UNRAVELING PLANT GENE REGULATORY NETWORKS USING MULTILAYER DATA INTEGRATION

By

Fabio Andrés Gómez Cano

A DISSERTATION

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

Biochemistry and Molecular Biology - Doctor of Philosophy

2023

ABSTRACT

The translation of genotype into phenotype largely depends on genes being expressed in the appropriate cell types at the correct time. These expression patterns are largely determined by transcription factors (TFs) controlling specific gene sets which together result in gene regulatory networks (GRN). GRNs may be elucidated using TF-centered approaches, such as DNA-affinity purification and chromatin immunoprecipitation sequencing (DAP- & ChIP-seq, respectively). Alternatively, the generation of thousands of gene expression samples has allowed the implementation of robust methods for TF-target inference. As part of my research, I developed strategies that integrate several high-throughput data types to identify transcription factor regulators of a broad spectrum of metabolic pathways in several plant systems. Specifically, I established frameworks for the analysis of *Camelina sativa*, maize (*Zea mays*), and *Arabidopsis thaliana* with species-specific tailored pipelines. Data resources availability by species-guided pipeline differed between species. In *Camelina*, I combined expression and DAP-seq assays to identify transcriptional regulators of lipid metabolism. In maize, I integrated expression variation, expression quantitative loci (eQTLs), and DAP- & ChIP-seq to build a multiple-layer network predicting regulators of phenylpropanoid, lipids, and carbon metabolism. Lastly, for Arabidopsis, utilizing a vast collection of RNA-seq samples, protein-DNA interactions (PDI), and protein-protein interactions (PPI), I tested co-regulation models that incorporate the influence of TF physical interactors on TF-target co-expression profiles. This comprehensive analysis also enabled the prediction of high-level TF complexes, providing valuable insights for refining models of TF regulation based on co-expression. Together, my studies contributed new knowledge to the regulatory hypotheses of specific metabolic pathways in plants, establishing a framework for elucidating GRN in other systems.

ACKNOWLEDGEMENTS

I wish to extend sincere gratitude to my esteemed advisor, Dr. Erich Grotewold, and all the distinguished members – both present and past – of the Grotewold Lab. The remarkable contributions and fruitful collaborations bestowed upon this work have been invaluable, making this endeavor possible through their unwavering support and steadfast dedication. I would also like to express my appreciation to all the committee members for their willingness to actively participate and continually contribute, further enhancing the quality and impact of this work. Moreover, I am deeply thankful for the support from Michigan State University and the NRT-IMPACTS fellowship program during my doctoral studies.

I would also like to acknowledge collaborators who played a crucial role in my research, such as Dr. Danny Schnell, Dr. Shin-Han Shiu, Dr. Patrick Edger, Dr. Arjun Krishnan, and Dr. Nathan Springer, all of whom were indispensable to my research outcomes. A special thanks goes to Dr. Arjun Krishnan, who consistently accompanied my journey through the PhD with a willingness to listen to any sort of unconventional ideas that come to mind, guiding them in the right direction.

Although this PhD took five years, it is a dream that started many years ago involving a lot of people who supported and brought light to me in the darkest moments. I would like to thank especially Angela Natalia Cano Garavito and Nubia Suárez Rincón, who were like angels for me during my early days in high school and as an undergraduate student back in Colombia. Luis Angel Cano Garavito, friend and big brother who was always there for give me hand when need it. Fabio Aldemar Gomez Sierra, mentor who guided me in nurturing my scientific passion. I will also express my gratitude to Juan Gonzalez and Julieta Manrique, who were more than friends; they were like my second parents in the US.

I want to extend my deep appreciation to my cherished parents and siblings for their unwavering support, invaluable guidance, and selfless sacrifices. Without them, none of my achievements would have been possible. I'd also like to thank my dear Isaac and Zoe, who are my oxygen and reason to persevere each day. Lastly, I express my gratitude to my beloved wife – Mariel – who serves as my guiding light in the storm, my oasis in the desert, and my faithful companion on this journey. This PhD is dedicated to her!

TABLE OF CONTENTS

LIST OF ABBREVIATIONS

ABA	Abscisic acid
ACR	Accessible chromatin regions
ATAC	Assay for transposase accessible chromatin
ChIP-seq	Chromatin immunoprecipitation sequencing
CRE	*Cis*-regulatory element
CRMs	*Cis*-regulatory modules
CUT&RUN	Cleavage under targets and release using nuclease
CUT&Tag	Cleavage under targets and tagmentation
DEG	Differentially expressed gene
DNA	Deoxyribonucleic acid
eQTL	expression quantitative loci
GRN	Gene regulatory network
GWAS	Genome-wide association studies
HCG	Highly co-expressed gene
HCT	Highly co-expressed target
LCT	Low co-expressed target
PDI	Protein-DNA interaction
PPI	Protein-protein interaction
RNA	Ribonucleic acid
SNP	Single-nucleotide polymorphism
TF	Transcription factor
TE	Transposable element

TRAP	Translating ribosome affinity purification
WiDiv	Wisconsin diversity

CHAPTER ONE: INTRODUCTION

1.1 GENE REGULATORY NETWORKS (GRNs)

Plants, unlike many other organisms, are sessile but account for over 80% of biomass on Earth (Bar-On et al., 2018). Their remarkable success can be attributed to their physiological diversity, which is governed by complex molecular networks. Therefore, a plant phenotype, whether it is morphological or physiological, can be defined as an emergent property of the molecular interactions that underlie it. Within these intricate molecular networks, transcription factor (TF) proteins play a crucial role as they are positioned at the end of signaling pathways and guide the transcription machinery responsible for the activation or repression of other genes (referred to as target genes of the corresponding TFs) (Gupta et al., 2021). The mechanistic basis of TF function lies in their ability to form protein-DNA interactions (PDI) by recognizing specific *cis*-regulatory elements (CREs) located near or distant from their target genes. Such interactions guide the recruitment of the transcriptional machinery. The collection of TFs and their corresponding target genes constitutes a gene regulatory network (GRN). In plants, as in other organisms, the structure of these GRNs determines spatiotemporal gene expression patterns (Swift and Coruzzi, 2017). Consequently, the wiring of a GRN has implications for phenotypic variation (Deplancke et al., 2016), plant responses to abiotic and biotic stress (Nakashima et al., 2014; Birkenbihl et al., 2017), speciation (Mack and Nachman, 2017), adaptation, and diversification (Mack and Nachman, 2017; Bowles et al., 2020), highlighting and justifying any effort to understand its structure and dynamics.

CRE sequence variation, primarily located in the non-coding regions of the genome, drives rewiring changes in GRNs (Sullivan et al., 2014). Single-nucleotide polymorphisms (SNPs) and small insertions/deletions within CREs can affect TF binding affinity, altering the interaction between TFs and their corresponding CRE (Marand et al., 2023). However, transposable elements

(TEs), which are highly abundant in non-coding sequences (Bennetzen et al., 2005) and constitute up to 85% of the plant genome, such as maize (Schnable et al., 2009), are among the major contributors of genomic variability. TEs can impact gene function through various mechanisms, such as gene inactivation, gene expression reprogramming, deletions, rearrangements, gene transposition, and protein exaptation (Lisch, 2013; Schmitz et al., 2022). In terms of expression variation, TEs can induce gene expression reprogramming by inserting, removing, or establishing new regulatory connections (Greene et al., 1994; Butelli et al., 2012). Moreover, TE insertions can modify the epigenetic landscape surrounding a gene, leading to changes in gene expression through chromatin modifications.

1.2 GRN CHARACTERIZATION

Wet lab approaches. Approaches to establish PDI can be categorized as gene-centered and TF-centered methods, which correspond to strategies focused on identifying TF regulators for specific genes and target genes for specific TFs, respectively (Arda and Walhout, 2010; Mejia-Guerra et al., 2012; Yang et al., 2017). The yeast one-hybrid (Y1H) assay and the electrophoretic mobility shift assay (EMSA) are frequently employed gene-centered methods (Arda and Walhout, 2010; Yang et al., 2016). Among the diverse array of TF-centered strategies, Chromatin Immunoprecipitation Sequencing (ChIP-seq) is a highly utilized assay for the identification of TF binding sites (TFBS) *in vivo*. Variations of ChIP-seq include Cleavage Under Targets and Release Using Nuclease (CUT&RUN) (Skene and Henikoff, 2017) and Cleavage Under Targets and Tagmentation (CUT&Tag) (Kaya-Okur et al., 2019), which overcome challenges associated with crosslinking and solubilization. These methods also require minimal sample material, offering significant advantages in experimental applications. Within the *in vitro* techniques, systematic evolution of ligands by exponential enrichment (SELEX), protein binding microarrays (PBM), and

DNA affinity purification sequencing (DAP-seq) are within the most widely used methods (Yang et al., 2016; O'Malley et al., 2016). Limitations to consider for EMSA and ChIP-seq include restrictions on the number of sequences and TFs that can be tested, respectively. Additionally, ChIP-seq captures numerous indirect binding events, making it challenging to identify direct targets. Similarly, DAP-seq, SELEX, and PBM can produce a high number of non-functional PDIs, primarily due to the lack of a native chromatin environment (Yang et al., 2016). Therefore, TF-target gene associations determined by these methods always require further experimental validation.

Given the inherent presence of false positives and the large number of interactions obtained through these experimental approaches, complementary analyses have been employed to identify high-confidence TF-target gene associations. The most widely used strategy is the identification of differentially expressed genes (DEGs) - after the perturbation of the corresponding TF - which identifies downstream genes affected by the perturbation of the corresponding TF. The perturbation itself also recovers a large number of indirect changes, such as cellular responses associated with the perturbation itself. However, the combination of PDI and DEG analyses allows for the identification and differentiation between direct target genes and indirect effects of the perturbation, respectively. Shockingly, this approach has shown that the overlap between DEGs and PDIs is overall low and may vary between 5-30% (Zeller et al., 2006; Morohashi and Grotewold, 2009; Morohashi et al., 2012; Eveland et al., 2014; Liu et al., 2015), indicating that a large fraction of the PDI may not lead to expression changes of the corresponding target genes. In yeast, the low fraction of overlapping DEG and PDI was associated with paralog TFs backing-up the function of knocked-out TFs (Gitter et al., 2009). This phenomenon has not yet been investigated in the context of plants. As an alternative to perturbation analysis, the identification

of co-expression networks has gained significant attention for narrowing down target genes to those that exhibit high co-expression with the corresponding TF (Eisen et al., 1998; Allocco et al., 2004; Vandepoele et al., 2009; Haynes et al., 2013; Wu and Ji, 2013; Angelini and Costa, 2014; Jiang and Mortazavi, 2018; Haque et al., 2019). Thus, the integration of DEGs under TF perturbations and co-expression networks provides opportunities to improve predictions obtained from experiments like DAP-seq. This approach is particularly valuable for systems in which ChIP-seq presents technical challenges, such as to generate mutants or antibodies for the corresponding TF. Additionally, it also allows for scalability in the number of TFs that can be tested (O'Malley et al., 2016; Ricci et al., 2019).

Numerous systematic and genome-wide endeavors have led to the discovery of millions of PDIs in various model organisms (Harbison et al., 2004; Deplancke et al., 2006; Zhu et al., 2009; Gerstein et al., 2010; Consortium et al., 2010; Négre et al., 2011; ENCODE Project Consortium, 2012). In the case of plants, specifically *Arabidopsis thaliana* and maize (*Zea maize* L.), similar efforts have been undertaken on a smaller scale and within specific biological contexts. These include the regulation of the root stele (Brady et al., 2011), secondary cell wall synthesis (Taylor-Teeples et al., 2015), phenolic metabolism (Yang et al., 2017), flower development (Chen et al., 2018), as well as responses to ABA (Song et al., 2016) and nitrogen (Gaudinier et al., 2018) among others. It is also noteworthy to highlight the significant contributions made in the identification of TF binding motifs (TFBMs) for over 640 TFs in Arabidopsis (O'Malley et al., 2016; Weirauch et al., 2014; Franco-Zorrilla et al., 2014) and more than 30 TFs in maize (Ricci et al., 2019; Galli et al., 2018). Invaluable source of information for the construction of regulatory models based on multi-omic data integrations (Song et al., 2020; Pérez et al., 2023).

Computational approaches. Technological advances in RNA sequencing have allowed the generation of thousands of expression samples, enabling the implementation of methods for TF-target inference. All these methods utilize the idea of identifying co-expressed genes as a means of inferring regulation without prior knowledge of the regulatory network. Among the various forms of co-expression, the most commonly employed approach is the inference of gene regulatory networks (GRNs) through the analysis of expression variations in spatial (e.g., different organs), temporal (e.g., developmental trajectory), perturbation, or genetic background contexts (Haque et al., 2019; Zhou et al., 2020). In all scenarios, the construction of a co-expression network involves three key steps: data processing and normalization, network reconstruction, and network evaluation (Haque et al., 2019; Johnson and Krishnan, 2022). While all three steps are important, the reconstruction method is particularly critical due to the constraints/assumptions it imposes on the network and the ability to differentiate between association and causation associations (Haque et al., 2019). The strategies for network reconstruction can be classified into four categories, including correlation and information-theoretic approaches, Boolean network approaches, Bayesian network approaches, and regression and differential equation-based models (Banf and Rhee, 2017). Each approach has its strengths and limitations, especially when considering the network's scale and the number of samples. However, common practices to enhance their strength and reduce limitations include restricting tested interactions, incorporating known interactions to improve threshold identification during the prediction process, and utilizing background models based on randomly assigned expression datasets (Banf and Rhee, 2017).

1.3 COMBINATORIAL GENE REGULATION (CGR)

A defining characteristic of GRNs is their combinatorial nature, where a single TF can regulate multiple sets of target genes through interactions with other proteins. These interactions can be

direct or indirect, for example mediated by DNA, and involve multiple regulatory proteins. This phenomenon is known as combinatorial gene regulation (CGR). From a practical perspective, CGR presents unique challenges for the prediction or identification of transcriptional regulation of specific processes, as a single TF may be linked to multiple processes. Additionally, multiple TFs may be linked to the same process simultaneously. Consequently, CGR contributes to the expansion and diversification of the regulatory repertoire of TFs (Reményi et al., 2004; Brkljacic and Grotewold, 2017). At the molecular level, implications of CGR include that TFs may form different protein complexes and/or bind to DNA in modular fashion to *cis*-regulatory modules (CRMs) (Brkljacic and Grotewold, 2017). In general, TF binding to a CRM can be categorized into three models: independent binding, competitive binding, and cooperative binding. In independent binding, TFs bind to separate CREs without any physical interaction between them. Competitive binding occurs when different TFs compete for the binding of the same CREs, potentially involving physical interactions. Cooperative binding, on the other hand, requires the formation of a TF complex to bind a CRE (Reiter et al., 2017; Colinas and Goossens, 2018). Major advances has been main to understand the molecular mechanisms behind the CGR, including TFs spatiotemporal expression variation, TFs post-translational modification, splicing of different TFs isoforms, TF conformational changes trigger by the interaction of small molecules, as well as histone modifications and chromatin structure (Brkljacic and Grotewold, 2017; Reiter et al., 2017). However, there is currently no single model that comprehensively predicts the CGR landscape of a gene or biological process, i.e., the combination of TFs that may exert control over the corresponding gene or biological process.

1.4 UNRAVELING GENE REGULATION: THE ROLE OF MULTI-OMICS

The incorporation of diverse genomic information has enhanced the accuracy of GRN models (Qian and Huang, 2020). Common sources of information include accessible chromatin regions (ACRs), histone marks, and DNA methylation patterns, enabling the construction of cell/tissue/condition-specific regulatory circuits (ENCODE Project Consortium, 2012; Neph et al., 2012; Baur et al., 2020). The integration of additional layers of information offers several advantages, such as uncovering novel regulatory principles and the identification of new combinations of *cis*-regulatory elements (i.e., novel CRMs) (Neph et al., 2012; Sullivan et al., 2014). Furthermore, the inclusion of protein-protein interactions (PPIs) between TFs and their corresponding PDIs associates highly connected TFs with stronger expression effects (Gerstein et al., 2012). Additionally, genes targeted by multiple TFs exhibit broader expression windows and at the same time collection of co-binding events enables the identification of TF complexes (Heyndrickx et al., 2014). These co-binding events have demonstrated specificity to particular biological processes, such as development-specific gene expression patterns (Chen et al., 2018). In addition to the PDI-related datasets, the construction of GRNs based on transcriptomics and proteomics has shown to complement each other, recovering more interactions together than individually when compared to GRNs built from ChIP-seq assays (Walley et al., 2016). Similarly, the integration of multiple layers of genomic information, ranging from chromatin to translation changes, enables the identification of species- and layer-specific responses to submergence, as well as the CRE architectures responsible for submergence-induced expression changes (Reynoso et al., 2019). The inclusion of marks that capture epigenetic features, along with chromatin accessibility (Yan et al., 2019) and chromatin interaction data (Ricci et al., 2019), enables the identification of development-specific enhancers and the association of distal ACR with target

genes through the formation of chromatin loops. Additionally, the incorporation of genetics, i.e. molecular trait information at population level, has demonstrated an outstanding potential to uncover the molecular mechanisms underlying complex traits (Li et al., 2013; Wen et al., 2014, 2016; Mizrachi et al., 2017; Kremling et al., 2018; Zhou et al., 2019; Mazaheri et al., 2019; Shrestha et al., 2022). Notably, with a few exceptions (Yang et al., 2022; Schaefer et al., 2018; Mizrachi et al., 2017), the integration of data through the identification of common features or patterns have been a common denominator in the described studies. However, significant progress has been made in other systems (Lee et al., 2019; Krassowski et al., 2020; Subramanian et al., 2020; Qian and Huang, 2020; Kang et al., 2022; Vahabi and Michailidis, 2022), and approaches used in these studies still need to be tested in plant systems.

1.5 WORKING SYSTEM AND CHAPTERS DISTRIBUTION

Analyzing multi-omics data has unveiled valuable insights in gene regulation, yet also introduced unique challenges. My research addresses some of these challenges by implementing and establishing strategies to integrate multiple-omic data and predicting GRNs, uncovering the regulatory circuits associated with specific plant biological processes. Specifically, I focused on predicting GRNs involved in the regulation of lipid metabolism in *Camelina Sativa* (Chapter 2), as well as other biological processes in Maize (*Zea mays*) (Chapter 3), and *Arabidopsis thaliana* (Chapter 4), using computational techniques. Due to the unique characteristics and data availability of each species, I have developed customized strategies for their analysis.

Camelina sativa, is a winter oilseed annual crop, member of the *Brassicaceae* family. *Camelina* oilseed crop that has gained attention for its potential use in biofuel production (Bansal and Durrett, 2016; Carlsson, 2009). However, despite its growing popularity, the available gene expression datasets for Camelina are limited to a few tens of samples (Gomez-Cano et al., 2020),

which is ~150 and ~500 times less than the data available for maize and Arabidopsis, respectively. Thus, I used co-expression-based prediction with hard filters to build a GRN associated with the control of lipid metabolism. My work represented the first lipid-related GRN described for Camelina.

Maize is one of the most widely grown cereal crops in the world, its grain and maize stover is a source of biomass for liquid fuel and is also extensively used as a major forage component (Khan et al., 2015; Trivedi et al., 2015). Unlike Camelina, maize has a wealth of genomic and genetic resources, which favor the generation of regulatory models based on more sophisticated strategies. Specifically, I used several multi-omic datasets to build multiple molecular networks, which were integrated using three different approaches. After a systematic evaluation of the integrations, I selected the best method to describe TF regulators of a diverse set of biological processes in maize. These resources are crucial for guiding the design of future experiments and laying the foundation for integrating multi-omic datasets in maize and other plant systems.

Arabidopsis, like Camelina, belongs to the Brassicaceae family and is one of the most extensively studied plant species. This makes Arabidopsis an appealing system for exploring the co-expression relationships between TFs and their experimentally identified target genes. Leveraging the vast collection of expression and PDI datasets in Arabidopsis, I uncovered previously unknown combinations of TFs that contribute to the regulation of diverse biological processes. These findings carry significant implications for the empirical understanding of complex gene regulatory networks, the function of transcription factors, and the significance of co-expression in protein-protein and protein-DNA interactions.

REFERENCES

Allocco, D.J., Kohane, I.S., and Butte, A.J. (2004). Quantifying the relationship between co-expression, co-regulation and gene function. BMC Bioinformatics **5**: 18.

Angelini, C. and Costa, V. (2014). Understanding gene regulatory mechanisms by integrating ChIP-seq and RNA-seq data: Statistical solutions to biological problems. Frontiers in Cell and Developmental Biology **2**: 1–8.

Arda, H.E. and Walhout, A.J.M. (2010). Gene-centered regulatory networks. Brief. Funct. Genomics **9**: 4–12.

Banf, M. and Rhee, S.Y. (2017). Computational inference of gene regulatory networks: Approaches, limitations and opportunities. Biochim. Biophys. Acta Gene Regul. Mech. **1860**: 41–52.

Bansal, S. and Durrett, T.P. (2016). Camelina sativa: An ideal platform for the metabolic engineering and field production of industrial lipids. Biochimie **120**: 9–16.

Bar-On, Y.M., Phillips, R., and Milo, R. (2018). The biomass distribution on Earth. Proc. Natl. Acad. Sci. U. S. A. **115**: 6506–6511.

Baur, B., Shin, J., Zhang, S., and Roy, S. (2020). Data integration for inferring context-specific gene regulatory networks. Curr Opin Syst Biol **23**: 38–46.

Bennetzen, J.L., Ma, J., and Devos, K.M. (2005). Mechanisms of recent genome size variation in flowering plants. Ann. Bot. **95**: 127–132.

Birkenbihl, R.P., Liu, S., and Somssich, I.E. (2017). Transcriptional events defining plant immune responses. Curr. Opin. Plant Biol. **38**: 1–9.

Bowles, A.M.C., Bechtold, U., and Paps, J. (2020). The Origin of Land Plants Is Rooted in Two Bursts of Genomic Novelty. Curr. Biol. **30**: 530–536.e2.

Brady, S.M. et al. (2011). A stele-enriched gene regulatory network in the Arabidopsis root. Mol. Syst. Biol. **7**: 1–9.

Brkljacic, J. and Grotewold, E. (2017). Combinatorial control of plant gene expression. Biochim. Biophys. Acta **1860**: 31–40.

Butelli, E., Licciardello, C., Zhang, Y., Liu, J., Mackay, S., Bailey, P., Reforgiato-Recupero, G., and Martin, C. (2012). Retrotransposons control fruit-specific, cold-dependent accumulation of anthocyanins in blood oranges. Plant Cell **24**: 1242–1255.

Carlsson, A.S. (2009). Plant oils as feedstock alternatives to petroleum - A short survey of potential oil crop platforms. Biochimie **91**: 665–670.

Chen, D., Yan, W., Fu, L.Y., and Kaufmann, K. (2018). Architecture of gene regulatory networks controlling flower development in Arabidopsis thaliana. Nat. Commun. **9**:1–13.

Colinas, M. and Goossens, A. (2018). Combinatorial Transcriptional Control of Plant Specialized Metabolism. Trends Plant Sci. **23**: 324–336.

Consortium, M. et al. (2010). Identification of Functional Elements and Regulatory Circuits by Drosophila modENCODE. Science **330**: 1787–1797.

Deplancke, B. et al. (2006). A gene-centered C. elegans protein-DNA interaction network. Cell **125**: 1193–1205.

Deplancke, B., Alpern, D., and Gardeux, V. (2016). The Genetics of Transcription Factor DNA Binding Variation. Cell **166**: 538–554.

Eisen, M.B., Spellman, P.T., Brown, P.O., and Botstein, D. (1998). Cluster analysis and display of genome-wide expression patterns. Proc. Natl. Acad. Sci. U. S. A. **95**: 14863–14868.

ENCODE Project Consortium (2012). An integrated encyclopedia of DNA elements in the human genome. Nature **489**: 57–74.

Eveland, A.L. et al. (2014). Regulatory modules controlling maize inflorescence architecture. Genome Res. **24**: 431–443.

Franco-Zorrilla, J.M.M., López-Vidriero, I., Carrasco, J.L.L., Godoy, M., Vera, P., and Solano, R. (2014). DNA-binding specificities of plant transcription factors and their potential to define target genes. Proc. Natl. Acad. Sci. U. S. A. **111**: 2367–2372.

Galli, M., Khakhar, A., Lu, Z., Chen, Z., Sen, S., Joshi, T., Nemhauser, J.L., Schmitz, R.J., and Gallavotti, A. (2018). The DNA binding landscape of the maize AUXIN RESPONSE FACTOR family. Nat. Commun. **9**: 4526.

Gaudinier, A. et al. (2018). Transcriptional regulation of nitrogen-associated metabolism and growth. Nature **563**: 259–264.

Gerstein, M.B. et al. (2012). Architecture of the human regulatory network derived from ENCODE data. Nature **489**: 91–100.

Gerstein, M.B. et al. (2010). Integrative Analysis of the Caenorhabditis elegans Genome by the modENCODE Project. Science **330**: 1775–1787.

Gitter, A., Siegfried, Z., Klutstein, M., Fornes, O., Oliva, B., Simon, I., and Bar-joseph, Z. (2009). Backup in gene regulatory networks explains differences between binding and knockout results. Mol. Syst. Biol. **5**: 1–7.

Gomez-Cano, F., Carey, L., Lucas, K., García Navarrete, T., Mukundi, E., Lundback, S., Schnell, D., and Grotewold, E. (2020). CamRegBase: a gene regulation database for the biofuel crop, Camelina sativa. Database **2020**.

Greene, B., Walko, R., and Hake, S. (1994). Mutator insertions in an intron of the maize knotted1 gene result in dominant suppressible mutations. Genetics **138**: 1275–1285.

Gupta, O.P., Deshmukh, R., Kumar, A., Singh, S.K., Sharma, P., Ram, S., and Singh, G.P. (2021). From gene to biomolecular networks: a review of evidences for understanding complex biological function in plants. Curr. Opin. Biotechnol. **74**: 66–74.

Haque, S., Ahmad, J.S., Clark, N.M., Williams, C.M., and Sozzani, R. (2019). Computational prediction of gene regulatory networks in plant growth and development. Curr. Opin. Plant Biol. **47**: 96–105.

Harbison, C.T. et al. (2004). Transcriptional regulatory code of a eukaryotic genome. Nature **431**: 99–104.

Haynes, B.C., Maier, E.J., Kramer, M.H., Wang, P.I., Brown, H., and Brent, M.R. (2013). Mapping functional transcription factor networks from gene expression data. Genome Res. **23**: 1319–1328.

Heyndrickx, K.S., Van de Velde, J., Wang, C., Weigel, D., and Vandepoele, K. (2014). A functional and evolutionary perspective on transcription factor binding in Arabidopsis thaliana. Plant Cell **26**: 3894–3910.

Jiang, S. and Mortazavi, A. (2018). Integrating ChIP-seq with other functional genomics data. Brief. Funct. Genomics **17**: 104–115.

Johnson, K.A. and Krishnan, A. (2022). Robust normalization and transformation techniques for constructing gene coexpression networks from RNA-seq data. Genome Biol. **23**: 1.

Kang, M., Ko, E., and Mersha, T.B. (2022). A roadmap for multi-omics data integration using deep learning. Brief. Bioinform. **23**.

Kaya-Okur, H.S., Wu, S.J., Codomo, C.A., Pledger, E.S., Bryson, T.D., Henikoff, J.G., Ahmad, K., and Henikoff, S. (2019). CUT&Tag for efficient epigenomic profiling of small samples and single cells. Nat. Commun. **10**: 1930.

Khan, N.A., Yu, P., Ali, M., Cone, J.W., and Hendriks, W.H. (2015). Nutritive value of maize silage in relation to dairy cow performance and milk quality. J. Sci. Food Agric. **95**: 238–252.

Krassowski, M., Das, V., Sahu, S.K., and Misra, B.B. (2020). State of the Field in Multi-Omics Research: From Computational Needs to Data Mining and Sharing. Front. Genet. **11**: 610798.

Kremling, K.A.G., Chen, S.-Y., Su, M.-H., Lepak, N.K., Romay, M.C., Swarts, K.L., Lu, F., Lorant, A., Bradbury, P.J., and Buckler, E.S. (2018). Dysregulation of expression correlates with rare-allele burden and fitness loss in maize. Nature **555**: 520–523.

Lee, B., Zhang, S., Poleksic, A., and Xie, L. (2019). Heterogeneous Multi-Layered Network Model for Omics Data Integration and Analysis. Front. Genet. **10**: 1381.

Li, H. et al. (2013). Genome-wide association study dissects the genetic architecture of oil biosynthesis in maize kernels. Nat. Genet. **45**: 43–50.

Lisch, D. (2013). How important are transposons for plant evolution? Nat. Rev. Genet. **14**: 49–61.

Liu, S., Kracher, B., Ziegler, J., Birkenbihl, R.P., and Somssich, I.E. (2015). Negative regulation of ABA signaling by WRKY33 is critical for Arabidopsis immunity towards Botrytis cinerea 2100. Elife **4**: e07295.

Mack, K.L. and Nachman, M.W. (2017). Gene Regulation and Speciation. Trends Genet. **33**: 68–80.

Marand, A.P., Eveland, A.L., Kaufmann, K., and Springer, N.M. (2023). cis-Regulatory Elements in Plant Development, Adaptation, and Evolution. Annu. Rev. Plant Biol. **74**: 111–137.

Mazaheri, M. et al. (2019). Genome-wide association analysis of stalk biomass and anatomical traits in maize. BMC Plant Biol. **19**: 45.

Mejia-Guerra, M.K., Pomeranz, M., Morohashi, K., and Grotewold, E. (2012). From plant gene regulatory grids to network dynamics. Biochimica et Biophysica Acta (BBA) - Gene Regulatory Mechanisms **1819**: 454–465.

Mizrachi, E., Verbeke, L., Christie, N., Fierro, A.C., Mansfield, S.D., Davis, M.F., Gjersing, E., Tuskan, G.A., Van Montagu, M., Van de Peer, Y., Marchal, K., and Myburg, A.A. (2017). Network-based integration of systems genetics data reveals pathways associated with lignocellulosic biomass accumulation and processing. Proc. Natl. Acad. Sci. U. S. A. **114**: 1195–1200.

Morohashi, K. et al. (2012). A genome-wide regulatory framework identifies maize pericarp color1 controlled genes. Plant Cell **24**: 2745–2764.

Morohashi, K. and Grotewold, E. (2009). A systems approach reveals regulatory circuitry for Arabidopsis trichome initiation by the GL3 and GL1 selectors. PLoS Genet. **5**: e1000396.

Nakashima, K., Yamaguchi-Shinozaki, K., and Shinozaki, K. (2014). The transcriptional regulatory network in the drought response and its crosstalk in abiotic stress responses including drought, cold, and heat. Front. Plant Sci. **5**: 1–7.

Négre, N. et al. (2011). A cis-regulatory map of the Drosophila genome. Nature **471**: 527–531.

Neph, S. et al. (2012). An expansive human regulatory lexicon encoded in transcription factor footprints. Nature **489**: 83–90.

O'Malley, R.C., Huang, S.S.C., Song, L., Lewsey, M.G., Bartlett, A., Nery, J.R., Galli, M., Gallavotti, A., and Ecker, J.R. (2016). Cistrome and Epicistrome Features Shape the Regulatory DNA Landscape. Cell **165**: 1280–1292.

Pérez, N.M., Ferrari, C., Engelhorn, J., Depuydt, T., Nelissen, H., Hartwig, T., and Vandepoele, K. (2023). MINI-AC: Inference of plant gene regulatory networks using bulk or single-cell accessible chromatin profiles. bioRxiv: 2023.05.26.542269.

Qian, Y. and Huang, S.-S.C. (2020). Improving plant gene regulatory network inference by integrative analysis of multi-omics and high resolution data sets. Current Opinion in Systems Biology **22**: 8–15.

Reiter, F., Wienerroither, S., and Stark, A. (2017). Combinatorial function of transcription factors and cofactors. Curr. Opin. Genet. Dev. **43**: 73–81.

Reményi, A., Schöler, H.R., and Wilmanns, M. (2004). Combinatorial control of gene expression. Nat. Struct. Mol. Biol. **11**: 812.

Reynoso, M.A. et al. (2019). Evolutionary flexibility in flooding response circuitry in angiosperms. Science **365**: 1291–1295.

Ricci, W.A. et al. (2019). Widespread long-range cis-regulatory elements in the maize genome. Nature Plants **5**: 1237–1249.

Schaefer, R.J., Michno, J.-M., Jeffers, J., Hoekenga, O., Dilkes, B., Baxter, I., and Myers, C.L. (2018). Integrating Coexpression Networks with GWAS to Prioritize Causal Genes in Maize. Plant Cell **30**: 2922.

Schmitz, R.J., Grotewold, E., and Stam, M. (2022). Cis-regulatory sequences in plants: Their importance, discovery, and future challenges. Plant Cell **34**: 718–741.

Schnable, P.S. et al. (2009). The B73 maize genome: complexity, diversity, and dynamics. Science **326**: 1112–1115.

Shrestha, V., Yobi, A., Slaten, M.L., Chan, Y.O., Holden, S., Gyawali, A., Flint-Garcia, S., Lipka, A.E., and Angelovici, R. (2022). Multiomics approach reveals a role of translational machinery in shaping maize kernel amino acid composition. Plant Physiol. **188**: 111–133.

Skene, P.J. and Henikoff, S. (2017). An efficient targeted nuclease strategy for high-resolution mapping of DNA binding sites. Elife **6**.

Song, L., Huang, S.S.C., Wise, A., Castanoz, R., Nery, J.R., Chen, H., Watanabe, M., Thomas, J., Bar-Joseph, Z., and Ecker, J.R. (2016). A transcription factor hierarchy defines an environmental stress response network. Science **354**.

Song, Q., Lee, J., Akter, S., Rogers, M., Grene, R., and Li, S. (2020). Prediction of condition-specific regulatory genes using machine learning. Nucleic Acids Res. **48**: e62.

Subramanian, I., Verma, S., Kumar, S., Jere, A., and Anamika, K. (2020). Multi-omics Data Integration, Interpretation, and Its Application. Bioinform. Biol. Insights **14**: 1177932219899051.

Sullivan, A.M. et al. (2014). Mapping and dynamics of regulatory DNA and transcription factor networks in A. thaliana. Cell Rep. **8**: 2015–2030.

Swift, J. and Coruzzi, G.M. (2017). A matter of time - How transient transcription factor

interactions create dynamic gene regulatory networks. Biochimica et Biophysica Acta (BBA) - Gene Regulatory Mechanisms **1860**: 75–83.

Taylor-Teeples, M. et al. (2015). An Arabidopsis gene regulatory network for secondary cell wall synthesis. Nature **517**: 571–575.

Trivedi, P., Malina, R., and Barrett, S.R.H. (2015). Environmental and economic tradeoffs of using corn stover for liquid fuels and power production. Energy Environ. Sci. **8**: 1428–1437.

Vahabi, N. and Michailidis, G. (2022). Unsupervised Multi-Omics Data Integration Methods: A Comprehensive Review. Front. Genet. **13**: 854752.

Vandepoele, K., Quimbaya, M., Casneuf, T., De Veylder, L., and Van de Peer, Y. (2009). Unraveling transcriptional control in Arabidopsis using cis-regulatory elements and coexpression networks. Plant Physiol. **150**: 535–546.

Walley, J.W., Sartor, R.C., Shen, Z., Schmitz, R.J., Wu, K.J., Urich, M.A., Nery, J.R., Smith, L.G., Schnable, J.C., Ecker, J.R., and Briggs, S.P. (2016). Integration of omic networks in a developmental atlas of maize. Science **353**: 814–818.

Weirauch, M.T. et al. (2014). Determination and inference of eukaryotic transcription factor sequence specificity. Cell **158**: 1431–1443.

Wen, W., Li, D., Li, X., Gao, Y., Li, W., Li, H., Liu, J., Liu, H., Chen, W., Luo, J., and Yan, J. (2014). Metabolome-based genome-wide association study of maize kernel leads to novel biochemical insights. Nat. Commun. **5**: 3438.

Wen, W., Liu, H., Zhou, Y., Jin, M., Yang, N., Li, D., Luo, J., Xiao, Y., Pan, Q., and Tohge, T. (2016). Combining quantitative genetics approaches with regulatory network analysis to dissect the complex metabolism of the maize kernel. Plant Physiol. **170**: 136–146.

Wu, G. and Ji, H. (2013). ChIPXpress: Using publicly available gene expression data to improve ChIP-seq and ChIP-chip target gene ranking. BMC Bioinformatics **14**.

Yang, F. et al. (2017). A Maize Gene Regulatory Network for Phenolic Metabolism. Mol. Plant **10**: 498–515.

Yang, F., Ouma, W.Z., Li, W., Doseff, A.I., and Grotewold, E. (2016). Establishing the Architecture of Plant Gene Regulatory Networks. Methods Enzymol. **576**: 251–304.

Yang, Z., Xu, G., Zhang, Q., Obata, T., and Yang, J. (2022). Genome-wide mediation analysis: an empirical study to connect phenotype with genotype via intermediate transcriptomic data in maize. Genetics **221**.

Yan, W., Chen, D., Schumacher, J., Durantini, D., Engelhorn, J., Chen, M., Carles, C.C., and Kaufmann, K. (2019). Dynamic control of enhancer activity drives stage-specific gene expression during flower morphogenesis. Nat. Commun. **10**: 1705.

Zeller, K.I. et al. (2006). Global mapping of c-Myc binding sites and target gene networks in human B cells. Proceedings of the National Academy of Sciences **103**: 17834.

Zhou, P., Li, Z., Magnusson, E., Gomez Cano, F., Crisp, P.A., Noshay, J.M., Grotewold, E., Hirsch, C.N., Briggs, S.P., and Springer, N.M. (2020). Meta Gene Regulatory Networks in Maize Highlight Functionally Relevant Regulatory Interactions. Plant Cell **32**: 1377–1396.

Zhou, S., Kremling, K.A., Bandillo, N., Richter, A., Zhang, Y.K., Ahern, K.R., Artyukhin, A.B., Hui, J.X., Younkin, G.C., Schroeder, F.C., Buckler, E.S., and Jander, G. (2019). Metabolome-Scale Genome-Wide Association Studies Reveal Chemical Diversity and Genetic Control of Maize Specialized Metabolites. Plant Cell **31**: 937–955.

Zhu, C. et al. (2009). High-resolution DNA-binding specificity analysis of yeast transcription factors. Genome Res. **19**: 556–566.

CHAPTER TWO: CAMREGBASE: A GENE REGULATION DATABASE FOR BIOFUEL CROP, *CAMELINA SATIVA*[1]

[1]This chapter has been published in the following manuscript:

Gomez-Cano F., Carey L., Lucas K., García Navarrete T., Mukundi E., Lundback S., Schnell S., Grotewold E., (2020), CamRegBase: a gene regulation database for the biofuel crop, *Camelina sativa*, Database, baaa075, https://doi.org/10.1093/database/baaa075

2.1 ABSTRACT

Camelina is an annual oilseed plant from the Brassicaceae family that is gaining momentum as a biofuel winter cover crop. However, a significant limitation in further enhancing its utility as a producer of oils that can be used as biofuels, jet fuels or bio-based products is the absence of a repository for all the gene expression and regulatory information that is being rapidly generated by the community. Here, we provide CamRegBase (https://camregbase.org/) as a one-stop resource to access Camelina information on gene expression and co-expression, transcription factors, lipid associated genes and genome wide orthologs in the close-relative reference plant Arabidopsis. We envision this as a resource of curated information for users, as well as a repository of new gene regulation information.

2.2 INTRODUCTION

Camelina sativa is an emerging biofuel crop (Carlsson, 2009; Iskandarov et al., 2014). With a low economic input requirement (Iskandarov et al., 2014), early season growth habit (Allen et al., 2014; Chaturvedi et al., 2018), genetic similarity to the model plant Arabidopsis (Liang et al., 2013) and relatively high oil composition in the seed (Moser, 2010; Berti et al., 2016), it has gained traction as a potential target for jet fuel and biodiesel production. Camelina's genome has been sequenced and annotated, has a hexaploid genome structure harboring 89 418 protein-coding genes organized in 20 chromosomes (Liang et al., 2013; Kagale et al., 2014) and is relatively easy to genetically transform (Liu et al., 2012).

A challenge, albeit not unique to Camelina, is how to best utilize the burgeoning genomic information for predictive metabolic engineering of seed oil production (Chappell and Grotewold, 2008; Grotewold, 2008). Clearly, knowing how much and where gene expression takes place is necessary, as recently demonstrated by recent studies aimed at increasing oil production in

Camelina using the co-expression of select genes (Chhikara et al., 2018). While RNA-Seq is a very powerful tool to determine global levels of gene expression, each analysis yields a large amount of data and therefore is non-trivial to curate and analyze for potential targets. To take advantage of all the currently available RNA-Seq data for Camelina, a relational database is the most ideal resource. Currently, Camelina genomics resources are part of the Brassica database BRAD (http://brassicadb.org) together with 11 Brassicaceae genomes. BRAD has a comparative approach to make plots of syntenic genomic regions and search the orthologs genes (Wang et al., 2015), but the most information available in BRAND is directed to *Brassica rapa*. In particular, the Camelina Genome Portal (camelinagenomics.org) allows a user to browse the whole Camelina genome assembly, conduct BLAST analyses to the Camelina genome, and view any of 15,946 (current number at date of publication) contig scaffolds on the sequenced genome. The University of Toronto has developed an electronic fluorescent pictograph browser (http://bar.utoronto.ca/) for Camelina sativa, which allows quick visual representation of expression data from a large developmental set. The Camelina Genomic Resources (camelinagenome.org) contains transcript data on protein and lipids but is only restricted to the developing embryo. Many databases also exist that provide information on TFs for one or multiple plants (Davuluri et al., 2003; Guo et al., 2005; Gao et al., 2006; Guo et al., 2008; Rushton et al., 2008; Wang et al., 2010; Yilmaz et al., 2009; Kagale et al., 2016). AGRIS (https://agris-knowledgebase.org/), for example, provides a useful resource for the knowledgebase described here, because it provides a comprehensive collection of Arabidopsis TFs and other regulatory components, that can be easily translated to Camelina based on the close relationship between these plants. Here, we introduce the Camelina Gene Regulation Database (https://camregbase.org/), which is intended as a one-stop resource for aspects related to Camelina gene regulation. CamRegBase v1.0 harbors all RNA-Seq experiments

available to-date with read abundance and the corresponding metadata, tissue-specific gene expression visualization and gene co-expression analyses. Additionally, CamRegBase 1.0 offers information on the orthologous relationships between Camelina and Arabidopsis genes along with the reported syntelog data (Kagale et al., 2016). Finally, as a valuable resource to researchers interested in studying the control of gene regulation, CamRegBase 1.0 provides a comprehensive catalog of transcription factors (TFs) and co-activators identified by our own analyses and those previously reported (Kagale et al., 2016) (http://planttfdb.cbi.pku.edu.cn/). With all the above-mentioned information integrated as one resource, CamRegBase is poised to become a primary resource for Camelina gene expression analyses.

2.3 RESULTS AND DISCUSSION

2.3.1 Database structure

The utilization of the open source Tripal toolkit for the construction of the database web portal ensures that it can be expanded by the addition of compatible extension modules, and it ensures interoperability with a number of widely used biological knowledgebases (Spoor et al., 2019). The overall database organization is schematized in Figure 2.1, with the search functionality of the site relying on the underlying database tables shown in the entity relationship diagram. All the records in Drupal are stored in the 'node' table, which is queried in relation to the other tables on the search term provided by the end user. The lines shown in the diagram show how the tables are related when a search is run. For example, when a search is run using the Gene Search page, the 'Search Data' table is queried to return data matching the search term in the 'title', 'name' or 'category' fields. That table contains a consolidation of data from the 'Node' and 'Taxonomy Data' tables along with a 'category' value based on the presence of the record in any of the 'Goslim Term', 'Aralip Pathway' and/or 'TF Family' tables. The consolidation was done to improve the

performance of the search function. Other searches query the tables directly. In the case of the Syntelogs search, the 'Homolog' is examined, and the data are returned along with related results from the 'Csa_g1', 'Csa_g2', 'Csa_g3' and 'Taxonomy Data' tables using the relationships shown in the entity relationship diagram.

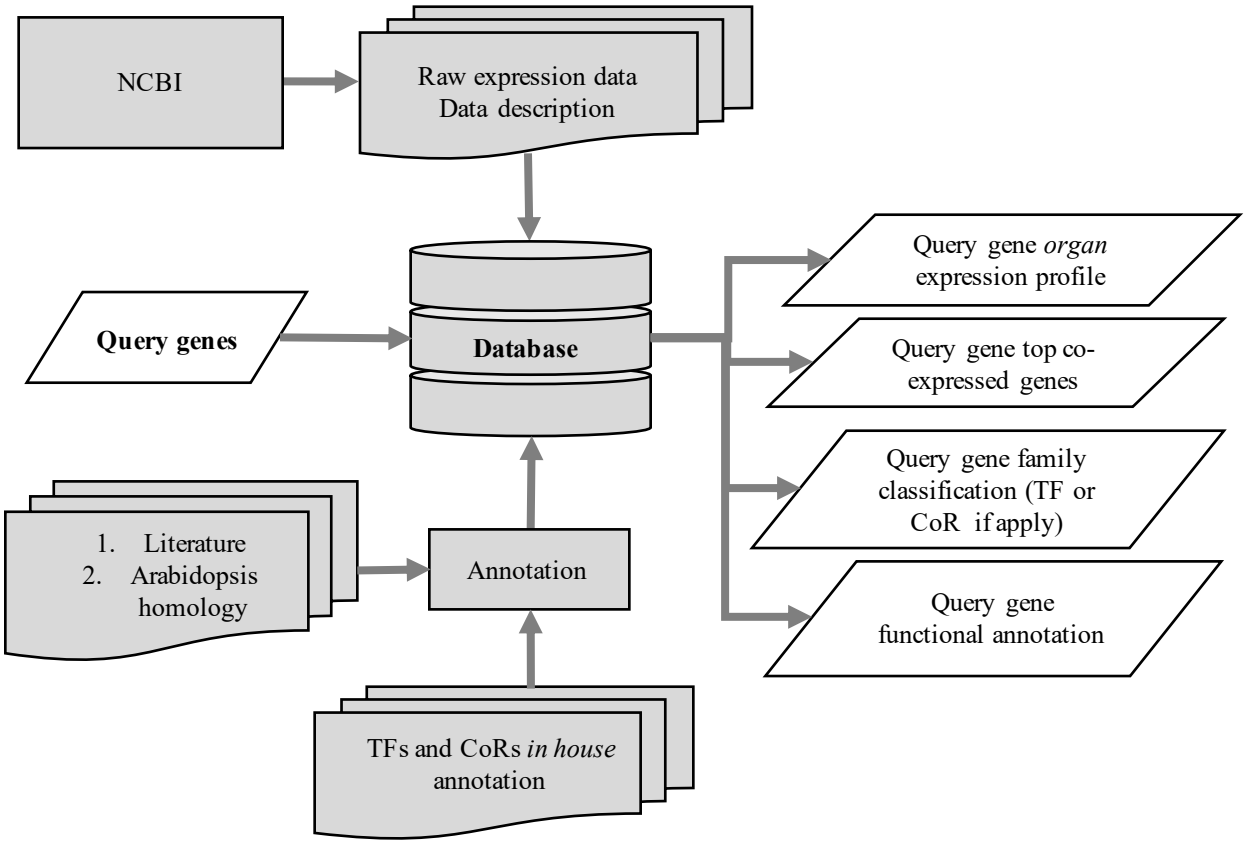

Figure 2.1 Schematic diagram outlining the architecture of CamRegBase 1.0

2.3.2 Expression database content

The gene expression database was built based on 131 publicly available Camelina RNA-seq experiments (See 'Materials and methods'). The data correspond to gene expression information from five different Camelina 'varieties', with DH55 and Suneson being the varieties with the largest number of samples (Figure 2.2a). Out of the 131 samples, 28 had no details regarding the variety and thus were labeled as unknown and utilized solely for the co-expression analyses (See below). Data were classified based on variety and further grouped on the basis of plant organs and

seed development stages. In total, we analyzed data from 12 different 'organs', including whole plant pools (referred as 'Plant'), and samples without 'organ' specification (defined as 'Unknown') (Figure 2.2b). Notably, seeds and roots represented the majority of samples available (38.8% and 16.8%, respectively) (Figure 2.2b). In terms of 'seed developmental stages', samples were analyzed that covered a range of 36 days post-anthesis (DPAs), with 14 different time points from 4 to 40 days post-anthesis. Overall, approximately four billion reads were analyzed, with an average of 29.9 million reads per sample and with 95.7% of the reads mapping to the genome.

To characterize the transcriptome at the sample level, the top 5% of genes with highest expression variation (TPMs) across all 131 samples were selected and a principal component analysis (PCA) was performed. The first two principal components explained 54.7% of the variation of the samples and allowed us to separate the 12 organs into 7 groups (Figure 2.2c). The 'root' and 'seed' samples grouped closest together and were the most distinct from the other samples. As expected, some samples aligned closely with others such as 'embryo' with 'seed' samples, 'cotyledons' with 'young leaf', and 'buds' with 'flowers' (dashed circles, Figure 2.2c). The observed separation suggests that, at least for the major groups, the data collected and presented here capture relevant biological information.

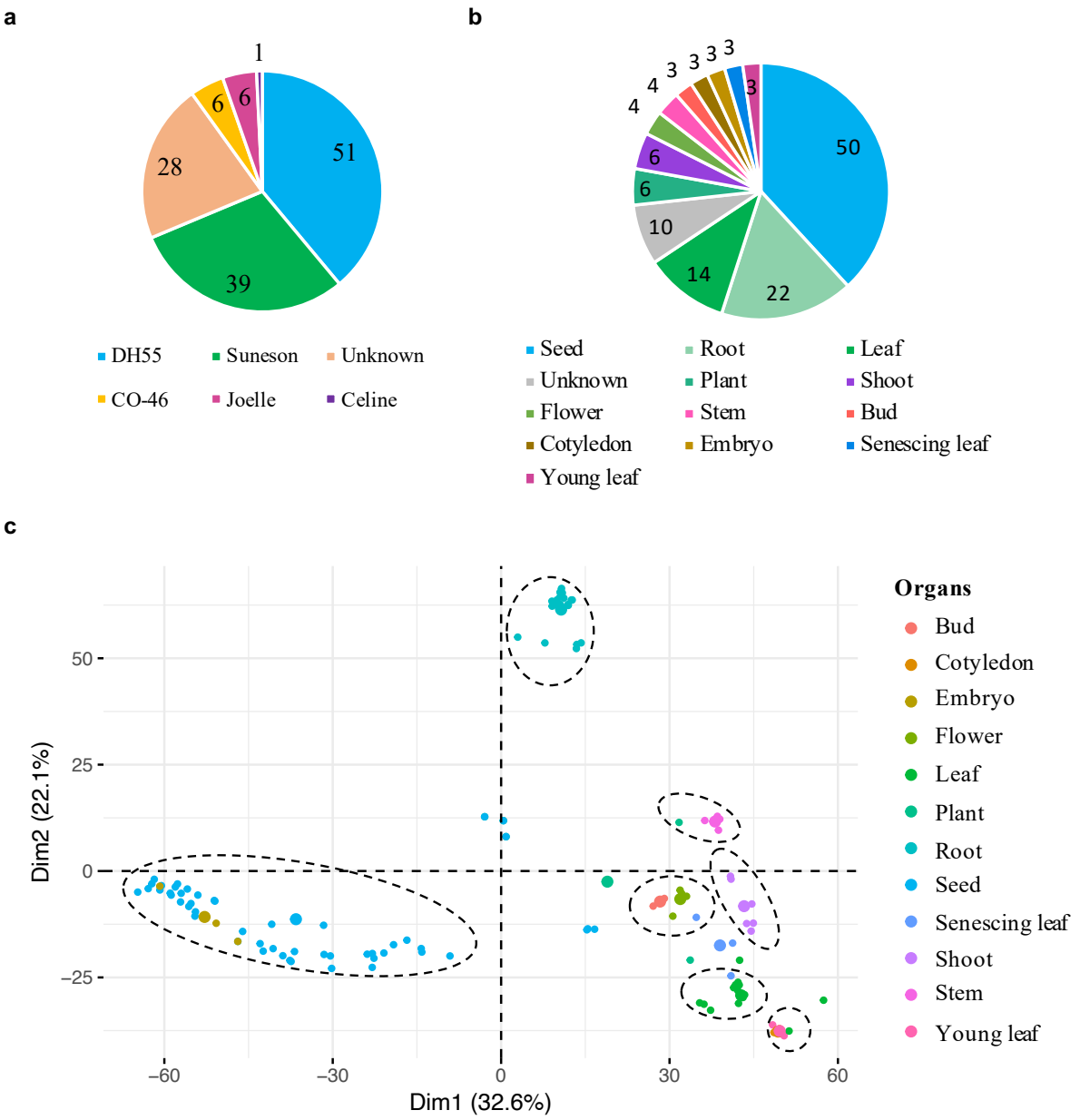

Figure 2.2 Gene expression data hosted on CamRegBase 1.0
Summary of expression data available on CamRegBase at the level of **a** Camelina varieties and **b** organ-specific samples. **c.** PCA of the Camelina transcriptome using log_2TPMs. Dotted ovals indicate major groups of samples identified by visual inspection of the PCA results.

2.3.3 Annotation of TFs

TFs and CoRs play central roles in controlling gene expression, and they provide powerful tools to manipulate developmental or metabolic pathways for biotechnological purposes (Grotewold, 2008; Gray and Grotewold, 2011). Thus, to characterize Camelina TFs and CoRs,

advantage was taken of the current literature in this regard (Kagale et al., 2016) and used to expand the previous collection using pipelines based on protein domain characterization and family classifications that worked well before in other plants (Yilmaz et al., 2009, 2011) (See 'Materials and methods'). In total, 4,619 TFs and 805 CoRs were identified of which 1,075 TFs and 793 CoRs had not been previously reported. Our analysis, however, failed to identify 971 TFs previously reported based on homology (Kagale et al., 2016). Currently, CamRegBase harbors information on 5,590 TFs classified into 81 families, and 805 CoR, classified into 25 different families (Figure 2.3).

