Jordano, The modularity of pollination networks. Proc. Natl. Acad. Sci. U.S.A. **104**, 19891–19896 (2007).

20. M. T. Dittrich, G. W. Klau, A. Rosenwald, T. Dandekar, T. Müller, Identifying functional modules in protein-protein interaction networks: an integrated exact approach. Bioinformatics. **24**, i223–31 (2008).

21. A. Clauset, M. E. J. Newman, C. Moore, Finding community structure in very large networks. Phys Rev E Stat Nonlin Soft Matter Phys. **70**, 066111 (2004).

22. M. T. Waters et al., GLK transcription factors coordinate expression of the photosynthetic apparatus in Arabidopsis. The Plant Cell. **21**, 1109–1128 (2009).

23. P. Wang, S. Kelly, J. P. Fouracre, J. A. Langdale, Genome-wide transcript analysis of early maize leaf development reveals gene cohorts associated with the differentiation of C4 Kranz anatomy. Plant J. **75**, 656–670 (2013).

24. M. K. Mejia-Guerra, E. S. Buckler, A k-mer grammar analysis to uncover maize regulatory architecture. BMC Plant Biology. **19**, 103 (2019).

25. R. C. O'Malley et al., Cistrome and Epicistrome Features Shape the Regulatory DNA Landscape. Cell. **165**, 1280–1292 (2016).

26. J. H. Friedman, B. E. Popescu, Predictive learning via rule ensembles. The Annals of Applied Statistics. **2**, 916–954 (2008).

27. S. Picelli, Å. K. Björklund, B. Reinius, S. Sagasser, G. Winberg, R. Sandberg, Tn5 transposase and tagmentation procedures for massively scaled sequencing projects. Genome Res. **24**, 2033–2040 (2014).

28. K.-Y. Han, J.-A. Song, K.-Y. Ahn, J.-S. Park, H.-S. Seo, J. Lee, Solubilization of aggregation-prone heterologous proteins by covalent fusion of stress-responsive Escherichia coli protein, SlyD. Protein Eng Des Sel. **20**, 543–549 (2007).

29. B. Burdo, J. Gray, M. P. Goetting-Minesky, B. Wittler, M. Hunt, T. Li, D. Velliquette, J. Thomas, I. Gentzel, M. dos S. Brito, M. K. Mejía-Guerra, L. N. Connolly, D. Qaisi, W. Li, M. I. Casas, A. I. Doseff, E. Grotewold, The Maize TFome – development of a transcription factor open reading frame collection for functional genomics. Plant J. **80**, 356–366 (2014).

30. B. Langmead, S. L. Salzberg, Fast gapped-read alignment with Bowtie 2. Nature Methods. **9**, 357–359 (2012).

31. H. Li, B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, R. Durbin, The Sequence Alignment/Map format and SAMtools. Bioinformatics. **25**, 2078–2079 (2009).

32. F. Ramírez, D. P. Ryan, B. Grüning, V. Bhardwaj, F. Kilpert, A. S. Richter, S. Heyne, F. Dündar, T. Manke, deepTools2: a next generation web server for deep-sequencing data analysis. Nucleic Acids Res. **44**, W160–W165 (2016).

33. P. V. Kharchenko, M. Y. Tolstorukov, P. J. Park, Design and analysis of ChIP-seq experiments for DNA-binding proteins. Nat Biotechnol. **26**, 1351–1359 (2008).

34. A. R. Quinlan, I. M. Hall, BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics. **26**, 841–842 (2010).

35. Q. Li, J. B. Brown, H. Huang, P. J. Bickel, Measuring reproducibility of high-throughput experiments. Ann. Appl. Stat. **5**, 1752–1779 (2011).

36. J. Feng, T. Liu, B. Qin, Y. Zhang, X. S. Liu, Identifying ChIP-seq enrichment using MACS. Nat Protoc. **7**, 1728–1740 (2012).

37. B. Gel, A. Díez-Villanueva, E. Serra, M. Buschbeck, M. A. Peinado, R. Malinverni, regioneR: an R/Bioconductor package for the association analysis of genomic regions based on permutation tests. Bioinformatics. **32**, 289–291 (2016).

