Ghent University
Faculty of Sciences
Department of Plant Biotechnology and Bioinformatics

Applications of Network-based and Integrative Approaches in Plant Systems Biology

Toepassingen van Netwerkgebaseerde en Integratieve Methodes in de Planten Systeembiologie

By

Razgar Seyed Rahmani

Supervisors:

Prof. Kathleen Marchal

Dr. Giles Miclotte

A dissertation submitted to Ghent University in partial fulfilment of the requirements for the degree of Doctor of Science, Bioinformatics

Academic year: 2021-2022

Acknowledgments

"If you want to lift yourself up, lift up someone else"
- Booker T. Washington

As a biotechnologist who pursued bioinformatics, my Ph.D. journey was full of challenges and experiences. I owe my success in completing this challenging task to the many people who helped me from the very beginning to the end. So, it is time to step back and acknowledge the support of those who helped me and made this journey possible, memorable, and enjoyable.

First and foremost, I would like to express my appreciation to both my supervisors Prof. Kathleen Marchal and Dr. Giles Miclotte for their consistent support and guidance. Thank you both for all your constant encouragement, profound belief in my abilities, friendly meetings, and personal support whenever I needed it. Your deep insights in this field were the fuel for my inspiration and determination. I feel very lucky to have such kind and supportive mentors.

I am deeply indebted to both of my advisors from Iran: Prof. Seyed Abolghasem Mohammadi and Prof. Mahmoud Toorchi, for introducing me to bioinformatics and for their constant motivation and kind advice. Allow me to extend my special thanks to Prof. Mohammad Moghadam and all other members of the Department of Plant Breeding and Biotechnology, University of Tabriz, Iran, for all their kind help and motivation.

A significant part of this thesis is based on the collaborations with Wuhan Botanical Garden, The Chinese Academy of Sciences. I am thankful to all my Chinese colleagues for sharing wonderful datasets and projects. Special thanks to Dr. Tao Shi, without our successful collaboration the completion of my dissertation would not have been possible. I am also so grateful to Prof. Yves Van de Peer for setting up those collaborations.

I would also like to offer gratitude to Prof. Jan Fostier and Dr. Dries Decap for the wonderful collaboration and for their insights during the final project. It was a pleasure working with you in such a friendly and constructive atmosphere.

A very special thanks to Prof. Steven Meare, for chairing my Ph.D. examination board, and to the rest of the jury members: Prof. Hilde Nelissen, Prof. Pieter De Bleser, Prof. Jan Fostier, and

Dr. Tao Shi for their constructive comments, friendly discussion, and saving me from some embarrassing mistakes.
This thesis has been generously supported by the Ministry of Science, Research and Technology of Iran (42 months) and Ghent University (12 months). I would like to thank again Prof. Seyed Abolghasem Mohammadi for helping me to acquire the first grant and Prof. Kathleen Marchal for the second grant. I am also thankful to IDLab-imec for hosting me during my stay at Ghent University and for all their generous help and support.
I also had the pleasure of working with great colleagues and friends: Lieven, Mahdi, Camilo, Maarten, Aranka, Louise, David, Patricio, Pieter-Paul, Simon, Luca, Lore, and Marie. Thank you all for your company and for being so friendly. Special thanks to all my Iranian friends for their support and for spending an enjoyable time together.
I also thank my parents for their endless love and nonstop support throughout my life. Thank you both for setting me off on the road and giving me the strength to follow my dreams. Thank you Klsum, Chiman, Nazdar, and Kamal for being the best sisters and brother ever.
Last but not least, I would like to thank my loving and awesome wife, Sayma, for patiently being there for me and standing by me during every struggle and all my successes. I am so lucky to have you in my life.

Razgar
May 2022, Ghent

Summary

The field of molecular biology has been revolutionized thanks to progress in sequencing technologies, resulting in the availability of an increasingly large number of omics datasets. As a result, the analysis and interpretation of big data created new challenges in animal and plant research communities. Various types of data are being produced to answer biological questions, for example (epi)genomics, transcriptomics, proteomics, and metabolomics data. Given the high availability of omics data sets and prior molecular information, processing these data with a single statistical model seems promising for obtaining useful insights concerning the biological process under consideration. However, due to the inherent noise of the underlying technologies that are used to generate high-throughput data, data interpretation requires vast biological expertise on top of methods that can efficiently process those data. Moreover, when analyzing a novel dataset, the findings from previous studies must be considered, since each dataset may convey unique information. The combination of novel and earlier data sets provides a more comprehensive understanding of the biology underlying the process of interest. To this end, we distinguish between two types of integrative approaches: (1) exploratory data analysis (post-analysis integration), and (2) model-based data integration.

Exploratory data analysis refers to an intuitive way of first analyzing each omics dataset separately, and then combining the resulting key features. The limitation of exploratory data analysis is that crosstalk between molecular entities at different molecular layers cannot properly be explored.

Model-based data integration is applicable if all layers of information can be captured in a single model. In our setting, we make use of methods that rely on a network model to represent and interpret different data sources. Network-based approaches have become popular in the field of system biology, since these methods allow for an intuitive way to integrate and interpret large-scale data sets. The application of network-based methods in this thesis can be classified into two categories: i) methods that infer networks from omics data to visualize/represent the data as gene-gene interaction network (e.g., coexpression networks); ii) methods that use the inferred gene-gene interaction network to drive the analysis. By mapping omics data on the inferred network prior, the molecular mechanism that drives the phenotype of interest can be revealed.

In this thesis, we apply integrative approaches in plant system biology in the context of different collaborative studies in order to solve outstanding biological problems.

In the first collaborative case study, we assess how different expression, epigenetic regulation, and functional constraints are associated with the fate of genes that underwent a whole genome duplication in sacred lotus (*Nelumbo nucifera*). For this study, dedicated expression and methylation data in lotus were available. These data were analyzed and their results were complemented with network-based analyses to study the behavior of genes of different duplication origins. We found that, after a whole genome duplication, genes that returned to a single copy state show the highest levels and breadth of expression, gene body methylation, and intron numbers. On the other hand, long-retained duplicates exhibit the lowest methylation in gene flanking regions and the highest degrees of protein-protein interactions and protein lengths. Our results highlight the impact of different functional constraints on gene fate and duplicate divergence, following a single whole genome duplication in lotus.

In a second case study, we used an integrative approach to systematically study brassinosteroid signaling in *Arabidopsis thaliana* using mutant lines. The mutant lines were identified through a genetic screening process, including a line carrying a loss-of-function mutation in the *BRI1* gene along three activation-tag suppressor lines. By using subnetwork inference methods in order to combine molecular prior information on gene-gene interactions with routine genomics analysis, we gained valuable insight into the studied process. Our results showed that network-based methods can provide a clear global picture of the molecular mechanisms underlying brassinosteroid signaling.

In the final chapter, we identify functional cis-regulatory elements by supplementing a computational method, BLSSpeller, with an integrative approach. BLSSpeller is a computational method to predict cis-regulatory elements by incorporating information from closely related species. On top of redesigning BLSSpeller to cope with larger datasets, we showed how a multi-omics approach can be applied to comprehensively complement the prediction results in order to extract the most reliable predictions. The results of this analysis are useful to infer gene functions and to complement gene regulatory networks in maize.

Nederlandstalige samenvatting

Dankzij de vooruitgang in de ontwikkeling van sequentietechnologieën is het domein van de moleculaire biologie hervormd, met als gevolg een immer groeiend aantal omics datasets. Hierdoor werden de analyse en de interpretatie van big data de nieuwe uitdagingen in het onderzoek rond plant en dier. Er worden verschillende soorten data geproduceerd om biologische vragen te beantwoorden, bijvoorbeeld (epi)genomische, transcriptomische, proteomische en metabolische data. Gezien de grote hoeveelheid omics-datasets en reeds gekende moleculaire informatie, lijkt het interessant om al deze data te verwerken met één enkel statistisch model, om zo tot nuttige inzichten te komen over het betreffende biologisch proces. Door de inherente ruis van de onderliggende technologieën, vereist de interpretatie van deze data enerzijds een enorme biologische expertise en anderzijds methoden die deze data efficiënt kunnen verwerken. Bovendien moeten bij het analyseren van een nieuwe dataset de bevindingen uit eerdere studies ook in overweging genomen worden, omdat elke dataset unieke informatie kan overbrengen. De combinatie van nieuwe en eerdere datasets biedt een bredere kijk op het onderliggende biologisch proces. Hiertoe onderscheiden we twee integratieve aanpakken: (1) verkennende data-analyse (integratie-na-analyse), en (2) data-integratie gebaseerd op een model.

De verkennende data-analyse, is een intuïtieve manier waar elke omics-dataset eerst afzonderlijk geanalyseerd wordt, waarna vervolgens de resulterende bevindingen gecombineerd worden. De beperking van deze aanpak is dat eventuele crosstalk tussen de moleculaire entiteiten in de verschillende moleculaire niveaus moeilijk te onderzoeken is.

De data-integratie gebaseerd op een model, kan gebruikt worden indien er een gepast model is waarin alle lagen informatie bevat kunnen worden. In onze setting steunen we op een netwerkmodel om verschillende databronnen weer te geven en te interpreteren. Netwerkgebaseerde modellen zijn erg populair in de systeembiologie, omdat deze toelaten om grootschalige datasets op een intuïtieve manier te integreren en te interpreteren. In dit proefschrift onderscheiden we twee categorieën van e toepassingen van netwerkgebaseerde modellen: i) methodes die netwerken opstellen op basis van omics-data, hier worden de data gevisualiseerd/gerepresenteerd door gen-gen-interacties (bijv. een coexpressienetwerk); ii) methodes waarin een gen-gen-interactienetwerk de analyse aandrijft. Hier wordt nieuwe data

afgebeeld op een gekend netwerk, waardoor het onderliggende moleculaire mechanisme onthuld wordt.

In dit proefschrift passen we integratieve aanpakken toe in de systeembiologie, dit doen we in de context van verschillende wetenschappelijke samenwerkingen, met als doel openstaande biologische problemen op te lossen.

In de eerste casestudy beoordelen we hoe afwijkende expressie, epigenetische regulering en functionele beperking geassocieerd worden met de bestemming van genen na een volledige genoomduplicatie in heilige lotus (Indische lotus; *Nelumbo nucifera*). Voor deze studie was er specifiek expressie- en methyleringsdata beschikbaar. Deze data werden geanalyseerd en de resultaten werden vervolgens aangevuld met netwerkgebaseerde analyses om het gedrag van genen met verschillende duplicatieorigines te onderzoeken. We ontdekten dat, genen die na een volledige genoomduplicatie terugkeerden naar enkelvoudige kopieën de hoogste niveaus en breedte van expressie, genlichaammethylering en intronengetallen tonen. Anderzijds, vertonen de blijvende duplicaten de laagste methylering in regio's grenzend aan genen en de hoogste graad van proteïne-proteïne-interacties en proteïnelengte. Onze resultaten benadrukken het effect van verschillende functionele beperkingen op de bestemming van genen en dubbele divergentie, na een enkele volledige genoomduplicatie in lotus.

In de tweede casestudy gebruikten we een integratieve aanpak om systematisch de brassinosteroïdesignalering in zandraket (*Arabidopsis thaliana*) te bestuderen met behulp van gemuteerde lijnen. De gemuteerde lijnen werden geïdentificeerd met genetische screening, en bevatten een lijn die een verlies-van-functieverandering in het gen BRI1 droeg langs drie activatie-label-onderdrukkingslijnen. Met behulp van netwerkgebaseerde methodes, verkregen we waardevolle inzichten in het onderzochte proces door gekende moleculaire informatie op gen-gen-interacties te combineren met een klassieke genomische analyse. Onze resultaten toonden aan dat netwerkgebaseerde methoden een duidelijk beeld kunnen geven van de moleculaire mechanismen die brassinosteroïdesignalering aandrijven.

In de laatste casestudy identificeren we functionele cis-regulerende elementen door een computationele methode, BLSSpeller, aan te vullen met een integratieve aanpak. BLSSpeller is een computationele methode om cis-regulerende elementen te voorspellen door informatie van nauw verwante soorten in rekening te nemen. Naast het herontwerpen van BLSSpeller om grotere datasets te kunnen verwerken, toonden we aan hoe een multi-omics aanpak de voorspellingsresultaten kan complementeren om zo de betrouwbaarste voorspellingen te verkrijgen. De resultaten van deze analyse zijn nuttig om genfuncties te bepalen en om genregulerende netwerken voor mais *(Zea mays)* aan te vullen.

Table of Contents

CHAPTER 2

List of Figures

List of acronyms

ABA: Abscisic acid
ACR: Active chromatin reigns
ATAC-seq: Assay for Transposase-Accessible Chromatin using sequencing
BAK1: BRI1-Associated receptor kinase
BES: Br-Insensitive-Ems-Suppressor
BIN: Brassinosteroid insensitive
BLS: Branch length score
BR: Brassinosteroid
BRI: BR insensitive

BRL1: BRI1-like homologs
BRS: BRI suppressor
BRZ: Brassinazole-Resistant
CDS: Coding DNA sequence
ChIP-seq: chromatin immunoprecipitation- sequencing
CRE: cis-regulatory elements
ECS: Extra Carpels and seeds
FDR: False discovery rate
FLS: Flagellin-Sensitive
FPKM: Fragments per kilobase of transcript per million mapped reads
GRN: Gene regulatory network
GBA: Guilt by association
GBH: gene balance hypothesis
GCN: Gene coexpression network
GO: Gene ontology
Hi-C: Throughput chromosome conformation capture
LF: Less fractionated
LSD: large-scale duplication
MF: more fractionated
NGS: Next-generation sequencing
PCA: Principal component analysis
PIN: Protein interaction network
ROS: Reactive oxygen species
RT-qPCR: Real time quantitative polymerase chain reaction
SCP: Serine carboxypeptidase
SNP: single nucleotide polymorphism
SSD: small-scale gene duplication
TD: Tandem duplication
TE: Transposable elements
TFBS: Transcript factor binding site
TF: Transcription Factor
TSS: transcriptional start sites
WGD: Whole genome duplication

1 Chapter 1

Introduction and background

"The best and most beautiful things in the world cannot be seen or even touched — they must be felt with the heart"
- Helen Keller

1.1 Introduction

In the last decade, biology has increasingly become data-driven. Large datasets generated by high-throughput technologies which containing information on various omics layers such as (epi)genomics, transcriptomics, proteomics, and metabolomics have been submitted to the public domain (Figure 1.1). The availability of such large-scale datasets brought new challenges to the scientific community in data processing, correction, and interpretation perspective. As more data become available, the integration of these datasets provides a holistic view of complex biological systems and of the interplay between different components of those systems. Methods to integrate omics data and prior molecular knowledge can be subdivided into exploratory data analysis and model-based data integration [1]. Different integrative methods rely on the use of a network model to represent different information sources.
In this thesis, we tried to answer a diverse set of fundamental research questions in plant biology by integrating different sources of omics datasets and prior information. Below we first introduce some general biological concepts. Subsequently, we provide an overview of the omics data sources that are used in the thesis, and finally, we describe why network-based methods offer an ideal approach to represent the prior molecular information and omics datasets in the integrative approaches.

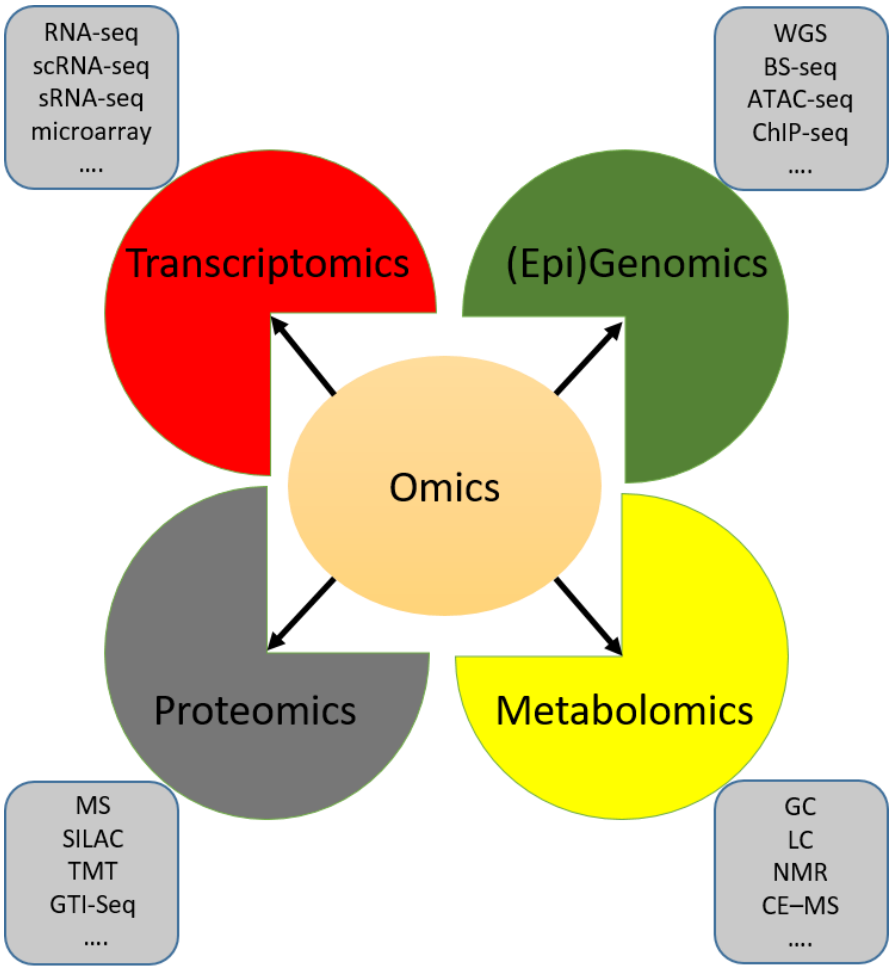


Figure 1.1 Large-scale datasets at different omics levels are becoming available that provide information on the structure, quantification, and modification of molecular entities at different omics levels. Examples of advanced technologies to produce omics data are shown next to each omics level.

1.2 Basic biological concepts used in the thesis

All the genetic information encoded in DNA consists of a four-letter code: A, T, C, G; standing for Adenine, Thymine, Cytosine, and Guanine, respectively. These are the nucleotide bases of DNA and their combination encodes all the information necessary to perform required biological functions during growth, development, and in response to internal and external signals. In order to translate the code to biological function, the code has to be transcribed into messenger RNA (mRNA) by the transcription machinery. Subsequently, the translation machinery, the ribosome, can read the mRNA and turn it into proteins. The genetic code in mRNA molecules is read in triplets, each representing a code for a single amino acid. Finally, chains of amino acids form proteins which play many critical roles in the cells and translate the genetic codes to phenotypes (Figure 1.2).

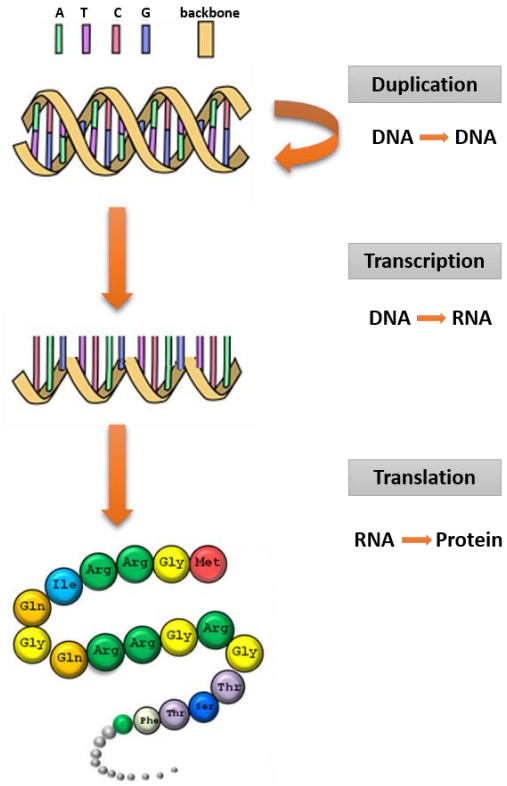


Figure 1.2 Simplified picture of complex processes by which the genetic information stored in DNA is translated into RNA and then RNA is translated into proteins as the functional agents in the cell.

Changes to the genetic code (mutation) can potentially lead to changes in proteins being produced and thus to distinct phenotypes. Although the majority of the mutations are either neutral or deleterious, they are a great source of phenotype variation that is at the heart of evolution and adaptation in response to environmental changes. In the plant genomics community, quite some effort has been invested in identifying and predicting the relationship between the genetic code and phenotype of the plant as understanding this relationship facilitates phenotypic engineering, a process in which the genomic information is altered to obtain traits desired by human needs. High-throughput technology has boosted our understanding of information flow from the genetic code to phenotypes. In the next section, we give an overview of the different types of omics data and the online resources containing these data in model plants.

1.3 Sources of large-scale omics datasets

The entire genome of an organism can be sequenced in less than a single day thanks to massively parallel sequencing technologies, so-called next-generation sequencing (NGS). We will refer interested readers to the history of NGS development and its connection to the traditional Sanger sequencing [2-5]. Eventually, NGS platforms generate millions of small fragments of (c)DNA, so-called short reads. Piecing together this fragmented information to recover the original information contained in the molecules provided to the sequencing

machines, is referred to as the assembly process. To this end, short reads are mapped to a reference genome, if this is available [6]. Alternatively, the reads can be assembled without prior genome information using *de novo* assembly methods [7]. NGS can be applied to sequence the entire genome or transcriptome, or constrained to target regions of interest. For example, active chromatin regions can be captured using ChIP-seq or ATAC-seq protocols and subjected to sequencing [8].

Availability of cost-efficient NGS technologies and highly efficient software to process the data obtained from NGS machines, along with rapid development and enormous progress in other omics fields to profile proteins and metabolites have made it possible to obtain large-scale molecular data within a tissue, and more recently even at single cell level [9]. Data at different molecular levels can be generated using those technologies. The profiling of genomes, RNA, metabolites, proteins, and epigenetic marks on DNA or histones is referred to as respectively genomic, transcriptomic, metabolomic, proteomic, and epigenetic profiling. Processing those data to extract meaningful and unbiased biological results requires specialized bioinformatics expertise (Figure 1.3). Many methods and tools have been specifically developed to analyze each type of high-throughput data as it is not trivial handle to the biological and technical noise that goes together with the generation of omics data [10, 11]. Below we will review the most commonly available large-scale and high-throughput omics data types for plant species.

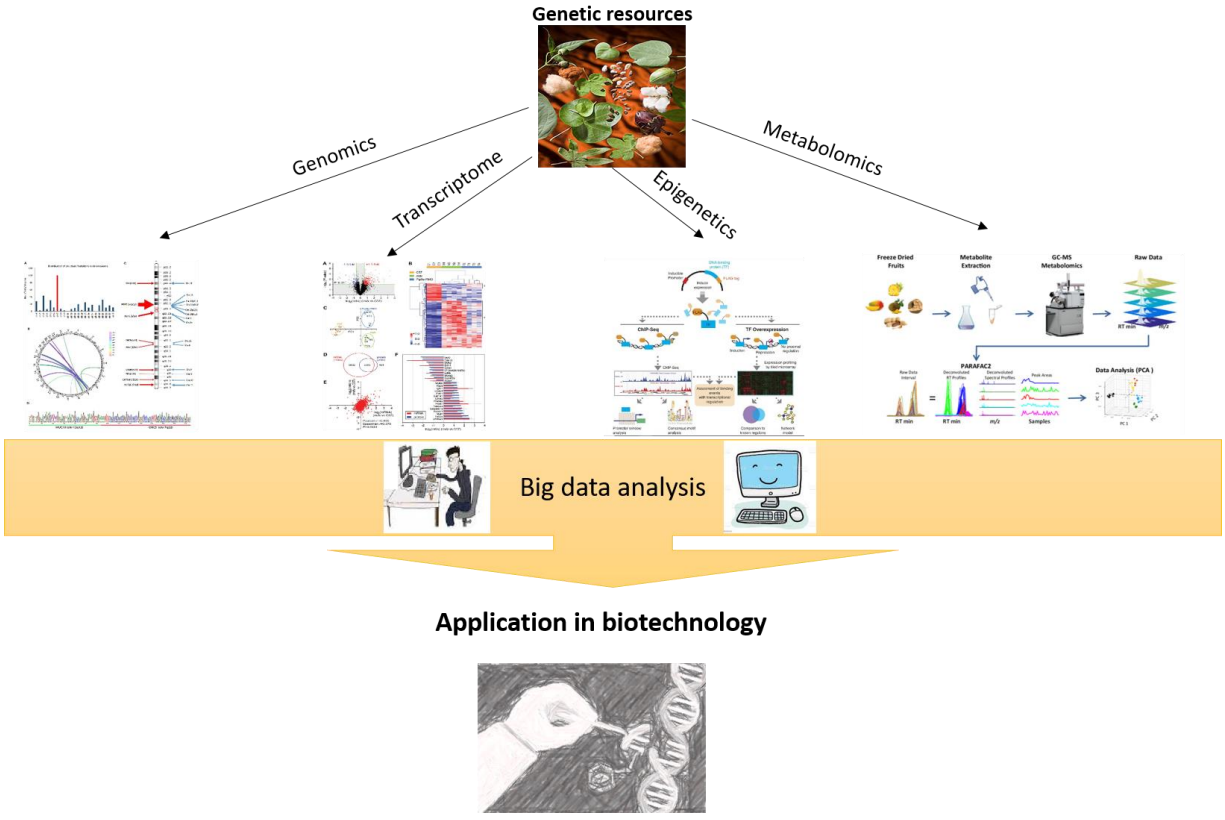


Figure 1.3 The Figure shows how information from different omics sources can be used to increase our fundamental understanding of plant responses. Such understanding can pave the way for biotechnological programs that focus on improving crop performance, quality, or response to stress. Sources [12-16]

1.3.1 Genomics data

Each organism has unique genomic information that differs from the genomic information of other individuals of the same species by differences in single nucleotide polymorphisms (SNP), indels (insertion or deletion), copy number variations, chromatin rearrangements, etc. Identifying those differences within and between species is the focus of genomics studies as it can explain differences in observed phenotypes. To facilitate uncovering those differences, more than 600 plant reference genomes have been sequenced and made available in public repositories [17]. Availability of the reference genome would facilitate generating and processing other high-throughput data at the DNA level (genomics data). Genomics data are generated to characterize the content, structure, organization, and dynamics of DNA. Using the genomics data and genome sequence, hidden information such as functional regions can be uncovered [18]; adaptive responses that took place during evolution (natural selection) can be reconstructed [19]; and the adaptive behavior of an organism in response to environmental constraints or artificial selection can be predicted [20]. Furthermore, within-species and cross-species comparisons and comparative genomics can establish which genes are common to all plants or specific to a particular species, or they can even determine differences at the single nucleotide level [21, 22]. For example, the difference in a single nucleotide in the regulatory region in the DNA sequence can change the binding site of a transcription factor (TF). As a result, individuals with a different nucleotide in that position might undergo different regulation of a specific biological process [23]. Besides cultivars, remarkable progress in NGS technologies has enabled the characterization of wild types and landraces to study the genomes among a large number of diverse individuals. Regarding crop improvement, genome information can facilitate exploring new genes and traits for potential application in breeding and biotechnological programs. Those resources can be used in advanced crop improvement programs for making critical decisions, i.e., genomics selection or identifying the candidate regions for targeted gene editing [24]. Many tools and databases have been developed for visualizing, mining, and analyzing plant genomics data such as Ensembl Plants [25], PLAZA [26], Phytozome [27], and many more.

1.3.2 Epigenetics data

Epigenetics aims at studying the reversible modifications on the DNA without altering the underlying DNA sequence. Such modifications include for instance DNA methylation, histone modifications, chromatin accessibility, and transcription factor (TF) binding sites [28, 29]. Epigenetic modifications influence the accessibility of chromatin to the transcriptional machinery and hence result in transcriptional activation or gene silencing. Hence, epigenetics data is advancing our understanding of gene expression regulation and complement expression data. The levels of epigenetic features (e.g., the level of DNA methylation and chromatin accessibility) can be measured using NGS technologies by capturing the regions of interest, followed by sequencing [30-32]. Below different types of epigenetics data are described in more detail.

1.3.2.1 DNA methylation

DNA methylation occurs at the 5′ position of cytosine which is conserved in plants and mammals. This methylation can be profiled using Bisulfite sequencing (BS-seq) or other advanced NGS-based methods [33]. Plant DNA methylation is a dynamic process mediated by methylating and demethylating enzymes and regulatory factors, occurs in cytosine sequence contexts: CG, CHG, and CHH (H represents A, T or C), mostly in heterochromatin regions [34]. DNA methylation is the main player in transposon silencing, gene expression regulation, chromosome interactions, and genome stability [35]. As a result, the DNA methylation state is crucial for development and environmental stress responses and can serve as environmental memory for plants to adapt to environmental changes [36-38].

1.3.2.2 Chromatin accessibility

The structure of DNA at the chromatin level is highly dynamic and this property influences the interaction of the transcriptional machinery with binding regions on the DNA. The chromatin accessibility landscape can be profiled to determine the active regions of the DNA to which nuclear macromolecules or TFs can bind. The genome-wide chromatin accessibility landscape can be quantified by enzymatically (DNase I, MNase, Tn5 transposase) or chemically (Formaldehyde) fragmenting the genome and isolating either the accessible or protected locations which are subsequently sequenced [31]. ATAC-seq is one of the commonly used methods for profiling accessible chromatin. It uses a hyperactive Tn5 transposase to insert Illumina sequencing adaptors into accessible chromatin regions. Publicly available ATAC-seq data provide a great resource to locate putative TF binding sites [39].

1.3.2.3 TF binding sites

Chromatin immunoprecipitation (ChIP-seq) assays are powerful methods to capture and sequence the DNA binding sites of a specific TF at the genome level (ChIP-seq). Unlike open chromatin assays, ChIP-seq can locate TF binding sites with higher resolution. To rapidly and inexpensively interrogate large numbers of TFs, DNA affinity purification sequencing (DAP-seq) was introduced that uses in-vitro-expressed TF to interrogate naked gDNA fragments to establish binding locations (peaks) and sequence motifs [40]. DAP-seq has been successfully applied to uncover the regulatory binding landscape of several plant species [40-44]. However, like other high-throughput methods, ChIP-seq and DAP-seq come with technical limitations including limited capacity to detect the binding sites of poorly expressed proteins (TF), and the dependence on the availability of specific antibody against the TF. Despite those limitations, ChIP-seq and DAP-seq have become indispensable tools for studying gene regulation and epigenetics.

1.3.3 Transcriptome data

NGS technologies can also be applied to profile the expression level of genes by determining the amount of RNA that is produced in a specific tissue or cell. The amount of RNA is used as

a proxy of the levels of proteins present in a cell. Hence, determining the differences in gene expression between individuals can help explain differences in observed phenotypes. Transcriptome analysis is commonly used to study the adaptive behavior of plants to environmental or developmental signals. In addition, it has been widely used to annotate reference genomes. If accompanied by genomics data, the transcriptomics data will provide valuable information to link changes in gene expression to genomic changes (aka eQTL mapping) [45, 46]. Two high-throughput methods coexist to quantify the amount of (m)RNA transcribed for every gene in the genome: microarrays and RNA-seq. Microarray is a hybridization-based approach for high-throughput expression analysis that only profiles predefined genes through probe hybridization. Specifically, a microarray is a laboratory tool that consists of a collection of microscope probes attached on a solid surface. In the hybridization step, strands (stranded cDNA) with the complement sequence of the probe will bind to the corresponding probe. Hybridization can be detected by linking probes with fluorescent or radioactive labels to be quantified. Although microarrays are increasingly being replaced by RNA-seq, it was commonly used for transcriptome profiling due to its cost efficiency and its standardization for model organisms, such as Arabidopsis. In addition, large-scale microarray databases that have been produced over the years are still being used in meta-analyses and comparative studies. Overall, it is considered an accurate, cost-effective, sensitive, and reliable technique, especially under high sample size conditions.

RNA-seq has emerged as an alternative method to profile gene expression which is very similar to genome sequencing and allows for full sequencing of the whole transcriptome [47]. Briefly, by breaking the cDNA into fragments, the nucleotide content of each fragment, i.e., sequencing reads, can be determined using next-generation sequencing machines. After prepossessing and quality filtering, the reads are mapped to the corresponding reference genome for gene expression quantification. Compared to microarrays, RNA-seq identifies more modulated protein-coding genes, allows for detecting splice variants, and non-coding transcripts, and provides a wider quantitative range of expression level changes [48, 49]. Due to the drop in cost and other advantages, it has become the standard and preferred technology over microarrays for gene expression profiling.

1.4 Represent biological information as a network

Powerful and scalable technologies enabled the generation of genome-wide datasets in genomics, epigenetics, metagenomics, proteomics, metabolomics, from which the interaction among molecular entities can be inferred [50-52]. Interaction information is typically represented as a graph or biological network consisting of nodes and edges. A network is a collection of nodes linked by edges. In the biological context, nodes represent molecular entities (genes, transcripts, proteins, metabolites, etc.), while edges represent pairwise relations (physical interaction, similarity in function or expression, activation or suppression, etc.) between the entities represented by the nodes. A set of attributes or features can be associated

with nodes or edges in the network to provide extra information. Some important terminologies related to network analysis used in this thesis are explained below.

1.4.1 Exploiting networks

In a **directed network**, the edges are valid only in one direction (direction from the source node to target node) while in an **undirected network** the direction is valid both ways (source and target nodes are interchangeable). In a **weighted network**, a confidence score (weight) is associated with each edge reflecting the degree of dependency between source and target (to what extent they are functionally dependent), or the degree of confidence on the inferred edge using prediction methods (e.g., consistency in detection or prediction from different methods or datasets). In contrast, the edges in an unweighted network are binary. This reflects that there is no reason to believe one interaction is stronger or more reliable than other edges. In some cases, the weighted network is converted to an unweighted one for computational convenience (this is also true for directed and undirected networks). **Negative weights** (e.g., derived from a negative correlation between the expression profiles of two genes) are normally converted to positive weights as popular network algorithms expect positive values as edge weights. For the sake of simplicity, we simplified networks by filtering self-loops and multiple edges between two nodes.