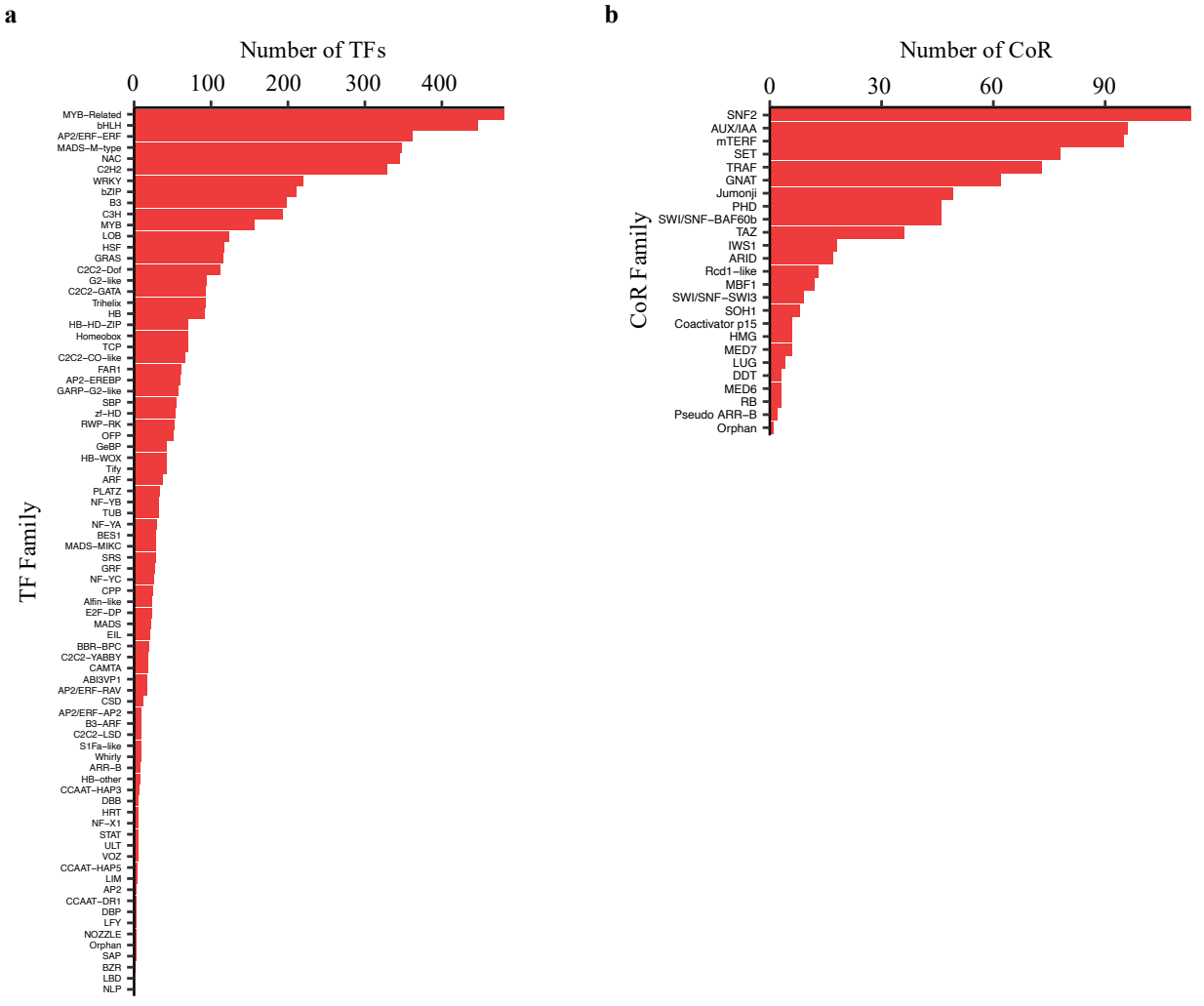

Figure 2.3 Distribution of the (a) TF and (b) CoR genes according to families as currently

Figure 2.3 (cont'd)
present in CamRegBase 1.0

2.3.4 Database functionalities

CamRegBase 1.0 consists of quick-buttons and tabs for navigation. The buttons are redundancies of the navigation tab. A unified search function within the 'Gene Search' tab was implemented, where a user may query Camelina genes by gene accession number, Arabidopsis GO Slim term or pathways from the Aralip database to find a gene of choice. Once a gene is selected, the resulting page provides gene information, a link to explore gene expression and a list of the top 50 co-regulated genes with their associated PCCs. When gene expression is explored, an expression analysis chart is displayed showing expression values across biosample numbers. Hovering over a data point will show the complete data information. Charts can also be downloaded in CSV format. Under the 'Regulation' tab a user may find a group of genes within a TF family, or they can go directly to the gene information page by searching with a Camelina gene accession number. Under the 'Gene Expression' tab, a user can go directly to gene expression information, or click on 'Heat Map' to view the selection of genes in a heat map, which can be sorted by gene name, annotation, or blast description. A drop-down selection of the samples permits to visualize just a few, or all the gene expression samples in the database. Alternatively, sample selection can also be done on the created heatmap by highlighting the desired samples; the heatmap will adjust accordingly. Finally, on the 'Syntelogs' tab, a user can query a Camelina or Arabidopsis gene accession number to see how they relate to one another.

2.4 METHODS

2.4.1 Gene expression data source

Expression data present in CamRegBase 1.0 was retrieved from the Gene Expression Omnibus. All samples collected corresponded to RNA-Seq experiments generated using the Illumina

platform. RNA-Seq results for a total of 131 experiments (including replicates) were collected, 40 of which corresponded to single-end libraries and 91 to paired-end libraries. These 131 experiments corresponded to a total of 16 different projects. All samples were subject to quality control using FastQC (http://www.bioinformatics.babraham.ac.uk/proje cts/fastqc/, V0.11.5). Libraries with adapters and reads with low quality (Phred < 20) were removed using Cutadapt (-a and -u, respectively) (http://cutadapt.readthedocs.io/ en/stable/index.html, V1.9). Clean reads were mapped to the reference genome (V2.0, http://camelinadb.ca) using HISAT2 (2.0.4) (Kim et al., 2019) with default parameters. Reads aligned to genes were counted with the R package Rsubread (V1.32.2), using default parameters and allowing multi-mapping reads (Liao et al., 2019), and the transcript abundance estimated as transcripts per kilobase million (TPM).

2.4.2 Database and web platform construction

The website sits on an Ubuntu 18.04 operating system, the current long-term support release, using a PostgreSQL database instance for backend storage and the Apache webserver for displaying pages. It was built on top of that base using the Drupal content management system with the Tripal and Tripal Analysis Expression modules along with their dependencies (Ficklin et al., 2011; Sanderson et al., 2013; Spoor et al., 2019). The data were loaded into the Chado and Drupal database schemas using importers constructed using Tripal, and custom PHP codes were written to provide the functionality seen on the site today. The software and versions currently in use are PostgreSQL (v10.12), PHP (v7.1), Apache (v2.4.41), Tripal (v3.2), Tripal Analysis Expression (v3.0) and Drupal (v7.69).

2.4.3 *Camelina sativa* gene annotation

All the functional annotations of *C. sativa* genes analyzed here, except for TFs (see below), were based on homology with Arabidopsis thaliana obtained by performing reciprocal BLAST

analyses on 'all proteins against all', and from literature (Kagale et al., 2014). The characterization and annotation of TFs and co-regulatory proteins (CoRs) assigned to the two databases harboring TFs and co-regulators (CsTFDB and CsCoTFDB, respectively) was carried out based on the identification of proteins that contain domains distinctive of these groups of proteins, as previously described (Yilmaz et al., 2009, 2011).

2.4.4 Gene regulation data collection and analysis

To identify potential TFs, we utilized already existing knowledge of known and identified TF protein domains from published literature sources, particularly AGRIS and GRASSIUS (grassius.org) (Yilmaz et al., 2009, 2011). The data obtained were used in conjunction with Pfam's Hidden Markov Models (HMM) to perform a domain search using the HMMER(v3) software against the predicted Camelina proteins sequences (Kagale et al., 2014). Hit scores were only retained if they were considered significant, where the threshold used was a gathering score greater than the reported HMM for domains that are found in the Pfam database.

Once potential TFs were identified, they were classified based on already established domain rules. The rules consist of which protein domain or domains are required for a TF to be part of a certain family. In some instances, it involves not having a specific domain or set of domains (forbidden domains) to be classified as part of the specified family. The co-regulators were classified based on rules previously established (Burdo et al., 2014). A modified version of the iTAK Perl script (Zheng et al., 2016) was utilized to assign the proteins to families based on hits obtained from the hmmscan application in the HMMER Program.

2.4.5 Gene co-expression analyses

The co-expression analyses between pairs of genes was calculated using the log2 of the TPMs as input data and the weighted Pearson correlation coefficient (PCC) as a metric for co-expression using the R

package wCorr (Version 1.9.1) (Emad and Bailey, 2017), with an optimal threshold of 0.4 to weight samples similarities.

REFERENCES

Allen, B.L., Vigil, M.F., and Jabro, J.D. (2014). Camelina growing degree hour and base temperature requirements. Agron. J. **106**: 940–944.

Berti, M., Gesch, R., Eynck, C., Anderson, J., and Cermak, S. (2016). Camelina uses, genetics, genomics, production, and management. Ind. Crops Prod. **94**: 690–710.

Burdo, B. et al. (2014). The Maize TFome--development of a transcription factor open reading frame collection for functional genomics. Plant J. **80**: 356–366.

Carlsson, A.S. (2009). Plant oils as feedstock alternatives to petroleum - A short survey of potential oil crop platforms. Biochimie **91**: 665–670.

Chappell, J. and Grotewold, E. (2008). Plant biotechnology - Predictive, green and quantitative. Curr. Opin. Biotechnol. **19**: 129–130.

Chaturvedi, S., Bhattacharya, A., Khare, S.K., and Kaushik, G. (2018). Camelina sativa: An emerging biofuel crop. In Handbook of Environmental Materials Management, C. Hussain, ed (Sringer: Switzerland), pp. 1–38.

Chhikara, S., Abdullah, H.M., Akbari, P., Schnell, D., and Dhankher, O.P. (2018). Engineering Camelina sativa (L.) Crantz for enhanced oil and seed yields by combining diacylglycerol acyltransferase1 and glycerol-3-phosphate dehydrogenase expression. Plant Biotechnol. J. **16**: 1034–1045.

Davuluri, R.V., Sun, H., Palaniswamy, S.K., Matthews, N., Molina, C., Kurtz, M., and Grotewold, E. (2003). AGRIS: Arabidopsis gene regulatory information server, an information resource of Arabidopsis cis-regulatory elements and transcription factors. BMC Bioinformatics **4**: 25.

Emad, A. and Bailey, P. (2017). wCorr: weighted correlations.–R package ver. 1.9. 1.

Ficklin, S.P., Sanderson, L.-A., Cheng, C.-H., Staton, M.E., Lee, T., Cho, I.-H., Jung, S., Bett, K.E., and Main, D. (2011). Tripal: a construction toolkit for online genome databases. Database **2011**.

Gao, G., Zhong, Y., Guo, A., Zhu, Q., Tang, W., Zheng, W., Gu, X., Wei, L., and Luo, J. (2006). DRTF: a database of rice transcription factors. Bioinformatics **22**: 1286–1287.

Gray, J. and Grotewold, E. (2011). Transcription factors, gene regulatory networks and agronomic traits. In Sustainable Agriculture and New Biotechnologies (CRC Press), pp. 65–94.

Grotewold, E. (2008). Transcription factors for predictive plant metabolic engineering: are we there yet? Curr. Opin. Biotechnol. **19**: 138–144.

Guo, A., He, K., Liu, D., Bai, S., Gu, X., Wei, L., and Luo, J. (2005). DATF: a database of

Arabidopsis transcription factors. Bioinformatics **21**: 2568–2569.

Guo, A.Y., Chen, X., Gao, G., Zhang, H., Zhu, Q.H., Liu, X.C., Zhong, Y.F., Gu, X., He, K., and Luo, J. (2008). PlantTFDB: a comprehensive plant transcription factor database. Nucleic Acids Res. **36**: D966–9.

Iskandarov, U., Kim, H.J., and Cahoon, E.B. (2014). Camelina: An emerging oilseed platform for advanced biofuels and bio-based materials. In Plants and BioEnergy, MC McCann, M.S. Buckeridge, and N.C. Carpita, eds (Springer: New York), pp. 131–140.

Kagale, S. et al. (2014). The emerging biofuel crop Camelina sativa retains a highly undifferentiated hexaploid genome structure. Nat. Commun. **5**: 1–11.

Kagale, S., Nixon, J., Khedikar, Y., Pasha, A., Provart, N.J., Clarke, W.E., Bollina, V., Robinson, S.J., Coutu, C., Hegedus, D.D., Sharpe, A.G., and Parkin, I.A.P. (2016). The developmental transcriptome atlas of the biofuel crop Camelina sativa. Plant J. **88**: 879–894.

Kim, D., Paggi, J.M., Park, C., Bennett, C., and Salzberg, S.L. (2019). Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. Nat. Biotechnol. **37**: 907–915.

Liang, C., Liu, X., Yiu, S.-M., and Lim, B.L. (2013). De novo assembly and characterization of Camelina sativa transcriptome by paired-end sequencing. BMC Genomics **14**: 146.

Liao, Y., Smyth, G.K., and Shi, W. (2019). The R package Rsubread is easier, faster, cheaper and better for alignment and quantification of RNA sequencing reads. Nucleic Acids Res. **47**: e47–e47.

Liu, X., Brost, J., Hutcheon, C., Guilfoil, R., Wilson, A.K., Leung, S., Shewmaker, C.K., Rooke, S., Nguyen, T., and Kiser, J. (2012). Transformation of the oilseed crop Camelina sativa by Agrobacterium-mediated floral dip and simple large-scale screening of transformants. In Vitro Cellular & Developmental Biology-Plant **48**: 462–468.

Moser, B.R. (2010). Camelina (Camelina sativa L.) oil as a biofuels feedstock: Golden opportunity or false hope? Lipid technology **22**: 270–273.

Rushton, P.J., Bokowiec, M.T., Laudeman, T.W., Brannock, J.F., Chen, X., and Timko, M.P. (2008). TOBFAC: the database of tobacco transcription factors. BMC Bioinformatics **9**: 53.

Sanderson, L.-A., Ficklin, S.P., Cheng, C.-H., Jung, S., Feltus, F.A., Bett, K.E., and Main, D. (2013). Tripal v1.1: a standards-based toolkit for construction of online genetic and genomic databases. Database **2013**.

Spoor, S. et al. (2019). Tripal v3: an ontology-based toolkit for construction of FAIR biological community databases. Database **2019**.

Wang, X., Wu, J., Liang, J., Cheng, F., and Wang, X. (2015). Brassica database (BRAD) version 2.0: integrating and mining Brassicaceae species genomic resources. Database **2015**.

Wang, Z., Libault, M., Joshi, T., Valliyodan, B., Nguyen, H.T., Xu, D., Stacey, G., and Cheng, J. (2010). SoyDB: a knowledge database of soybean transcription factors. BMC Plant Biol. **10**: 14.

Yilmaz, A., Mejia-Guerra, M., Kurz, K., Liang, X., Welch, L., and Grotewold, E. (2011). AGRIS: Arabidopsis Gene Regulatory Information Server, an update. Nucleic Acids Res. **39**: D1118–1122.

Yilmaz, A., Nishiyama, M.Y., Jr, Fuentes, B.G., Souza, G.M., Janies, D., Gray, J., and Grotewold, E. (2009). GRASSIUS: a platform for comparative regulatory genomics across the grasses. Plant Physiol. **149**: 171–180.

Zheng, Y. et al. (2016). iTAK: A Program for Genome-wide Prediction and Classification of Plant Transcription Factors, Transcriptional Regulators, and Protein Kinases. Mol. Plant **9**: 1667–1670.

CHAPTER THREE: EXPLORING *CAMELINA SATIVA* LIPID METABOLISM REGULATION BY COMBINING GENE CO-EXPRESSION AND DNA AFFINITY PURIFICATION ANALYSES[1]

[1]This chapter has been published in the following manuscript:

Gomez-Cano F., Chu Y.-H., Cruz-Gomez M., Abdullah H.M., Lee Y.S., Schnell D., Grotewold E. (2022), Exploring *Camelina sativa* lipid metabolism regulation by combining gene co-expression and DNA affinity purification analyses. Plant J, 110: 589-606. https://doi.org/10.1111/tpj.15682

3.1 ABSTRACT

Camelina is an annual oilseed plant that is gaining momentum as a biofuel cover crop. Understanding gene regulatory networks (GRNs) is essential to deciphering plant metabolic pathways, including lipid metabolism. Here, we take advantage of a growing collection of gene expression datasets to predict transcription factors (TFs) associated with the control of Camelina lipid metabolism. We identified ~350 TFs highly co-expressed with lipid-related genes (LRGs). These TFs are highly represented in the MYB, AP2/ERF, bZIP, and bHLH families, including a significant number of homologs of well-known Arabidopsis lipid and seed developmental regulators. After prioritizing the top 22 TFs for further validation, we identified DNA-binding sites and predicted target genes for 16 out of the 22 TFs tested using DNA affinity purification sequencing (DAP-seq). Enrichment analyses of targets supported the co-expression prediction for most TF candidates, and the comparison to Arabidopsis revealed some common themes, but also aspects unique to Camelina. Within the top potential lipid regulators, we identified CsaMYB1, CsaABI3AVP1-2, CsaHB1, CsaNAC2, CsaMYB3, and CsaNAC1 as likely involved in the control of seed fatty acid elongation; and CsaABI3AVP1-2 and CsabZIP1 as potential regulators of the synthesis and degradation of triacylglycerols (TAGs), respectively. Altogether, the integration of co-expression data and DNA-binding assays permitted us to generate a high-confidence and short list of Camelina TFs involved in the control of lipid metabolism during seed development.

3.2 INTRODUCTION

The Brassicaceae Camelina (*Camelina sativa* L. Crantz) annual plant is gaining increasing attention as a potential oilseed crop with characteristics that make it alluring as a renewable feedstock for biofuels and biobased products, among many other applications (Carlsson, 2009; Iskandarov et al., 2014). Camelina has a hexaploid genome that harbors ~90,000 genes organized

into 20 chromosomes (Liang et al., 2013; Kagale et al., 2014). When compared with Arabidopsis (*Arabidopsis thaliana*), Camelina genes were classified into three types, including syntenic orthologs (syntelogs, ~70% of all genes), tandem duplicates (~12%), and non-syntenic (~18%) genes (Kagale et al., 2014). Within the set of syntelogs (a.k.a. paralogs), 10% of them are defined as fractionated because not all the three copies are conserved (Kagale et al., 2014). Remarkably, in addition to having a low rate of fractionation, the majority of Camelina's paralogs (in the case of triplicated genes) display no significant differences in expression levels (Kagale et al., 2014). Despite the challenges imposed by its polyploid genome, extensive gene expression analyses performed on developing Camelina seeds provided a transcriptome reference for this emerging crop, which includes 26 different datasets obtained at 13 different time points during seed development, and one immediately after germination, expression data that is available at CamRegBase (Gomez-Cano et al., 2020). Yet, despite the growing collection of mRNA accumulation data, expression information from early time points during seed development, a critical stage for lipid biosynthesis (Rodríguez-Rodríguez et al., 2013; Pollard et al., 2015), is largely missing. Another important available resource in Camelina, given its biotechnological implications, is the growing list of genes associated with fatty acid (FA) and oil biosynthesis (Nguyen et al., 2013; Mudalkar et al., 2014; Abdullah et al., 2016; Gomez-Cano et al., 2020). This was to a large extent possible thanks to the close phylogenetic relationship of Camelina with Arabidopsis, reflected in the high sequence similarity of their genomes (Nikolov et al., 2019; Mandáková et al., 2019).

Camelina seeds are ~50 times larger than those of Arabidopsis, they are rich in triacylglycerols (TAGs) containing mainly long unsaturated FAs, including linoleic acid (C18:3), which are excellent sources of omega-3 FAs (Gugel and Falk, 2006; Berti et al., 2016). Depending on the

ecotype, Camelina oil may represent up to 40% of the total seed dry weight, which also contains high levels of vitamin E and antioxidants responsible for extending the lifetime of Camelina oil-containing products (Berti et al., 2016; Malik et al., 2018; Chaturvedi et al., 2019). As in other plants, Camelina TAG synthesis starts with the synthesis of FAs in plastids (Voelker and Kinney, 2001). In Camelina embryos, the maximum rate of oil synthesis is at mid-maturation ("green cotyledon"), i.e., between 14-20 days post anthesis (DPA), while the mid-point for oil deposition is around 17-18 DPA. Consistently, C18:3 reaches its highest accumulation rate at 22 DPA (Pollard et al., 2015). In addition to C18:3, Camelina TAG also contain significant amounts of very long chain FAs (VLCFA) (C20-C24) with similar accumulation rates to C18:3 (maximum rate ~ 22-24 DPA), and detected as early as 11 DPA (Pollard et al., 2015).

A growing number of Camelina genes involved in FA and TAG biosynthesis are being identified (Nguyen et al., 2013; Abdullah et al., 2016; Morineau et al., 2017; Ozseyhan et al., 2018; Neumann et al., 2021). However, Camelina transcription factors (TFs) that control the expression of the corresponding enzymatic genes remain largely unknown. In higher plants, the synthesis of FA and TAG in seeds is tightly coordinated with development. In Arabidopsis, there is a growing number of TFs involved in seed development with direct or indirect effects on FA/TAG synthesis (Le et al., 2010; Baud and Lepiniec, 2010; Leprince et al., 2016; Tian et al., 2020). Major regulators include the ABI3VP1 proteins LEAFY COTYLEDON 2 (LEC2), ABSCISIC ACID INSENSITIVE 3 (ABI3), and FUSCA3 (FUS3), which besides controlling seed development-related processes, are also positive regulators of FA/TAG synthesis (Giraudat et al., 1992; Bäumlein et al., 1994; Stone et al., 2001). Other important regulators include LEAFY COTYLEDON 1 (LEC1) and LEAFY COTYLEDON1-LIKE (L1L), which are CCAAT-HAP3 proteins (Lotan et al., 1998; Kwong et al., 2003), the basic leucine zipper 53 (bZIP53) (Alonso et

al., 2009), AGAMOUS-Like15 (AGL15) (Zheng et al., 2009), the MYB proteins MYB115, MYB118, MYB107, and MYB9 (Wang et al., 2009; Troncoso-Ponce et al., 2016; Lashbrooke et al., 2016), and the homeobox GLABRA2 (Shen et al., 2006). Also, VP1/ABSCISIC ACID INSENSITIVE3-LIKE1, 2, and 3 (VAL1, 2, 3), all members of the ABI3VP1 family, are known for their roles in repressing the seed maturation program before germination (Tsukagoshi et al., 2007; Suzuki and McCarty, 2008; Guerriero et al., 2009). Downstream of some of these developmental regulators are several TFs that modulate specific aspects of lipid metabolism, including WRINKLED1 (WRI1), which controls carbon flux from sucrose to FA biosynthesis (Cernac and Benning, 2004). WRI1 is regulated at the transcriptional level by LEC1 and MYB89, and at the post-translational level by KIN10 and TEOSINTE BRANCHED1/CYCLOIDEA/PROLIFERATING CELL FACTOR 4 (TCP4) (Li et al., 2017; Zhai et al., 2017; Kong et al., 2020; Pelletier et al., 2017). Some of these regulatory associations are conserved between species (Kong and Ma, 2018; Kong, et al., 2020; Devic and Roscoe, 2016). In Camelina, the overexpression of Arabidopsis MYB96 led to a significant increase in epicuticular and total wax (Lee et al., 2014), resembling the functions that MYB96 has in Arabidopsis under drought conditions (Seo et al., 2011). To what extent these regulatory networks are conserved between Arabidopsis and Camelina remains unknown.

Despite the close phylogenetic relationship of Arabidopsis and Camelina, they accumulate different quantities and types of seed oils (Li et al., 2006). Thus, understanding the regulatory processes associated with these differences provides opportunities for further enhancing seed oil production. Here, we describe the use of gene co-expression analyses to identify several TF candidates associated with the regulation of lipid biosynthesis in Camelina. These predictions were confirmed using DNA affinity purification followed by sequencing (DAP-seq) analysis of the

corresponding TF candidates. Altogether, we identify and associate different TF candidates with specific aspects of the lipid-related process, including key players in regulating lipid accumulation during seed development in Camelina.

3.3 RESULTS

3.3.1 Expression analysis of genes involved in lipid accumulation during Camelina seed development

To complement the sparse gene expression information available for early stages of Camelina seed development, we collected seeds from Suneson Camelina plants at 5, 8, and 11 days post-anthesis (DPA). The sampling was performed for three biological replicates and RNA was extracted from seeds at the corresponding developmental stages and then used to perform RNA-seq analyses (see *Methods*). To characterize the expression of genes involved in lipid metabolism, we first collated lipid-related genes (LRGs) from CamRegBase (https://camregbase.org/) (Gomez-Cano et al., 2020) and classified them according to the information provided by AraLip (http://aralip.plantbiology.msu.edu/) (Li-Beisson et al., 2013). In accordance with these criteria, a total of 2,765 Camelina LRGs were identified, which were then classified into 25 different groups according to their role in different aspects of lipid metabolism, and because of their homology to well-described Arabidopsis lipid regulators.

We used the publicly available developing-seed gene expression datasets and the RNA-seq information generated here from 5-11 DPA seeds to analyze mRNA accumulation patterns of the annotated LRGs. Overall, we identified four major types of genes based on their mRNA abundance during seed development. The smallest group (121 genes) corresponded to genes expressed at high levels [average transcripts per million (TPM) ~380] across all the developing-seed stages tested. These genes were largely associated with functions such as TAG and FA synthesis (Figure 3.1a,

b). The second highest-expressed group (average TPM ~20) consisted of 553 genes with predominant functions associated with lipid synthesis, desaturation, and export from plastids. Most of the genes were associated with two groups with medium-low (average TPM 4.5, ~5,847 genes) and low (average TPM ~0.5, 1,244 genes), primarily related to functions associated with the biosynthesis of membrane lipids, waxes, and suberins (Figure 3.1a, b).

Genes highly expressed in developing seeds corresponded to functions associated with I) TAG synthesis, II) FA synthesis, and III) FA elongation & desaturation (Figure 3.1b), and this is why we analyzed the mRNA accumulation dynamics of these major groups of lipid metabolic genes across the various developmental stages. TAG synthesis genes peak at 18 - 29 DPA with expression values (TPM) several times (>10 times) higher than the other two processes (Figure 3.1c). Partially consistent with metabolite data (Pollard et al., 2015), FA synthesis, elongation, and desaturation genes peak at 10-11 DPA (Figure 3.1c). The value of the newly-added RNA-seq datasets (indicated red in Figure 3.1a), particularly for 5 and 8 DPA is evident from the high level of expression of several LRGs early during seed development (a few examples indicated with asterisks in Figure 3.1a). Taken together, our analyses provide a comprehensive overview of the expression of LRGs during Camelina seed development, featuring specific gene sets with potential major lipid metabolism roles, providing an opportunity to uncover key regulators.

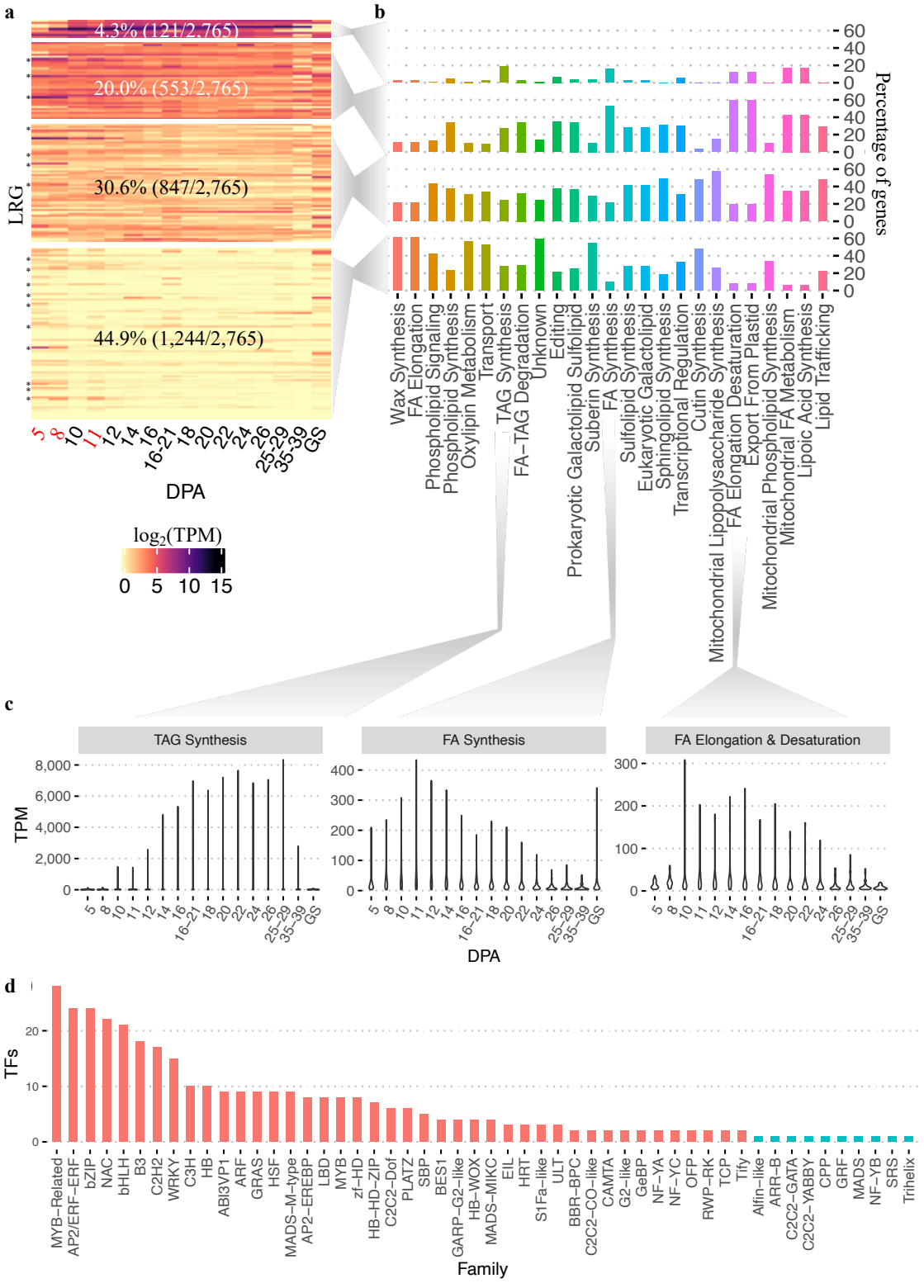

Figure 3.1. Expression dynamics of LRGs during seed development

a. Heatmap representing mRNA accumulation information data highlighting four LRG clusters (rows). The clusters were generated based on the expression level of the corresponding genes

Figure 3.1 (cont'd)
during sixteen timepoints across Camelina seed development including samples immediately after germination (columns). In total we analyzed 2,765 LRGs collected from CamRegBase. DAP: Days after pollination. GS: Germinated seed. **b**. Bar graph indicating the percentage of LRGs assigned to different lipid-related processes by each of the clusters of expression presented in (**a**). The lipid-related processes were defined based on homology with Arabidopsis and following the AraLip classification. **c**. Expression variation across seed development of three major LRG groups. **d**. Bar graph indicating the number of TF classified by families identified as potential lipid-metabolism regulators in Camelina. Red color indicates TF families significantly enriched (FDR < 0.05, Fisher's Exact Test).

3.3.2 Identification of candidate lipid transcriptional regulators by co-expression analysis

To identify candidate genes encoding TFs potentially associated with the regulation of Camelina LRGs, we estimated the mutual information (MI) between each of the 5,590 TFs annotated in CamRegBase and each gene in the genome using all the available Camelina gene expression data. For each TF, we extracted the highest 200 genes (average MI $\geq$ 1) as corresponding to the co-expressed genes of the corresponding TF. We then evaluated whether LRGs were statistically overrepresented [False Discovery Rate (FDR) < 0.05, Fisher's Exact Test] within these 200 genes. From the 5,590 TFs analyzed, we identified 350 TFs that met the criteria. The 350 TFs belonged to 52 different TF families and those with the highest representation corresponded to MYB, AP2/ERF, bZIP, and bHLH families (Figure 3.1d).

We compared our list of TF candidates with 36 Arabidopsis TFs known to participate in the regulation of lipid and/or seed development. The 36 Arabidopsis TF corresponded to 105 Camelina homologous genes, as reported in CamRegBase (Gomez-Cano et al., 2020), consistent with the hexaploid nature of the Camelina genome. We excluded ten out of the 105 Camelina TFs because of the absence of evidence for expression in the available Camelina expression data. We found a significant overlap between the TFs annotated by homology as Arabidopsis lipid regulators and those TFs predicted by our analysis (28 TFs overlapped, *P*-value < 0.05, Hypergeometric test), providing confidence in our approach. These 28 TFs included homologs of WRI1, WRI4, ABI3,

FUS3, LEC2, MYB9, MYB41, MYB107, MYB94, AGL15, VAL2, EEL, and DEWAX (Meinke *et al.*, 1994; Focks and Benning, 1998; Bensmihen *et al.*, 2002; Cernac and Benning, 2004; Tsukagoshi *et al.*, 2007; Braybrook and Harada, 2008; Zheng *et al.*, 2009; To *et al.*, 2012; Go *et al.*, 2014; Kosma *et al.*, 2014; Lee and Suh, 2015; Lashbrooke *et al.*, 2016; Lee *et al.*, 2016; Zhang *et al.*, 2016; Pouvreau *et al.*, 2020). Noteworthy, not all the Camelina paralogs were co-expressed with the same number of LRGs. For example, one of the three Camelina homologs of Arabidopsis AtMYB94, AtMYB41, AtVAL2, AtWRI4, and AtDEWAX were not co-expressed significantly with LRGs. Similarly, only one of the three Camelina paralogs of AtAGL15 and AtLEC2 were present in the list of 350 Camelina TFs (Figure 3.2a).

To prioritize Camelina TF candidates for functional studies, we ranked the 350 identified TFs based on the number of co-expressed LRGs. Notably, the top candidates also showed preferential expression in seeds, as indicated by the seed Z-scores (Kryuchkova-Mostacci and Robinson-Rechavi, 2017) (See *Methods*). From the ranked list, we selected the top 35 TFs, which included 13 pairs of paralogs. From the paralog pairs, we selected only the TF with the largest number of co-expressed LRGs and the highest expression, resulting in a final list of 22 TFs that were subjected to further analyses. Four of these TFs were homologs of known seed development and/or lipid metabolism regulators in Arabidopsis, corresponding to ABI3, FUS3, MYB9, and MYB107 (Giraudat et al., 1992; Keith et al., 1994; Lashbrooke et al., 2016).

To further characterize the TF candidates, we evaluated the conservation of the predicted TF-LRG associations between Camelina and Arabidopsis. For this, we re-analyzed >250 publicly available Arabidopsis RNA-seq experiments using identical pipeline and metrics as for Camelina, selecting datasets similar to the samples used for the Camelina co-expression analyses. We focused specifically on our list of 22 Camelina TFs. Arabidopsis homologs of CsaMYB1 and CsaMYB3

were not expressed on the analyzed data and therefore were excluded from this analysis. In total, within the remaining 20 TFs, ten showed a conserved significant co-expression with LRGs (Figure 3.2b). Substantiating our analyses, the three well-described Arabidopsis lipid regulators AtABI3 (CsaABI3VP1-1), AtFUS3 (CsaABI3VP1-2), and AtMYB9 (CsaMYB2), were identified as part of the conserved co-expression associations. This co-expression analysis identified seven Camelina TFs (and their Arabidopsis homologs) that had not been previously associated with lipid metabolism, including CsaNAC1, CsaNAC2, Csazf-HD1, CsaB3-1, CsaAP2/B3-like-1, CsaULT1, and CsaLBD1 (Figure 3.2b). The remaining ten Camelina TFs that did not show a conserved co-expression with Arabidopsis LRGs are likely to correspond to Camelina-specific lipid regulators, or alternatively they are not involved in the control of lipid metabolism.

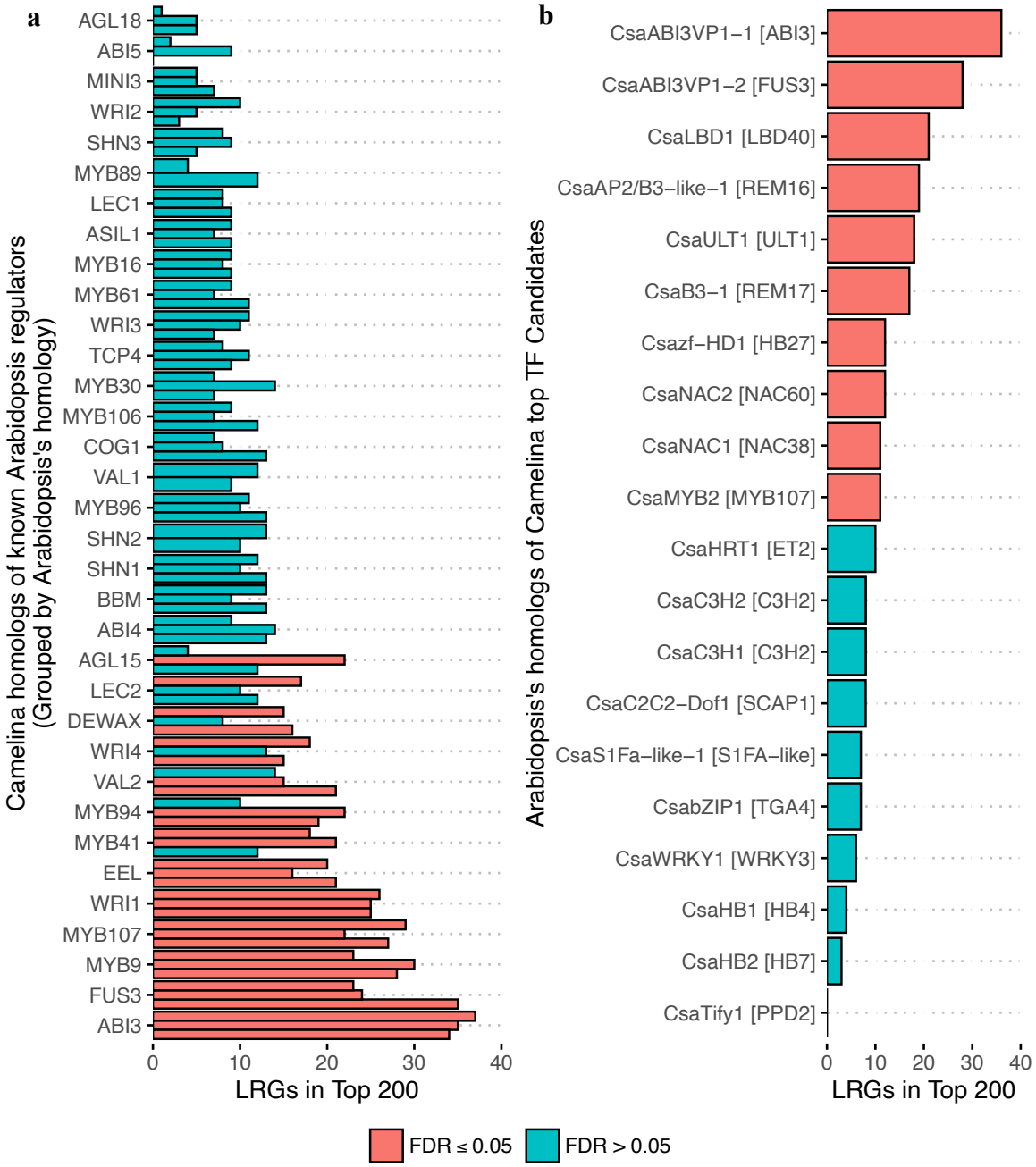

Figure 3.2 Co-expression of known lipid/seed development regulators and LRGs in Camelina and Arabidopsis

The bar graphs show the total number of LRG co-expressed with (**a**) Camelina homologs of each Arabidopsis TF (note that there are three bars for each Arabidopsis regulator because of the hexaploid nature of the Camelina genome), or (**b**) Arabidopsis homologs (names in square brackets) for the Camelina top TFs. The color of the bar indicates the significance of the number LRG co-expressed (light-red, FDR ≤ 0.05; turquoise, FDR > 0.05).

3.3.3 Establishing the DNA-binding landscape of the candidate transcription factors

To further characterize the 22 TFs and to identify potential target genes, we applied DAP-seq (O'Malley et al., 2016). We synthesized and cloned the corresponding open reading frames (ORFs) for the 22 TFs in a vector that permitted expression of the protein fused at the N-terminus to a Halo-tag (Bartlett et al., 2017). We also generated a Camelina unmethylated DAP-seq DNA library (ampDAP-seq) from green tissues of mature plants (see *Methods*). We reasoned that unmethylated DNA better captures the majority of the PDIs in which these TFs are likely to participate (O'Malley et al., 2016), and eliminates variations in methylation patterns between cell types or tissues. We performed DAP-seq in duplicate for each Halo-TF, and with the Halo-tag alone as the control. We obtained on average 25.5 million reads per sample, out of which about half mapped uniquely to the available Camelina genome (v2, cv. DH55) (Kagale et al., 2014). To assess the variance and reproducibility of the experiments, we performed a principal component analysis (PCA) using uniquely mapped reads. The first two PC showed all TFs well separated from the control (Halo). However, we also observed five TFs with strikingly different replicates, indicating low reproducibility between them. For each TF, we also analyzed the similarity of the uniquely mapped reads between each pair of replicates, which confirmed the differences observed on the PCA analysis for replicates of the five TFs. Based on these observations, we discarded the DAP-seq results obtained for CsaABI3VP1-1 and CsaB3-1 (because of its high correlation with the HALO control), and settled on analyzing the replicates of CsaMYB2, CsaULT1, and CsaTify1 independently (replicates with PCC < 0.7). For the remaining 17 TFs, DNA-binding regions (peaks) were called using both replicates. Thus, in total, 20 TFs were tested for the presence of peaks. The number of identified peaks varied greatly between the TFs, with CsaC2C2-Dof1 showing >100,000 peaks, and four TFs having less than 500 peaks (CsaTify1, CsaS1Fa-like-1,

CsaMYB2, and CsaULT1), which were not further used. In consequence, a total of 16 TFs were kept for further analyses.

The analysis of the distance between the peak summit and the closest annotated transcription start sites (TSSs) indicated that, on average, ~63% of the total peak summits are within 3 kbs of the TSSs. Thus, our results are in agreement with the peak genomic distribution patterns previously observed in DAP-seq experiments for Arabidopsis and maize (O'Malley et al., 2016; Galli et al., 2018). We compared, in terms of successful identification of TF binding motifs, all our DAP-seq results (including those which failed to pass the quality controls) with those performed in Arabidopsis (O'Malley et al., 2016) and determined that 17 common TFs were tested (TF homologs). To note, 3/17 TFs did not work in either plant, 7/17 TFs worked in Camelina but not in Arabidopsis, and 6/17 TFs worked in both plants. The remaining TF (AtMYB107 homolog of CsaMYB2) worked only in Arabidopsis, likely related to the lack of MYB domains on the Camelina annotated transcript. Finally, the corresponding genes for CsaMYB1, CsaNAC2, and CsaC3H2 were not previously tested in Arabidopsis. In summary, we provide here high-confidence DNA-binding data for 16 TFs, of which 10 were previously unknown in Arabidopsis.

To evaluate the quality of the predicted DAP-seq peaks of the corresponding 16 TFs, we determined the $\log_2$ fold change of the binding ($\log_2$FC, See *Methods*). We defined high-confidence peaks for further analyses as those showing $\log_2FC > 0.5$ in both replicates, which represented ~32.5% of the total peaks called (Figure 3.3). One additional criteria that we applied to decide whether DAP-seq provided meaningful information or not was the enrichment for particular TF-binding motifs (TFBM) within the recovered peaks, a widely accepted characteristic of the DNA fragments recognized by TFs (Lambert et al., 2018). To identify the TFBMs associated with each TF, we ranked all the high-quality peaks based on their $\log_2$FC, selected the top 1,000

peaks for each TF and identified the motif consensus using MEME-ChIP (Machanick and Bailey, 2011). To evaluate the relevance of the predicted TFBMs in the context of the identified peaks, we searched each TFBM across the full set of peaks for each TF, focusing on two specific aspects: (1) The fraction of peaks that harbored the motif, and (2) the localization of the motif within the peak (distance to the summit). We carried out this analysis by extending each peak 50 bps around the summit (Figure 3.4). The most significant motifs identified for each of the 16 TFs corresponded to those with the largest abundance and which displayed a clear accumulation close to the summit of each peak (Motif 1 in Figure 3.4). Thus, for the rest of this study, we considered high-confidence peaks those that harbored such a motif, corresponding to ~92% of all the peaks evaluated.

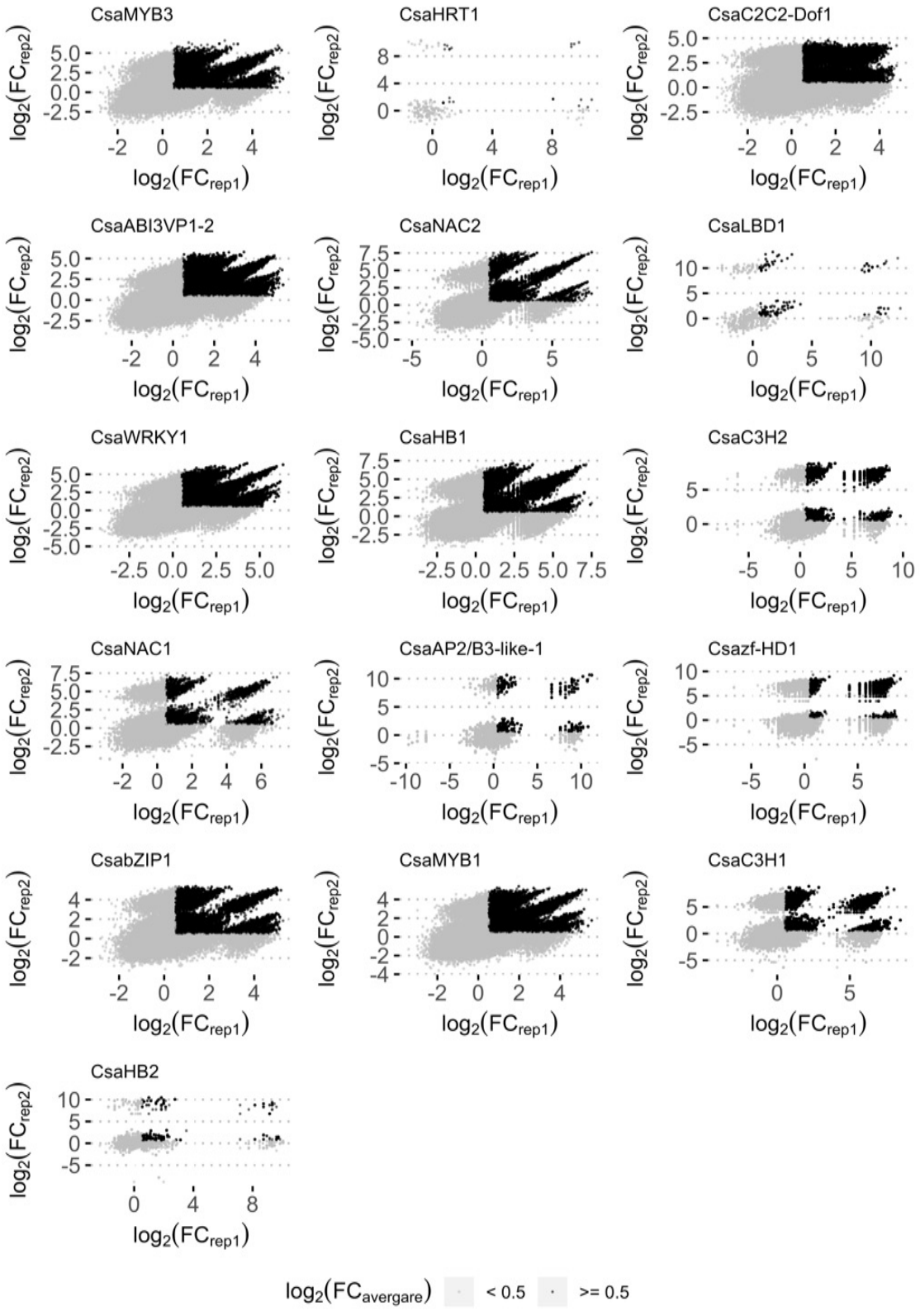

Figure 3.3 Reproducibility analysis between TF replicates based on DNA-binding fold changes

Figure 3.3 (cont'd)
We calculated the $\log_2$ of the binding fold change ($\log_2$FC) for the total predicted peaks for each TF dividing the number of reads obtained for each peak with Halo-TF by the number of reads obtain for the same peak for the Halo control. Peaks with $\log_2$FC $\geq$ 0.5 in both replicates were defined as highly reproducible peaks.

We compared the DNA-binding specificities provided by DAP-seq between the corresponding six Camelina and Arabidopsis homologs. We re-analyzed all six Arabidopsis DAP-seq using the same pipeline employed in the current study. Five TF pairs (AtNAC38 and CsaNAC1; AtFUS3 and CsaABI3VP1-2; AtMYB67 and CsaMYB3; AtTGA4 and CsabZIP1; AtWRKY3 and CsaWRKY1) showed almost identical DNA-binding preferences, suggesting that the amino acid residues that distinguish the Arabidopsis and Camelina homologs are not significantly affecting *in vitro* DNA-binding specificities. The only exception was CsaAP2/B3-like-1 for which none of the top motifs identified matched the TTTGGCGGGAA sequence consensus predicted for AtREM1. This result puzzled us, hence we decided to re-check if the Arabidopsis and Camelina genes were properly annotated. Indeed, we determined that one of the B3 domains that characterizes the DNA-binding domain of AtREM1 (Romanel et al., 2009) was absent in the cloned CsaAP2/B3-like-1 ORF, because of a likely error in the current Camelina genome annotation. Taken together, we identified the DNA-binding patterns for 16 Camelina TFs, and determined a similar correspondence with the Arabidopsis homolog, when available.

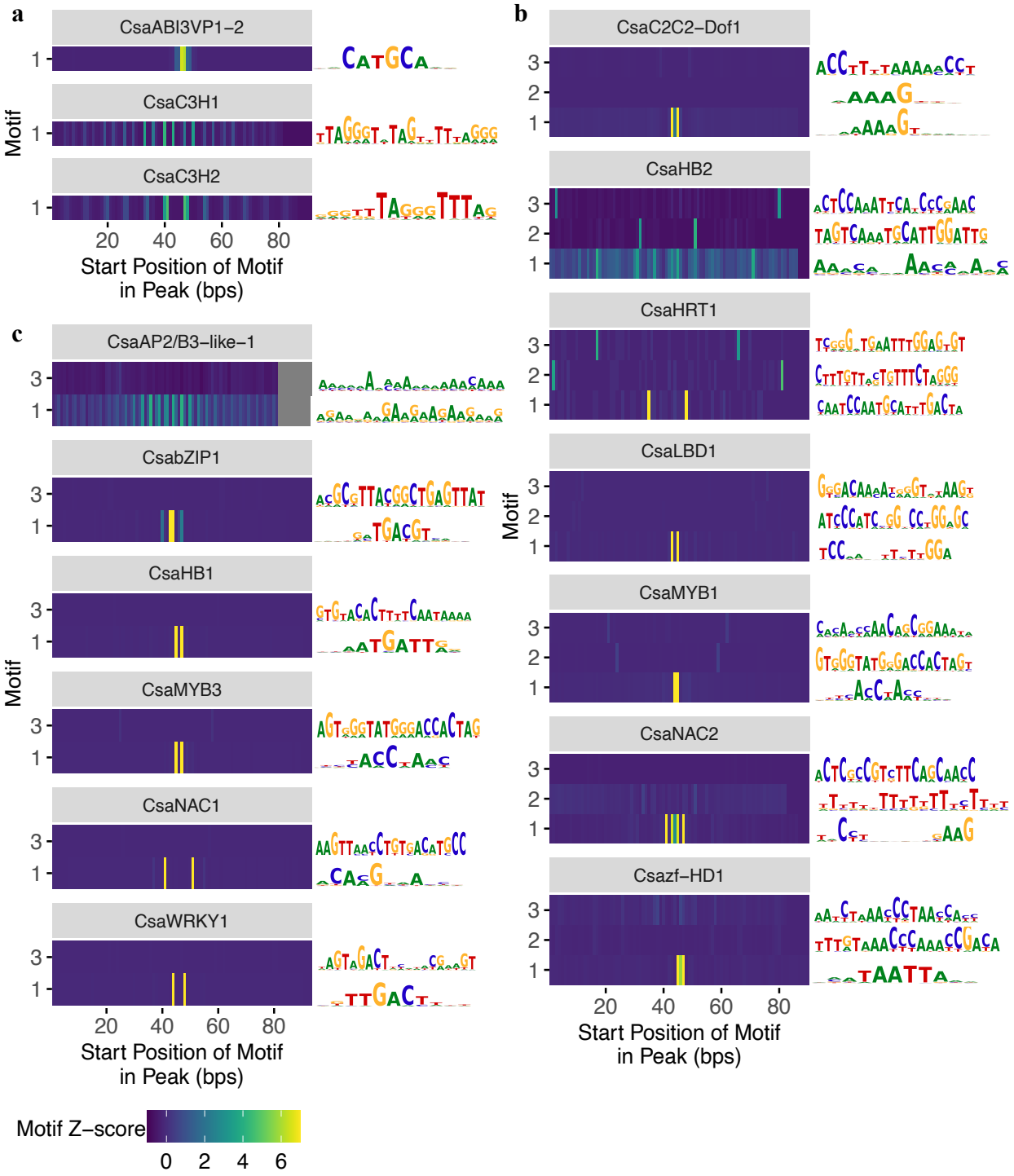

Figure 3.4 Distribution of predicted TF binding motifs (TFBMs) in the predicted peaks
The prediction of TFBMs resulted in up to three different motifs for some of the TFs. To identify the main motif, we counted the frequency and location of each of the predicted TFBMs in all the predicted peaks. The frequency of each TFBMs is presented as the motif Z-score in a heatmap indicating the start position of the TFBMs on the peak. Each TFBM was tested independently regardless of whether the TFs have a single (a), two (b), or three (c) TFBMs.