38. R. Bukowski, X. Guo, Y. Lu, C. Zou, B. He, Z. Rong, B. Wang, D. Xu, B. Yang, C. Xie, L. Fan, S. Gao, X. Xu, G. Zhang, Y. Li, Y. Jiao, J. F. Doebley, J. Ross-Ibarra, A. Lorant, V. Buffalo, M. C. Romay, E. S. Buckler, D. Ware, J. Lai, Q. Sun, Y. Xu, Construction of the third-generation Zea mays haplotype map. Gigascience. **7**, 1–12 (2018).

39. P. Danecek, A. Auton, G. Abecasis, C. A. Albers, E. Banks, M. A. DePristo, R. E. Handsaker, G. Lunter, G. T. Marth, S. T. Sherry, G. McVean, R. Durbin, The variant call format and VCFtools. Bioinformatics. **27**, 2156–2158 (2011).

40. A. Gaudinier, J. Rodriguez-Medina, L. Zhang, A. Olson, C. Liseron-Monfils, A.-M. Bågman, J. Foret, S. Abbitt, M. Tang, B. Li, D. E. Runcie, D. J. Kliebenstein, B. Shen, M. J. Frank, D. Ware, S. M. Brady, Transcriptional regulation of nitrogen-associated metabolism and growth. Nature. **563**, 259–264 (2018).

41. A. Clauset, C. R. Shalizi, M. E. J. Newman, Power-law distributions in empirical data. SIAM Rev. **51**, 661–703 (2009).

42. C. S. Gillespie, Fitting Heavy Tailed Distributions: The poweRlaw Package. Journal of Statistical Software. **64**, 1–16 (2015).

43. M. Van Bel, T. Diels, E. Vancaester, L. Kreft, A. Botzki, Y. Van de Peer, F. Coppens, K. Vandepoele, PLAZA 4.0: an integrative resource for functional, evolutionary and comparative plant genomics. Nucleic Acids Res. **46**, D1190–D1196 (2018).

44. M. Bastian, S. Heymann, M. Jacomy, Gephi: An Open Source Software for Exploring and Manipulating Networks. ICWSM. **8**, 361–362 (2009).

45. F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, É. Duchesnay, Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research. **12**, 2825–2830 (2011).

46. D. Müllner, fastcluster: Fast Hierarchical, Agglomerative Clustering Routines for R and Python. Journal of Statistical Software. **53**, 1–18 (2013).

47. E. Tanaka, T. Bailey, C. E. Grant, W. S. Noble, U. Keich, Improved similarity scores for comparing motifs. Bioinformatics. **27**, 1603–1609 (2011).

48. G. Cuellar-Partida, F. A. Buske, R. C. McLeay, T. Whitington, W. S. Noble, T. L. Bailey, Epigenetic priors for identifying active transcription factor binding sites. Bioinformatics. **28**, 56–62 (2012).

Acknowledgments: We thank B. Du, R. Foo, J. Li and M. Huang for technical support, G. Ramstein for advice in the analysis of GWAS data, and S. Miller for proofreading the manuscript. **Funding:** This work was funded by National Science Foundation of China 31871313, National Key Research and Development Program of China 2016YFD0101003, National Science Foundation of China 31871313, Taishan Pandeng program, USDA-ARS, NSF PGRP #1238014, as well as Hong Kong GRF-14108117 and AoE/M-403/16. J.A. Valdes Franco acknowledges support from CONACyT-I2T2 through scholarship for graduate studies; **Author contributions:** S. Zhong, E.S. Buckler, and P. Li designed and supervised the research; X. Tu. developed techniques, performed experiments and processed the raw data with assistance from D. Tzeng, P. Chu and X. Dai. M.K. Mejia-Guerra designed and performed computational analysis, J.A.Valdes Franco performed population genetics analysis; M.K. Mejia-Guerra wrote the paper. All the authors read the paper and agree with the final version; **Competing interests:** Authors declare no competing interests.