Neighbors (first-order neighbors) refer to two nodes in the network that are connected by at least one edge. **Second-order neighbors** consist of all nodes that are connected to the node of interest by one intermediate node. Similarly, higher-order neighbors (third-order, fourth-order, etc.) are defined by the number of intermediate nodes needed to connect the two nodes under consideration. A **path** is a sequence of edges which joins a source to a target node. **Length** of a path is the number of edges in the path that are needed to reach the target node from the source node. A network is **connected** if for each node at least one path exists to any of the other nodes in the network; otherwise, the network is disconnected and consists of a set of sub-networks or connected components. The **degree** of a node is the number of edges connecting that node to its adjacent nodes. In a directed graph, the **outdegree** is the number of edges leaving the node of interest, while the **indegree** is the number of incoming edges ending at the node of interest. A **hub** is a node with a high degree. A network can be represented as an **adjacency matrix**, i.e., square matrix with binary entries (or edges weights) indicating whether an edge exists between nodes in row ‘i’ and column ‘j’. The **adjacency** matrix is the algebraic representation of a graph and can easily be used to extract structural information.

1.4.2 Homogeneous biological networks

Different types of biological networks exist that capture interactions at different molecular layers. **Homogeneous networks** typically capture one specific level of information. For instance, protein interaction networks correspond to interactions among proteins, regulatory networks contain interactions between TFs and their targets, metabolic networks represent

interactions between metabolites and enzymes, and coexpression networks represent mutual similarity in expression profiles between genes (Figure 1.4).

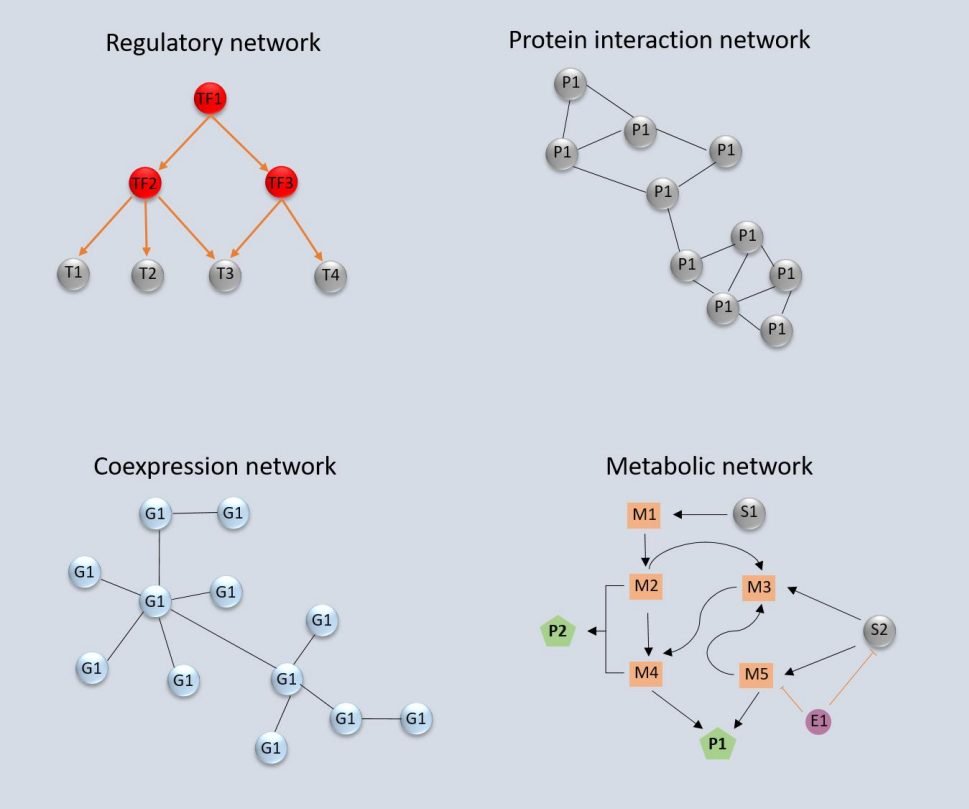


Figure 1.4 Examples of biological networks. Nodes represent biological units, and edges represent the type of connection among nodes. An edge can be directed, and its associated weight indicates its reliability. Regulatory networks are an example of directed networks that represent the flow of signals from TFs to targets. Protein interaction networks are often undirected and indicate the physical interaction among proteins. Coexpression networks are inferred from expression data with undirected edges. Metabolic networks are more complex including directed, undirected, activator, and suppressor edges.

1.4.2.1 Gene regulatory networks

Gene regulation is a complex process in which TFs bind to the regulatory regions of their target genes. These regulatory regions or TF binding sites are non-coding genomic regions that are either located in the vicinity of the target genes (cis-regulatory elements) or many kilobases away from their target genes (distal regulatory elements). Next to transcriptional regulation, many additional processes such as phosphorylation, and other protein modifications such as carbonylation, glycosylation, and S-nitrosylation affect signaling [53, 54]. These different regulatory mechanisms do not act independently. The crosstalk between these mechanisms forms an interconnected regulatory system that senses and integrates endogenous and environmental signals and converts them into altered gene expression. The complex regulatory

connections from regulators to their target genes can be summarized in a gene regulatory network.

PlantTFDB (http://planttfdb.cbi.pku.edu.cn/) provides a comprehensive, high-quality resource for both plant transcription factors (TFs) and their targets for 165 plant species. It includes: (i) a set of high-quality, non-redundant TF binding motifs derived from experiments; (ii) multiple types of regulatory elements identified from high-throughput sequencing data; and (iii) regulatory interactions curated from the literature and inferred by combining TF binding motifs and regulatory elements [55, 56].

1.4.2.2 Protein interaction network

Understanding protein interactions is crucial to unveiling the mechanisms underlying plant response to internal and external signals. The protein interaction network is a graph representation of the physical binding of proteins to each other in the cells derived from protein binding assays [57, 58]. The edges in the protein interaction network can represent stable (formed in protein complexes) or transient (to make protein modifications) bindings [59]. Unlike stable interactions, transient or dynamic interactions are more challenging to identify by experimental or computational methods [60]. Protein interactions can be directed or undirected, with the majority of them being indirect. Protein interaction networks have only been experimentally assessed for model species such as Arabidopsis, and even for those model species, only a subset of the protein interactions has been characterized experimentally [61]. Most interactions are extrapolations of interactions from other model organisms (interologs) [62, 63]. For instance, Geisler-Lee et al. [64] identified 19,979 interactions for 3617 proteins in *Arabidopsis thaliana* by aligning with *S. cerevisiae, Caenorhabditis elegans, Drosophila melanogaster*, and *Homo sapiens*. In addition, protein interactions are often biased toward well-studied genes, which often have high numbers of associated functional annotations ("multifunctional") [65]. Therefore, the protein interaction networks from public databases such as String [66] must be used and interpreted with care, especially for non-model plant species.

1.4.2.3 Metabolic network

A metabolic network comprises enzymes and metabolites as nodes and chemical reactions between nodes as edges which represents the most important physiological and biochemical processes in a cell [67]. Unlike protein interactions, the majority of metabolic interactions are directed, starting from precursors to intermediate and end-point metabolites. As metabolites represent the endpoint of gene-environment interactions, metabolic interactions are considered the most reliable source of interactions between molecular entities. Hence, metabolic networks are at the heart of system biology methods and have been extensively used to complement, validate, and interpret different omics datasets [68-70]. Upon genome sequencing, reconstructing the metabolic network of the sequenced genome is common practice in all organisms. This is because metabolic reactions are highly conserved among species and the

available information can be extrapolated to a newly sequenced genome [71, 72]. However, this extrapolation is only possible for highly conserved metabolic processes, which only cover a small set of genes. Identifying lineage and species-specific metabolic processes, and changes in the rate and levels of metabolites in response to internal and external signals is a challenge. As a result, a metabolic network covers relatively fewer genes than a network derived from the transcriptome and proteome [73]. Several databases and tools have been developed for metabolic pathway analysis in plants such as KEGG [67], Plant Metabolic Network or PMN [71], MetaCyc [74], and MapMan [75].

1.4.2.4 Gene coexpression network

Gene coexpression networks are inferred from gene expression measurements using microarray or RNA-seq data. Each gene is represented by its expression profile (a vector containing the expression of the gene in many different samples, genotypes, or conditions). Subsequently, the pairwise similarity between these gene expression vectors is determined. In the simplest form, the similarity is calculated using the absolute value of the pairwise Pearson correlation coefficient. Those pairwise gene-gene similarities are converted into an undirected network, after filtering the weak correlations. The resulting network consists of edges with associated correlation values greater than a user-specified cutoff. In this network, nodes correspond to genes and edges represent the mutual similarity in expression profiles between the connected nodes. Many improvements have been proposed to avoid hard-thresholding [76], to alleviate the impact of outliers [77], and to reduce the number of false associations [78]. Taken together, the choice of an appropriate similarity measure and some careful preprocessing of the expression data is the key to constructing a reliable coexpression network [79, 80]. Like other networks, statistical machine learning methods can be applied to coexpression networks to associate genes of unknown function with biological processes, prioritize candidate genes, infer gene regulatory interactions, and to compare properties of the gene groups in the network. Similarly, a coexpression network can be used for clustering nodes into groups such that nodes belonging to the same group are functionally related and share common properties. For most plant species, coexpression networks have been constructed and sorted in repositories such as PlaNet [81], ATTED-II [82], and AraNet [83].

1.4.3 Heterogeneous interaction network

Each homogeneous biological network is derived from their cognate experimental data sources to capture interactions at a specific molecular layer. Unfortunately, such experimental approaches are limited in their breadth and resolution for technical and budgetary reasons, which makes the inferred biological networks incomplete [30, 84, 85]. Integrating different data sources can increase the breadth and depth of these homogeneous networks [86, 87]. For instance, coexpression information can be used to support inferred protein interactions [88] or to infer TF target relation [89]. Alternatively, different levels of biological networks can be combined into a single heterogeneous interaction network. In the word of Aristotele: the whole

is greater than the sum of its parts, indeed an interaction network built from different information layers can better explain systematic mechanisms that control complex traits, and it can predict the behavior of biological systems with higher precision [87, 90, 91]. However, constructing interaction networks is not trivial. To this end, interactions from different experiments performed in different conditions are typically combined. This often results in networks that are over-connected but at the same time incomplete and may not accurately represent the 'biological truth'. Therefore, sophisticated biological knowledge and advanced computational methods are required to extract relevant information from such noisy networks [89].

1.5 Integrative analysis in plants systems biology

As mentioned before, the increase in high-throughput sequencing data from different layers and availability of prior molecular information allowed researchers to move from studying the cellular response at a single omics level toward exploring the behavior of an entire system at different molecular layers (DNA, RNA, proteins, etc.) (Figure 1.3). However, integrating the heterogeneous prior molecular information and the data generated from high-throughput techniques requires systematic approaches, which is proven to be a difficult task [92]. Advanced and well-defined methods are needed to ensure accurate large-scale data analysis and extract meaningful results by integrating multiple types of quantitative and qualitative molecular data. Two classes of approaches for multi-omics data analysis are relevant to this dissertation: i) exploratory data analysis (post-analysis data integration), and ii) model-based data integration (Figure 1.5).

The exploratory data analysis refers to an intuitive way of first analyzing each omics dataset in isolation, and then stitching the key feature results together in a post-processing step, to search for patterns that are supported by all complementary datasets. More specifically, it aims to find the underlying relationship between datasets in a descriptive manner or to predict a certain response using one or more explanatory datasets [93]. Approaches based on the aforementioned sequential analysis depend on in-depth biological insight. New candidate genes involved in complex phenotypes can be introduced using exploratory data analysis approaches and can be provided to wet-laboratory researchers for further investigation [94]. Prioritizing candidate genes and suggesting the mechanism of a gene's effect on a particular phenotype requires knowledge from literature and system-level information of all molecular entities. For example, transcriptome, proteome, and metabolome data have been explored to identify stage-specific biomarkers for biological processes in several plant species [95-98].

Exploring prior molecular knowledge, omics datasets, and phylogenetic data is also a common approach. For example, De Smet et al [99] associated gene function with gene copy number state by exploring sequence information and expression datasets. Similar studies have explored the combination of molecular prior information, sequencing, expression, and epigenetics data to provide deeper insight into genes function and uncover genetic components that shaped observed evolutionary patterns in plants [100-103].

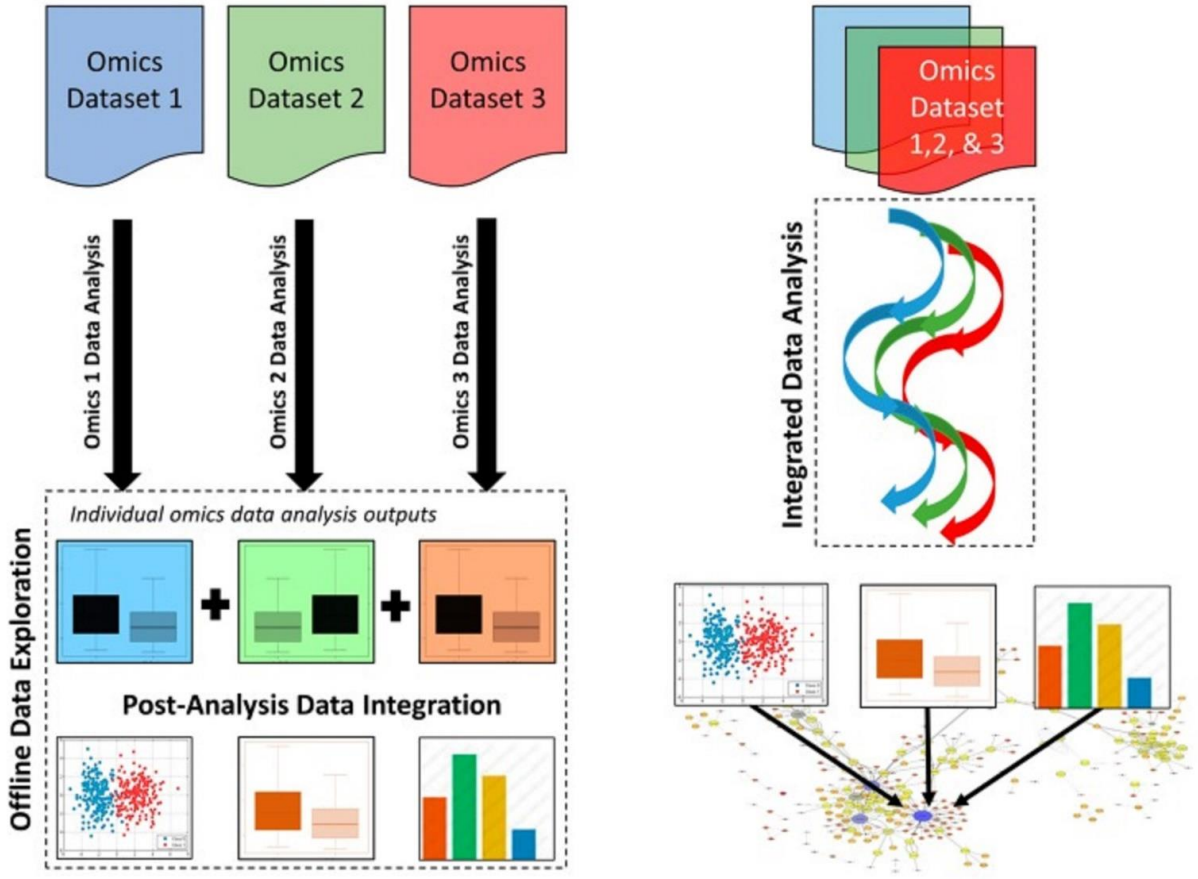


Figure 1.5 Principal differences between exploratory data analysis (left) and model-based data integration (right). In the exploratory data analysis, each dataset is analyzed in isolation, and conclusions are made based on collective information from different levels. In model-based data integration, data are integrated into a single model, typically a network, to be explored at once. Source [73].

Although it is possible to explore each omics dataset separately, the crosstalk between the molecular entities at different molecular layers cannot be properly assessed by analyzing each dataset in isolation [104]. Model-based integration methods are algorithmically more complicated but can only be applied if all layers of information can be captured in a single appropriate model. Different integrative methods rely on the use of a network model to represent different information sources. This is because networks provide an intuitive way to represent biological information and molecular interactions. Networks can be used or explored with graph-based algorithms and mathematical models that have been developed or adopted from other domains [105]. The way a network is used largely depends on the research question at hand. Their applications are widespread, ranging from inferring gene or protein function through the guilt by association principle [106]; predicting interactions among biological units [107]; prioritization of gene lists [108], and inference of active subnetworks that could best explain the observed phenotype of interest in a specific experiment [17]. Because of their popularity, we will review some network-based methods that are used in plant systems biology studies.

A very popular approach is the representation of single omics data as a network to identify functional relationships between molecular entities. This is achieved by assuming that genes associated or interacting in a network are more likely to share function. This process is known as guilt by association (GBA) which relies on module inference. GBA has been widely used in biology to assign a function to unknown genes [109, 110]. Practically, GBA can also be used to prioritize genes that control phenotypes of economic importance, such as tolerance to environmental stress and resistance to disease using information available on known genes [111, 112].

With the rapid accumulation of large expression data for many plant species, coexpression network analysis is one of the most popular methods to detect gene modules that show high transcriptional coordination across a variety of experimental conditions [79]. Coexpression analysis has been successfully applied to infer transcriptional regulatory networks for nitrogen metabolism [113], to identify and associate new genes to metabolic pathways [114], to transfer gene functional annotation from model plants to non-model plants using comparative coexpression analysis across plant species with homology information [115], and to prioritize candidate SNP-associated genes [116].

Gene regulatory networks (GRN) are another popular and powerful tool that reflects a blueprint of the molecular interactions underlying plant responses. GRN has been employed to identify new regulatory connections for metabolic pathways [117] or in response to environmental and intrinsic signals [118, 119].

Despite being powerful, protein interaction networks (PIN) are less characterized compared to gene regulatory networks and coexpression networks, due to the limited availability of cost-effective and statistically powerful methods for characterizing protein functions and interactions in plants [120].

However, network-based methods that rely solely on a single level of information and that do not use any additional information have a poor performance to prioritize genes for a specific trait or biological process [121, 122]. This is because the network is treated as a static reference, which does not reflect the dynamic nature of molecular networks [111]. In addition, the association between GBA and some biological network properties, e.g., a “scale-free” degree distribution is not well explored [121]. In some cases, prediction performance could be attributed entirely to node degree effects [65]. However, networks constructed from single omics data are helpful in assessing the quality of other networks, and the performance of computational methods, by assuming that a high-quality network or computational results should map well onto known gene function information and other molecular information [121, 123].

Studies that integrate more than one level of information go a step beyond. A pioneer algorithm for this class is ENDEAVOUR [124] which calculates gene-wise statistics from heterogeneous genome-wide data sources (including molecular interactions) and ranks genes according to their similarity to known genes involved in the biological process under analysis. For example,

Lysenko et al. (2013) identify candidate virulence genes in the fungus *Fusarium graminearum* by integrating gene coexpression, protein-protein interactions, and sequence similarity [125].

1.6 Network-based omics data interpretation

Networks derived from publicly available molecular interaction data (PIN, GRN, GCN, etc.) are often used as a scaffold or roadmap to interpret an in-house generated omics dataset. This application is referred to as **network-based data interpretation** [90, 108]. Analysis of omics data reveals a list of genes (or other biological entities) of which the behavior or state is significantly altered between individuals or conditions, e.g., a list of genes that are differentially expressed, genes with different copy number states, differentially methylated genes, etc. Although gene lists identify entities that are activated or inhibited, the direct casual and functional associations between the entities in the gene list and the observed phenotype cannot be established [89, 126]. Mapping the gene list on a network prior can help to systematically explain how listed genes interact to drive the phenotype of interest and recover the unobserved mechanisms at upstream or intermediate molecular levels [69, 73]. By leveraging candidate genes with known interaction information, spuriously identified candidate genes can be removed as they will not be part of the subnetworks. In addition, genes relevant to the process of interest that were not measured due to technical limitations or their biological properties, are indirectly identified by being part of a connected subnetwork to which also many of the candidate genes belong. Hence, these subnetworks reflect the pathways of relevance triggered by the studied process. This is of particular importance for complex traits such as plant response to stress in which many genes with small effects interact in a complex system to drive the response [112]. The advantage of subnetwork analysis is that the backbone interaction network used to drive the analysis contains next to well-annotated connections between genes also less well-documented edges, derived from the large body of publicly available omics datasets. This allows to also assigning genes without GO annotation to a process of which their neighbors involved. It is also possible to infer causality relationships by identifying the upstream programs of the modules to describe the cause of observed changes in molecular entities behavior.

Advanced methods have been developed to overlay the omics data on the interaction scaffold. Some clustering-based methods search for subnetworks in the weighted interaction network using (multi)omics data. The weight on the edges can be derived from either the node properties (such as expression, copy number, etc.) or from the structure of the graph. The former uses node features and assigns a higher weight to the edges that connect nodes sharing common features. For example, an edge between two genes that have a similar copy number or mutation state between different individuals gets a high weight. Examples of those methods include iOmicsPASS [127], and SteinerNet [128]. The weight of an edge can also be derived from the structure of the graph. Two nodes are more similar and hence will be assigned a higher weight if they are structurally close. Examples of algorithms that incorporate the structure of the network to assign weight to the edges of the interaction network are nuChart [129] and SNF [106]. Ignoring the edge directionality and treating the interaction network as an undirected

graph is the main limitation of those methods. PheNetic [90] is unique as it can incorporate both the weights and the directionalities of the edges. In PheNetic, candidate genes identified through an omics study are mapped on the interaction network, and subsequently, the algorithm searches for subnetworks that connect as many candidate genes as possible using the least number of edges. The edge weights can be derived from both features on the nodes (coming from omics data) and the structure of the graph. Features on the nodes allow for integrating data from a specific omics study with the network scaffold [127, 128], while the structure of the network is handy to alleviate the effect of hubs in the downstream analysis (Figure 1.6) [90, 108, 130].

Network-based interpretation is powerful but also comes with certain assumptions. Casting the data in a way they can be modeled together is not trivial and also not feasible for each application. Other limitations of network-based methods are discussed in the last chapter.

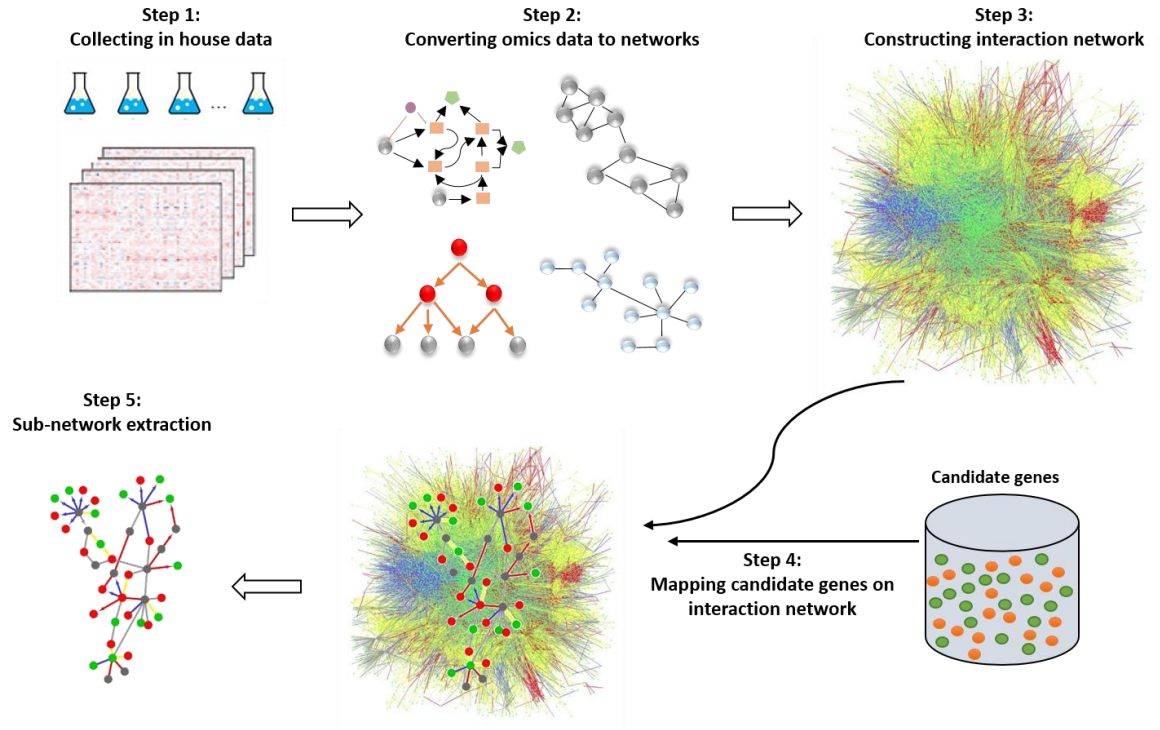


Figure 1.6 The flow of mapping candidate genes resulting from an omics study on the interaction network. First, relevant datasets are collected, processed, and converted to molecular networks at different molecular layers. Then, the molecular networks from different sources are combined in an integrated interaction network. The gene list containing candidate genes from an omics study is mapped on the interaction network.

1.7 Integrate sequence data with other omics data

As the amount of generated sequence data grows, new biological questions can be answered by mining thousands of genomic datasets that span different technological platforms, genotypes, tissues, and developmental stages (Curtis Huttenhower, Oliver Hofmann, 2010). Due to its

complexity, gene regulation is a good candidate to be investigated by such large-scale datasets. Upon binding of the TFs to cis-regulatory elements (CREs), the expression of the target genes can be activated or repressed in response to developmental and environmental signals. However, each gene can be regulated by multiple TFs and each TF can target thousands of downstream genes which form a complex system of regulatory networks. Uncovering the details of this regulatory network and CREs identification is essential for understanding the cell response to the internal and external stimuli and can facilitate synthetic biology and metabolic engineering [131, 132]. Integrating omics datasets with other molecular data to identify transcriptional regulatory mechanisms is an active research subject [133]. The pioneering algorithms to integrate molecular information and omics datasets to infer transcriptional regulation are GRAM [131], DISTILLER [134], SEREND [135], DeMAND [136], SIRENE [137], and GeneXPress [126]. We refer interested readers to reviews on integrative methods to infer gene regulatory interaction [89, 132].

Discovering TF binding motifs is a key step in inferring gene regulatory networks [43, 138, 139]. The most reliable way to discover binding sites is through experimental approaches [40, 140]. Charting the entire regulatory landscape is becoming a routine practice thanks to the recent progress in genomics. This includes experimental high-throughput approaches to characterize the binding sites for a specific cluster of TFs through ChIP-seq and DAP-seq, open (accessible) chromatin regions through DNase-seq and ATAC-seq, chromatin interacting sites using Hi-C, and epigenetic marks such as DNA methylation and histone modification [30-32, 40, 141].

Unfortunately, such experimental approaches are limited in their breadth and resolution for technical and budgetary reasons [30, 31, 40]. In addition, because the data generation process is noisy and assumptions are made during data analysis, the final binding regions cannot be located exactly but often cover hundreds to a thousand bp. In addition, for non-model organisms, the application of the experimental approaches is limited by the lack of genome information [30, 141]. Computational approaches can complement experimental approaches in order to validate or to pinpoint the accurate location of the experimentally identified TF binding sites. Phylogenetic footprinting is a comparative approach that integrates information from closely related species to overcome spurious predictions in binding site identification [142]. Functional binding sites should be well conserved in closely related species as most mutations in functionally important regions mostly interfere with their functionality and would be pruned through selective pressure over the time [18]. Therefore, conserved regions in the homologs from closely related species can be candidate binding sites. Among the tools for phylogenetic footprinting, BLSSpeller is unique in the sense that it explores the full sequence space and allows for alignment-free search by easing the constraint that a binding site must be well aligned [143]. The computational prediction can be integrated with a wide range of genomics and functional molecular data to identify the best candidates for further investigation.

1.8 Research goals and outline

This dissertation is a compilation of collaborative research papers in which we applied integrative analyses to answer biological questions on specific plant systems. We specifically focus on combining expression data with other information sources. We hereby made use of both exploratory data analysis and network-based approaches.

Chapter 2 studies the fate of genes following an ancient whole gene duplication (WGD) in sacred lotus. Lotus is the only species that experienced a single WGD after diverging from the basal-most angiosperm, *Amborella*, and for which a well-assembled genome is available. We studied the expression, epigenetic regulation, and functional constraints that shape gene fates following a WGD. This is important to provide deeper insight into genes function as gene function is not independent of gene copy number state. In addition to standard genomics analysis, a network-based method was adopted to explore and compare the behavior of genes with different duplication states in more detail. More specifically, we tested the gene balance hypothesis (GBH) on the network. The GBH states that the fate of duplicates in transcription factors or kinases after WGD is largely shaped by their hub-like properties in the network and those genes can only be retained or deleted together with their "interactors". This hypothesis was tested in the physical interaction network inferred from Arabidopsis. We also classified duplicate gene pairs based on their properties in the coexpression network and linked the expression divergence of these gene pairs with their divergence in sequence and function.

In chapter 3 we employed an Arabidopsis mutant line carrying a loss-of-function mutation in the *BRI1* gene (involved in brassinosteroid (BR) signaling pathway) along with three activation-tag suppressor lines that were able to partially revert the BRI1 mutant phenotype (dwarf) to a wild-type phenotype. The "activation tagging" falls into the inserted DNA category that involves generating random genomic insertions of a transgene that contains transcriptional enhancers capable of increasing the expression of nearby genes. Mutants are powerful in functional genomics to study the function of genes. However, assessing the function of every gene through mutations is not possible due to the existence of gene homologs which create functional redundancy and difficulty in technical processes in developing mutant lines. Hence, a systematic approach is required to understand the components of the process of interest, here BR signaling, using information on the mutant lines. Using an integrative approach, in which the expression data of the four BR signaling mutant lines and the wild-type was combined with the prior molecular network could identify a more holistic view on BR signaling and its crosstalk with other hormones signaling. A weighting system based on expression data was adopted to make the interaction network specific to the conditions used in the study. The results demonstrate the advantage of using network-based analysis over the traditional methods such as differential expression and GO enrichment analysis. Differential expression analysis of one mutant compared to the wild-type could explain part of the observed phenotype only.

Chapter 4 shows how regulatory regions can be predicted using sequence information from closely related species and how datasets from different omics levels can be employed to validate the results. First, BLSSpeller was redesigned to enable the analysis of larger datasets. Then, it was used to identify conserved motifs in *Zea mays* in a comparative genomics setting. Combining predictions with available genomics and functional data allowed further elucidating transcriptional regulation in *Zea mays*.

References

1. Noor E, Cherkaoui S, Sauer U: **Biological insights through omics data integration.** *Current Opinion in Systems Biology* 2019, **15:**39-47.
2. Barba M, Czosnek H, Hadidi A: **Historical perspective, development and applications of next-generation sequencing in plant virology.** *Viruses* 2014, **6:**106-136.
3. Goodwin S, McPherson JD, McCombie WR: **Coming of age: ten years of next-generation sequencing technologies.** *Nature Reviews Genetics* 2016, **17:**333-351.
4. Liu L, Li Y, Li S, Hu N, He Y, Pong R, Lin D, Lu L, Law M: **Comparison of next-generation sequencing systems.** *Journal of Biomedicine and Biotechnology* 2012, **2012**.
5. Schuster SC: **Next-generation sequencing transforms today's biology.** *Nature methods* 2008, **5:**16-18.
6. Pop M: **Genome assembly reborn: recent computational challenges.** *Briefings in bioinformatics* 2009, **10:**354-366.
7. Baker M: **De novo genome assembly: what every biologist should know.** *Nature methods* 2012, **9:**333-337.
8. Mamanova L, Coffey AJ, Scott CE, Kozarewa I, Turner EH, Kumar A, Howard E, Shendure J, Turner DJ: **Target-enrichment strategies for next-generation sequencing.** *Nature methods* 2010, **7:**111-118.
9. Rich-Griffin C, Stechemesser A, Finch J, Lucas E, Ott S, Schäfer P: **Single-cell transcriptomics: a high-resolution avenue for plant functional genomics.** *Trends in plant science* 2020, **25:**186-197.
10. Grabowski P, Rappsilber J: **A primer on data analytics in functional genomics: how to move from data to insight?** *Trends in biochemical sciences* 2019, **44:**21-32.
11. Reel PS, Reel S, Pearson E, Trucco E, Jefferson E: **Using machine learning approaches for multi-omics data analysis: A review.** *Biotechnology Advances* 2021, **49:**107739.
12. Ban Y, Fischer JV, Maniar KP, Guo H, Zeng C, Li Y, Zhang Q, Wang X, Zhang W, Bulun SE: **Whole-genome sequencing and target validation analysis of müllerian adenosarcoma: a tumor with complex but specific genetic alterations.** *Frontiers in oncology* 2020, **10:**538.
13. Roberts TC, Johansson HJ, McClorey G, Godfrey C, Blomberg KEM, Coursindel T, Gait MJ, Smith CE, Lehtiö J, El Andaloussi S: **Multi-level omics analysis in a murine model of dystrophin loss and therapeutic restoration.** *Human molecular genetics* 2015, **24:**6756-6768.
14. Ahmad S, Wei X, Sheng Z, Hu P, Tang S: **CRISPR/Cas9 for development of disease resistance in plants: recent progress, limitations and future prospects.** *Briefings in Functional Genomics* 2020, **19:**26-39.
15. Turkarslan S, Peterson EJ, Rustad TR, Minch KJ, Reiss DJ, Morrison R, Ma S, Price ND, Sherman DR, Baliga NS: **A comprehensive map of genome-wide gene regulation in Mycobacterium tuberculosis.** *Scientific data* 2015, **2:**1-10.
16. Khakimov B, Mongi RJ, Sørensen KM, Ndabikunze BK, Chove BE, Engelsen SB: **A comprehensive and comparative GC–MS metabolomics study of non-volatiles in Tanzanian grown mango, pineapple, jackfruit, baobab and tamarind fruits.** *Food chemistry* 2016, **213:**691-699.