3.3.4 Predicting gene targets for the selected TFs

To identify potential gene targets for the 16 TF candidates, we determined which genes were located within 3 kbps of each high-confidence peak summit, since 3 kbps capture many of the biologically-relevant TF-target gene interactions (Springer et al., 2019). We identified a total of 31,898 potential targets for these 16 TFs, with CsaMYB1 and CsaHRT1 showing the largest (6,816 genes) and lowest (9 genes) number of target genes, respectively. As a first step towards assessing the biological significance of the DAP-seq results and its concordance with the co-expression prediction, we tested if the predicted targets were enriched in LRGs and/or in TFs associated with the control of LRGs and seed development (TF-LRG/development). Tellingly, 4/16 and 6/16 sets of targets showed significant enrichment (P-value ≤ 0.05, Fisher-exact test) on LRGs and TF-LRG/development targets, respectively (Figure 3.5a). Moreover, CsaABI3VP1-2 (Camelina homolog of AtFUS3) showed enrichment on both sets of genes, suggesting an important role in lipid metabolism. Thus, in total, 9 out of the 16 TFs tested showed a significant enrichment for target genes associated with lipid metabolism in Camelina.

Previously, we showed that half of the candidate TF homologs in Arabidopsis were enriched in LRGs by co-expression (Figure 3.5b). Thus, we tested if they were also enriched in target genes annotated as LRGs, as we found for the Camelina TFs (Figure 3.5a). We performed the analysis with the six Arabidopsis TFs for which we previously evaluated TFBMs. The analysis of the Arabidopsis DAP-seq data was performed using the same pipeline and controls as we used for the Camelina data. Out of the six Arabidopsis TFs, five showed enrichment for target genes annotated as LRGs and TF-LRG/development (Figure 3.5b). This finding, along with the conservation of the corresponding TFBMs, suggests conservation of the corresponding regulatory functions. Curiously, the five Arabidopsis TFs showed target enrichment for both type of genes: LRGs and

TF-LRGs (Figure 3.5b). This contrasts with what we found for the corresponding Camelina homologs which showed in all cases but one either enrichment for LRGs or TF-LRGs but not in both (Figure 3.5a). Finally, neither CsaAP2/B3-like-1 nor AtREM16 showed targets enriched in LRGs or TF-LRG/development. While this is consistent with the possibility that we used a truncated protein for CsaAP2/B3-like-1, our results suggest that AtREM16 plays a secondary role as a lipid metabolism regulator.

To characterize other functional roles of the set of predicted target genes associated with the corresponding TFs, we investigated enrichment for Gene Ontology (GO) terms. All the TFs tested have at least one GO term enriched that is lipid-related. After removing redundant and general terms, we clustered all the TFs based on the top 10 GOs for each (based on *P* values), allowing us to separate them into two main clusters. One cluster (indicated in green) was associated with a wide range of GO terms, including regulation of development and several metabolic processes, particularly lipid metabolism-related functions, as well as phenylpropanoid and carboxylic acid biosynthesis processes. Members of the other cluster (indicated in orange, Figure 3.6) have in common the terms signal transduction, defense responses, regulation of gene expression, and regulation of nitrogen compounds. When all the data is considered together, these analyses provide additional evidence that the DAP-seq results bore biologically meaningful targets and support the initial co-expression predictions, including the discovery of previously unrecognized candidate regulators of lipid-related genes.

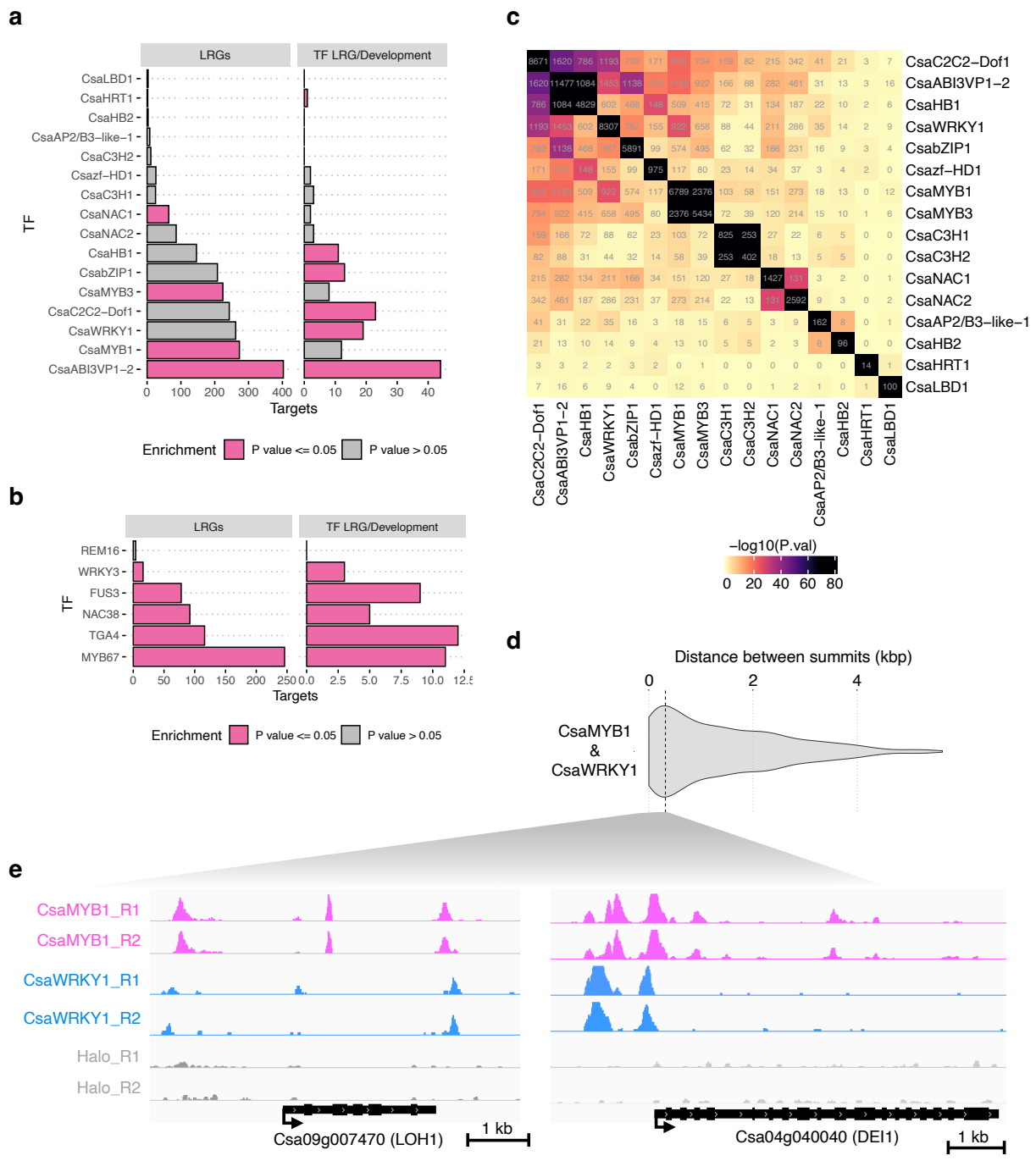

Figure 3.5 Regulatory landscape of predicted lipid-related regulators based on DAP-seq
The bar graph indicates the number of predicted target genes annotated as (**a**) Camelina or (**b**) Arabidopsis LRGs. Arabidopsis TFs correspond to homologs of the Camelina predicted candidates. The red color indicates the significance of the overlap of target genes annotated as LRG vs the total number of predicted target genes for each of the tested TF (*P*-value ≤ 0.05, Fisher's Exact Test). **c**. Heatmap indicating the number of common targets between pairs of TFs. Color scale indicates the *P*-value associated with the corresponding number of common targets (Fisher's Exact Test). **d**. Violin plot showing the distribution of distances (in bps) between summits

Figure 3.5 (cont'd)
(peak centers) of CsaMYB1 and CsaWRKY1 mapped to common targets. The vertical dashed line indicates the most frequent distance between summits. **e**. IGV plots with co-binding profiles (peak) generated from the DAP-seq experiments of CsaMYB1 and CsaWRKY1 highlighting two shared targets with the respective gene models obtained from Camelina V2.0 at the bottom. Peaks heights correspond to the number of reads by bins (10 bp) per million mapped reads.

Figure 3.6 Heatmap and hierarchical clustering of TF candidates based on the top 10 GO terms significantly enriched

Figure 3.6 (cont'd)
GO terms not significantly enriched are shown in white. The color indicates the percentage of targets annotated within the corresponding GO term.

Similarities in the functional annotation of target genes for TF pairs may indicate that the corresponding TFs share common targets. Alternatively, the TFs could regulate different genes in the same process/pathway. To distinguish between these two possibilities, we evaluated the overlap in targets between the 16 TFs. Almost half of the comparisons showed significant target overlaps (*P*-value < 0.05, Fisher's Exact Test) (Figure 3.5c, darker colors indicate smaller *P*-values). As anticipated, TFs from the same family (CsaMYB1 and CsaMYB3; CsaNAC1 and CsaNAC2) had the largest number of shared target genes, likely driven by the very similar *in vitro* DNA-binding consensus of the corresponding TFs. Noteworthy, while significant, the overlap comprises only a subset of all the targets for each of these TFs, suggesting that outside the shared core motif, each TF has specific DNA-binding preferences (Figure 3.4). Many of the TF pairs have overlapping targets (*e.g.,* CsaMYB1 and CsaMYB3; CsaHB1, CsaABI3VP1-2, and CsaC2C2-Dof1; CsaNAC1, CsaNAC2, and CsaAP2/B3-like-1), indicating that they function in the control of related biological processes. We explored this hypothesis by comparing two of the non-homologous TF pairs with the highest number of common targets, corresponding to CsaMYB1-CsaWRKY1 and CsabZIP1-CsaHB1, which had 922 and 468 common targets, respectively. For the CsaMYB1-CsaWRKY1 pair, we found that shared targets were enriched in multiple lipid-related GO terms at several levels of the GO hierarchy, including carboxylic acid biosynthesis and very long-chain fatty acid biosynthesis. Contrary to the pattern observed for CsaMYB1-CsaWRKY1, common targets of CsabZIP1-CsaHB1 were enriched in a more diverse list of biological processes not observed on the corresponding individual list of enriched GO terms, including flavone biosynthesis, regulation of transcription, activation of protein kinase activity,

root hair cell tip growth, and leaf senescence, suggesting that their role in lipid metabolism control is not linked to common target genes in the pathway.

To further understand the potential participation of CsaMYB1 and CsaWRKY1 in gene co-regulation of their common targets, we evaluated the distribution of binding sites in the 922 shared targets. For most of them, the binding sites were within a few hundred base pairs apart from each other (the average distance was 320 bps; Figure 3.5d), highlighting a possible cooperative work at the DNA level (post-DNA binding) (Reiter et al., 2017). The proximity and potential significance for transcriptional regulation is exemplified by the two shared targets Csa03g002110 and Csa04g040040 (Figure 3.5e), Arabidopsis homologs of 3-KETOACYL-COA SYNTHASE (KCS1, At1g01120) and PASTICCINO 1 (PAS1/DEI1, At3g54010), which are involved in FA and VLCFA synthesis (Shang et al., 2016; Roudier et al., 2010), respectively, further underscoring the potential regulatory role of CsaMYB1 and CsaWRKY1 on lipid metabolism.

3.3.5 Identified TFs associate with distinct aspects of lipid metabolism

To better understand the specific aspects of lipid metabolism that each of the identified TFs might be involved with, we scored how many targets of each TF corresponded to each of the lipid pathway categories (as presented in Figure 3.1b). In total, 11/16 TFs showed significant enrichment for targets annotated across several lipid-related processes (P-value < 0.05, Fisher's Exact Test). As examples, CsaMYB3, CsaMYB1, CsaWRKY1, and CsaABI3VP1-2 were enriched in more than four different processes, with their top target processes being suberin synthesis (18.2%), cutin synthesis (25.3%), and transcriptional regulation (18.1% and 41.9%), respectively. Remarkably, several combinations of TFs showed significant enrichment for the same processes. Finally, we also observed that the targets for CsaABI3VP1-2 and CsaWRKY1

were significantly enriched in genes associated with TAG synthesis (22.6%) and FA-TAG degradation (15.7%), respectively, which are core processes in the accumulation of seed oil.

In parallel, to evaluate the biological significance of the regulatory interactions predicted at the pathway co-expression level, we applied the gene set enrichment analysis (GSEA) algorithm (Subramanian et al., 2005) using the Pearson Correlation Coefficient (PCC) as the scoring metric. Thus, significant positive and negative enrichment values indicate association of the corresponding TF with a metabolic pathway in a positive or negative fashion, respectively. Also, under these conditions, GSEA permits the identification of TF-process relationships that have significant co-expression signals at the pathway rather than as individual target gene levels (Subramanian et al., 2005). Eight out of the sixteen TFs tested showed significant enrichment (P-value < 0.05) for at least one of the processes tested. CsaABI3VP1-2 showed the largest number of significant associations (up to ten), including FA elongation and desaturation, FA and TAG synthesis, and transcriptional regulation. The second and third TFs with most enriched processes were CsaWRYK1 and CsaMYB1, with seven each. We also observed eleven TF-process associations with negative enrichment scores, indicating enrichment for negative co-expression values, within which CsaWRKY1, CsaMYB1, CsaMYB3, and CsabZIP1 are included. The former showed enrichment for negative scores on its corresponding targets annotated under FA synthesis, transcriptional regulation, and transport, while the latter with targets annotated under cutin synthesis, wax synthesis, and FA elongation. These results suggest major roles of these TFs as negative regulators of the mentioned pathways.

Finally, we combined both sets of results (target enrichment and GSEA results) to identify high-confidence TF-process associations. Six of the eleven TFs analyzed showed significant associations in both tests with at least six different processes, to a total of ten TF-pathway

associations (pink edges in Figure 3.7). Transcriptional regulation was the process with the largest number of connections. CsaNAC1, CsaWRKY1, and CsaABI3VP1-2 were the three TFs with the largest number of associations (two for each of them, Figure 3.7). Three out of the ten TF-pathway associations showed significant negative enrichment (Figure 3.7), indicating transcriptional repression roles of the corresponding TFs on the respective pathways. Also, it is worth noting that one of the main processes enriched for the targets of CsaMYB1 and CsaNAC2 was cutin synthesis (Figure 3.7). CsaABI3VP1-2 was the only TF significantly enriched in TAG synthesis- and transcriptional regulation-related targets (Figure 3.7), and remarkably we found that the large majority of the targets that we predicted for CsaABI3VP1-2 were also TF targets previously identified for AtFUS3 by either chromatin immunoprecipitation-DNA microarray (ChIP-chip) (Wang and Perry, 2013) or DAP-seq assays (O'Malley et al., 2016), uncovering potential Camelina-specific interactions as well as unreported Arabidopsis targets. Altogether, these analyses underscore CsaABI3VP1-2 as a good candidate playing a major role in lipid metabolism in Camelina, similar to AtFUS3 (Yamamoto et al., 2010; Wang and Perry, 2013; Zhang et al., 2016).

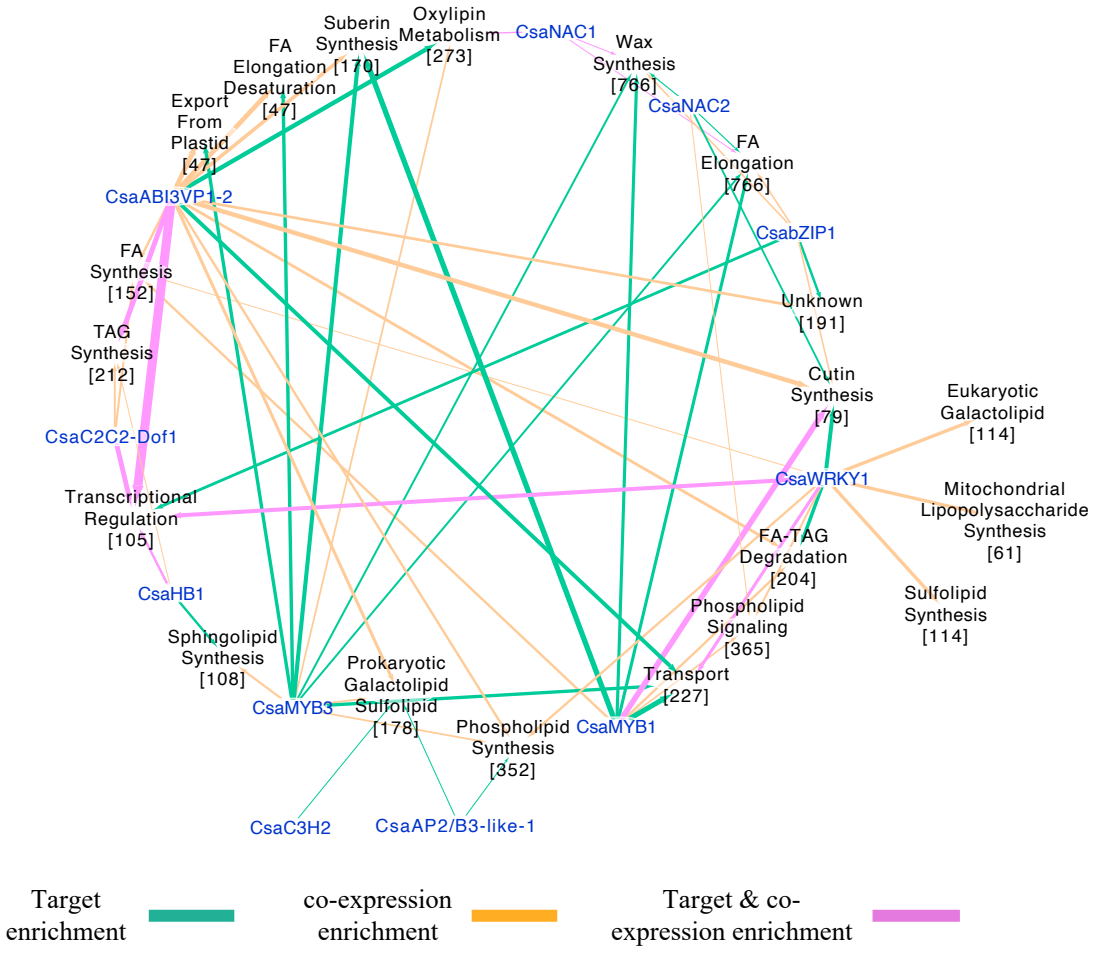

Figure 3.7. High-confidence TF-process network
Associations predicted based on target enrichment and GSEA using TF-target PCC as score metric are indicated by lines joining TFs (blue) and specific processes associated with lipid metabolism (black). The thickness of the edges represents the fraction of lipid-related genes in the pathway that is being targeted by the corresponding TF. The total number of genes annotated for each of the corresponding lipid-related processes are indicated inside square brackets.

3.3.6 Dynamic behavior of the predicted networks during seed development

To gain further insights on the regulatory effect of the identified TF-target interactions in Camelina seeds, we performed a second co-expression analysis with GENIE3 (Huynh-Thu et al., 2010) using only expression data from seeds. GENIE3 uses a regression tree and random forest algorithm to make regulatory prediction implying causality (Huynh-Thu et al., 2010). Thus, we assumed that predictions identified by GENIE3 and supported by DAP-seq are highly confident

regulatory interactions occurring specifically in seeds. The significance of the predicted score was assayed using a permutations test (FDR ≤ 0.001, 1,000 permutations). Overall, 35% of the targets identified by DAP-seq were also predicted as targets of the corresponding TFs by GENEI3 (Figure 3.8a). The highest percentage of DAP-seq seed co-expressed targets was observed for CsaNAC2 and CsabZIP1 (~54% each, Figure 3.8a). These results suggest that many of the predicted TF-target associations have a regulatory effect in the context of seed development.

To parse TFs involved in controlling FA and TAG-related genes in seeds, we combined the target enrichment and the GSEA results (Figure 3.7) to select TFs associated with the corresponding pathways. Consequently, we reduced the TF-target DAP-seq network to only targets co-expressed in seeds (as predicted by GENIE3) (Figure 3.8a). With this subset of TF-target interactions, we tested the enrichment for targets on the corresponding pathways once again to determine if the reduced TF-target network still had a significant number of targets associated with FA and TAG-related processes. Seven TFs showed enrichment for seed co-expressed targets associated with at least three different pathways (FDR < 0.05, Fisher's Exact Test) (Figure 3.8b). FA elongation was the pathway most frequently targeted, with six different TFs associated with it (Figure 3.8b). CsaABI3VP1-2 and CsaMYB1 were the two TFs with most seed co-expressed targets annotated under FA elongation. However, TAG synthesis and FA-TAG degradation were significantly targeted by just one TF each, CsaABI3VP1-2 and CsabZIP1, respectively (Figure 3.8b).

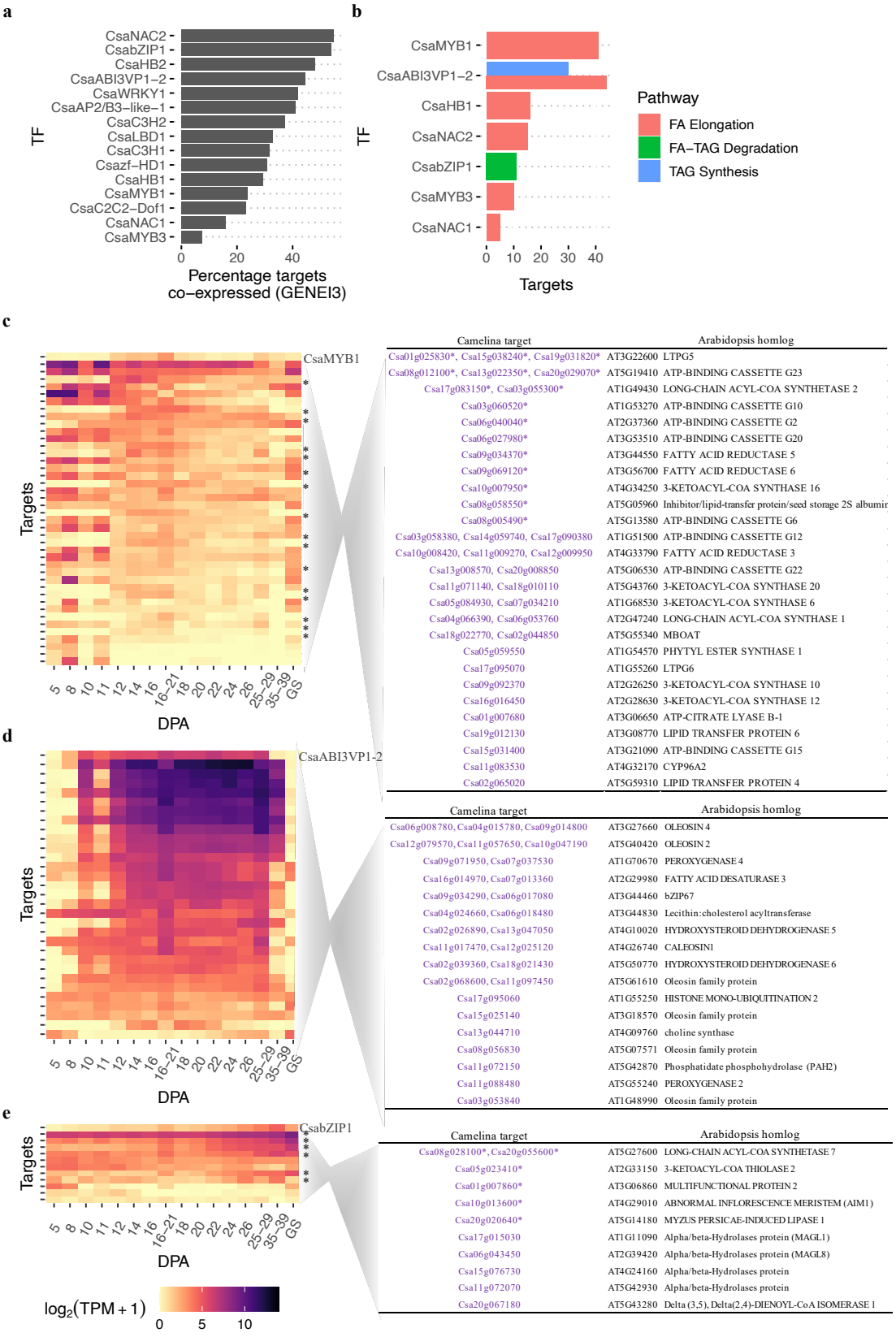

Figure 3.8. Integration of seed co-expression and DNA-binding information

Figure 3.8 (cont'd)
a. Bar graph indicating the percentage of DAP-seq targets supported by the co-expression associations predicted with GENEI3 using seed expression data. **b**. Bar graph of the seven most significant TF-lipid-related process interactions that passed the enrichment test after incorporation of seed co-expressed targets. For each TF, the total number of target genes annotated for the corresponding FA and TAG related processes are indicated. **c**. Heatmap representing the expression dynamics of targets of CsaMYB1 associated with FA elongation during seed development. **d**. Heatmap representing the expression dynamics of targets of CsaABI3VP1 associated with TAG synthesis during seed development. **e**. Heatmap representing the expression dynamics of targets of CsabZIP1 associated with FA/TAG degradation during seed development. The right panel list the gene IDs for the Camelina genes represented in the heatmaps and the gene IDs for the corresponding Arabidopsis homologs.

We selected three of seven TF-pathway interactions (CsaMYB1 & FA elongation, CsaABI3VP1 & TAG synthesis, and CsabZIP1 & FA-TAG degradation, Figure 4b) to analyze the expression dynamics of the corresponding TFs and targets during seed development. CsaMYB1 showed two expression windows, one at 12 - 21 DPA and a second at 25 - 29 DPA (Figure 3.8c). These expression profiles are in concordance with the reported peaks of FA synthesis and TAG accumulation (11-24 DPA) (Pollard *et al.*, 2015). CsaABI3VP1-2 showed a broader expression window, starting at 8 DPA with constant expression until 25 - 29 DPA (Figure 3.8d). Finally, CsabZIP1 is mainly expressed during the later stages of seed development (expression peak ~35 - 39 DPA) (Figure 3.8e), consistent with the expected pattern for controlling TF/TAG degradation right before seed germination.

As for the corresponding target genes, we observed several expression patterns consistent with activation or repression by the respective TFs, as exemplified for the targets of CsaMYB1 and CsabZIP1 (Figure 3.8c, e). Within the set of the CsabZIP1 targets, it is worth mentioning multiple Arabidopsis homologs involved in FA beta-oxidation during seed germination (Fulda et al., 2004; Footitt et al., 2006; Jiang *e*t al., 2011; Richmond and Bleecker, 1999), and homeostasis of phospholipid and neutral lipids (Ghosh et al., 2009) (Figure 3.8e). Finally, most (28/30) of the CsaABI3VP1-2 targets showed a similar expression to the corresponding TF (Figure 3.8d). Within

these targets, it is worth noting the Csa16g014970/Csa07g013360 and Csa09g034290/Csa06g017080 gene pairs which are homologs of Arabidopsis FATTY ACID DESATURASE 3 (AtFAD3) (At2g29980) and AtbZIP67 (At3g44460), respectively. The former is involved in linolenic acid synthesis (O'Neill et al., 2011), while the latter is a known regulator of AtFAD3 (Mendes et al., 2013), highlighting a potential feedforward loop between CsaABI3VP1-2, Csa09g034290/Csa06g017080, and Csa16g014970/Csa07g013360 in Camelina.

3.4 DISCUSSION

Camelina is an oilseed crop that has emerged as a prominent feedstock for biofuels and industrial oils during the past decade. Its polyploid genome makes it challenging to identify genes involved in the biosynthesis or regulation of seed oils by classical loss-of-function approaches. The homology to Arabidopsis has permitted to translate knowledge gained in this model plant to Camelina, exemplified in the manipulation of epicuticular and total wax production by the overexpression of AtMYB96 (Lee et al., 2014), or in the increase of seed oil by the overexpression of Arabidopsis WRI1 (An and Suh 2015). However, homology-based approaches are unlikely to reveal the regulators that make Camelina such a good oil producer. Moreover, techniques such as ChIP-seq, classically used to discover TF targets, can be challenging to implement because of the difficulties associated with developing antibodies that recognize a single homolog in a polyploid, and the use of epitope-tagged version of the TF for ChIP experiments is questionable because the function of the epitope-tagged TF cannot be tested unless a mutant is available (which again is difficult to obtain in a polyploid).

We present here a co-expression guided approach aimed at identifying candidate Camelina TFs involved in the control of seed oils, followed by the evaluation of TF target genes based on DAP-seq. While not perfect, this strategy overcomes many of the limitations imposed by a

polyploid genome, providing a small set of candidate TFs that can be used for metabolic engineering efforts (Grotewold 2008). Co-expression analyses identified 22 TFs strongly co-expressed with LRGs, which were further reduced to 16 after several quality control steps. Furthermore, co-expression analyses with seed expression data allowed us to identify specific metabolic processes targeted by our regulators, including the control of FA- and TAG-related genes during Camelina seed development.

Evidence of the robustness of our co-expression analysis is provided by the inclusion in our list of candidate TFs homologs of well-known regulators of lipid-related metabolism in Arabidopsis, including the Camelina homologs of ABI3, FUS3, MYB9, and MYB107 (Giraudat et al., 1992; Keith et al., 1994; Lashbrooke et al., 2016). Our list of Camelina TFs also includes homologs of TFs indirectly associated with lipid metabolism in Arabidopsis, such as the Arabidopsis homolog of CsaNAC2 (AtNAC60). AtNAC60 was shown to play a role in sugar sensing (Li et al., 2014), as a negative regulator of AtABI5 (Yu et al., 2020), and is a target of AtABI4 (Li et al., 2014). Both AtABI5 and AtABI4 are known regulators of sugar-responsive expression, seed germination, and lipid metabolism (Chandrasekaran et al., 2020; Skubacz et al., 2016). While AtULT1 (homolog of CsaULT1) has been implicated in various Arabidopsis plant developmental processes (Fletcher, 2001; Pires et al., 2015; Ornelas-Ayala et al., 2020), a recent transcriptome analysis of loss of AtUTL1 function (Tyler et al., 2019) showed a significant enrichment (P-value < 0.05) for LRGs among the differentially expressed genes, indicating a participation of AtUTL1 in the control of lipids. We could not test the co-expression of AtMYB67 (homolog of CsaMYB3) with Arabidopsis LRGs because it is expressed at very low levels. However, supporting a potential role of CsaMYB3 in lipid-related metabolism, AtMYB67 physically interacts with the known negative regulator of cuticular wax biosynthesis AtDEWAX

(Trigg et al., 2017), which is a target of AtAGL15, a regulator of embryogenesis and gibberellic acid catabolism (Zheng et al., 2013). However, our analysis also identified 13 Camelina TFs (including CsaC3H2, CsaHB2, Csazf-HD1, CsaC3H1, CsaAP2/B3-like-1, CsaC2C2-Dof1, CsaB3-1, CsaS1Fa-like-1, CsaNAC1, CsaWRKY1, CsaTify1, CsaHB1, and CsaLBD1) that were previously not associated with the regulation of lipids.

To further elucidate the role of these 22 Camelina TFs in lipid regulation and to identify potential targets of these TFs, we applied DAP-seq to them. DAP-seq has many limitations, chief among them is that it is performed in a chromatin-free context, resulting in the identification of binding sites and potential targets that might not be accessible *in vivo*. However, it is easy to implement, and it is not affected by a polyploid genome, as ChIP techniques are. DAP-seq permitted us to identify TFBMs and potential target genes for 16 out of the 22 TFs identified. When we compared our DAP-seq results with those derived from a large scale analysis conducted for Arabidopsis TFs (O'Malley et al., 2016), we determined that DNA-binding properties for ten out of the 16 TFs are not available for the corresponding Arabidopsis orthologs, either because were not tested (homologs of CsaMYB1, CsaNAC2, CsaC3H2), or because the corresponding experiments for Arabidopsis TFs did not result in meaningful results (homologs of CsaC2C2-Dof1, CsaC3H1, CsaHB1, CsaHB2, CsaHRT1, CsaLBD1, and Csazf-HD1).

The analysis of TF target enrichment for LRGs and GO terms within the sets of predicted targets provided additional validation for the results obtained from the co-expression analyses for nine TFs (Figure 3.5a). Of interest, in several instances, the GO enrichment analysis also exposed biological processes previously reported for the corresponding Arabidopsis homologs. For example, the DAP-seq results for CsaMYB1 and CsaWRKY1 showed enrichment for targets associated with suberin biosynthesis and defense responses, which are known functions of

AtMYB9 and AtWRKY3, respectively (Lai et al., 2008; Lashbrooke et al., 2016; Birkenbihl et al., 2018). The targets of CsaNAC2 were also enriched in genes associated with fruit ripening, hormone biosynthesis processes, exit from dormancy, inositol lipid-mediated signaling, and lipid homeostasis terms (Figure 3.6), which are in good agreement with the functions attributed to AtANAC60, the Arabidopsis homolog (Li et al., 2014; Yu et al., 2020). The targets of Csazf-HD1 showed enrichment for several GO terms, including cell death, cell wall organization, cellular response to endogenous stimulus, and cellular response to hormone stimulus, matching the predicted functions of AtZHD13/RHD1 (Liu et al., 2021). CsaHB1's targets were enriched in GO terms related to leaf development and light responses, response to auxin, response to abscisic acid, and post-embryonic plant organ development, among others, similar to the known functions of AtHB4 (Carabelli et al., 1993; Sorin et al., 2009; Bou-Torrent et al., 2012). Finally, the targets of CsaABI3VP1-2 showed enrichment for GO terms that cover the full spectrum of the functions known for its Arabidopsis homolog, AtFUS3 (Curaba et al., 2004; Kagaya et al., 2005; Tiedemann et al., 2008; Lumba et al., 2012; Tang et al., 2017). When considered together, these results not only provide evidence of the biological significance of the associations identified here, but also reveal the intertwined connections between lipid metabolism and other biological processes in Camelina.

Regulators of plant metabolism often regulate multiple genes in a pathway, making them attractive for metabolic engineering (Broun 2004; Grotewold 2008). We took advantage of this characteristic of metabolic regulators to identify TF-process relationships with significant co-expression signals at the pathway, rather than as individual target gene level by applying GSEA (Subramanian et al., 2005). This permitted us to identify previously unstated associations that further support several of the identified TFs as important lipid regulators (Figure 3.7). Finally, we

took advantage of a computational method (GENIE3) that involves causality (rather than simply correlation) to further support the role of several of the identified TFs in controlling particular aspects of lipid metabolism in seeds. When taken together with the results from the other methods applied in this study, our results suggest that CsabZIP1 is involved in controlling FA and TAG degradation just before seed germination, CsaMYB1 regulates FA elongation, and CsaABI3VP1 controls the synthesis of TAG (Figures 3 & 4). Our results also implicate CsaMYB1 and CsaNAC2 as participating in the regulation of cutin biosynthesis (Figure 3.7).

Gene regulation is at the core of many important agronomic attributes and TFs have a large potential to modify complex traits (Century et al. 2008; Springer et al. 2019). Identifying TFs that control specific metabolic or developmental processes in polyploids is challenging because the effect of mutations is often masked by redundancy, and traditional approaches to investigate TF function are limited by high sequence identity between homologs. Our strategy to identify Camelina candidate TFs involved in the regulation of lipid metabolism was based on a combination of co-expression analyses and target identification using DAP-seq. These resulted in the identification of a set of 16 TFs. The presence among these 16 TF of several that were previously shown in Arabidopsis to participate in different aspects of lipid accumulation furnished a validation for the approach. Incorporating into our pipeline co-expression analyses that imply causality and that take into consideration that TFs often control multiple genes in a metabolic pathway further provided a better picture of the regulatory events involving the identified TFs in seed oil accumulation in Camelina. Similar combination of approaches could significantly contribute to identify key regulators for important agronomic traits in other polyploids.

3.5 METHODS

3.5.1 Plant materials and growth conditions

Camelina sativa cultivar Suneson was grown in the plant biology greenhouse at Michigan State University. RNA-seq and DAP-seq experiments were performed on plants grown for one month at 22 °C and under 16:8 -h light/dark cycles. For seed RNA-seq, total RNA was extracted from seedpods harvested at 5, 8, and 11 DPAs. For DAP-seq, a pool of ten leaves from six mature (two-month-old) plants was collected.

3.5.2 Cloning and expression of transcription factors for DAP-seq

A set of 22 full-length Camelina TF-ORFs were annotated using the *Camelina sativa* cultivar DH55 (reference genome, V2.0, http://camelinadb.ca). Coding regions were assembled (when required) using expression data available for the *Camelina* cultivar Suneson. TF-ORFs Gateway compatible were synthesized by Genewiz (https://www.genewiz.com/Public/Services/Gene-Synthesis/Standard). Clones were recombined using LR clonase II (Life Technologies) into the pIX-Halo expression vector containing both T7 and SP6 promoters (pIX-Halo:ccdB) 6xHis-tag at C-terminus, no stop codon but has T7 terminator.

3.5.3 RNA-seq library preparation

Total RNA from fresh seed, after removing pod covers, was extracted using Spectrum Plant Total RNA Kit (Sigma-Aldrich) according to the manufacturer's protocol. The total RNA was prepared with three biological replicates, each with ~100 mg seeds. The quality of total RNA was determined by TapeStation4200 (Agilent), and cDNA library was generated with 1 μg of total RNA using TruSeq stranded mRNA (Illumina). The pooled libraries were sequenced with a pair-ended read length of 150 bp by Illumina HiSeq 4000 at the Research Technology Support Facility Genomics Core at Michigan State University.

3.5.4 DAP-seq library preparation

Camelina sativa genomic DNA were extracted using urea buffer (7M urea, 350mM NaCl, 50mM Tris-HCl pH8, 20mM EDTA, 1% N-lauroyl sarcosine) and mixed with phenol:chloroform:isoamyl alcohol 25:24:1. The supernatants containing DNA were further precipitated using 3M NaOAc (pH 5.2) and isopropanol followed by 70% ethanol wash. The DNA pellet was resuspended in UltraPure™ DNase/RNase-Free Distilled Water (Invitrogen) followed by RNase A (Roche) treatment and ethanol precipitation. DAP-seq gDNA libraries were constructed following the protocol of Bartlett et al. (2017) with minor modifications. Extracted gDNA were fragmented to the size range between 200-400 bp using Diagenode's Bioruptor® 300 for 40 cycles with 30 seconds on/off at high energy. The fragmented DNA was further used for end repair and adapter ligation. To create modification-free DNA, additional 11 cycles of PCR amplifications were performed using the adapter ligated libraries, followed by ethanol precipitation. Finally, the amplified gDNA libraries (ampDAP) were used for all protein-DNA interaction procedures. The Halo-tagged TFs were expressed in the wheat germ *in vitro* transcription/translation SP6 promoter system (Promega). All the buffers and procedures for DNA-protein interaction were as published (Bartlett et al., 2017), except that the input gDNA library amount and the final step of library size selection. About 200 ng of ampDAP gDNA library were added as an input to mix with each pIX-HALO-TF protein. Finally, to perform double-size selection targeting 300-400 bp fragments, 0.7 volume of Agencourt AMPure XP beads (35 µl) to 1 volume of sample (50 µl) were mixed for 5 minutes and the bead was discarded to remove fragments with size larger than 400 bp. Next, the supernatant (85 µl) containing < 400 bp fragments were added to 0.2 volume of Agencourt AMPure XP beads (10 µl) and mixed for 15 minutes. The bound fragments were eluted from the beads by adding 18 ml UltraPure™ DNase/RNase-Free

Distilled Water (Invitrogen). The concentrations of eluted DNA were measured using the Qubit HS dsDNA assay kit, and approximate 5-20 ng/µl final concentrations were obtained. The fragment size and binding capacity to the flow cell were further examined on the agarose gel by six cycles of PCR using 2 µl of eluted ampDAP-seq library with Illumina P5 and P7 primers. Twelve libraries were pooled in one lane and sequenced by Illumina HiSeq 4000 SE50 at RTSF Genomics core at Michigan State University.

3.5.5 Data processing, quantification, and statistical analyses

RNA-seq. Sample quality control was performed using FastQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/, V0.11.5). Adapters and low quality reads were trimmed with Trimmomatic (Bolger et al., 2014) using the following parameters: ILLUMINACLIP:Adapter.fastq:2:40:15 SLIDINGWINDOW:4:20 MINLEN:30. Cleaned reads were mapped to the reference genome (V2.0, http://camelinadb.ca) using HISAT2 (2.0.4) (Kim *et al.*, 2019) with default parameters. Reads aligned to genes were counted with the R package Rsubread v1.32.2 (Liao et al., 2019), using default parameters and counting only uniquely-mapped reads (Liao et al., 2019), and the transcript abundance estimated as transcripts per kilobase million (TPM). Arabidopsis RNA-seq samples were re-analyzed using the same pipeline. Cleaned reads were mapped to the TAIR10 Arabidopsis genome (https://www.arabidopsis.org/).

Selection of TF candidates based on co-expression analyses. Camelina lipid-related genes (LRGs), TFs and the whole genome expression data were collected from CamRegbase (https://camregbase.org/, Gomez-Cano et al., 2020). Specifically, we retrieved 2,765 LRGs, 5,590 TF, and TPMs values for 131 publicly available RNA-seq experiments. Mutual information (MI) was used as the co-expression metric and estimated with the R package Parmigene v1.0.2 (Sales and Romualdi, 2011). The top 200 genes with the highest MI (MI value ≥ 1) were assumed as the

co-expressed genes of the corresponding TFs. The significance of the common LRGs and co-expressed genes by TF was tested with a Fisher-Exact Test. TF family enriched on TFs significantly co-expressed with LRGs were characterized using the R package GeneOverlap v1.26.0 (Shen and Sinai, 2020). TF-target genes co-expression analysis was performed using the $\log_2$ of TPMs+1 collected from CamRegBase and generated in this work from seed samples (Table S1). The co-expression was estimated as the Pearson coefficient of the expression of the corresponding TF and target expression profiles and was calculated with the *cor* function implemented in R v3.6.0 (https://www.r-project.org/). Arabidopsis homologs of the corresponding Camelia candidates were collected from CamRegbase (https://camregbase.org/, Gomez-Cano et al., 2020), and the co-expression analysis was performed with the sample pipeline and filters implemented in Camelina's analyzes.

DAP-seq read mapping, filtering, and peak calling. Sample quality control was performed using FastQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/, V0.11.5). Adapters and low quality reads were trimmed with Trimmomatic (Bolger et al., 2014) using the following parameters: ILLUMINACLIP:Adapter.fastq:2:40:15 SLIDINGWINDOW:4:20 MINLEN:30. Cleaned reads were mapped to the reference genome (V2.0, http://camelinadb.ca) with Bowtie2 v2.3.4.1 (Langmead and Salzberg, 2012) and only using nuclear chromosomes. Multi-mapping reads were filtered with Samtools v1.9 (Li et al., 2009): samtools view -q 30. Peaks were called using GEM v3.4 (Guo *et al.*, 2012) using the HALO vector as negative control, and the following parameters: --d Read_Distribution_default.txt --k_min 6 --k_max 15 --k_seqs 2000 --outNP --sl. For TFs with replicates, GEM was called with the replicate mode. Only TFs with >500 predicted peaks were used for further analysis. Arabidopsis DAP-seq samples were re-analyzed using the

same pipeline. Cleaned reads were mapped to the TAIR10 Arabidopsis genome (https://www.arabidopsis.org/).

Peak quality control, and motif enrichment. Summit peaks were extended 50 bps and formatted into SAF files to count mapped reads by peak using the R package Rsubread v1.32.2 (Liao et al., 2019). Read abundance was estimated as counts per kilobase million (CPM) by peak, which were then used to estimate the log_2FC (TF/Halo) of each peak. Peaks with $log_2FC > 0.5$ were used for further analysis. TF binding motifs were identified with the meme-chip tool (-meme-minw 6 -meme-maxw 20) from the MEME suite v5.1.1 (Machanick and Bailey, 2011), using top 1,000 peaks with largest log_2FC. DNA sequences of the corresponding peaks used as input on meme-chip were extracted from the reference genome (V2.0, http://camelinadb.ca) using the *getfasta* function from bedtools v2.26.0 (Quinlan and Hall, 2010). MEME's predicted motifs frequency and distribution were assayed with the FIMO tool from the MEME suite v5.1.1 (Grant et al., 2011). Fimo analysis was performed with default parameters using the total set of peaks selected after the log_2FC filter as the fasta database. Motif Z-scores were estimated as follow: $Z_{motif}score = (X_{mi} - \underline{X_m})/sd(X_m)$, such that $\underline{X_m}$ and $sd(X_m)$ are the average and standard deviation of the total number of significant hits of the motif *m* (Fimo q-value <=0.05) for the corresponding TF *X*. And, X_{mi} represents the total number of significant hits (Fimo q-value <=0.05) of the TF *X* at the peak's position *i* for the corresponding motif *m*. The position *i* was defined as the position at which the motif's first nucleotide did match within the corresponding peak sequence (i.e., position 1 to 100, having 50 as the peak's summit). Peaks without a motif match were filtered out from further analysis. Motif logos were generated using MotifStack (Ou et al., 2018). Peaks were visualized with the Integrated Genome Browser (IGV) (Robinson et al., 2011), for which bam files were converted into bigwig files normalizing mapped reads by bins per million mapped reads (BPM)

using bins 10 bps long. Bigwig files were generated using the bamCoverage tool from deepTools v3.5.0 (Ramírez et al., 2016).

Target genes identification, lipid-related target enrichment, and GO enrichment analysis. Targets genes were assigned based on the peak's summit distance to the closest annotated transcription start site (TSS). We use 3 kpbs around the TSS as the maximum distance threshold. Summit-TSS distances were estimated using the *closest* function from bedtools v2.26.0 (Quinlan and Hall, 2010) as follows: closestBed -a Summit.file.bed -b Cs_TSS.bed -D "b". The TSSs bed file was generated using the current genome annotation available (V2.0, http://camelinadb.ca). Lipid-related target gene enrichment was carried out using the R package GeneOverlap v2.28.0 (Shen and Sinai, 2020), with the total number of Camelina genes annotated as background. GO enrichment on predicted target genes was performed using the R package topGO v2.38.1 (Alexa and Rahnenfuhrer, 2010). The top 10 GO terms were selected based on the P-value filtering out general and redundant terms. Genes GO annotation was retrieved from CamRegBase, which is based on homolog with Arabidopsis (Gomez-Cano et al., 2020).

Target enrichment at pathway level and gene set enrichment analysis (GSEA) algorithm. The identification of TF enriched on target genes associated to specific lipid-related pathways we performed using the gene-pathway annotation introduced in Figure 3.1, and using the R package GeneOverlap v2.28.0 (Shen and Sinai, 2020), with the total number of Camelina genes annotated as background. The GSEA assay was performed with the list of pathways/genes presented in Figure 3.1 using the R package FGSEA v1.18.0 (Korotkevich et al., 2021), and with Pearson Correlation Coefficients (PCCs) as scoring metric. The PCC was estimated as weighted PCC (wPCC) between the corresponding TFs and the current annotated genes in Camelina (V2.0, http://camelinadb.ca). We use in this analysis the same list of expression samples analyzed during

the prediction of TF candidates. Expression values (TPMs) were $\log_2$ transformed, and wPCCs were calculate using the R package wCorr (Version 1.9.1) (Emad and Bailey, 2017)with an optimal threshold of 0.4, which reduces overestimation of the PCC because of similar samples (*e.g.,* biological replicates).

Seed co-expression analysis based on GENIE3. The estimation of potential target genes based on expression was performed with only seed expression data using the GENIE3 algorithm, implemented on the R package GENEI3 v1.14.0 (https://bioconductor.org/packages/release/bioc/html/GENIE3.html) (Huynh-Thu *et al.*, 2010). To identify significant scores, we re-assigned the gene IDs randomly at the expression matrix to then recalculate the corresponding GENIE3 score. This process was replicated 1,000 times in order to generate a null distribution for each potential target gene. The significance of the observed score vs the null distribution was estimated calculated as the significance of the Z-score observed, which calculated as follows:

$$Z_{score} = \frac{Score_{observed} - Average(Score_{random[1,..,1000]})}{\frac{sd(Score_{random})}{\sqrt{total\ random\ values}}}$$

Given the number of comparisons, estimated *P*-values were corrected for multiple testing by Benjamini-Hochberg method (Yoav Benjamini and Yosef Hochberg, 1995).

3.5.6 Data availability and accession numbers

The supporting the findings of this work are available on the supplementary files. Raw sequencing data generated can be found in the NCBI SRA databased under the accession number PRJNA763897. Processed data have been deposited in the NCBI GEO databased under the accession number GSE184283.

REFERENCES

Abdullah, H.M., Akbari, P., Paulose, B., Schnell, D., Qi, W., Park, Y., Pareek, A. and Dhankher, O.P. (2016) Transcriptome profiling of Camelina sativa to identify genes involved in triacylglycerol biosynthesis and accumulation in the developing seeds. *Biotechnol. Biofuels*, **9**, 136.

Alexa, A. and Rahnenfuhrer, J. (2010) topGO: enrichment analysis for gene ontology. *R package version* 2

Alonso, R., Oñate-Sánchez, L., Weltmeier, F., Ehlert, A., Diaz, I., Dietrich, K., Vicente-Carbajosa, J. and Dröge-Laser, W. (2009) A pivotal role of the basic leucine zipper transcription factor bZIP53 in the regulation of Arabidopsis seed maturation gene expression based on heterodimerization and protein complex formation. *Plant Cell*, **21**, 1747–1761.

An, D. and Suh, M.C. (2015) Overexpression of ArabidopsisWRI1 enhanced seed mass and storage oil content in Camelina sativa. *Plant Biotechnol. Rep*., **9**, 137–148.

Bartlett, A., O'Malley, R.C., Huang, S.-S.C., Galli, M., Nery, J.R., Gallavotti, A. and Ecker, J.R. (2017) Mapping genome-wide transcription-factor binding sites using DAP-seq. *Nat. Protoc.*, **12**, 1659–1672.

Baud, S. and Lepiniec, L. (2010) Physiological and developmental regulation of seed oil production. *Prog. Lipid Res.*, **49**, 235–249.

Bäumlein, H., Miséra, S., Luerßen, H., Kölle, K., Horstmann, C., Wobus, U. and Müller, A.J. (1994) The FUS3 gene of Arabidopsis thaliana is a regulator of gene expression during late embryogenesis. *Plant J.*, **6**, 379–387.

Bensmihen, S., Rippa, S., Lambert, G., Jublot, D., Pautot, V., Granier, F., Giraudat, J. and Parcy, F. (2002) The homologous ABI5 and EEL transcription factors function antagonistically to fine-tune gene expression during late embryogenesis. *Plant Cell*, **14**, 1391–1403.

Berti, M., Gesch, R., Eynck, C., Anderson, J. and Cermak, S. (2016) Camelina uses, genetics, genomics, production, and management. *Ind. Crops Prod.*, **94**, 690–710.

Birkenbihl, R.P., Kracher, B., Ross, A., Kramer, K., Finkemeier, I. and Somssich, I.E. (2018) Principles and characteristics of the Arabidopsis WRKY regulatory network during early MAMP-triggered immunity. *Plant J.*, **96**, 487–502.

Bou-Torrent, J., Salla-Martret, M., Brandt, R., Musielak, T., Palauqui, J.-C., Martínez-García, J.F. and Wenkel, S. (2012) ATHB4 and HAT3, two class II HD-ZIP transcription factors, control leaf development in Arabidopsis. *Plant Signal. Behav.*, **7**, 1382–1387.

Bolger, A.M., Lohse, M. and Usadel, B. (2014) Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*, **30**, 2114–2120.

Braybrook, S.A. and Harada, J.J. (2008) LECs go crazy in embryo development. *Trends Plant Sci.*, **13**, 624–630.

Broun, P. (2004) Transcription factors as tools for metabolic engineering in plants. *Curr. Opin. Plant Biol.*, **7**, 202–209.

Carabelli, M., Sessa, G., Baima, S., Morelli, G. and Ruberti, I. (1993) The Arabidopsis Athb-2 and -4 genes are strongly induced by far-red-rich light. *Plant J.*, **4**, 469–479.

Carlsson, A.S. (2009) Plant oils as feedstock alternatives to petroleum – A short survey of potential oil crop platforms. *Biochimie*, **91**, 665–670. Available at: http://dx.doi.org/10.1016/j.biochi.2009.03.021.

Century, K., Reuber, T.L. and Ratcliffe, O.J. (2008) Regulating the regulators: the future prospects for transcription-factor-based agricultural biotechnology products. *Plant Physiol.*, **147**, 20–29.

Cernac, A. and Benning, C. (2004) WRINKLED1 encodes an AP2/EREB domain protein involved in the control of storage compound biosynthesis in Arabidopsis. *Plant J.*, **40**, 575–585.

Chandrasekaran, U., Luo, X., Zhou, W. and Shu, K. (2020) Multifaceted Signaling Networks Mediated by Abscisic Acid Insensitive 4. *Plant Commun*, **1**, 100040.

Chaturvedi, S., Bhattacharya, A., Khare, S.K. and Kaushik, G. (2019) Camelina sativa: An Emerging Biofuel Crop. In C. M. Hussain, ed. *Handbook of Environmental Materials Management*. Cham: Springer International Publishing, pp. 2889–2925.

Curaba, J., Moritz, T., Blervaque, R., Parcy, F., Raz, V., Herzog, M. and Vachon, G. (2004) AtGA3ox2, a key gene responsible for bioactive gibberellin biosynthesis, is regulated during embryogenesis by LEAFY COTYLEDON2 and FUSCA3 in Arabidopsis. *Plant Physiol.*, **136**, 3660–3669.

Devic, M. and Roscoe, T. (2016) Seed maturation: Simplification of control networks in plants. *Plant Sci.*, **252**, 335–346.

Emad, A. and Bailey, P. (2017) wCorr: weighted correlations. *R package version*. 1.9. 1.

Fletcher, J.C. (2001) The ULTRAPETALA gene controls shoot and floral meristem size in Arabidopsis. *Development*, **128**, 1323–1333.

Focks, N. and Benning, C. (1998) wrinkled1: a novel, low-seed-oil mutant of Arabidopsis with a deficiency in the seed-specific regulation of carbohydrate metabolism. *Plant Physiol.*, **118**, 91–101.

Footitt, S., Marquez, J., Schmuths, H., Baker, A., Theodoulou, F.L. and Holdsworth, M. (2006) Analysis of the role of COMATOSE and peroxisomal beta-oxidation in the determination of germination potential in Arabidopsis. *J. Exp. Bot.*, **57**, 2805–2814.

Fulda, M., Schnurr, J., Abbadi, A., Heinz, E. and Browse, J. (2004) Peroxisomal Acyl-CoA synthetase activity is essential for seedling development in Arabidopsis thaliana. *Plant Cell*, **16**, 394–405.