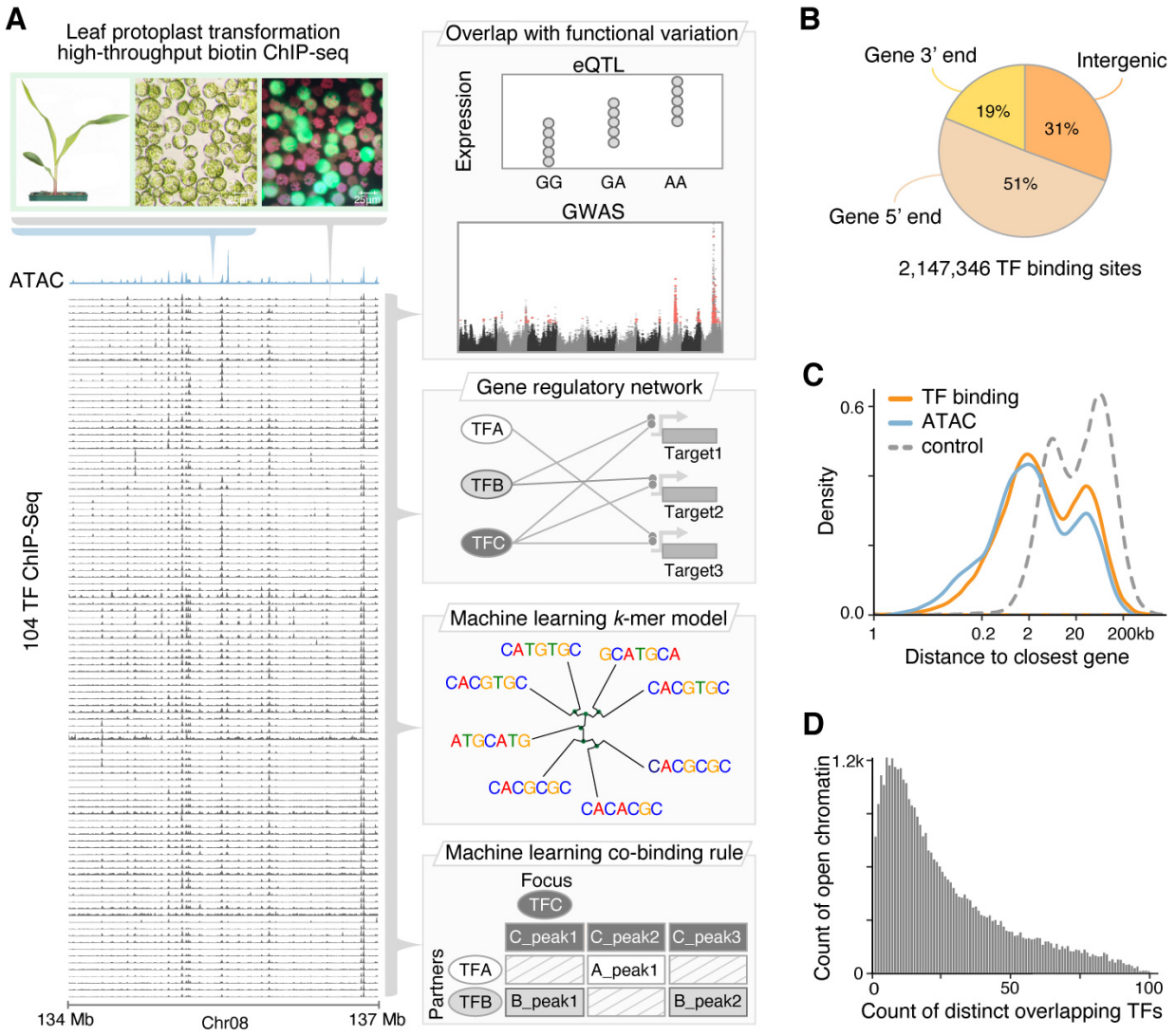

Fig. 1. The regulatory landscape of the maize leaf. (A). Overview of the analytical approaches used to decode the regulatory architecture of the maize leaf. TF ChIP-Seq and ATAC-Seq were performed using maize leaf protoplasts. The dataset was used in the following analysis (*i*) overlapped TF binding loci with phenotype-associated SNPs for an assortment of traits; (*ii*) reshaped TF binding data into the regulatory network; used TF binding regions to train models that can predict binding from either (*iii*) sequence information or (*iv*) TF co-localization information. **(B)** Genome-wide distribution of TF binding loci nearby (+/- 2.5 kb) to the annotated genes, and remaining distal regions. **(C)** Density plots corresponding to distances between annotated genes (grey), distances of TF binding loci (pink) and open chromatin regions (blue) to closest annotated gene after excluding regions within or overlapping with gene bodies. **(D)** Distribution of the count of distinct TFs that overlap with open chromatin regions.