17. Kersey PJ: **Plant genome sequences: past, present, future.** *Current opinion in plant biology* 2019, **48:**1-8.
18. Hardison RC: **Comparative genomics.** *PLoS biology* 2003, **1:**e58.
19. Anderson JT, Willis JH, Mitchell-Olds T: **Evolutionary genetics of plant adaptation.** *Trends in Genetics* 2011, **27:**258-266.
20. Mrode R, Ojango JM, Okeyo A, Mwacharo JM: **Genomic selection and use of molecular tools in breeding programs for indigenous and crossbred cattle in developing countries: current status and future prospects.** *Frontiers in genetics* 2019, **9:**694.
21. Chen JY, Liu C, Gui YJ, Si KW, Zhang DD, Wang J, Short DP, Huang JQ, Li NY, Liang Y: **Comparative genomics reveals cotton-specific virulence factors in flexible genomic regions in Verticillium dahliae and evidence of horizontal gene transfer from Fusarium.** *New Phytologist* 2018, **217:**756-770.
22. Kono TJ, Lei L, Shih C-H, Hoffman PJ, Morrell PL, Fay JC: **Comparative genomics approaches accurately predict deleterious variants in plants.** *G3: Genes, Genomes, Genetics* 2018, **8:**3321-3329.
23. Zhu G, Wang S, Huang Z, Zhang S, Liao Q, Zhang C, Lin T, Qin M, Peng M, Yang C: **Rewiring of the fruit metabolome in tomato breeding.** *Cell* 2018, **172:**249-261. e212.
24. Pazhamala LT, Kudapa H, Weckwerth W, Millar AH, Varshney RK: **Systems biology for crop improvement.** *The Plant Genome (TSI)* 2021**:**1-23.
25. Kersey PJ, Allen JE, Allot A, Barba M, Boddu S, Bolt BJ, Carvalho-Silva D, Christensen M, Davis P, Grabmueller C: **Ensembl Genomes 2018: an integrated omics infrastructure for non-vertebrate species.** *Nucleic acids research* 2018, **46:**D802-D808.
26. Van Bel M, Diels T, Vancaester E, Kreft L, Botzki A, Van de Peer Y, Coppens F, Vandepoele K: **PLAZA 4.0: an integrative resource for functional, evolutionary and comparative plant genomics.** *Nucleic acids research* 2018, **46:**D1190-D1196.
27. Goodstein DM, Shu S, Howson R, Neupane R, Hayes RD, Fazo J, Mitros T, Dirks W, Hellsten U, Putnam N: **Phytozome: a comparative platform for green plant genomics.** *Nucleic acids research* 2012, **40:**D1178-D1186.
28. Slotkin RK, Martienssen R: **Transposable elements and the epigenetic regulation of the genome.** *Nature reviews genetics* 2007, **8:**272-285.
29. Pikaard CS, Scheid OM: **Epigenetic regulation in plants.** *Cold Spring Harbor perspectives in biology* 2014, **6:**a019315.
30. Park PJ: **ChIP–seq: advantages and challenges of a maturing technology.** *Nature reviews genetics* 2009, **10:**669-680.
31. Klemm SL, Shipony Z, Greenleaf WJ: **Chromatin accessibility and the regulatory epigenome.** *Nature Reviews Genetics* 2019, **20:**207-220.
32. Kurdyukov S, Bullock M: **DNA methylation analysis: choosing the right method.** *Biology* 2016, **5:**3.
33. Barros-Silva D, Marques CJ, Henrique R, Jerónimo C: **Profiling DNA methylation based on next-generation sequencing approaches: new insights and clinical applications.** *Genes* 2018, **9:**429.
34. Zhang X, Yazaki J, Sundaresan A, Cokus S, Chan SW-L, Chen H, Henderson IR, Shinn P, Pellegrini M, Jacobsen SE: **Genome-wide high-resolution mapping and functional analysis of DNA methylation in Arabidopsis.** *Cell* 2006, **126:**1189-1201.
35. Zhang H, Lang Z, Zhu J-K: **Dynamics and function of DNA methylation in plants.** *Nature reviews Molecular cell biology* 2018, **19:**489-506.
36. Agorio A, Vera P: **ARGONAUTE4 is required for resistance to Pseudomonas syringae in Arabidopsis.** *The Plant Cell* 2007, **19:**3778-3790.
37. Rambani A, Rice JH, Liu J, Lane T, Ranjan P, Mazarei M, Pantalone V, Stewart Jr CN, Staton M, Hewezi T: **The methylome of soybean roots during the compatible interaction with the soybean cyst nematode.** *Plant Physiology* 2015, **168:**1364-1377.
38. Eichten SR, Springer NM: **Minimal evidence for consistent changes in maize DNA methylation patterns following environmental stress.** *Frontiers in plant science* 2015, **6:**308.
39. Marand AP, Chen Z, Gallavotti A, Schmitz RJ: **A cis-regulatory atlas in maize at single-cell resolution.** *Cell* 2021, **184:**3041-3055. e3021.

40. O'Malley RC, Huang S-sC, Song L, Lewsey MG, Bartlett A, Nery JR, Galli M, Gallavotti A, Ecker JR: **Cistrome and epicistrome features shape the regulatory DNA landscape.** *Cell* 2016, **165:**1280-1292.
41. Li M, Huang S-SC: **DNA Affinity Purification Sequencing (DAP-Seq) for Mapping Genome-Wide Transcription Factor Binding Sites in Plants.** In *Accelerated Breeding of Cereal Crops.* Springer; 2022: 293-303
42. Zhang Y, Li Z, Lin K, Peng Y, Ye L, Zhuang Y, Wang M, Xie Y, Guo J, Teng W: **Evolutionary rewiring of the wheat transcriptional regulatory network by lineage-specific transposable elements.** *Genome research* 2021, **31:**2276-2289.
43. Sonawane AR, DeMeo DL, Quackenbush J, Glass K: **Constructing gene regulatory networks using epigenetic data.** *NPJ Systems Biology and Applications* 2021, **7:**1-13.
44. Franco-Zorrilla JM, Prat S: **DAP-Seq Identification of Transcription Factor-Binding Sites in Potato.** In *Solanum tuberosum.* Springer; 2021: 123-142
45. Kliebenstein D: **Quantitative genomics: analyzing intraspecific variation using global gene expression polymorphisms or eQTLs.** *Annual review of plant biology* 2009, **60:**93-114.
46. Cubillos FA, Coustham V, Loudet O: **Lessons from eQTL mapping studies: non-coding regions and their role behind natural phenotypic variation in plants.** *Current opinion in plant biology* 2012, **15:**192-198.
47. Merrick BA, Phadke DP, Auerbach SS, Mav D, Stiegelmeyer SM, Shah RR, Tice RR: **RNA-Seq profiling reveals novel hepatic gene expression pattern in aflatoxin B1 treated rats.** *PloS one* 2013, **8:**e61768.
48. Wang J, Vasaikar S, Shi Z, Greer M, Zhang B: **WebGestalt 2017: a more comprehensive, powerful, flexible and interactive gene set enrichment analysis toolkit.** *Nucleic acids research* 2017, **45:**W130-W137.
49. Xu X, Zhang Y, Williams J, Antoniou E, McCombie WR, Wu S, Zhu W, Davidson NO, Denoya P, Li E: **Parallel comparison of Illumina RNA-Seq and Affymetrix microarray platforms on transcriptomic profiles generated from 5-aza-deoxy-cytidine treated HT-29 colon cancer cells and simulated datasets.** *BMC bioinformatics* 2013, **14:**1-14.
50. Liu Z-P, Chen L: **Proteome-wide prediction of protein-protein interactions from high-throughput data.** *Protein & cell* 2012, **3:**508-520.
51. Yuan Y, Bar-Joseph Z: **GCNG: graph convolutional networks for inferring gene interaction from spatial transcriptomics data.** *Genome Biology* 2020, **21:**1-16.
52. Raza K: **Fuzzy logic based approaches for gene regulatory network inference.** *Artificial intelligence in medicine* 2019, **97:**189-203.
53. Salazar C, Brümmer A, Alberghina L, Höfer T: **Timing control in regulatory networks by multisite protein modifications.** *Trends in cell biology* 2010, **20:**634-641.
54. Csizmok V, Forman-Kay JD: **Complex regulatory mechanisms mediated by the interplay of multiple post-translational modifications.** *Current opinion in structural biology* 2018, **48:**58-67.
55. Jin J, Tian F, Yang D-C, Meng Y-Q, Kong L, Luo J, Gao G: **PlantTFDB 4.0: toward a central hub for transcription factors and regulatory interactions in plants.** *Nucleic acids research* 2016**:**gkw982.
56. Tian F, Yang D-C, Meng Y-Q, Jin J, Gao G: **PlantRegMap: charting functional regulatory maps in plants.** *Nucleic acids research* 2020, **48:**D1104-D1113.
57. Braun P, Aubourg S, Van Leene J, De Jaeger G, Lurin C: **Plant protein interactomes.** *Annual review of plant biology* 2013, **64:**161-187.
58. Von Mering C, Krause R, Snel B, Cornell M, Oliver SG, Fields S, Bork P: **Comparative assessment of large-scale data sets of protein–protein interactions.** *Nature* 2002, **417:**399-403.
59. Perkins JR, Diboun I, Dessailly BH, Lees JG, Orengo C: **Transient protein-protein interactions: structural, functional, and network properties.** *Structure* 2010, **18:**1233-1243.
60. Byrum S, Smart SK, Larson S, Tackett AJ: **Analysis of stable and transient protein–protein interactions.** In *Chromatin Remodeling.* Springer; 2012: 143-152
61. Hao T, Peng W, Wang Q, Wang B, Sun J: **Reconstruction and application of protein–protein interaction network.** *International journal of molecular sciences* 2016, **17:**907.

62. Keskin O, Tuncbag N, Gursoy A: **Characterization and prediction of protein interfaces to infer protein-protein interaction networks.** *Current pharmaceutical biotechnology* 2008, **9:**67-76.
63. Ma C-Y, Chen Y-PP, Berger B, Liao C-S: **Identification of protein complexes by integrating multiple alignment of protein interaction networks.** *Bioinformatics* 2017, **33:**1681-1688.
64. Geisler-Lee J, O'Toole N, Ammar R, Provart NJ, Millar AH, Geisler M: **A predicted interactome for Arabidopsis.** *Plant physiology* 2007, **145:**317-329.
65. Gillis J, Pavlidis P: **The impact of multifunctional genes on" guilt by association" analysis.** *PloS one* 2011, **6:**e17258.
66. Szklarczyk D, Franceschini A, Wyder S, Forslund K, Heller D, Huerta-Cepas J, Simonovic M, Roth A, Santos A, Tsafou KP: **STRING v10: protein–protein interaction networks, integrated over the tree of life.** *Nucleic acids research* 2015, **43:**D447-D452.
67. Kanehisa M: **The KEGG database.** In *Novartis found symp*. 2002: 91-103.
68. Kell DB, Brown M, Davey HM, Dunn WB, Spasic I, Oliver SG: **Metabolic footprinting and systems biology: the medium is the message.** *Nature reviews microbiology* 2005, **3:**557-565.
69. Zampieri M, Sauer U: **Metabolomics-driven understanding of genotype-phenotype relations in model organisms.** *Current Opinion in Systems Biology* 2017, **6:**28-36.
70. Altaf-Ul-Amin M, Hirose K, Nani JV, Porta LC, Tasic L, Hossain SF, Huang M, Ono N, Hayashi MA, Kanaya S: **A system biology approach based on metabolic biomarkers and protein–protein interactions for identifying pathways underlying schizophrenia and bipolar disorder.** *Scientific reports* 2021, **11:**1-11.
71. Schläpfer P, Zhang P, Wang C, Kim T, Banf M, Chae L, Dreher K, Chavali AK, Nilo-Poyanco R, Bernard T: **Genome-wide prediction of metabolic enzymes, pathways, and gene clusters in plants.** *Plant physiology* 2017, **173:**2041-2059.
72. Vallenet D, Calteau A, Dubois M, Amours P, Bazin A, Beuvin M, Burlot L, Bussell X, Fouteau S, Gautreau G: **MicroScope: an integrated platform for the annotation and exploration of microbial gene functions through genomic, pangenomic and metabolic comparative analysis.** *Nucleic Acids Research* 2020, **48:**D579-D589.
73. Pinu FR, Beale DJ, Paten AM, Kouremenos K, Swarup S, Schirra HJ, Wishart D: **Systems biology and multi-omics integration: viewpoints from the metabolomics research community.** *Metabolites* 2019, **9:**76.
74. Caspi R, Billington R, Keseler IM, Kothari A, Krummenacker M, Midford PE, Ong WK, Paley S, Subhraveti P, Karp PD: **The MetaCyc database of metabolic pathways and enzymes-a 2019 update.** *Nucleic acids research* 2020, **48:**D445-D453.
75. Thimm O, Bläsing O, Gibon Y, Nagel A, Meyer S, Krüger P, Selbig J, Müller LA, Rhee SY, Stitt M: **MAPMAN: a user-driven tool to display genomics data sets onto diagrams of metabolic pathways and other biological processes.** *The Plant Journal* 2004, **37:**914-939.
76. Zhang B, Horvath S: **A general framework for weighted gene co-expression network analysis.** *Statistical applications in genetics and molecular biology* 2005, **4**.
77. Obayashi T, Hayashi S, Saeki M, Ohta H, Kinoshita K: **ATTED-II provides coexpressed gene networks for Arabidopsis.** *Nucleic acids research* 2009, **37:**D987-D991.
78. Deihimi T, Niazi A, Ebrahimi M, Kajbaf K, Fanaee S, Bakhtiarizadeh MR, Ebrahimie E: **Finding the undiscovered roles of genes: an approach using mutual ranking of coexpressed genes and promoter architecture-case study: dual roles of thaumatin like proteins in biotic and abiotic stresses.** *SpringerPlus* 2012, **1:**1-10.
79. Rao X, Dixon RA: **Co-expression networks for plant biology: why and how.** *Acta biochimica et biophysica Sinica* 2019, **51:**981-988.
80. Emamjomeh A, Robat ES, Zahiri J, Solouki M, Khosravi P: **Gene co-expression network reconstruction: a review on computational methods for inferring functional information from plant-based expression data.** *Plant biotechnology reports* 2017, **11:**71-86.
81. Mutwil M, Klie S, Tohge T, Giorgi FM, Wilkins O, Campbell MM, Fernie AR, Usadel B, Nikoloski Z, Persson S: **PlaNet: combined sequence and expression comparisons across plant networks derived from seven species.** *The Plant Cell* 2011, **23:**895-910.

82. Obayashi T, Aoki Y, Tadaka S, Kagaya Y, Kinoshita K: **ATTED-II in 2018: a plant coexpression database based on investigation of the statistical property of the mutual rank index.** *Plant and Cell Physiology* 2018, **59:**e3-e3.
83. Lee T, Yang S, Kim E, Ko Y, Hwang S, Shin J, Shim JE, Shim H, Kim H, Kim C: **AraNet v2: an improved database of co-functional gene networks for the study of Arabidopsis thaliana and 27 other nonmodel plant species.** *Nucleic acids research* 2015, **43:**D996-D1002.
84. Sun MG, Kim PM: **Evolution of biological interaction networks: from models to real data.** *Genome biology* 2011, **12:**1-11.
85. Alanis-Lobato G: **Mining protein interactomes to improve their reliability and support the advancement of network medicine.** *Frontiers in genetics* 2015, **6:**296.
86. Qian Y, Huang S-sC: **Improving plant gene regulatory network inference by integrative analysis of multi-omics and high resolution data sets.** *Current Opinion in Systems Biology* 2020, **22:**8-15.
87. Peng Z, He S, Gong W, Xu F, Pan Z, Jia Y, Geng X, Du X: **Integration of proteomic and transcriptomic profiles reveals multiple levels of genetic regulation of salt tolerance in cotton.** *BMC plant biology* 2018, **18:**1-19.
88. Vella D, Zoppis I, Mauri G, Mauri P, Di Silvestre D: **From protein-protein interactions to protein co-expression networks: a new perspective to evaluate large-scale proteomic data.** *EURASIP Journal on Bioinformatics and Systems Biology* 2017, **2017:**1-16.
89. De Smet R, Marchal K: **Advantages and limitations of current network inference methods.** *Nature Reviews Microbiology* 2010, **8:**717-729.
90. De Maeyer D, Weytjens B, Renkens J, De Raedt L, Marchal K: **PheNetic: network-based interpretation of molecular profiling data.** *Nucleic acids research* 2015, **43:**W244-W250.
91. Misra BB, Langefeld C, Olivier M, Cox LA: **Integrated omics: tools, advances and future approaches.** *Journal of molecular endocrinology* 2019, **62:**R21-R45.
92. Jamil IN, Remali J, Azizan KA, Nor Muhammad NA, Arita M, Goh H-H, Aizat WM: **Systematic multi-omics integration (MOI) approach in plant systems biology.** *Frontiers in plant science* 2020, **11:**944.
93. Rajasundaram D, Selbig J: **More effort—more results: recent advances in integrative 'omics' data analysis.** *Current opinion in plant biology* 2016, **30:**57-61.
94. Hassani-Pak K, Castellote M, Esch M, Hindle M, Lysenko A, Taubert J, Rawlings C: **Developing integrated crop knowledge networks to advance candidate gene discovery.** *Applied & Translational Genomics* 2016, **11:**18-26.
95. Zamboni A, Di Carli M, Guzzo F, Stocchero M, Zenoni S, Ferrarini A, Tononi P, Toffali K, Desiderio A, Lilley KS: **Identification of putative stage-specific grapevine berry biomarkers and omics data integration into networks.** *Plant Physiology* 2010, **154:**1439-1459.
96. Ji X, Ren J, Zhang Y, Lang S, Wang D, Song X: **Integrated Analysis of the Metabolome and Transcriptome on Anthocyanin Biosynthesis in Four Developmental Stages of Cerasus humilis Peel Coloration.** *International journal of molecular sciences* 2021, **22:**11880.
97. Maia M, Ferreira AE, Nascimento R, Monteiro F, Traquete F, Marques AP, Cunha J, Eiras-Dias JE, Cordeiro C, Figueiredo A: **Integrating metabolomics and targeted gene expression to uncover potential biomarkers of fungal/oomycetes-associated disease susceptibility in grapevine.** *Scientific reports* 2020, **10:**1-15.
98. Amiour N, Imbaud S, Clément G, Agier N, Zivy M, Valot B, Balliau T, Armengaud P, Quilleré I, Cañas R: **The use of metabolomics integrated with transcriptomic and proteomic studies for identifying key steps involved in the control of nitrogen metabolism in crops such as maize.** *Journal of Experimental Botany* 2012, **63:**5017-5033.
99. De Smet R, Adams KL, Vandepoele K, Van Montagu MC, Maere S, Van de Peer Y: **Convergent gene loss following gene and genome duplications creates single-copy families in flowering plants.** *Proceedings of the National Academy of Sciences* 2013, **110:**2898-2903.
100. Defoort J, Van de Peer Y, Carretero-Paulet L: **The evolution of gene duplicates in angiosperms and the impact of protein–protein interactions and the mechanism of duplication.** *Genome biology and evolution* 2019, **11:**2292-2305.

101. Flagel LE, Wendel JF: **Evolutionary rate variation, genomic dominance and duplicate gene expression evolution during allotetraploid cotton speciation.** *New Phytologist* 2010, **186:**184-193.
102. Coate JE, Song MJ, Bombarely A, Doyle JJ: **Expression-level support for gene dosage sensitivity in three Glycine subgenus Glycine polyploids and their diploid progenitors.** *New Phytologist* 2016, **212:**1083-1093.
103. Lehti-Shiu MD, Uygun S, Moghe GD, Panchy N, Fang L, Hufnagel DE, Jasicki HL, Feig M, Shiu S-H: **Molecular evidence for functional divergence and decay of a transcription factor derived from whole-genome duplication in Arabidopsis thaliana.** *Plant physiology* 2015, **168:**1717-1734.
104. Chen R, Mias GI, Li-Pook-Than J, Jiang L, Lam HY, Chen R, Miriami E, Karczewski KJ, Hariharan M, Dewey FE: **Personal omics profiling reveals dynamic molecular and medical phenotypes.** *Cell* 2012, **148:**1293-1307.
105. Fouss F, Saerens M, Shimbo M: *Algorithms and models for network data and link analysis.* Cambridge University Press; 2016.
106. Wang B, Mezlini AM, Demir F, Fiume M, Tu Z, Brudno M, Haibe-Kains B, Goldenberg A: **Similarity network fusion for aggregating data types on a genomic scale.** *Nature methods* 2014, **11:**333-337.
107. Glass K, Huttenhower C, Quackenbush J, Yuan G-C: **Passing messages between biological networks to refine predicted interactions.** *PloS one* 2013, **8:**e64832.
108. De Maeyer D, Renkens J, Cloots L, De Raedt L, Marchal K: **PheNetic: network-based interpretation of unstructured gene lists in E. coli.** *Molecular BioSystems* 2013, **9:**1594-1603.
109. Consortium AIM, Dreze M, Carvunis A-R, Charloteaux B, Galli M, Pevzner SJ, Tasan M, Ahn Y-Y, Balumuri P, Barabási A-L: **Evidence for network evolution in an Arabidopsis interactome map.** *Science* 2011, **333:**601-607.
110. Tian W, Zhang LV, Taşan M, Gibbons FD, King OD, Park J, Wunderlich Z, Cherry JM, Roth FP: **Combining guilt-by-association and guilt-by-profiling to predict Saccharomyces cerevisiae gene function.** *Genome biology* 2008, **9:**1-21.
111. Hou L, Chen M, Zhang CK, Cho J, Zhao H: **Guilt by rewiring: gene prioritization through network rewiring in genome wide association studies.** *Human molecular genetics* 2014, **23:**2780-2790.
112. Lee I, Seo Y-S, Coltrane D, Hwang S, Oh T, Marcotte EM, Ronald PC: **Genetic dissection of the biotic stress response using a genome-scale gene network for rice.** *Proceedings of the National Academy of Sciences* 2011, **108:**18548-18553.
113. Gaudinier A, Rodriguez-Medina J, Zhang L, Olson A, Liseron-Monfils C, Bågman A-M, Foret J, Abbitt S, Tang M, Li B: **Transcriptional regulation of nitrogen-associated metabolism and growth.** *Nature* 2018, **563:**259-264.
114. Wei H, Persson S, Mehta T, Srinivasasainagendra V, Chen L, Page GP, Somerville C, Loraine A: **Transcriptional coordination of the metabolic network in Arabidopsis.** *Plant physiology* 2006, **142:**762-774.
115. Movahedi S, Van Bel M, Heyndrickx KS, Vandepoele K: **Comparative co-expression analysis in plant biology.** *Plant, cell & environment* 2012, **35:**1787-1798.
116. Schaefer RJ, Michno J-M, Jeffers J, Hoekenga O, Dilkes B, Baxter I, Myers CL: **Integrating coexpression networks with GWAS to prioritize causal genes in maize.** *The Plant Cell* 2018, **30:**2922-2942.
117. Redekar N, Pilot G, Raboy V, Li S, Maroof S: **Inference of transcription regulatory network in low phytic acid soybean seeds.** *Frontiers in plant science* 2017, **8:**2029.
118. Penfold CA, Buchanan-Wollaston V, Denby KJ, Wild DL: **Nonparametric Bayesian inference for perturbed and orthologous gene regulatory networks.** *Bioinformatics* 2012, **28:**i233-i241.
119. Vermeirssen V, De Clercq I, Van Parys T, Van Breusegem F, Van de Peer Y: **Arabidopsis ensemble reverse-engineered gene regulatory network discloses interconnected transcription factors in oxidative stress.** *The Plant Cell* 2014, **26:**4656-4679.

120. McCormack ME, Lopez JA, Crocker TH, Mukhtar MS: **Making the right connections: network biology and plant immune system dynamics.** *Current Plant Biology* 2016, **5:**2-12.
121. Gillis J, Pavlidis P: **"Guilt by association" is the exception rather than the rule in gene networks.** *PLoS computational biology* 2012, **8:**e1002444.
122. Gunning M, Pavlidis P: **"Guilt by association" is not competitive with genetic association for identifying autism risk genes.** *Scientific reports* 2021, **11:**1-15.
123. Lee I, Ambaru B, Thakkar P, Marcotte EM, Rhee SY: **Rational association of genes with traits using a genome-scale gene network for Arabidopsis thaliana.** *Nature biotechnology* 2010, **28:**149-156.
124. Aerts S, Lambrechts D, Maity S, Van Loo P, Coessens B, De Smet F, Tranchevent L-C, De Moor B, Marynen P, Hassan B: **Gene prioritization through genomic data fusion.** *Nature biotechnology* 2006, **24:**537-544.
125. Lysenko A, Urban M, Bennett L, Tsoka S, Janowska-Sejda E, Rawlings CJ, Hammond-Kosack KE, Saqi M: **Network-based data integration for selecting candidate virulence associated proteins in the cereal infecting fungus Fusarium graminearum.** *PloS one* 2013, **8:**e67926.
126. Segal E, Shapira M, Regev A, Pe'er D, Botstein D, Koller D, Friedman N: **Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data.** *Nature genetics* 2003, **34:**166-176.
127. Koh HW, Fermin D, Vogel C, Choi KP, Ewing RM, Choi H: **iOmicsPASS: network-based integration of multiomics data for predictive subnetwork discovery.** *NPJ systems biology and applications* 2019, **5:**1-10.
128. Tuncbag N, McCallum S, Huang S-sC, Fraenkel E: **SteinerNet: a web server for integrating 'omic'data to discover hidden components of response pathways.** *Nucleic acids research* 2012, **40:**W505-W509.
129. Merelli I, Liò P, Milanesi L: **NuChart: an R package to study gene spatial neighbourhoods with multi-omics annotations.** *PLoS One* 2013, **8:**e75146.
130. Voordeckers K, Kominek J, Das A, Espinosa-Cantu A, De Maeyer D, Arslan A, Van Pee M, van der Zande E, Meert W, Yang Y: **Adaptation to high ethanol reveals complex evolutionary pathways.** *PLoS genetics* 2015, **11:**e1005635.
131. Bar-Joseph Z, Gerber GK, Lee TI, Rinaldi NJ, Yoo JY, Robert F, Gordon DB, Fraenkel E, Jaakkola TS, Young RA: **Computational discovery of gene modules and regulatory networks.** *Nature biotechnology* 2003, **21:**1337-1342.
132. Saint-André V: **Computational biology approaches for mapping transcriptional regulatory networks.** *Computational and Structural Biotechnology Journal* 2021, **19:**4884-4895.
133. Muhammad II, Kong SL, Akmar Abdullah SN, Munusamy U: **RNA-seq and ChIP-seq as complementary approaches for comprehension of plant transcriptional regulatory mechanism.** *International journal of molecular sciences* 2020, **21:**167.
134. Lemmens K, De Bie T, Dhollander T, De Keersmaecker SC, Thijs IM, Schoofs G, De Weerdt A, De Moor B, Vanderleyden J, Collado-Vides J: **DISTILLER: a data integration framework to reveal condition dependency of complex regulons in Escherichia coli.** *Genome biology* 2009, **10:**1-13.
135. Ernst J, Beg QK, Kay KA, Balázsi G, Oltvai ZN, Bar-Joseph Z: **A semi-supervised method for predicting transcription factor–gene interactions in Escherichia coli.** *PLoS computational biology* 2008, **4:**e1000044.
136. Woo JH, Shimoni Y, Yang WS, Subramaniam P, Iyer A, Nicoletti P, Martínez MR, López G, Mattioli M, Realubit R: **Elucidating compound mechanism of action by network perturbation analysis.** *Cell* 2015, **162:**441-451.
137. Mordelet F, Vert J-P: **SIRENE: supervised inference of regulatory networks.** *Bioinformatics* 2008, **24:**i76-i82.
138. Delgado FM, Gómez-Vela F: **Computational methods for gene regulatory networks reconstruction and analysis: a review.** *Artificial intelligence in medicine* 2019, **95:**133-145.
139. Zheng G, Huang T: **The reconstruction and analysis of gene regulatory networks.** *Computational Systems Biology* 2018**:**137-154.

140. Weirauch MT, Yang A, Albu M, Cote AG, Montenegro-Montero A, Drewe P, Najafabadi HS, Lambert SA, Mann I, Cook K: **Determination and inference of eukaryotic transcription factor sequence specificity.** *Cell* 2014, **158:**1431-1443.
141. Inukai S, Kock KH, Bulyk ML: **Transcription factor–DNA binding: beyond binding site motifs.** *Current opinion in genetics & development* 2017, **43:**110-119.
142. Blanchette M, Tompa M: **FootPrinter: a program designed for phylogenetic footprinting.** *Nucleic acids research* 2003, **31:**3840-3842.
143. De Witte D, Van de Velde J, Decap D, Van Bel M, Audenaert P, Demeester P, Dhoedt B, Vandepoele K, Fostier J: **BLSSpeller: exhaustive comparative discovery of conserved cis-regulatory elements.** *Bioinformatics* 2015, **31:**3758-3766.

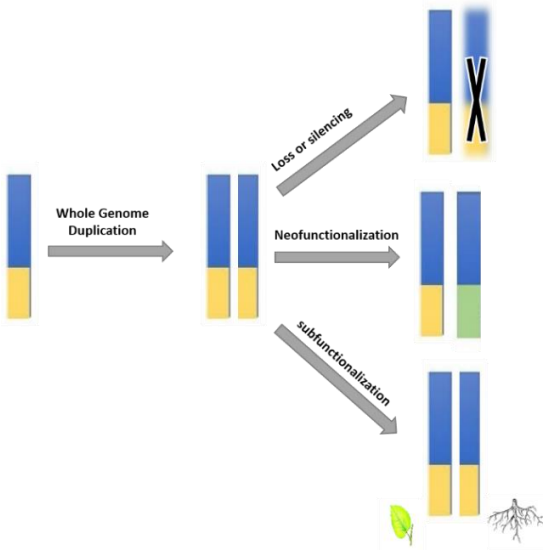
2 Chapter 2

Distinct Expression and Methylation Patterns for Genes with Different Fates following a Single Whole-Genome Duplication in Flowering Plants

"Believing as I do in evolution, I merely believe that it is the method by which God created, and is still creating, life on earth'

- Rachel Carson

Flowering plant genomes are characterized by large variations of genome size and level of ploidy, which is mostly driven by gene duplication. Interestingly, the fate of duplicated regions (loss, neofunctionalization, subfunctionalization) is not independent of their genetic characteristic and their function. One of the well-known hypotheses is the gene (dosage) balance hypothesis (GBH) which suggests that altering the stoichiometric balance of members of macromolecular complexes results in dosage-dependent phenotypes. Hence, preserving the balance by retaining every single gene in the complex is required and genes in complexes can only be deleted together with their 'interactors'. In line with this hypothesis, multiple studies showed that genes with retained copies following a WGD event are enriched for regulatory and signaling functions whereas single-copy genes are associated with housekeeping functions. Describing the difference in functional constraints on expression, epigenetic regulation, and behavior in biological networks that shape the fate of genes after WGD can shed light on the function of genes and the biological processes they involved. The study described below is a collaborative effort with CAS Key Laboratory of Aquatic Botany and Watershed Ecology, Wuhan Botanical Garden. In the framework of the study, the candidate performed the quality-control on the expression data, preprocessed the data for analysis, performed the expression and network analysis, and contributed to the writing of the paper. The candidate was not involved in the experimental design, the wet-lab data generation, and the genome analysis. Page 2-29 includes an overview of the contributions of all authors.



Distinct Expression and Methylation Patterns for Genes with Different Fates following a Single Whole-Genome Duplication in Flowering Plants [1]

Tao Shi, Razgar Seyed Rahmani, Paul F. Gugger, Muhua Wang, Hui Li, Yue Zhang, Zhizhong Li, Qingfeng Wang, Yves Van de Peer*, Kathleen Marchal*, Jinming Chen*

Published in *Molecular Biology and Evolution*, 37(8), August 2020

Abstract

For most sequenced flowering plants, multiple whole-genome duplications (WGDs) are found. Duplicated genes following WGD often have different fates that can quickly disappear again, be retained for long(er) periods, or subsequently undergo small-scale duplications. However, how different expression, epigenetic regulation, and functional constraints are associated with these different gene fates following a WGD still requires further investigation due to successive WGDs in angiosperms complicating the gene trajectories. In this study, we investigate lotus (*Nelumbo nucifera*), an angiosperm with a single WGD during the K–pg boundary. Based on improved intraspecific-synteny identification by a chromosome-level assembly, transcriptome, and bisulfite sequencing, we explore not only the fundamental distinctions in genomic features, expression, and methylation patterns of genes with different fates after a WGD but also the factors that shape post-WGD expression divergence and expression bias between duplicates. We found that after a WGD genes that returned to single copies show the highest levels and breadth of expression, gene body methylation, and intron numbers, whereas the long-retained duplicates exhibit the highest degrees of protein–protein

interactions and protein lengths and the lowest methylation in gene flanking regions. For those long-retained duplicate pairs, the degree of expression divergence correlates with their sequence divergence, degree in protein–protein interactions, and expression level, whereas their biases in expression level reflecting subgenome dominance are associated with the bias of subgenome fractionation. Overall, our study on the paleopolyploid nature of lotus highlights the impact of different functional constraints on gene fate and duplicate divergence following a single WGD in plant.