Galli, M., Khakhar, A., Lu, Z., Chen, Z., Sen, S., Joshi, T., Nemhauser, J.L., Schmitz, R.J. and Gallavotti, A. (2018) The DNA binding landscape of the maize AUXIN RESPONSE FACTOR family. *Nat. Commun.*, **9**, 4526.

Ghosh, A.K., Chauhan, N., Rajakumari, S., Daum, G. and Rajasekharan, R. (2009) At4g24160, a soluble acyl-coenzyme A-dependent lysophosphatidic acid acyltransferase. *Plant Physiol.*, **151**, 869–881.

Giraudat, J., Hauge, B.M., Valon, C., Smalle, J., Parcy, F. and Goodman, H.M. (1992) Isolation of the Arabidopsis ABI3 gene by positional cloning. *Plant Cell*, **4**, 1251–1261.

Grant, C.E., Bailey, T.L. and Noble, W.S. (2011) FIMO: scanning for occurrences of a given motif. *Bioinformatics*, **27**, 1017–1018.

Gomez-Cano, F., Carey, L., Lucas, K., García Navarrete, T., Mukundi, E., Lundback, S., Schnell, D. and Grotewold, E. (2020) CamRegBase: a gene regulation database for the biofuel crop, Camelina sativa. *Database* , **2020**. Available at: http://dx.doi.org/10.1093/database/baaa075.

Go, Y.S., Kim, H., Kim, H.J. and Suh, M.C. (2014) Arabidopsis Cuticular Wax Biosynthesis Is Negatively Regulated by the DEWAX Gene Encoding an AP2/ERF-Type Transcription Factor. *Plant Cell*, **26**, 1666–1680.

Grotewold, E. (2008) Transcription factors for predictive plant metabolic engineering: are we there yet? *Curr. Opin. Biotechnol*., **19**, 138–144.

Guerriero, G., Martin, N., Golovko, A., Sundström, J.F., Rask, L. and Ezcurra, I. (2009) The RY/Sph element mediates transcriptional repression of maturation genes from late maturation to early seedling growth. *New Phytol.*, **184**, 552–565.

Gugel, R.K. and Falk, K.C. (2006) Agronomic and seed quality evaluation of Camelina sativa in western Canada. *Can. J. Plant Sci.*, **86**, 1047–1058.

Guo, Y., Mahony, S. and Gifford, D.K. (2012) High resolution genome wide binding event finding and motif discovery reveals transcription factor spatial binding constraints. *PLoS Comput. Biol.*, **8**, e1002638.

Huynh-Thu, V.A., Irrthum, A., Wehenkel, L. and Geurts, P. (2010) Inferring regulatory networks from expression data using tree-based methods. *PLoS One*, **5**. Available at: http://dx.doi.org/10.1371/journal.pone.0012776.

Iskandarov, U., Kim, H.J. and Cahoon, E.B. (2014) Camelina: An emerging oilseed platform for advanced biofuels and bio-based materials. In MC McCann, M. S. Buckeridge, and N. C. Carpita, eds. *Plants and BioEnergy*. New York: Springer, pp. 131–140.

Jiang, T., Zhang, X.-F., Wang, X.-F. and Zhang, D.-P. (2011) Arabidopsis 3-ketoacyl-CoA thiolase-2 (KAT2), an enzyme of fatty acid β-oxidation, is involved in ABA signal transduction. *Plant Cell Physiol.*, **52**, 528–538.

Kagale, S., Koh, C., Nixon, J., et al. (2014) The emerging biofuel crop Camelina sativa retains a highly undifferentiated hexaploid genome structure. *Nat. Commun.*, **5**, 1–11.

Kagaya, Y., Okuda, R., Ban, A., Toyoshima, R., Tsutsumida, K., Usui, H., Yamamoto, A. and Hattori, T. (2005) Indirect ABA-dependent regulation of seed storage protein genes by FUSCA3 transcription factor in Arabidopsis. *Plant Cell Physiol.*, **46**, 300–311.

Keith, K., Kraml, M., Dengler, N.G. and McCourt, P. (1994) fusca3: A Heterochronic Mutation Affecting Late Embryo Development in Arabidopsis. *Plant Cell*, **6**, 589–600.

Kim, D., Paggi, J.M., Park, C., Bennett, C. and Salzberg, S.L. (2019) Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nat. Biotechnol.*, **37**, 907–915.

Kong, Q. and Ma, W. (2018) WRINKLED1 transcription factor: How much do we know about its regulatory mechanism? *Plant Science*, **272**, 153–156. Available at: http://dx.doi.org/10.1016/j.plantsci.2018.04.013.

Kong, Q., Singh, S.K., Mantyla, J.J., Pattanaik, S., Guo, L., Yuan, L., Benning, C. and Ma, W. (2020) TEOSINTE BRANCHED1/CYCLOIDEA/PROLIFERATING CELL FACTOR4 Interacts with WRINKLED1 to Mediate Seed Oil Biosynthesis. *Plant Physiol.*, **184**, 658–665.

Kong, Q., Yang, Y., Guo, L., Yuan, L. and Ma, W. (2020) Molecular Basis of Plant Oil Biosynthesis: Insights Gained From Studying the WRINKLED1 Transcription Factor. *Front. Plant Sci.*, **11**, 24.

Korotkevich, G., Sukhov, V., Budin, N., Shpak, B., Artyomov, M.N. and Sergushichev, A. (2021) Fast gene set enrichment analysis. *bioRxiv*, 060012.

Kosma, D.K., Murmu, J., Razeq, F.M., Santos, P., Bourgault, R., Molina, I. and Rowland, O. (2014) AtMYB41 activates ectopic suberin synthesis and assembly in multiple plant species and cell types. *Plant J.*, **80**, 216–229.

Kryuchkova-Mostacci, N. and Robinson-Rechavi, M. (2017) A benchmark of gene expression tissue-specificity metrics. *Brief. Bioinform.*, **18**, 205–214.

Kurdyukov, S., Faust, A., Trenkamp, S., et al. (2006) Genetic and biochemical evidence for involvement of HOTHEAD in the biosynthesis of long-chain alpha-,omega-dicarboxylic fatty acids and formation of extracellular matrix. *Planta*, **224**, 315–329.

Kwong, R.W., Bui, A.Q., Lee, H., Kwong, L.W., Fischer, R.L., Goldberg, R.B. and Harada, J.J. (2003) LEAFY COTYLEDON1-LIKE defines a class of regulators essential for embryo development. *Plant Cell*, **15**, 5–18.

Lai, Z., Vinod, K., Zheng, Z., Fan, B. and Chen, Z. (2008) Roles of Arabidopsis WRKY3 and

WRKY4 transcription factors in plant responses to pathogens. *BMC Plant Biol.*, **8**, 68.

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

Lambert, S.A., Jolma, A., Campitelli, L.F., et al. (2018) The Human Transcription Factors. *Cell*, **175**, 598–599.

Lashbrooke, J., Cohen, H., Levy-Samocha, D., et al. (2016) MYB107 and MYB9 Homologs Regulate Suberin Deposition in Angiosperms. *Plant Cell*, **28**, 2097–2116.

Le, B.H., Cheng, C., Bui, A.Q., et al. (2010) Global analysis of gene activity during Arabidopsis seed development and identification of seed-specific transcription factors. *Proc. Natl. Acad. Sci. U. S. A.*, **107**, 8063–8070.

Lee, S.B., Kim, H., Kim, R.J. and Suh, M.C. (2014) Overexpression of Arabidopsis MYB96 confers drought resistance in Camelina sativa via cuticular wax accumulation. *Plant Cell Rep.*, **33**, 1535–1546.

Lee, S.B., Kim, H.U. and Suh, M.C. (2016) MYB94 and MYB96 Additively Activate Cuticular Wax Biosynthesis in Arabidopsis. *Plant Cell Physiol.*, **57**, 2300–2311.

Lee, S.B. and Suh, M.C. (2015) Cuticular wax biosynthesis is up-regulated by the MYB94 transcription factor in Arabidopsis. *Plant Cell Physiol.*, **56**, 48–60.

Leprince, O., Pellizzaro, A., Berriri, S. and Buitink, J. (2016) Late seed maturation: drying without dying. *J. Exp. Bot.*, **68**, 827–841. Available at: [Accessed July 7, 2021].

Liao, Y., Smyth, G.K. and Shi, W. (2019) The R package Rsubread is easier, faster, cheaper and better for alignment and quantification of RNA sequencing reads. *Nucleic Acids Res.*, **47**, e47–e47.

Liang, C., Liu, X., Yiu, S.-M. and Lim, B.L. (2013) De novo assembly and characterization of Camelina sativa transcriptome by paired-end sequencing. *BMC Genomics*, **14**, 146.

Li-Beisson, Y., Shorrosh, B., Beisson, F., et al. (2013) Acyl-lipid metabolism. *Arabidopsis Book*, **11**, e0161.

Li, H., Handsaker, B., Wysoker, A., et al. (2009) The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, **25**, 2078–2079.

Li, D., Jin, C., Duan, S., et al. (2017) MYB89 Transcription Factor Represses Seed Oil Accumulation. *Plant Physiol.*, **173**, 1211–1225.

Li, P., Zhou, H., Shi, X., et al. (2014) The ABI4-induced Arabidopsis ANAC060 transcription factor attenuates ABA signaling and renders seedlings sugar insensitive when present in the nucleus. *PLoS Genet.*, **10**, e1004213.

Li, Y., Beisson, F., Pollard, M. and Ohlrogge, J. (2006) Oil content of Arabidopsis seeds: the influence of seed anatomy, light and plant-to-plant variation. *Phytochemistry*, **67**, 904–915.

Liu, P., Nie, W.-F., Xiong, X., et al. (2021) A novel protein complex that regulates active DNA demethylation in Arabidopsis. *J. Integr. Plant Biol.*, **63**, 772–786.

Lotan, T., Ohto, M., Yee, K.M., et al. (1998) Arabidopsis LEAFY COTYLEDON1 is sufficient to induce embryo development in vegetative cells. *Cell*, **93**, 1195–1205.

Lumba, S., Tsuchiya, Y., Delmas, F., Hezky, J., Provart, N.J., Shi Lu, Q., McCourt, P. and Gazzarrini, S. (2012) The embryonic leaf identity gene FUSCA3 regulates vegetative phase transitions by negatively modulating ethylene-regulated gene expression in Arabidopsis. *BMC Biol.*, **10**, 8.

Machanick, P. and Bailey, T.L. (2011) MEME-ChIP: motif analysis of large DNA datasets. *Bioinformatics*, **27**, 1696–1697.

Malik, M.R., Tang, J., Sharma, N., Burkitt, C., Ji, Y., Mykytyshyn, M., Bohmert-Tatarev, K., Peoples, O. and Snell, K.D. (2018) Camelina sativa, an oilseed at the nexus between model system and commercial crop. *Plant Cell Rep.*, **37**, 1367–1381.

Mandáková, T., Pouch, M., Brock, J.R., Al-Shehbaz, I.A. and Lysak, M.A. (2019) Origin and Evolution of Diploid and Allopolyploid Camelina Genomes Were Accompanied by Chromosome Shattering. *Plant Cell*, **31**, 2596–2612.

Meinke, D.W., Franzmann, L.H., Nickle, T.C. and Yeung, E.C. (1994) Leafy Cotyledon Mutants of Arabidopsis. *Plant Cell*, **6**, 1049–1064.

Mendes, A., Kelly, A.A., Erp, H. van, Shaw, E., Powers, S.J., Kurup, S. and Eastmond, P.J. (2013) bZIP67 regulates the omega-3 fatty acid content of Arabidopsis seed oil by activating fatty acid desaturase3. *Plant Cell*, **25**, 3104–3116.

Morineau, C., Bellec, Y., Tellier, F., Gissot, L., Kelemen, Z., Nogué, F. and Faure, J.-D. (2017) Selective gene dosage by CRISPR-Cas9 genome editing in hexaploid Camelina sativa. *Plant Biotechnol. J.*, **15**, 729–739.

Mudalkar, S., Golla, R., Ghatty, S. and Reddy, A.R. (2014) De novo transcriptome analysis of an imminent biofuel crop, Camelina sativa L. using Illumina GAIIX sequencing platform and identification of SSR markers. *Plant Mol. Biol.*, **84**, 159–171.

Neumann, N.G., Nazarenus, T.J., Aznar-Moreno, J.A., Rodriguez-Aponte, S.A., Mejias Veintidos, V.A., Comai, L., Durrett, T.P. and Cahoon, E.B. (2021) Generation of camelina mid-oleic acid seed oil by identification and stacking of fatty acid biosynthetic mutants. *Ind. Crops Prod.*, **159**, 113074.

Nguyen, H.T., Silva, J.E., Podicheti, R., et al. (2013) Camelina seed transcriptome: a tool for meal and oil improvement and translational research. *Plant Biotechnol. J.*, **11**, 759–769.

Nikolov, L.A., Shushkov, P., Nevado, B., Gan, X., Al-Shehbaz, I.A., Filatov, D., Bailey, C.D. and Tsiantis, M. (2019) Resolving the backbone of the Brassicaceae phylogeny for investigating trait diversity. *New Phytol.*, **222**, 1638–1651.

O'Malley, R.C., Huang, S.S.C., Song, L., Lewsey, M.G., Bartlett, A., Nery, J.R., Galli, M., Gallavotti, A. and Ecker, J.R. (2016) Cistrome and Epicistrome Features Shape the Regulatory DNA Landscape. *Cell*, **165**, 1280–1292.

O'Neill, C.M., Baker, D., Bennett, G., Clarke, J. and Bancroft, I. (2011) Two high linolenic mutants of Arabidopsis thaliana contain megabase-scale genome duplications encompassing the FAD3 locus. *Plant J.*, **68**, 912–918.

Ornelas-Ayala, D., Vega-León, R., Petrone-Mendoza, E., Garay-Arroyo, A., García-Ponce, B., Álvarez-Buylla, E.R. and Sanchez, M. de la P. (2020) ULTRAPETALA1 maintains Arabidopsis root stem cell niche independently of ARABIDOPSIS TRITHORAX1. *New Phytol.*, **225**, 1261–1272.

Ozseyhan, M.E., Kang, J., Mu, X. and Lu, C. (2018) Mutagenesis of the FAE1 genes significantly changes fatty acid composition in seeds of Camelina sativa. *Plant Physiol. Biochem.*, **123**, 1–7.

Ou, J., Wolfe, S.A., Brodsky, M.H. and Zhu, L.J. (2018) motifStack for the analysis of transcription factor binding site evolution. *Nat. Methods*, **15**, 8-9.

Pelletier, J.M., Kwong, R.W., Park, S., et al. (2017) LEC1 sequentially regulates the transcription of genes involved in diverse developmental processes during seed development. *Proc. Natl. Acad. Sci. U. S. A.*, **114**, E6710–E6719.

Pires, H.R., Shemyakina, E.A. and Fletcher, J.C. (2015) The ULTRAPETALA1 trxG factor contributes to patterning the Arabidopsis adaxial-abaxial leaf polarity axis. *Plant Signal. Behav.*, **10**, e1034422.

Pollard, M., Martin, T.M. and Shachar-Hill, Y. (2015) Lipid analysis of developing Camelina sativa seeds and cultured embryos. *Phytochemistry*, **118**, 23–32.

Pouvreau, B., Blundell, C., Vohra, H., Zwart, A.B., Arndell, T., Singh, S. and Vanhercke, T. (2020) A Versatile High Throughput Screening Platform for Plant Metabolic Engineering Highlights the Major Role of ABI3 in Lipid Metabolism Regulation. *Front. Plant Sci.*, **11**, 288.

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

Ramírez, F., Ryan, D.P., Grüning, B., Bhardwaj, V., Kilpert, F., Richter, A.S., Heyne, S., Dündar, F. and Manke, T. (2016) deepTools2: a next generation web server for deep-sequencing data analysis. *Nucleic Acids Res.*, **44**, W160–5.

Reiter, F., Wienerroither, S. and Stark, A. (2017) Combinatorial function of transcription factors and cofactors. *Curr. Opin. Genet. Dev.*, **43**, 73–81.

Richmond, T.A. and Bleecker, A.B. (1999) A defect in beta-oxidation causes abnormal inflorescence development in Arabidopsis. *Plant Cell*, **11**, 1911–1924.

Robinson, J.T., Thorvaldsdóttir, H., Winckler, W., Guttman, M., Lander, E.S., Getz, G. and Mesirov, J.P. (2011) Integrative genomics viewer. *Nat. Biotechnol.*, **29**, 24–26.

Rodríguez-Rodríguez, M.F., Sánchez-García, A., Salas, J.J., Garcés, R. and Martínez-Force, E. (2013) Characterization of the morphological changes and fatty acid profile of developing Camelina sativa seeds. *Ind. Crops Prod.*, **50**, 673–679.

Romanel, E.A.C., Schrago, C.G., Couñago, R.M., Russo, C.A.M. and Alves-Ferreira, M. (2009) Evolution of the B3 DNA binding superfamily: new insights into REM family gene diversification. *PLoS One*, **4**, e5791.

Roudier, F., Gissot, L., Beaudoin, F., et al. (2010) Very-long-chain fatty acids are involved in polar auxin transport and developmental patterning in Arabidopsis. *Plant Cell*, **22**, 364–375.

Sadler, C., Schroll, B., Zeisler, V., Waßmann, F., Franke, R. and Schreiber, L. (2016) Wax and cutin mutants of Arabidopsis: Quantitative characterization of the cuticular transport barrier in relation to chemical composition. *Biochim. Biophys. Acta*, **1861**, 1336–1344.

Saez, A., Rodrigues, A., Santiago, J., Rubio, S. and Rodriguez, P.L. (2008) HAB1-SWI3B interaction reveals a link between abscisic acid signaling and putative SWI/SNF chromatin-remodeling complexes in Arabidopsis. *Plant Cell*, **20**, 2972–2988.

Sales, G. and Romualdi, C. (2011) Parmigene-a parallel R package for mutual information estimation and gene network reconstruction. *Bioinformatics*, **27**, 1876–1877.

Sarnowski, T.J., Ríos, G., Jásik, J., et al. (2005) SWI3 subunits of putative SWI/SNF chromatin-remodeling complexes play distinct roles during Arabidopsis development. *Plant Cell*, **17**, 2454–2472.

Seo, P.J., Lee, S.B., Suh, M.C., Park, M.-J., Go, Y.S. and Park, C.-M. (2011) The MYB96 transcription factor regulates cuticular wax biosynthesis under drought conditions in Arabidopsis. *Plant Cell*, **23**, 1138–1152.

Shang, B., Xu, C., Zhang, X., Cao, H., Xin, W. and Hu, Y. (2016) Very-long-chain fatty acids restrict regeneration capacity by confining pericycle competence for callus formation in Arabidopsis. *Proc. Natl. Acad. Sci. U. S. A.*, **113**, 5101–5106.

Shen, B., Sinkevicius, K.W., Selinger, D.A. and Tarczynski, M.C. (2006) The homeobox gene GLABRA2 affects seed oil content in Arabidopsis. *Plant Mol. Biol.*, **60**, 377–387.

Shen, L. and Sinai, I. (2020) GeneOverlap: Test and visualize gene overlaps. *R package 1.26.0*

Skubacz, A., Daszkowska-Golec, A. and Szarejko, I. (2016) The Role and Regulation of ABI5 (ABA-Insensitive 5) in Plant Development, Abiotic Stress Responses and Phytohormone Crosstalk. *Front. Plant Sci.*, **7**, 1884.

Sorin, C., Salla-Martret, M., Bou-Torrent, J., Roig-Villanova, I. and Martínez-García, J.F. (2009) ATHB4, a regulator of shade avoidance, modulates hormone response in Arabidopsis seedlings. *Plant J.*, **59**, 266–277.

Springer, N., León, N. de and Grotewold, E. (2019) Challenges of Translating Gene Regulatory Information into Agronomic Improvements. *Trends Plant Sci.*, **Dic**, 1075–1082.

Stone, S.L., Kwong, L.W., Yee, K.M., Pelletier, J., Lepiniec, L., Fischer, R.L., Goldberg, R.B. and Harada, J.J. (2001) LEAFY COTYLEDON2 encodes a B3 domain transcription factor that induces embryo development. *Proc. Natl. Acad. Sci. U. S. A.*, **98**, 11806–11811.

Subramanian, A., Tamayo, P., Mootha, V.K., et al. (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. U. S. A.*, **102**, 15545–15550.

Suzuki, M. and McCarty, D.R. (2008) Functional symmetry of the B3 network controlling seed development. *Curr. Opin. Plant Biol.*, **11**, 548–553.

Tang, L.P., Zhou, C., Wang, S.S., Yuan, J., Zhang, X.S. and Su, Y.H. (2017) FUSCA3 interacting with LEAFY COTYLEDON2 controls lateral root formation through regulating YUCCA4 gene expression in Arabidopsis thaliana. *New Phytol.*, **213**, 1740–1754.

Tian, R., Paul, P., Joshi, S. and Perry, S.E. (2020) Genetic activity during early plant embryogenesis. *Biochem. J*, **477**, 3743–3767.

Tiedemann, J., Rutten, T., Mönke, G., et al. (2008) Dissection of a complex seed phenotype: novel insights of FUSCA3 regulated developmental processes. *Dev. Biol.*, **317**, 1–12.

To, A., Joubès, J., Barthole, G., Lécureuil, A., Scagnelli, A., Jasinski, S., Lepiniec, L. and Baud, S. (2012) WRINKLED Transcription Factors Orchestrate Tissue-Specific Regulation of Fatty Acid Biosynthesis in Arabidopsis. *The Plant Cell*, **24**, 5007–5023. Available at: http://dx.doi.org/10.1105/tpc.112.106120.

Trigg, S.A., Garza, R.M., MacWilliams, A., et al. (2017) CrY2H-seq: a massively multiplexed assay for deep-coverage interactome mapping. *Nat. Methods*, **14**, 819–825.

Troncoso-Ponce, M.A., Barthole, G., Tremblais, G., To, A., Miquel, M., Lepiniec, L. and Baud, S. (2016) Transcriptional Activation of Two Delta-9 Palmitoyl-ACP Desaturase Genes by MYB115 and MYB118 Is Critical for Biosynthesis of Omega-7 Monounsaturated Fatty Acids in the Endosperm of Arabidopsis Seeds. *Plant Cell*, **28**, 2666–2682.

Tsukagoshi, H., Morikami, A. and Nakamura, K. (2007) Two B3 domain transcriptional repressors prevent sugar-inducible expression of seed maturation genes in Arabidopsis seedlings. *Proc. Natl. Acad. Sci. U. S. A.*, **104**, 2543–2547.

Yoav Benjamini and Yosef Hochberg (1995) Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *J. R. Stat. Soc. Series B Stat. Methodol.*, **57**, 289–300.

Tyler, L., Miller, M.J. and Fletcher, J.C. (2019) The Trithorax Group Factor ULTRAPETALA1 Regulates Developmental as Well as Biotic and Abiotic Stress Response Genes in Arabidopsis. *G3*, **9**, 4029–4043.

Voelker, T. and Kinney, A.J. (2001) VARIATIONS IN THE BIOSYNTHESIS OF SEED-STORAGE LIPIDS. *Annu. Rev. Plant Physiol. Plant Mol. Biol.*, **52**, 335–361.

Wang, F. and Perry, S.E. (2013) Identification of direct targets of FUSCA3, a key regulator of Arabidopsis seed development. *Plant Physiol.*, **161**, 1251–1264.

Wang, X., Niu, Q.-W., Teng, C., Li, C., Mu, J., Chua, N.-H. and Zuo, J. (2009) Overexpression of PGA37/MYB118 and MYB115 promotes vegetative-to-embryonic transition in Arabidopsis. *Cell Res.*, **19**, 224–235.

Yamamoto, A., Kagaya, Y., Usui, H., Hobo, T., Takeda, S. and Hattori, T. (2010) Diverse roles and mechanisms of gene regulation by the Arabidopsis seed maturation master regulator FUS3 revealed by microarray analysis. *Plant Cell Physiol.*, **51**, 2031–2046.

Yu, B., Wang, Y., Zhou, H., Li, P., Liu, C., Chen, S., Peng, Y., Zhang, Y. and Teng, S. (2020) Genome-wide binding analysis reveals that ANAC060 directly represses sugar-induced transcription of ABI5 in Arabidopsis. *Plant J.*, **103**, 965–979.

Zhai, Z., Liu, H. and Shanklin, J. (2017) Phosphorylation of WRINKLED1 by KIN10 Results in Its Proteasomal Degradation, Providing a Link between Energy Homeostasis and Lipid Biosynthesis. *Plant Cell*, **29**, 871–889.

Zhang, M., Cao, X., Jia, Q. and Ohlrogge, J. (2016) FUSCA3 activates triacylglycerol accumulation in Arabidopsis seedlings and tobacco BY2 cells. *Plant J.*, **88**, 95–107.

Zheng, Q., Zheng, Y. and Perry, S.E. (2013) AGAMOUS-Like15 promotes somatic embryogenesis in Arabidopsis and soybean in part by the control of ethylene biosynthesis and response. *Plant Physiol.*, **161**, 2113–2127.

Zheng, Y., Ren, N., Wang, H., Stromberg, A.J. and Perry, S.E. (2009) Global identification of targets of the Arabidopsis MADS domain protein AGAMOUS-Like15. *Plant Cell*, **21**, 2563–2577.

CHAPTER FOUR: MULTI-NETWORK INTEGRATION TO PRIORITIZE REGULATORY GENES OF METABOLISM IN MAIZE

4.1 ABSTRACT

Elucidating gene regulatory networks (GRNs) is a major area of study within plant systems biology. Phenotypic traits are intricately linked to specific gene expression profiles. These expression patterns primarily arise from regulatory connections between sets of transcription factors (TFs) and their target genes. In this study, I integrated publicly available co-expression networks encompassing over 6,000 RNA-seq samples, approximately 16 million SNPs, and around 300 protein-DNA interaction assays, which comprised 245 ChIP-seq and 38/ DAP-seq assays. Overall, I constructed four distinct types of TF-target networks, including co-expression, protein-DNA interaction (PDI), *trans*-expression quantitative loci (*trans*-eQTL), and *cis*-eQTL combined with PDIs. In total, I analyzed ~4.6M interactions. I implemented three different strategies to integrate these four types of networks, and performed evaluation of the method based on knockouts and random networks. These results identify transcriptional regulators of different biological processes, including hormone-, metabolic- and development-related processes. Finally, using the topological properties of the full integrated network I identify potentially functional redundant TF paralogs. Our findings retrieve functions previously documented for numerous TFs and reveal novel functions that are crucial for informing the design of future experiments. Moreover, I am laying the foundation for the integration of multi-omic datasets in maize and other plant systems.

4.2 INTRODUCTION

Plant cells, like those of other organisms, use multiple interconnected molecular layers which collaboratively coordinate every cellular process, from cellular division to metabolite synthesis and adaptation to environment changes. Among these molecular layers, transcription factor (TF) proteins play a vital role in controlling the expression of other genes (known as target genes)

(Gupta et al., 2021). The regulation requires the direct protein-DNA interactions (PDI) between TFs and specific *cis*-regulatory elements (CREs) located in close (promoters) or distal (enhancers/silencers) regulatory regions of the corresponding target genes (Schmitz et al., 2022). Furthermore, TFs can also regulate gene expression indirectly by engaging in protein-protein interactions (PPI) with other proteins. Together, collections of PDIs form highly interconnected gene regulatory networks (GRNs). Overall, GRN are characterized by gene-centered and TF-centered approaches (Arda and Walhout, 2010; Mejia-Guerra et al., 2012; Yang et al., 2016). Common gene-centered methods include yeast one-hybrid (Y1H) assay and electrophoretic mobility shift assay (EMSA) (Arda and Walhout, 2010; Yang et al., 2016). TF-centered strategies involve techniques like ChIP-seq for *in vivo* TF binding site discovery, and SELEX, PBM, and DAP-seq for *in vitro* analysis (Yang et al., 2016; O'Malley et al., 2016). The organization of GRNs has implications for phenotypic variation (Deplancke et al., 2006), plant responses to abiotic and biotic stress (Nakashima et al., 2014; Birkenbihl et al., 2017; Sun et al., 2022), development (Marand et al., 2023), speciation (Mack and Nachman, 2017), as well as adaptation and diversification (Mack and Nachman, 2017; Bowles et al., 2020; Marand et al., 2023), among others. Therefore, it is crucial to comprehend the structure and dynamics of these networks, emphasizing the significance and rationale behind such endeavors.

Maize holds great agricultural significance due to its versatility and wide range of applications. It serves as a staple food, particularly in sub-Saharan Africa and Latin America, and is also used for animal feed, meat production, dairy, and poultry products (Erenstein et al., 2022). Part of maize versatility underlines its extraordinary maize metabolic diversity (Riedelsheimer et al., 2012; Wen et al., 2014, 2016; Zhou et al., 2019), which is established by its genetic diversity (Schnable et al., 2009; Hufford et al., 2021; McMullen et al., 2009), and varies as a function of endogenous

variables (e.g. organs or development stages) (Zhou et al., 2019) and environment factors (Wen et al., 2014; Kusmec et al., 2017). Irrespective of the methodology employed and the traits analyzed, maize has consistently shown a complex genetic architecture, as indicated by the number of loci potentially linked to a single trait and their minor contribution to the variance of the corresponding traits (Riedelsheimer et al., 2012; Wen et al., 2014, 2016; Xiao et al., 2017; Mazaheri et al., 2019; Zhou et al., 2019). These challenges present two primary obstacles in comprehending the molecular mechanisms underlying these traits: the involvement of multiple genes in a single phenotypic trait and the influence of additional genetic factors that determine and modulate the genetic contribution to phenotypic variations.

A distinctive characteristic of maize is its genome itself, which has undergone a recent whole genome duplication (WGD) event ~5-12 Mya, and is currently defined as an ancient tetraploid (Wei et al., 2007). It has an abundance of tandem duplicated genes (Kono et al., 2018), and is highly-enriched in transposable elements (~85%) (Schnable et al., 2009). Furthermore, the WGD event resulted in the formation of two subgenomes (maize1 and maize2), exhibiting unequal gene loss and expression patterns, primarily driven by the subgenome with a lower fraction rate (Schnable et al., 2011). The dominant subgenome (i.e., maize1) was shown to have a larger contribution to phenotypic variations (Renny-Byfield et al., 2017). Nevertheless, the precise molecular mechanisms underlying the asymmetric contributions of each subgenome remain largely unknown and are likely orchestrated at multiple molecular levels, including regulation, signaling, and interactome level, as evidenced by examining co-expression and multi-network comparisons of homeologs (Li et al., 2016; Han et al., 2023). Understanding these mechanisms holds significant implications for modeling, prioritizing, and unraveling the principal factors

behind agriculturally relevant traits, while also advancing our fundamental comprehension of maize evolution.

The multi-omic data sets have gained attention as alternatives to address the complexity of genetic and phenotypic variation observed in biological systems, offering insights at various levels of a biological process (Tolani et al., 2021). Integrating these diverse datasets is an evolving field, with four main approaches: conceptual integration (overlapping observations), statistical integration, network-based integration, and machine learning-based integration (Depuydt et al., 2023). In maize, like in other organisms, the advancements in technology have led to the rapid generation of genomic data. These data encompass various molecular layers at different scales, such as TF binding profiles (Galli et al., 2018; Ricci et al., 2019; Tu et al., 2020), accessible chromatin regions (ACRs) (Rodgers-Melnick et al., 2016; Ricci et al., 2019; Marand et al., 2021), expression and co-expression atlases (Sekhon et al., 2011; Stelpflug et al., 2016; Hoopes et al., 2019; Zhou et al., 2020), and transcriptomic, proteomic, and metabolic at population-level (Li et al., 2013; Wen et al., 2014, 2016; Kremling et al., 2018; Zhou et al., 2019; Mazaheri et al., 2019; Shrestha et al., 2022). Consequently, there has been a growing focus on integrating multi-omic datasets (Liu et al., 2016; Walley et al., 2016; Wen et al., 2016; Jin et al., 2017; de Abreu E Lima et al., 2018; Lee et al., 2019; Schaefer et al., 2018; Wen et al., 2018). However, most integration efforts primarily involve the verification of each layer with one another (i.e., conceptual integration) (Depuydt et al., 2023). There are a few exceptions where the layers are leveraged to enhance the integration (Schaefer et al., 2018; Yang et al., 2022) or to learn from their combined information (Han et al., 2023). Therefore, emphasizing the need for a comprehensive assessment of integration strategies and its effectiveness to prioritize gene-specific processes.

In this study, I analyzed genetic and gene expression variation in 304 maize inbred lines. I utilized data from over 300 publicly available ChIP- and DAP-seq experiments, along with 45 previously analyzed co-expression networks (Zhou et al., 2020). Combining these datasets, I built four molecular networks and employed three integration methods. I sought to annotate transcription factors (TFs) based on their predicted target genes. I combined published knockouts and created random networks as strategies to evaluate the corresponding functional predictions. This allowed me to identify the integration strategy that made functional predictions more similar to those observed in knockout assays, while minimizing the chance of random predictions. In essence, it allowed me to recover predictions rarely predicted by a random network. I provided evidence that these predictions recovered TF-process associations previously linked to specific biological processes. The compiled predictions enabled the creation of a TF-process association list, which, when combined with TF-target networks, facilitates the identification of regulators for processes like abscisic acid (ABA), lipid, phenylpropanoid, and leaf-related processes. Finally, I demonstrated that employing the generated embedding post-integration of all four networks, which recovers pattern of connectivity within the full combined network, allows distinguishing homologous (aka, paralogs) with potential redundancy in maize. Collectively, these findings offer a remarkable amalgamation of TF-process associations and lay the foundation for prospective network-based functional prediction in maize. Moreover, this invaluable tool facilitates the linkage of previously identified genetic markers with clusters of functionally associated genes, utilizing their connectivity patterns within the presented networks.

4.3 RESULTS

4.3.1 Construction of a maize regulatory network based on multiple layers

To build a multi-layer TF-function association network, I collected previously published co-expression networks, single-nucleotide polymorphisms (SNPs), and reanalyzed publicly available expression, DAP-seq, and ChIP-seq datasets in maize. In total, I included several co-expression networks, genetic variation data for 304 maize inbred lines, and 289 DNA-binding assays (DAP-seq and ChIP-seq) associated with 144 TFs (Figure 4.1a). In total, I identified ~3.4M, ~155.1K, ~1.18M, and 112.46K TF-gene associations derived from the co-expression networks (CENs), a *trans*-eQTL association network (GAN), a gene-regulatory network (GRN), and *cis*-eQTLs overlapped with GRN interactions (eGRN), respectively. The GRN was built based on DAP/ChIP-seq assays (Figure 4.1b). Construction details of the corresponding networks are described below.

Coexpression network (CEN). To build the CEN layer, I started by collecting 45 CENs previously published (Zhou et al., 2020), and added an additional network constructed with a subset of expression datasets associated with 304 inbred lines [Wisconsin Diversity (WiDiv) panel (Mazaheri et al., 2019)]. The 304 lines were selected based on availability of high-density whole genome sequencing derived SNPs (Bukowski et al., 2018), following consistent methods with previously reported CENs (Zhou et al., 2020) (see *Methods*). Thus, in total, I utilized 46 different co-expression networks to define the TF-target CEN layer. Each network was reduced to only maize genes in synteny with *Sorghum bicolor* (Schnable, 2019) to avoid potential bias towards non-functional genes when conducting gene enrichment analyses. The syntenic gene filter was also applied to all other network types (*i.e.*, GRN, eGRN, and GAN). On average, the collected CENs showed ~1,055 TFs co-expression network (Figure S4.1a) with ~74 predicted target genes (a.k.a, targets) per TF (Figure S4.1b). Combining all 46 CENs, I identified ~3.4M TF-target

associations involving 1,852 TFs and 23,788 targets (on average, ~1,350 targets per TF; targets can include other TF genes). To note, some of these TFs had several times more targets than the average TF (Figure S4.1b). For example, the ABI3VP1-7 (ABI7) and the C2C2-CO-like-transcription factor 8 (COL8), showed >400 targets in five and four CENs, respectively (Figure S3.1b); or COL13 and the bHLH-transcription factor 127 (bHLH127), which in total showed > 6,000 targets each (Figure S4.1c).

Gene association network based on *trans-e*QTLs (*GAN)*. This layer was built based on *trans*-eQTLs identified in eight distinct tissues encompassing several developmental stages. Overall, after quality control and data preprocessing (see *Methods*), I tested between 15.5M - 16.7M SNPs against the expression of 15.3K - 26.4K genes across the eight tissue types. Thus, after discarding non-significant eQTLs (See *Methods*) and non-syntenic genes (Schnable, 2019), I obtained a total of ~22.9M eQTL-gene associations including ~10M and ~26.4k different SNPs and target genes, respectively. These associations were classified as *cis*-eQTL overlapped with its target genes (*cis*-eQTLt), *trans/cis*-eQTL, *cis*-eQTL, *trans*-eQTL, and unassigned eQTL according to the distance between each eQTL and its corresponding target gene and eQTL-gene co-location (Figure S4.2a). Under this classification schema, I identified 10.2M, 6.7M, 1.20M, 1.18M, and 3.5M unassigned eQTL, *trans*-eQTL, *trans/cis*-eQTL, *cis*-eQTLt, and *cis*-eQTL, respectively (Figure S4.2a). Within them, *trans*-eQTLs (eQTLs overlapped with annotated genes and located >50 kbs far away from their corresponding target genes) were used to define the GAN.

After removing redundant links (*e.i.*, multiple eQTL supporting the same gene-to-gene connection), the resultant GAN harbored ~155k associations, including 23.9K and 18.9K source and target genes, respectively. Here, “source gene” was defined as a gene overlapped with the corresponding eQTL. To better understand the nature of the genes captured on the predicted GAN,

I classified source and target genes into five functional categories including transcription factors (TFs), co-regulatory factors (CoReg), mediators, kinases, enzymes, and others (Yilmaz et al., 2009; Zheng et al., 2016; Mathur et al., 2011). Interestingly, "Kinase" and "Enzyme" were the top two classes with the highest number of target genes, even larger than the "TF" class (Figure S4.2b). Similarly, "Enzyme" class was the most frequent target class followed by "Mediator" and co-regulators ("CoReg") (Figure S4.2c). I counted the interaction frequency between the corresponding classes, and after "Other", "Enzyme" was the functional class with more interactions (13.7K) (Figure S4.2d), highlighting "Enzymes" as one of the functional classes more interconnected within the predicted GAN. Finally, I also note that the GAN recovered gene-gene interactions that capture both typical TF-target interactions, but also physical protein-protein interactions. An example is provided by the the HSF-transcription factor 20 (HSF20) which showed 354 targets, including 27 genes previously reported as heat-response related genes (Zhou et al., 2021), as well as five known physical interactors (Zhu et al., 2016). This highlights a typically unexplored set of regulatory connections among genes at several hierarchical levels.

Gene-regulatory network (GRN), and *cis-eQTLs* overlapped with *GRN interactions (eGRN)*. To construct the GRN, I collected and reanalyzed 283 PDI experiments associated with 142 different TFs. All the reanalyzed assays corresponded to TF-centered approaches, including 215 ChIP-seqs in protoplast (pChIP-seq), 30 classic ChIP-seq, and 38 DAP-seq. A single data analysis pipeline was used to process all PDI assays to reduce pipeline-specific bias (See *Methods*). On average, I obtained ~52k peaks per TF which, in total, represented ~7.6M PDIs. Most of the predicted peaks were contributed by pChIP-seq (Tu et al., 2020), which represented 75% of the data analyzed (on average, ~55k peaks by TF) (Figure S4.3a). To identify high-confidence peaks, I applied two filtering criteria. First, I gathered accessible chromatin regions (ACRs) from the

recently published single-nuclei ATAC-seq (snATAC-seq) atlas (Marand et al., 2021), retaining only TF's peaks that overlapped with ACRs. Therefore, I compared all the DAP-seq and ChIP-seq datasets to a shared regulatory maize space. Second, I removed peaks with low counts per million (CPM) (as defined by a Z-score ≤ -0.5) for each PDI assay. Overall, I filtered out ~3.8M and ~1.1M peaks using the ACR and CPM criteria, respectively (Figure S4.3b, c). As expected, DAP-seq assays, in both filters, have the largest percentage of discarded peaks (Figure S4.3b, c). Interestingly, comparing low-coverage and ACR co-location peaks and their distance to the closest annotated transcription start site (TSS), I find that peaks with the highest Z-values mapped largely to ACRs near TSSs (~10 kbs around) (Figure S4.3d). These last patterns were observed in all data types (DAP-, ChIP-, and pChIP-seq), thus, supporting the biological relevance of the high-confidence peaks retained. After filtering, I ended with a set of ~3.6M of peaks that were used for further analyses.

To define target genes, I integrated the peak-TSS distance and their overlap with *cis*-eQTLs (declared when a peak summit and a *cis*-eQTL were at ≤ 20 bp away). Combining these metrics, I classify the peaks into three types of peaks close to TSSs (≤ 3 kb) and two types of peaks far away from TSSs (> 3 kbs and ≤ 50 kb). Specifically, peaks in close proximity (≤ 3 kb) were defined as follows: peaks without *cis*-eQTL support (Figure S4.3e, light purple peaks), with *cis*-eQTL support and similar target prediction (Figure S4.3e, light green peaks), and with *cis*-eQTL support and different target prediction (Figure S4.3e, yellow peaks). These categories represented the 54.9%, 1.9%, and 0.1% of the total analyzed peaks, respectively. Similarly, peaks located far away were classified as peaks with (3.3%) and without (39.5%) *cis*-eQTL support (Figure S4.3e; light blue and gray peaks, respectively). Overall, I did not observe differences in peak categories among PDI data types (Figure S4.3e, bottom panel). Thus, after discarding peaks located far away and without

cis-eQTL support, I build a gene regulatory network (GRN) and *cis*-eQTL supporting GRN (eGRN) combined all peaks by TF irrespectively of the PDI source. In total, I captured ~1.12M (GRN) and ~1123.46K (eGRN) TF-target interactions, including 138 TFs and ~23.9K and 13.9K target genes, respectively (Figure 4.1a, b).

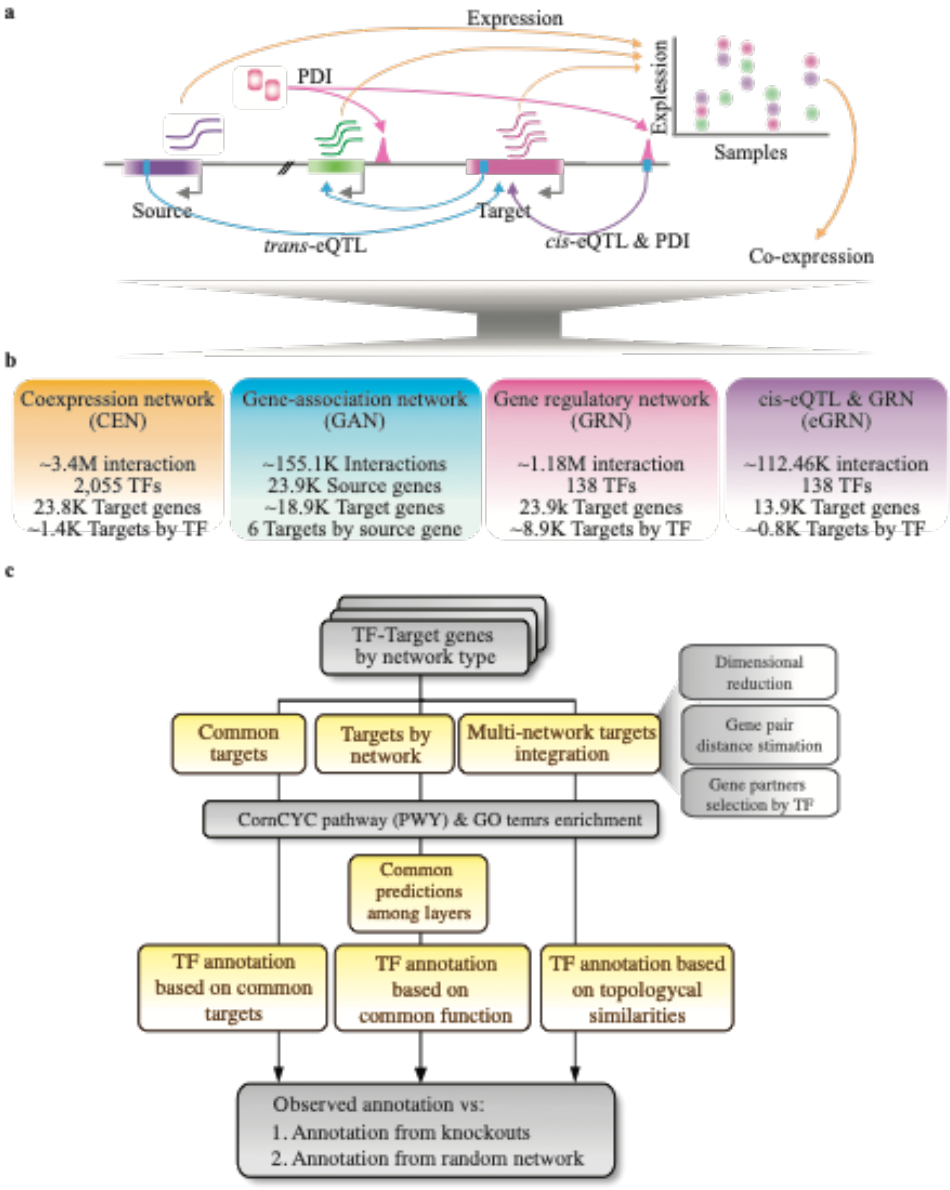

Figure 4.1 Construction of maize gene regulatory network based on multiple data types
a. Model indicating the different types of TF-gene associations used to define the network types analyzed in this work. **b**. Summary of the metrics of the four types of network layers. **c**. Schematic

Figure 4.1 (cont'd)
representation of the pipeline implemented to annotate and evaluate the corresponding functional predictions.

4.3.2 TF Functional annotation

A major difference between the networks constructed is the number of TFs and their corresponding target/associated genes (here, indistinguishably called target genes), which hinders comparisons between layers. For instance, all four networks have 111 TFs with at least one target gene (Figure S4.4a, b), however, this number is reduced to only 17 TFs when comparing TFs with at least ten different target genes by network (Figure S4.4a, c). This reduction is largely caused by the low number of predicted targets on the GAN layer (on average, ~6.5 targets by gene). In consequence, I implemented three different strategies (*common interactions, common integrations,* and *network-based*) to functionally annotate the TFs present in the corresponding networks. In all three approaches, the annotation was performed based on enrichment of metabolic pathways (PWYs) (Andorf et al., 2016) and GO terms (Wimalanathan et al., 2018) (Figure 4.1c) (See *Methods*). Briefly, the most conservative approach, *common interactions*, assumes that only common TF-target interactions between layers (i.e., GAN, GRN, eGRN, and CEN networks) capture true targets of the corresponding TF, and by extension its function.

Common function, assumes that a TF function is most accurately captured by those functions commonly enriched across different network types. Thus, it prioritizes functions commonly enriched for the corresponding TF across layers. Finally, *network-based* combines all layers to then extract topological properties for each gene. It assumes that each interaction type bore equally valid information about the function of the corresponding TFs. Specifically, it combines all four layers (GAN, CEN, eGRN, and GRN) creating a denser network (combined network) to then extract physical parameters - embeddings - from each gene in the combined network (See

methods). The transformation of the networks into a matrix of genes and embeddings allows the grouping of genes based on the similarity of their embeddings. Here, I used the mutual rank of the mutual information as the metric to identify highly similar genes in the embedding matrix, to then test for enrichment with PWYs and GO terms between the corresponding genes. Ultimately, this strategy allowed me to make functional annotation of TFs independently of the number of target genes predicted at the individual layer types (*i.e*., GAN, CEN, eGRN, and GRN).

Common interactions. To identify common interactions, I compared all layers with each other (Figure S4.4a) and obtained ~4.6M TF-target interactions. As expected, GRN and eGRN were the layers with the largest number of overlapping interactions (~112.5K), followed by GRN and CEN (~102.7K) (Figure S4.4d). After identifying common interactions, I keep 206.2k out of the 4.6M interactions, including 934 and ~20.6K different TFs and target genes (Figure S4.4d). Using target genes as a proxy to annotate the TFs function, I test the enrichment of common target genes with PWYs and GO terms by TF (See *Methods*). Also, given the similarities in their molecular functions, I included co-regulators in the analysis and treated them without distinction from TF. In total, I found 2,812 TF-PWY and 8,550 TF-GO significant associations [False Discovery Rate (FDR) ≤ 0.1, Fisher's Exact Test] (Figure 4.2a), which on average represented ~8 and ~80 PWYs and GO terms by TFs (Figure 4.2b). Combining PWY and GO term results, I annotated 347 TFs, out of which 235 TFs showed enrichment only in the PWYs analysis. The remaining 112 TFs showed enrichment with both PWYs and GO terms (Figure 4.2c).

Common function. To identify common functions, I initially tested the enrichment of target genes with PWYs and GO terms for each TF in each layer. I retained TFs that had at least one PWY/GO term enriched in the last two different layers. This allowed me to explore common predictions between layers for the corresponding TFs (Figure S4.5a). I observed a variable number

of TFs enriched with PWYs (ranging from 120 to 2,019 TFs) and GO terms (ranging from 72 to 1,777 TFs) across the different layers. Between the layers, eGRN had the fewest annotations, while CEN had the largest number of annotations (Figure S4.5b). Thus, after selecting TFs with at least one PWY and/or GO enrichment, I ended with 966 TFs and 245 TFs, respectively. When considering PWY annotations, the layer pair of CEN & GAN had the highest number of TFs annotated (888 TFs), while the layer pair of GRN & GAN had the lowest (59 TFs). Similarly, for GO term annotations, CEN & GAN had the highest number (130 TFs), while GRN & GAN had the lowest number (59 TFs) (Figure S4.5c, d). Regarding common predictions, I identified overlapping PWYs by evaluating gene overlap among all PWYs enriched per TF between layers (P-value ≤ 0.05, Fisher's Exact) (Figure S4.5a). A similar approach was used to identify common functions at the GO term level. However, due to the hierarchical and redundant nature of the GO terms, I employed semantic similarity rather than gene overlap to determine common GO terms per TF between layers (FDR ≤ 0.1) (See *Methods*). These two annotation analyses together yielded 7,081 TF-function annotations (727 TF-PWY and 6,354 TF-GO) (Figure 4.2a). On average, this corresponds to 3.5 different PWYs and 57.7 different GO term associations per TF (Figure 4.2b). In terms of TFs, these associations encompass annotations for 204 TFs through PWY enrichment and 110 TFs through GO term enrichment (Figure 4.2c).

Network-based. I combined all four layers, *i.e*., CEN, GEN, GRN, and eGRN, to then scale the interaction frequencies from 0.5 to 1, being 0.5 and 1 the weight for interactions observed in one and all four layers, respectively. With the scale version of the combined networks, I proceeded to identify low-dimensional representations (embeddings) for each gene/node in the combined network (Figure S4.6a) (See *Methods*). The combined network included 4.6M interactions associated with 36.4K genes. Unlike the previous two strategies, this method generated an equal

number of descriptors (embeddings) for each gene in the network, thus allowing the identification of genes with similar properties, including TFs present in the CEN and/or GAN layers without data on the GRN/eGRN layers. The distance of the embedding vector between genes was defined as the decay function of the mutual rank of the mutual information of the embedding (See *Methods*) (Figure S4.6a). On average, I found 235 highly similar genes per TF [Distance (D), ≤ 0.05, See *Methods*] (Figure S4.6b). As in previous approaches, I annotated the corresponding TFs by assaying the enrichment with PWYs and GO terms of their highly similar genes (Figure S4.6a). In total, I found 23,796 and 7,722 TF-PWY and TF-GO significant associations (FDR ≤ 0.1, Fisher's Exact) (Figure 4.2a), which on average captured ~7 and ~8 PWYs and GO terms per TF, respectively (Figure 4.2b). Combining both assays, I annotated 2,910 different TFs, out of which 1,030 TFs showed enrichment with both PWYs and GO terms (Figure S4.6c). To note, these 1,030 TFs belong to 82 different TFs (including co-regulator) families capturing and representing - on average - 34% of the total proteins annotated in the corresponding families (Figure S4.6d). This highlights the potential of the method to annotate TFs with unobserved layers.

Comparing all three methods, *network-based* allowed me to identify the largest and lowest total number of TF-PWYs and TF-GOs associations, respectively (Figure 4.2a). Also, it has the lowest average of PWYs and GO terms per TF (Figure 4.2b). Unexpectedly, *network-based* and *common target* methods predicted a similar number of PWYs per TFs, which contrasts with the significantly lower number of GOs between *network-based* and the other two methods (Figure 4.2b). Importantly, the number of TF annotated by the *network-based* is >2.5 times larger than the other two methods (Figure 4.2c, left panel). Finally, combining all results, I functionally annotated 2,917 TFs. However, 94 (Figure 4.2c, violet plus green labels) and 32 (light blue plus green labels)

out of the 2,243 TFs showed at least a PWY and a GO term enrichment in all three methods, respectively.

4.3.3 Evaluation of functional prediction with knockouts

TF perturbation experiments enable the understanding of the TF regulatory landscape by unraveling the direct and indirect effects of expression changes induced by the expression variation of the corresponding TFs. Here, I used 21 previously published knockouts associated with 13 different TFs (Zhou et al., 2020; Ellison et al., 2023) to assay the accuracy of each of the three methods by two independent strategies. Specifically, I questioned the overlap between predicted and observed PWYs/GO terms within DEGs for the corresponding knockouts. In parallel, I also tested the gene set enrichment analysis (GSEA) of the predicted PWYs/GO terms within the corresponding TF knockouts (Subramanian et al., 2005) (See *Methods*). Unexpectedly, predicted PWYs - without distinction of the methods - showed poor overlapping with PWYs observed at the knockout's assays, as well as low recovering of PWY significantly enriched within DEGs as estimated by the GSEA (Figure S4.7). Conversely, comparisons between predicted and observed GO analysis showed similarities [measured by the GO semantic similarity (GSS)] different than the expected by chance (P-value ≤ 0.05) (Figure S4.8). Overall, the GO terms from knockouts and the GO terms predicted by *network-based* and *common function* are significantly more similar than the *common targets* predictions (higher GSS values, P-value ≤ 0.05, Wilcoxon test) (Figure 4.2d). Remarkably, when a prediction is available, the *network-based* method recovers the GO terms with the highest GSS values among all the methods (Figure 4.2d, TB1 and FEA4 results). Additionally, I observed that seven knockouts lacked predictions from *network-based* methods, while eight others had predictions only with *network-based* methods (Figure 4.2d, e). This variability in predictions can be partly attributed to the low number of target genes (when the

prediction is absent; Figure S4.9a, TFs with Z-score ≤ 0) and the absence of data in at least one of the four layers (i.e., GRN, eGRN, CEN, and GAN) (when the *network-based* method is the only one making the prediction, Figure S4.9b). Consistently with the GSS analysis, the GSEA results indicate that GO terms recovered with the *network-based* and the *common function* are more consistently identified across the different knockouts (Figure 4.2e). Thus, all together, my results suggest that *network-based* predictions are resilient to the presence-absence variation of layer's data, although susceptible to the number of targets by TF. By extension, they also indicate that *common targets* and *function* predictions are more sensitive to the absence of data in at least one of the layers.