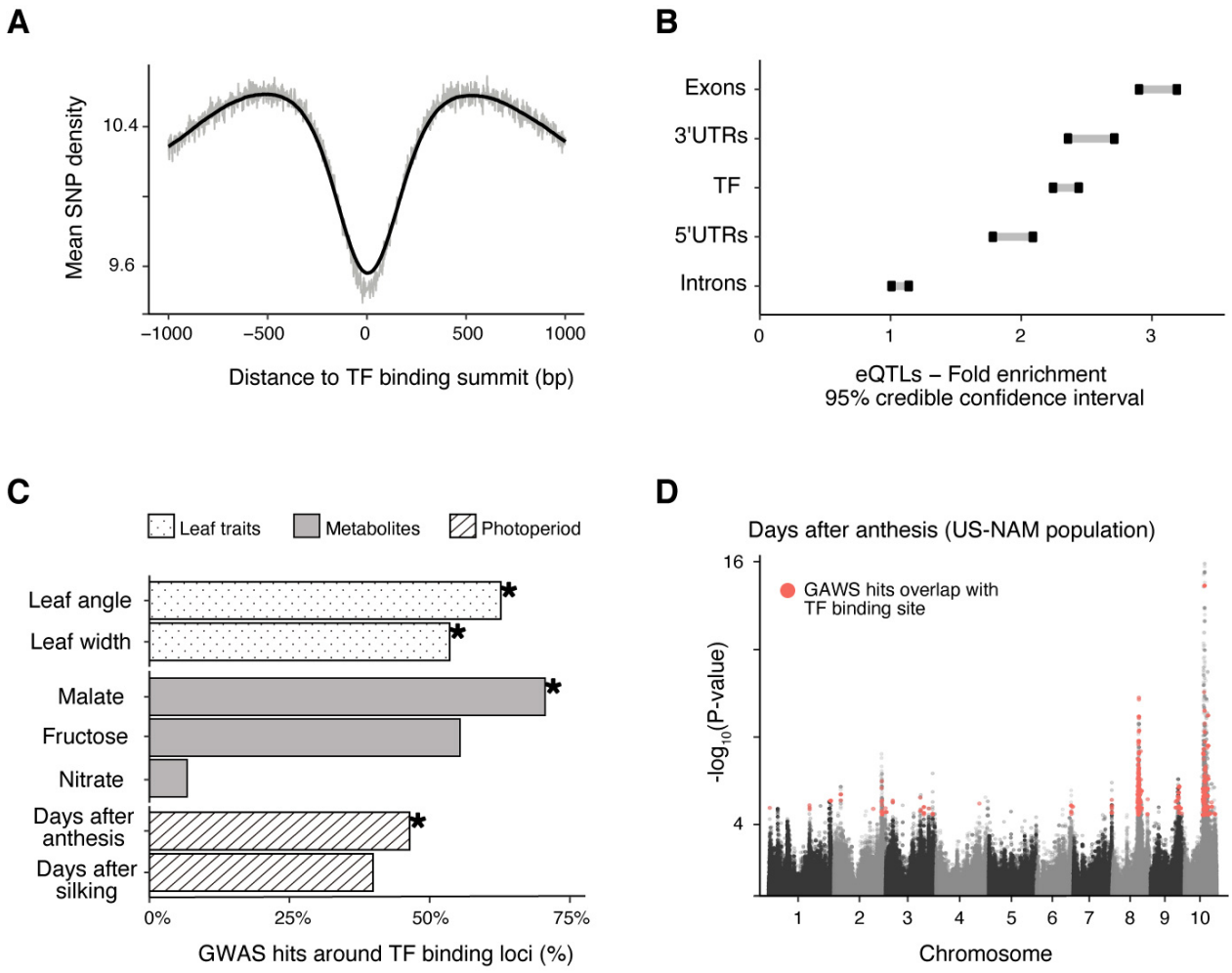

Fig. 2. Characterization of nucleotide diversity under regulatory regions. (A) Mean SNPs density calculated in sliding windows of 100 bp bins flanking TF binding summits, and normalized by mappability rate. **(B)** 95% confidence interval for the enrichment of SNPs associated to variation in mRNA levels (eQTLs) vs. non eQTLs SNPs, and relative to control regions for different sets of genomic regions **(C)** Proportion of phenotype-associated GWAS hits for an assortment of traits overlapping to TF binding loci. Traits in which the enrichment was statistically significant are labelled with an asterisk **(D)** Manhattan plot of GWAS for days after anthesis. Highlighted GWAS hits overlap with binding regions for a group of TFs (PRR5, ELF3, COL3, COL7, COL18 and DOF3/PBF1) associated to photoperiod variation (*10*, *11*).