2.1 Introduction

Gene duplication is one of the most important drivers of eukaryotic evolution. Indeed, by increasing the amount of raw genetic material on which evolution can work, gene duplication generates the genetic redundancy through which processes such as subfunctionalization and neofunctionalization can create functional novelty [2-7]. Apart from small-scale gene duplication (SSD), also whole-genome duplication (WGD), whereby thousands of novel genes are created at once, has been frequently observed during evolution, especially in flowering plants [8-11]. Interestingly, the fate of genes duplicated through such large-scale duplication (LSD) events often seems to be different from that of genes duplicated in small-scale events, and previous studies have shown that the chance of survival and maintenance of genes duplicated in a WGD is very much dependent on their function. On the one hand, despite repeated WGDs in angiosperms, many genes were found that convergently revert to single-copy status, and in Arabidopsis, they exhibit more constitutive and higher expression than duplicate genes in general and are enriched in housekeeping functions [12, 13]. One explanation is that the deletion of duplicates is needed to prevent copies with dominant-negative mutations, which might interfere with the correct functioning of the wild type copy [12, 13]. On the other hand, there are those genes that are retained in excess following WGD for a longer time. For these retained duplicate genes, gene balance hypothesis (GBH) states that maintaining stoichiometric balance is crucial, and genes can only be deleted together with their “interactors” where losing or further duplication of part of the network or complex is detrimental because the stoichiometry is challenged [14-18]. Genes that underwent SSDs, such as tandemly duplicated genes, in contrast were found to be selected for either increased gene dosage or rapid gene turnover in order to confer lineage-specific adaptation because they are mostly insensitive to dosage imbalance [3, 19]. Although these theories explain how different mechanisms that potentially affect gene fate after WGD, we still do not know the difference in functional constraints including quantifiable features such as expression, epigenetic regulation, and protein–protein interactions (PPIs) imposed on those genes with different fates after a WGD (single-copy, WGD, and SSD genes).

Studies including a recent investigation on WGDs across plants including 134 sequenced angiosperms suggest that after diverging from the extant basal-most angiosperm (Amborella), only lotus and seagrass (*Zostera marina*) experienced a single WGD (4x), whereas the other angiosperms experienced at least a genome triplication (6x) or sequential WGDs [20].

However, the scaffold-level genome assembly of seagrass provides limited information on synteny to study the gene fates after its WGD [21]. Case studies of recently released genomes also show that columbine, Liriodendron and water lily experienced a single WGD [22-24]. Therefore, the genome of sacred lotus *(Nelumbo nucifera* Gaertn.) is one of the few angiosperms carrying a well-retained intraspecific synteny reflecting only a single ancient WGD coincided with the K–pg boundary [25-28]. Because of its relatively simple and ancient WGD history, lotus genome facilitates comparing genes with different fates (duplication status) following a single WGD. In addition, because long-retained duplicate pairs descending from the same WGD event can be easily tracked in species such as lotus, the (functional) factors, including dosage-balance constraint, that shape the expression pattern divergence of duplicate gene pairs can also be well investigated. Yet, in Arabidopsis, poplar, soybean, tomato, or maize, the fact that multiple different rounds of WGDs occurred makes it difficult to study the fate of the most ancient duplicates [29-31]. Other than divergence in expression pattern, many duplicate pairs might have bias in expression level [32]. Often, this expression bias between the two copies is associated with subgenome dominance which is a phenomenon that was initially defined in allopolyploid cotton and later in other (presumed) paleo-allopolyploids: copies residing in one less fractionated (LF; parental) subgenome tend to show higher expression than those in the other (parental) subgenome [33-40].

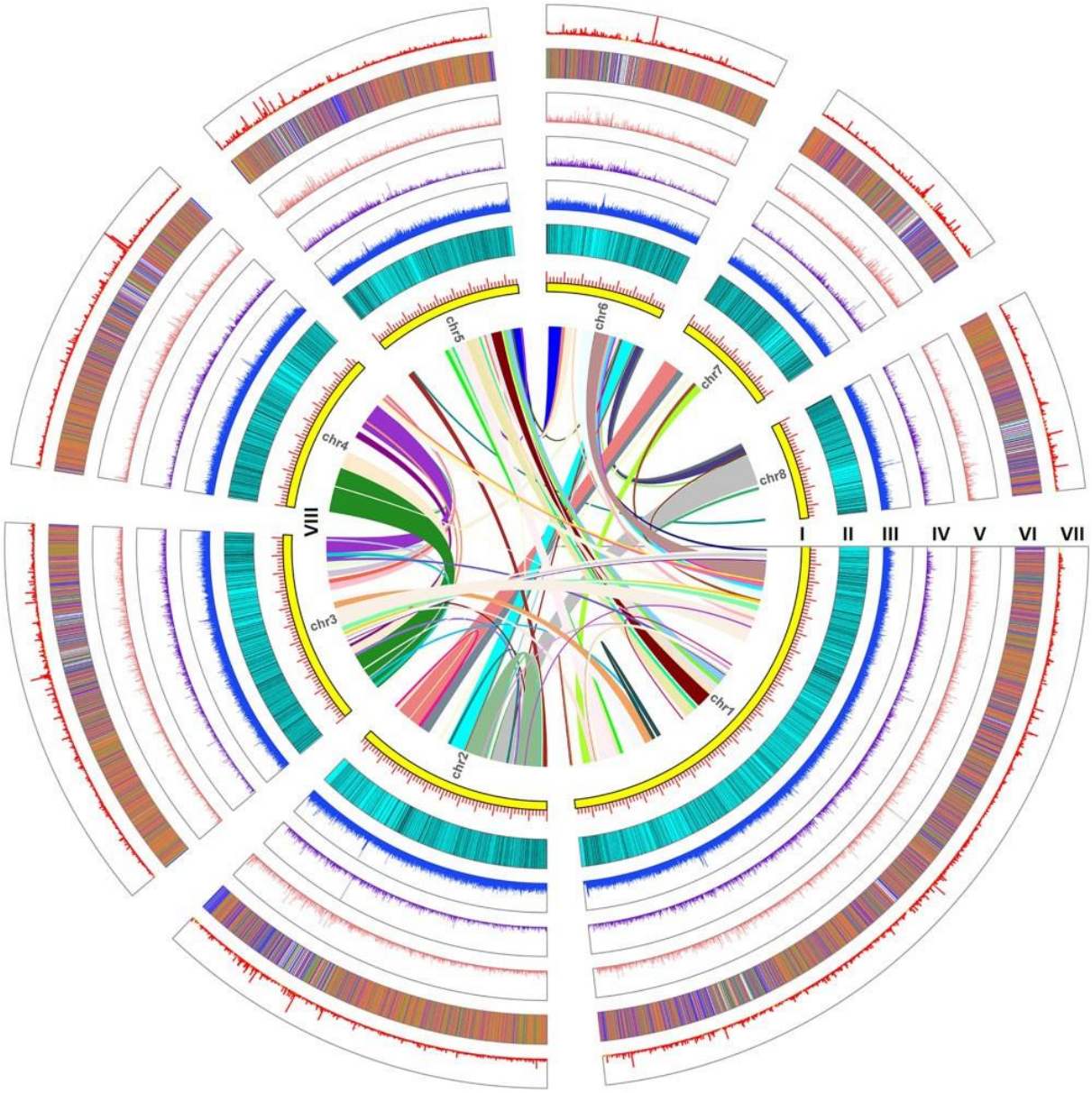


Figure 2.1 Circos plot of lotus genome assembly. From inside to outside rings: (I) size (Mb) of the assembly for each chromosome; (II) density distribution of genes; (III) density distribution of sRNA - TEs; (IV) density distribution of sRNA + TEs; (V) dot plot of nucleotide diversity of CDS for each gene; (VI) methylation level of genes and flanking regions; (VII) gene expression level (log-transformed FPKM value); and (VIII) syntenic paralogs are linked by colored lines.

Therefore, understanding the mechanisms such as epigenetic regulation and subgenome dominance underlying the divergence in expression pattern and level after a WGD in lotus will improve our understanding of how a duplicate pair diverges in function. To better address the questions as mentioned above, we build an improved genome assembly of the lotus var. “China Antique” by PacBio long-read sequencing and scaffolding using high-throughput chromosome conformation capture (Hi-C). This can optimally identify the genomic relics from both ancient SSD and WGD events. Complementing this chromosome-level assembly with further whole-genome bisulfite (methylation) sequencing, RNA-seq, and genome resequencing data, not

only allow us to study the mechanisms, such as expression and epigenetic regulation that coordinate and maintain the functional integrity of genes displaying different evolutionary fates, but also provide further insight into the genetic mechanisms that create functional divergence of duplicates retained after a WGD.

2.2 Results

2.2.1 A chromosome-level assembly of lotus

Based on newly generated data, we obtained an improved assembly and annotation of the lotus genome. Combining PacBio Sequel subreads (11.9 G; 1,330,739 subreads with a mean length of 8.8 kb and N50 of 12.7 kb) with previously published Illumina paired-end (PE) reads (94.2 Gb) [25], resulted in a hybrid assembly, containing contigs with an N50 length of 484.3 kb. This assembly is about 12.5 times the length of previously assembled contigs (v2013) (N50 38.8 kb) (Figure S2.1). The final 4,709 contigs cover about 807.6 Mb. Using genome-wide HI-C, overall, 4,248 contigs (799.7 Mb) were anchored and ordered into eight different pseudomolecules (chromosomes) (**Figure 2.2**). Further optimization of the assembly by gap filling and polishing (error correction using accurate Illumina reads) resulted in a final assembly consisting of eight pseudochromosomes (813.2 Mb) and 456 unanchored contigs (8.0 Mb) (**Figure 2.1** and Table S2.1).

The newly assembled genome contains 58.5% repetitive sequences, of which 48.7% of the total assembly consists of known transposable elements (TEs) and 9.1% of unknown repeats (Figure 2.1 and Table S2.2). Gene annotation based on a repeat-masked genome yielded a total of 32,124 protein-coding genes (Figure 2.1). The accuracy of the new assembly was assessed by a previous single-nucleotide polymorphism (SNP)-based linkage map of lotus [41]. The majority of uniquely mapped SNP markers from a given linkage group aligned within the same pseudochromosome in the new assembly, whereas in the old assembly these markers showed a partitioned and mosaic distribution over different megascaffolds (v2013) (Figure S2.3). To assess the completeness of the assembly, we investigated to what extent the 1,440 plant conserved gene set of BUSCO was recovered: 94.6% (1,362) of the gene set was completely retrieved, 3.1% (44) was partially retrieved, and 2.3% (34) was “missing.” This shows that our assembly is the most complete lotus assembly to date when comparing to the other lotus assemblies (Table S2.3) (Gui et al. 2018). This is supported by the fact that the number of syntenic orthologs, for instance in relation to monocots, is substantially higher in our new assembly than in an older version: 5,421 *Brachypodium distachyon* genes and 5,922 rice genes showed a collinear relationship in the new assembly, whereas in the old assembly the numbers were 3,690 and 4,040, respectively (v2013) (Figure S2.4). Comparing eudicot genomes from the Plant Genome Duplication Database (PGDD) and our lotus assembly to both *B. distachyon* and rice learns that both the new and old assemblies of lotus share more collinear orthologs with the two monocot genomes than the other eudicots (Figure S2.4). Although lotus and the other eudicots in the PGDD together form a sister group to monocots, the genome architecture

(at least considering synteny) of lotus seems to resemble that of monocots most, probably because most eudicots present in the PGDD have undergone at least one triplication or further rounds of WGDs subsequent to eudicot radiation (Ming et al. 2013).

2.2.2 Classification of single-copy and duplicated lotus genes

To define different classes of lotus duplicates [42, 43], first, within-species syntenic blocks were identified (see Materials and Methods). Such blocks, showing conservation in gene content and order, and thus potentially representing remnants of a WGD, were found across all chromosomes (**Figure 2.1** and Table S2.4). Comparison of peaks in 4dTv (4-fold degenerate site transversion) distances which represent age distributions formed by the divergence of syntenic duplicates (4dTv median 0.158) and divergence of orthologs between lotus and *Macadamia ternifolia* (the other sequenced Proteales species) (4dTv median 0.405) suggests that most syntenic duplicates (WGD) have been derived from a duplication event after the split between Macadamia and lotus (Mann–Whitney U test, $P < 0.01$) (Figure S2.5).

Next to 2,353 orphan genes (defined as genes in lotus that have no homolog in any other considered plant species), we identified 29,771 genes with homologs in other species (non-orphan genes) (Table S2.5). Among these lotus genes, so-called dispersed duplicates are the most abundant (13,235), followed by duplicates resulting from WGD (referred to WGD) (9,482), tandemly duplicated genes (2,622), single-copy genes (2,261), proximal duplicated genes (1,566), and finally duplicates that underwent both WGD and tandem duplication (WGD&TD) (605), as classified by MCscanX (Figure S2.6A and Table S2.5,). Orphan genes are mostly either single-copy (62.14%) dispersed duplicates (33.81%) (Figure S2.6B and Table S2.5). The above-defined gene groups were used to further study how the fate of genes, for instance after WGD, correlates with functional constraints, reflected by PPIs, gene expression, and epigenetic and sequence properties. Lotus-specific orphan genes were analyzed separately.

2.2.3 Single-copy genes and WGD-derived duplicates of lotus show conservation in copy number in related taxa

Here, we estimated the extent to which dosage sensitivity (copy number conservation) of lotus genes depends on their duplication status. Hereto, we first grouped lotus genes according to their duplication status in lotus (as defined above, “single-copy genes,” “WGD,” “tandem duplicates,” and others) and subsequently assessed whether the orthologs of these lotus genes retained the same copy number status in two related eudicot species, namely *M. ternifolia* and *Vitis vinifera*. Macadamia was chosen because it is the sequenced Proteales species that is closest to lotus, whereas *Vitis*, with only one eudicot genome triplication, was also chosen because of its relatively conserved genome architecture compared with the other core eudicots [44]. To assess the variation in copy number across the studied species, we used the coefficient of variation (CV). The average copy number among the three species (as shown in the violin plot) varies largely among the genes of different duplication status, and therefore standard

deviation cannot serve to assess the variation in this case (Figure 2.2A). Single-copy genes (grouped according to their single-copy status in lotus) have a median of the average copy number among the three species close to one, indicating that, for genes grouped as single-copy in lotus, there is a general strong selection against gene redundancy in the related species as well (Figure 2.2A).

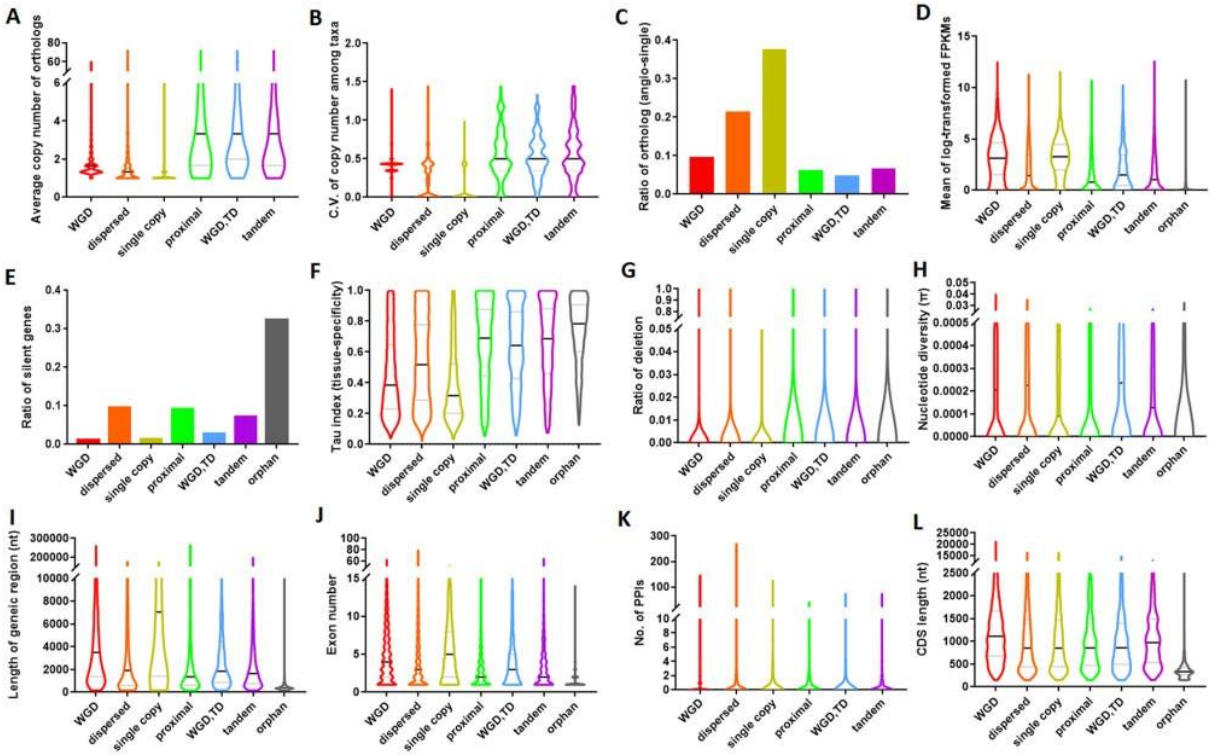


Figure 2.2 Violin plots of expression, functional, and genomic features of genes from different gene groups (based on duplication status). (*A*) The average copy number of orthologs. (*B*) Coefficient of variation (CV) of copy number among taxa. (*C*) Ratio of orthologs as "angio-singles." (*D*) The mean of log-transformed FPKM. (*E*) The ratio of silent genes. (*F*) Tissue specificity index (based on Tau index). (*G*) The average portion of the deleted genic sequence in tropical lotus comparing to the reference genome (ratio of deletion). (*H*) Nucleotide diversity (π). (*I*) Length of the genic region. (*J*) Exon number. (*K*) The number of protein–protein interactions inferred from the closest orthologs in *Arabidopsis*. (*L*) CDS length. Black line: median; gray line: quantile.

For genes classified as lotus WGD-derived duplicates, a median of the average copy number between 1 and 2 was found, suggesting that genes belonging to this group also tend to display a limited level of gene redundancy in the three studied taxa (Figure 2.2A). Interestingly, dispersed and WGD-derived duplicates show, after single-copy genes, respectively the second and the third-lowest CV for variation in copy number, and therefore presumably exhibit higher dosage sensitivity than local duplicates (tandem, proximal, and WGD&TD) (Kruskal–Wallis test, all P values < 0.01) (Figure 2.2B). This is in line with the GBH, which states that WGD-derived duplicates are more dosage-sensitive or more strict in preserving their copy numbers than local duplicates [19, 45]. For the group of the dispersed duplicates, the interpretation is

less trivial as these genes contain WGD-derived duplicates that lost collinearity, local duplicates that lost "proximity" to other duplicates, transposed duplicates, or "angiosperm-conserved single-copy genes" ("angio-singles") that were created by earlier preangiosperm duplications but stopped duplicating during angiosperm radiation. By examining the proportion of "angio-singles" in each of the studied gene groups using annotations described in a previous publication [12], we found that next to the group of single-copy genes, the group of dispersed duplicates contains the second-highest enrichment of "angio-singles" (Figure 2.2C). Greater 4dTv distances between the most similar dispersed duplicates than between corresponding orthologs (*Nelumbo* vs. *Amborella*) (Kruskal–Wallis test, all P values < 0.01) (Figure S2.7) suggest that "angio-singles" in dispersed duplicates were mostly created by early duplications prior to angiosperm radiation. As those early duplicates stopped duplicating during angiosperm radiation, they were classified as so-called single-copy genes in angiosperms. This explains why the group of dispersed duplicates also shows a low CV in copy number.

2.2.4 Single-copy genes and WGD-derived duplicate genes have high expression level and breadth

To understand why single-copy genes and WGD-derived duplicates are more highly constrained in copy number, we compared the level and breadth of gene expression for the above-defined gene groups. This is because genes expressed at higher levels tend to be under stronger selective pressure [46-49]. Average gene expression levels (log-transformed FPKMs), observed in 41 samples representing a variety of tissue-types, varied substantially among the studied gene groups. Single-copy genes showed on average the highest expression level (Kruskal–Wallis test, all P values < 0.01) (Figure 2.2D and Table S2.6). This result is consistent with a previous finding in Arabidopsis showing that the angiosperm-conserved single-copy genes generally show higher expression than duplicated genes [12]. This larger expression ubiquity also implies that single-copy genes are more likely involved in housekeeping functions than genes belonging to the other groups. When focusing on the duplicated genes, genes retained after WGD show on average a significantly higher expression level than genes from groups representing other types of duplicates (Kruskal–Wallis test, all P values < 0.01) (Figure 2.2D). Because essential genes are found to be highly expressed in Arabidopsis and other plants [50], this suggests that both single-copy and WGD-derived duplicates might constitute the more essential genes in lotus. Therefore, the strong purifying selection from gene essentiality of these two groups of genes might play an important role in constraining their dosage sensitivity (copy number change among taxa).

Further, we found that in lotus the largest gene group, namely the dispersed duplicates, possesses the highest ratio of silent genes (genes that are not expressed in any of the investigated samples) (9.61%), followed by proximal duplicates (9.20%) and tandem duplicates (7.29%), whereas genes resulting from WGD&TD (2.81%), from WGD (1.15%) and single-copy genes (1.42%) display much lower ratios of silent genes (Figure 2.2E). This

explains that even though dispersed duplicates contain a large portion of "angiosperm-conserved single-copy genes," they do not show a higher expression level than duplicates retained from WGD because they also contain a substantial number of silent (likely pseudogenized) duplicate genes. We further showed that compared with the expressed dispersed duplicates, the silent dispersed duplicates generally have younger ages (measured by 4dTv), lower number of introns, smaller protein length, and lower selective pressure, suggesting that they might be recent retrotransposed duplicates (Figure S2.8). Overall, these comparisons further confirm that losing function by gene silencing is not a random phenomenon and that single-copy genes and duplicates retained after a WGD are the least likely to be silenced.

Moreover, using the Tau index to measure expression specificity across different lotus tissues, we revealed that single-copy genes (mean Tau index of 0.38) show the lowest expression specificity of all gene groups (Kruskal–Wallis test, all P-values < 0.01). In addition, WGD duplicates (mean Tau index 0.45) exhibit significantly lower expression specificity than other types of duplicates (Kruskal–Wallis test, all P val- ues < 0.01) (Figure 2.2F). Both single-copy genes and genes retained from a WGD tend to have a wider "expression breadth" than small-scale duplicates, and hence their expression might be essential in most tissues as is supported by findings in Arabidopsis [50]. By showing higher expression level and breadth, both single-copy genes and WGD-derived genes might expose themselves to stronger purifying selection. This is supported by lotus genome resequencing data that show significantly lower ratios of sequence deletion and nucleotide diversity (π) for single-copy genes and WGD-derived duplicates than for small-scale duplicates (Kruskal–Wallis test, all P values < 0.01) (Figure 2.2G and H).

2.2.5 Differences in expression might be associated with differences in methylation level and TE distribution

Most cis-regulatory elements reside in gene flanking regions, which play profound roles in gene regulation. Given the impact of epigenetic regulation on gene expression, we assessed whether the above-mentioned differences in expression among different gene groups could be associated with differences in methylation level on gene flanking regions [51-53]. Hereto, we used methylation data obtained from leaf, petal, stamen petaloid, and stamen. Cytosine methylation levels at CG, CHG, and CHH sites along the gene (upstream, genic and downstream region) generally display a curved "W" shape with the lowest methylation level being observed close to the gene start and stop sites; Note that similar "W"-like shapes were observed when using an alternative definition of flanking regions (see Materials and Methods) (Figure 2.3 and Figures S2.9 and S2.10). These patterns in which the lowest methylation level is observed near the flanking regions agree with the finding that methylation can inhibit the binding of RNA polymerase II and transcriptional initiation [51]. Among CG, CHG, and CHH sites, the methylation level is the strongest at CG (mean ML = 0.458) (Figure 2.3A). The

average methylation level in flanking regions (promoters and downstream regions) of genes retained after a WGD is significantly lower than the methylation levels of genes belonging to other groups, indicating that duplicates retained after a WGD are transcriptionally less repressed by methylation in flanking regions. This is displayed in figure 2.3 for methylation levels observed in leaf. Similar figures were obtained for the methylation data obtained from other tissues (Figure S2.9 and S2.10). This average lower methylation level in flanking regions for genes that were retained after a WGD is in line with their relatively higher expression level and breadth. In contrast, the higher expression level and breadth of single-copy genes as compared with genes from other groups seem not to be associated with relatively lower methylation levels of flanking regions: single-copy genes display a higher methylation level in their promoters than genes belonging to the other groups (Kruskal–Wallis test, P values < 0.01).

In plants, (24-nt) RNA-directed DNA methylation is frequent in regions containing TEs, likely because most TEs need to be silenced to reduce TE activity and maintain genome stability. Hence, we assessed the degree to which differences in methylation level in gene flanking regions can be associated with the presence of TEs, including both TEs with 24-nt small (interfering) RNA (sRNA TE) and those without (sRNA TE) [54] (see Materials and Methods). Interestingly, the differences in TE density, especially of sRNA TEs, between the different gene groups resembles the distribution pattern of the overall CG and CHG methylation levels, where the gene group representing duplicates retained after a WGD shows the lowest average TE density in gene flanking regions and concomitantly also the lowest average methylation levels in these flanking regions (Figure S2.11).

Unlike gene flanking regions, the methylation level along the gene body (gene region) seems to be more related to differences in gene expression among the different gene groups. Whereas gene flanking DNA methylation is generally believed to repress gene expression [55-57], we found that higher gene body methylation level tends to occur in the gene groups with higher expression level and breadth, that is, single-copy and WGD duplicates. Interestingly, we found that for the group of single-copy genes, on average, the higher methylation level in the gene body seems to correlate with their greater gene length and exon number (Kruskal–Wallis test, all P values < 0.01) (Figure 2.2I and J). The fact that introns often contain TEs which are often associated with higher methylation levels might explain why single-copy genes also display the highest TE density in their gene body (Kruskal–Wallis test, all P values < 0.01) (Figure 2.3 and Table S2.6) [58, 59].

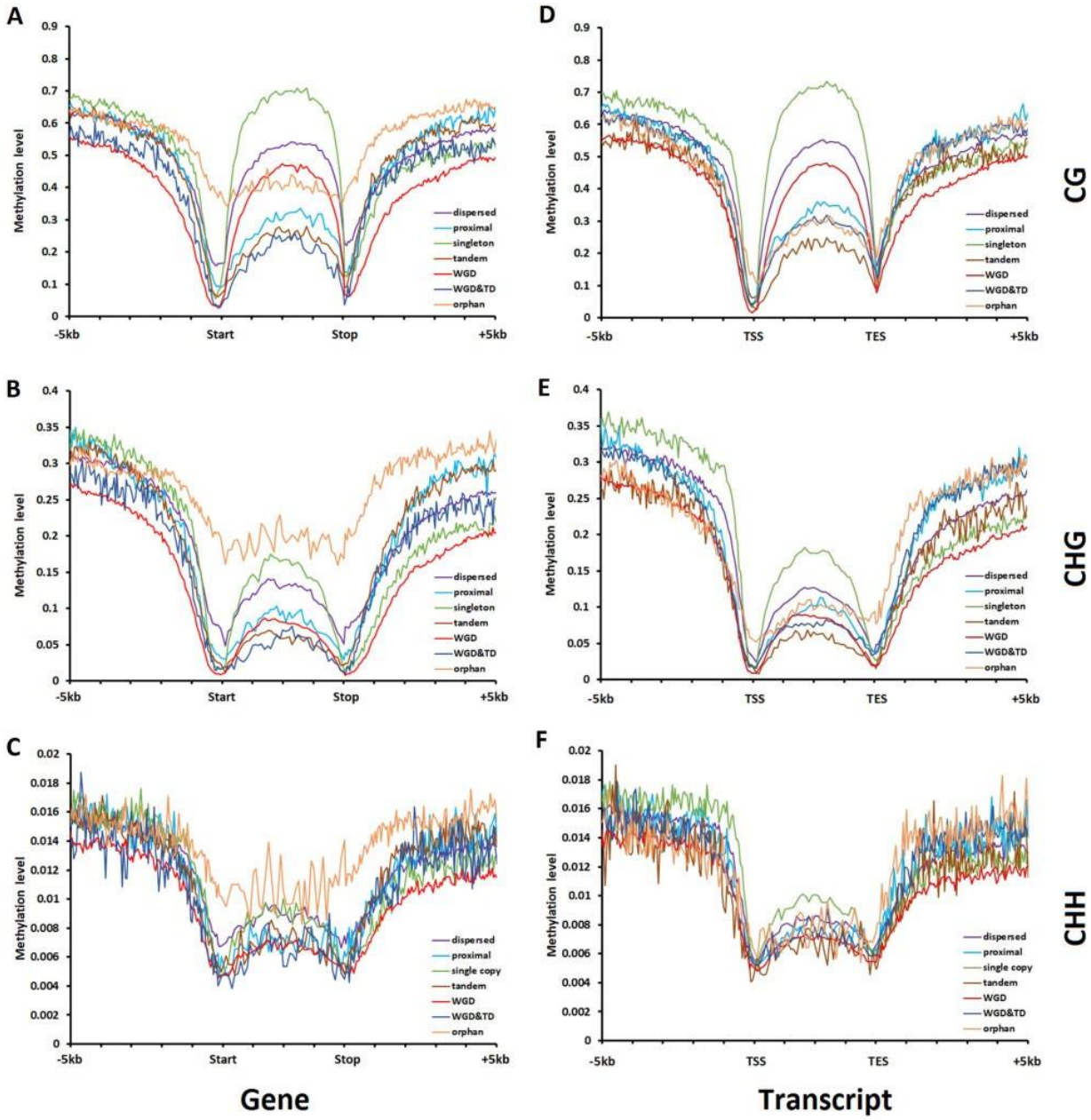


Figure 2.3 Differences in average CG, CHG, and CHH methylation level (ML) in lotus leaf along the gene and flanking regions among different gene groups based on the duplication status. (A–C) Methylation of all annotated genes. (D–F) Methylation of the genes with RNA-seq evidence.

2.2.6 WGD-derived duplicates are constrained by gene dosage balance

The evolutionary fate of duplicates is often explained employing the GBH: Genes with regulatory or signaling functions such as transcription factors or kinases will largely impact the regulatory network after a duplication because of their hub-like properties. Such duplicates are preferentially retained because the loss of one copy might disrupt many genes to which they directly or indirectly connect [18, 60]. If gene balance plays a role in the preferential retention of duplicates after a WGD, this should be reflected in the topological properties of these WGD

duplicates [17]. To assess the effect of gene balance, we analyzed the topological properties of genes belonging to each of the studied gene groups in the physical interaction network. As 27,458 out of 32,124 lotus genes (85.5%) can have the closest ortholog to corresponding Arabidopsis genes, the protein–protein interactome map from the "Arabidopsis Interactome Map" was used as a scaffold for the lotus (PPI) network (see Materials and Methods) (Arabidopsis Interactome Mapping Consortium 2011). We found that indeed genes retained after a WGD show the highest average number of PPIs (mean PPIs 1.31) (Kruskal–Wallis test, all P values < 0.01), whereas genes belonging to the other groups only differ marginally in the number of PPIs in which they tend to be involved (Figure 2.2K). Even though the analyses above suggest that, based on their relatively high expression level and breadth, single-copy genes are likely to be the more essential genes, these single-copy genes are not involved in more PPIs than genes from other groups. It appears that single-copy genes tend to immediately return to their single-copy status after a WGD because little dosage-balance constraint is imposed by the interaction network and a strong selection against gene redundancy is present [12]. Larger protein length for genes is often found to be associated with the possibility of increased interfacing with different interactors [61, 62]. Intriguingly, we also found that genes retained from a WGD have the largest average coding sequences (CDS) or protein length (Kruskal–Wallis test, all P values < 0.01), whereas genes retained after SSDs show a comparably smaller protein length, which further supports the stronger constraint of dosage balance on genes retained from a WGD (Figure 2.2L).

For the different groups of genes, we also assessed the bias in which genes are retained following duplication by calculating their Gene Ontology (GO) enrichment (K–S test with P value < 0.01). We showed that the top 30 most significantly enriched GO terms for gene groups with different duplication status have no overlapping functionalities (GO terms) (Figure S2.12). In line with the GBH, we observed that genes retained after a WGD are mostly enriched in biological terms relating to protein phosphorylation and regulation of transcription (Figure S2.12). In addition, we found that duplicates from the lotus WGD were significantly enriched in genes related to trehalose biosynthesis, polyamine biosynthesis, xylem, and phloem development (Figure S2.12). These duplications might have contributed to unique features of lotus: Because both trehalose and polyamine (metabolites) help plants to survive in stresses such as drought and cold [63-65], the unique longevity of lotus seeds and their survival during K–pg boundary might have benefited from the duplication of these biosynthesis genes. Also, the well-developed aerenchyma in stem and rhizome of lotus might have benefited from the duplication of genes related to xylem and phloem [66]. In contrast, small-scale duplicates (groups of tandem and proximal duplicates) are mostly enriched in metabolic processes, whereas genes resulting from a combination of WGD&TD are enriched in transport processes (Figure S2.12). Thus, both the PPI network and GO functional enrichment analyses suggest that gene-balance-driven selection determines the retention of duplicates after a WGD.

2.2.7 Orphan genes in lotus display unique properties

Orphan genes, comprising 7.32% of all lotus genes, are either single-copy genes or form dispersed duplicates, suggesting they are either not retained after lotus WGD or appeared after the lotus WGD (Figure S2.6A and B). They show a much lower average expression level, an elevated ratio of silent genes, and a higher expression specificity than genes with homology to known proteins (nonorphan genes) (Kruskal–Wallis test, P values of all pairwise comparisons <0.01) (Figure 2.2D–F). The relatively higher average π and the ratio of sequence deletion of orphan genes suggest that they are under more relaxed selection than genes from other groups (Figure 2.2G and H). Moreover, they have on average a shorter CDS, a shorter gene length and the lowest number of exons, implying that they are shorter and have a less complex gene structure (Figure 2.2I, J, and L). Additionally, orphan genes only display small differences in ML and TE density between their flanking regions and gene bodies (Figure 2.3 and Figure S2.9). Meanwhile, with much higher ML and TE density in gene flanking regions than non-orphan genes, it is more likely that most dispersed orphan genes were created by transposed duplications mediated by TEs (Figure 2.3 and Figure S2.11). Hence, as orphan genes exhibit features that reflect their relatively weaker functional relevancy, especially weak expression and rapid sequence turnover within lotus populations, than all nonorphan genes, they were not used to study the fate of genes after a WGD.