Combining all GSEA results, I find that, on average, only 25% of total GO predictions show significant GSEA scores (P-value ≤ 0.05), denoting a low recovering rate of GO terms (Figure 4.2e). I argue that the characteristic indirect effects of the knockout can explain these low recovery rates, combined with the tissue/condition/genotype differences between the knockout assays and the data used in the corresponding predictions. I used the TFs expression-specificity as a proxy to understand the relationship between the low fraction of GO recovered by GSEA and the tissue/condition/genotype variation among the corresponding TFs. Including all the TFs for which I obtained at least PWY/GO term prediction and using the Tau index as a metric (Kryuchkova-Mostacci and Robinson-Rechavi, 2017), I find a bi-modal expression distribution with ~55% of the TFs trending into a sample-specific expression fashion (Figure S4.9c, Tau ≥ 0.65). Interestingly, only four out of the 13 TFs tested in the knockout analysis are expressed in a sample-specific fashion (P-value ≤ 0.05) (Figure S4.9d, labeled in green). The top four included RA1 (Tau 0.99) and TB1 (Tau 0.96), which also are the top two TFs with the largest fraction of GO term supported by the GSEA (Figure 4.2e). Hence, the results support the notion that a portion of the

low fraction of GO terms recovered can be attributed to the differences in conditions used on the knockout and prediction analyses. Thus, I predict that perturbation analyses conducted under conditions that mitigate tissue/condition effects may lead to a higher overlap.

4.3.4 Evaluation of functional prediction by comparing with random networks

Despite recovering GO terms similar to - and enrichment with - GO terms from knockouts (Figure 4.2d,e), which method generates fewer false positives still needs to be determined. In consequence, I assayed the identification of GO terms from ~3,000 random networks to establish which method recovered the lower fraction of false positives (See *Methods*). I counted the number and significance of the GO terms enriched in random networks as a proxy for the precision, and the similarity of observed GOs (true TF-target interactions) with the GOs from random networks as proxy for the accuracy of the corresponding methods. Also, to compare predictions across methods for the same TF, I reduced our analysis to only the 32 TFs with GO predictions in all three methods (Figure 4.2c, green and blue intersection). I posited that methods with fewer GO terms, less significant P-values (FDR), and GO terms from random networks less similar to observed GO terms are indicative of better predictions. Remarkably, *network-based* identified significantly enriched GO terms in only ~12% of random networks tested, which contrasts with the ~28% and ~72% obtained with the *common function* and the *common target* methods, respectively (Figure 4.2f). Concordantly, *Network-based* predicted significantly fewer GO terms (Figure 4.2g), with less significant P-values (Figure 4.2h) and GO terms less similar (lowest GSS values) to the predicted from true interactions per TF than those observed with *common function* and *common target* methods (Figure 4.2i), highlighting *Network-based* as the method with the highest precision and accuracy. To be noted, *the common function* predicted fewer GO terms per TFs than the *common target*; although its predictions have P-values more significant and with GSS

values equally similar to those observed in the *common target* (Figure 4.2g-i), patterns that were consistently observed also at individual TF level (Figure S4.10), positionings the *common functional* and the common *target* as equally noisy methods.

Overall, the *network-based* method detected a lower number of GO terms per TF (Figure 4.2b) and had GO terms enriched in significantly fewer random networks (Figure 4.2f). This suggests limitations in the method's ability to identify GO term associations, as it inherently identifies fewer GO terms per TF. To examine this possibility, I investigated whether the number of GO terms observed with true interactions could be attributed to chance. Remarkably, 30 out of 32 tested TFs exhibited a significantly higher total number of GO terms compared to those expected by chance (P-value <= 0.05) (Figure S4.11). Thus, despite the *network-based* approach yielding fewer GO terms per TF, these identified GO terms contain valuable biological information that is unlikely to occur randomly. In conclusion, within the given data context used in this work, I affirm the *network-based* method as the superior approach. Consequently, I exclusively relied on *network-based* predictions for subsequent analyses.

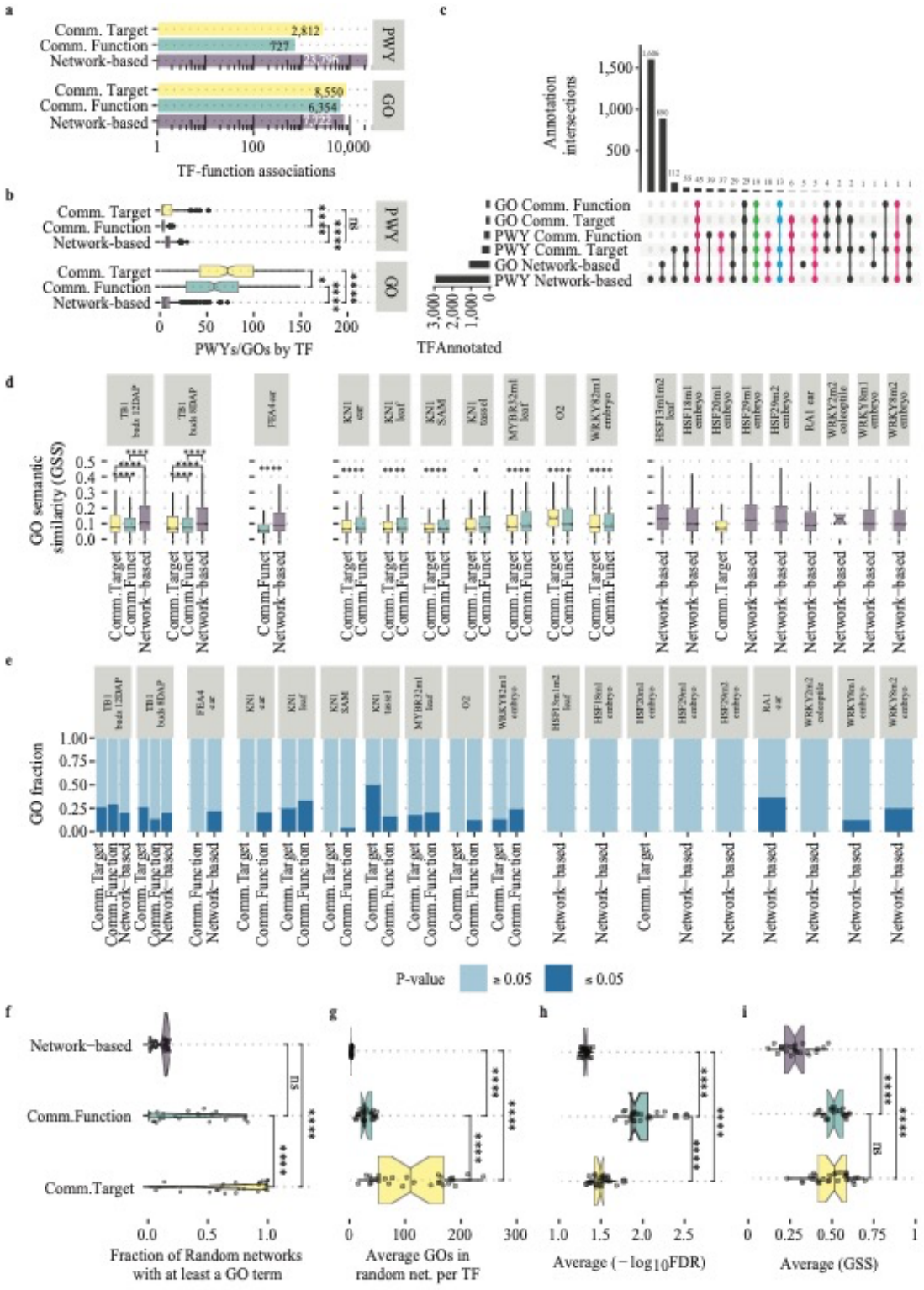

Figure 4.2 Annotation and evaluation of TF functional annotation by contrasting predictions with knockout assays and random networks

Figure 4.2 (cont'd)
Total PWYs and GO terms predicted per integration method after combining all TFs predictions (**a**) and per TF (**b**). **c**. Upset plot comparing total TFs annotated by each method and annotation system. Colors indicate the groups of TFs functionally annotated by all three methods by enrichment with PWYs (fuchsia), GO terms (blue), and both PWYs and GO terms (green). Black groups indicate TFs annotated by at least one of the methods and annotation systems. **d**. Boxplot of the GO semantic similarity for the top 10 most similar GO terms observed in knockout assays for each of the predicted GO terms per TF and methods. **e**. Stacked barplot indicating the fraction of the GO terms predicted and significantly enriched - by GSEA analysis - in the knockouts. **f**. Violin plot showing the fraction of random networks with at least one significant (FDR ≤ 0.1, Fisher exact test) GO term by TF. **g**, **h**, and **i.** Boxplot showing the average number of GO terms (**g**), -log10FDR (h), and GSS (**i**) observed in 3000 random networks by method. The GSS values were calculated by comparing each random network with the observed GO terms from the true TF-target interactions. Asterisks indicate P-value significance (*: $p \leq 0.05$, **: 633 $p \leq 0.01$, ***: $p \leq 0.001$, ****: $p \leq 0.0001$, two-sided t-test). "TFm" denotes multiple mutant lines for the same TF.

4.3.5 Prioritization of regulators by biological process

The *network-based* method detected approximately 7.7K TF-GO associations, encompassing 1,036 TFs and 2,219 GO terms (Figure 4.2a, c & Figure S4.6b, c). For ease of TF comparison, I retained associations involving GO terms within the biological process (BP) category and having fewer than eight hundred associated genes (when a more specific GO term association was present). Additionally, to minimize GO term redundancy, I mapped GO terms with a small number of annotated genes (less than 50 genes) to their nearest GO term parent. After applying the filters, I continued with 4,337 TF-GO associations, including 902 TFs and 559 GO terms. The distribution TF-GO associations obtained hold a scale-free distribution (Figure 4.3a, b). Typically, highly interconnected GO terms and TFs suggest a greater number of annotated and targeted genes, respectively. Nevertheless, I did not discover any evidence linking the gene count per GO term or the target gene count per TF to their respective degrees (Figure 4.3c, d). Therefore, these analyses highlight GO terms whose regulation may depend on multiple TFs, and TFs that may contribute to regulating several biological processes.

From the perspective of gene regulation, when multiple TFs are associated with multiple GO terms, it suggests that the regulatory impact of a TF on a GO term is influenced by the presence of other TFs. To assess the contribution of individual TFs to their respective GO terms, I calculated a scaled enrichment score (Z-score of the enrichment) for each TF and GO term (See *Methods*). Utilizing the scores as an indicator to assess the significance of individual TF-GO associations relative to all other TF and GO term associated, I observed that only 3.4% (151/4,337) of the TF-GO associations exhibit high enrichment scores (Z-score ≥ 1) (Figure 4.3e, top right corner), indicating that only a reduced fraction of the TFs and GOs analyzed have strong enriched scores for the corresponding association. Consequently, this implies that most of the analyzed TFs/GO terms have multiple associations of comparable significance. I combined both Z-scores (per TF and GO term) into a reciprocated Z-score (rZ, See *Methods*) to rank TFs by GO term using a single metric. I evaluated the ranking after grouping GO terms into specific biological processes (Figure 4.3f, and Figure S4.12). I highlight here, 46, 62, 47, and 50 abscisic acid (ABA)-, lipid-, phenylpropanoid-, and leaf-related TF-GO associations that were targeted by 44, 55, 47, and 50 different TFs, respectively (Figure 4.3f). Using the rZ score as filter (rZ ≥ 0.5), I narrowed down the list to 25, 27, 15, and 19 TF candidates to control the corresponding processes (Figure 4.3f, dots with name label included). Some examples included the top two TFs ABA-related, NAC56 and WRINKLED2 (WRI2). Additionally, WRI2 was also on the top three of the TFs related to lipid-related metabolism (Figure 4.3f, second panel). Finally, five (WOX9a, OFP39, Zm00001d024353, EREB149, and LBD24) out of the initial 47 TFs phenylpropanoid-related were previously identified as maize regulators of phenolic-related genes by yeast-one hybrid assays (Y1H) (Yang et al., 2017). Altogether, this highlights the biological relevance of the associations this analysis predicted.

Apart from controlling specific enzymatic or signaling-related genes, TFs can also regulate biological processes by targeting other TFs. This leads to the formation of regulatory circuits with multiple hierarchical levels and network motifs. To identify TFs that play a higher-level role in controlling a biological process, I calculated the ratio of TFs targeted by other TFs within specific GO terms to the total number of TFs targeted by the corresponding TF. Given that I looked for TF associated with a common function, this ratio represents the weighted proportion of feed-forward loops associated at the level of biological process compared to the overall TF targets of each TF. This measure is referred to as the upstream regulator score (URS). I calculated the URS for the twenty different processes, including eight hormones-, seven metabolic-, and five developmental-related processes (Figure 4.3g). Cytokinin- and shoot-related GO terms were the top two processes with the highest score, with C2C2-CO-like 13 (COL13) and RAMOSA1 (RA1) as their top regulators, respectively (Figure 4.3g). To note, COL13 was previously associated with carbon metabolism (Tu et al., 2020), and is also differential expressed on the *indeterminate1* (*1d1*) loss of function mutant (Minow et al., 2018). ID1 is a maize regulator of autonomous floral induction (Colasanti et al., 1998). Thus, our results suggest a role of COL13 in the connection among cytokinin, carbon metabolism, and flowering; mechanistic association previously reported in other plant systems (Bartrina et al., 2011; Wahl et al., 2013). Similarly, RA1 was predicted as the top upstream regulator of shot-related processes, and RA1 itself was linked with shoot system development (Figure 4.3g, process number 17), both of them functions previously associated with RA1 (Eveland et al., 2014). To further understand the regulatory landscape of the four processes described previously (Figure 4.3f), I selected the top two predictions - URS score - for each process and traced their TF targets back into the original networks (*i.e.,* GRN, eGRN, CEN, and GAN) (Figure 4.3h). Specifically, I looked for regulators directly upstream of any of the top TFs as

predicted by the reciprocal Z-score (rZ) analysis (Figure 4.3f). Without exception, I found at least an upstream regulator directly targeting (GRN network) at least one of the top three TFs from the rZ analysis, i.e., a top regulator of the corresponding biological process (Figure 4.3h). To highlight an example within the ABA-related process network, EREB17 targeted NAC56 and WRI2 (tops TFs in rZ analysis), and bHLH43 (URS top one) targeted WRI2 and EREB17 (Figure 4.3f, h, first panel). This configuration forms a feed-forward loop with bHLH43 on top (i.e., bHLH43 targets EREB17 and WRI2, and EREB17 targets WRI2). Within the lipid-related network, ARF14 targeted WRI2 and PRH65 (top two and three by rZ score) (Figure 4.3f, h, second panel), as so did HB33 targeting LBD23 (top in rZ) in the phenylpropanoid-related network (Figure 4.3f, h, third panel). Finally, WRKY25 (top USR) targeted the MYBR4 and EREB126 (top two TFs in rZ), as well as BZR2 (top two in URS analysis) on the leaf-related network. Thus, it highlights specific regulatory hypotheses to further experiment validations.

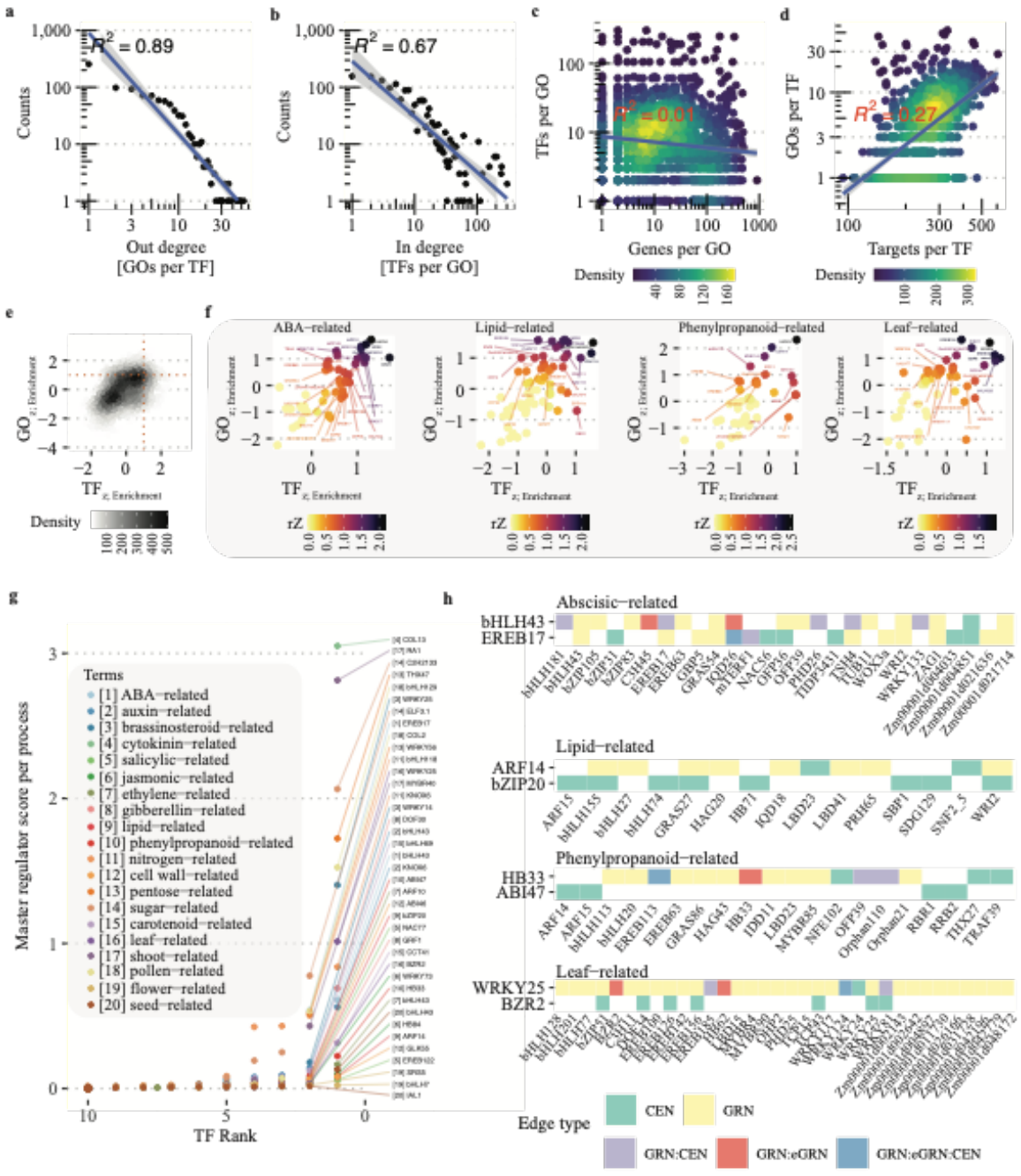

Figure 4.3 Prioritization of regulators by biological process using network-based prediction
a Out degree and **b** in degree distributions of the TF-GO term predictions obtained from the *network-based* integration analysis. **c** and **d**, scatter plots indicating the frequency - as density - of the number of TFs by GO terms (in degree) and GOs per TF (out degree) as a function of the number gene annotated per GO term (**c**) and target genes per TF (**d**), respectively. **e**, Scatter plot indicating the frequency - as density - of the TF-GO term enrichment scores scaled, which allows to rank GOs highly enriched with specific TF (GO_z) and TFs highly enriched with specific GO term (TF_z). The enriched was calculated only with TF-GO term associations already predicted in

Figure 4.3 (cont'd)
previous analysis. Dotted line orange indicates TF-GO term associations with enrichment score a standard deviation over the observed average for the corresponding TF and GO term (Z-score of enrichment ≥ 1 for both GO term and TF). **f**. Scatter plot with reciprocal Z score (rZ) of four different biological process mapped into the GO and TF scaled score coordinates as presented in **e**; GO terms were grouped as follow: ABA-related (GO:0009737, GO:0009738, GO:0009688, and GO:0009788), Lipid-related (GO:0031408, GO:0006099, GO:0006635, GO:0019915, GO:0006629, GO:0019375, GO:0044255, GO:0016042, GO:0051790, GO:0008610, GO:0009062, and GO:0045332), phenylpropanoid-related (GO:0009963, GO:0009698, GO:2000762, and GO:0009699), and leaf-related (GO:0009965, GO:0048366, GO:0010305, GO:0010150). TF name/gene id labels are included for TFs with rZ ≥ 0.5. **g.** Scatter plot with TF ranked their upstream regulator score (URS) by biological process. TF name labels are included for TFs with rank ≤ 2. Square brackets indicate an arbitrary biological process index which matches with the number in square brackets of the corresponding TF names. All URS scores are calculated based on the original GRN, eGRN, CEN and GAN networks. **h**. Heatmap with top two TFs (*y* axes) from the URS analysis (**g**) for the four biological processes presented in **f.** *X* axes indicate the corresponding TF targets**.** Colors indicating the network(s) source of the corresponding interactions.

4.3.6 Topological properties predict TF homeologs redundancy

Although substantial efforts have been made to comprehend and anticipate the functional redundancy between maize paralogs in subgenomes (Schnable et al., 2011; Li et al., 2016; Kono et al., 2018; Han et al., 2023), the problem remains far from complete comprehension. I anticipate that if a pair of paralogs exhibit functional redundancy, these differences may manifest in their topological properties, i.e., functional redundant paralogs would display similar properties indicating a comparable network arrangement. To assess the similarity between paralogs, I generated a distance matrix from the embeddings using the mutual rank (MR) of mutual information as metric (Figure 4.4a). Next, I mapped TF paralogs (Schnable, 2019) and analyzed their MR and the similarity of their MR profiles with all the genes in the embeddings matrix as a proxy for understanding the similarity of their embeddings and the similarity of their resemblance with other TFs examined. I also differentiated between paralogs located on the same chromosome, serving as a proxy for pre-speciation tandem duplicates. In total, I tested 932 TF pairs, and regardless of the metric used, TF paralogs situated on the same chromosome demonstrated greater

similarity compared to TF paralogs on different chromosomes (Figure 4.4b, c). I combined both scaled metrics to identify highly similar TF pairs (Figure 4.4d). As expected, both metrics were correlated (Pearson correlation 0.68), yet they effectively served the purpose of identifying TF paralog pairs that were highly similar. I tallied the number of interactions after the embedding integrations (Figure S4.6a) for the top ten TF pairs that were most and less similar (Figure 4.4e), all top ten TF pairs more similar have several common interactions (Figure 4.4e, light brown TF pairs highlighted), contrary to those observed within the top less similar which have none (Figure 4.4e, gray brown TF pairs highlighted). Additionally, I find seven TF pairs mapped to the same chromosome out of the top ten TF pairs (Figure 4.4e, TF pairs with asterisk). Thus, for additional assessment of tandem duplicates distribution and the shared interactions between TF homologs in the context of the embedding similarities, I categorize all TF pairs into nine bins using both scaled similarity metrics (Figure 4.4d, dashed black lines). The bins are structured to include the most dissimilar TF pairs in the first bin (I) and the most similar pairs in the last bin (IX) (Figure 4.4f, internal box). Confirming the observation from the top ten TFs (Figure 4.4e), the bin IX contained 5-7 times more tandem duplicates than the other bins (Figure 4.4f). I quantify shared interactions using the Jaccard index. By considering bin I as a reference, I detect significant differences ($p < 0.05$, two-sided t-test) across five distinct bins (Figure 4.4g), primarily categorized based on the scaled correlation between TF pairs (Figure 4.4d, x axes). Furthermore, bin IX exhibits the utmost values, validating the predictive capacity of embedding similarities for functional redundancy in TF paralogs.

Considering variations and similarities in topological properties as indications of function divergence and conservation, respectively, I expect that the protein sequence or expression variation of the TFs in bin I will be greater. Focusing exclusively on TF pairs from bin I and IX

(representing a TF pair with contrasting embedding similarities), I calculated the Hamming distance of the amino acid sequences and co-expression as proxies to understand the observed differences in topological dissimilarities. Unexpectedly, I did not notice any differences in the Hamming distance between TF pairs highly similar or dissimilar at the topological level, as evidenced by TF pairs highly conserved (low Hamming distance) in both groups of TFs (Figure 4.4h). However, TF pairs in bin I (PCC = 0.44) showed slightly lower average co-expression values compared to those observed for TFs in bin IX (PCC = 0.5). Interestingly, when co-expression values are mapped in the context of TFs' Hamming distance, it allows me to differentiate TF paralog pairs that may be undergoing neofunctionalization/subfunctionalization due to variations in their protein sequences or its regulation. A striking example of the former is observed in MYBR1 and MYBR81, which have significantly different sequences (Hamming distance close to 1), distinct embedding profiles (bin I), and yet display high co-expression (PCC > 0.9) (Figure 4.4i). In contrast, HAG1 and HAG38, as well as GRAS14 and GRAS82, which also belong to bin I and have high similarity in peptide sequences (Hamming distance close to 0), show variation only in their co-expression (PCC < 0.3) suggesting variation at the regulation level (Figure 4.4i). Additionally, within the groups of TFs sharing similar embedding profiles (bin IX), I identified TFs exhibiting high conservation (Hamming distance close to 0) and similar expression, implying a significant degree of redundancy (e.g., MADS73 and TU1) (Figure 4.4j). Furthermore, I observed TFs with limited co-expression but high conservation (Hamming distance close to 0, e.g., C3H53 and C3H36), as well as TFs with fairly poor conserved peptide sequences (hamming distance close to 1, ABI5 and ABI4), indicating differences in its regulation (Figure 4.4j). Thus, altogether, the combination of embedding similarity, protein amino acid similarities, and co-expression enables the identification of TFs that are clearly variable or redundant, which is a key

observation for understanding function redundancy (in terms of GO enrichment) as described previously (Figure 4.3).

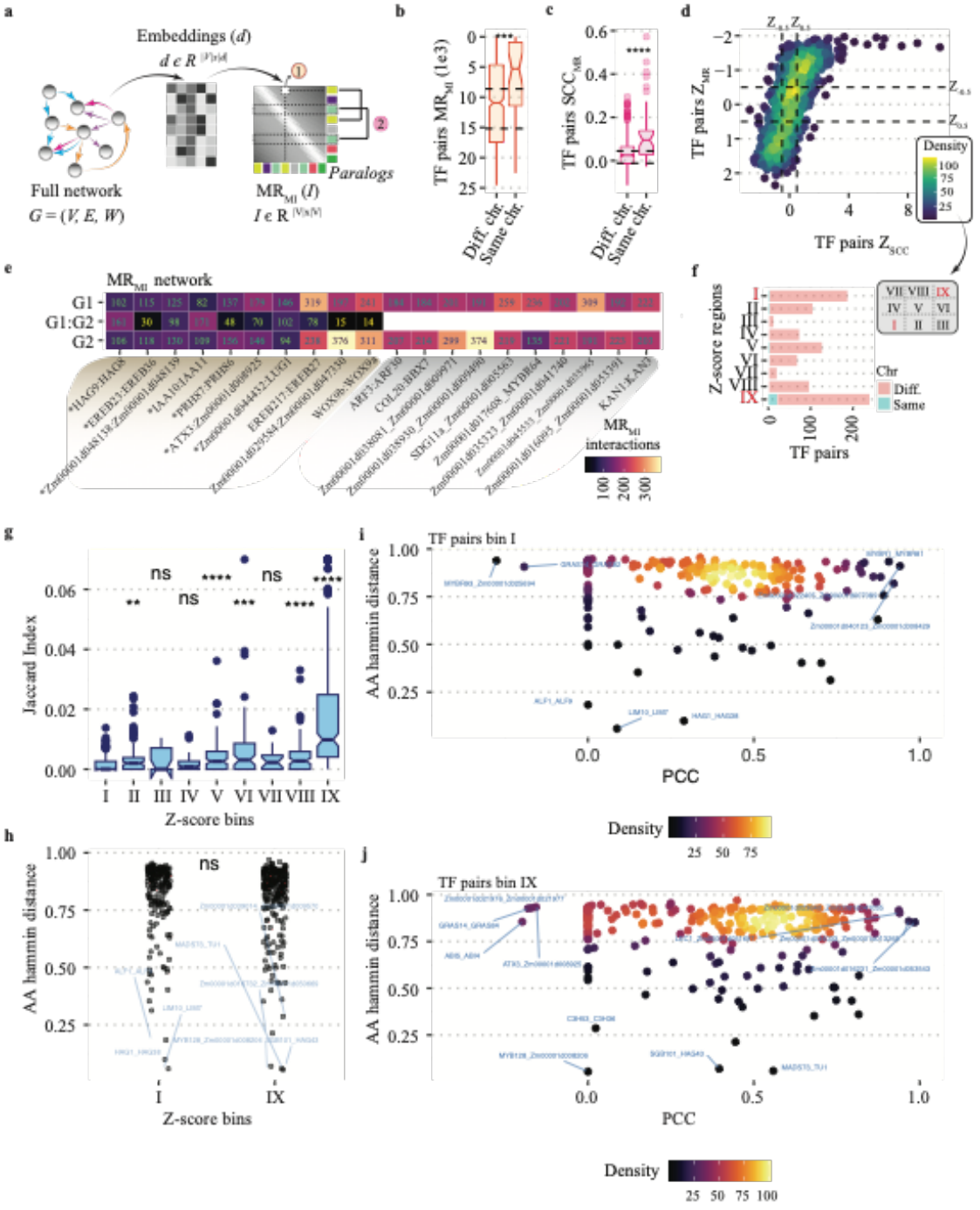

Figure 4.4 Network embedding as predictor of TF paralogs with functional variation

Figure 4.4 (cont'd)
a. Diagram illustrating the key stages of comparing TF paralogs through embedding similarities. **b** and **c**. Box plots displaying the MR_{MI} of TF pairs (**b**) and the Spearman correlations (SCC) of the observed MR_{MI} profiles (**c**) derived from the embedding. **d**. Combined scaled scores of the MRMI and SCC for TF pairs. Black dashed lines with Z-scores of -0.5 and 0.5 indicate values below and above the average observed standard deviation. **e**. Heatmap indicating the total number of associated genes for the top ten TFs, on the top right corner and the bottom left corner TF pair (**d**). G1 and G2 represent the number of unique genes associated with the first and second TFs in the corresponding pair. G1:G2 indicates the common associations between the corresponding pair. **F**. Bar plot indicating the total number of TF pairs by bin. Bins are indicated on the interval box, which is a map of the zones in the plot in (**d**). **g**. Box plot with Jaccard index (as an approximation of common associated genes) by TF pair by bin (as presented in **f**). **h**. Jitter plot displaying amino acid (AA) differences (Hamming distance) between TF pairs in bins I and IX. **i** and **j**, Jitter density of points representing AA Hamming distance and co-expression (measured as PCC) for TF pairs in bin I (**i**) and IX (**j**). Asterisks indicate P-value significance (*: $p \leq 0.05$, **: 633 $p \leq 0.01$, ***: $p \leq 0.001$, ****: $p \leq 0.0001$, two-sided t-test). "TFm" denotes multiple mutant lines for the same TF.

4.4 DISCUSSION

Cells utilize complex networks of proteins to integrate and synergistically regulate their activities. Capturing the full extent of biological complexity requires the integration of multiple-omic disciplines that generate layers of information from the cell. From a technical perspective, the integration of multiple network types allows one to verify and complement one another (Tolani et al., 2021; Shen et al., 2023; Depuydt et al., 2023). Maize, as many other plants, accumulates a vast and diverse type of metabolites (Riedelsheimer et al., 2012; Wen et al., 2014; Zhou et al., 2019). Despite major advances in the understanding of the genetic and external factors that influence metabolic variation and accumulation in maize, transcriptional regulators of many of these metabolic pathways are largely unknown. This represents a knowledge gap that could be bridged by integrating multiple networks, which leverages the continuously growing multi-omic data available in maize (Liu et al., 2016; Walley et al., 2016; Wen et al., 2016; Jin et al., 2017; de Abreu E Lima et al., 2018; Lee et al., 2019; Schaefer et al., 2018; Wen et al., 2018). In this study, I analyzed three distinct data types (PDI, expression, and natural variation) and constructed four

different molecular networks (layers). I utilized various integration methods to prioritize transcriptional regulators associated with specific biological processes, which allows me to gained valuable insights into potential regulatory mechanisms underlying maize metabolism - as well as developmental-related processes - paving the way for designing specific experiments aimed at crop improvement, metabolic engineering, and basic gene regulation understanding of the of the corresponding processes.

The rapid generation of multi-omic genomic data in maize has led to growing efforts to implement integration strategies (Liu et al., 2016; Walley et al., 2016; Wen et al., 2016; Jin et al., 2017; de Abreu E Lima et al., 2018; Lee et al., 2019; Schaefer et al., 2018; Wen et al., 2018). However, the large majority of the studies relies on the idea of verifying each layer with one another (i.e., conceptual integration) (Depuydt et al., 2023), with a few exceptions where layers are used to level up each layer with one another (Schaefer et al., 2018; Yang et al., 2022) or to learn from the combination of them (Han et al., 2023). Here, I implemented three different integration strategies to make functional annotations of the TFs (Figure 4.1). Our findings indicate that the integration of multiple layers based on *common targets* and *functions*, although more intuitive, does not effectively recover observed GO terms in knockouts (Figure 4.2d, e). Instead, it frequently yields results that can be readily attributed to chance, as demonstrated by the number of times that a GO term may be retrieved from random networks, as well as their high similarity with the GO terms from the true network (Figure 4.2f-i). Surprisingly, the *common targets* strategy predicts a similar number (Figure 4.2a, g) and category (Figure 4.2i) of GO terms in random networks as in the true/observed networks (Figure 4.2a, g). From a technical perspective, this suggests that the initial set of interactions contains a significant number of false positives, which explains why random networks can recover similar sets of GO terms. Overall, the number of GO

terms and their significance (P values) in the enrichment analysis with random networks (Figure 4.2g, h), suggest that *common target* and *common function* are more lenient than *network-based*. This drawback is compounded by the inherent technical noise associated with corresponding layers (e.g., PDI without transcriptional effect). Furthermore, it indicates that even when a TF-target interaction is highly reliable (due to its presence in multiple layers), it alone is insufficient to provide an accurate representation of the biological landscape associated with the corresponding TF (Figure 4.2d). Interestingly, unlike the first two methods, the network-based approach proved to predict GO terms that are less likely to be observed from a random network. I interpreted this as a sign of robustness in the identification of genes truly functionally related (Figure 4.2g-i). My contention is that this robustness is rooted in the inherent nature of the embedding generation process, as it is highly improbable to observe similar wiring patterns across layers, despite the expected presence of potential false positives within each respective layer. Additionally, of equal significance, the *network-based* approach facilitated predictions for a considerably larger number of TFs (Figure 3.4c), thereby influencing the design of future experiments aimed at uncovering and validating specific TF functions in maize.

In general, TF expand their regulatory repertoire through functional or physical interactions with other TFs (Reményi et al., 2004; Brkljacic and Grotewold, 2017), as evidence the formation of regulatory cluster both at the level of TF-target genes (Tu et al., 2020) and in the organization cis-regulatory elements across cell types (Marand et al., 2021). Here, combining all the TF-function predictions made by our *network-based* integration I find a network-like structure independent of the number of genes by GO term or targets by TF (Figure 4.3a-d). Using a scaled enrichment score for each TF and GO term, I showed that only ~3% of our predictions had a single TF as the primary regulator of the corresponding GO term, which indicated that most TFs

contributes to the regulation of multiple functions, and that the regulation of a biological process requires the involvement of multiple TFs. These similarities between the patterns observed here and previously reported ones are interpreted as validation of the presented results. Additionally, I prioritized TFs by biological process combining scaled scores from both TF and GO terms and built two tier regulatory models for specific biological processes. Noteworthy, the top two TFs ABA-related (NAC56 and WRI2) are differentially expressed under drought and cold conditions (Hoopes et al., 2019), both conditions trigger the accumulation of ABA (Cutler et al., 2010; Waadt et al., 2022). Interestingly, WRI2 - a homeolog of the lipid metabolism master regulator WRI1 (Pouvreau et al., 2011) - was also in the top three of the TFs related to lipid-related metabolism (Figure 3.3f, second panel) highlighting molecular connections between ABA signaling/control and lipid metabolism (Guschina et al., 2002; Chen et al., 2020). Similarly, five out of the previously identified maize regulators (Yang et al., 2017) and predicted by here as regulator of phenolic-related genes included LBD24, which has the higher enrichment score (rZ = 0.08) confirming its previous identification as a highly-connected TF within the phenolic metabolism Y1H network (Yang et al., 2017). Finally, within the leaf-related predictions, I find MYBR4 as the top prediction linked with leaf morphogenesis term (GO:0009965) (Figure 4.3f, four panel). The closest MYBR4's protein in Arabidopsis is AtMYB46 (AT5G12870, 68% identity and 29% coverage), which is a direct target of SECONDARY WALL-ASSOCIATED NAC DOMAIN PROTEIN1 (SND1) and works as regulator of secondary wall biosynthesis in fibers and vessels in Arabidopsis (Zhong et al., 2007). Thus, it provides a plausible mechanism for the association of MYBR4 with leaf morphology in maize. Altogether, this highlights the biological relevance of the associations this analysis predicted. I showed the presence of feed-forward motifs within my results, which are known as a mechanism for reinforcement of regulatory signals (Alon, 2007). Together, these

discoveries expand the anticipation of TF regulators of metabolic pathways to encompass a wider array of biological processes, bearing significant implications for forthcoming biotechnological applications, such as precise modifications of developmental processes, for instance.

In summary, I built four different gene networks in maize, which included the re-analysis of almost 300 PDI assays under the same pipeline, and the associations of ~15M with public expression in a population of >300 inbred lines. I integrated these datasets with co-expression networks from our previous work (Zhou et al., 2020). Considering the inherent challenge posed by the variations in each respective network, I examined three distinct integration methods and employed two different strategies to functionally annotate TFs. Our findings demonstrated that integrating all layers, followed by the identification of highly-similar genes based on their embeddings, enabled the identification of genes which functionally allows the annotation of >1,000 TFs. Notably, the embedding similarities create a network of gene-gene associations involving over 24K genes. This study focused exclusively on regulatory-related genes, such as TFs and coregulators. Nevertheless, I foresee that the resources provided here for the remaining unexplored ~22k genes will also offer a wider array of functionalities.

4.5 METHODS

4.5.1 Genetic markers

A set of 304 diverse inbred lines with publicly available SNP and gene expression information were included in our eQTL analysis (Bukowski et al., 2018; Kremling et al., 2018; Mazaheri et al., 2019). SNP marker data from whole genome sequencing along with RNA-seq were combined between studies based on physical positions. In the case that an overlap was observed between the two datasets, the RNA-seq marker was preferentially kept. The expression datasets capture variation both at the genotypic and tissue level.

4.5.2 RNA-seq and co-expression data

All the RNA-seq and co-expression datasets utilized here were previously published (Zhou et al., 2020), except for co-expression network 46, which was constructed using the 304 inbred lines analyzed for genetic markers (referred to as n304). Specifically, pre-mapped CPM values were gathered for the respective inbred lines and employed the exact strategy outlined by Zhou et al. (2020) to construct the corresponding co-expression network, ensuring comparability among all 46 networks. All co-expression networks were based on RandomForestRegressor and using the top 100K association by TF.

4.5.3 eQTL identification and classification

eQTLs were identified using eight distinct tissue types encompassing different developmental stages from germination to plant maturity (GRoot, Gshoot, Kern, L3Base, L3Tip, LMAD, LMAN, and seedling) (Bukowski et al., 2018; Kremling et al., 2018; Mazaheri et al., 2019). SNPs were filtered by removing non-biallelic markers, and those with a minor allele frequency < 0.05. Each of the tissue-specific expression datasets were filtered independently by retaining genes with ≥ 6 reads in ≥ 20% of samples and ≥ 0.1 TPM in ≥ 20% of samples. After filtering, it was tested between 15.5M - 16.7M SNPs against the expression of 15.3K - 26.4K genes across the eight tissue types. Briefly, to test SNP-gene associations, a series of eight candidate linear models were fitted beginning with a naive T-test then progressively controlling for different levels of kinship and population structure in a mixed linear model. For each model tested, the association was deemed significant if the observed P-value surpassed the 10K permutation threshold computed for each gene. Non-significant eQTLs were discarded when the association was supported by fewer than two of the candidate linear models and when the association involved non-syntenic genes (Schnable, 2019). The significant associations were classified as cis-eQTLt, trans/cis-eQTL, cis-

eQTL, trans-eQTL, and unassigned eQTL according to the distance between each eQTL and its corresponding target gene as well as its co-location with annotated maize genes (genome B73-V4, Figure S4.2a).

4.5.4 Protein-DNA interactions data analysis

Raw reads from classic ChIP- (Bolduc et al., 2012; Morohashi et al., 2012; Eveland et al., 2014; Pautler et al., 2015; Li et al., 2018; Zhan et al., 2018; Dong et al., 2019), ChIP-seq from protoplast (pChIP-seq) (Tu et al., 2020), and DAP-seq (Ricci et al., 2019; Dong et al., 2020) were collected from publicly available dataset. Reads quality control and peaks identification was performed as reported previously (Gomez-Cano et al., 2022). Briefly, read quality control was performed using FastQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/, V0.11.5). Adapters and low quality reads were trimmed with Trimmomatic (Bolger et al., 2014) using the following parameters: ILLUMINACLIP:Adapter.fastq:2:40:15 SLIDINGWINDOW:4:20 MINLEN:30. Cleaned reads were mapped to the maize genome (V4) (Jiao et al., 2017) with Bowtie2 v2.3.4.1 (Langmead and Salzberg, 2012) and only using nuclear chromosomes. Multi-mapping reads were filtered with Samtools v1.9 (Li et al., 2009) (q 30). Peaks were called using GEM v3.4 (Guo et al., 2012). In DAP-seq assays, HALO vector was used as a control. Peaks from classic ChIP-seq were called including duplicates and using the corresponding mutants or tag-protein as control. Finally, ChIP from prototals were called including replicates. All of them used the following parameters: --d Read_Distribution_default.txt --k_min 6 --k_max 15 --k_seqs 2000 --outNP --sl. Only TFs with >500 predicted peaks were used for further analysis.

Peak quality control. Peaks were filtered by testing overlapping with ACRs, and by scaling the number of Counts per Million (CPMs) by peaks per assay. Briefly, peaks with a Z-score larger than -0.5 and mapping to ACR were kept for further analysis. The CPMs were obtained after

extending the peaks' summit 50 bps and converted to SAF files for counting mapped reads per peak using Rsubread v1.32.2 (Liao et al., 2019). Available accessible chromatin regions (ACRs) were collected from previously published maize ATLAS (Marand et al., 2021). Peaks with a Z-score ≤ -0.5 and mapping to ACRs were retained for further analysis.

4.5.5 Functional annotation

All functional annotations were performed after discarding non syntenic genes. PWYs were collected from CornCYC (Andorf et al., 2016), and GO terms were obtained from GAMER (Wimalanathan et al., 2018)**.** Syntenic genes were defined based on Schnable et al. (2019). Enrichment analysis for PWYs and GO terms was conducted in R using the GeneOverlap (v1.30.0) and topGO (v2.46.0) packages, respectively. GO term semantic similarity was calculated using the GOSemSim (v2.20.0) package in R, employing the "*Wang*" method. For *common function* analysis, all GO comparisons were performed on the original set of enriched GO terms, and after the comparisons (GO semantic similarity), all GO terms were mapped to their closest parent GO terms using the R package Rrvgo (v1.6) (Sayols, 2023). Similarly, all GO terms significantly enriched from *common target* and *network-based* analyses were mapped to its closed parent before any description to reduce redundancy.

4.5.6 Network integration

Common interactions. I compared TF-target associations across all layers (i.e., GAN, GRN, eGRN, and CEN networks) and considered interactions as common when they are present in at least two different layers for the same corresponding TF. Subsequently, I assessed the enrichment of PWYs and GO terms using the common TF interactions. Any significant GO terms were then mapped to their closest parent terms.

Common function. Common functions were identified by testing the enrichment of target genes with PWYs and GO terms for each TF in each layer. TFs that had at least one PWY/GO term enriched in at least two different layers were retained to assess common predictions across layers (Figure S3.5a). The similarity in PWY predictions among layers was performed by comparing all PWYs between layers for each TF using a Fisher exact test. Enrichment test was used because a single gene could be annotated in multiple PWYs. Hence, PWYS were considered overplayed only if they exhibited a significant number of overlapped genes (P-value < 0.05). The similarity in GO term predictions were performed by measuring the semantic similarity between the corresponding terms. Significant GO terms are then mapped to their closest parent terms.

Network-based. All four layers were combined and then scaled the interaction frequencies from 0.5 to 1, as follow 0.5 + (0.5/4)*N, being N the number of times that same interaction was observed. Embeddings of the scaled network were identified with PecanPy (Liu and Krishnan, 2021) using the following parameters: --weighted --dimensions 50 --walk-length 80 --num-walks 10 --directed. Gene similarity was assessed by computing the mutual rank (MR) of the mutual information (MI) using the following formula: $MR_{MI} = sqrt(MI\_rank * tMI\_rank)$, where MI_rank represents the rank of the MI matrix and tMI_rank represents the transpose matrix of MI_rank. The MI was calculated using the R package Parmigene (Sales and Romualdi, 2011). To select highly similar genes by TF based on its MR_{MI} I used a decay function as follows: $D = e^{-(MRMI\ -1)/50}$. D values ≤ 0.05 were taken as highly similar (Wisecaver et al., 2017). After identifying genes highly similar per TF, I proceed to test the enrichment of PWY and GO terms. Significant GO terms are then mapped to their closest parent terms.

4.5.7 Knockout and random network validation

PWY and GO term predictions for TFs were contrasted with enriched PWY and GO terms identified in knockout analysis using DEGs and their corresponding log2FC values. I used data from previous studies for KN1 (Bolduc et al., 2012), RA1 (Eveland et al., 2014), FEA4 (Pautler et al., 2015), O2 (Zhan et al., 2018), bZIP22 (Li et al., 2018), and TB1 (Dong et al., 2019) reanalyzed in (Zhou et al., 2020). Additional knockout data (MYBR32_m1, WRKY82_m1, HSF13m1m2, HSF18m1, HSF20m1, HSF29m1, HSF29m2, WRKY2m2, WRKY8m1, and WRKY8m2) were collected from (Ellison et al., 2023). The enrichment of PWY and GO terms were performed with DEG selected based on adjusted P-value as reported by DESeq2 (Padj ≤ 0.05) (Love et al., 2014) and following indication described above (Methods session 4.5.5). The similarities between PWYs and GO terms predicted by each integration method and those observed in the corresponding knockout were estimated using PWY overlapping and GO semantic similarity, as previously described (Methods sections 4.5.5**).** Gene set enrichment analyses were conducted using the R package FGSEA (v1.20) (Korotkevich et al., 2021) with the parameters: minSize = 5, maxSize = 1000, and eps = 0. The gene sets tested were defined based on the predicted PWYs and GO terms for each TF, considering the available knockout data (Methods section 4.5.6). The fraction of recovered predictions was calculated by determining the number of significant (P-value ≤ 0.05) PWYs and GO terms out of the total tested.

The comparison of each method's prediction against the random networks was conducted by randomizing each of the four initial networks (GRN, eGRN, CEN, and GAN) 3,000 times, generating 3,000 random versions of each layer. Subsequently, I annotated and integrated each set of random networks following the procedures described in Methods sections 4.5.5 and 4.5.6, similar to the original networks. All random networks were generated using the “rewire” function

from the R package Igraph (v1.2.4.1), with the following parameters: avoided loops and with niter = NodesInNetwork * 10000).

4.5.8 Prioritization of transcriptional regulators-process associations

All prioritization analyses were conducted using *network-based* results. GO terms with less than 800 genes were retained, and after mapping excessively specific GO terms (≤ 50 genes) to their corresponding GO terms parent. Mapping to parent terms was performed following the procedures described in Methods section 3.5.5. Then, I proceeded to calculate the enrichment score associated with each TF-GO association as follow:

$$E_{ij} = \mathrm{Log}_2[(c/t)/(p/u)]$$

Where E_{ij} is the enrichment score of the TF*i* with the GO*j*, *c* is the intersection of target genes of TF*i* and annotated genes on GO*j*, *t* is the total number of target genes of TF*i*, *p* is the total number of genes annotated on GO*j*, and *u* is the total number of genes in maize, which in this case refers to the total number of syntenic genes with sorghum (Schnable, 2019). All Eij values were subsequently normalized by each TFi and GOj as follows:

$$Zi = (Eij - Ui)/\sigma i$$

and

$$Zj = (Eij - Uj)/\sigma j$$

Here, Ui and Uj represent the average enrichment score value for all the GOj associated with TFi and all the TFi associated with GOj, respectively. Similarly, σi and σj represent the standard deviation of the enrichment score value for all the GOj associated with TFi and all the TFi associated with GOj, respectively. Finally, I calculated the reciprocal Z-score (rZ) as follow:

$$rZij = \mathrm{sqrt}\ (\max(0, Zi)^{\wedge 2} + \max(0, Zj)^{\wedge 2})$$

4.5.9 Similarities in sequence among TF paralogs

Sequences for all peptides associated with the corresponding pair of paralogs were collected from MaizeGDB (https://maizegdb.org/) using genome v4 (Jiao et al., 2017). TFs' similarities were calculated by averaging the Hamming distance between all amino acid sequences associated with the respective TFs. The Hamming distance was computed using the R package DECIPHER (v2.22) (Wright, 2016) and the "DistanceMatrix" function with the following parameters: includeTerminalGaps = TRUE, penalizeGapLetterMatches = TRUE, and correction = "none".

REFERENCES

de Abreu E Lima, F., Li, K., Wen, W., Yan, J., Nikoloski, Z., Willmitzer, L., and Brotman, Y. (2018). Unraveling lipid metabolism in maize with time-resolved multi-omics data. Plant J. **93**: 1102–1115.

Alon, U. (2007). Network motifs: theory and experimental approaches. Nat. Rev. Genet. **8**: 450–461.

Andorf, C.M. et al. (2016). MaizeGDB update: new tools, data and interface for the maize model organism database. Nucleic Acids Res. **44**: D1195–201.

Arda, H.E. and Walhout, A.J.M. (2010). Gene-centered regulatory networks. Brief. Funct. Genomics **9**: 4–12.

Bartrina, I., Otto, E., Strnad, M., Werner, T., and Schmülling, T. (2011). Cytokinin regulates the activity of reproductive meristems, flower organ size, ovule formation, and thus seed yield in Arabidopsis thaliana. Plant Cell **23**: 69–80.

Birkenbihl, R.P., Liu, S., and Somssich, I.E. (2017). Transcriptional events defining plant immune responses. Curr. Opin. Plant Biol. **38**: 1–9.

Bolduc, N., Yilmaz, A., Mejia-Guerra, M.K., Morohashi, K., O'Connor, D., Grotewold, E., and Hake, S. (2012). Unraveling the KNOTTED1 regulatory network in maize meristems. Genes Dev. **26**: 1685–1690.

Bolger, A.M., Lohse, M., and Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics **30**: 2114–2120.

Bowles, A.M.C., Bechtold, U., and Paps, J. (2020). The Origin of Land Plants Is Rooted in Two Bursts of Genomic Novelty. Curr. Biol. **30**: 530–536.e2.

Brkljacic, J. and Grotewold, E. (2017). Combinatorial control of plant gene expression. Biochim. Biophys. Acta **1860**: 31–40.

Bukowski, R. et al. (2018). Construction of the third-generation Zea mays haplotype map. Gigascience **7**: 1–12.

Chen, K., Li, G.-J., Bressan, R.A., Song, C.-P., Zhu, J.-K., and Zhao, Y. (2020). Abscisic acid dynamics, signaling, and functions in plants. J. Integr. Plant Biol. **62**: 25–54.

Colasanti, J., Yuan, Z., and Sundaresan, V. (1998). The indeterminate gene encodes a zinc finger protein and regulates a leaf-generated signal required for the transition to flowering in maize. Cell **93**: 593–603.

Cutler, S.R., Rodriguez, P.L., Finkelstein, R.R., and Abrams, S.R. (2010). Abscisic acid: emergence of a core signaling network. Annu. Rev. Plant Biol. **61**: 651–679.

Deplancke, B. et al. (2006). A gene-centered C. elegans protein-DNA interaction network. Cell

125: 1193–1205.

Depuydt, T., De Rybel, B., and Vandepoele, K. (2023). Charting plant gene functions in the multi-omics and single-cell era. Trends Plant Sci. **28**: 283–296.

Dong, Z., Xiao, Y., Govindarajulu, R., Feil, R., Siddoway, M.L., Nielsen, T., Lunn, J.E., Hawkins, J., Whipple, C., and Chuck, G. (2019). The regulatory landscape of a core maize domestication module controlling bud dormancy and growth repression. Nat. Commun. **10**: 3810.

Dong, Z., Xu, Z., Xu, L., Galli, M., Gallavotti, A., Dooner, H.K., and Chuck, G. (2020). Necrotic upper tips1 mimics heat and drought stress and encodes a protoxylem-specific transcription factor in maize. Proc. Natl. Acad. Sci. U. S. A. **117**: 20908–20919.

Ellison, E.L., Zhou, P., Hermanson, P., Chu, Y.-H., Read, A., Hirsch, C.N., Grotewold, E., and Springer, N.M. (2023). Mutator transposon insertions within maize genes often provide a novel outward reading promoter. bioRxiv: 2023.06.05.543741.

Erenstein, O., Jaleta, M., Sonder, K., Mottaleb, K., and Prasanna, B.M. (2022). Global maize production, consumption and trade: trends and R&D implications. Food Security **14**: 1295–1319.

Eveland, A.L. et al. (2014). Regulatory modules controlling maize inflorescence architecture. Genome Res. **24**: 431–443.

Galli, M., Khakhar, A., Lu, Z., Chen, Z., Sen, S., Joshi, T., Nemhauser, J.L., Schmitz, R.J., and Gallavotti, A. (2018). The DNA binding landscape of the maize AUXIN RESPONSE FACTOR family. Nat. Commun. **9**: 4526.