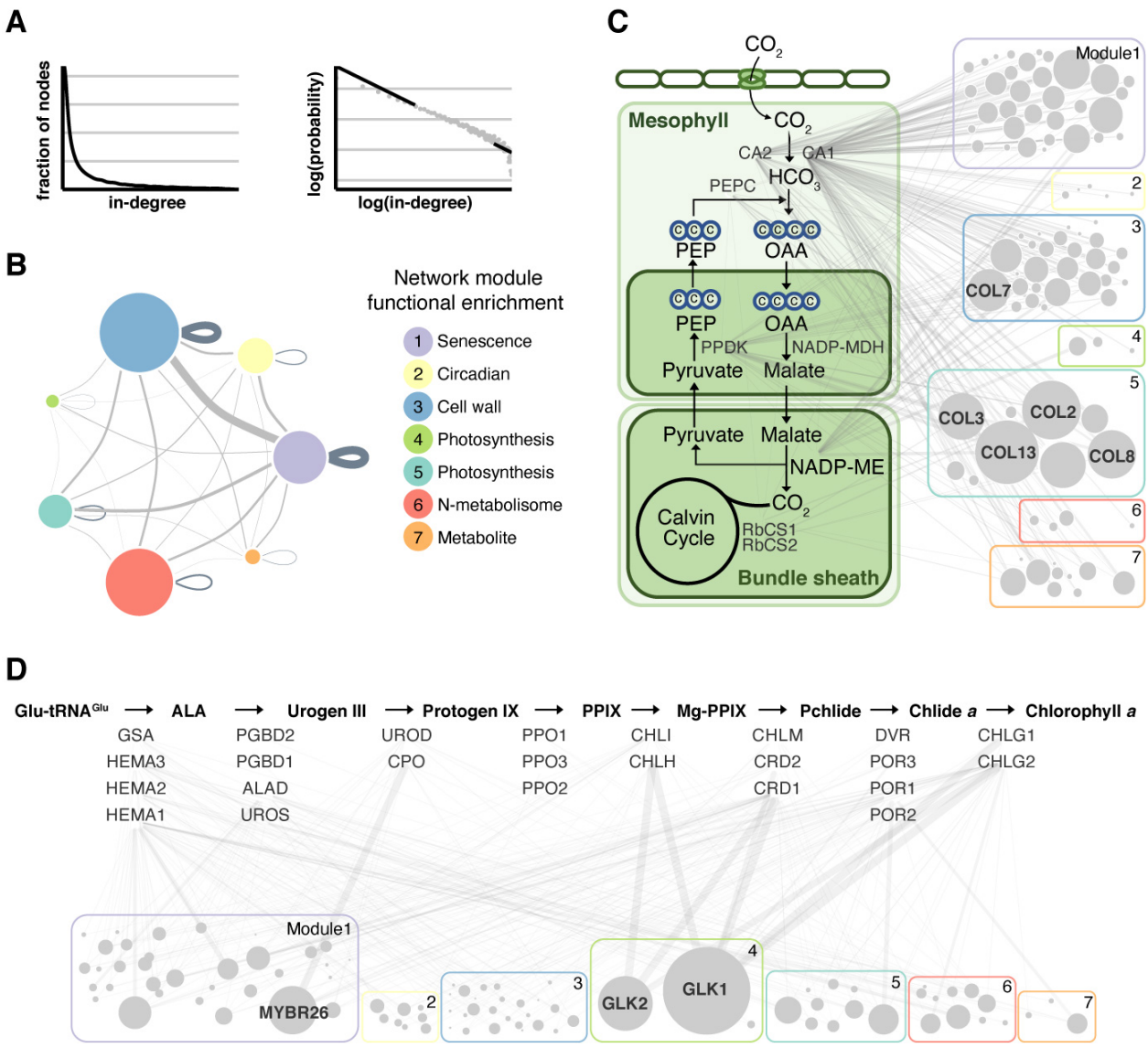

Fig. 3. Properties of the regulatory network. (A) Distribution of node *in-degree* values and the QQ-plot version of the same distribution showing the power-law fit (dashed line). **(B)** A graph diagram to summarize the modular structure of the network. Each module is represented as a meta-node (circles) with size proportional to the number of nodes within. Meta-nodes are linked with meta-edges with the width proportional to the number of edges within (~40%) or across modules (~60%). Modules are named according to their