2.2.8 WGD-derived duplicates that have diverged in function

WGD-derived duplicates can subfunctionalize and/or neofunctionalize due to changes in the protein-coding domain, or because of regulatory changes causing divergence of expression. Here, we focused on the latter phenomenon and assessed the degree to which duplicate pairs retained from a WGD diverged in gene expression behavior. Hereto, we relied on the interconnectivity score calculated based on the coexpression network [67] (Figure 2.4A). Based on the interconnectivity score, duplicates retained after a WGD were subdivided into five groups: Gene duplicates belonging to group A (connectivity > 0.5 with a P value < 0.01) tend to share many neighbors in the coexpression network and are unlikely to have subfunctionalized or neofunctionalized. The degree of connectivity gradually decreases for duplicates belonging to groups B and C but still is larger than what can be expected by chance, given the local connectivity of the duplicate pairs under study. In contrast, duplicate pairs belonging to group D share no coexpressed neighbors and the absence of shared neighbors is significant given the local connectivity of the genes in a pair (connectivity < 0.15 and P value > 0.99). These genes diverged in expression pattern are more likely to have subfunctionalized or neofunctionalized (Figure 2.4A). Genes belonging to group E (with connectivity < 0.15 and $0.99 > x > 0.1$) show detectable connectivity in the coexpression network but this connectivity is not higher than what can be expected by chance. As for these gene pairs, it is difficult to decide whether they share coexpression neighbors, they were not considered for further analyses.

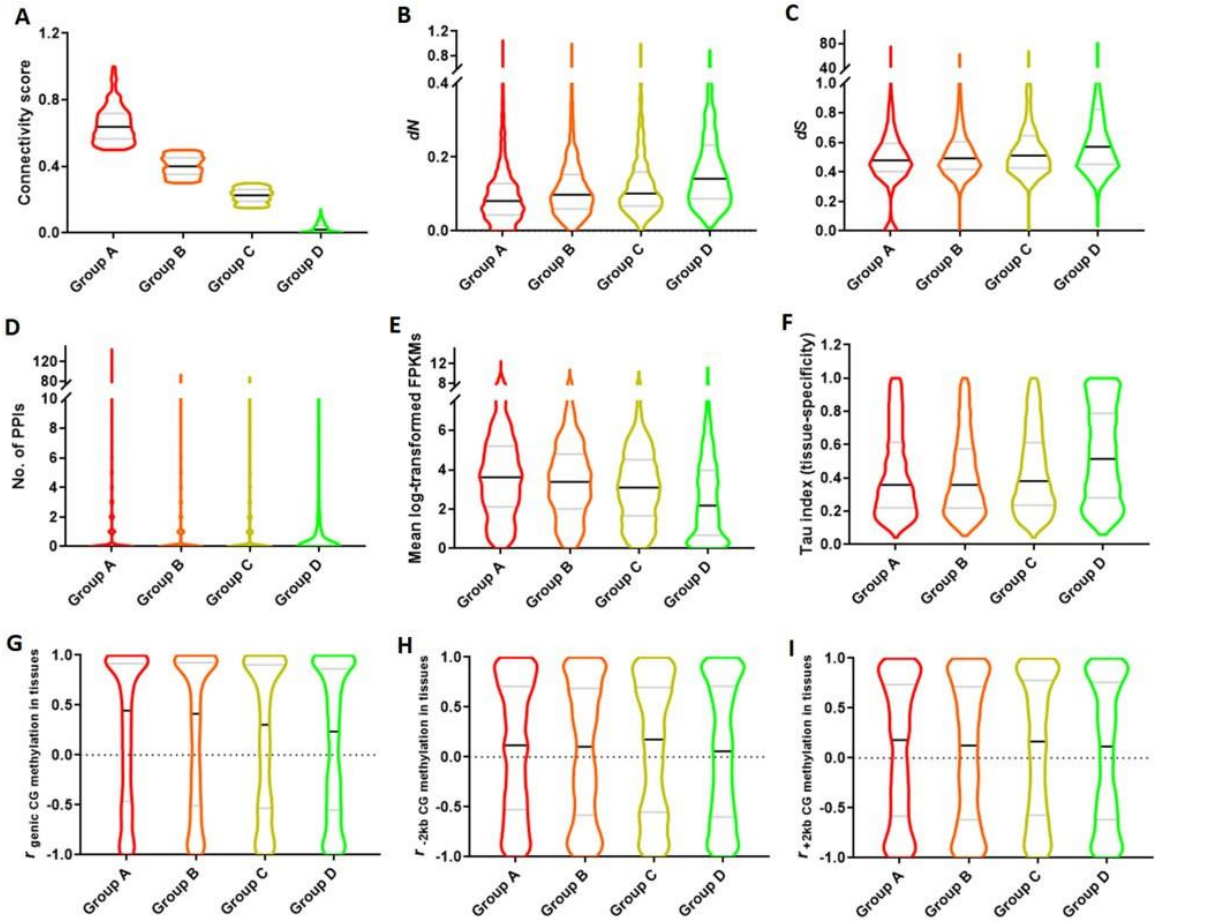


Figure 2.4 Violin plots of expression, functional, methylation, and evolutionary features of WGD-derived duplicate genes with different levels of expression divergence (group A, group B, group C, and group D). (*A*) Connectivity score. (*B*) d*N*, nonsynonymous mutation. (*C*) d*S*, synonymous mutation. (*D*) The number of protein–protein interactions inferred from the closest orthologs in *Arabidopsis*. (*E*) The mean of log-transformed FPKM. (*F*) Tissue specificity index (based on Tau index). (*G–I*) *r* (correlation coefficient) of CG methylation levels in tissues between duplicates for gene body (*G*), upstream (*H*), and downstream region (*I*). Black line: median; gray line: quantile.

To compare the degree of functional constraint on duplicates with different levels of expression divergence, we further assessed sequence and expression related characteristics for gene pairs belonging to each of the different groups (excluding group E). In line with the observed increase in expression divergence, also both the number of nonsynonymous substitutions (dN) and the number of synonymous substitutions (dS) in group A (the group with duplicates that display the most conserved expression behavior) are significantly lower than those in group D (the group most diverged in expression behavior) (Kruskal–Wallis test, all P values < 0.05), which further shows a gradual increase from group A to group D (Figure 2.4B and C). Thus, duplicate pairs that show little expression divergence tend to retain their sequence similarity (especially groups A and B). This indicates that these genes are conserved and under higher functional constraint which might be related to a relatively stronger dosage balance. We indeed also observed that duplicates that displayed the largest sequence and expression conservation

(group A) are also more frequently interacting in the PPI network than duplicates that display the most divergent expression behavior (group D) (as assessed by the degree of the duplicate genes in the PPI network) (Kruskal–Wallis test, all P values < 0.01), and accordingly a gradual decrease from group A to group D was observed, which seems in line with a previous study on WGD-derived duplicates and small-scale duplicates in Arabidopsis, tomato, and maize (Figure 2.4D) [29]. Moreover, both the average gene expression level and expression breadth (expressed as the opposite of the Tau index) in group A are significantly higher than group D (Kruskal–Wallis test, all P value < 0.01), which also exhibit a gradual change from group A to group D (Figure 2.4E and F). This indicates that duplicate pairs more conserved in their expression behavior are involved in more generic functions, whereas as expected, the duplicates more divergent in expression behavior tend to have more specialized functions. The small difference of tissue specificity (Tau index) between group A and group B might indicate they are both still under strong functional constraints (Kruskal–Wallis test, all P value 0.141). However, we did not observe that the degree of expression divergence between duplicated gene pairs belonging to different groups exhibits any significant association with overall methylation level (in tissues) or TE density (Figures S2.13–S2.15). This suggests that the gradual increase in gene expression level of duplicates from group D (less conserved in expression behavior) to group A (most conserved in expression behavior) is not related to a decline in methylation level. Because the methylation level of a gene can change in different tissues, we also calculated how the methylation pattern between duplicates is different in a well-defined region of the gene (gene body, upstream or downstream) by using correlation coefficient (r). A gene's methylation pattern is here defined as the variable of methylation levels in the four tissues on a defined region of the gene (see Materials and Methods). This analysis was performed for CG, CHG, and CHH methylation, and for each genic region separately. We found that duplicates belonging to group A (the group most conserved in expression behavior) display significantly more correlated CG methylation patterns in their genic region (with the highest r) than those of group D (Kruskal–Wallis test, all P value < 0.01), with a gradual decline from group A to group D (Figure 2.4G). This trend was not visible for the CHG nor CHH sites in upstream and downstream regions of duplicates (Figure 2.4H and I and Figure S2.16A–F). This suggests that the level to which CG methylation occurs in different tissues tends to be more conserved for duplicates that are more conserved in expression behavior. Subfunctionalized genes tend to display more differences in CG methylation level across tissues in their genic regions.

The duplicates with the most conserved expression behavior (group A) are enriched in GO terms related to protein translation (ribosome) and regulation of transcription, both functions which are known to be dosage-sensitive (Figure S2.17) [30, 68]. In contrast, the duplicates that are most diverged in expression (group D) are mainly enriched in transport mechanisms (e.g., transmembrane transport, spermine biosynthetic process, and anion transport), which are not typical dosage-sensitive functions. As a reference, we also analyzed duplicates from the *Arabidopsis* K–pg boundary WGD (At-*b*) and the recent WGD (At-*a*) with a similar strategy (using a similar grouping based on their degree of expression divergence) (Figure S2.18). In line

with our results in lotus, also here GO terms related to ribosome synthesis and regulation of transcription and biological processes are enriched in the groups representing the genes that displayed the least expression divergence after duplication (respectively group A of At-b and At-a) (Figures S2.19 and S2.20). For the duplicates from At-b (group D) that diverged most in expression, GO terms related to response to chemicals, hormone, and stimulus were most enriched, whereas for the diverged genes of At-a (group A) enriched, GO terms related to membrane, transferase activity, and oligopeptide transporter activity (Figures S2.19 and S2.20). This analysis shows that both in lotus and Arabidopsis duplicates that display the least expression divergent are related to dosage-sensitive functions, whereas the duplicated most divergent in expression (subfunctionalized) tend to have lineage-specific functions. For example, group D in lotus was enriched in "circadian regulation of calcium ion oscillation." This enrichment could be associated with the presence of four lotus genes (namely, *Nn-CRY1 a*, *b* and *Nn-CRY2 a*, *b*) being homologous to respectively Arabidopsis *Cryptochrome 1* (*CRY1*) and *Cryptochrome 2* (*CRY2*) (Figure 2.4A and Figures S2.17 and S2.21). Although *CRY1* is a flavin-type blue-light photoreceptor, participating in blue-light induced stomatal opening and thermomorphogenesis, *CRY2* is a blue/UV-A photoreceptor controlling flowering time and cotyledon expansion [69-71]. Therefore, these four circadian rhythm related genes that underwent post-WGD subfunctionalization might be associated with the lineage-specific adoption of lotus-specific characteristics related to the rigorous rhythm of flower opening and closure.

2.2.9 Subgenome dominance and fractionation

Subgenome dominance is a phenomenon in polyploids, particularly allopolyploids, in which genes are preferentially lost from one parental subgenome and for which the genes that are retained on this parental subgenome are also expressed at lower levels than their corresponding copies on the alternative parental subgenome [72]. Here, we wanted to assess whether we could find evidence for subgenome dominance in lotus. For most syntenic blocks, there are many more non-anchor genes (singlets) than anchor genes (collinear genes), suggesting that there has been extensive gene loss and genome rearrangement after the lotus WGD (Figure 2.5A). Most of the syntenic genome fragments are different in the degree to which gene duplicates are retained (retention of gene numbers), and all pairs of the syntenic regions are different in length (Figure 2.5A). Only 19 out of the 130 syntenic regions with at least six ancestral genes are significantly biased in gene retention (v2 test, $P < 0.05$), rendering it is difficult to partition syntenic genomic fragments based on the significance of gene retention (Table 2.S7). Hence, to study subgenome dominance, we instead grouped the detected syntenic genomic fragments into two groups based on their number of retained ancestral genes and length of the syntenic fragments: We distinguished a group of respectively the LF and the more fractionated (MF) regions (Figure 2.5A). Duplicated genes of which one copy has an FPKM that is twice as high as that of the alternative copy were identified. The copy with the higher FPKM was referred to as the dominant copy. Interestingly, LF fragments always have a higher ratio of copies with

dominant gene expression (mean 34.49%, SD 1.16%) than MF fragments (mean 29.97%, SD 1.16%). This subgenome dominance can be congruently observed for all 41 surveyed RNA-seq samples obtained from different tissues (Figure 2.5B). In addition, by investigating the CG, CHG, and CHH methylation and the ratio of sRNA TE and sRNA TE in both genic and flanking regions, we found that methylation level and TE density are significantly lower in the LF fragments than in the more fragmented ones (Mann–Whitney U test, all P-value < 0.01). This association between subgenome dominance and differential methylation might underly the expression bias between the two copies (Figure 2.5C–H and Figures S2.23 and S2.24). Next, we wondered whether the association between subgenome dominance and differential methylation would still hold if we would focus on the subgroups of genes that are respectively more or less subfunctionalized (where the level of subfunctionalization is proxied by the degree to which the duplicates diverged in expression behavior, see above). We noticed in the analysis performed above that duplicate pairs with more conserved expression behavior across tissues (group A) tend to have mutually more similar patterns of CG methylation levels on gene body across tissues than duplicates with more divergent expression behavior (group D). Because of the aforementioned observation, we would expect that duplicates with more conserved expression behavior would possibly display a smaller difference in methylation level between the MF and LF regions than the duplicates with more divergent expression behavior (group D), that is, group A might be less likely show subgenome dominance. However, the (most) subfunctionalized duplicate pairs (group D) do not show any remarkable differences in methylation level as compared with pairs from the other groups (Figures S2.25–S2.30). This indicates that subgenome dominance is likely a phenomenon that acts independently from subfunctionalization (as defined in this work).

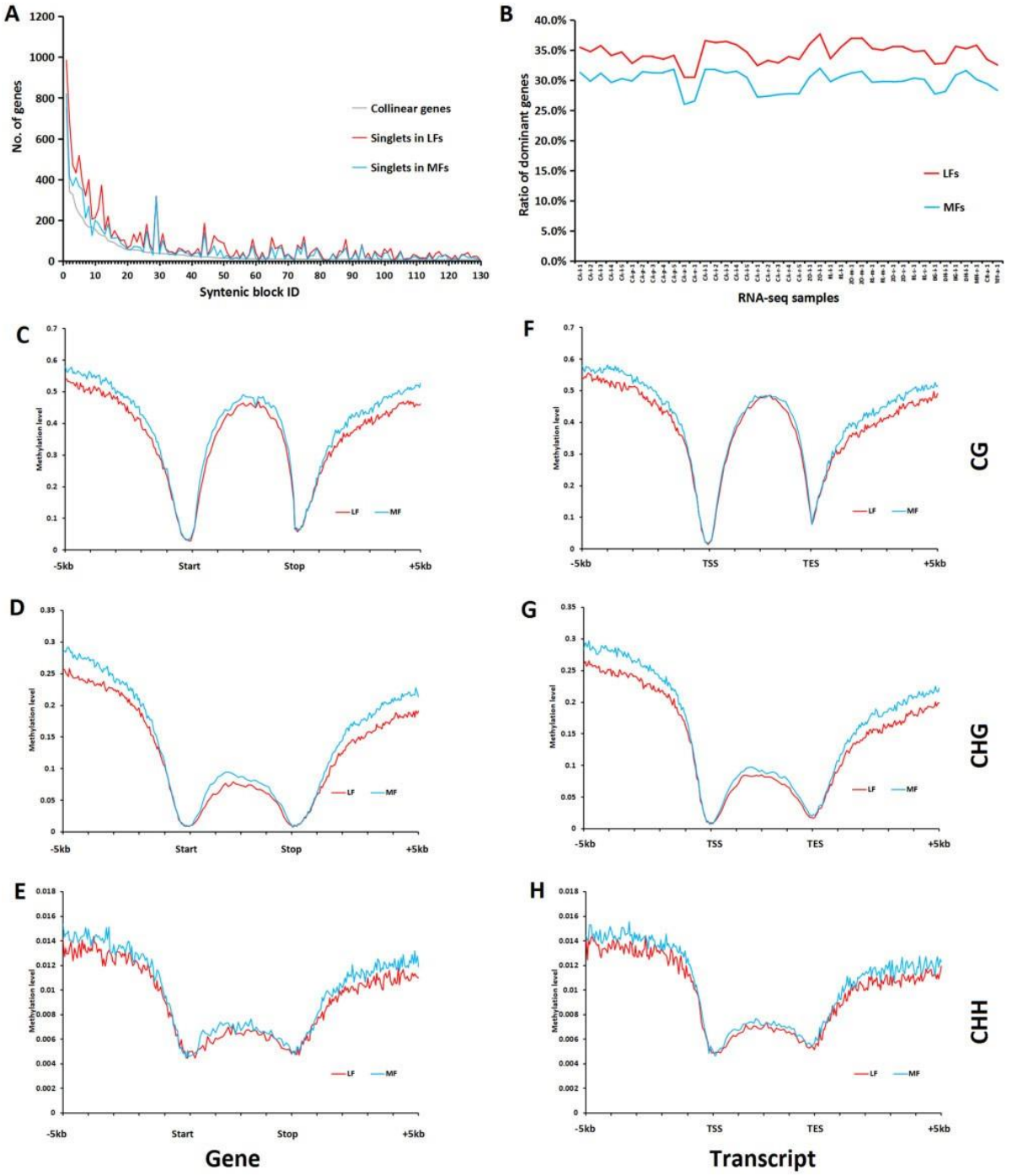


Figure 2.5 Subgenome fractionation and dominance in lotus. (*A*) Differences in the number of singlets (noncollinear) genes across 130 pairs of duplicate syntenic blocks. (*B*) The ratios of dominant copies in collinear genes between LF blocks and MF blocks across 41 RNA-seq samples. (*C–H*) Differences in average CG, CHG, and CHH methylation level in leaf along gene and flanking regions between duplicates that belong to LF blocks and MF blocks. (*C–E*) Methylation of all annotated genes. (*F–H*) Methylation of the genes with RNA-seq evidence.

2.3 Discussion

Since WGD is frequent and common during plant evolution [8, 27, 73], understanding how different genes evolve after a WGD is important for evolutionary biology. In this study, we

updated the assembly and annotation of the lotus var. “China Antique” genome by using long-read sequencing data and HI-C. This updated reference assembly largely improved the detection of collinearity to the other species, as well as within genome collinearity (relics of WGD). Notably, we performed integrative methylation and expression analyses which, when combined with all relevant genomic analyses, provide a unique opportunity to study how functional constraints and dosage balance may determine the fate of genes after a single round of WGD. We observed that single-copy genes display the highest expression level and breadth and do not show a hub-like behavior by having few protein interactors. In line with a previous study, also in lotus single-copy genes maintain their single-copy status regardless of a WGD because there appears to be a strong selection against gene redundancy [12, 13, 74]. The observed differences in expression behavior and the observed functional bias among duplicates after the WGD in lotus are in line with the GBH [14-18]. Duplicates retained after a WGD are on average more highly expressed, show a functional bias toward conservative functions shared among plant lineages such as gene transcription and signaling, have the highest number of PPIs, and are the greatest in CDS length by having the longest proteins potentially providing more interface(s) for interacting proteins. However, in keeping with previous studies, local duplicates in lotus show lower and more condition-dependent expression and are enriched in lineage-specific functions such as metabolism, disease-resistance, and other dosage-insensitive functions [31, 37, 75, 76].

The above observations are further supported by evolutionary patterns observed at the sequence level (nucleotide diversity and the ratio of sequence deletion in gene coding regions). Single-copy genes show the highest sequence conservation which is consistent with studies in Arabidopsis [12]. In addition, WGD duplicates exhibit relatively higher sequence conservation than local duplicates, agreeing with what has been observed in Arabidopsis, rice, and Populus [28, 31]. The degree to which genes display sequence conservation seems to be correlated to their expression breath rather than to their expression level. Genes that have been retained following multiple ways of duplications such as TD and WGD have been suggested to have undergone strong selection for higher dosage [77]. For instance, in lotus, the expansion of the *LPR1/2* gene by TD and WGD resulted in adaptation to a low-phosphate aquatic environment [25]. Other examples of multiple duplication events in certain gene families in *Arabidopsis* and *Brassica* have been associated with increased immunity [78]. Interestingly, among all locally duplicated genes detected in our study, genes that underwent both “WGD&TD” show significantly higher average expression levels, lower methylation levels, and lower TE densities in promoters than proximal and tandem duplicates. This suggests that also in lotus, genes that underwent both WGD and tandem duplication are selected for the higher overall gene products not only through multiple duplication events but also by other mechanisms such as transcriptional and epigenetic regulation.

Notably, we could show that the relatively higher expression level of genes retained after WGD might be associated with a differential epigenetic regulation. Cytosine methylation in genic and

flanking regions affects gene expression [55]. We observed that indeed genes that were retained after a WGD showed decreased methylation levels in gene flanking regions as compared with other gene groups explaining their higher expression level. In addition, as was observed in other studies, increased methylation was associated with a higher presence of TEs [55-57, 79-81]. In contrast to what is generally expected for flanking region methylation to repress gene expression, we found that lotus single-copy genes which are the most abundantly expressed were also the most abundantly methylated in their gene bodies. This has also been observed in rice (Wang Y, Wang X, et al. 2013). So in lotus, it appears that gene body methylation of single-copy genes seems to induce expression rather than repressing it [82-84]. In lotus, the observed gene body methylation pattern of single-copy genes is also associated with the presence of TEs. The abundant methylation on the gene bodies (genic regions) for single-copy genes could be associated with a similar TE distribution and the presence of abundant introns, indicating that methylation is involved in silencing TEs. Alternatively, it has been shown that gene body methylation can enhance splicing accuracy by improving the distinction of exon–intron boundaries [51, 52]. This might be particularly relevant in maintaining the functional integrity of single-copy genes, given their high intron number [51, 52]. However, future functional and genetic studies on TEs and introns of single-copy genes are necessary to support these hypotheses.

Lotus orphan genes were treated separately in the current study because of their evolutionary transience. The lotus WGD occurred 66 Ma after the split with its closest sequenced relative, *Macadamia*, about 111 Ma [25, 85]. Their low expression level, high expression specificity, and high methylation level imply that orphan genes tend to be transcriptionally repressed to avoid producing nonfunctional peptides (proteins) and that they are not required in most tissues or organs. Their small protein size, gene length, and exon number are consistent with observations in *Drosophila* and *Arabidopsis* [86-88]. Although their high nucleotide diversity suggests relatively low functional importance, their functionality cannot be excluded [89, 90].

Given that long-retained duplicates from a WGD are important genetic material for plant innovation and evolution, our current study further focused on how those retained duplicates diverge in expression pattern and level across different lotus tissues. Whereas maintaining gene balance plays right after WGD, subfunctionalization and neofunctionalization explain the long-term evolution of duplicates retained from WGD [3, 14, 30, 91-93]. In lotus, WGD duplicates displayed a continuous spectrum of expression divergence where some duplicates share largely the same coexpression partners, whereas other duplicates display a completely distinct expression pattern. Lotus duplicates that display lower expression divergence tend to correspond to the hubs of PPI networks, have relatively longer protein and gene lengths, display higher average expression levels and breadth, more similar pattern of changeof CG methylation in gene bodies across different tissues between duplicate pairs, relatively low pairwise amino acid sequence divergence and low nucleotide diversity, which all support they are under a stronger gene balance constraints [16]. Many of these observations are in

accordance with studies in, for instance, Arabidopsis, maize, and tomato [29]. Yet, in contrast to lotus, these plants underwent sequential rounds of WGDs which makes it difficult to study the fate of the most ancient duplicates. Hence, the fact that the same findings made in these other species are also observed in lotus indicates that they are truly associated with the fate of ancient duplicates. Subgenome dominance can be an important source of bias in expression level between duplicated gene pairs retained from a WGD and can result in significant differences in gene retention (content), the intensity of TE insertion, methylation, and population-level polymorphisms between subgenomes [34, 40, 85, 94, 95]. Depending on the studied species, the level of subgenome fractionation that occurs after a WGD can be significantly different, ranging from extensive fractionation in, for example, monkeyflower (WGD estimated at 140 Ma), maize (11.9 Ma), Brassica (13–17 Ma), Arabidopsis (40 Ma), and cotton (60 Ma) [34, 40, 85, 94, 95] to little sub-genome fractionation in, for example, soybean (5–13 Ma), banana (65 Ma), and poplar (8 Ma) fractionation [95, 96]. In our study, about 14.6% of syntenic block pairs in lotus show significant bias in fractionation, a level which is in between the fraction observed in the paleoautopolyploid soybean (5.4%) and the paleoallopolyploid maize (31%) [95]. The LF blocks show on average about 4.52% more (expression) dominant copies than the MF blocks, which is a difference that is higher than what is observed in soybean (0–1%) but lower than in maize (~10%) [95]. As the extent of biases in lotus (66 Ma) is between a paleoautopolyploid and a paleoallopolyploid, likely, its two ancestral parental genomes had already diverged to some extent before the formation of ancient polyploid. So far, lotus shows evidence of one of the oldest appearances of subgenome dominance among the above-mentioned plant genomes.

2.4 Materials and Methods

2.4.1 Plant material, PacBio Sequel, and Hi-C sequencing

Sacred lotus ‘China Antique’ was grown and collected from Wuhan Botanical Garden of the Chinese Academy of Sciences. DNA from young leaves of ‘China Antique’ was extracted using Plant DNA Isolation Reagent (Tiangen, China). Two DNA libraries (insert sizes of 10,123 bp and 10,157 bp, respectively) were constructed according to the PacBio library preparation protocol and sequenced on a PacBio Sequel platform (Pacific Biosciences, USA) at Annoroad Genomics (Beijing, China). Subreads with quality score under 0.8 were discarded. The HI-C DNA library of ‘China Antique’ was prepared at Annoroad Genomics (Beijing, China) under a previously published protocol [97]. Briefly, nuclear DNA of young lotus leaves was cross-linked inside the tissue cell sample. Then, the extracted DNA was digested with the restriction enzyme (HindIII/MboI). Biotinylation was tagged at the sticky ends of the digested DNA fragments, and then mutually ligated at random after dilution. The library of condensed, sheared, and biotinylated DNA fragments were prepared for paired-end (PE) sequencing with 150 bp reads on Illumina HiSeq platform.

2.4.2 Chromosome-level assembly

All contigs were assembled using PacBio and Illumina reads. SparseAssembler was applied to assemble Illumina PE reads of lotus ‘China Antique’ into short but accurate Illumina contigs [98, 99]. These Illumina contigs and PacBio Sequel reads were co-assembled into longer contigs with the hybrid assembly tool DBG2OLC [100]. Errors in these hybrid contigs were further polished with all Illumina PE reads using BWA-MEM and Pilon 2.10 [101, 102]. The HI-C sequencing reads were mapped on the ‘China Antique’ hybrid assembly contigs using BWA-MEM [101]. Finally, the chromosome-level scaffolding of these contigs were performed with LACHESIS [103]. Additional gaps in pseudochromosomes were filled with PacBio subreads using Jelly and polished with Illumina reads using Pilon 2.10 [104].

2.4.3 Repeat annotation

Repeats including transposable elements on the new ‘China Antique’ assembly were annotated following a previously published protocol [105]. Generally, MITEs (miniature inverted repeat transposable elements) were predicted using MITE-Hunter under default settings [106]. The most abundant plant TEs (transposable elements), LTRs (long terminal repeat retrotransposons) were collected, false-positives were filtered, and redundancy was reduced using LTR-harvest, LTR-digest and the Perl scripts provided by the protocol ‘Repeat Library Construction-Advanced’ (http://weatherby.genetics.utah.edu/MAKER/wiki/index.php) [107, 108]. Other repeats were collected by RepeatModeler (http://www.repeatmasker.org). Gene fragments in all collected repeats were excluded by searching against all plant protein sequences from Plant Plaza 4.0 [109]. After collecting and building the lotus repeat database, the genome was further annotated using RepeatMasker (http://www.repeatmasker.org).

2.4.4 Gene annotation

Protein-coding genes on the ‘China Antique’ assembly were annotated based on (1) RNA-seq mapping, (2) protein homology searches and (3) ab initio prediction. For gene prediction with transcriptional evidence, 41 publicly available RNA-seq data from leaf, petioles, rhizome, root and apical bud of lotus were downloaded from NCBI SRA database and mapped on the genome using HISAT2-StringTie pipeline [110, 111]. Transcript coordinates from different RNA-seq samples were further merged using TACO [112]. Coding region of transcripts were annotated using Transdecoder (https://github.com/TransDecoder). Homology-based gene prediction was performed using GeMoMa with genome sequences and gene coordinates from *Arabidopsis thaliana*, *Carica papaya*, *Vitis vinifera*, *Macadamia ternifolia* (Proteales) and *Brachypodium distachyon* as input [109, 113, 114]. *Ab initio* gene prediction was performed using Braker2 which took in intron hints from transcript coordinates of RNA-seq based assemblies [115]. The final consensus gene annotations were produced by EVidenceModeler with weights of ‘RNA-seq > gene homology > *ab initio*’ [116]. Gene ontology annotations were further performed using the 'non-redundant' database of plants via BLAST2GO with default settings [117]. The

lotus interactome was inferred using PPI data from Arabidopsis by top BLAST hit (best homologs) (Arabidopsis Interactome Mapping Consortium) [118].

2.4.5 Validation of genome assembly

Accuracy and structural completeness of the new genome assembly were assessed using (1) previously published SNP markers from genetic linkage groups, (2) ratio of genome collinearity with other species and (3) conserved single copy genes of plant from BUSCO. For comparison, SNP markers from a high density lotus genetic map from a previous study were downloaded and mapped onto the new and old 'China Antique' assemblies [41] using bowtie allowing no mismatch other than SNP site [119]. Collinearity between the genetic map and the genome assembly was anchored by SNP markers. Distributions of SNP markers on genome assemblies were inspected by bar plots, and collinearity was visualized by dot plots. To assess the genome architecture using genome collinearity, we searched homologous genomic blocks in genomes of two monocots, *Oryza sativa* (rice) and *Brachypodium distachyon*, against the new (v2018) and old (v2013) 'China Antique' assemblies using MCScanX [120]. First, potential anchors between the two genomes were identified using BLASTP (E <1e−5). Then, MCscanX found all orthologous syntenies with at least six anchor points. For further comparisons, orthologous syntenies between other eudicots species and the two monocot representatives were downloaded from Plant Genome Duplication Database [121]. To assess the completeness of the gene regions in the assembly, 1440 conserved plant single copy genes as benchmark were searched using BUSCO v2 [122].

2.4.6 Classification of genes by duplication status

Duplicated genes in extant genomes typically originated through different duplication events. Depending on the size of the genomic regions involved in the duplication event, a distinction is made between small (SSD) and large-scale duplications (LSD). LSD can be maintained as syntenies which likely are retained from WGD. Within SSD, a distinction is made between local (tandem and proximal duplication) versus dispersed duplications [17, 43]. Tandem duplicates lead to a cluster of two or more consecutive paralogous sequences while proximal having one or a few intervening genes. Dispersed duplicates are mainly unclassified duplicates. Genes that underwent both WGD and tandem duplications often exist, which we refer to as 'WGD&TD' [123, 124].

To identify ancient genome duplication of lotus, homologs were first identified by all-against-all BLASTP for syntenic anchors (E <1e−5). Intra-specific syntenic blocks were identified with same approach as the one used for the identification of orthologous synteny described above using MCscanX [120]. To identify WGDs, raw 4dTv (the number of transversion 4-fold degenerative sites) of all syntenic paralogous pairs were calculated and further corrected for possible multiple transversions at the same site based on a previous method [125]. A histogram of 4dTv for all syntenic paralogs was plotted with a bin size of 0.01. To classify syntenic blocks according to WGDs, the median of 4dTv of each syntenic block were used. Syntenic blocks

with less than six duplicate pairs with valid 4dTv after correction were classified as syntenies of uncertain origin. Other divergence parameters including *dS* or K_s (synonymous substitution rate), *dN* or K_a (nonsynonymous substitution rate) and *dN/dS* for all syntenic paralogs were calculated using codeML from PAML package [126]. Further, 4dTv of orthologous divergence were also plotted in histograms. For the fragmented genome assembly *Macadamia ternifolia*, orthologous pairs were predicted using OrthoMCL [127, 128]. The chronological order of WGDs and species split (*Nelumbo* versus *Macademia*) were confirmed by Mann-Whitney *U* test based on rate calibrated 4dTv.

Single copy genes and genes of other duplication status including those originating from dispersed duplication, tandem duplication and proximal duplication events, WGD&TD were also detected by MCscanX [120]. All lotus genes without homology to other sequenced species were defined as orphan genes, while the rest was regarded as non-orphan genes whose ancestral proteins appeared at least before the split of lotus and *Macademia* (111 mya). The family Nelombonaceae (in Proteales) is a species-poor clade with only two closely related *Nelumbo* species. To obtain lotus orphan genes, *Macadamia ternifolia* (the other only sequenced Proteales genome) and PlantPlaza 3.0 database were used in phylostratigraphic analyses [129]. The groups of genes of different duplication status were used for subsequent comparative analyses. As most orphan genes are evolutionarily transient, they were analyzed separately [129].