Gomez-Cano, F., Chu, Y.-H., Cruz-Gomez, M., Abdullah, H.M., Lee, Y.S., Schnell, D.J., and Grotewold, E. (2022). Exploring Camelina sativa lipid metabolism regulation by combining gene co-expression and DNA affinity purification analyses. Plant J.

Guo, Y., Mahony, S., and Gifford, D.K. (2012). High resolution genome wide binding event finding and motif discovery reveals transcription factor spatial binding constraints. PLoS Comput. Biol. **8**: e1002638.

Gupta, O.P., Deshmukh, R., Kumar, A., Singh, S.K., Sharma, P., Ram, S., and Singh, G.P. (2021). From gene to biomolecular networks: a review of evidences for understanding complex biological function in plants. Curr. Opin. Biotechnol. **74**: 66–74.

Guschina, I.A., Harwood, J.L., Smith, M., and Beckett, R.P. (2002). Abscisic acid modifies the changes in lipids brought about by water stress in the moss Atrichum androgynum. New Phytol. **156**: 255–264.

Han, L. et al. (2023). A multi-omics integrative network map of maize. Nat. Genet. **55**: 144–153.

Hoopes, G.M., Hamilton, J.P., Wood, J.C., Esteban, E., Pasha, A., Vaillancourt, B., Provart, N.J., and Buell, C.R. (2019). An updated gene atlas for maize reveals organ-specific and

stress-induced genes. Plant J. **97**: 1154–1167.

Hufford, M.B. et al. (2021). De novo assembly, annotation, and comparative analysis of 26 diverse maize genomes. Science **373**: 655–662.

Jiao, Y. et al. (2017). Improved maize reference genome with single-molecule technologies. Nature **546**: 524–527.

Jin, M. et al. (2017). Integrated genomics-based mapping reveals the genetics underlying maize flavonoid biosynthesis. BMC Plant Biol. **17**: 17.

Kono, T.J.Y., Brohammer, A.B., McGaugh, S.E., and Hirsch, C.N. (2018). Tandem Duplicate Genes in Maize Are Abundant and Date to Two Distinct Periods of Time. G3 **8**: 3049–3058.

Korotkevich, G., Sukhov, V., Budin, N., Shpak, B., Artyomov, M.N., and Sergushichev, A. (2021). Fast gene set enrichment analysis. bioRxiv: 060012.

Kremling, K.A.G., Chen, S.-Y., Su, M.-H., Lepak, N.K., Romay, M.C., Swarts, K.L., Lu, F., Lorant, A., Bradbury, P.J., and Buckler, E.S. (2018). Dysregulation of expression correlates with rare-allele burden and fitness loss in maize. Nature **555**: 520–523.

Kryuchkova-Mostacci, N. and Robinson-Rechavi, M. (2017). A benchmark of gene expression tissue-specificity metrics. Brief. Bioinform. **18**: 205–214.

Kusmec, A., Srinivasan, S., Nettleton, D., and Schnable, P.S. (2017). Distinct genetic architectures for phenotype means and plasticities in Zea mays. Nat Plants **3**: 715–723.

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

Lee, T., Lee, S., Yang, S., and Lee, I. (2019). MaizeNet: a co-functional network for network-assisted systems genetics in Zea mays. Plant J. **99**: 571–582.

Liao, Y., Smyth, G.K., and Shi, W. (2019). The R package Rsubread is easier, faster, cheaper and better for alignment and quantification of RNA sequencing reads. Nucleic Acids Res. **47**: e47–e47.

Li, C., Yue, Y., Chen, H., Qi, W., and Song, R. (2018). The ZmbZIP22 Transcription Factor Regulates 27-kD γ-Zein Gene Transcription during Maize Endosperm Development. Plant Cell **30**: 2402–2424.

Li, H. et al. (2013). Genome-wide association study dissects the genetic architecture of oil biosynthesis in maize kernels. Nat. Genet. **45**: 43–50.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., and 1000 Genome Project Data Processing Subgroup (2009). The Sequence Alignment/Map format and SAMtools. Bioinformatics **25**: 2078–2079.

Li, L., Briskine, R., Schaefer, R., Schnable, P.S., Myers, C.L., Flagel, L.E., Springer, N.M., and Muehlbauer, G.J. (2016). Co-expression network analysis of duplicate genes in maize

(Zea mays L.) reveals no subgenome bias. BMC Genomics **17**: 875.

Liu, H. et al. (2016). MODEM: multi-omics data envelopment and mining in maize. Database **2016**.

Liu, R. and Krishnan, A. (2021). PecanPy: a fast, efficient, and parallelized Python implementation of node2vec. Bioinformatics **37**: 3377–3379.

Love, M.I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. **15**: 550.

Mack, K.L. and Nachman, M.W. (2017). Gene Regulation and Speciation. Trends Genet. **33**: 68–80.

Marand, A.P., Chen, Z., Gallavotti, A., and Schmitz, R.J. (2021). A cis-regulatory atlas in maize at single-cell resolution. Cell **184**: 3041–3055.e21.

Marand, A.P., Eveland, A.L., Kaufmann, K., and Springer, N.M. (2023). cis-Regulatory Elements in Plant Development, Adaptation, and Evolution. Annu. Rev. Plant Biol. **74**: 111–137.

Mathur, S., Vyas, S., Kapoor, S., and Tyagi, A.K. (2011). The Mediator complex in plants: structure, phylogeny, and expression profiling of representative genes in a dicot (Arabidopsis) and a monocot (rice) during reproduction and abiotic stress. Plant Physiol. **157**: 1609–1627.

Mazaheri, M. et al. (2019). Genome-wide association analysis of stalk biomass and anatomical traits in maize. BMC Plant Biol. **19**: 45.

McMullen, M.D. et al. (2009). Genetic properties of the maize nested association mapping population. Science **325**: 737–740.

Mejia-Guerra, M.K., Pomeranz, M., Morohashi, K., and Grotewold, E. (2012). From plant gene regulatory grids to network dynamics. Biochimica et Biophysica Acta (BBA) - Gene Regulatory Mechanisms **1819**: 454–465.

Minow, M.A.A., Ávila, L.M., Turner, K., Ponzoni, E., Mascheretti, I., Dussault, F.M., Lukens, L., Rossi, V., and Colasanti, J. (2018). Distinct gene networks modulate floral induction of autonomous maize and photoperiod-dependent teosinte. J. Exp. Bot. **69**: 2937–2952.

Morohashi, K. et al. (2012). A genome-wide regulatory framework identifies maize pericarp color1 controlled genes. Plant Cell **24**: 2745–2764.

Nakashima, K., Yamaguchi-Shinozaki, K., and Shinozaki, K. (2014). The transcriptional regulatory network in the drought response and its crosstalk in abiotic stress responses including drought, cold, and heat. Front. Plant Sci. **5**: 1–7.

O'Malley, R.C., Huang, S.S.C., Song, L., Lewsey, M.G., Bartlett, A., Nery, J.R., Galli, M., Gallavotti, A., and Ecker, J.R. (2016). Cistrome and Epicistrome Features Shape the

Regulatory DNA Landscape. Cell **165**: 1280–1292.

Pautler, M., Eveland, A.L., LaRue, T., Yang, F., Weeks, R., Lunde, C., Je, B.I., Meeley, R., Komatsu, M., Vollbrecht, E., Sakai, H., and Jackson, D. (2015). FASCIATED EAR4 encodes a bZIP transcription factor that regulates shoot meristem size in maize. Plant Cell **27**: 104–120.

Pouvreau, B., Baud, S., Vernoud, V., Morin, V., Py, C., Gendrot, G., Pichon, J.-P., Rouster, J., Paul, W., and Rogowsky, P.M. (2011). Duplicate maize Wrinkled1 transcription factors activate target genes involved in seed oil biosynthesis. Plant Physiol. **156**: 674–686.

Reményi, A., Schöler, H.R., and Wilmanns, M. (2004). Combinatorial control of gene expression. Nat. Struct. Mol. Biol. **11**: 812.

Renny-Byfield, S., Rodgers-Melnick, E., and Ross-Ibarra, J. (2017). Gene Fractionation and Function in the Ancient Subgenomes of Maize. Mol. Biol. Evol. **34**: 1825–1832.

Ricci, W.A. et al. (2019). Widespread long-range cis-regulatory elements in the maize genome. Nature Plants **5**: 1237–1249.

Riedelsheimer, C., Lisec, J., Czedik-Eysenberg, A., Sulpice, R., Flis, A., Grieder, C., Altmann, T., Stitt, M., Willmitzer, L., and Melchinger, A.E. (2012). Genome-wide association mapping of leaf metabolic profiles for dissecting complex traits in maize. Proc. Natl. Acad. Sci. U. S. A. **109**: 8872–8877.

Rodgers-Melnick, E., Vera, D.L., Bass, H.W., and Buckler, E.S. (2016). Open chromatin reveals the functional maize genome. Proc. Natl. Acad. Sci. U. S. A. **113**: E3177–84.

Sales, G. and Romualdi, C. (2011). parmigene—a parallel R package for mutual information estimation and gene network reconstruction. Bioinformatics **27**: 1876–1877.

Sayols, S. (2023). rrvgo: a Bioconductor package for interpreting lists of Gene Ontology terms. MicroPubl Biol **2023**.

Schaefer, R.J., Michno, J.-M., Jeffers, J., Hoekenga, O., Dilkes, B., Baxter, I., and Myers, C.L. (2018). Integrating Coexpression Networks with GWAS to Prioritize Causal Genes in Maize. Plant Cell **30**: 2922.

Schmitz, R.J., Grotewold, E., and Stam, M. (2022). Cis-regulatory sequences in plants: Their importance, discovery, and future challenges. Plant Cell **34**: 718–741.

Schnable, J. (2019). Pan-Grass Syntenic Gene Set (sorghum referenced) with both maize v3 and maize v4 gene models. figShare.

Schnable, J.C., Springer, N.M., and Freeling, M. (2011). Differentiation of the maize subgenomes by genome dominance and both ancient and ongoing gene loss. Proc. Natl. Acad. Sci. U. S. A. **108**: 4069–4074.

Schnable, P.S. et al. (2009). The B73 maize genome: complexity, diversity, and dynamics. Science

326: 1112–1115.

Sekhon, R.S., Lin, H., Childs, K.L., Hansey, C.N., Robin Buell, C., de Leon, N., and Kaeppler, S.M. (2011). Genome-wide atlas of transcription during maize development. Plant J. **66**: 553–563.

Shen, S., Zhan, C., Yang, C., Fernie, A.R., and Luo, J. (2023). Metabolomics-centered mining of plant metabolic diversity and function: Past decade and future perspectives. Mol. Plant **16**: 43–63.

Shrestha, V., Yobi, A., Slaten, M.L., Chan, Y.O., Holden, S., Gyawali, A., Flint-Garcia, S., Lipka, A.E., and Angelovici, R. (2022). Multiomics approach reveals a role of translational machinery in shaping maize kernel amino acid composition. Plant Physiol. **188**: 111–133.

Stelpflug, S.C., Sekhon, R.S., Vaillancourt, B., Hirsch, C.N., Buell, C.R., de Leon, N., and Kaeppler, S.M. (2016). An Expanded Maize Gene Expression Atlas based on RNA Sequencing and its Use to Explore Root Development. Plant Genome **9**.

Subramanian, A., Tamayo, P., Mootha, V.K., Mukherjee, S., Ebert, B.L., Gillette, M.A., Paulovich, A., Pomeroy, S.L., Golub, T.R., Lander, E.S., and Mesirov, J.P. (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc. Natl. Acad. Sci. U. S. A. **102**: 15545–15550.

Sun, Y., Oh, D.-H., Duan, L., Ramachandran, P., Ramirez, A., Bartlett, A., Tran, K.-N., Wang, G., Dassanayake, M., and Dinneny, J.R. (2022). Divergence in the ABA gene regulatory network underlies differential growth control. Nat Plants **8**: 549–560.

Tolani, P., Gupta, S., Yadav, K., Aggarwal, S., and Yadav, A.K. (2021). Big data, integrative omics and network biology. Adv. Protein Chem. Struct. Biol. **127**: 127–160.

Tu, X., Mejía-Guerra, M.K., Valdes Franco, J.A., Tzeng, D., Chu, P.-Y., Shen, W., Wei, Y., Dai, X., Li, P., Buckler, E.S., and Zhong, S. (2020). Reconstructing the maize leaf regulatory network using ChIP-seq data of 104 transcription factors. Nat. Commun. **11**: 5089.

Waadt, R., Seller, C.A., Hsu, P.-K., Takahashi, Y., Munemasa, S., and Schroeder, J.I. (2022). Plant hormone regulation of abiotic stress responses. Nat. Rev. Mol. Cell Biol. **23**: 680–694.

Wahl, V., Ponnu, J., Schlereth, A., Arrivault, S., Langenecker, T., Franke, A., Feil, R., Lunn, J.E., Stitt, M., and Schmid, M. (2013). Regulation of flowering by trehalose-6-phosphate signaling in Arabidopsis thaliana. Science **339**: 704–707.

Walley, J.W., Sartor, R.C., Shen, Z., Schmitz, R.J., Wu, K.J., Urich, M.A., Nery, J.R., Smith, L.G., Schnable, J.C., Ecker, J.R., and Briggs, S.P. (2016). Integration of omic networks in a developmental atlas of maize. Science **353**: 814–818.

Wei, F. et al. (2007). Physical and genetic structure of the maize genome reflects its complex evolutionary history. PLoS Genet. **3**: e123.

Wen, W. et al. (2018). An integrated multi-layered analysis of the metabolic networks of different tissues uncovers key genetic components of primary metabolism in maize. Plant J. **93**: 1116–1128.

Wen, W., Li, D., Li, X., Gao, Y., Li, W., Li, H., Liu, J., Liu, H., Chen, W., Luo, J., and Yan, J. (2014). Metabolome-based genome-wide association study of maize kernel leads to novel biochemical insights. Nat. Commun. **5**: 3438.

Wen, W., Liu, H., Zhou, Y., Jin, M., Yang, N., Li, D., Luo, J., Xiao, Y., Pan, Q., and Tohge, T. (2016). Combining quantitative genetics approaches with regulatory network analysis to dissect the complex metabolism of the maize kernel. Plant Physiol. **170**: 136–146.

Wimalanathan, K., Friedberg, I., Andorf, C.M., and Lawrence-Dill, C.J. (2018). Maize GO Annotation-Methods, Evaluation, and Review (maize-GAMER). Plant Direct **2**: e00052.

Wisecaver, J.H., Borowsky, A.T., Tzin, V., Jander, G., Kliebenstein, D.J., and Rokas, A. (2017). A Global Coexpression Network Approach for Connecting Genes to Specialized Metabolic Pathways in Plants. Plant Cell **29**: 944–959.

Wright, E. (2016). Using DECIPHER v2.0 to analyze big biological sequence data in R. R J. **8**: 352.

Xiao, Y., Liu, H., Wu, L., Warburton, M., and Yan, J. (2017). Genome-wide Association Studies in Maize: Praise and Stargaze. Mol. Plant **10**: 359–374.

Yang, F. et al. (2017). A Maize Gene Regulatory Network for Phenolic Metabolism. Mol. Plant **10**: 498–515.

Yang, F., Ouma, W.Z., Li, W., Doseff, A.I., and Grotewold, E. (2016). Establishing the Architecture of Plant Gene Regulatory Networks. Methods Enzymol. **576**: 251–304.

Yang, Z., Xu, G., Zhang, Q., Obata, T., and Yang, J. (2022). Genome-wide mediation analysis: an empirical study to connect phenotype with genotype via intermediate transcriptomic data in maize. Genetics **221**.

Yilmaz, A., Nishiyama, M.Y., Garcia-Fuentes, B., Souza, G.M., Janies, D., Gray, J., and Grotewold, E. (2009). GRASSIUS: A platform for comparative regulatory genomics across the grasses. Plant Physiol. **149**: 171–180.

Zhan, J., Li, G., Ryu, C.H., Ma, C., Zhang, S., Lloyd, A., Hunter, B.G., Larkins, B.A., Drews, G.N., Wang, X., and Yadegari, R. (2018). Opaque-2 Regulates a Complex Gene Network Associated with Cell Differentiation and Storage Functions of Maize Endosperm. Plant Cell **30**: 2425–2446.

Zheng, Y. et al. (2016). iTAK: A Program for Genome-wide Prediction and Classification of Plant Transcription Factors, Transcriptional Regulators, and Protein Kinases. Mol. Plant **9**: 1667–1670.

Zhong, R., Richardson, E.A., and Ye, Z.H. (2007). The MYB46 transcription factor is a direct

target of SND1 and regulates secondary wall biosynthesis in Arabidopsis. Plant Cell **19**: 2776–2792.

Zhou, P., Enders, T.A., Myers, Z.A., Magnusson, E., and Crisp, P.A. (2021). Applying cis-regulatory codes to predict conserved and variable heat and cold stress response in maize. bioRxiv.

Zhou, P., Li, Z., Magnusson, E., Gomez Cano, F., Crisp, P.A., Noshay, J.M., Grotewold, E., Hirsch, C.N., Briggs, S.P., and Springer, N.M. (2020). Meta Gene Regulatory Networks in Maize Highlight Functionally Relevant Regulatory Interactions. Plant Cell **32**: 1377–1396.

Zhou, S., Kremling, K.A., Bandillo, N., Richter, A., Zhang, Y.K., Ahern, K.R., Artyukhin, A.B., Hui, J.X., Younkin, G.C., Schroeder, F.C., Buckler, E.S., and Jander, G. (2019). Metabolome-Scale Genome-Wide Association Studies Reveal Chemical Diversity and Genetic Control of Maize Specialized Metabolites. Plant Cell **31**: 937–955.

Zhu, G., Wu, A., Xu, X.-J., Xiao, P.-P., Lu, L., Liu, J., Cao, Y., Chen, L., Wu, J., and Zhao, X.-M. (2016). PPIM: A Protein-Protein Interaction Database for Maize. Plant Physiol. **170**: 618–626.

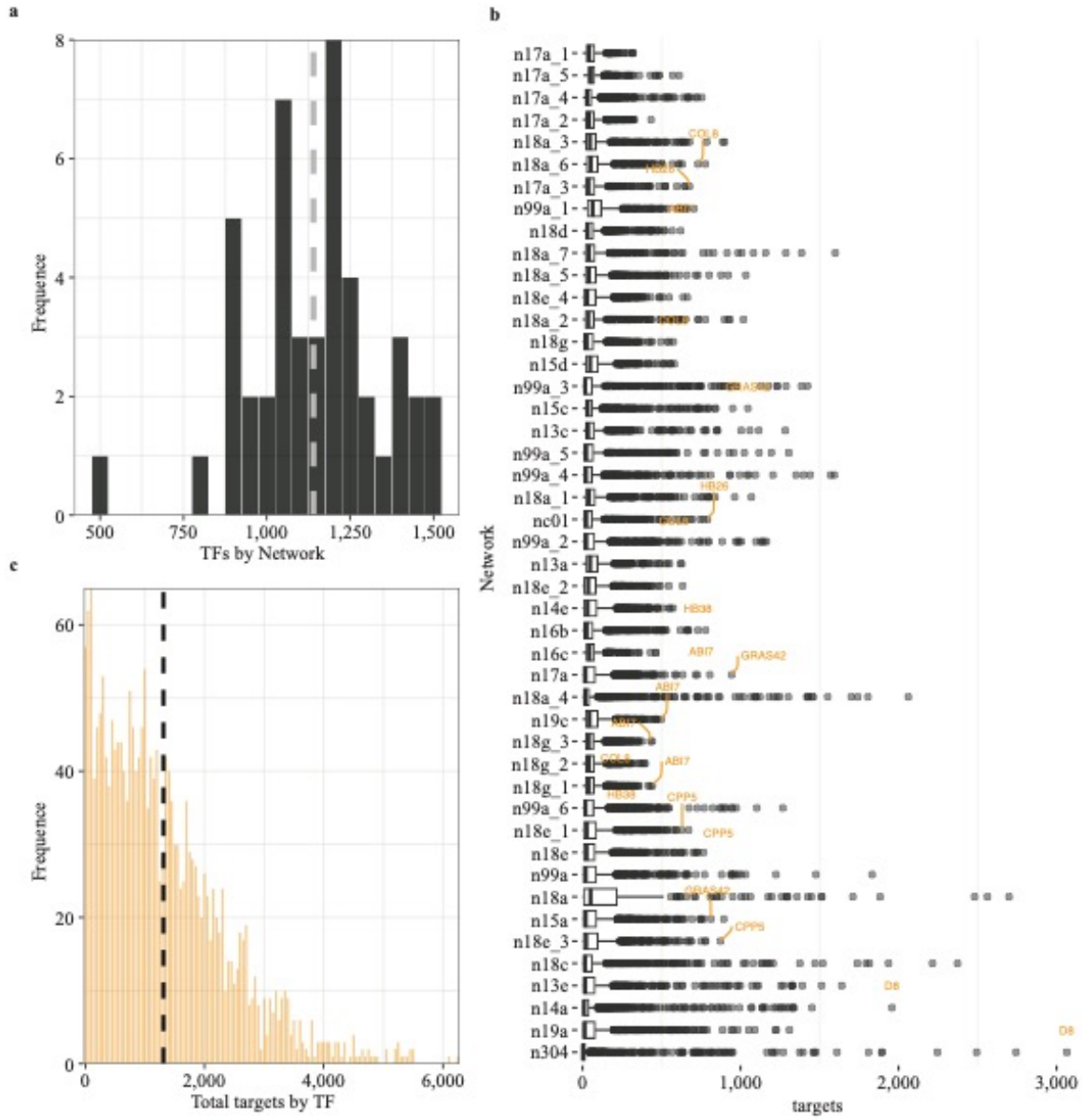

Figure S4.1 TFs and interactions used in the layer of co-expression network (CEN)

a. Histogram showing the frequency of TFs with at least a target gene per CEN. Dotted gray line indicates average TFs in all 46 CEN. **b.** Boxplot indicating total target genes per TFs across the different CENs. CEN are named following Zhou *et al*., (2022) nomenclature. Orange labels highlight TFs with the largest number of target genes in several CEN. **c**. Histogram with the frequency of total target genes per TF after combined results from all 46 CENs.

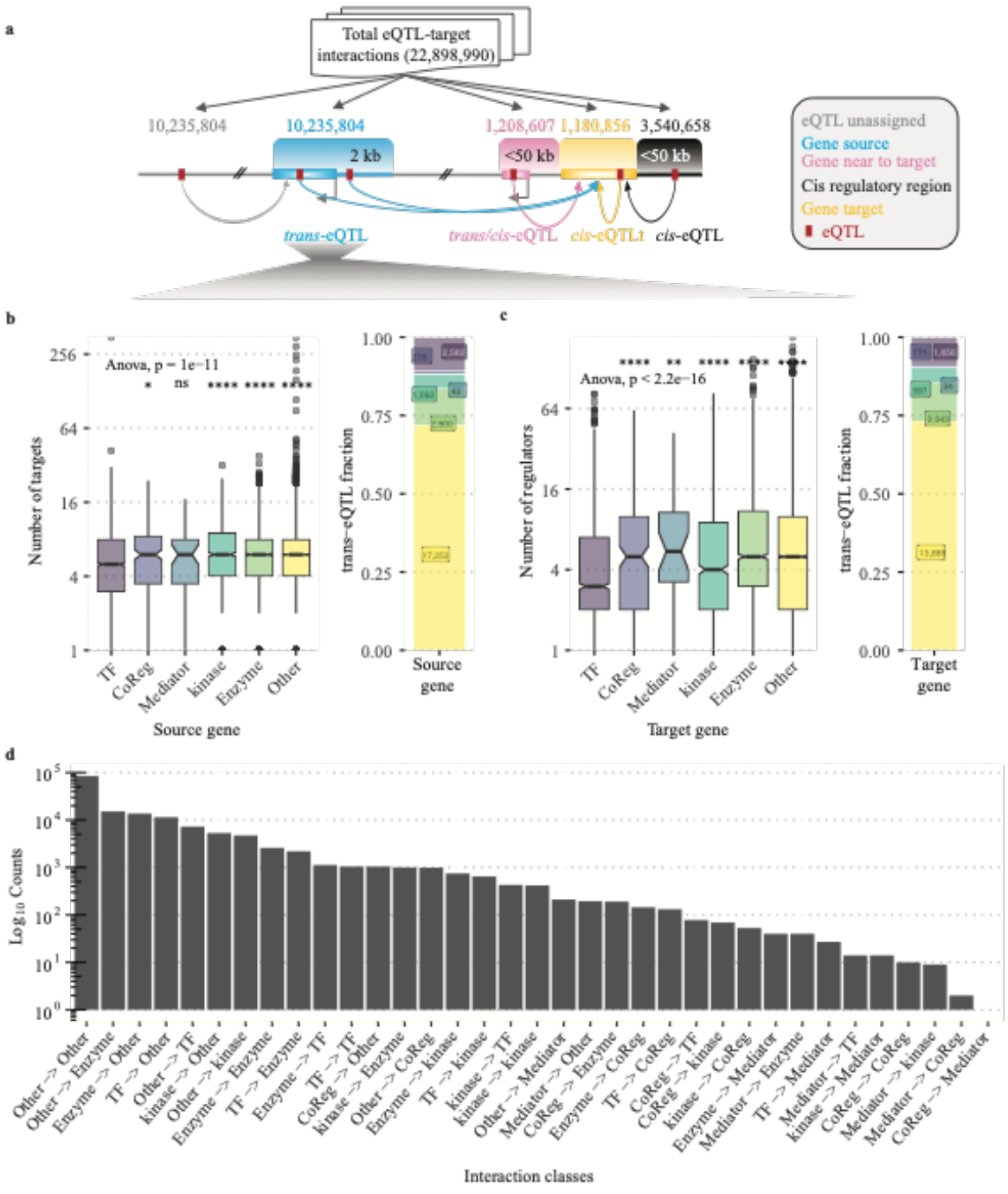

Figure S4.2 Defining a gene association network (GAN) based on trans-eQTL

a. model indicating total eQTLs identified and the classification schema used to define trans-eQTL, trans/cis-eQTL, cis-eQTLt, and cis-eQTL. Within them, trans-eQTLs were used to define the GAN. In the context of trans-eQTLs, a source gene (in blue) was defined as a gene whose promoter (2kb upstream from TSS) or gene baby overlapped with an eQTL. Genes whose expression is explained by the SNP variation were defined as gene targets (gene in yellow). **b** and **c**. I classify each source and target gene into five functional categories to count the number of

Figure S4.2 (cont'd)
associations by category (unclassified genes defined as other). **Left panel**, Boxplot indicating the number of targets (**b**) and source (**c**) genes by each gene category. **Right panel**, Stacked bar plots indicate the fraction of each gene category over the total genes in GAN. **d**. Bar plot indicating total interactions by gene category pair.

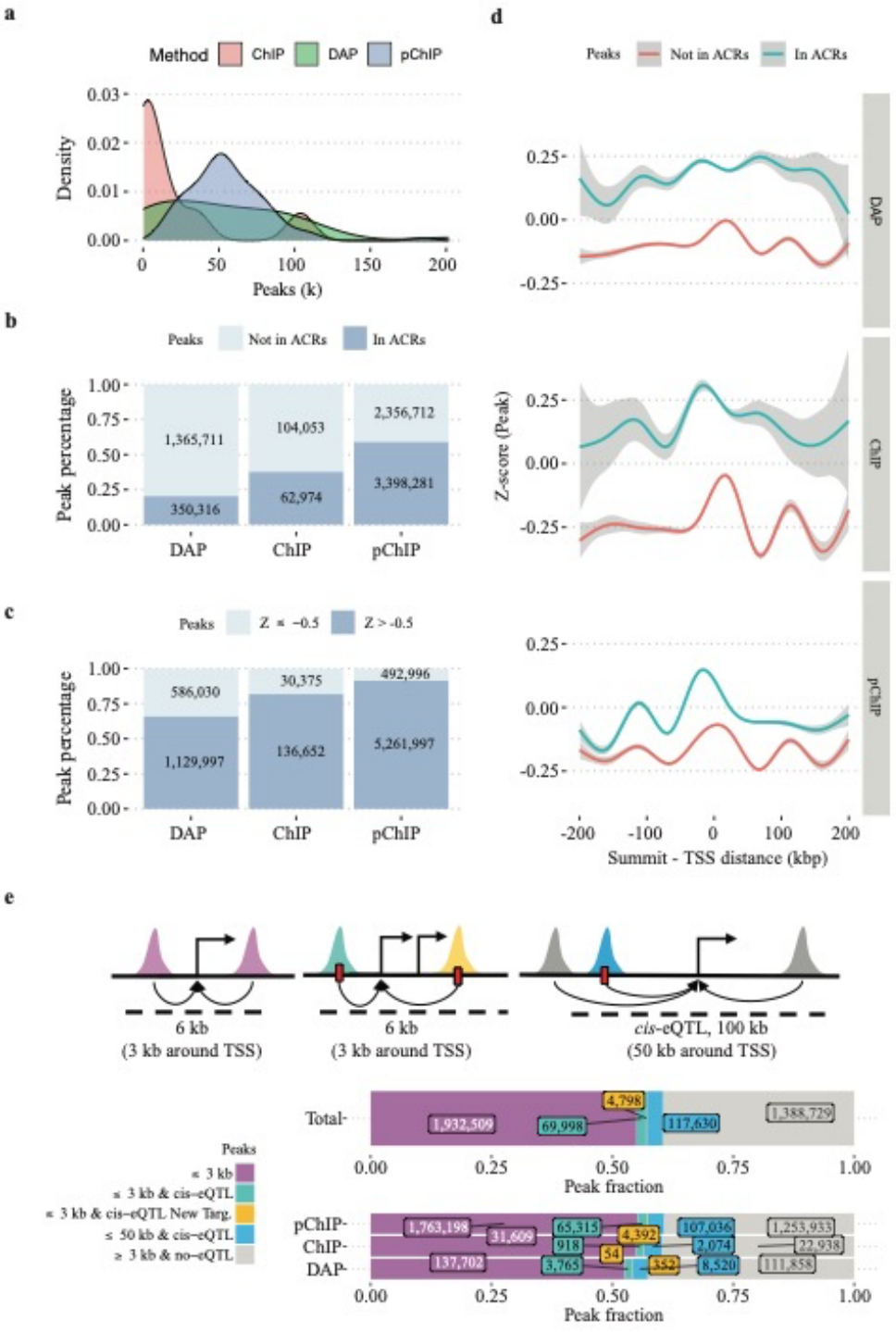

Figure S4.3 Establishing the maize gene regulatory network (GRN) layer based on protein-DNA interaction data

a. Density plot with distribution of peaks by PDI data type. **b** and **c**. Stacked bar plot with fraction of peaks mapped to accessible chromatin region (ACR) (**b**) and with low peak coverage (**c**) (CPM scaled and filtered; Z ≤ -0.5). **d**. Locally weighted scatterplot smoothing (LOESS) line plot of Z-scores by peak in 10 kb bins around 200 kb of the closest transcription start site (TSS).

Figure S4.3 (cont'd)
e. Classification schema (top) and corresponding proportion of total combined peaks (bottom, first stacked bar plot) and peaks by method (bottom, second stacked bar plot) utilized for determining target genes.

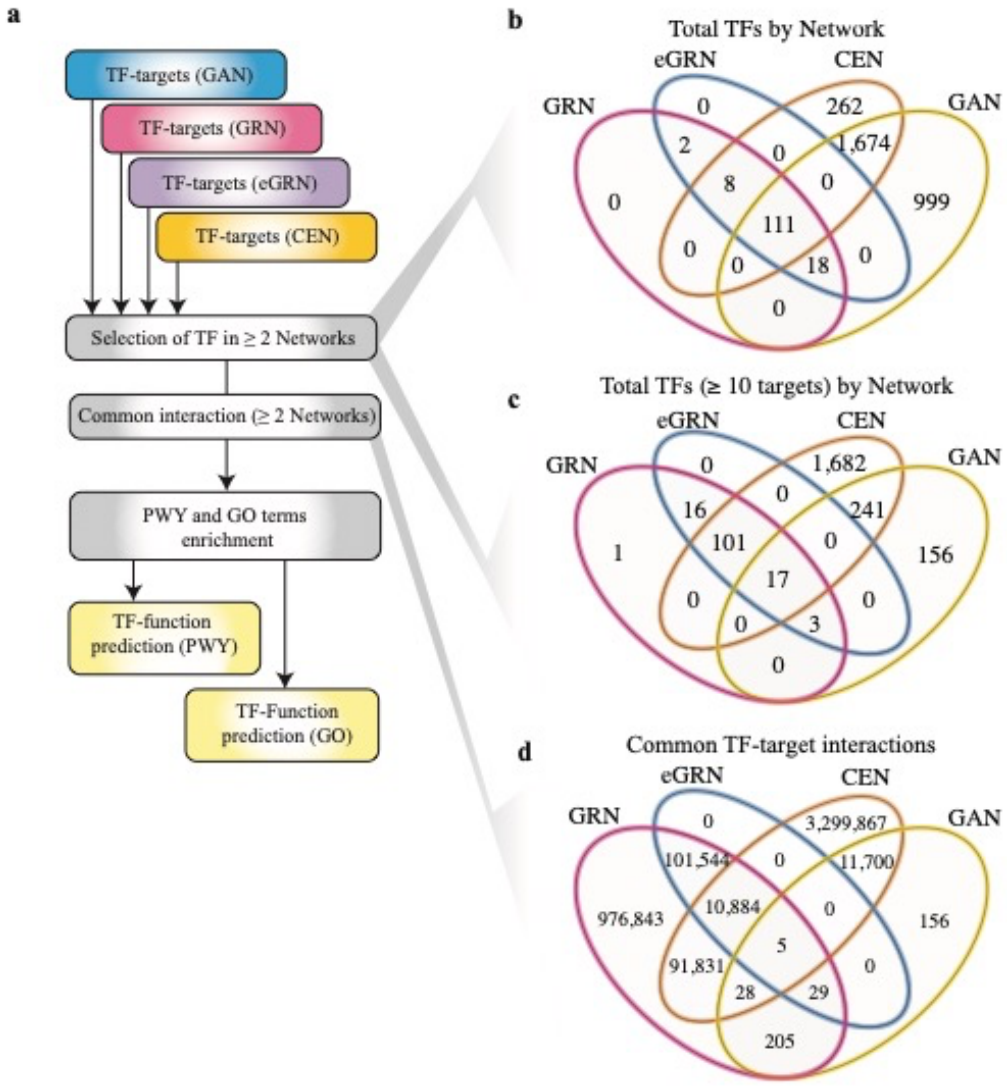

Figure S4.4 Strategy to annotate TFs based on common targets
a. Schema of pipeline used to annotate TFs based on common target genes amount layer (GAN, GRN, eGRN, and CEN). **b** and **c**. Venn diagram indicating the number of common TFs with at least one and ten target genes. **d**. Venn diagram indicating total common interactions (TF-target gene) among layers.

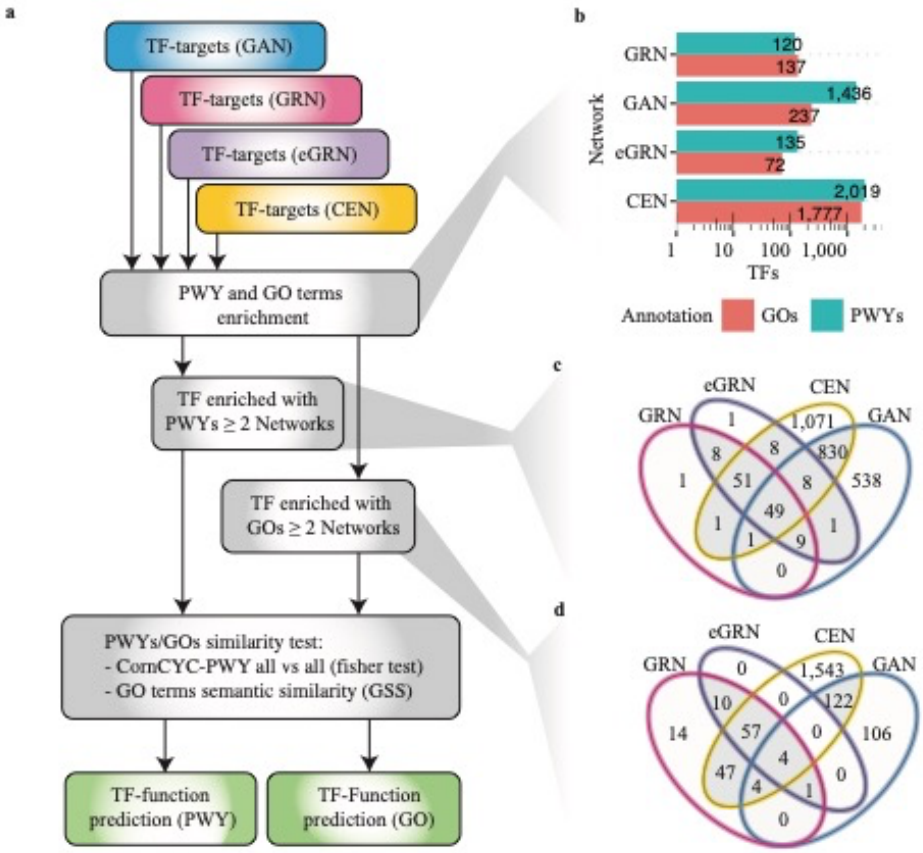

Figure S4.5 Strategy to annotate TFs based on common functions

a. Schema of pipeline used to annotate TFs based on *common function* amount layer (GAN, GRN, eGRN, and CEN). **b**. Bar plot indicating total TFs annotated by layer and by type of function. **c** and **d.** Venn diagram indicating the number of TFs with at least a PWY (**c**) and GO term (**d**) commonly enriched among the corresponding layers.

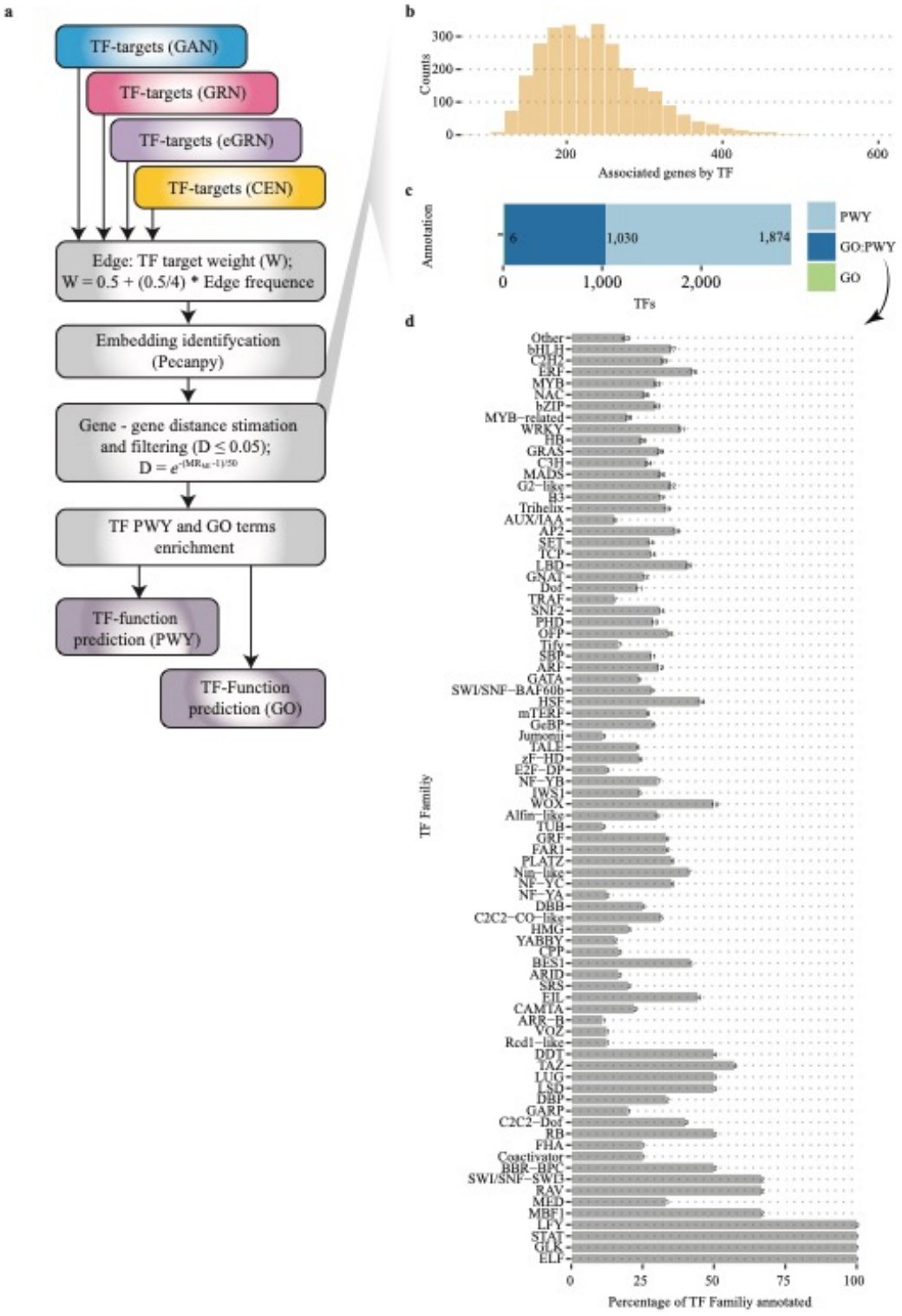

Figure S4.6 Network-based strategy to annotate TFs

a. Schema of pipeline used on the integration of layers to identify TF with similar topological properties, defined here as network-based TF annotation. **b**. Histogram plot indicating the distribution of genes associated per TF. **c**. Stacked bar plot with total TFs annotated by enrichment

Figure S4.6 (cont'd)
with PWYs and GO terms. **d**. Bar plot indicating the percentage of TFs annotated for the 82 TF families (and co-regulator) with at least a TF annotated (**c**).

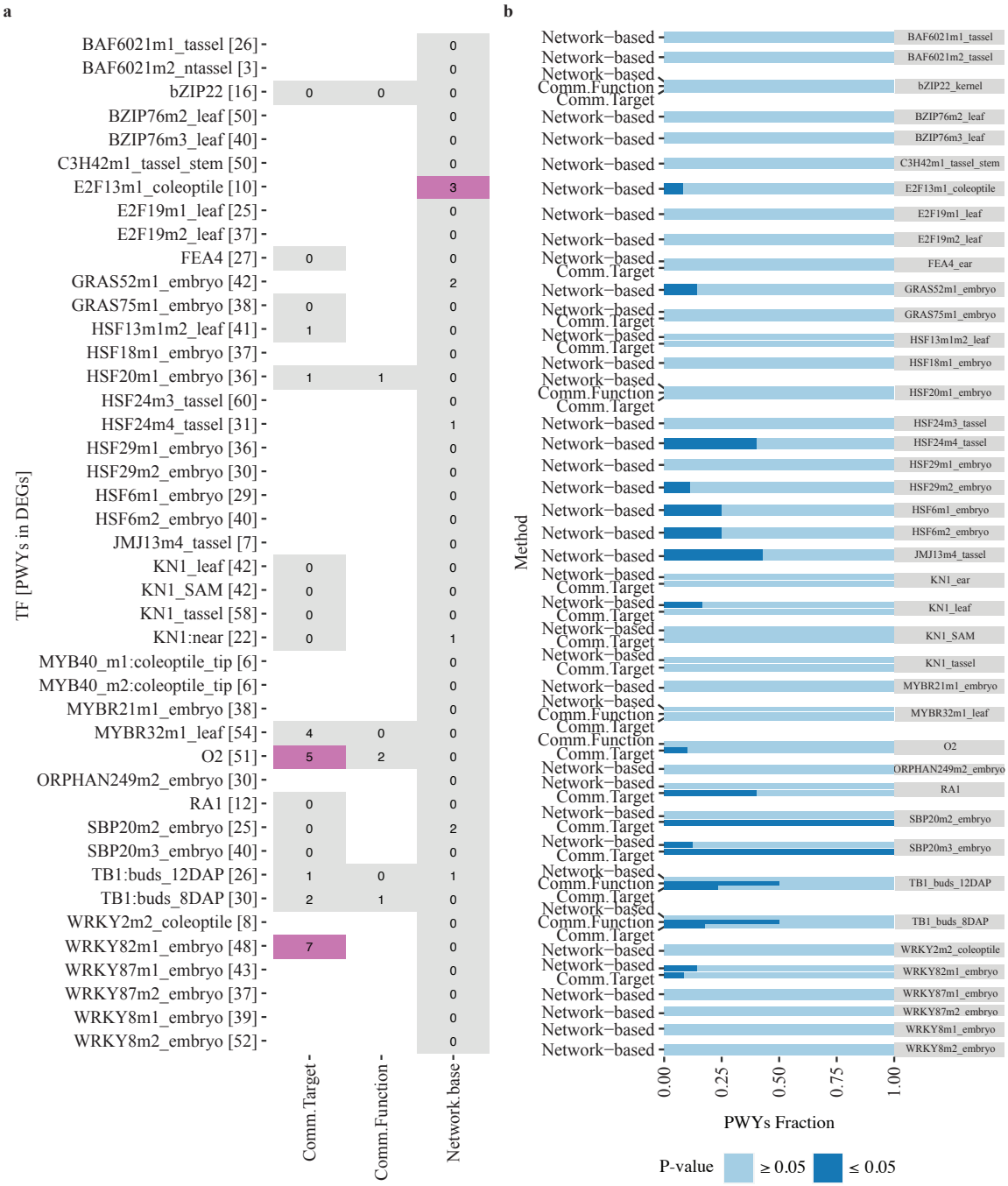

Figure S4.7 Predicted PWY overlapped poorly with PWY observed in knockouts assays

a. Heatmap shows the count of overlapped PWYs between predicted and observed PWY in knockouts per method. The Violet box signifies significant overlap (P-value 0.05, Fisher test). An empty box (white) denotes no predicted PWYs for the corresponding TF. Square braking indicates the number of PWYs significantly enriched in the corresponding knockout (P-value 0.05, Fisher test). **b**. Stacked bar plot indicating the fraction of predicted PWY significantly enriched on DEGs per knockout assay and method.

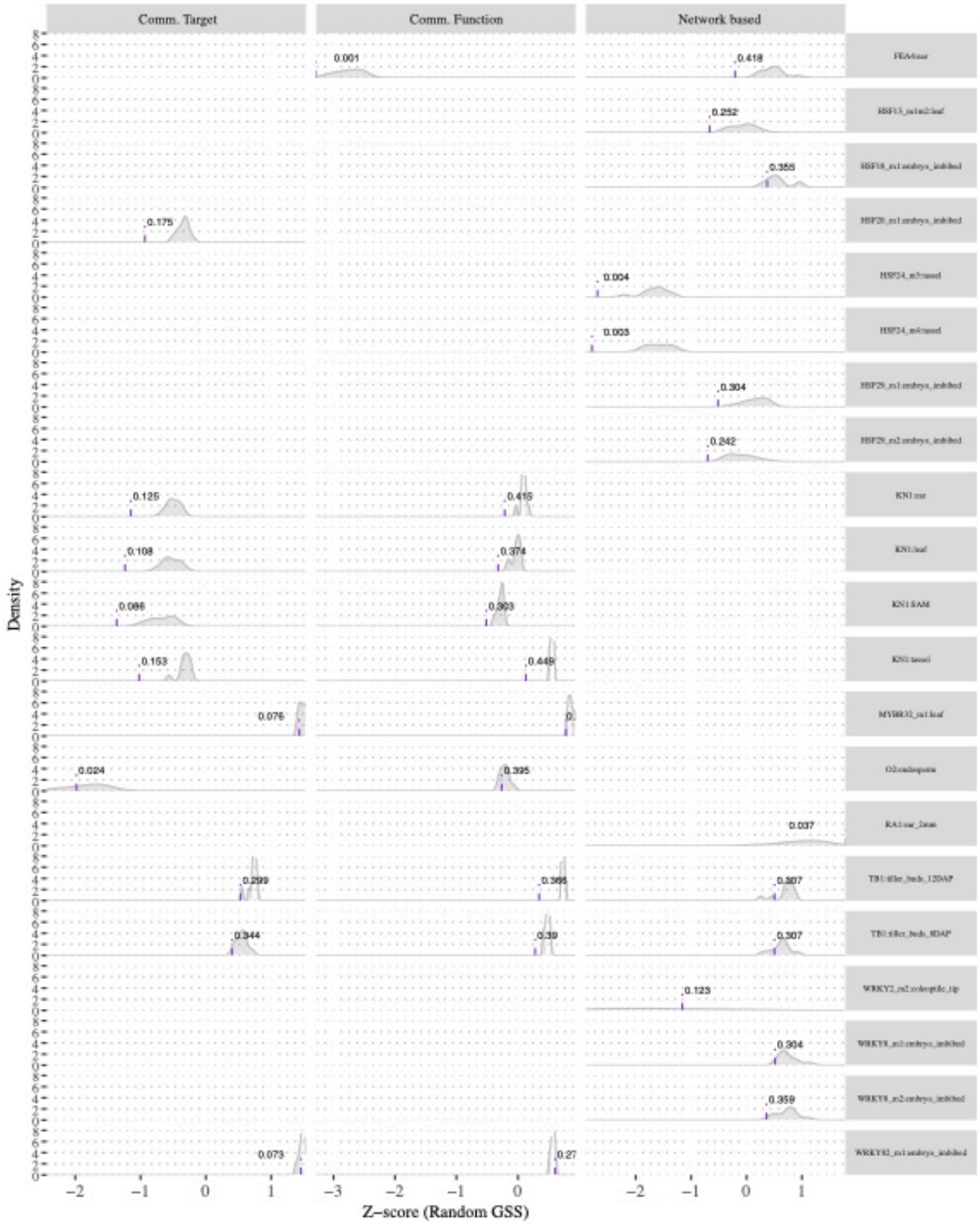

Figure S4.8 GO semantic similarities observed between predicted GO terms and enriched GO terms in knockout are not occurring by chance

Figure S4.8 (cont'd)
Density plot displaying the distribution of random GO term semantic similarity. The observed value on real GO term enrichment, along with its corresponding P-value concerning the random distribution, is highlighted by the horizontal purple line.

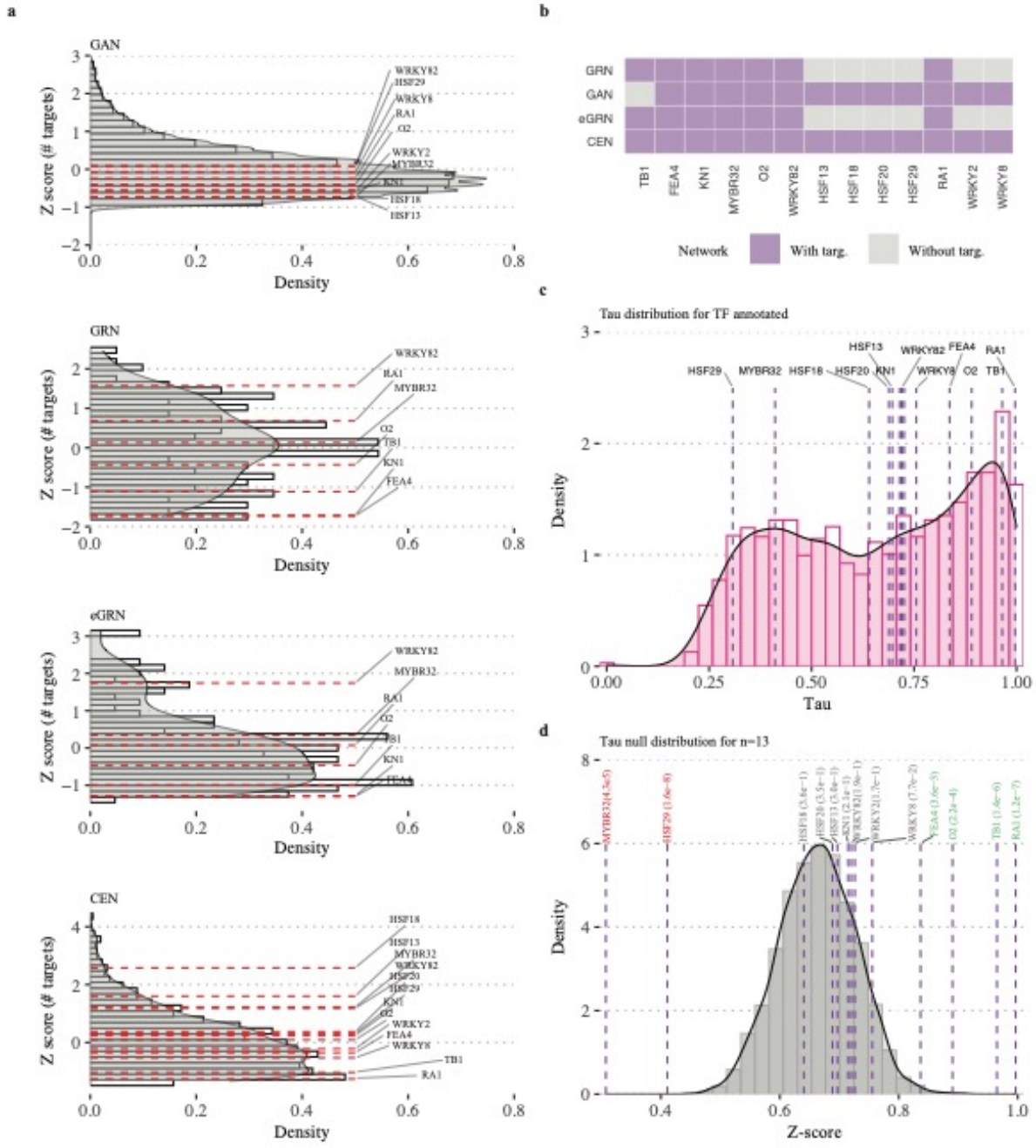

Figure S4.9 Target genes and expression distribution of TFs compared with knockout results

a. Histogram and density plot display the scaled (Z-score) number of target genes for each layer. TFs utilized in the knockout analysis are indicated by dotted red lines. **b**. Heatmap shows the presence or absence of target genes in each of the four layers for every TF analyzed in the knockout assays. **c**. Histogram and density plot of Tau index distribution for TF 2,910 TFs annotated with

Figure S4.9 (cont'd)
at least PWY/GO term. TFs utilized in the knockout analysis are indicated by dotted purple lines. **d**. Histogram and density plot illustrate the null distribution of Tau after randomly sampling 13 TFs a thousand times. TFs used in the knockout analysis are represented by dotted purple lines. P-values were calculated using the null distribution as a reference.