enriched GO terms (**C-D**) Detailed view of two photosynthesis pathways enriched in modules 5 and 4, respectively.

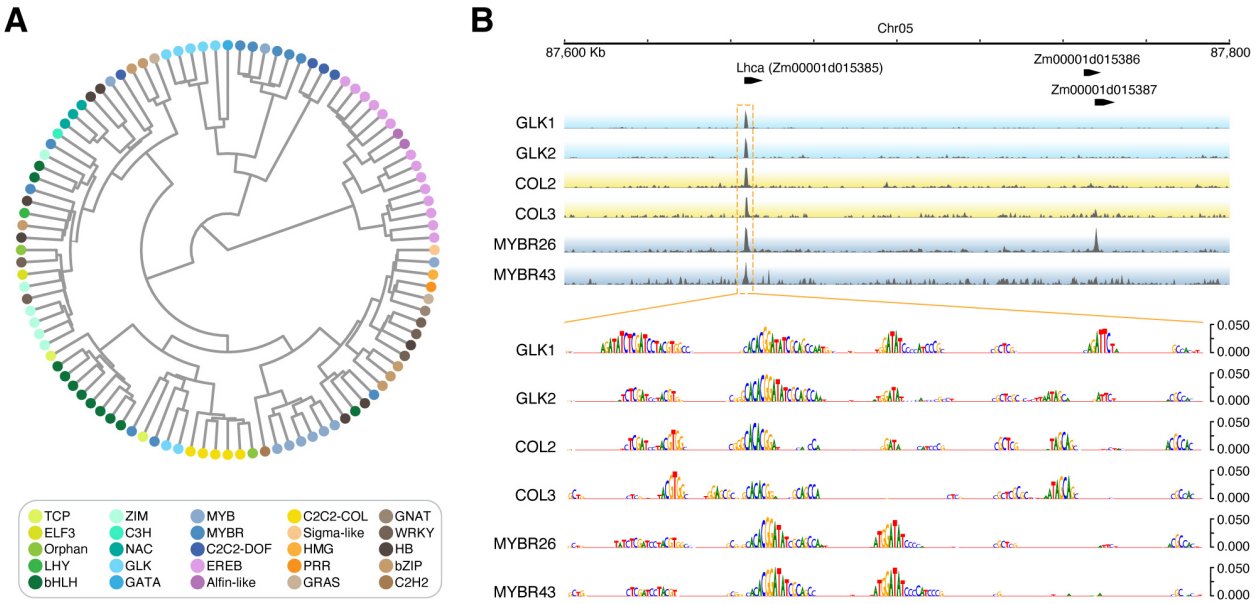

Fig. 4. Models of TF binding. (A) 104 TF clustered by sequence binding similarity (predicted *k*-mer weights) derived from the sequence models. **(B)** Genome browser view of the promoter region of the light-harvesting complex of PSI gene (*Lhca*, Zm00001d015385). Occlusion analysis for TFs targeting *Lhca* showing how the bag-of-k-mers model could highlight putative regulatory positions at base pair resolution.