To explore the dosage sensitivity for our studied groups (subdivided as described above), we defined for each lotus gene its orthologs in *Macademia* and *Vitis vinifera* using OrthoMCL [127, 128]. For each gene we calculated the average copy number of its orthologs in the three taxa. For each lotus gene, its coefficient of variation (C.V.) in the number of copies observed in the three species was used to estimate the dosage sensitive. For each of the gene groups studied, the average copy number and C.V. were reported.

2.4.7 Nucleotide diversity and ratio of sequence deletion of lotus genes

Illumina re-sequencing data from 18 Asian lotus individuals including rhizome lotus, flower lotus, seed lotus, wild lotus, and Thai lotus were downloaded from NCBI (Supplementary Data) [130]. Illumina reads were mapped to the new ‘China Antique’ assembly using BWA-mem [101]. Mapped files were processed by Picard (https://broadinstitute.github.io/picard/). The SNP variants were called by HaplotypeCaller of GATK 3.7 with further hard filtering of ‘QD < 2 || FS > 60 || MQ < 30’ [131]. Nucleotide diversity (π) of each CDS from each annotated gene was estimated using Popgenome [132] in R. The ratio of sequence deletion for each gene is calculated by the ratio of InDel length in each CDS. A Mann-Whitney *U* test or (non-parametric) Kruskal-Wallis test was applied to compare average π and average ratio of sequence deletion of gene CDS between (among) different gene groups in Graphpad PRISM 7.

2.4.8 Expression analysis

All 41 RNA-seq samples used for gene annotation were also used for expression analyses. The average expression level per gene for each gene group showing different fates after the lotus WGD was estimated by the average log-transformed FPKM values of the genes in a group. Tissue specific expression was assessed by the Tau index [133] To define 'tissue' we clustered the 41 samples using the log transformed FPKM data (Euclidean distance). Samples clustered in 8 distinct tissue groups: leaf, petiole, apical, rhizome internode, root, rhizome (later swelling), rhizome (middle swelling), rhizome (stolon). The Tau index was calculated as follows:

$Tau = \frac{\sum_{i=1}^{n}(1-\hat{x_i})}{n-1}$;

where $\hat{x} = \frac{x_i}{max_{1 \le i \le n}(x_i)}$ and $x_i$is the expression for tissue i.

To build the coexpression network, the genes with an expression value in at least in 2 samples were retained (28578 are present in the network with 480849 edges). The 'rank of correlation coefficient' (Obayashi and Kinoshita, 2009) was used to determine the degree of pairwise coexpression. To calculate the rank-based correlation the gene-gene Pearson correlation matrix derived from the log2-transformed FPKM values was transformed into a rank matrix. For every gene-gene combination, the Mutual Rank score was calculated using the following formula:

$$MR(AB) = \sqrt{rank_{(A \to B)} * rank_{(B \to A)}}$$

where $Rank_{(A \to B)}$ is the rank of the correlation of gene B with gene A as compared to its correlation with all other genes [134]. Smaller MR scores correspond to a higher degree of pairwise correlation between two genes and can be converted to a network edge weight using the following formula

$$weight_{(A \to B)} = e^{-(MR_{(A \to B)}-1)/10}$$

guaranteeing that the range of edge weights in the coexpression network scales between 0 and 1. A small value (one) was added to the FPKM values prior to the log transformation in order to avoid having undefined values of the rank-based correlation for zero values.

2.4.9 Grouping WGD genes based on their expression behavior

Post-WGD duplicate pairs were subdivided into groups based on their expression divergence. To assess the degree to which duplicate pairs diverged in expression, we used an interconnectivity score [67]. The interconnectivity between a pair of duplicated genes assesses the degree to which two duplicate genes share neighbors in the coexpression network. The higher the connectivity score, the more the duplicates are assumed to share the same expression profile.

$$CN(i,j) = \frac{N(i) \cap N(j)}{\sqrt{N(i) * N(j)}}$$

where N(i) and N(j) describe the number of neighbors that are located at most three edges distance of respectively the nodes i and j in the duplicate pair (i, j). The number of shared neighbors between the genes of the duplicate pair is normalized by the total number of neighbors of the two genes in the duplicate pair.

In addition, we determined for each duplicate gene pair whether the number of shared neighbors that contributed to the connectivity measure is statistically significant using the hypergeometric test: for every duplicate pair, the number of up to third order neighbors for one gene N_A was determined and used to calculate the chance of a success p = Na/N where N is the total number of genes in the genome. The number of third order neighbors for the second gene in the pair (B) was used as the number of trials and the number of neighbors shared between A and B was considered the number of successes. Using this parameters, the cumulative mass function was calculated to calculate the p-value i.e., observing the same number of shared neighbors between two genes just by chance. Based on the combination of the hypergeometric P value and the connectivity score the duplicates were subdivided in 5 groups: group A with connectivity >0.5 and p-value <0.01, group B with connectivity 0.5>x>0.3 and p-value <0.01, group C with 0.3>x>0.15 and p-value <0.01, group D with connectivity <0.15 and p-value >0.99 and group E with connectivity <0.15 and 0.99>x>0.1. Group E contains the genes that show a certain but insignificant connectivity. This category was not retained for further analysis. Duplicate genes that belong to Group A, B and C share coexpressed neighbors (more for groupA >B>C) and they share more neighbors in the coexpression network than can be expected by chance given the local connectivity of the genes in the pair. Genes belonging to group D show a significant low to no connectivity in the coexpression network.

2.4.10 Comparisons of different gene features

Genomic traits including the length of the CDS, the number of exons and the gene length were directly obtained from the lotus genome annotation. Given that there is currently no protein interactome map for lotus, for those lotus genes displaying homology to Arabidopsis genes, their number of PPIs were inferred from the closest homolog in Arabidopsis (Arabidopsis Interactome Mapping Consortium, 2011) [118]. Genomic traits (CDS length, gene length, exon number) and evolutionary parameters (dN, dS, dN/dS, π) were summarized and compared between different genes of different groups using (non-parametric) Kruskal-Wallis test in Graphpad PRISM 7.

2.4.11 sRNA+ Transposable Element (TE), sRNA−TE and methylation level analyses of gene duplicates

To test whether TE insertion and methylation level differences might contribute to duplicate gene expression difference, firstly, small RNAs of ‘China Antique’ were mapped to the genome

using bowtie with zero tolerance of mismatch [26]. Only uniquely mapped sRNAs were used to define TEs [34]. TEs were classified into sRNA+ TEs and sRNA-TEs based on whether there was any sRNA aligning to them. Gene flanking regions (± 5 kb) and gene bodies (translation start to translation stop site) were analyzed using sliding window. The overlapping region(s) between flanking region and the neighboring gene body were excluded from flanking region analyses. For each gene, a 100-bp sliding window with 10-bp step of each 5′ and 3′ flanking region and the 40 evenly divided windows of gene body (from translation start site to translation stop site) was used [135]. For each sliding window, the proportion of the sequence being composed by sRNA+ TE or sRNA-TEs was calculated. The average proportion in each sliding window was calculated for each gene group under investigation. These averaged proportions were then used to estimate the TE density in the flanking regions and gene body of different gene groups. Meanwhile, whole genome methylation was analyzed based on bisulfite sequencing (BS-seq) on young leaves from a wild lotus (Khabarovsk, Russia), flanking region was evenly divided into 100 50-bp windows, and the gene body (from translation start site to translation stop site) was evenly divided into 40 windows [135]. We included both exon and intron for methylation level of gene body because pre-mRNAs, being transcribed from DNA, contain introns. Methylation level including CG, CHG and CHH sites of different gene groups was estimated using BS-Seeker2 and cgmaptools for each window [136, 137].

2.4.12 Subgenome fractionation and dominance

Subgenome fractionation bias was analyzed as outlined previously [96.]. Numbers of collinear genes and non-collinear genes for pairs of syntenic blocks were tested for significant fractionation bias (χ^2 Test). Differences in TE ratio and methylation between collinear genes in less fractionated (LF) and more fractionated (MF) syntenic blocks were analyzed with the same approach described above. Subgenome fractionation bias is often associated with subgenome dominance. To test subgenome dominance, all 41 RNA-seq samples were used. For each RNA-seq sample, the dominant copy was defined as the one showing an expression that was more than two fold higher than the expression level of the alternative copy (FPKM). Further, for each RNA-seq sample, the ratios of dominant copies in LFs and MFs were summarized and compared.

Author contributions

T.S., K.M., Y.V.d.P., Q.W., and J.C. conceptualized the study; T.S., R.S.R., and K.M. performed the methodology; T.S., H.L., Y.Z., and Z.L. performed the investigation; T.S. and R.S.R. performed the formal analysis; T.S. wrote the original draft; K.M., Y.V.d.P., M.W., P.F.G., and J.C. reviewed and edited the manuscript; Q.W., Y.V.d.P., K.M., and J.C. supervised the study.

Supplementary Data: Supplementary data are available at:

https://academic.oup.com/mbe/article/37/8/2394/5826357?login=true#supplementary-data

2.4.13 References

1. Shi T, Rahmani RS, Gugger PF, Wang M, Li H, Zhang Y, Li Z, Wang Q, Van de Peer Y, Marchal K: **Distinct expression and methylation patterns for genes with different fates following a single whole-genome duplication in flowering plants.** *Molecular biology and evolution* 2020, **37:**2394-2413.
2. Blanc G, Wolfe KH: **Functional divergence of duplicated genes formed by polyploidy during Arabidopsis evolution.** *The Plant Cell* 2004, **16:**1679-1691.
3. Gout J-F, Lynch M: **Maintenance and loss of duplicated genes by dosage subfunctionalization.** *Molecular biology and evolution* 2015, **32:**2141-2148.
4. Ohno S: *Evolution by gene duplication.* Springer Science & Business Media; 2013.
5. Sandve SR, Rohlfs RV, Hvidsten TR: **Subfunctionalization versus neofunctionalization after whole-genome duplication.** *Nature genetics* 2018, **50:**908-909.
6. Shiu S-H, Bleecker AB: **Receptor-like kinases from Arabidopsis form a monophyletic gene family related to animal receptor kinases.** *Proceedings of the National Academy of Sciences* 2001, **98:**10763-10768.
7. Zhang J: **Evolution by gene duplication: an update.** *Trends in ecology & evolution* 2003, **18:**292-298.
8. Cui L, Wall PK, Leebens-Mack JH, Lindsay BG, Soltis DE, Doyle JJ, Soltis PS, Carlson JE, Arumuganathan K, Barakat A: **Widespread genome duplications throughout the history of flowering plants.** *Genome research* 2006, **16:**738-749.
9. Fawcett JA, Maere S, Van De Peer Y: **Plants with double genomes might have had a better chance to survive the Cretaceous–Tertiary extinction event.** *Proceedings of the National Academy of Sciences* 2009, **106:**5737-5742.
10. Jiao Y, Wickett NJ, Ayyampalayam S, Chanderbali AS, Landherr L, Ralph PE, Tomsho LP, Hu Y, Liang H, Soltis PS: **Ancestral polyploidy in seed plants and angiosperms.** *Nature* 2011, **473:**97-100.
11. Ruprecht C, Lohaus R, Vanneste K, Mutwil M, Nikoloski Z, Van de Peer Y, Persson S: **Revisiting ancestral polyploidy in plants.** *Science Advances* 2017, **3:**e1603195.
12. De Smet R, Adams KL, Vandepoele K, Van Montagu MC, Maere S, Van de Peer Y: **Convergent gene loss following gene and genome duplications creates single-copy families in flowering plants.** *Proceedings of the National Academy of Sciences* 2013, **110:**2898-2903.
13. Paterson AH, Chapman BA, Kissinger JC, Bowers JE, Feltus FA, Estill JC: **Many gene and domain families have convergent fates following independent whole-genome duplication events in Arabidopsis, Oryza, Saccharomyces and Tetraodon.** *Trends in genetics* 2006, **22:**597-602.
14. Bekaert M, Edger PP, Pires JC, Conant GC: **Two-phase resolution of polyploidy in the Arabidopsis metabolic network gives rise to relative and absolute dosage constraints.** *The Plant Cell* 2011, **23:**1719-1728.
15. Birchler JA, Veitia RA: **Gene balance hypothesis: connecting issues of dosage sensitivity across biological disciplines.** *Proceedings of the National Academy of Sciences* 2012, **109:**14746-14753.
16. De Smet R, Van de Peer Y: **Redundancy and rewiring of genetic networks following genome-wide duplication events.** *Current opinion in plant biology* 2012, **15:**168-176.
17. Freeling M: **Bias in plant gene content following different sorts of duplication: tandem, whole-genome, segmental, or by transposition.** *Annual review of plant biology* 2009, **60:**433-453.

18. Tasdighian S, Van Bel M, Li Z, Van de Peer Y, Carretero-Paulet L, Maere S: **Reciprocally retained genes in the angiosperm lineage show the hallmarks of dosage balance sensitivity.** *The Plant Cell* 2017, **29:**2766-2785.
19. Lan T, Renner T, Ibarra-Laclette E, Farr KM, Chang T-H, Cervantes-Pérez SA, Zheng C, Sankoff D, Tang H, Purbojati RW: **Long-read sequencing uncovers the adaptive topography of a carnivorous plant genome.** *Proceedings of the National Academy of Sciences* 2017, **114:**E4435-E4441.
20. Qiao X, Li Q, Yin H, Qi K, Li L, Wang R, Zhang S, Paterson AH: **Gene duplication and evolution in recurring polyploidization–diploidization cycles in plants.** *Genome biology* 2019, **20:**1-23.
21. Olsen JL, Rouzé P, Verhelst B, Lin Y-C, Bayer T, Collen J, Dattolo E, De Paoli E, Dittami S, Maumus F: **The genome of the seagrass Zostera marina reveals angiosperm adaptation to the sea.** *Nature* 2016, **530:**331-335.
22. Aköz G, Nordborg M: **The Aquilegia genome reveals a hybrid origin of core eudicots.** *Genome biology* 2019, **20:**1-12.
23. Chen J, Hao Z, Guang X, Zhao C, Wang P, Xue L, Zhu Q, Yang L, Sheng Y, Zhou Y: **Liriodendron genome sheds light on angiosperm phylogeny and species–pair differentiation.** *Nature plants* 2019, **5:**18-25.
24. Zhang L, Chen F, Zhang X, Li Z, Zhao Y, Lohaus R, Chang X, Dong W, Ho SY, Liu X: **The water lily genome and the early evolution of flowering plants.** *Nature* 2020, **577:**79-84.
25. Ming R, VanBuren R, Liu Y, Yang M, Han Y, Li L-T, Zhang Q, Kim M-J, Schatz MC, Campbell M: **Genome of the long-living sacred lotus (Nelumbo nucifera Gaertn.).** *Genome biology* 2013, **14:**1-11.
26. Shi T, Wang K, Yang P: **The evolution of plant microRNAs: insights from a basal eudicot sacred lotus.** *The plant journal* 2017, **89:**442-457.
27. Vanneste K, Baele G, Maere S, Van de Peer Y: **Analysis of 41 plant genomes supports a wave of successful genome duplications in association with the Cretaceous–Paleogene boundary.** *Genome research* 2014, **24:**1334-1347.
28. Wang Y: **Locally duplicated ohnologs evolve faster than nonlocally duplicated ohnologs in Arabidopsis and rice.** *Genome biology and evolution* 2013, **5:**362-369.
29. Defoort J, Van de Peer Y, Carretero-Paulet L: **The evolution of gene duplicates in angiosperms and the impact of protein–protein interactions and the mechanism of duplication.** *Genome biology and evolution* 2019, **11:**2292-2305.
30. Jiang W-k, Liu Y-l, Xia E-h, Gao L-z: **Prevalent role of gene features in determining evolutionary fates of whole-genome duplication duplicated genes in flowering plants.** *Plant physiology* 2013, **161:**1844-1861.
31. Rodgers-Melnick E, Mane SP, Dharmawardhana P, Slavov GT, Crasta OR, Strauss SH, Brunner AM, DiFazio SP: **Contrasting patterns of evolution following whole genome versus tandem duplication events in Populus.** *Genome research* 2012, **22:**95-105.
32. Lehti-Shiu MD, Uygun S, Moghe GD, Panchy N, Fang L, Hufnagel DE, Jasicki HL, Feig M, Shiu S-H: **Molecular evidence for functional divergence and decay of a transcription factor derived from whole-genome duplication in Arabidopsis thaliana.** *Plant physiology* 2015, **168:**1717-1734.
33. Bottani S, Zabet NR, Wendel JF, Veitia RA: **Gene expression dominance in allopolyploids: hypotheses and models.** *Trends in Plant Science* 2018, **23:**393-402.
34. Cheng F, Sun C, Wu J, Schnable J, Woodhouse MR, Liang J, Cai C, Freeling M, Wang X: **Epigenetic regulation of subgenome dominance following whole genome triplication in Brassica rapa.** *New Phytologist* 2016, **211:**288-299.
35. Edger PP, Smith R, McKain MR, Cooley AM, Vallejo-Marin M, Yuan Y, Bewick AJ, Ji L, Platts AE, Bowman MJ: **Subgenome dominance in an interspecific hybrid, synthetic allopolyploid, and a 140-year-old naturally established neo-allopolyploid monkeyflower.** *The Plant Cell* 2017, **29:**2150-2167.
36. Flagel LE, Wendel JF: **Evolutionary rate variation, genomic dominance and duplicate gene expression evolution during allotetraploid cotton speciation.** *New Phytologist* 2010, **186:**184-193.

37. Langham RJ, Walsh J, Dunn M, Ko C, Goff SA, Freeling M: **Genomic duplication, fractionation and the origin of regulatory novelty.** *Genetics* 2004, **166:**935-945.
38. Rapp RA, Udall JA, Wendel JF: **Genomic expression dominance in allopolyploids.** *BMC biology* 2009, **7:**1-10.
39. Vicient CM, Casacuberta JM: **Impact of transposable elements on polyploid plant genomes.** *Annals of Botany* 2017.
40. Woodhouse MR, Cheng F, Pires JC, Lisch D, Freeling M, Wang X: **Origin, inheritance, and gene regulatory consequences of genome dominance in polyploids.** *Proceedings of the National Academy of Sciences* 2014, **111:**5283-5288.
41. Liu Z, Zhu H, Liu Y, Kuang J, Zhou K, Liang F, Liu Z, Wang D, Ke W: **Construction of a high-density, high-quality genetic map of cultivated lotus (Nelumbo nucifera) using next-generation sequencing.** *BMC genomics* 2016, **17:**1-11.
42. Wang X, Zhang Z, Fu T, Hu L, Xu C, Gong L, Wendel JF, Liu B: **Gene-body CG methylation and divergent expression of duplicate genes in rice.** *Scientific reports* 2017, **7:**1-11.
43. Wang Y, Wang X, Lee TH, Mansoor S, Paterson AH: **Gene body methylation shows distinct patterns associated with different gene origins and duplication modes and has a heterogeneous relationship with gene expression in Oryza sativa (rice).** *New Phytologist* 2013, **198:**274-283.
44. Jaillon O, Aury J-M, Noel B, Policriti A, Clepet C, Cassagrande A, Choisne N, Aubourg S, Vitulo N, Jubin C: **The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla.** *nature* 2007, **449:**463-467.
45. Coate JE, Song MJ, Bombarely A, Doyle JJ: **Expression-level support for gene dosage sensitivity in three Glycine subgenus Glycine polyploids and their diploid progenitors.** *New Phytologist* 2016, **212:**1083-1093.
46. Akashi H: **Gene expression and molecular evolution.** *Current opinion in genetics & development* 2001, **11:**660-666.
47. Drummond DA, Bloom JD, Adami C, Wilke CO, Arnold FH: **Why highly expressed proteins evolve slowly.** *Proceedings of the National Academy of Sciences* 2005, **102:**14338-14343.
48. Jovelin R, Phillips PC: **Expression level drives the pattern of selective constraints along the insulin/Tor signal transduction pathway in Caenorhabditis.** *Genome biology and evolution* 2011, **3:**715-722.
49. Song H, Gao H, Liu J, Tian P, Nan Z: **Comprehensive analysis of correlations among codon usage bias, gene expression, and substitution rate in Arachis duranensis and Arachis ipaënsis orthologs.** *Scientific reports* 2017, **7:**1-12.
50. Lloyd JP, Seddon AE, Moghe GD, Simenc MC, Shiu S-H: **Characteristics of plant essential genes allow for within-and between-species prediction of lethal mutant phenotypes.** *The Plant Cell* 2015, **27:**2133-2147.
51. Lorincz MC, Dickerson DR, Schmitt M, Groudine M: **Intragenic DNA methylation alters chromatin structure and elongation efficiency in mammalian cells.** *Nature structural & molecular biology* 2004, **11:**1068-1075.
52. Luco RF, Pan Q, Tominaga K, Blencowe BJ, Pereira-Smith OM, Misteli T: **Regulation of alternative splicing by histone modifications.** *Science* 2010, **327:**996-1000.
53. Zhang J, Liu Y, Xia E-H, Yao Q-Y, Liu X-D, Gao L-Z: **Autotetraploid rice methylome analysis reveals methylation variation of transposable elements and their effects on gene expression.** *Proceedings of the National Academy of Sciences* 2015, **112:**E7022-E7029.
54. Zhai J, Liu J, Liu B, Li P, Meyers BC, Chen X, Cao X: **Small RNA-directed epigenetic natural variation in Arabidopsis thaliana.** *PLoS genetics* 2008, **4:**e1000056.
55. Hirsch CD, Springer NM: **Transposable element influences on gene expression in plants.** *Biochimica et Biophysica Acta (BBA)-Gene Regulatory Mechanisms* 2017, **1860:**157-165.
56. Stroud H, Greenberg MV, Feng S, Bernatavichute YV, Jacobsen SE: **Comprehensive analysis of silencing mutants reveals complex regulation of the Arabidopsis methylome.** *Cell* 2013, **152:**352-364.
57. Weber M, Hellmann I, Stadler MB, Ramos L, Pääbo S, Rebhan M, Schübeler D: **Distribution, silencing potential and evolutionary impact of promoter DNA methylation in the human genome.** *Nature genetics* 2007, **39:**457-466.

58. Lisch D, Bennetzen JL: **Transposable element origins of epigenetic gene regulation.** *Current opinion in plant biology* 2011, **14:**156-161.
59. Swinburne IA, Silver PA: **Intron delays and transcriptional timing during development.** *Developmental cell* 2008, **14:**324-330.
60. Rody HV, Baute GJ, Rieseberg LH, Oliveira LO: **Both mechanism and age of duplications contribute to biased gene retention patterns in plants.** *BMC genomics* 2017, **18:**1-10.
61. Caffrey DR, Somaroo S, Hughes JD, Mintseris J, Huang ES: **Are protein–protein interfaces more conserved in sequence than the rest of the protein surface?** *Protein Science* 2004, **13:**190-202.
62. Jones S, Thornton JM: **Principles of protein-protein interactions.** *Proceedings of the National Academy of Sciences* 1996, **93:**13-20.
63. Montilla-Bascón G, Rubiales D, Hebelstrup KH, Mandon J, Harren FJ, Cristescu SM, Mur LA, Prats E: **Reduced nitric oxide levels during drought stress promote drought tolerance in barley and is associated with elevated polyamine biosynthesis.** *Scientific reports* 2017, **7:**1-15.
64. Zentella R, Mascorro-Gallardo JO, Van Dijck P, Folch-Mallol J, Bonini B, Van Vaeck C, Gaxiola R, Covarrubias AA, Nieto-Sotelo J, Thevelein JM: **A Selaginella lepidophylla trehalose-6-phosphate synthase complements growth and stress-tolerance defects in a yeast tps1 mutant.** *Plant Physiology* 1999, **119:**1473-1482.
65. Zhao D-Q, Li T-T, Hao Z-J, Cheng M-L, Tao J: **Exogenous trehalose confers high temperature stress tolerance to herbaceous peony by enhancing antioxidant systems, activating photosynthesis, and protecting cell structure.** *Cell Stress and Chaperones* 2019, **24:**247-257.
66. Casto AL, McKinley BA, Yu KMJ, Rooney WL, Mullet JE: **Sorghum stem aerenchyma formation is regulated by SbNAC_D during internode development.** *Plant Direct* 2018, **2:**e00085.
67. Hsu C-L, Huang Y-H, Hsu C-T, Yang U-C: **Prioritizing disease candidate genes by a gene interconnectedness-based approach.** In *BMC genomics*. BioMed Central; 2011: 1-12.
68. Edger PP, Pires JC: **Gene and genome duplications: the impact of dosage-sensitivity on the fate of nuclear genes.** *Chromosome Research* 2009, **17:**699-717.
69. Endo M, Mochizuki N, Suzuki T, Nagatani A: **CRYPTOCHROME2 in vascular bundles regulates flowering in Arabidopsis.** *The Plant Cell* 2007, **19:**84-93.
70. Wang Q, Zuo Z, Wang X, Gu L, Yoshizumi T, Yang Z, Yang L, Liu Q, Liu W, Han Y-J: **Photoactivation and inactivation of Arabidopsis cryptochrome 2.** *Science* 2016, **354:**343-347.
71. Zhou Y, Xun Q, Zhang D, Lv M, Ou Y, Li J: **TCP transcription factors associate with PHYTOCHROME INTERACTING FACTOR 4 and CRYPTOCHROME 1 to regulate thermomorphogenesis in Arabidopsis thaliana.** *Iscience* 2019, **15:**600-610.
72. Liang Z, Schnable JC: **Functional divergence between subgenomes and gene pairs after whole genome duplications.** *Molecular plant* 2018, **11:**388-397.
73. Zwaenepoel A, Li Z, Lohaus R, Van de Peer Y: **Finding evidence for whole genome duplications: a reappraisal.** *Molecular plant* 2019, **12:**133-136.
74. Li Z, Defoort J, Tasdighian S, Maere S, Van de Peer Y, De Smet R: **Gene duplicability of core genes is highly consistent across all angiosperms.** *The Plant Cell* 2016, **28:**326-344.
75. Denoeud F, Carretero-Paulet L, Dereeper A, Droc G, Guyot R, Pietrella M, Zheng C, Alberti A, Anthony F, Aprea G: **The coffee genome provides insight into the convergent evolution of caffeine biosynthesis.** *science* 2014, **345:**1181-1184.
76. Wu H-J, Zhang Z, Wang J-Y, Oh D-H, Dassanayake M, Liu B, Huang Q, Sun H-X, Xia R, Wu Y: **Insights into salt tolerance from the genome of Thellungiella salsuginea.** *Proceedings of the National Academy of Sciences* 2012, **109:**12219-12224.
77. Katju V, Bergthorsson U: **Copy-number changes in evolution: rates, fitness effects and adaptive significance.** *Frontiers in genetics* 2013, **4:**273.
78. Hofberger JA, Nsibo DL, Govers F, Bouwmeester K, Schranz ME: **A complex interplay of tandem-and whole-genome duplication drives expansion of the L-type lectin receptor kinase gene family in the brassicaceae.** *Genome biology and evolution* 2015, **7:**720-734.

79. He X-J, Chen T, Zhu J-K: **Regulation and function of DNA methylation in plants and animals.** *Cell research* 2011, **21:**442-465.
80. Park J, Xu K, Park T, Yi SV: **What are the determinants of gene expression levels and breadths in the human genome?** *Human molecular genetics* 2012, **21:**46-56.
81. Zemach A, McDaniel IE, Silva P, Zilberman D: **Genome-wide evolutionary analysis of eukaryotic DNA methylation.** *Science* 2010, **328:**916-919.
82. Bewick AJ, Ji L, Niederhuth CE, Willing E-M, Hofmeister BT, Shi X, Wang L, Lu Z, Rohr NA, Hartwig B: **On the origin and evolutionary consequences of gene body DNA methylation.** *Proceedings of the National Academy of Sciences* 2016, **113:**9111-9116.
83. Su Z, Han L, Zhao Z: **Conservation and divergence of DNA methylation in eukaryotes: new insights from single base-resolution DNA methylomes.** *Epigenetics* 2011, **6:**134-140.
84. Takuno S, Gaut BS: **Body-methylated genes in Arabidopsis thaliana are functionally important and evolve slowly.** *Molecular biology and evolution* 2012, **29:**219-227.
85. Hedges SB, Marin J, Suleski M, Paymer M, Kumar S: **Tree of life reveals clock-like speciation and diversification.** *Molecular biology and evolution* 2015, **32:**835-845.
86. Guo W, Fiziev P, Yan W, Cokus S, Sun X, Zhang MQ, Chen P-Y, Pellegrini M: **BS-Seeker2: a versatile aligning pipeline for bisulfite sequencing data.** *BMC genomics* 2013, **14:**1-8.
87. Neme R, Tautz D: **Phylogenetic patterns of emergence of new genes support a model of frequent de novo evolution.** *BMC genomics* 2013, **14:**1-13.
88. Palmieri N, Kosiol C, Schlötterer C: **The life cycle of Drosophila orphan genes.** *elife* 2014, **3:**e01311.
89. Li L, Foster CM, Gan Q, Nettleton D, James MG, Myers AM, Wurtele ES: **Identification of the novel protein QQS as a component of the starch metabolic network in Arabidopsis leaves.** *The Plant Journal* 2009, **58:**485-498.
90. McLysaght A, Hurst LD: **Open questions in the study of de novo genes: what, how and why.** *Nature Reviews Genetics* 2016, **17:**567-578.
91. Duarte JM, Cui L, Wall PK, Zhang Q, Zhang X, Leebens-Mack J, Ma H, Altman N, DePamphilis CW: **Expression pattern shifts following duplication indicative of subfunctionalization and neofunctionalization in regulatory genes of Arabidopsis.** *Molecular biology and evolution* 2006, **23:**469-478.
92. Lynch M, Force A: **The probability of duplicate gene preservation by subfunctionalization.** *Genetics* 2000, **154:**459-473.
93. Teufel AI, Liu L, Liberles DA: **Models for gene duplication when dosage balance works as a transition state to subsequent neo-or sub-functionalization.** *BMC evolutionary biology* 2016, **16:**1-8.
94. Li F, Fan G, Lu C, Xiao G, Zou C, Kohel RJ, Ma Z, Shang H, Ma X, Wu J: **Genome sequence of cultivated Upland cotton (Gossypium hirsutum TM-1) provides insights into genome evolution.** *Nature biotechnology* 2015, **33:**524-530.
95. Zhao M, Zhang B, Lisch D, Ma J: **Patterns and consequences of subgenome differentiation provide insights into the nature of paleopolyploidy in plants.** *The Plant Cell* 2017, **29:**2974-2994.
96. Garsmeur O, Schnable JC, Almeida A, Jourda C, D'Hont A, Freeling M: **Two evolutionarily distinct classes of paleopolyploidy.** *Molecular biology and evolution* 2014, **31:**448-454.
97. Lieberman-Aiden E, Van Berkum NL, Williams L, Imakaev M, Ragoczy T, Telling A, Amit I, Lajoie BR, Sabo PJ, Dorschner MO: **Comprehensive mapping of long-range interactions reveals folding principles of the human genome.** *science* 2009, **326:**289-293.
98. Ye C, Ma ZS, Cannon CH, Pop M, Douglas WY: **Exploiting sparseness in de novo genome assembly.** In *BMC bioinformatics*. BioMed Central; 2012: 1-8.
99. Ye C, Ma ZS, Cannon CH, Pop M, Yu DW: **SparseAssembler: de novo Assembly with the Sparse de Bruijn Graph.** *arXiv preprint arXiv:11062603* 2011.
100. Ye C, Hill CM, Wu S, Ruan J, Ma ZS: **DBG2OLC: efficient assembly of large genomes using long erroneous reads of the third generation sequencing technologies.** *Scientific reports* 2016, **6:**1-9.
101. Li H, Durbin R: **Fast and accurate short read alignment with Burrows–Wheeler transform.** *bioinformatics* 2009, **25:**1754-1760.

102. Walker BJ, Abeel T, Shea T, Priest M, Abouelliel A, Sakthikumar S, Cuomo CA, Zeng Q, Wortman J, Young SK: **Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement.** *PloS one* 2014, **9:**e112963.
103. Burton JN, Adey A, Patwardhan RP, Qiu R, Kitzman JO, Shendure J: **Chromosome-scale scaffolding of de novo genome assemblies based on chromatin interactions.** *Nature biotechnology* 2013, **31:**1119-1125.
104. English AC, Richards S, Han Y, Wang M, Vee V, Qu J, Qin X, Muzny DM, Reid JG, Worley KC: **Mind the gap: upgrading genomes with Pacific Biosciences RS long-read sequencing technology.** *PloS one* 2012, **7:**e47768.
105. Campbell MS, Law M, Holt C, Stein JC, Moghe GD, Hufnagel DE, Lei J, Achawanantakun R, Jiao D, Lawrence CJ: **MAKER-P: a tool kit for the rapid creation, management, and quality control of plant genome annotations.** *Plant physiology* 2014, **164:**513-524.
106. Han Y, Wessler SR: **MITE-Hunter: a program for discovering miniature inverted-repeat transposable elements from genomic sequences.** *Nucleic acids research* 2010, **38:**e199-e199.
107. Ellinghaus D, Kurtz S, Willhoeft U: **LTRharvest, an efficient and flexible software for de novo detection of LTR retrotransposons.** *BMC bioinformatics* 2008, **9:**1-14.
108. Steinbiss S, Willhoeft U, Gremme G, Kurtz S: **Fine-grained annotation and classification of de novo predicted LTR retrotransposons.** *Nucleic acids research* 2009, **37:**7002-7013.
109. Van Bel M, Diels T, Vancaester E, Kreft L, Botzki A, Van de Peer Y, Coppens F, Vandepoele K: **PLAZA 4.0: an integrative resource for functional, evolutionary and comparative plant genomics.** *Nucleic acids research* 2018, **46:**D1190-D1196.
110. Kim D, Langmead B, Salzberg SL: **HISAT: a fast spliced aligner with low memory requirements.** *Nature methods* 2015, **12:**357-360.
111. Pertea M, Pertea GM, Antonescu CM, Chang T-C, Mendell JT, Salzberg SL: **StringTie enables improved reconstruction of a transcriptome from RNA-seq reads.** *Nature biotechnology* 2015, **33:**290-295.
112. Niknafs YS, Pandian B, Iyer HK, Chinnaiyan AM, Iyer MK: **TACO produces robust multisample transcriptome assemblies from RNA-seq.** *Nature methods* 2017, **14:**68-70.
113. Keilwagen J, Wenk M, Erickson JL, Schattat MH, Grau J, Hartung F: **Using intron position conservation for homology-based gene prediction.** *Nucleic acids research* 2016, **44:**e89-e89.
114. Nock CJ, Baten A, Barkla BJ, Furtado A, Henry RJ, King GJ: **Genome and transcriptome sequencing characterises the gene space of Macadamia integrifolia (Proteaceae).** *BMC genomics* 2016, **17:**1-12.
115. Hoff KJ, Lange S, Lomsadze A, Borodovsky M, Stanke M: **BRAKER1: unsupervised RNA-Seq-based genome annotation with GeneMark-ET and AUGUSTUS.** *Bioinformatics* 2016, **32:**767-769.
116. Haas BJ, Salzberg SL, Zhu W, Pertea M, Allen JE, Orvis J, White O, Buell CR, Wortman JR: **Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments.** *Genome biology* 2008, **9:**1-22.
117. Conesa A, Götz S, García-Gómez JM, Terol J, Talón M, Robles M: **Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research.** *Bioinformatics* 2005, **21:**3674-3676.
118. Yang J, Osman K, Iqbal M, Stekel DJ, Luo Z, Armstrong SJ, Franklin FCH: **Inferring the Brassica rapa interactome using protein–protein interaction data from Arabidopsis thaliana.** *Frontiers in plant science* 2013, **3:**297.
119. Langmead B: **Aligning short sequencing reads with Bowtie.** *Current protocols in bioinformatics* 2010, **32:**11.17. 11-11.17. 14.
120. Wang Y, Tang H, DeBarry JD, Tan X, Li J, Wang X, Lee T-h, Jin H, Marler B, Guo H: **MCScanX: a toolkit for detection and evolutionary analysis of gene synteny and collinearity.** *Nucleic acids research* 2012, **40:**e49-e49.
121. Lee T-H, Tang H, Wang X, Paterson AH: **PGDD: a database of gene and genome duplication in plants.** *Nucleic acids research* 2012, **41:**D1152-D1158.
122. Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM: **BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs.** *Bioinformatics* 2015, **31:**3210-3212.