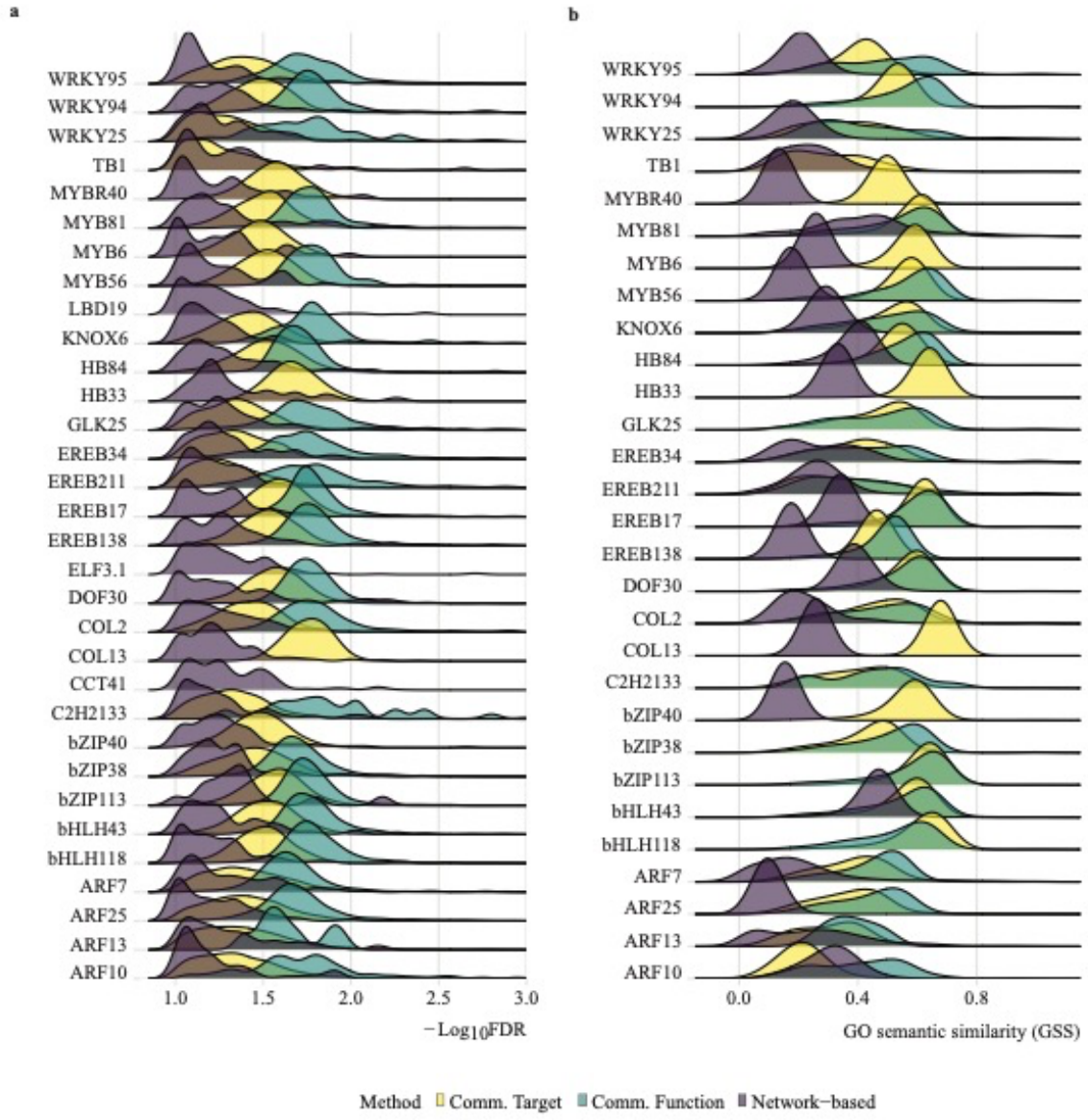

Figure S4.10 GO term significance and similarity distributions from random networks per TF

a and **b**. Density ridges plot showing the average -log10FDR (**a**) and GSS (**b**) distributions in 3,00 random networks for each TF. The GSS values were calculated by comparing each random network with the observed GO terms from the true TF-target interactions.

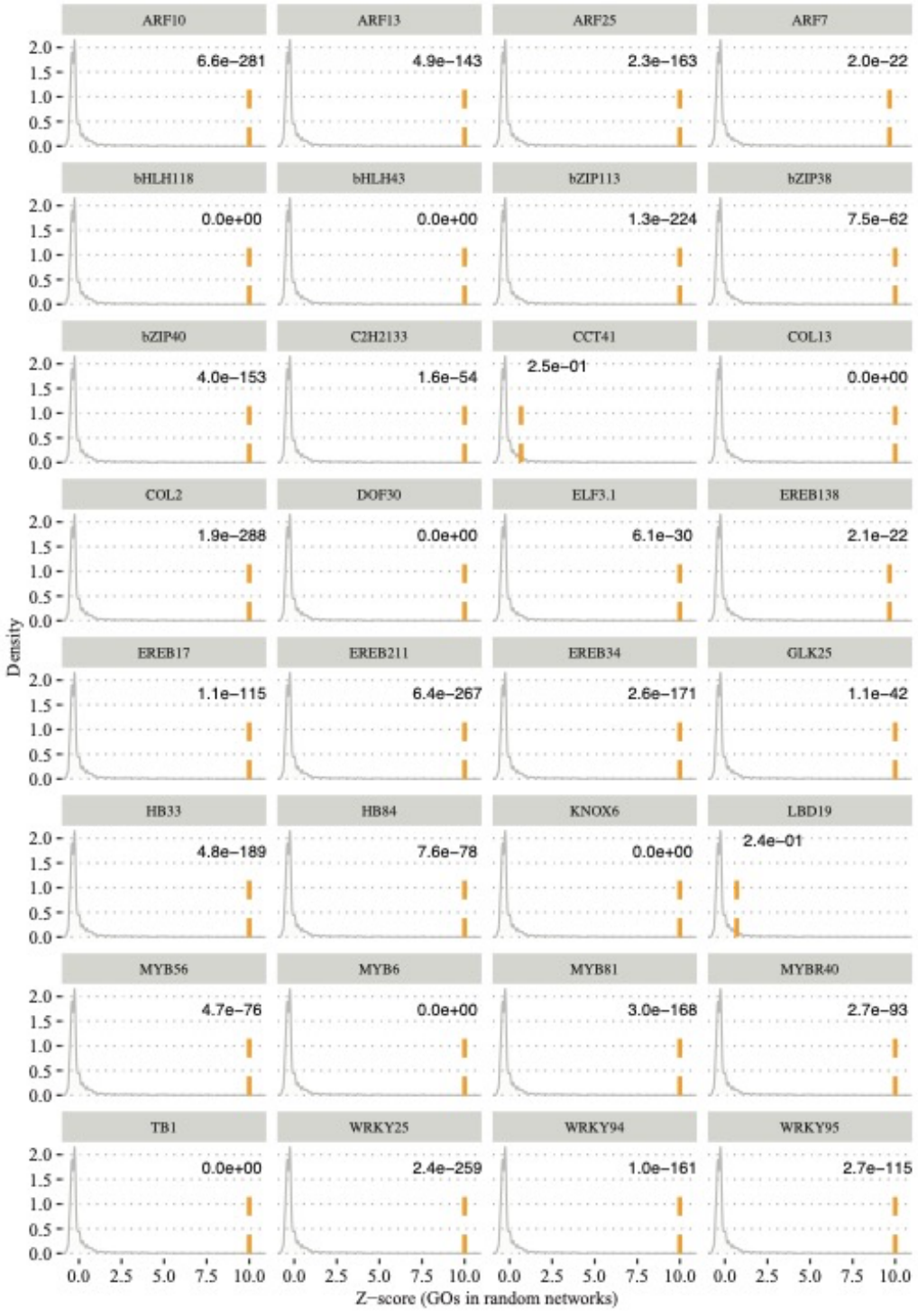

Figure S4.11 Scale count of GO terms in random networks

Density plot displaying the distribution of GO terms in random networks predicted by the *network-based* method for each corresponding TF. The observed number of significantly enriched GO terms for the corresponding TF is indicated by a dotted orange line. The p-value was calculated using the random distribution as the null distribution.

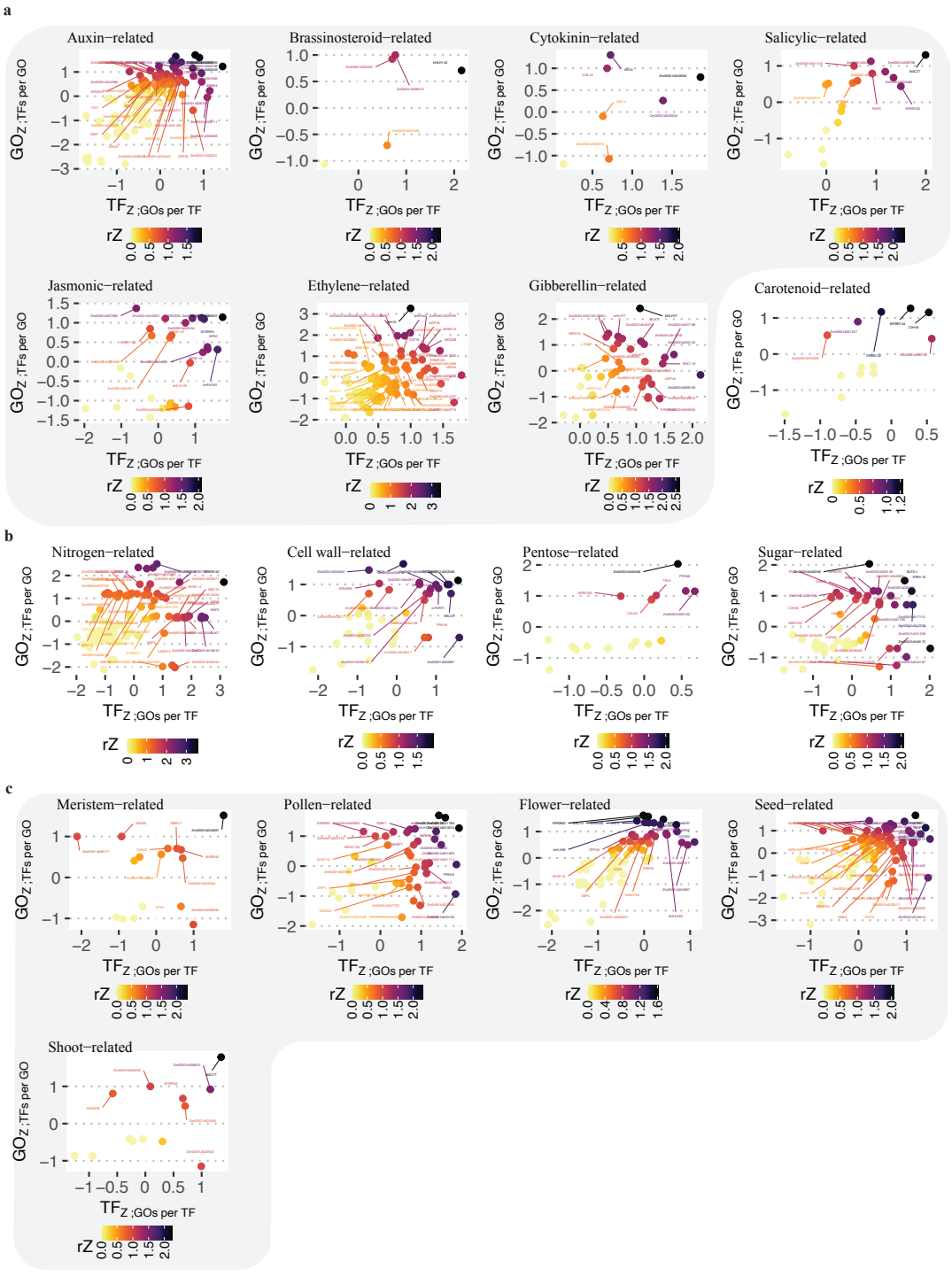

Figure S4.12 Enrichment score for TF and GO term in several biological processes
a, **b**, and **c**. Scatter plot with reciprocal Z score (rZ) of hormones- (**a**), metabolism- (**b**), and development-related (**c**) process.

CHAPTER FIVE: ARABIDOPSIS CO-EXPRESSION SIGNATURES OF COMBINATORIAL GENE REGULATION

5.1 ABSTRACT

Gene co-expression analyses provide a powerful tool to determine gene associations. The interaction of transcription factors (TFs) with their target genes is an essential step in gene regulation, yet to what extent TFs-target gene associations are recovered in co-expression studies remains unclear. Using the wealth of data available for Arabidopsis, I show here that protein-DNA interactions are overall poor indicators of TF-target co-expression, yet the inclusion of TF-TF interaction information significantly enhances co-expression signals. These results highlight the impact of combinatorial gene control on such gene association networks. I integrated this information to predict higher-order regulatory complexes, which are difficult to identify experimentally. I demonstrate that genes strongly co-expressed with a TF are also enriched in indirect targets. These results have significant implications on the empirical understanding of complex gene regulatory networks and transcription factor function, and the significance of co-expression from the perspective of protein-protein and protein-DNA interactions

5.2 INTRODUCTION

The translation of genotype into phenotype is largely dependent on genes being expressed in the appropriate cell types at the correct time (Swift and Coruzzi, 2017). Such expression is mainly controlled by transcription factors (TFs) recognizing specific *cis*-regulatory regions in the genes that they regulate resulting in protein-DNA interaction (PDI) which together define a gene regulatory network (GRN) (Gupta et al., 2021). PDIs are experimentally identified using combinations of gene- and TF-centered approaches; gene-centered approaches result in the identification of TF regulators for specific genes, while TF-centered approaches permit identifying target genes of a particular TF (Arda and Walhout, 2010; Yang et al., 2017; Mejia-Guerra et al., 2012). Within the most commonly used TF-centered strategies include chromatin-

immunoprecipitation (ChIP) and DNA-affinity purification (DAP) methods, often coupled with high-throughput sequencing (ChIP-Seq and DAP-Seq, respectively) (Park, 2009; O'Malley et al., 2016).

Identification of PDIs is particularly important in the context of the effect that a TF has on the expression of its target genes. Often, however, identified TF targets show no changes in expression when the activity of the corresponding TF is perturbed (Zeller et al., 2006; Morohashi and Grotewold, 2009; Morohashi et al., 2012; Eveland et al., 2014; Liu et al., 2015). While in some instances technical artifacts are responsible, the low overlap between TF targets and differentially expressed genes are more often due to redundancy in the activity of the TF (Gitter et al., 2009; Hu et al., 2007), the timing of the PDI interactions (Para et al., 2014; Swift and Coruzzi, 2017; Brooks et al., 2019), the ability of some master regulators to bind closed chromatin regions (Pajoro et al., 2014; Sayou et al., 2016; Tao et al., 2017; Jin et al., 2021; Lai et al., 2021), and/or regulation of the target gene by the TF in only a fraction of the cells sampled (Nolan et al., 2023). For these reasons, the tethering of a TF to the regulatory region of a gene without a clear contribution to the control of the gene's expression is often considered of limited biological significance (Banks et al., 2016; Jiang and Mortazavi, 2018). Additionally, TF are also known by their combinatorial nature, where a single TF can regulate multiple sets of target genes through interactions with other proteins, defined as combinatorial gene regulation (CGR) (Reményi et al., 2004; Brkljacic and Grotewold, 2017). However, despite CGR being a well-documented phenomenon in plant systems (Reményi et al., 2004; Heyndrickx et al., 2014a; Brkljacic and Grotewold, 2017; Colinas and Goossens, 2018; Lacchini and Goossens, 2020), there is no single study that attempts to predict the contribution of CGR to the low overlap in expression changes observed after perturbation and target genes observed in PDI assays.

In general, it is assumed that genes with very similar expression patterns are regulated by similar mechanisms, involving shared TFs (Eisen et al., 1998; Vandepoele et al., 2009; Haynes et al., 2013; Zhou et al., 2020; Geng et al., 2021; Burks et al., 2022). Similar patterns of gene expression can be captured by gene co-expression networks (Eisen et al., 1998; Stuart et al., 2003; Haynes et al., 2013; Wisecaver et al., 2017; Rao and Dixon, 2019; Zhou et al., 2020; Geng et al., 2021; Burks et al., 2022). Multiple examples of implementation of co-expression networks or specific TF-target co-expression patterns have allowed the prioritization of PDIs (Wu and Ji, 2013; Jiang and Mortazavi, 2018; Zhou et al., 2020; Furuya et al., 2021; Geng et al., 2021; Burks et al., 2022; Gomez-Cano et al., 2022). Here, I took advantage of data-rich *Arabidopsis thaliana* (Arabidopsis), which provides an attractive system to investigate the co-expression relationships between TFs and their corresponding predicted target genes, and how the co-expression patterns are affected by the formation of TF-TF complexes. Specifically, I obtained expression and co-expression data from ATTED-II (http://atted.jp/), a database that provides co-expression information obtained from various gene expression analyses (Obayashi et al., 2018). The co-expression data was combined with over five million PDIs identified through ChIP-chip, ChIP-seq, and DAP-Seq. All of these PDIs are accessible via AGRIS (http://agris-knowledgebase.org/) (Palaniswamy et al., 2006; Yilmaz et al., 2011). Additionally, I included 9,503 experimentally established PPI for Arabidopsis TFs that can be accessed through the BioGRID database (Oughtred et al., 2019). Combining the expression and co-expression from ATTED-II, I determined that about half of the TFs are globally co-expressed with their targets as a set, with this number increasing to 85% when local co-expression patterns are considered. I show that a small fraction (in average ~5%) of the direct targets are robustly co-expressed with the corresponding TFs. However, when TF complexes deduced from available PPI data are considered, the number of targets co-expressed

with a TF significantly increases. By integrating PDIs, PPIs, and co-expression information, I predicted the formation of ternary TF complexes, some with strong support from experimental data. Finally, I determined the TFs most highly co-expressed are largely represented by direct and indirect TF targets. These findings have significant implications on the empirical understanding of complex gene regulatory networks, and the meaning of co-expression from the standpoint of PPIs and PDIs.

5.3 RESULTS

5.3.1 Transcription factors and their targets show varying levels of co-expression

To investigate the co-expression of Arabidopsis TFs and their corresponding target genes, I collected existing PDI data involving 555 TFs and 25,255 target genes (see *Methods*). The target genes were determined based on the proximity, when coordinates of peak were available, between the peak of the respective TF and the target genes. It is worth noting that the majority of PDIs used were derived from DAP-seq, which, due to the absence of chromatin context, may contain a higher proportion of non-functional TF-target associations (O'Malley et al., 2016). With these datasets, I built a PDI network that included 2,271,066 interactions that were then used to interrogate the co-expression relationships between each TF and its targets, using the mutual rank (MR) of the PCC (MR-PCC), as reported by ATTED-II (Obayashi et al., 2018), and the mutual rank of the mutual information (MR-MI) (See *Methods*). I used PCC and MI capturing linear and non-linear relationships, respectively (Banf and Rhee, 2017), and the corresponding MR value in order to reduce dataset-dependent associations and to improve the predictive power of the correlation (Obayashi and Kinoshita, 2009; Obayashi et al., 2018).

To assess the significance of co-expression between each TF and its corresponding set of target genes, I conducted two distinct analyses for each TF: (1) I compared the average MR of a TF with

its targets to the average MR of the TF with a randomly selected gene set of similar size. TFs that exhibited significant differences compared to the random set were classified as 'co-expressed by average MR' (see *Methods*). (2) I examined differences in the distributions of MRs between a TF and its target genes versus all non-target genes. TF-target pairs that demonstrated significant differences ($P < 0.05$, Kolmogorov-Smirnov test) compared to the distribution of TF-non-target pairs were categorized as 'co-expressed by MR distribution' (see *Methods*). It should be noted that the analyses based on MR-PCC values were performed separately for negative and positive correlation values. Hence, based on the results of the statistical tests, I determined that 231/555 TFs (using MR-PCC) and 172/555 TFs (using MR-MI) showed significant co-expression with their respective target genes (Figure S5.1a, b). Additionally, by comparing both co-expression metrics (MR-PCC and MR-MI), I identified 124 TFs that were common to both analyses (Figure 5.1a). In total, I identified 279 (172 + 231 - 124) TFs that exhibited significant co-expression with their corresponding target gene sets, while the remaining 276 TFs did not show significant co-expression. A closer look into only the MR-PCC results allowed us to establish that 186/231 TFs showed significant co-expression (either by MR distribution and/or MR average tests) only with positively co-expressed targets (potential transcriptional activators), and 23/231 only with negatively co-expressed targets (potential transcriptional repressors) (Figure S5.1c). Remarkably, 22 TFs showed significant co-expression with different sets of both positively and negatively associated target genes, indicating that they can function both as transcriptional activators or repressors, depending on the target gene subset (Figure S5.1c).

To further characterize the TF-target genes co-expression profiles observed, I classified the TFs into four co-expression categories: TFs co-expressed with their targets based on MR-PCC (107 TFs), TFs co-expressed based on both MR-PCC and MR-MI (124 TFs), TFs co-expressed

based on MR-MI alone (48 TFs), and TFs that did not display significant co-expression with their corresponding targets (276 TFs) (Figure 5.1a). Next, I grouped the MR distribution into bins, ranging from the smallest to the largest rank, to analyze the proportion of targeted genes in each bin (~250 MR values per bin) per TF. Consequently, smaller and larger MR-PCC values correspond to more positive and negative co-expression values, respectively. In the MR-PCC distribution, TFs that displayed significant co-expression with their targets were predominantly distributed within the first 25 bins (i.e. within around the first 6,250 genes most co-expressed per TF) (Figure 5.1b). Conversely, TFs that did not show significant co-expression with their respective targets demonstrated a distinct pattern in the MR-PCC distribution (Figure 5.1b, gray panel). I observed similar patterns in the MR-MI distribution as well (Figure S5.2). Notably, MI does not differentiate between positive and negative associations. Thus, all significant values, when present, are captured in the left tail of the distribution. Additionally, there was a consistent ~1% presence of targets across all bins in the distribution (Figure 5.1b, indicated by line plot with target % beneath each heatmap). These findings validate earlier observations in Arabidopsis (Zaborowski and Walther, 2020), corroborating the absence or low co-expression relationship between TFs and their respective target genes.

Given that many TF functions are often highly cell-type, tissue, or stress specific, I analyzed the co-expression at different scales (Zhou et al., 2020; Lee et al., 2023; Nolan et al., 2023). Specifically, I introduced a new category called "local co-expression," which involved analyzing subsets of expression datasets obtained after clustering similar samples. These subsets served as proxies for organ- and condition-specific co-expression (see *Methods*). In total, I identified twelve distinct sample clusters representing potential conditions (Figure S5.3). Similar to the previous global co-expression analysis, I employed two statistical methods (average MR and MR

distribution) and two metrics (MR-PCC and MR-MI). To explore the presence of local co-expression patterns in the 276 TFs that did not exhibit significant global co-expression with their target genes, I kept these sets separate. Overall, I discovered that 199 out of 276 TFs displayed significant co-expression with their target genes in at least one of the clusters (Figure 5.1c). As expected, TFs with global co-expression patterns were found to exhibit co-expression with targets in multiple local clusters (Figure 5.1c), with the exception of seven TFs (WIP5, MYB1, PLT1, ERF109, HHO5, NAC4, and AT5G47660). These seven TFs showed significant global co-expression, but no evident local co-expression in any of the clusters. The reason for this intriguing behavior is not yet clear.

I explored the distinguishing characteristics of TFs that do not exhibit global or local co-expression with their target genes. I observed a significant difference in the connectivity within the network between TFs showing co-expression and those that do not. TFs lacking co-expression with their alleged targets displayed significantly smaller in-degree (representing the number of TFs binding to a specific promoter region of the corresponding TF) and out-degree (representing the number of target genes bound by a TF) compared to co-expressed TFs ($P < 0.05$, Mann-Whitney U test; Figure S5.4). These findings suggest that TFs with lower connectivity in the network may have distinct co-expression relationships with their targets. However, I cannot dismiss the possibility that the identified clusters may not be sufficiently resolved for these TFs.

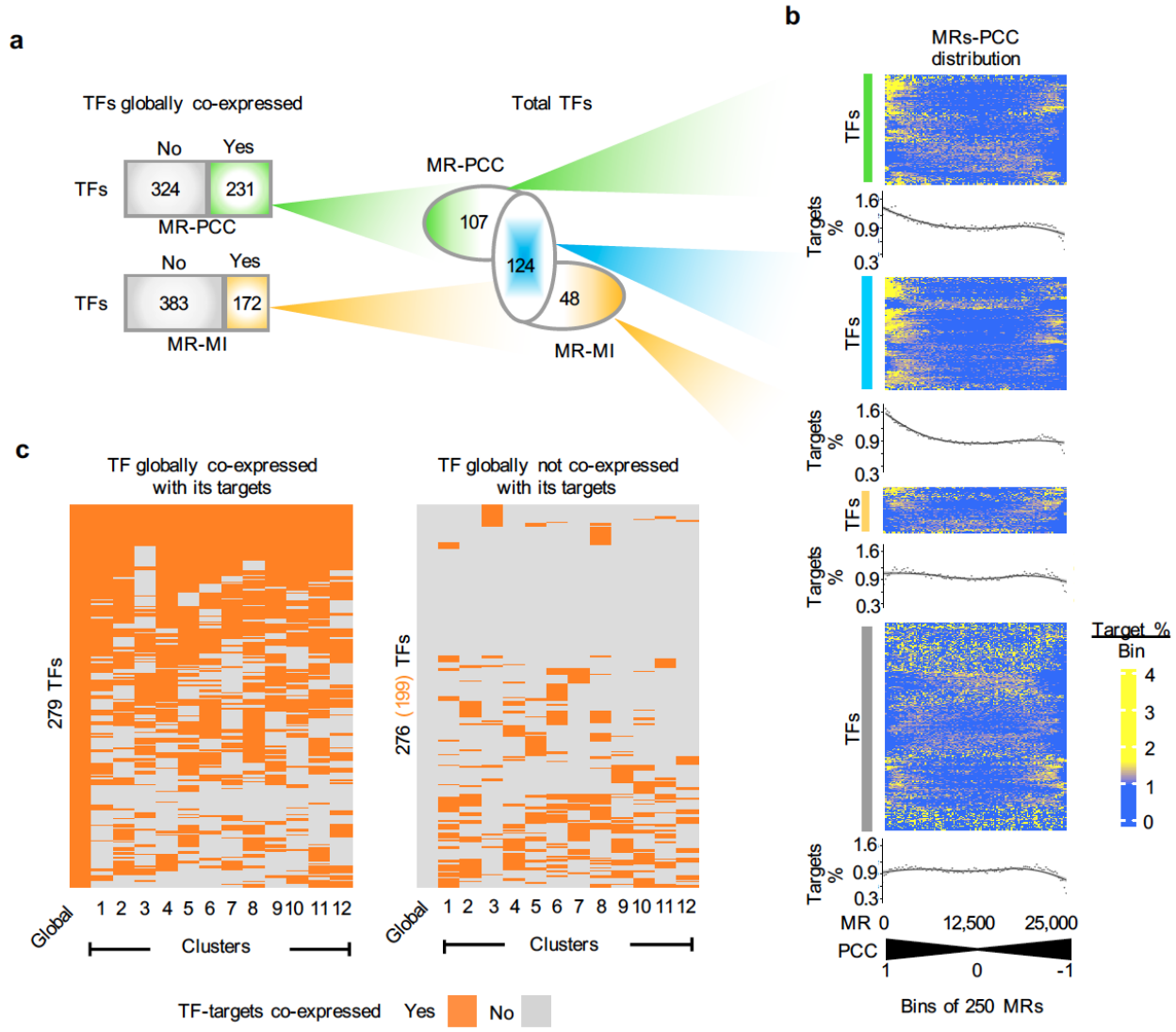

Figure 5.1 Patterns of co-expression between TFs and their direct target genes
a. Total number of TFs globally co-expressed with their corresponding targets across all tissues and conditions based on MR-PCC and MR-MI. The Venn diagrams show the overlap between the two metrics. **b**. Heatmaps displaying the distribution of MR-PCC values across 25,296 Arabidopsis genes. TFs are divided into four co-expression groups: TFs co-expressed with their targets based on MR-PCC (107 TFs), on both MR-PCC and MR-MI (124 TFs), MR-MI only (48 TFs), and TFs that do not show significant co-expression with their targets (276 TFs). The colors indicate the percentage of TF targets within each bin of 250 MRs. There are 101 bins along the PCC distribution, representing the co-expression values of each TF with the 25,296 Arabidopsis genes. Small MR values correspond to positive PCC values, while large MR values represent negative PCC values. The line-dot plots below each heatmap display the average percentage of targets for all TFs in each bin. **c**. Heatmap illustrates the local co-expression profiles of each TF analyzed across 12 different expression clusters. The color indicates whether there is co-expression (orange) or no co-expression (gray). The left panel shows TFs that are globally co-expressed with their targets, while the right panel shows those that are not. The number in brackets represents the count of TFs with significant co-expression in at least one of the local clusters.

5.3.2 Few targets are highly co-expressed with their respective TFs

The distribution of target genes along the MR-PCC range mentioned earlier (Figure 5.1b) reveals a limited presence of targets among the genes exhibiting the highest co-expression with each TF. Specifically, the maximum proportion of targets within a bin containing 250 co-expressed genes is approximately 5% (Figure 5.1b). Moreover, the percentage of targets gradually decreases beyond the first 5,000 MRs, capturing a maximum of 25% of the total identified direct targets for each TF. To assess the proportion of highly co-expressed targets (HCT) for each TF, I defined the top and bottom 2.5% of the MR-PCC distribution as the set of highly co-expressed genes (HCGs) and tallied the total number of targets within these intervals. Among all TFs, ARABIDOPSIS PSEUDO-RESPONSE REGULATOR 9 (PRR9) exhibited the highest percentage (36%) of target genes identified as HCTs according to the defined criteria. However, on average, only 4.7% of the targets qualified as HCTs (Figure 5.2a), indicating that, on average, the remaining 95.3% of the targets were classified as low co-expressed targets (LCTs).

5.3.3 PPIs condition TF co-expression with direct targets

To gain insights into the limited co-expression between TFs and their target genes, I explored how the presence of multiple physically interacting TFs regulating a gene could influence the observed co-expression pattern. I obtained 815 experimentally determined protein-protein interactions (PPIs) involving 313 out of the 555 TFs analyzed in this study from BioGRID. Specifically, using this PPI information, I assessed the extent to which the formation of TF complexes (e.g., TFx-TFz) could account for the high fraction of low co-expressed targets (LCTs) associated with each TFx. To do this, I calculated the partial co-expression correlation of TFx with all LCTs, conditioned on the presence of TFz (de la Fuente et al., 2004; Kim, 2015; Uygun et al., 2016). This analysis allowed me to examine the co-expression of TFx target genes with a TFx

complex (TxCC). It is important to note that these correlations are not symmetric, meaning that TxCC may differ from TzCC. Additionally, TCC refers specifically to correlations conditioned by already reported TF heterodimers. I performed the correlation analysis using all Arabidopsis genes and identified the top 2.5% highly co-expressed genes (at each tail of the correlation distribution as cut-off) for each TFx-TFz complex. I found that, on average, 5% of the LCTs of a TF are co-expressed with the complexes in which the TF is involved (i.e., TxCC) (Figure 5.2b). Furthermore, I calculated the percentage of TxCC based on the number of interactions, revealing that the average of TxCC is not influenced by the total number of targets associated with the respective TF (Spearman Correlation, rs = 0.02) (Figure 5.2c, color scale distribution). However, when considering all interactions for each TF, it became evident that the percentage of targets co-expressed with a complex increased proportionally with the number of known interactors that a TF possesses (correlation, r_s = 0.69) (Figure 5.2c), indicating that a significant proportion of the LCTs described previously can be explained by considering complexes of interacting TFs.

Even among TFs with a similar number of analyzed complexes, there is notable variation in the proportion of TxCC (Figure 5.2d). For instance, within the subset of TFs that have a single known partner, I observed distinct cases represented by DEHYDRATION RESPONSE ELEMENT-BINDING PROTEIN 26 (DREB26) and ethylene response factor (ERF) (AT4G18450). These TFs interact with BASIC HELIX LOOP HELIX PROTEIN 10 (BHLH010) and GT-1, respectively, and the corresponding complexes explain 5.5% and 1.6% of the LCTs (Figure 5.2d). This finding highlights the specific and unique impact of each TF complex on the percentage of co-expressed target genes, potentially reflecting functional aspects of combinatorial gene regulation.

Thus far, I have demonstrated that incorporating regulatory complexes can enhance the co-expression of TFs with their targets. Despite the variable number of common targets shared by these interacting TFs (TFx-TFz in Figure 5.2e), only a small fraction of these shared targets exhibit co-expression with the complex (TxCC-Tz, Figure 5.2e). Therefore, to gain a deeper understanding of the co-expression patterns among the shared targets of TFx and TFz, I compared the proportion of these targets that co-expressed with the TFx-TFz complex and also exhibited high co-expression with TFz (Figure 5.2f, blue box), with the TFz-TFx complex (TzCC) (Figure 5.2f, orange box), or show low co-expression with TFz (Figure 5.2f, gray box). Overall, 91% of the shared targets that are also TxCC were found to have modest co-expression with TFz (LCTz, gray in Figure 5.2g). Only 3.9% of the shared targets exhibited high co-expression with TFz (Figure 5.2g, blue box), and 4.1% co-expressed with both complexes (TxCC and TzCC) (orange in Figure 5.2g). These findings emphasize the significance of considering TF complexes when interpreting the co-expression between TFs and their targets.

To assess the biological significance of the co-expression observed between targets and TF complexes, I examined specific examples. HHO2 (HRS1 HOMOLOG2) and HHO3 (HRS1 HOMOLOG3) are MYB-related TFs involved in phosphate homeostasis, lateral root development (Nagarajan et al., 2016), and nitrogen responses (Varala et al., 2018). Our analysis revealed that the HHO2-HHO3 complex co-expressed with 43 targets. Notably, HHO2, HHO3, and six of their targets exhibited differential expression in response to different nitrogen growth conditions (Figure 5.2h), supporting the functional relevance of complex formation and its associated targets.

I also examined the SVP (SHORT VEGETATIVE PHASE) - GBF2 (G-BOX BINDING FACTOR 2) complex. SVP acts as a flowering repressor (Chen et al., 2018) and is also involved in drought responses (Bechtold et al., 2015), while GBF2 is associated with abscisic acid (ABA)

responses (Song et al., 2016). My results identified 429 shared co-expressed targets for the SVP-GBF2 complex (Figure 5.2i), of which 130 genes were differentially expressed under drought conditions (Harb et al., 2010; Wilkins et al., 2010; Bechtold et al., 2015). These findings support the notion that TF targets, which lack significant co-expression with the TFs individually, do exhibit co-expression when considering TF complexes.

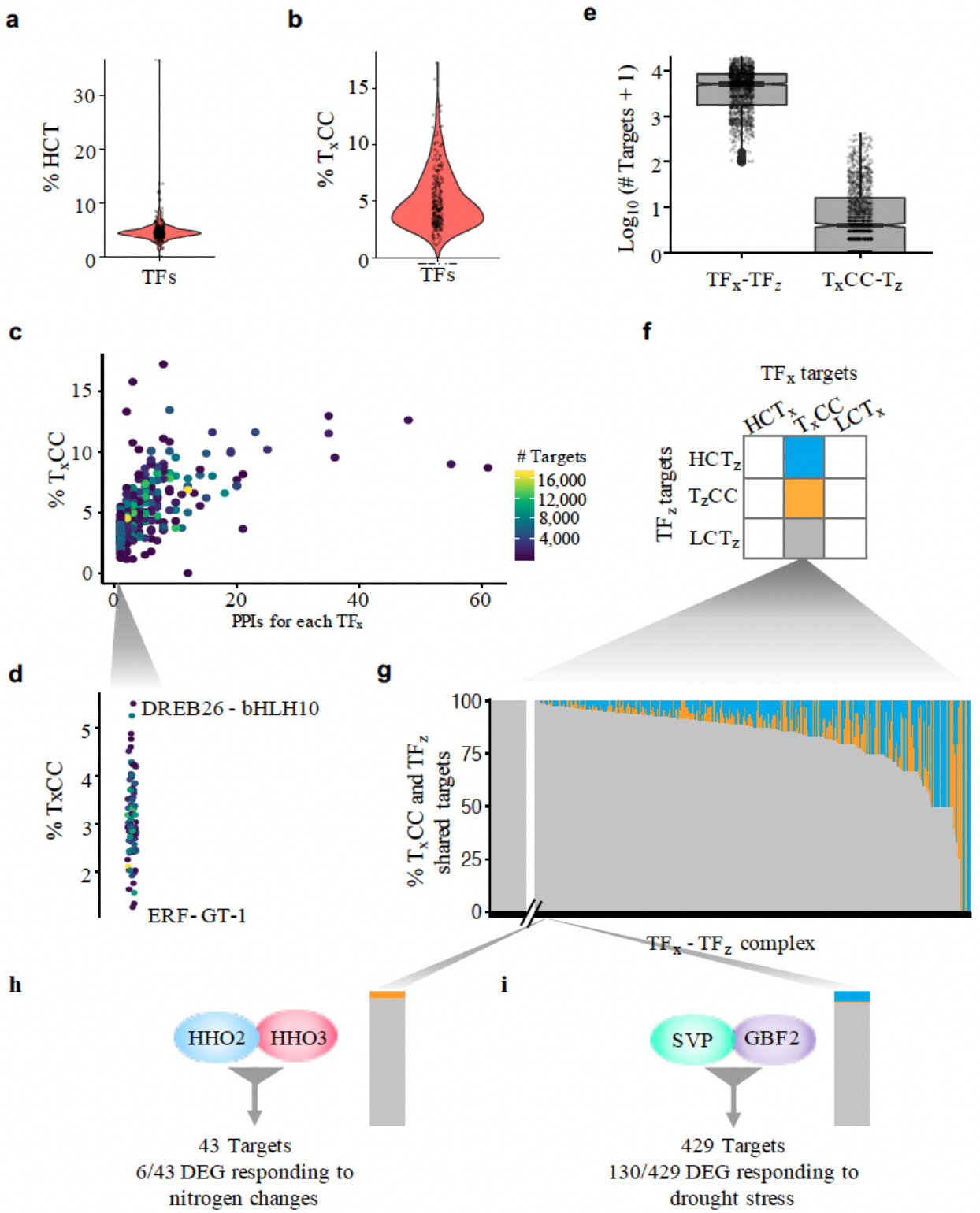

Figure 5.2 Targets are more frequently co-expressed with TF complexes than with individual TFs

a. Violin plot displaying the proportion of highly co-expressed targets (HCT) for 313 TFs. **b**. Boxplot illustrating the percentage of low co-expressed targets (LCTs) that coincide with targets

Figure 5.2 (cont'd)
co-expressed with a TFx complex (TxCC). **c**. Percentage of TxCCs in relation to the total number of PPIs involving each TF. **d**. Enlarged view of the section in (**c**) depicting TFs with only one interacting partner. DREB26-bHLH10 and ERF (At4g18450)-GT-1 represent extreme cases in the distribution. The color scale in (**c**) and (**d**) indicates the number of targets for each TF. **e**. Boxplot presents the number of shared targets between the 815 analyzed TF complexes (TFx-TFz) or the number of targets of a given TFx co-expressed with the TFx-TFz complex (TxCC) that are also targets of TFz. **f**. Schematic representation of the comparison made among target genes of TFz and targets of TFx categorized as HCTs, TCCs, or LCTs of TFx, denoted by blue, orange, and yellow, respectively. **g**. Distribution of targets based on the comparison in (**f**) for the 815 analyzed TFx-TFz complexes. Complexes are shown on the x-axis, while the y-axis represents the frequency of overlap. The HHO2-HHO3 (**h**) and SVP-GBF2 (**i**) TF complexes serve as representative examples from the analyzed TF complexes. **h**. The numbers indicate the differentially expressed genes (DEGs) under various nitrogen growth conditions. **i**. The numbers indicate DEGs, also identified as targets of the corresponding complexes, under drought stress in three different studies. The sidebar plot provides a zoomed-in view of the HHO2-HHO3 and SVP-GBF2 positions on the shared target distribution shown in **g**.

5.3.4 Co-expressed targets shared by binary TF complexes suggest higher-order arrangements

The results presented so far indicate that the integration of co-expression and physical interaction information contributes to the identification of TFs that control gene expression working as part of complexes. There are many instances in which Arabidopsis TF pairs interact and control shared sets of target genes (Brkljacic and Grotewold, 2017; Bemer et al., 2017). However, the experimental identification of higher-order (beyond binary) TF complexes is not without challenges (Lambert et al., 2018). To investigate whether the combination of co-expression, PPI, and PDI information might provide insights on higher-order TF complexes, I started by describing the complexes made up by TGA10 (TGACG MOTIF-BINDING PROTEIN 10), TCP14 (TGA10 with TEOSINTE BRANCHED, cycloidea and PCF 14), and a homeodomain-like TF (AT2G40260) (Trigg et al., 2017). The TGA10-TCP14 and TGA10-AT2G40260 complexes share 80% of targets co-expressed with each complex (Figure 5.3a, black nodes). Moreover, shared targets had similar expression correlation with both heterodimers (either positive

or negative), indicating that both complexes potentially activate or repress the same sets of genes (Figure 5.3a). These results, combined with the information that TCP14 and AT2G40260 physically interact with each other (Trigg et al., 2017), provide strong evidence that TGA10, TCP14, and AT2G40260 form a ternary complex that controls the expression of all targets indicated in Figure 5.3a.

I proceeded to examine the presence of other triple-binary (tri-bi) TF combinations in Arabidopsis, similar to the TGA10-TCP14 and TGA10-AT2G40260 complexes. To do this, I initially identified 47 TFs that had at least two interacting partners and PDI information. I then determined the percentage of shared target genes between these pairs (Figure 5.3b, orange) and compared it to the percentage of targets unique to each pair (Figure 5.3b, gray). In certain cases, all targets were shared by both binary complexes (indicated by the orange columns in Figure 5.3c), while only around 8% were shared by binary complexes with minimal overlap (columns on the right in Figure 5.3c). Notably, 13 out of the 47 tri-bi combinations tested showed experimental evidence for all three binary interactions (indicated by black arrows in Figure 5.3c), supporting the existence of higher-order (ternary) complexes. However, I was unable to establish a statistically significant correlation between the number of shared targets and experimental evidence confirming the formation of ternary complexes. This lack of correlation likely stems from the limited availability of PPI data for many of the TF pairs involved, rather than the shared percentage of co-expressed targets being an inadequate indicator of ternary complex formation.

I next investigated how frequently TFs involved in tri-bi interactions share common targets. Unlike the previous analysis, I now considered TFs with more than two PPIs. I identified a total of 2,013 true tri-bi instances (i.e., with evidence of physical interaction for all pairs of the tri-bi) involving 140 TFs. In approximately 90% of these instances, the TFs showed a significant overlap

of target genes (false discovery rate < 0.01, Fisher's exact test). This indicates that TFs involved in tri-bi interactions often share a substantial number of targets, making them strong candidates for the formation of tertiary, or even higher-order, complexes. To assess whether the fraction of shared targets differs from random tri-bi complexes, I compared the co-expressed shared targets of TF complexes from experimentally demonstrated tri-bi instances to those from tri-bi instances obtained through a randomized binary interactome approach for each TF (see *Methods*).

Among the 104 TFs analyzed, I identified 12 TFs involved in tri-bi instances with a significantly larger fraction of shared targets compared to the background model (Figure S5.5a). An illustrative case is ABI5 (ABA INSENSITIVE5), which participates in eight tri-bi instances and exhibits a median shared fraction of targets of 0.77 (Figure 5.3d). Remarkably, six out of the eight tri-bi instances involving ABI5 consist of a combination of four TFs from the ABF (ABSCISIC ACID RESPONSIVE ELEMENTS-BINDING PROTEIN) family (Figure 5.3e). The number of target genes varies across the tri-bi instances, ranging from 258 for ABF2-ABI5-ABF4 to 290 for ABF3-ABI5-ABF4 (Figure 5.3e). The 290 ABF3-ABI5-ABF4 gene targets include 46 genes differentially expressed in *abi5* mutant seeds (Bi et al., 2017). Remarkably, ABF2, ABI5, and ABF4 also interact with SnRK2.2 (SNF1-RELATED PROTEIN KINASE 2), PP2CA (PROTEIN PHOSPHATASE 2CA) (Yoshida et al., 2010; Lynch et al., 2012), and AHG1 (ABA-HYPERSENSITIVE GERMINATION 1) (Lynch et al., 2012), which are key known post-translational regulators of ABI5 (Skubacz et al., 2016). I found 41 TFs involved in tri-bi instances with a significantly reduced fraction of shared targets compared to the expected background model (Figure S5.5b). These findings suggest that these TFs may participate at least in dimeric complexes where they bind overlapping sets of target genes.

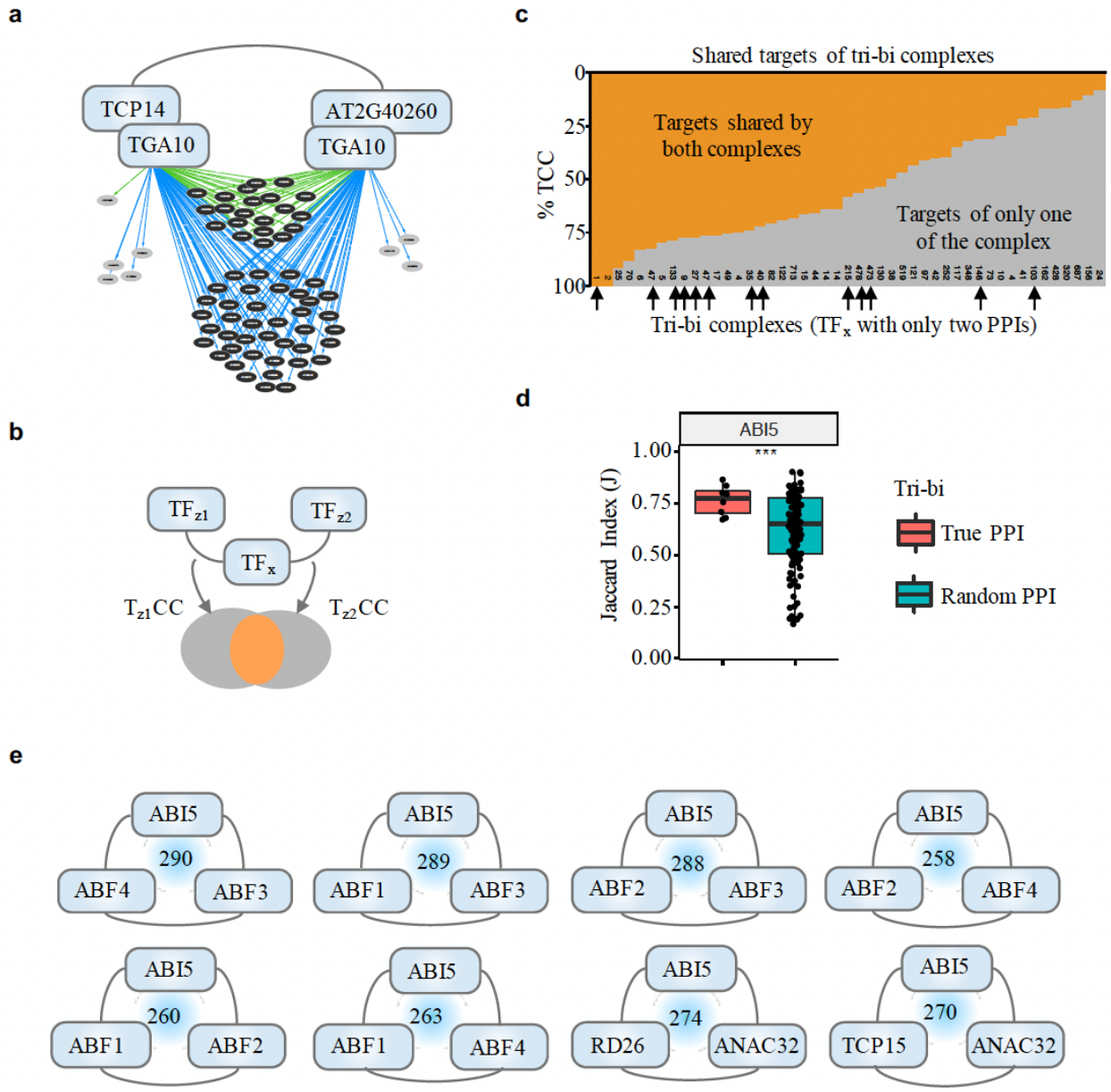

Figure 5.3 Common co-expressed targets of TF complexes suggest higher-order TF arrangements

a. Co-expressed targets shared by the TGA10-TCP14 and TGA10-AT2G40260 TF complexes are represented. Black nodes indicate common targets for both complexes, while light gray nodes represent targets controlled by one complex but not the other. Green arrows indicate positive co-expression correlation (activation), and blue arrows indicate negative co-expression correlation (repression) with the respective TF complexes. **b**. The strategy used to identify shared targets by comparing TxCC between pairs of dimers is illustrated schematically. **c**. The percentage of total targets bound by both complexes (orange) or only by one complex (gray) is shown. Black arrows indicate tri-bi complexes with experimental evidence for all three binary interactions. **d**. ABI5

Figure 5.3 (cont'd)
serves as an example of 12 TFs with significantly larger fractions of shared targets in tri-bi complexes compared to randomly formed tri-bi complexes (two-sided t-test $P < 0.05$). Similarity between the sets of target genes for corresponding dimers was measured using Jaccard indices. **e**. Tri-bi complexes involving ABI5 are depicted, with experimentally verified interactions shown as lines and the numbers in blue indicating targets of the complexes.

5.3.5 Genes highly co-expressed with TFs are enriched in indirect TF targets

In previous sections, I focused on the co-expression patterns between TFs and their direct targets. However, a question that remains unanswered is whether there is a relationship between a TF and the genes that are most highly co-expressed with that TF. To explore this, I examined how many target genes of a TF also belong to the top 5% most highly co-expressed genes (HCG) with that TF. Surprisingly, for the large majority of the TFs (80%), less than 30% of the HCG are among the target genes. There is one exception, NF-BY2 (nuclear factor Y, subunit B2), where this number is as high as 82% (Figure 5a). I explored the possibility that genes that are not direct targets of a TFx could be targets of a TFx partner (TFz), or that they could be targets of a second TF (TFy) that is itself a direct target of TFx.

To assess the impact of TF partners (TFz) on the highly co-expressed genes of TFx, I investigated the proportion of highly co-expressed genes that are targets of TFz but not of TFx itself. Our analysis revealed that out of the 313 tested TFs, 309 TFs had at least one highly co-expressed gene that was a target of one of its TFz partners. On average, approximately 10% of the highly co-expressed genes of a TF belonged to this category (Figure 5.4b). Similarly, to understand the contribution of downstream targets to the highly co-expressed genes of a downstream TFy in the regulatory hierarchy, I examined the same set of 313 TFs. Among these TFs, 306 TFs bound to a TFy that had at least one direct target gene highly co-expressed with the upstream TFx. On average, around 9.8% of the genes most highly co-expressed with TFx were indirect targets of TFy (Figure 5.4c). I also compared the actual set of highly co-expressed genes recovered using true

interactions with those obtained using random networks (PPI and PDI, respectively) (See *Methods*). The random TF PPIs yielded a similar number of highly co-expressed genes compared to the known PPIs ($P > 0.05$, Mann-Whitney U test) (Figure 5.6a). It is worth noting that the PPI network used in this analysis had an average path length of 3.5 edges between all TF nodes, indicating weak independence between the true and random PPIs. In contrast, the random target TFy resulted in a significantly smaller number of highly co-expressed genes compared to the true targets ($P < 0.05$, Mann-Whitney U test) (Figure S5.6b), suggesting that downstream hierarchical regulators play a crucial role in explaining the presence of highly co-expressed genes for the corresponding TF.

I computed the combined contribution of TFz interactors and downstream TFs (TFy) to the set of highly co-expressed genes for each of the 313 TFs. This allowed me to determine that, on average, 90% of the genes most highly co-expressed with a TF consist of its direct targets (~16%), targets of its TFz partners (~4%, after excluding partners that are also direct targets of TFx), and downstream targets (~70%, targets of a TF's target) (Figure 5.4d). Interestingly, I also found examples in which the partner for TFx is also a downstream target, participating in a feed-forward loop (FFL) (26% out of total TFs). FFLs are among the most highly represented regulatory motifs present in Arabidopsis (Chen et al., 2018) and other eukaryotes (Milo et al., 2004).