123. Liebrand TW, van den Burg HA, Joosten MH: **Two for all: receptor-associated kinases SOBIR1 and BAK1.** *Trends in plant science* 2014, **19:**123-132.
124. Matus JT, Aquea F, Arce-Johnson P: **Analysis of the grape MYB R2R3 subfamily reveals expanded wine quality-related clades and conserved gene structure organization across Vitis and Arabidopsis genomes.** *BMC plant biology* 2008, **8:**1-15.
125. Tang H, Wang X, Bowers JE, Ming R, Alam M, Paterson AH: **Unraveling ancient hexaploidy through multiply-aligned angiosperm gene maps.** *Genome research* 2008, **18:**1944-1954.
126. Yang Z: **PAML 4: phylogenetic analysis by maximum likelihood.** *Molecular biology and evolution* 2007, **24:**1586-1591.
127. Li L, Stoeckert CJ, Roos DS: **OrthoMCL: identification of ortholog groups for eukaryotic genomes.** *Genome research* 2003, **13:**2178-2189.
128. Neale DB, Wegrzyn JL, Stevens KA, Zimin AV, Puiu D, Crepeau MW, Cardeno C, Koriabine M, Holtz-Morris AE, Liechty JD: **Decoding the massive genome of loblolly pine using haploid DNA and novel assembly strategies.** *Genome biology* 2014, **15:**1-13.
129. Arendsee ZW, Li L, Wurtele ES: **Coming of age: orphan genes in plants.** *Trends in plant science* 2014, **19:**698-708.
130. Huang L, Yang M, Li L, Li H, Yang D, Shi T, Yang P: **Whole genome re-sequencing reveals evolutionary patterns of sacred lotus (Nelumbo nucifera).** *Journal of integrative plant biology* 2018, **60:**2-15.
131. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M: **The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data.** *Genome research* 2010, **20:**1297-1303.
132. Pfeifer B, Wittelsbürger U, Ramos-Onsins SE, Lercher MJ: **PopGenome: an efficient Swiss army knife for population genomic analyses in R.** *Molecular biology and evolution* 2014, **31:**1929-1936.
133. Yanai I, Benjamin H, Shmoish M, Chalifa-Caspi V, Shklar M, Ophir R, Bar-Even A, Horn-Saban S, Safran M, Domany E: **Genome-wide midrange transcription profiles reveal expression level relationships in human tissue specification.** *Bioinformatics* 2005, **21:**650-659.
134. Obayashi T, Kinoshita K: **Rank of correlation coefficient as a comparable measure for biological significance of gene coexpression.** *DNA research* 2009, **16:**249-260.
135. Wang H, Beyene G, Zhai J, Feng S, Fahlgren N, Taylor NJ, Bart R, Carrington JC, Jacobsen SE, Ausin I: **CG gene body DNA methylation changes and evolution of duplicated genes in cassava.** *Proceedings of the National Academy of Sciences* 2015, **112:**13729-13734.
136. Guo W, Zhu P, Pellegrini M, Zhang MQ, Wang X, Ni Z: **CGmapTools improves the precision of heterozygous SNV calls and supports allele-specific methylation detection and visualization in bisulfite-sequencing data.** *Bioinformatics* 2018, **34:**381-387.
137. Guo YL: **Gene family evolution in green plants with emphasis on the origination and evolution of a rabidopsis thaliana genes.** *The Plant Journal* 2013, **73:**941-951.

3 Chapter 3

Genome-wide expression and network analyses of mutants in key brassinosteroid signaling genes

"Biology is the study of the complex things in the Universe. Physics is the study of the simple ones"
-Richard Dawkins

The brassinosteroids (BRs) are a class of plant hormones regulating plant growth and development. Mutants are powerful in functional genomics to study gene function. Also the function of genes involved in BR biosynthesis and signaling has been established by developing mutant lines. Mutants that are defective in either BR biosynthesis or signaling show dwarfism. However, assessing the function of every gene through mutations is not feasible because of the existence of gene homologs which create functional redundancy and because of the technical difficulty of generating mutant lines. Better exploiting already existing data can help inferring the function of genes which have not yet been explored. In the present chapter, we present a computational approach to overlay the in-house data with already existing knowledge in order to understand the components of the BR signaling. We show how network-based integration of expression data with prior information can offer a comprehensive overview of the crosstalk between BR and other hormone signaling. The candidate analyzed the phenotypic and expression data, performed the network analyses, and analyzed the results from a biological perspective. The candidate summarized the results of the study in a coherent story. The candidate was not involved in the experimental design or in the wet-lab data generation. Page 3-23 includes an overview of the contributions of all authors.

Genome-wide expression and network analyses of mutants in key brassinosteroid signaling genes [1]

Razgar Seyed Rahmani, Tao Shi, Dongzhi Zhang, Xiaoping Gou, Jing Yi, Giles Miclotte, Kathleen Marchal*, Jia Li*

Published in *BMC Genomics*, 22.1, June 2021

Abstract

Background

Brassinosteroid (BR) signaling regulates plant growth and development in concert with other signaling pathways. Although many genes have been identified that play a role in BR signaling, the biological and functional consequences of disrupting those key BR genes still require detailed investigation. Here we performed phenotypic and transcriptomic comparisons of *A. thaliana* lines carrying a loss-of-function mutation in *BRI1* gene, *bri1*-5, that exhibits a dwarf phenotype and its three activation-tag suppressor lines that were able to partially revert the *bri1-5* mutant phenotype to a WS2 phenotype, namely *bri1-5/bri1-1D*, *bri1-5/brs1-1D,* and *bri1-5/bak1-1D*. Of the three investigated *bri1-5* suppressors, *bri1-5/bak1-1D* was the most effective suppressor at the transcriptional level. All three *bri1-5* suppressors showed altered expression of the genes in the abscisic acid (ABA signaling) pathway, indicating that ABA likely contributes to the partial recovery of the wild-type phenotype in these *bri1-5* suppressors. Network analysis revealed crosstalk between BR and other phytohormone signaling pathways, suggesting that interference with one hormone signaling pathway affects other hormone signaling pathways. In addition, differential expression analysis suggested the existence of a strong negative feedback from BR signaling on BR biosynthesis and also predicted that *BRS1*, rather than being directly involved in signaling, might be responsible for providing an optimal environment for the interaction between BRI1 and its ligand. Our study provides insights into the molecular mechanisms and functions of key brassinosteroid (BR) signaling genes, especially *BRS1*.

3.1 Background

Brassinosteroids (BRs) are essential plant hormones, regulating multiple processes amongst which plant growth, flowering, senescence, and seed germination [2]. BR biosynthetic and signaling mutants display aberrant morphological phenotypes such as dwarfism, reduced fertility, impaired photomorphogenesis, and altered vascular development [3, 4]. Whereas the phenotypes of mutants in BR biosynthetic genes can be rescued by the application of exogenous BRs, this is not the case for strains carrying mutations in genes responsible for BR signal

perception and transduction. Hence, these latter strains are referred to as BR insensitive (*bri*) mutants [2, 4]. In the BR signaling pathway (Figure 3.1), BRs are perceived by membrane-localized leucine-rich-repeat-receptor kinase BRI1 or by the BRI1-like homologs, BRL1 and BRL3 [5, 6]. After binding to BRs, BRI1 and its co-receptor BRI1-Associated Receptor Kinase 1 (**BAK1**) phosphorylate each other. This results in triggering a cytoplasmic phosphorylation/dephosphorylation signaling cascade which deactivates the GSK3-like kinase BRASSINOSTEROID INSENSITIVE 2 (**BIN2**) through dephosphorylating [7, 8]. Upon BIN2 deactivation, the downstream transcription factors, BRASSINAZOLE-RESISTANT1 **(BZR**1) and BR-INSENSITIVE-EMS-SUPPRESSOR1 (**BES1**) are dephosphorylated by PP2A (PHOSPHATASE 2A). This results in their disassociation from 14-3-3 proteins, causing them to get activated and regulating a range of downstream genes involved in various aspects of plant growth and development [9-11]. In the absence of BRs, BIN2 is active (phosphorylated) and it prevents the activation of BZR1 and BES1. Because BRI1 is the core receptor of BRs, mutants of *BRI1* have been used as genetic background to identify suppressors, i.e., other genes that when mutated, suppress the *bri1* phenotype and thus may play a role in BR signaling. For example, the function of *BZR1* has been unveiled by using the null allele of *BRI,* ***bri1-116*** [12]. The weak mutant of *bri1***,** ***bri1-5,*** can be rescued by overexpression of ***BAK1*** and *BRI* Suppressor 1 **(*BRS1*)** [4, 13]. BRS1 is a secreted member of the serine carboxypeptidase (SCP) family [4]. The fact that overexpression of *BRS1* can suppress two weak *BRI1* extracellular domain mutants, *bri1-5* and *bri1-9*, but not the strong cytoplasmic domain mutant *bri1-1*, implies that *BRS1*, unlike the downstream genes, *BZR1* or *BES1*, may function upstream of the BR signaling pathway or in a close regulatory relationship with *BRI1* [4]. Moreover, three of the five overexpressed *BRS1*'s homologs amongst which *ECS1* (Extra Carpels and Seeds 1) can also partially suppress the phenotype of the *bri1-5* mutant observed in leaves. Overexpression of *BRS1*'s homologs also increase the number of carpels and seeds, confirming the role of *BRS1* and its homologs in the BR signaling [14]. Yet, the detailed mechanism of how *BRS1* potentially interacts with other BR genes in order to maintain balance in BR signaling is still unknown.

Some genes involved in BR signaling are also involved in other processes, such as stress response, and can act independently of the presence of BRs. Several studies found that *bes1-1D* and *bzr1-1D* backgrounds are not responsive to exogenous BRs, suggesting that *BES1* and *BZRI* have also other functions than BR signaling [15, 16]. In another study, BAK1 was found to work together with Flagellin-Sensitive 2 (FLS2) during pathogen defense programmed cell death independently of BR signaling [15, 17-19]. In addition, SERK1 and SERK2, the homologs of BAK1 play a role in male microsporogenesis, also independently of BR signaling [20]. Some *bri1* mutants show in addition to reduced growth, an increased stress-tolerance, further confirming the complexity and dosage sensitivity of BR signaling and regulation [21, 22]. Transcriptomic studies and network analysis have been shown to be effective in uncovering the expression and biological consequences of gene mutants, and have successfully been applied to study several BR genes such as *BRI1* and *BES1* [23]. Therefore, in the present study,

we applied a similar strategy to elucidate the role of *BRI1*, *BAK1*, and *BRS1* in regulating/restoring the response to BRs and/or in other functions independent of BR signaling.

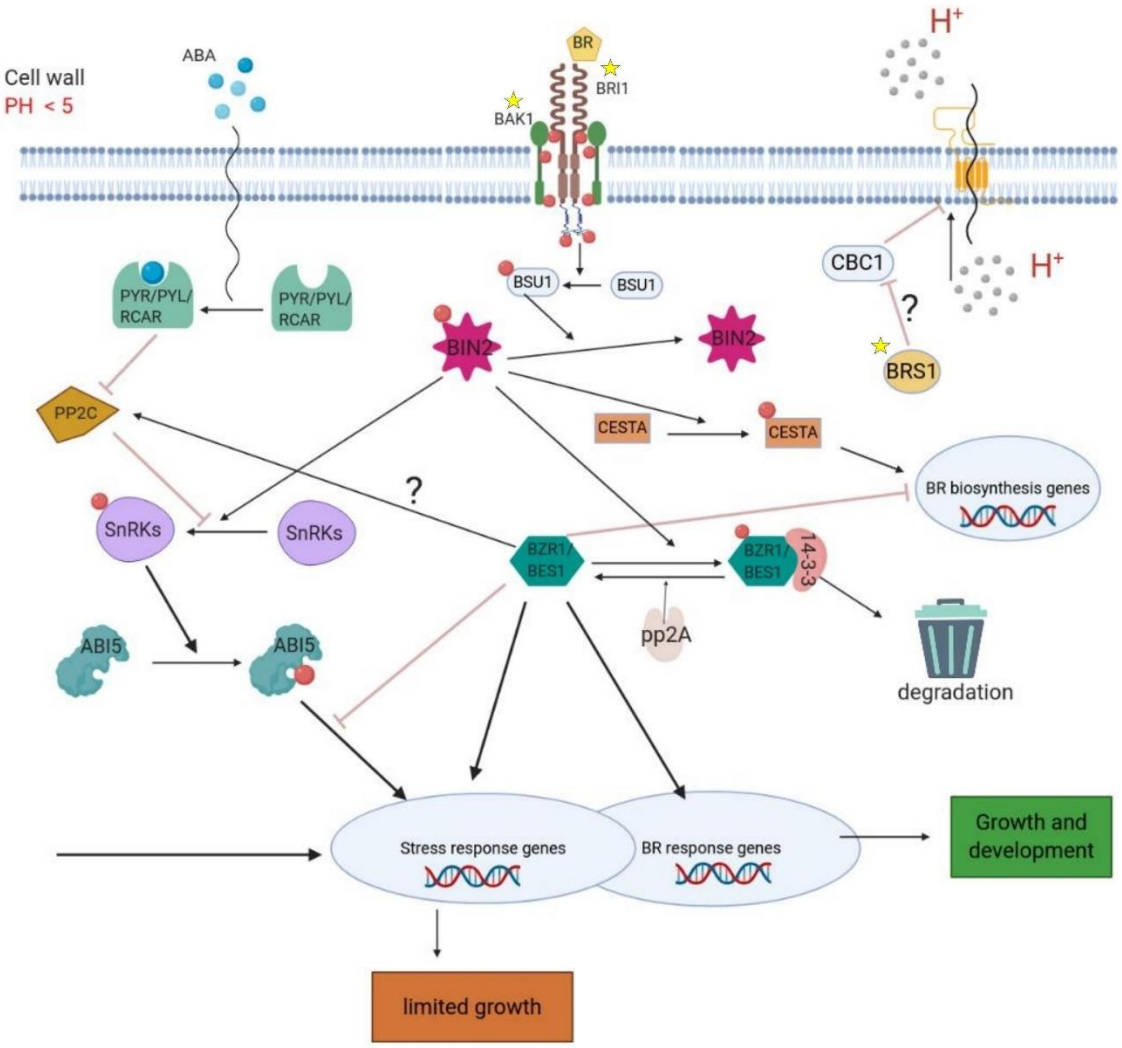


Figure 3.1 Schematic overview of the BR signaling cascade and the results of this study. The figure provides a simplified scheme of BR signaling based on [6, 7]. The genes studied in this work are indicated by a yellow star. Binding of BRs to the BRI1/BAK1 receptor triggers the phosphorylation/dephosphorylation signaling cascade that leads to the deactivation (dephosphorylation) of BIN2. The effects of BIN2 and BZR1/BES1 on BR-biosynthesis genes are depicted. The overlap between stress-response and BR response genes and the dual effect of BZR1/BES1 on stress response genes is also shown. The question marks indicate missing links that have been suggested based on the result of the present study. The hypothetical inferred role for BRS1 in providing a better condition for BRI/BAK1/BR binding by generating a more acidic environment is shown on the top right-hand side. The compensatory pathway resulting in the over-expressing expression of PP2C mediated by ABA is shown on the left-hand side. (Created with BioRender.com)

3.2 Results

3.2.1 *bri1-5/bak1-1D*, *bri1-5/brs1-1D* and *bri1-5/bri1-1D* partially reconstitute *bri1-5* gene expression

To better understand the molecular mechanisms of key BR signaling genes, we performed a phenotypic screening and expression analysis of *bri1-5* and its three activation-tag suppressors along with their corresponding wild-type, WS2. Two suppressor strains *bri1-5/bak1-1D* and *bri1-5/brs-1D* were obtained from [13]. An additional *bri1-5/bri1-1D* mutant was generated in the framework of the current study (see Methods). Sequencing the *BRI1* flanking region from the suppressor *bri1-5/bri1-1D* showed that the activation tag was inserted 534 bp downstream of the *BRI1* gene (Supplementary Figure 3.1-A). All suppressor mutants were shown to indeed overexpress the activation tagged gene as confirmed by Real-Time qPCR (RT-qPCR) (Table S3.1). Phenotypically, all *bri1-5* suppressors (*bri1-5*/*bak1-1D, bri1-5*/*brs1-1D,* and *bri1-5*/*bri1-1D)* lines displayed larger seedlings than the *bri1-5* mutant, but still significantly smaller than the WS2 (Figure 3.2). Of all suppressor mutants, the *bri1-5/bak1-1D* line best approximated the growth phenotype of the WS2, and its larger seedling seemed to be mainly the effect of its larger root length and to a lower extent of its larger hypocotyl length (both of which were significantly larger than the *bri1-5* mutant). The contribution of the epidermal cell length in recovering the *bri1-5* is marginal in the *bri1-5*/*bak1-1D* (line with the largest seeding) but seems much more pronounced in the *bri1-5/brs1-1D* (Figure 3.2, Supplementary Figure 1: B-F). This indicates that in the *bri1-5/brs1-1D* mechanisms other than those in the *bri1-5*/*bak1-1D* line play a role in alleviating the *bri1-5* phenotype.

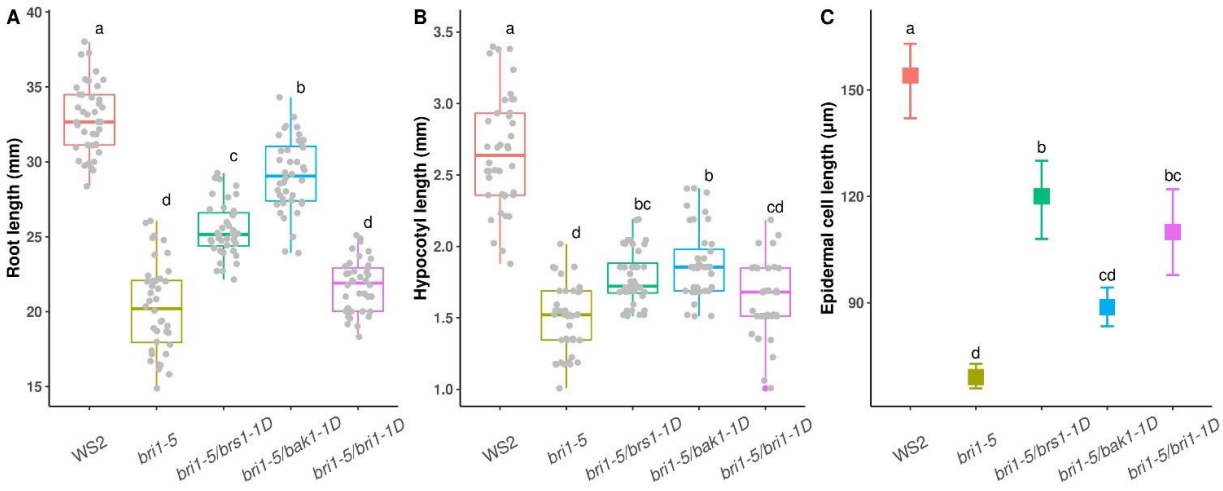


Figure 3.2 Root, hypocotyl, and epidermal cell length at seedling stage of plants used for expression profiling. Root (A), hypocotyl (B), and epidermal cell length (C) of WS2, *bri1-5*, *bri1-5/brs1-1D*, *bri1-5/bak1-1D,* and *bri1-5/bri1-1D*, measured 7 days after germination. For the root and hypocotyl length, the boxplot shows the distribution of data for 40 plants. For cell length, the bar represents the 95% confidence interval for the mean and the square indicates the location of the mean. Groups (different

plant lines) were statistically compared by ANOVA and Tukey tests. Groups are ranked based on their significance level where "a" represents the group with the highest mean and "d" the group with the lowest mean. Groups with different letters are significantly different.

To gain insight into which pathways in each of the studied lines were responsible for recovering the *bri1-5* growth phenotype to wild-type level, we performed gene expression analysis. All suppressor lines, together with the wild-type (WS2) and the *bri1-5* background were sampled at a 7-day seedling stage (whole seedling including hypocotyl and root). To assess the reproducibility of the expression analysis, we measured the extent to which the expression profiles of replicate samples were similar using Principal Component Analysis (PCA): PCA indeed showed that the largest fraction of the variation in gene expression between the samples could be assigned to differences in genetic background and not to differences between replicates of the same genetic background, confirming the reproducibility (Figure S3.2). In addition, microarray results were confirmed using RT-qPCR for a randomly selected set of differentially expressed genes (Figure S3.3).

We determined for each mutant line its differential expression versus the same common reference i.e., the expression state in WS2, resulting in a total of 1413 differentially expressed genes (Additional file1). The Venn diagram represented in Figure 3.3 shows to what extent the different lines share the same differentially expressed genes (aberrantly expressed versus the WS2 control). Figure 3.3 and the scatter-plots in Figure S3.4 (A-C) show that of all suppressor lines, *bri1-5/bak1-1D* could restore the largest number of genes that were affected in expression in *bri1-5* (about two-thirds of the genes that were differentially expressed in *bri1-5* were no longer differentially expressed in *bri1-5/bak1-1D*). This is in line with its observed phenotypic behavior as indeed *bri1-5/bak1-1D* seems to also phenotypically best compensate for the *bri1-5* mutation.

The Venn diagram in Figure 3.3 also shows that the *bri1-5/brs1-1D* and *bri1-5/bri1-1D* lines share the largest fraction of similarly affected genes. The latter is also illustrated in Figure S3.4 panel D-F which shows that from all pairwise comparisons between suppressor lines, the level of differential expression relative to the mutant *bri1-5* is most correlated between the suppressor lines *bri1-5/brs1-1D* and *bri1-5/bri1-1D* (i.e., $R^2 = 0.20$). This suggests a similar role for *BRI1* and *BRS1* in the BR signaling pathway. Note that in Figure S3.4 D-F, rather than performing a direct correlation analysis of the expression between two mutant lines, we performed correlation analysis with the expression of each mutant line relative to the same reference (expression in *bri1-5*). In this way, the correlation analysis is driven by the expression of the genes that change their expression relative to *bri1-5*. Although this results in lower correlation values than when directly comparing the expression values of the mutant lines, it better reflects the consistency between mutant lines in restoring genes affected in the *bri1-5* mutant.

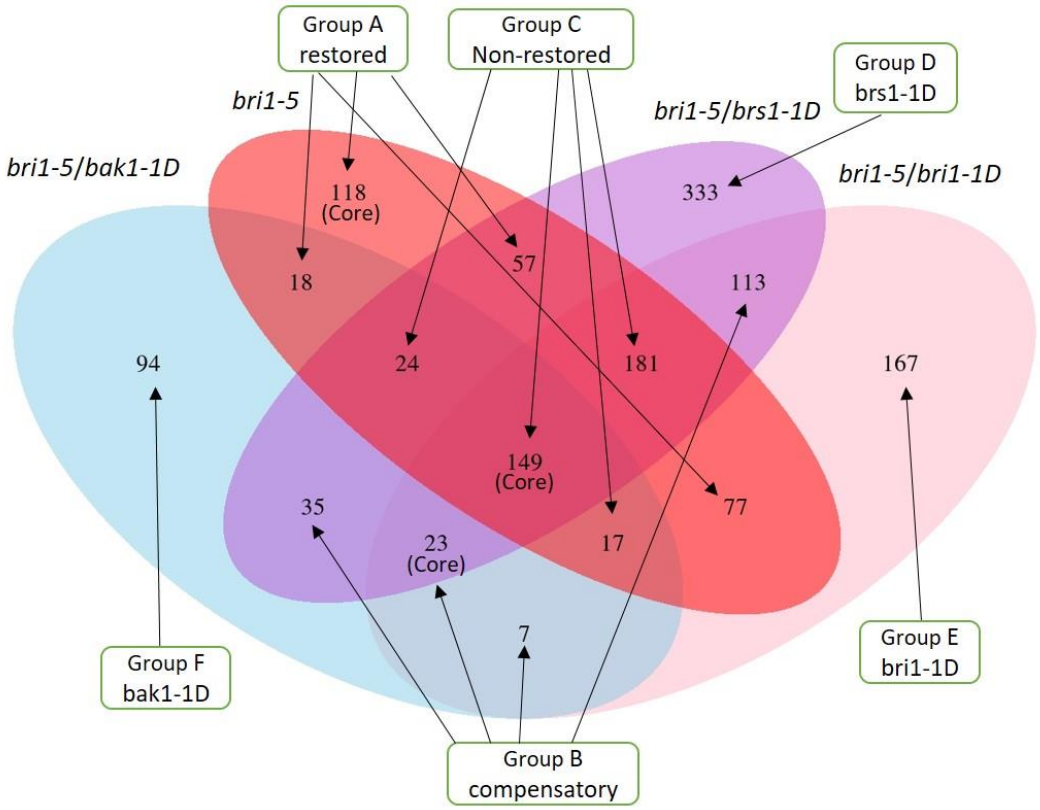


Figure 3.3 Differentially expressed genes (DEGs relative to WS2) being compared between *bri1-5* and its three suppressors. Group A (restored genes, 270 genes): genes differentially expressed in the *bri1-5* mutant but no longer in at least two of the suppressor lines; Group B (compensatory genes, 178 genes): genes that are differentially expressed in at least two suppressors but not in *bri1-5*; Group C (genes that were not restored, 371 genes): Genes that are aberrantly expressed in *bri1-5* and at least two of the suppressor lines. Group D (333 genes), E (167), and F (94 genes) contain genes that are exclusively differentially expressed in respectively the *bri1-5/brs1-1D*, *bri1-5/bri1-1D,* and *bri1-5/bak1-1D* suppressor lines. The "core" below the number indicates the most reliable set for the group. The total number of potentially interesting genes is 1430.

To confirm the extent to which the different suppressor strains molecularly restore the defects in the *bri1-5* mutation, we compiled a list of marker genes representative of downstream pathways affected by BR signaling (Additional file 2). This consisted of 233 marker genes that were according to literature regulated by BR signaling (genes that became up or down-regulated upon treatment with exogenous BRs or by overexpressing the BR signaling genes). Of those marker genes, only those that were significantly affected in the *bri1*-5 line were retained in order to identify the mutant line that best suppresses the *bri1-5* mutation (96 marker genes). Figure 3.4 and Figure S3.5 show how the expression of these genes is, as compared to WS2 affected in the *bri1-5* mutant and how some of those genes got restored in the suppressor mutants. These results confirm what we observed based on the global expression analysis, i.e., that the *bri1-5/bak1-1D* restored the *bri1-5* affected marker genes to the largest extent, and that

molecularly the *bri1-5/brs1-1D* and *bri1-5/bri1-1D* mutant tend to behave more similarly in restoring the same marker genes.

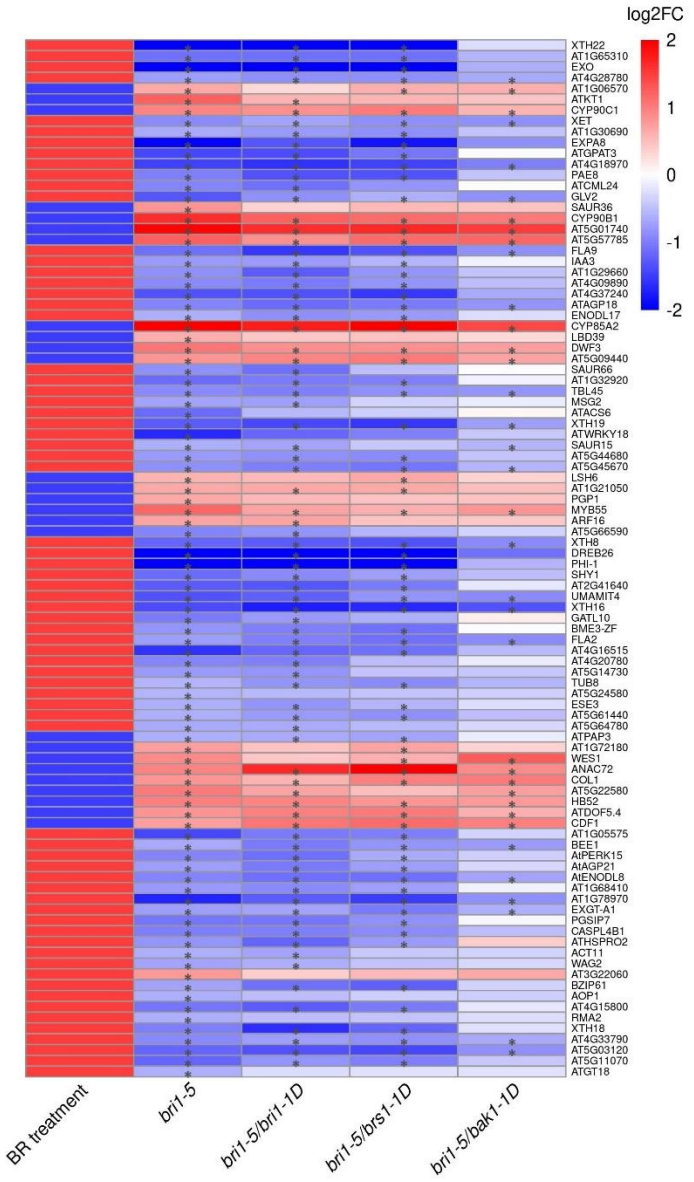


Figure 3.4 Expression behavior of marker genes representative of downstream BR signaling pathways. Column BR treatment: colors indicate whether a gene was reported to be up (red) or down (blue) regulated according to literature upon treatment with exogenous BRs or in a line containing a gain-of-function mutation in a BR signaling gene. Genes were only selected as representative for downstream BR signaling if the up/down regulation of their expression was confirmed by at least 5 independent references and also affected in the *bri1-5* line of our study (compared to WS2). Columns *bri1-5*, *bri1-5/bri1-1D*, *bri1-5/brs1-1D*, *bri1-5/bak1-1* indicate whether the genes were found to be up or down-regulated compared to WS2 according to our expression data. Color scale indicates whether a gene is up-regulated (red), down-regulated (blue), or not differentially expressed (white). A * indicates that the adjusted p-value < 0.05.

3.2.2 Identifying compensatory and restoring pathways

Pathway analysis (see Methods) unveiled the pathways overrepresented amongst the differentially expressed gene sets in each of the mutant lines. Figure S3.6-S8 and Table S3.2 show a number of pathways that are differentially expressed in both *bri1-5* and all of the suppressor lines. These represent the pathways that are responsible for the aberrant growth phenotype in the *bri1-5* mutant and that could not entirely be restored or compensated for in the suppressor lines. Among others, pathways related to cell wall synthesis (cell wall cellulose synthesis), protein and lipid metabolism can explain the residual discrepancy between the WS2 growth phenotype and the suppressors.

We assumed that if the suppressor strains alleviate the phenotype of the *bri1-5* mutant, they could do so because they either restore the pathways disrupted in the *bri1-5* mutant to wild-type levels or they induce genes that compensate for the *bri1-5* affected pathways. Both mechanisms are reflected in the expression data. Processes that are aberrantly expressed in the *bri1-5* mutant, but not in any of the suppressor lines, represent pathways that are restored to WS2 levels in all suppressor lines. This seems to be the case for some genes related to cytochrome P450 oxidase (Table S3.2). The fact that they are restored (or not significantly affected) in any of the suppressors indicates they might be essential for the recovery of the WS2 phenotype. Interestingly, the genes related to "glutathione S transferases" (Figure S3.7, Table S3.2) are largely down-regulated in the *bri1-5* mutant, restored to normal in *bri1-5/bri1-1D* and *bri1-5/bak1-1D* and up-regulated as compared to WS2 levels in the *bri1-5/brs1-1D* suppressor, indicating that some overcompensation is needed for this pathway in the *bri1-5/brs1-1D* background in order to restore the *bri1-5* phenotype. In addition, ABA-related metabolism (Figure S3.8 and Table S3.2) seems to have been affected by all suppressors, but at least not to a significant level in the *bri1-5* mutant. Therefore, ABA signaling seems to represent a compensatory pathway, i.e., a pathway that needs to be triggered in the suppressor strains in order to restore the *bri1-5* affected pathways and phenotype.