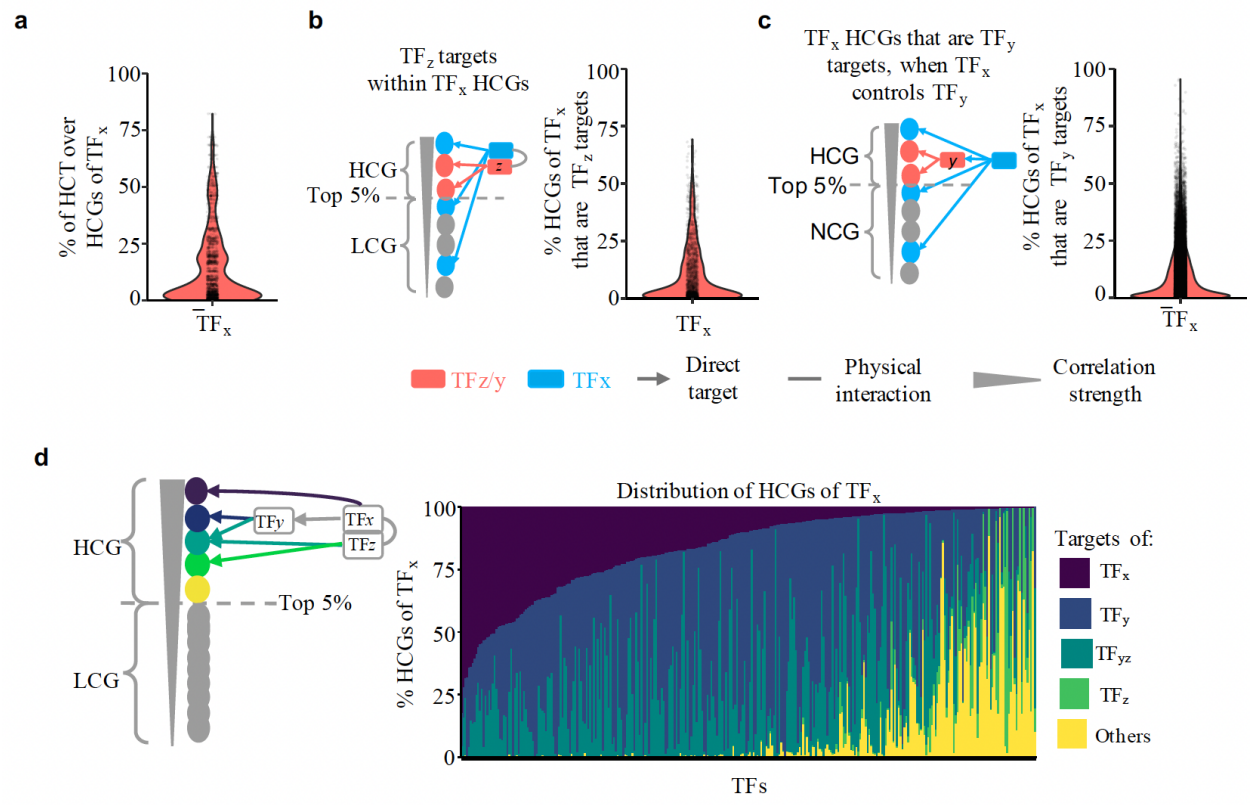

Figure 5.4 Genes highly co-expressed with TFs are enriched in indirect TF targets
a. Percentage of highly co-expressed genes (HCGs) of TFx that are confirmed targets of TFx. **b** and **c**. Model and percentage of highly co-expressed genes that are potential indirect targets of TFx through its TFz interactors (**b**) and a TFy downstream of the corresponding TFx (**c**). **d**. Percentage of HCGs attributed to direct or indirect targeting by TFx.

5.4 DISCUSSION

In this chapter, I examined the co-expression patterns between TFs and their targets using comprehensive PDI, PPI, and gene expression data for Arabidopsis. I found that approximately half (279) of the TFs studied exhibit global co-expression with their targets, while an additional 35% (199) display local-specific co-expression in at least one of the twelve sample clusters identified. Interestingly, for 77 Arabidopsis TFs with extensive PDI information, there is no conclusive evidence of co-expression with their identified targets beyond what would be expected by chance. This suggests that certain TFs only show co-expression under specific conditions, and it is possible that utilizing single-cell sequencing will uncover additional co-expression

relationships that are not apparent in organ-level gene expression experiments due to the complexity of cell populations. I show that only a small fraction (on average 4.7%; Figure 5.2a) of the direct targets are among the genes most highly co-expressed with a given TF. Conversely, direct targets are a small fraction of the genes highly co-expressed with a TF (in average 14.3%; Figure 5.4a). Considering that high co-expression is frequently employed as an additional measure to establish the biological importance of a PDI, my findings suggest that these comparisons involve a more intricate regulatory framework.

In the endeavor to uncover the co-expression connections between TFs and their targets, I observed that a significant proportion (up to 17%) of targets that are not highly co-expressed with a specific TF are indeed co-expressed with TF complexes. Interestingly, a substantial number of co-expressed targets (up to 100%, averaging around 22%) were shared by multiple members of the complex, even if they were not highly co-expressed with individual TFs. These findings align with extensive literature highlighting the concept of combinatorial gene regulation (Ravasi et al., 2010; Brkljacic and Grotewold, 2017; Colinas and Goossens, 2018; Droge-Laser and Weiste, 2018). To investigate the biological significance of co-expressed targets associated with two distinct TF complexes (HHO2-HHO3 and SVP-GBF2), I examined their expression changes under stress conditions. Remarkably, in both cases, I identified differentially expressed target genes and TF members within the complex. Our results emphasize the necessity of considering the combinatorial nature of gene regulation to fully harness the potential of co-expression analyses.

Identifying ternary TF complexes experimentally presents significant challenges. To address this, I employed a comprehensive approach combining co-expression data, protein-protein interactions (PPIs), and shared targets obtained from PDI data to analyze potential TF pairs that may form ternary complexes (Figure 5.3c). For instance, I discovered eight potential ABI5 ternary

complexes involving four TFs from the ABF family (ABF1/2/3/4). These findings align with experimental evidence suggesting functional redundancy between ABF3 and ABI5 (Finkelstein et al., 2005), as well as the regulatory role of ABI5 and ABF2/3/4 in the degradation of chlorophyll-related genes (Gao et al., 2016). Moreover, it is known that ABF3/4 and NF-YC (nuclear factor Y subunit C) form a complex that controls flowering in response to drought by regulating SOC1 (SUPPRESSOR OF OVEREXPRESSION OF CONSTANS1) expression (Hwang et al., 2019), which is also targeted by ABI5 during seedling development (O'Malley et al., 2016). These results strongly suggest the formation of a larger-order complex involving ABF3-ABF4-ABI5. Together, by integrating PPIs between TFs with co-expression studies, I predicted a number of potential ternary TF complexes, which could now be experimentally validated, an easier undertaking than carrying out *do novo* identification.

Another question addressed by this study regards the nature of the association of the other genes that are highly co-expressed with a TF, if they are not targets of the TF itself. I showed that, on average for the 313 TFs investigated, almost a third of the highly co-expressed genes are either indirect targets of the TF (targets of a TF target), direct targets of the TF or direct targets of a TF partner. Is important to note that in many instances this number was much larger, which to some extent justifies the wide-spread use of co-expression as a proxy to carry out functional association of TFs and different plant traits (Haque et al., 2019; Kulkarni and Vandepoele, 2019). However, what these studies also show is that the use of co-expression is a poor indicator of direct interactions between TFs and their target genes. Establishing the co-expression relationships of TFs and their target genes has wide implications for elucidating the architecture of gene regulatory networks in all organisms and establishing the meaning of co-expression as a tool to elucidate molecular interactions.

5.5 METHODS

5.5.1 Data collection

Expression and global co-expression data were collected from the ATTED-II database (http://atted.jp/, versions Ath-r.v15-08 and Ath-r.c2-0, respectively) (Obayashi et al., 2018). In total, I used 1,416 different RNA-Seq libraries with expression data associated for 25,296 different genes. I collected the protein-DNA interaction information as raw peaks (bed or narrowpeak files from ChIP-chip, ChIP-Seq, and DAP-Seq experiments) from the Gene Expression Omnibus (GEO) and/or supplementary material from reference source (Yant et al., 2010; Wang et al., 2010; Brandt et al., 2012; Gregis et al., 2013; Jensen et al., 2013; Merelo et al., 2013; ÓMaoiléidigh et al., 2013; Heyndrickx et al., 2014b; Verkest et al., 2014; Liu et al., 2015; Nagel et al., 2015; Li et al., 2016; Liu et al., 2016; O'Malley et al., 2016; Song et al., 2016; Van Leene et al., 2016; Albihlal et al., 2018; Besbrugge et al., 2018; Chen et al., 2018; Shanks et al., 2018; Xu et al., 2018). The assignment of a peak region to a gene was carried out assuming a promoter region of 2 kb upstream from the transcription start site (TSS) for each Arabidopsis gene (genome annotation TAIR10). I used all peak region sizes as reported originally. All protein-protein interactions (PPIs) used for the identification of complex co-expressed targets were collected from the BioGRID database for Arabidopsis (V3.5.169) (Oughtred et al., 2019).

5.5.2 Evaluation of co-expression and determination of mutual rank values

For the evaluation of the global co-expression between TFs and their corresponding targets, I used the mutual ranks (MRs) of the Pearson Correlation Coefficient (PCC) and the Mutual Information (MI) as co-expression metrics. MR were defined for each gene as follows: R_{ij} is the rank of the correlation of gene *i* with the gene *j*, and R_{ji} is the rank of the correlation of gene *j* with the gene *i*, with the lowest value as the best rank (close to 1). Then, MR is equal to the square root

of R_{ij} times R_{ji}. Global MRs from positive PCC were used as reported by ATTED-II, while global MRs from negative PCC values were transformed into a second MR by subtracting the original MR reported from the maximum possible MR (25,296) for each TF. For the calculation of local MRs-PCC, I used the expression normalized as reported by ATTED-II, parsing the samples into twelve expression conditions through a dimensional reduction of the total dataset, followed by a k-means analysis (see *Methods* **5.5.10**). Grouping these samples as expression conditions, I proceeded to calculate the PCC between genes. I employed a weighted PCC to accurately measure the correlation between genes. To avoid an inflated correlation influenced by replicates, I incorporated a weighting parameter based on the correlation of corresponding samples. This approach helps prevent overestimation of the gene correlation. The weighted PCC was calculated using the R package wCorr (Version 1.9.1) (Emad and Bailey, 2017), using the same optimal threshold (0.4) as in ATTED-II. All global and local co-expression analyses using MR-MI values were carried out with the same samples used for the calculation of the respective MR-PCC values. The correlation-based on MI was estimated using the R package Parmigene (Version 1.0.2) (Sales and Romualdi, 2011), and with 1e-12 as noise to break ties due to limited numerical precision.

5.5.3 Identification of TFs co-expressed with the corresponding target genes

The significance of the MRs between TFs and their corresponding targets was assayed using both MR-PCC and MR-MI correlation metrics, and two independent statistics tests. First, I compared for each TF the average MR value of the targets vs. a null distribution of average MRs values from 1,000 random sets of genes, referred to as co-expression by MR average. Each random sample was generated by sampling with replacement *N* random genes to the *N* number of direct targets of each TF. For the MR-PCC values, I compared separately MR distributions of positively and negatively PCC values. To define if average MRs of the target genes were significantly smaller

than the null distribution, I calculated the Z-score using the MR values of the true targets using the random set of genes as background (which follow a gaussian distribution). The significance (P-value) of corresponding Z-score was corrected for multiple testing (FDR < 0.05, Benjamini-Hochberg method) (Yoav Benjamini and Yosef Hochberg, 1995). Secondly, I evaluated the differences between target and non-target genes by comparing their empirical cumulative distributions. This was done using a one-sided Kolmogorov-Smirnov test, with the alternative hypothesis being that the target genes' distribution is greater than the non-target genes' distribution. This test determined if the MRs of the target genes deviated significantly from those of the non-target genes (FDR < 0.05). Both positive and negative correlations were tested independently for both the average-based and distribution-based co-expression assessments.

5.5.4 Identification of targets co-expressed with TF complexes

The identification of complex-co-expressed targets was carried out for TFs present in our list of TFs with PDI data and at least one protein-protein interaction (PPI) between them in BioGRID. In total, I found 815 protein-protein interactions (PPIs) associated with 313 different TFs. Using these PPIs, I evaluated the effect of the formation of a TF complex (TFx-TFz) over lowly co-expressed targets (LCTs) of TFx by: (1) Assuming TFx-TFz as a new protein, thus, I averaged their expression (TFx and TFz) and then re-calculated the co-expression of the complex with a target *y*. This co-expression analysis was carried out using the weighted PCC as described above. (2) I also calculated the partial correlation of TFx with genes *y* conditioned by TFz: p(TFx ~ y | TFz), such that TFx and TFz interact between them and y is a TFx target. The partial correlation was calculated using the R package PPCOR (Kim, 2015). In both cases, I calculated the co-expression of the complex against all genes in the genome to define the significant values on the distribution obtained (See below).

5.5.5 Definition of highly co-expressed targets

I defined highly co-expressed genes as those genes in the top 5% of the correlation distribution, assuming them as genes with correlation values significantly different from the average of correlation distribution ($P < 0.05$). For PCC values, I took the 2.5% from each tail (i.e., 5% in total), while for MI values I took the top 5%. This last, given that MI does not discriminate between positive and negative associations. The approach was also implemented to define highly targeted co-expressed with a complex (TCC).

5.5.6 Degree network connectivity

I defined the in-degree and the out-degree as the number of TFs that bound the promoter of a particular target gene and the number of targets of a particular TF, respectively. Differences in both degrees, in- & out-degree, between TF co-expressed with its corresponding targets and those than not were tested by a Mann-Whitney test.

5.5.7 Protein-Protein Interactions (PPIs) and Protein-DNA interactions (PDIs) network randomization

I created random PPIs and PDI networks to test the significance of the shared targets between dimers of the tri-bi and to test the significance of number the indirect targets within the set if genes highly co-expressed with a TFs, as well as significance of number the indirect targets by TFs in cascade. In all the cases I used the *rewire* function from the R package Igraph (v1.2.4.1) to generate the random network with similar degree by node and avoiding loops (niter=NodesInNetwork*1000). Random PPI network was built with the *directed* parameter as *FALSE* while the random PDI was set as *TRUE*, which allows the shuffling of edges between TF and target genes only.

5.5.8 Definition of tri-bi complexes with significant number of shared targets

In total, I selected 104 TFs after discarding tri-bi instances with no significant target overlap, as well as TFs involved in less than two tri-bi instances (to avoid comparison with few samples). To compute the differences between the random and true PPIs, I calculated the Jaccard index (J) between every pair of dimers involved in each tri-bi, and then I asked if the mean of the J values between true tri-bi instances was different from the J values mean of tri-bi instances derived from the random PPI collection (see randomization network description).

5.5.9 Counting the HCG of a TFx that are targeted by TFz partners and TFy downstream of the corresponding TFx

To test the significance of the percentages of HCG of TF_x explained because either they are targets of an interactor TFz or a target TFy; I compared the actual set of HCGs recovered based on true interaction versus random networks (of PPI and PDI, respectively). I measured the overlap (Jaccard index) of the HCGs of TF_x with the corresponding set of TF_z and TF_y targets.

5.5.10 Definition of local expression clusters

Given the heterogeneity of the annotation of the expression samples used in this work, I defined expression clusters based on the expression similarities between the samples analyzed. First, I downloaded from the ATTED-II database the normalized expression data (Ath-r.v15-08) (Obayashi et al., 2018) used for the construction of the global co-expression database analyzed here. Second, I dimensionally-reduced the expression data by means of t-distributed stochastic neighbor embedding (t-SNE) method, to then cluster the samples using the respective t-SNE 1 and t-SNE 2 values. The t-SNE analysis was performed using the R package Rtsne (V0.15) (https://cran.r-project.org/web/packages/Rtsne/index.html), with the following parameters: pca set TRUE, perplexity=30, theta=0.5, dims=2. The clustering was performed using the R "kmeans”

function with scale t-SNE values and number of clusters equal to 12. I choose 12 clusters based on the total within sum of square (wss) value calculated using the fviz_nbclust (nboot = 300, k.max = 25) function of the R packages factoextra (v1.0.5) (https://cran.rproject.org/web/packages/factoextra/index.html).

REFERENCES

Albihlal, W.S., Obomighie, I., Blein, T., Persad, R., Chernukhin, I., Crespi, M., Bechtold, U., and Mullineaux, P.M. (2018). Arabidopsis HEAT SHOCK TRANSCRIPTION FACTORA1b regulates multiple developmental genes under benign and stress conditions. J. Exp. Bot. **69**: 2847–2862.

Arda, H.E. and Walhout, A.J.M. (2010). Gene-centered regulatory networks. Brief. Funct. Genomics **9**: 4–12.

Banf, M. and Rhee, S.Y. (2017). Computational inference of gene regulatory networks: Approaches, limitations and opportunities. Biochim. Biophys. Acta Gene Regul. Mech. **1860**: 41–52.

Banks, C.J., Joshi, A., and Michoel, T. (2016). Functional transcription factor target discovery via compendia of binding and expression profiles. Sci. Rep. **6**: 20649.

Bechtold, U. et al. (2015). Time-series transcriptomics reveals that AGAMOUS-LIKE22 affects primary metabolism and developmental processes in drought-stressed arabidopsis. Plant Cell **28**: 345–366.

Bemer, M., van Dijk, A.D.J., Immink, R.G.H., and Angenent, G.C. (2017). Cross-family transcription factor interactions: an additional layer of gene regulation. Trends Plant Sci. **22**: 66–80.

Besbrugge, N. et al. (2018). GSyellow, a Multifaceted Tag for Functional Protein Analysis in Monocot and Dicot Plants. Plant Physiol. **177**: 447–464.

Bi, C., Ma, Y., Wu, Z., Yu, Y.T., Liang, S., Lu, K., and Wang, X.F. (2017). Arabidopsis ABI5 plays a role in regulating ROS homeostasis by activating CATALASE 1 transcription in seed germination. Plant Mol. Biol. **94**: 197–213.

Brandt, R. et al. (2012). Genome-wide binding-site analysis of REVOLUTA reveals a link between leaf patterning and light-mediated growth responses. Plant J. **72**: 31–42.

Brkljacic, J. and Grotewold, E. (2017). Combinatorial control of plant gene expression. Biochim. Biophys. Acta **1860**: 31–40.

Brooks, M.D., Cirrone, J., Pasquino, A.V., Alvarez, J.M., Swift, J., Mittal, S., Juang, C.-L., Varala, K., Gutiérrez, R.A., Krouk, G., Shasha, D., and Coruzzi, G.M. (2019). Network Walking charts transcriptional dynamics of nitrogen signaling by integrating validated and predicted genome-wide interactions. Nat. Commun. **10**: 1569.

Burks, D.J., Sengupta, S., De, R., Mittler, R., and Azad, R.K. (2022). The Arabidopsis gene co-expression network. Plant Direct **6**: e396.

Chen, D., Yan, W., Fu, L.Y., and Kaufmann, K. (2018). Architecture of gene regulatory networks controlling flower development in Arabidopsis thaliana. Nat. Commun. **9**: 1–13.

Colinas, M. and Goossens, A. (2018). Combinatorial Transcriptional Control of Plant Specialized Metabolism. Trends Plant Sci. **23**: 324–336.

Droge-Laser, W. and Weiste, C. (2018). The C/S1 bZIP Network: A Regulatory Hub Orchestrating Plant Energy Homeostasis. Trends Plant Sci. **23**: 422–433.

Eisen, M.B., Spellman, P.T., Brown, P.O., and Botstein, D. (1998). Cluster analysis and display of genome-wide expression patterns. Proc. Natl. Acad. Sci. U. S. A. **95**: 14863–14868.

Emad, A. and Bailey, P. (2017). wCorr: weighted correlations.–R package ver. 1.9. 1.

Eveland, A.L. et al. (2014). Regulatory modules controlling maize inflorescence architecture. Genome Res. **24**: 431–443.

Finkelstein, R., Gampala, S.S.L., Lynch, T.J., Thomas, T.L., and Rock, C.D. (2005). Redundant and distinct functions of the ABA response loci ABA-insensitive(ABI)5 and ABRE-binding factor (ABF)3. Plant Mol. Biol. **59**: 253–267.

de la Fuente, A., Bing, N., Hoeschele, I., and Mendes, P. (2004). Discovery of meaningful associations in genomic data using partial correlation coefficients. Bioinformatics **20**: 3565–3574.

Furuya, T., Saito, M., Uchimura, H., Satake, A., Nosaki, S., Miyakawa, T., Shimadzu, S., Yamori, W., Tanokura, M., Fukuda, H., and Kondo, Y. (2021). Gene co-expression network analysis identifies BEH3 as a stabilizer of secondary vascular development in Arabidopsis. Plant Cell **33**: 2618–2636.

Gao, S., Gao, J., Zhu, X., Song, Y., Li, Z., Ren, G., Zhou, X., and Kuai, B. (2016). ABF2, ABF3, and ABF4 Promote ABA-Mediated Chlorophyll Degradation and Leaf Senescence by Transcriptional Activation of Chlorophyll Catabolic Genes and Senescence-Associated Genes in Arabidopsis. Mol. Plant **9**: 1272–1285.

Geng, H., Wang, M., Gong, J., Xu, Y., and Ma, S. (2021). An Arabidopsis expression predictor enables inference of transcriptional regulators for gene modules. Plant J. **107**: 597–612.

Gitter, A., Siegfried, Z., Klutstein, M., Fornes, O., Oliva, B., Simon, I., and Bar-joseph, Z. (2009). Backup in gene regulatory networks explains differences between binding and knockout results. Mol. Syst. Biol. **5**: 1–7.

Gomez-Cano, F., Chu, Y.-H., Cruz-Gomez, M., Abdullah, H.M., Lee, Y.S., Schnell, D.J., and Grotewold, E. (2022). Exploring Camelina sativa lipid metabolism regulation by combining gene co-expression and DNA affinity purification analyses. Plant J.

Gregis, V. et al. (2013). Identification of pathways directly regulated by SHORT VEGETATIVE PHASE during vegetative and reproductive development in Arabidopsis. Genome Biol. **14**: R56.

Gupta, O.P., Deshmukh, R., Kumar, A., Singh, S.K., Sharma, P., Ram, S., and Singh, G.P. (2021). From gene to biomolecular networks: a review of evidences for understanding complex biological function in plants. Curr. Opin. Biotechnol. **74**: 66–74.

Haque, S., Ahmad, J.S., Clark, N.M., Williams, C.M., and Sozzani, R. (2019). Computational prediction of gene regulatory networks in plant growth and development. Curr. Opin. Plant Biol. **47**: 96–105.

Harb, A., Krishnan, A., Ambavaram, M.M.R., and Pereira, A. (2010). Molecular and physiological analysis of drought stress in arabidopsis reveals early responses leading to acclimation in plant growth. Plant Physiol. **154**: 1254–1271.

Haynes, B.C., Maier, E.J., Kramer, M.H., Wang, P.I., Brown, H., and Brent, M.R. (2013). Mapping functional transcription factor networks from gene expression data. Genome Res. **23**: 1319–1328.

Heyndrickx, K.S., Vandepoele, K., Weigel, D., de Velde, J.V., and Wang, C. (2014a). A Functional and Evolutionary Perspective on Transcription Factor Binding in Arabidopsis thaliana.

Heyndrickx, K.S., Van de Velde, J., Wang, C., Weigel, D., and Vandepoele, K. (2014b). A functional and evolutionary perspective on transcription factor binding in Arabidopsis thaliana. Plant Cell **26**: 3894–3910.

Hu, Z., Killion, P.J., and Iyer, V.R. (2007). Genetic reconstruction of a functional transcriptional regulatory network. Nat. Genet. **39**: 683–687.

Hwang, K., Susila, H., Nasim, Z., Jung, J.Y., and Ahn, J.H. (2019). Arabidopsis ABF3 and ABF4 Transcription Factors Act with the NF-YC Complex to Regulate SOC1 Expression and Mediate Drought-Accelerated Flowering. Mol. Plant **12**: 489–505.

Jensen, M.K., Lindemose, S., de Masi, F., Reimer, J.J., Nielsen, M., Perera, V., Workman, C.T., Turck, F., Grant, M.R., Mundy, J., Petersen, M., and Skriver, K. (2013). ATAF1 transcription factor directly regulates abscisic acid biosynthetic gene NCED3 in Arabidopsis thaliana. FEBS Open Bio **3**: 321–327.

Jiang, S. and Mortazavi, A. (2018). Integrating ChIP-seq with other functional genomics data. Brief. Funct. Genomics **17**: 104–115.

Jin, R., Klasfeld, S., Zhu, Y., Fernandez Garcia, M., Xiao, J., Han, S.-K., Konkol, A., and Wagner, D. (2021). LEAFY is a pioneer transcription factor and licenses cell reprogramming to floral fate. Nat. Commun. **12**: 626.

Kim, S. (2015). ppcor: An R Package for a Fast Calculation to Semi-partial Correlation Coefficients. Commun Stat Appl Methods **22**: 665–674.

Kulkarni, S.R. and Vandepoele, K. (2019). Inference of plant gene regulatory networks using data-driven methods: A practical overview. Biochim. Biophys. Acta Gene Regul. Mech.: 194447.

Lacchini, E. and Goossens, A. (2020). Combinatorial Control of Plant Specialized Metabolism: Mechanisms, Functions, and Consequences. Annu. Rev. Cell Dev. Biol. **36**: 291–313.

Lai, X. et al. (2021). The LEAFY floral regulator displays pioneer transcription factor properties. Mol. Plant **14**: 829–837.

Lambert, S.A., Jolma, A., Campitelli, L.F., Das, P.K., Yin, Y., Albu, M., Chen, X., Taipale, J., Hughes, T.R., and Weirauch, M.T. (2018). The Human Transcription Factors. Cell **175**: 598–599.

Lee, T.A., Nobori, T., Illouz-Eliaz, N., Xu, J., Jow, B., and Nery, J.R. (2023). A single-nucleus atlas of seed-to-seed development in Arabidopsis. bioRxiv.

Li, D. et al. (2016). FAR-RED ELONGATED HYPOCOTYL3 activates SEPALLATA2 but inhibits CLAVATA3 to regulate meristem determinacy and maintenance in Arabidopsis. Proc. Natl. Acad. Sci. U. S. A. **113**: 9375–9380.

Liu, S., Kracher, B., Ziegler, J., Birkenbihl, R.P., and Somssich, I.E. (2015). Negative regulation of ABA signaling by WRKY33 is critical for Arabidopsis immunity towards Botrytis cinerea 2100. Elife **4**: e07295.

Liu, T.L., Newton, L., Liu, M.-J., Shiu, S.-H., and Farré, E.M. (2016). A G-Box-Like Motif Is Necessary for Transcriptional Regulation by Circadian Pseudo-Response Regulators in Arabidopsis. Plant Physiol. **170**: 528–539.

Lynch, T., Erickson, B.J., and Finkelstein, R.R. (2012). Direct interactions of ABA-insensitive(ABI)-clade protein phosphatase(PP)2Cs with calcium-dependent protein kinases and ABA response element-binding bZIPs may contribute to turning off ABA response. Plant Mol. Biol. **80**: 647–658.

Mejia-Guerra, M.K., Pomeranz, M., Morohashi, K., and Grotewold, E. (2012). From plant gene regulatory grids to network dynamics. Biochimica et Biophysica Acta (BBA) - Gene Regulatory Mechanisms **1819**: 454–465.

Merelo, P., Xie, Y., Brand, L., Ott, F., Weigel, D., Bowman, J.L., Heisler, M.G., and Wenkel, S. (2013). Genome-wide identification of KANADI1 target genes. PLoS One **8**: e77341.

Milo, R., Itzkovitz, S., Kashtan, N., Levitt, R., Shen-Orr, S., Ayzenshtat, I., Sheffer, M., and Alon, U. (2004). Superfamilies of Evolved and Designed Networks. Science **303**: 1538–1542.

Morohashi, K. et al. (2012). A genome-wide regulatory framework identifies maize pericarp color1 controlled genes. Plant Cell **24**: 2745–2764.

Morohashi, K. and Grotewold, E. (2009). A systems approach reveals regulatory circuitry for Arabidopsis trichome initiation by the GL3 and GL1 selectors. PLoS Genet. **5**: e1000396.

Nagarajan, V.K., Satheesh, V., Poling, M.D., Raghothama, K.G., and Jain, A. (2016).

Arabidopsis MYB-Related HHO2 Exerts a Regulatory Influence on a Subset of Root Traits and Genes Governing Phosphate Homeostasis. Plant Cell Physiol. **57**: 1142–1152.

Nagel, D.H., Doherty, C.J., Pruneda-Paz, J.L., Schmitz, R.J., Ecker, J.R., and Kay, S.A. (2015). Genome-wide identification of CCA1 targets uncovers an expanded clock network in Arabidopsis. Proc. Natl. Acad. Sci. U. S. A. **112**: E4802–10.

Nolan, T.M. et al. (2023). Brassinosteroid gene regulatory networks at cellular resolution in the Arabidopsis root. Science **379**: eadf4721.

Obayashi, T., Aoki, Y., Tadaka, S., Kagaya, Y., and Kinoshita, K. (2018). ATTED-II in 2018: A Plant Coexpression Database Based on Investigation of the Statistical Property of the Mutual Rank Index. Plant Cell Physiol. **59**: e3–e3.

Obayashi, T. and Kinoshita, K. (2009). Rank of correlation coefficient as a comparable measure for biological significance of gene coexpression. DNA Res. **16**: 249–260.

O'Malley, R.C., Huang, S.S.C., Song, L., Lewsey, M.G., Bartlett, A., Nery, J.R., Galli, M., Gallavotti, A., and Ecker, J.R. (2016). Cistrome and Epicistrome Features Shape the Regulatory DNA Landscape. Cell **165**: 1280–1292.

ÓMaoiléidigh, D.S., Wuest, S.E., Rae, L., Raganelli, A., Ryan, P.T., Kwasniewska, K., Das, P., Lohan, A.J., Loftus, B., Graciet, E., and Wellmer, F. (2013). Control of reproductive floral organ identity specification in Arabidopsis by the C function regulator AGAMOUS. Plant Cell **25**: 2482–2503.

Oughtred, R. et al. (2019). The BioGRID interaction database: 2019 update. Nucleic Acids Res. **47**: D529–D541.

Pajoro, A. et al. (2014). Dynamics of chromatin accessibility and gene regulation by MADS-domain transcription factors in flower development. Genome Biol. **15**: R41.

Palaniswamy, K., James, S., Sun, H., Lamb, R., Davuluri, R.V., and Grotewold, E. (2006). AGRIS and AtRegNet: A platform to link cis-regulatory elements and transcription factors into regulatory networks. Plant Phyisiol. **140**: 818–829.

Para, A. et al. (2014). Hit-and-run transcriptional control by bZIP1 mediates rapid nutrient signaling in Arabidopsis. Proc. Natl. Acad. Sci. U. S. A. **111**: 10371–10376.

Park, P.J. (2009). ChIP–seq: advantages and challenges of a maturing technology. Nat. Rev. Genet. **10**: 669–680.

Rao, X. and Dixon, R.A. (2019). Co-expression networks for plant biology: Why and how. Acta Biochim. Biophys. Sin. **51**: 981–988.

Ravasi, T. et al. (2010). An atlas of combinatorial transcriptional regulation in mouse and man. Cell **140**: 744–752.

Reményi, A., Schöler, H.R., and Wilmanns, M. (2004). Combinatorial control of gene

expression. Nat. Struct. Mol. Biol. **11**: 812.

Sales, G. and Romualdi, C. (2011). Parmigene-a parallel R package for mutual information estimation and gene network reconstruction. Bioinformatics **27**: 1876–1877.

Sayou, C. et al. (2016). A SAM oligomerization domain shapes the genomic binding landscape of the LEAFY transcription factor. Nat. Commun. **7**: 11222.

Shanks, C.M., Hecker, A., Cheng, C.-Y., Brand, L., Collani, S., Schmid, M., Schaller, G.E., Wanke, D., Harter, K., and Kieber, J.J. (2018). Role of BASIC PENTACYSTEINE transcription factors in a subset of cytokinin signaling responses. Plant J. **95**: 458–473.

Skubacz, A., Daszkowska-Golec, A., and Szarejko, I. (2016). The Role and Regulation of ABI5 (ABA-Insensitive 5) in Plant Development, Abiotic Stress Responses and Phytohormone Crosstalk. Front. Plant Sci. **7**: 1884.

Song, L., Huang, S.-S.C., Wise, A., Castanon, R., Nery, J.R., Chen, H., Watanabe, M., Thomas, J., Bar-Joseph, Z., and Ecker, J.R. (2016). A transcription factor hierarchy defines an environmental stress response network. Science **354**.

Stuart, J.M., Segal, E., Koller, D., and Kim, S.K. (2003). A gene-coexpression network for global discovery of conserved genetic modules. Science **302**: 249–255.

Swift, J. and Coruzzi, G.M. (2017). A matter of time - How transient transcription factor interactions create dynamic gene regulatory networks. Biochimica et Biophysica Acta (BBA) - Gene Regulatory Mechanisms **1860**: 75–83.

Tao, Z., Shen, L., Gu, X., Wang, Y., Yu, H., and He, Y. (2017). Embryonic epigenetic reprogramming by a pioneer transcription factor in plants. Nature **551**: 124–128.

Trigg, S.A. et al. (2017). CrY2H-seq: a massively multiplexed assay for deep-coverage interactome mapping. Nat. Methods **14**: 819–825.

Uygun, S., Peng, C., Lehti-Shiu, M.D., Last, R.L., and Shiu, S.H. (2016). Utility and Limitations of Using Gene Expression Data to Identify Functional Associations. PLoS Comput. Biol. **12**: 1–27.

Vandepoele, K., Quimbaya, M., Casneuf, T., De Veylder, L., and Van de Peer, Y. (2009). Unraveling transcriptional control in Arabidopsis using cis-regulatory elements and coexpression networks. Plant Physiol. **150**: 535–546.

Van Leene, J. et al. (2016). Functional characterization of the Arabidopsis transcription factor bZIP29 reveals its role in leaf and root development. J. Exp. Bot. **67**: 5825–5840.

Varala, K. et al. (2018). Temporal transcriptional logic of dynamic regulatory networks underlying nitrogen signaling and use in plants. Proc. Natl. Acad. Sci. U. S. A. **115**: 6494–6499.

Verkest, A. et al. (2014). A generic tool for transcription factor target gene discovery in

Arabidopsis cell suspension cultures based on tandem chromatin affinity purification. Plant Physiol. **164**: 1122–1133.

Wang, C., Xu, J., Zhang, D., Wilson, Z.A., and Zhang, D. (2010). An effective approach for identification of in vivo protein-DNA binding sites from paired-end ChIP-Seq data. BMC Bioinformatics **11**: 81.

Wilkins, O., Bräutigam, K., and Campbell, M.M. (2010). Time of day shapes Arabidopsis drought transcriptomes. Plant J. **63**: 715–727.

Wisecaver, J.H., Borowsky, A.T., Tzin, V., Jander, G., Kliebenstein, D.J., and Rokas, A. (2017). A Global Coexpression Network Approach for Connecting Genes to Specialized Metabolic Pathways in Plants. Plant Cell **29**: 944–959.

Wu, G. and Ji, H. (2013). ChIPXpress: Using publicly available gene expression data to improve ChIP-seq and ChIP-chip target gene ranking. BMC Bioinformatics **14**.

Xu, C., Cao, H., Xu, E., Zhang, S., and Hu, Y. (2018). Genome-Wide Identification of Arabidopsis LBD29 Target Genes Reveals the Molecular Events behind Auxin-Induced Cell Reprogramming during Callus Formation. Plant Cell Physiol. **59**: 744–755.

Yang, F. et al. (2017). A Maize Gene Regulatory Network for Phenolic Metabolism. Mol. Plant **10**: 498–515.

Yant, L., Mathieu, J., Dinh, T.T., Ott, F., Lanz, C., Wollmann, H., Chen, X., and Schmid, M. (2010). Orchestration of the floral transition and floral development in Arabidopsis by the bifunctional transcription factor APETALA2. Plant Cell **22**: 2156–2170.

Yilmaz, A., Mejia-Guerra, M.K., Kurz, K., Liang, X., Welch, L., and Grotewold, E. (2011). AGRIS : the Arabidopsis Gene Regulatory Information Server , an update. Nucleic Acids Res. **39**: 1118–1122.

Yoav Benjamini and Yosef Hochberg (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. J. R. Stat. Soc. Series B Stat. Methodol. **57**: 289–300.

Yoshida, T., Fujita, Y., Sayama, H., Kidokoro, S., Maruyama, K., Mizoi, J., Shinozaki, K., and Yamaguchi-Shinozaki, K. (2010). AREB1, AREB2, and ABF3 are master transcription factors that cooperatively regulate ABRE-dependent ABA signaling involved in drought stress tolerance and require ABA for full activation. Plant J. **61**: 672–685.

Zaborowski, A.B. and Walther, D. (2020). Determinants of correlated expression of transcription factors and their target genes. Nucleic Acids Res. **48**: 11347–11369.

Zeller, K.I. et al. (2006). Global mapping of c-Myc binding sites and target gene networks in human B cells. Proceedings of the National Academy of Sciences **103**: 17834.

Zhou, P., Li, Z., Magnusson, E., Gomez Cano, F., Crisp, P.A., Noshay, J.M., Grotewold, E.,

Hirsch, C.N., Briggs, S.P., and Springer, N.M. (2020). Meta Gene Regulatory Networks in Maize Highlight Functionally Relevant Regulatory Interactions. Plant Cell **32**: 1377–1396.

APPENDIX

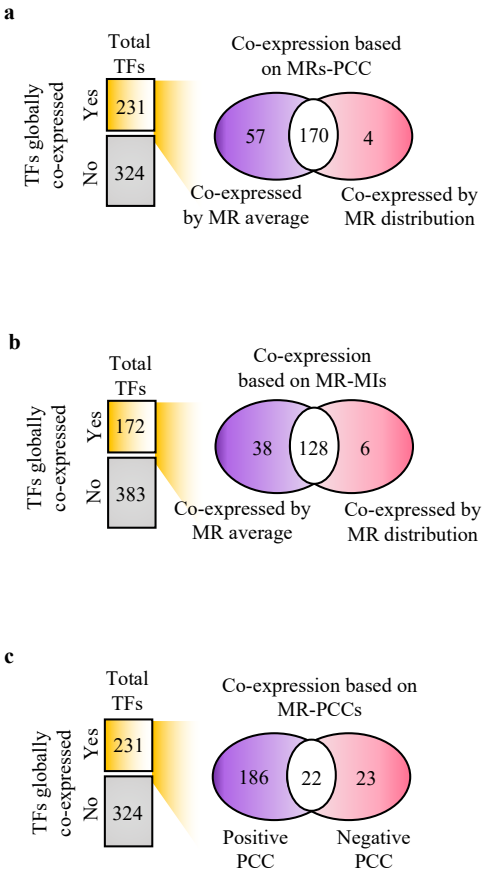

Figure S5.1 Evaluation of co-expression of TFs and corresponding target genes

a and **b**. Comparison of the two statistical approaches used to test differences in either average or distribution of MRs between targets and not targets genes by **(a)** PCC-MR or **(b)** MI-MR. **c.** Venn diagrams comparing the total number of positive and negatively co-expressed TFs with their targets based on PCC-MR.

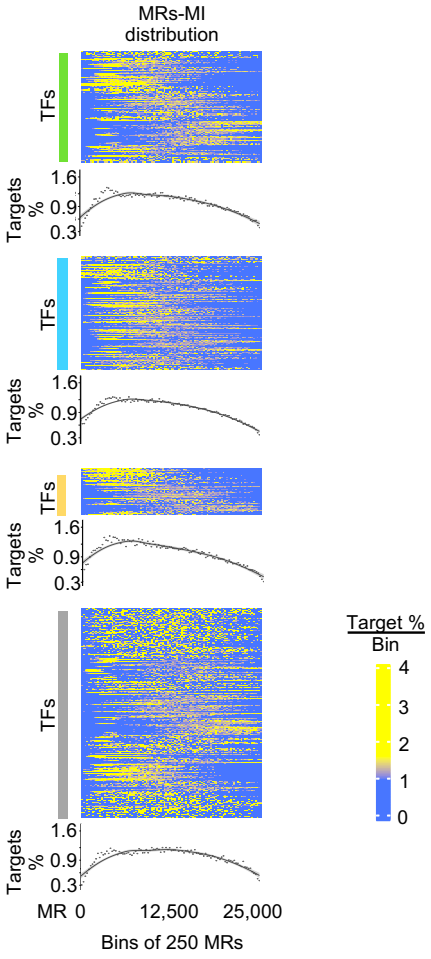

Figure S5.2 Heatmaps displaying the distribution of MR-MI values across 25,296 Arabidopsis genes

Colors represent the percentage of TF targets within bins of 250 MRs. In total, there are 101 bins along the PCC distribution corresponding to co-expression values of each TF with 25,296 genes (genes expressed in the dataset used, see *Methods*). Small MR represent larger MI, thus, better association between TF and genes in bin. Dot plots under each heat map represent the average percentage of targets for all the TFs along each bin. Color side bars represent TFs categories as presented in Figure 5.1.

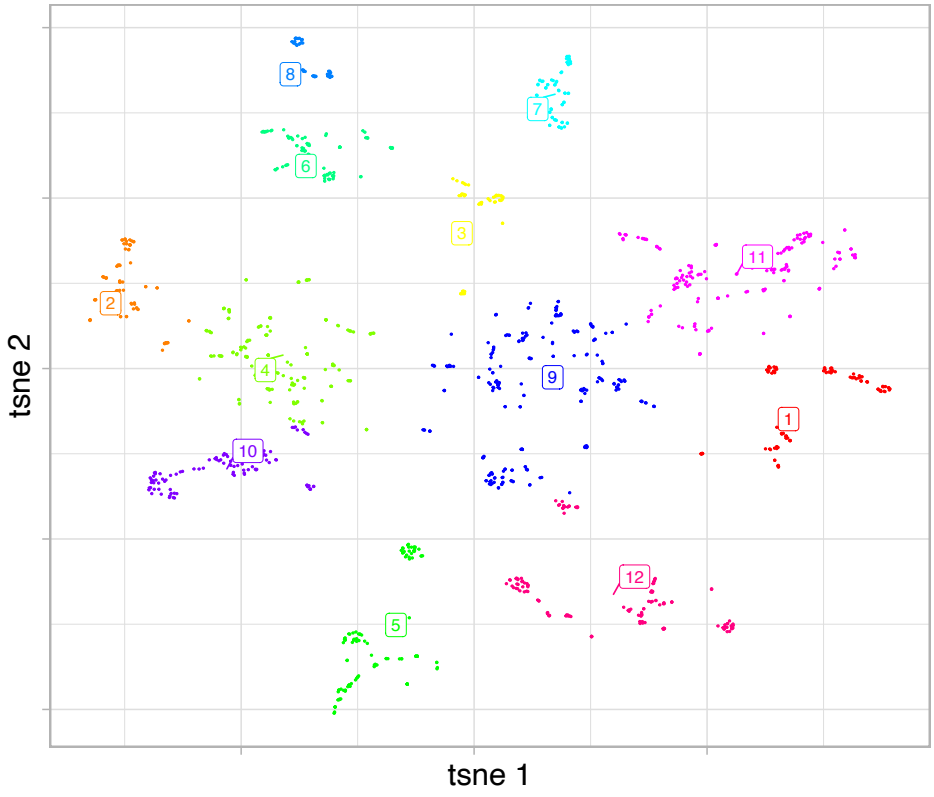

Figure S5.3 Sample expression clusters used to define local expression values

Clusters were defined by k-means clustering (k=12 defined by Elbow method) using the t-Distributed Stochastic Neighbor Embedding (t-SNE) 1 and 2 of the expression data.

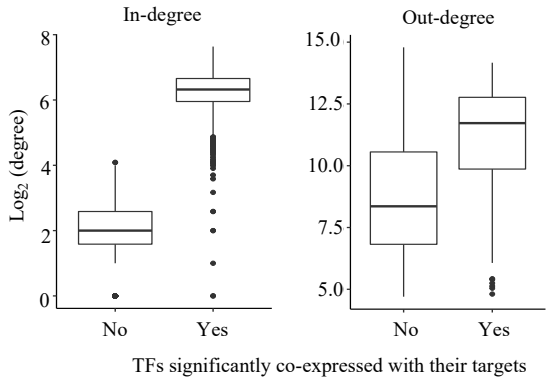

Figure S5.4 In- and out-degree differences between TFs co-expressed and not-co-expressed with their targets

This classification accounts for both globally and locally co-expression results. Both types of degree (in and out) showed statistically significant differences between TFs co-expressed or not co-expressed with its targets (Mann-Whitney U test, P. value < 0.05).

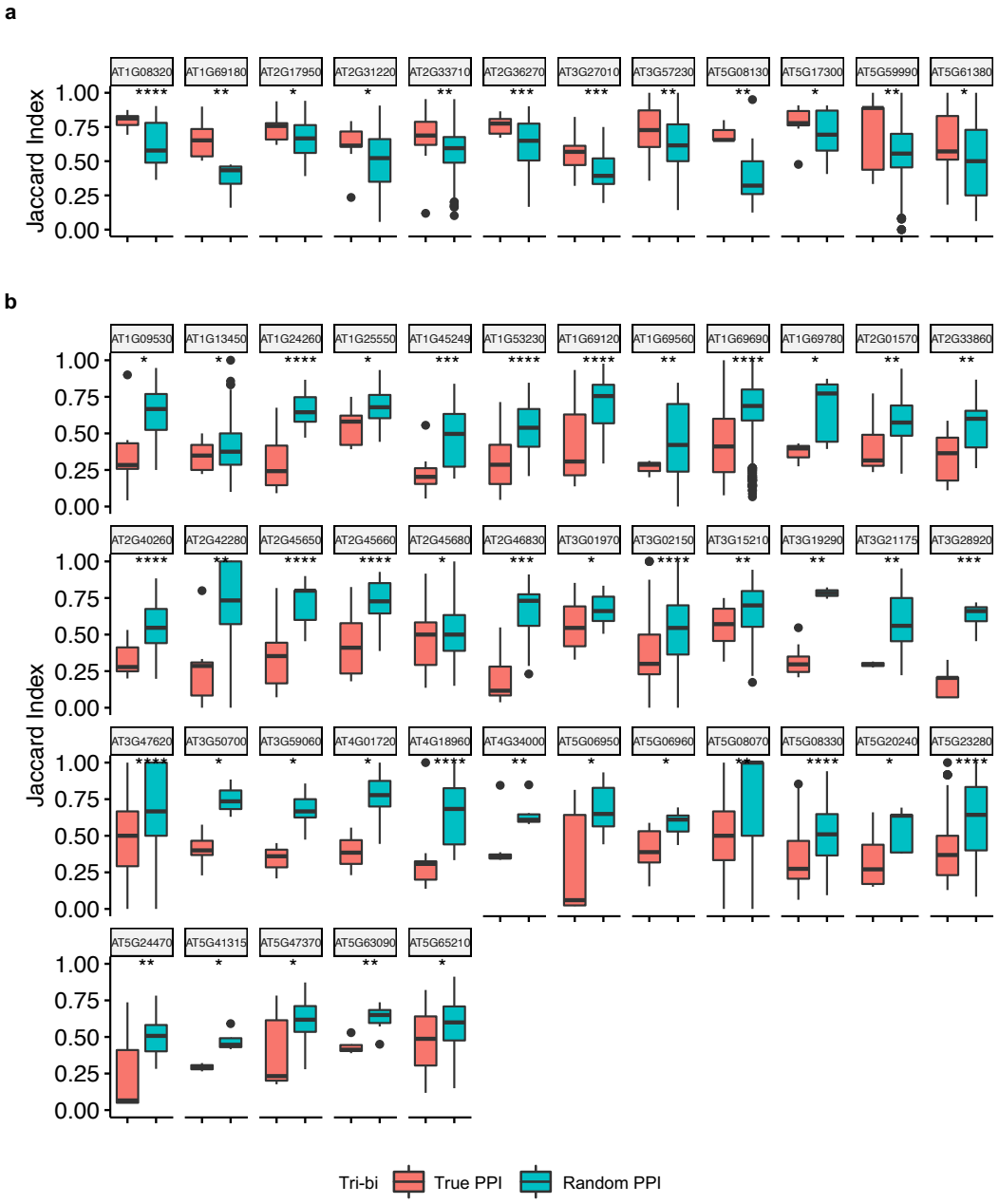

Figure S5.5 Target genes recovered for tri-bi complexes

a and **b**. Comparison of target genes recovered for tri-bi of 53 TFs with a shared fraction significantly larger **(a)** or smaller **(b)** than by random PPIs. The similarity of the recovery set of targets was measured as the Jaccard index between the set of targets of each pair of dimers that form a tri-bi complex. Asterisks indicate P-value significance (*: p <= 0.05, **: p <= 0.01, ***: p <= 0.001, ****: p <= 0.0001, two-sided *t*-test).

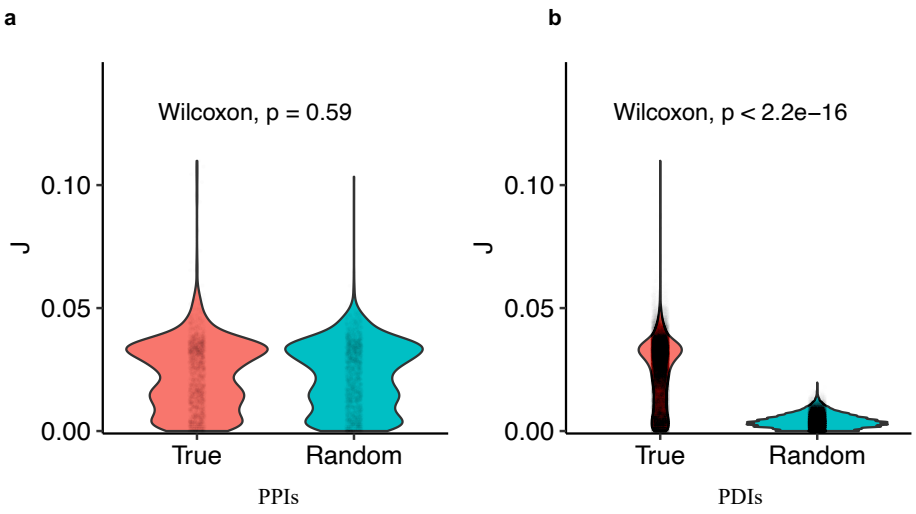

Figure S5.6 Evaluation of HCG not targets of TFx
a and **b**. Comparison of HCG which are not targets of TF_x recovered because they are either **(a)** a target of a TFz interactor of TFx, or **(b)** a target of a TFy regulated by TFx vs random interaction. Jaccard index (J) calculated as the number of TFz/TFy targets shared with the HCGs non-targets of TFx over the total TFz/y targets plus total HCGs no-targets.

CHAPTER SIX: CONCLUSIONS

Understanding gene regulatory networks (GRNs) has significant implications at various levels and in every biological system. However, the unraveling of GRNs, predicting their interactions, and prioritizing regulatory associations with phenotypic/biological consequences remains a long-standing and unsolved problem. I employed several strategies in this study to predict and comprehend the associations between transcription factors (TFs) and target genes in different plant systems using various data types. The results propose previously unknown regulatory associations specific to these plant systems and offer guidance for future research aimed at unraveling GRNs in a species-specific manner. It is important to note that the species-specific strategies presented here were primarily based on the availability and nature of the data.

I developed a co-expression system with highly stringent thresholds in Camelina, a plant known for its complex genome and limited data available (compared to maize and Arabidopsis). This analysis did not rely on any assumptions about TF-target gene relationships using a specific metabolic pathway as an example, simplifying the analysis of regulatory associations. Thus, my combined analysis of co-expression and PDI predicted six TFs involved with lipid metabolism in Camelina. Five of these TFs were not previously associated with lipid metabolism in any other plant system. Moving on to maize, the extensive genetic data and the increasing availability of PDI data sets enabled me to create a framework for evaluating various approaches for the integration of multi-omic data to predict TF regulatory function. In addition to finding the best strategy and specific regulatory hypotheses for further validation, these analyses identify numerous potential functional connections which showed enrichment with GO terms also observed in random networks. Importantly, this indicates that a substantial portion of the predicted associations involves a significant number of interactions that are either false positives or whose related GO terms are just too over-represented, resulting in enrichment even from random interactions

(random networks). In addition, I showed that the embedding representation of the network allows not only to identify TF-functionally redundant but also TF paralogous potentially redundant. Lastly, the established method allowed me to predict transcriptional regulators of different biological processes as well as potential upstream regulators of the corresponding TFs. Based on my analysis, I predicted a comprehensive list of regulators for twenty distinct biological processes. These processes range from development to metabolism and include associations of transcription factors that were identified in the past, thus validating our predictions.

Finally, the abundance of data available in Arabidopsis enabled me to establish a broader framework regarding the predictability of target genes for TFs based on co-expression models. My findings confirm a long-standing observation: many direct target genes do not exhibit significant co-expression with their corresponding TF. Additionally, many genes co-expressed with TFs are not direct targets of the respective TF. However, my analysis expanded on this observation by revealing the influence of physical TF-TF interactions and downstream TFs in explaining the occurrence of either low-expression targets or highly co-expressed genes that are not targets. Altogether, this work established tools and strategies and provided hypotheses to understand GRN in plants better.

My results predicted regulatory associations that were previously unknown, it's important to note that some of these associations are currently under validation experimentally by other researchers in the Grotewold lab. Yet, it’s also worth mentioning that these findings may have a fraction of false positives regulatory associations for the corresponding TFs. Determining the exact fraction of false positives is challenging due to the nature of the data available. While it's common to identify examples of potential regulatory associations in the literature, it is less common to find experimentally validated negative examples with nonregulatory effects. This creates uncertainty

regarding the false positive fraction. Yet, it is important to mention, one contributing factor to false positives is the heterogeneity of the data. For example, there are variations in the expression datasets analyzed using different pipelines, as well as differences in the way peaks from PDI assays are called and the technologies used to generate the peaks. Thus, my implementation of filters based on normalized counts and discarding of peaks without consensus TFBM was largely complemented by the utilization of the peaks in the context of open chromatin regions. In maize, this normalization allowed for comparisons across different types of experiments (i.e., DAP-seq and ChIP-seq) and reduced the assignment of potential false positive target genes. Specifically, I discarded at least 50% of the called peaks because of their overlap with nearby closed chromatin regions. Therefore, it is highly recommended to exclude this layer of information in future endeavors to predict and understand GRN in plants.

Aside from the methodological limitations, it is also important to highlight that all the analyses and results presented here examined TFs and genes as individual interactions. However, in several cases, my results keep leading back to situations where only the integration of the corresponding predictions makes sense. For example, this is evident in their interpretation as complexes (partial correlation results in Arabidopsis) and modules of TFs working together in specific biological processes (presence of multiple TFs with similar influence on individual GO terms, results in maize). Therefore, a major improvement to the models and analyses presented here would include the interrogation of multiple TFs in the same model or establishing predictive models that consider the modular and combinatorial nature of gene regulation. For instance, it is easy to envision the typical co-expression model for linear regression or a similar model (regardless of the model's assumptions) that considers the contribution of multiple TFs to the expression variation of the corresponding target genes. Additionally, with techniques such as DAP-seq and ATAC-seq, it is

undeniable that models aiming to build regulatory models in the context of open chromatin regions, not only for single TFs but also for co-expressed TFs, will enable the prediction of TF-target interactions from PDIs dataset with significantly higher accuracy.

When I combined the results of all three systems, it's important to highlight the effectiveness of integrating multiple layers to handle the complexity of gene regulatory networks (GRNs). Moreover, considering the identification of common associations across different data types offers a straightforward method to analyze data and discover highly reliable interactions. Yet, looking at the bigger picture, incorporating information from various layers into more robust prediction strategies reduces the likelihood of false positives, and consequently increasing the accuracy of the corresponding predictions. It's also worth noting that TFs generally have multiple regulatory functions and are significantly redundant, which has been extensively documented. In this study, I demonstrated that by integrating multiple data types, it's possible to narrow down this complexity and formulate specific testable regulatory hypotheses.