As less than 5000 genes can be mapped using pathway analysis, we performed a more elaborate analysis using a network-based approach. Network analysis provides an intuitive way of combining expression data with prior information on known molecular interactions or already available functional data [24, 25]. This approach first maps candidate genes, that are identified through expression analysis, on an integrated molecular interaction network. Then, it identifies subnetworks that connect as many candidate genes as possible [25]. By leveraging candidate genes identified through expression analysis with known interaction information, spuriously identified candidate genes can be removed as they will not be part of the subnetworks. In addition, genes relevant to the process of interest that are themselves not regulated at the level of expression are indirectly identified by being part of a connected component/subnetwork to which also many of the candidate genes belong. Such an integrated analysis provides a more comprehensive view of the process of interest. Here, we applied such an integrated network-based strategy to gain a more in-depth insight into the molecular mechanisms through which the suppressor lines can restore the *bri1-5* phenotype to WS2 levels. To perform this network analysis, we started from the gene sets depicted in Figure 3.3.

3.2.3 Involvement of hormone signaling in alleviating the *bri1-5* phenotype

To study the interaction between non-restored, restored, and compensatory pathways in more depth, we combined the following gene sets for network analysis (see Figure 3.3): i) genes that were most likely restored in the suppressors (genes of group A i.e., the genes with altered expression in the *bri1-5* mutant, but restored to WS2 level in at least 2 suppressors), ii) genes that were compensatory in most of the suppressors (genes of group B i.e., the genes not differentially expressed in the *bri1-5* mutant, but differentially expressed in at least two suppressor strains) and iii) genes altered in the *bri1-5* mutant that most likely were not restored in the suppressors (genes of group C i.e., the genes, differentially expressed in the *bri1-5* mutant and at least two of the suppressor strains). This combined set of genes (789 genes) is referred to as the set of seed genes or the genes we want to maximally connect on the interaction network.
Network analysis (see Material and Methods) identified 8 sub-networks (Figure 3.5) containing the set of seed genes that could be connected through the interaction network. These subnetworks contain not only seed genes, but also connector genes. These are genes that are not differentially expressed themselves, but that are still recovered by the network analysis, because of their high connectivity with seed genes. As they are needed to connect seed genes in the network, they are most likely involved in the same processes as the seed genes. The subnetworks were annotated based on their enrichment in known GO functions (being enriched in respectively negative regulation of ABA, response to auxin, fatty acid metabolism process, developmental process, oligopeptide transport, response to ROS, BR homeostasis, and ethylene activated signaling (Figure 3.5)). This indicates that these are the pathways that contribute to alleviating *bri1-5* signaling deficiency in the suppressor strains.

In-depth analysis shows that **the subnetwork enriched in ABA signaling** (Figure 3.5, subnetwork 1) contains several known negative regulators of ABA signaling: *HAI1*, *HAB1 ABI1*, *ABI2*, and *PP2CA* acted as compensatory genes: these were are up-regulated in at least 2 *bri1-5* suppressor lines compared to wild-type, but were not affected in the *bri1-5* mutant (*HAI1* and *HAB1* being significantly up-regulated in all suppressor mutants; *ABI1*, *ABI2*, *PP2CA*, being significantly up-regulated in two suppressors, see Figure S3.9); In addition, *HAI2* was affected in the *bri1-5* line, but could not be restored in at least two suppressors (non-restored gene), and *HAB2* was identified as a connector node.

Interestingly, several targets of the ABA signaling pathway (*DTX50, HVA22D, PUB19, COR15B, next to HAI1, HAB1)* were identified as differentially expressed in all three suppressors (identified based on a GO enrichment of the core of group B, but not in *bri1-5*). This indicates that ABA signaling has indeed been affected in the suppressor strains to compensate for the *bri1-5* signaling deficiency. Of these, *DTX50, HVA22D, PUB19, COR15B* could not be connected by PheNetic on the interaction network, implying they are either not annotated in the interaction network (*COR15B*) or quite distantly located from each other in the network.

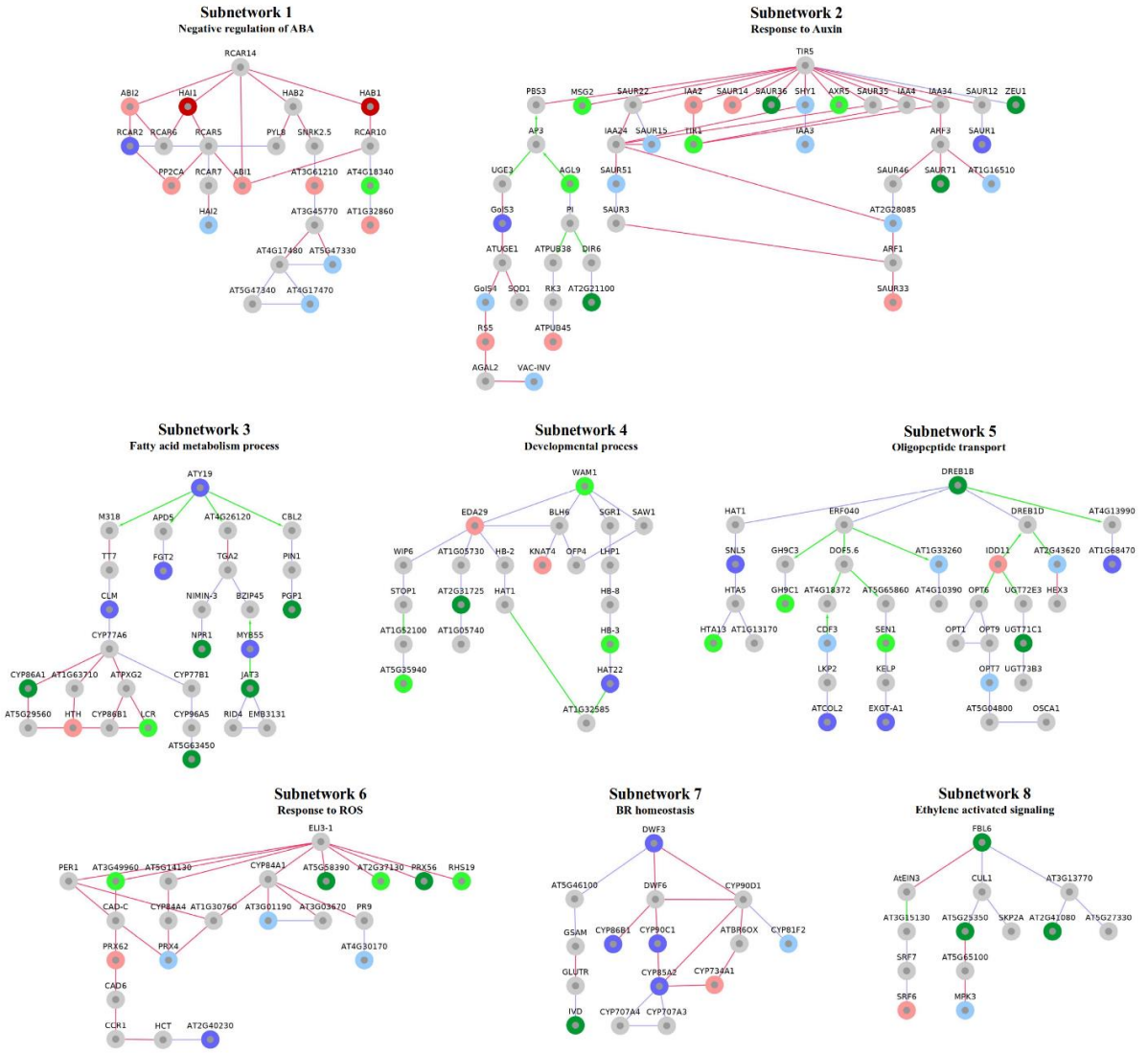


Figure 3.5 Subnetworks resulting from network analysis. Subnetworks identified by PheNetic representing different pathways that were identified by mapping and connecting the genes of group A (restored genes), B (compensatory genes), and C (not restored genes) on the interaction network. Node color: dark green, dark red, and dark blue indicate the core genes of groups A, B, and C, respectively. Likewise, the light green, light red, and light blue correspond to the non-core genes in groups A, B and C, respectively. Connector genes that were not identified as differentially expressed, but identified by PheNetic on the paths that connect the differentially expressed genes are shown in gray. Edge color: regulatory edges are shown in green, metabolic edges in red and protein-protein edges are shown in blue color.

The aforementioned negative regulators of ABA signaling in subnetwork 1 belong to the protein phosphatase 2C (PP2C) gene family which has nine members in total (*HAI2*, *HAB2*, *HAB1*, *HAI3*, *PP2CA*, *ABI1*, *AHG1*, *ABI2*, *AHI1*). PP2C is known to indirectly repress ABI5, the main activator of ABA signaling [26]. PP2C is also known to repress BIN2 activity [7, 8]: as BIN2 activates ABI5 by phosphorylating SnRKs [7, 8], repressing ABA signaling by PP2C via

blocking SnRKs phosphorylation seems to compensate for the deficiency in BRI1 mediated signaling (Figure 3.1). The subnetwork enriched in ABA signaling (network 1) also contains members of the *PYR/PYL/RCAR* family as connector genes (*RCAR5*, *RCAR6*, *RCAR7*, *RCAR10*, *RCAR14*, *PYL8*). The PYR/PYL/RCAR family constitutes the receptor of ABA signaling and promotes the activation of SnRKs by repressing PP2C [16, 27]. The fact that the SnRKs (SnRK2.5) and PYR/PYL/RCAR genes were identified as connector genes implies that they are likely involved in the pathways that connect the affected, restored, and compensatory genes of subnetwork 1. They are most likely not primarily regulated at the expression level, given their role in phosphorylation-mediated signaling [8, 28]. This explains why they were detected as connector genes and not retrieved by differential expression analysis.
We could not find any link in the literature to explain how *PP2C* can be up-regulated by BR signaling in order to repress ABA signaling. It seems that there exist some missing links between *BZR1/BES1* (or downstream TFs) and the *PP2C* gene family. By partially recovering BR signaling in the suppressor lines, one would expect that *PP2C* gene expression levels restore to WS2 expression level. However, they appear to become up-regulated in suppressors, indicating that further compensatory repression of the ABA signaling is required in order to restore the *bri1-5* phenotype.

Other than the ABA subnetwork, subnetworks related to other signaling processes like auxin signaling (subnetwork 2), ROS signaling (subnetwork 6), and ethylene signaling (subnetwork 8) were also detected. It is well known that crosstalk between these phytohormone signaling pathways exists [29, 30]. Hence interfering with one pathway e.g., ABA signaling through BR signaling might affect other phytohormone pathways as well. Based on the pathway analysis using MapMan and the network analysis we can conclude that ABA signaling is mostly a compensatory pathway (genes indicated in red color in Figure 3.5 represent compensatory genes), whereas auxin, ethylene, and ROS signaling pathways are at least partially restored to the WS2 level (genes indicated in green color in Figure 3.5 represent restored genes). However, restoring those pathways to the WS2 level seems to also depend on the presence of at least some compensatory genes (genes indicated in red color in Figure 3.5 represent compensatory genes). The remaining subnetworks are enriched for fatty acid metabolism (subnetwork 3), developmental processes (subnetwork 4), and oligopeptide transport (subnetwork 5) confirming that all suppressors could partially restore some affected primary metabolic pathways. This is in line with the MapMan results and the recovered phenotypes.

3.2.4 Negative feedback of BR signaling on BR biosynthesis:

The network result shows that the BR biosynthesis (subnetwork 7) is not well recovered in any of the suppressors. This subnetwork contains 4 genes (*CYP90C1*, *CYP90B1/DWF4, CYP85A2, CYP90A1/DWF3)* that are differentially expressed in the *bri1-5* lines and all of the suppressors. Those genes, belonging to the cytochrome P450 superfamily play a role in BR biosynthesis by converting the sterol “campesterol” to BRs [31]. *CYP708A3*, another BR biosynthesis gene was

found to be differentially expressed in all suppressors and the *bri1-5* mutant (Figure 3.6). However, as this gene was not present in the interaction network, it was missed by the network analysis. Unlike the four aforementioned cytochrome P450 genes, the molecular function of the *CYP708A3* gene is still unknown. Interestingly it shows an expression pattern that is anticorrelated to that of the other cytochrome P450 genes (Figure 3.6). This result along with the fact that *CYP708A3* is known to be up-regulated by exogenous brassinolide (BL) treatment [32, 33] suggests it acts as an inhibitor of BRs biosynthesis.

The fact that the expression of BR biosynthetic genes is affected by mutations in BR signaling genes points towards the existence of negative feedback of BR signaling on BR biosynthesis. If indeed negative feedback exists between BR signaling and biosynthesis, this feedback should be reflected in quantitative differences in overexpression of the BR signaling and biosynthesis genes in the *bri1-5* and suppressor mutants. The better the signaling can be restored in the suppressors (as reflected by the phenotype), the less we expect the expression of the BR biosynthesis to be aberrant. We indeed found that the expression of the BR-biosynthesis genes (*CYP90C1*, *CYP90A1*, *CYP85A2*, *CYP90B1, CYP708A3*) are less affected in the strains that better mimic the wild-type phenotype (see Figure 3.6, the best suppressor of *bri1-5*, *bri1-5*/*bak1-D,* shows the lowest expression change of the biosynthesis genes). This further supports the existence of negative feedback from BR regulation on BR biosynthesis: a more sustained BR signaling results in decreased BR biosynthesis, whereas suboptimal BR signaling is compensated for by higher transcriptional activity of BR biosynthetic genes.

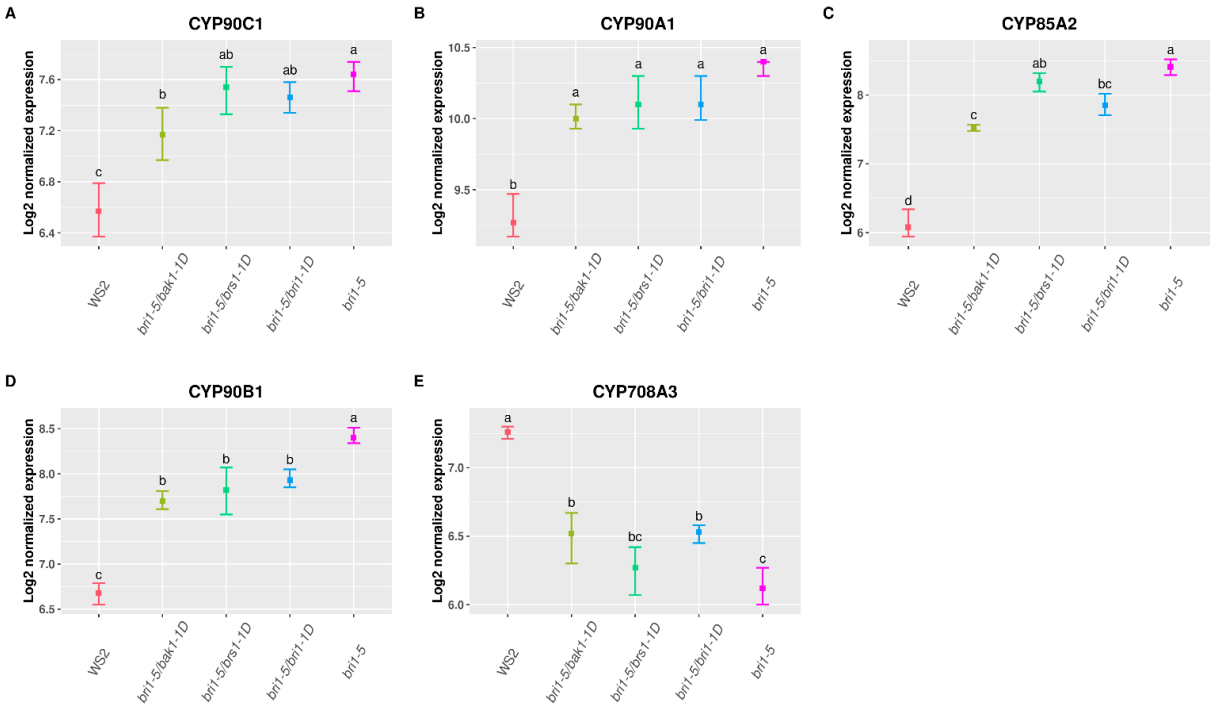


Figure 3.6 Comparing the mean expression values of BR-biosynthesis genes in the *bri1-5* mutant and suppressor lines. For each line (WS2, *bri1-5,* and suppressors) the average log2 expression values of

gene expression across replicates are given for the indicated BRs biosynthesis genes. The squares indicate the location of the mean and bars show the 95 percent confidence interval for the mean. The main BR-biosynthesis genes are affected in the *bri1-5* mutant and all suppressors. However, the plots show that BR-biosynthesis genes are less affected in the line (*bri1-5/bak1-1D)* that best suppresses the *bri1-5* phenotype. Pairwise comparisons between the average values were performed using Tukey's post hoc test. Groups are ranked based on their significance mean where "a" represents the group with the highest mean and "d" the group with the lowest mean. Groups indicated with different letters are significantly different.

3.2.5 Link between stress response and BR signaling

Next, we analyzed the genes that were uniquely altered in each of the suppressors, i.e., the suppressor-specific compensatory genes (genes of group D, group E, and group F, respectively). Based on GO analysis we mainly found stress related processes to be overrepresented in each of the groups. Genes involved in these stress related processes seem to be scattered over the interaction network as they could not be recovered as well delineated subnetworks, indicating that different stress-related genes are induced in the different lines. According to the literature, there is crosstalk between BR signaling and signaling by other hormones in response to stress, especially via ABA and auxin signaling [29]. In the absence of BRs (or low amounts of BRs), BIN2, affects ABA and auxin signaling, resulting in the induction of stress response genes [7, 8, 16]. On the other hand, some stress-response genes are known to be targets of *BZR1* and *BES1* [8, 16] (Figure 3.1), indicating that also when BRs levels are high, stress response genes can be activated. These observations show that balanced BRs levels are needed for normal growth and that deviation from the optimal levels (either too high or too low) would activate stress response mechanisms. We observed that by partially recovering *bri1-5* signaling deficiency by suppressor lines, the transcript level of some stress-response genes is restored to normal, but other stress-response genes become induced (Additional file 3: GO enrichment for genes exclusively differentially expressed in each suppressor, "GO_only_bri1-5", "GO_only_bri1-1D", "GO_only_bak1-1D", "GO_only_brs1-1D"). This observation is in line with this complex effect of BRs and BR signaling on stress response pathways.

3.2.6 Iron Ion homeostasis, ferroxidase, and glutathione transferase activity are identified as compensatory mechanisms unique to the *bri1-5/brs1-1d* suppressor

Unlike for BRI1 and BAK1, much less is known about the role of BRS1 in BR signaling. Therefore, we had a closer look at genes of group D which are exclusively differentially expressed in *bri1-5/brs1-1D* mutant and hence comprise compensatory pathways specific for *bri1-5/brs1-1D*. GO enrichment showed that the genes of this group (group D, 333 genes) are not only overrepresented in stress related processes (see above) but also in glutathione transferase (up-regulated), (Figure 3.7). This overrepresentation in glutathione transferase is in line with the MapMan results. These results showed how in the *bri1-5/brs1-1D* mutant the expression of the glutathione transferase was not only restored as compared to the other

suppressor lines, but even overcompensated as compared to WS2 levels (Figure S3.7). We also found that several members of the CCAAT-binding factor complex (CBC) (*NFYA2*, *NFYA3*, *NFYA6*, *NFYA10*) were uniquely up-regulated in *bri1-1D/brs1-1D* (Figure 3.7 and Figure S3.9). Members of this complex have been associated with the control of iron homeostasis in *Candida Glabrata* [34]. In addition, iron ion homeostasis/ferroxidase activity was also found to be down-regulated, specifically in the *bri1-5/brs1-1D*. In the ferroxidase reaction, four H^+ are used to catalyzes the oxidization of Fe^{2+} to Fe^{3+}, repressing this reaction results in the accumulation of H^+ which can be transported to the apoplast via plasma-membrane pumping mediated by ATPase (H^+-ATPase transporters) [35]. Accordingly, we also found that the main inhibitor of H^+-ATPase transporters, *CBC1*, was significantly down-regulated in *bri1-5/brs1-1D (*fold change -1.7, adj p-value 8.36e-06), but not in the other suppressors. This implies that H^+-ATPase transporters are more active in *bri1-5/brs1-1D* to export H^+ from cytosol into apoplast, making the apoplast more acidic (Figure 3.1). In line with this hypothesis, the up-regulated glutathione transferase activity in the *brs1-1D* mutant (Figure S3.9) might be essential to compensate for the more acidic environment in the *bri1-5/brs1-1D* and would be required for maintaining redox homeostasis. In addition, we hypothesize that the observed acidification could generate a cellular environment that improves BRI1-BRs binding or BRI1-BAK1 dimerization and hence contribute to restoring the *bri1-5* mutant phenotype.

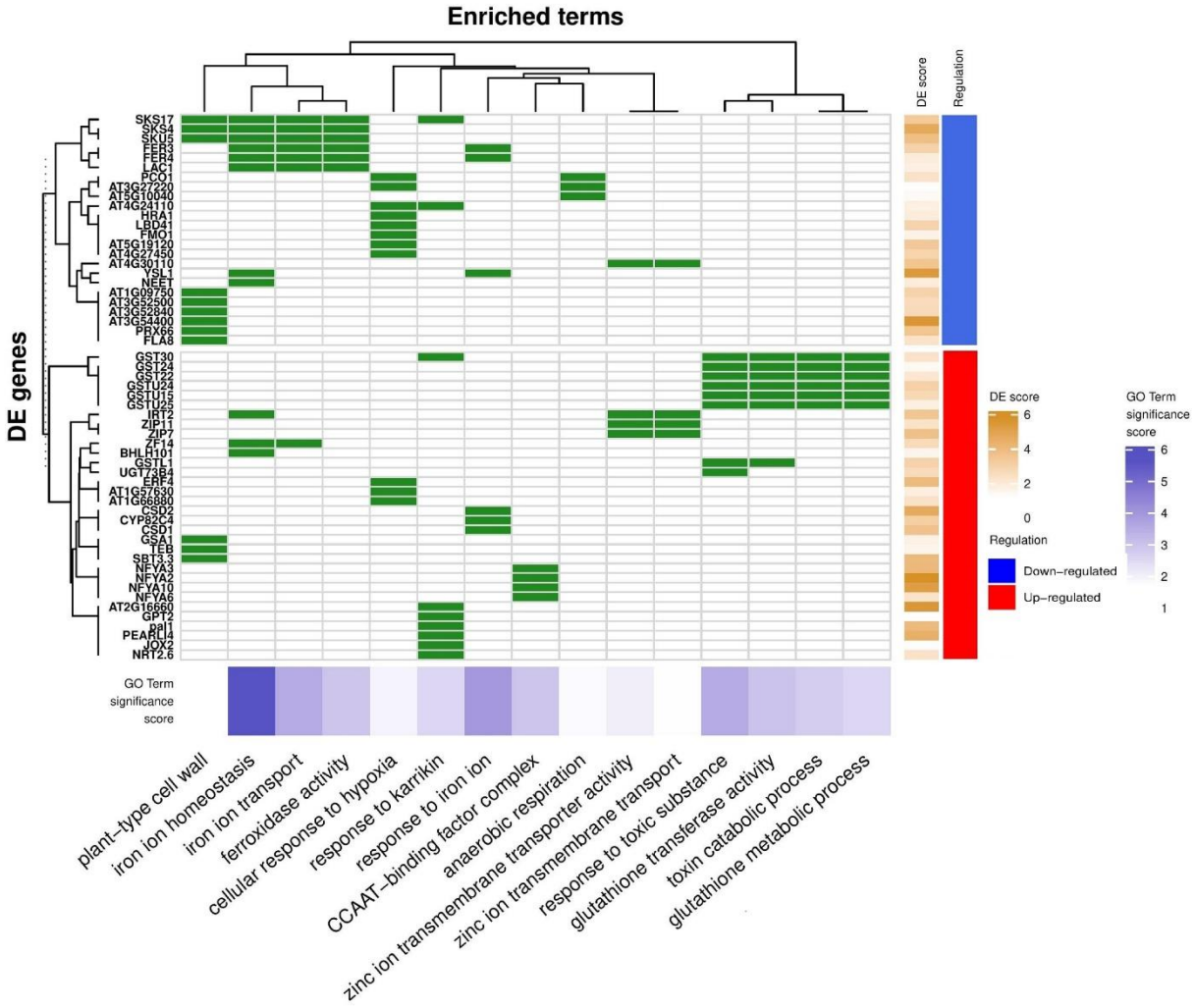


Figure 3.7 GO enrichment for differentially expressed genes (DEGs) exclusively in bri1-5/brs1-1D compared to WS2. The over-represented GO terms and DEG are shown on the x-axis and left-side y-axis, respectively. The green color shows the corresponding gene is present in the indicated GO term and white means it is not; "DE score" reflects the degree of log fold changes (differential expression compared to WS2); "Regulation" represents down (blue) and up (red) regulation for the corresponding gene. The small bottom heat map shows the significant over-representation value for GO terms based on the p-value in the hypergeometric test.

3.3 Discussion

In this study, we explored the alteration of gene expression in BR signaling mutants to better understand BR signaling and the functions of key BR genes.

3.3.1 Crosstalk between BR signaling, other hormone signaling pathways, and primary metabolites

Our analysis identified ABA signaling as a mainly compensatory pathway, and signaling pathways related to ethylene, auxin, and ROS, as pathways involved in partially restoring the *bri1-5* expression phenotype to WS2 levels. It is also obvious that restoring pathways can depend on the presence of compensatory genes in the same pathway. The crosstalk between BR signaling and signaling by ABA observed in our study is in line with the literature [36]. The serine-threonine kinases SnRK2.2 is the main positive regulator of ABA signaling by regulating key TFs such as *SLAC1*, *KAT1* [37, 38]. PHOSPHATASE 2C (PP2C) represses the SnRK2.2 by blocking its phosphorylation mediated by BIN2 (Figure 3.1). In the presence of ABA, the complex of PYR/PYL/RCAR inactivates the PP2C by blocking its substrate's entry [39]. The activated SnRK2 phosphorylate ABI5 leads to the activation of downstream ABA-dependent mechanisms. In the absence of BRs (or BR signaling deficiency like *bri1-5*), SnRK2.3 can mimic the presence of ABA in triggering ABA signaling, once it is phosphorylated by BIN2 [26]. This means that repressing ABA signaling (repressing SnRK2) can compensate for the BR signaling deficiency. In line with our result, a recent study showed that overexpression of *ABI1* or *ABI2,* which encodes the negative regulator of ABA signaling could promote BR signaling [36]. Network analysis suggested that all *bri1-5* suppressor strains suppressed the ABA signaling (SnRKs) through up-regulating its negative regulator, *PP2C* (Figure 3.1).

Links between BR signaling and auxin we observed are also supported by literature and in line with the observed phenotype of the *bri1-5* mutant. ARFs (auxin response factors) are transcription factors that affect root and shoot elongation [30]. In the presence of BRs, BZR1 and BES1 enhance the DNA-binding activity of the auxin response factors ARF6 and ARF7 to promote auxin response [40, 41]. This explains why the suppressor strains restore auxin-related pathways. However, it remains unclear whether the observed restored auxin signaling was the result of the activation of BZR1/BES1 or whether the suppressors employed other pathways to restore the auxin signaling. On the other hand, at low concentrations of BRs, BIN2 enhances the DNA-binding activity of the auxin response factors ARF2, ARF7, and ARF19 through phosphorylation. This results in growth and root elongation in the absence of BRs [38, 42]. Hence, optimal plant growth and development regulated by auxin signaling requires a balanced level of BR signaling. Like with ABA and auxin, crosstalk between BRs and respectively ethylene and cytokinin has also been reported [29].

BR signaling controls development and growth by regulating metabolic processes such as fatty acid metabolism. For example, exogenous BR treatment was shown to promote leaf senescence, likely via the alteration of fatty acid composition in *Pisum sativum* [43]. Exogenous BR treatment would increase the content of free fatty acids and decrease the content of fatty acids bound to galactolipids [43]. Similar studies showed the effect of exogenous BR on fatty acid composition under salt [44] and drought [45] stress. As expected, fatty acid metabolism (subnetwork 3) and developmental processes (subnetwork 4) have been partially restored by *bri1-5* suppressor lines, supporting the impact of BR signaling on the development, and the composition of fatty acids.

Transcription factors, regulators, oligopeptides, and proteins are an essential part of BR and other signaling pathways. Previous studies showed that BR treatment increases the protein levels in the nuclei of hypocotyl cells [12, 46]. Transport of transcription factors and regulators to the nucleus is essential to regulate gene expression by BR signaling as a downstream effect [47]. This is in line with our results: oligopeptide transport (subnetwork 5) is affected in the *bri1-5* mutant and is partially restored by suppressor lines.

3.3.2 Negative feedback between BR signaling and BR biosynthesis

In addition, our results provide evidence for negative feedback between BR signaling and BR biosynthesis. This hypothesis was already made by Noguchi et al. [3] who explained the observed accumulation of BR precursors in *bri1* mutants by the presence of a negative effect of the BR signaling proteins BZR1/BES1 on the BR-biosynthesis pathway. Although it cannot be excluded that the previously reported accumulation of BR precursors in *bri1-5* mutants results from the inability of the mutant lines to use available BRs, our observation in BR signaling mutants suggests that aberrant regulation of BR biosynthesis can also contribute to the accumulation of BRs precursors: it seems that non-aberrant BR signaling is required for homeostasis of appropriate levels of endogenous BRs. Our results also show that the level of negative feedback depends on the degree to which the suppressor could compensate for the phenotypic difference between the *bri1-5* and WS2. The better the defects in BR signaling were alleviated (as reflected by the phenotype), the less pronounced the observed effects on the BR biosynthetic genes. This observation also supports the hypothesis made by Gruszka et al [7] that BIN2 would regulate BR-biosynthesis through phosphorylating CESTA, a transcription factor that regulates BR-biosynthesis (Figure 3.1).

3.3.3 BR signaling and stress response

Because BR signaling regulates response to a wide spectrum of stresses [16], it is not unexpected we observed that stress response genes were affected in the mutants that interfered with BR signaling. At the low level of BRs (or BR signaling deficiency e.g., *bri1-5*) the activated SnRK2 by BIN2 mimics the presence of ABA, activates ABI5, and finally regulates stress response genes (Figure 3.1). In the presence of BRs, BZR1/BES1 inhibits ABI5 and can terminate the ABA signaling. On the other hand, BZR1/BES1 can regulate the expression of stress response independent of ABA (Figure 3.1). Therefore, some stress response genes are regulated by BIN2 at low levels of BRs through ABA signaling, while other stresses are controlled by BZR1/BES1 at high levels of BRs independent of ABA signaling. We observed that by partially recovering *bri1-5* signaling deficiency in *bri1-5* suppressor strains, the transcript level of some stress-response genes is restored to normal, but other stress-response genes become induced. The need for optimal BR homeostasis might also explain why some gain-of-function mutants (e.g., *bes1-D*) described in the literature or why treatment with exogenous BR give rises to a phenotypic response that is worse than the one observed in the *bri1-5* mutant (shorter root) [48]. This further confirms that an appropriate balance in BR

signaling is essential to guarantee coherent cross-talk between hormones signaling networks and any aberration of this optimal level leads to activation of stress response genes.

In addition, there is evidence that BAK1 plays a role in regulating stress-response pathways independently from BR signaling [17]. Since BAK1 usually works as a coreceptor and serves to promote cross-phosphorylation leading to downstream signaling, the existence of other stress-sensor receptors interacting with BAK1 cannot be excluded.

3.3.4 Acidification possibly involved in providing an optimal environment for BRI1 and ligand binding

Our analysis of the genes/pathways that are uniquely involved in the *bri1-5/brs1-1D* suppressor line to compensate for the *bri1-5* mutant showed that BRS1 seems involved in the acidification of the apoplast environment. We hypothesize that this acidification could contribute to an improved BRI1-BRs binding or BRI1-BAK1 dimerization and hence restoration of the *bri1-5* mutant phenotype. In vitro studies have indeed shown that BRs preferentially bind to BRI1 in an acidic cell wall environment (pH < 5) [27, 49]. It has also been suggested that changing the pH environment by endocytosis of BRI1 from the plasma membrane into the cytosol reduces the affinity of BRI1 to BRs and would terminate BR signaling [27]. The same ligand-receptor mechanism has been reported in animal cells [50]. In addition, acidification of the apoplast is the major requirement for increasing cell wall extensibility, which controls cell extension and can also be a compensatory pathway in *brs1-1D* [51]. This is confirmed by the phenotypic analysis which shows that indeed the *bri1-5/brs1-1D* line at least partially restores the epidermal cell length (Figure 3.2). These observations support that the *brs1-1D* mutant can restore BR signaling by creating an acidic environment and providing the optimal conditions for either BRI1-ligand binding or BRI1-BAK1 dimerization along with improving cell wall extensibility. This might also explain the very similar genome-wide expression impact of *brs1-1D* and *bri1-1D* and explains why overexpression of *BRS1* can suppress two weak BRI1 extracellular domain mutants, *bri1-5* and *bri1-9*, but not the strong cytoplasmic domain mutant *bri1-1*.

The *bri1-5/brs1-1D* suppressor line also induces glutathione transferase activity which is necessary for redox homeostasis. This link between BR and redox signaling is in line with the literature [52]. An oxidative environment induces BZR1 activity and promotes the interaction of BZR1 with downstream TFs, ARF6, and PIF4 [52]. Since a loss-of-function mutant of BRS1 shows no obvious phenotype, [4] but its gain-of-function mutant does and partially restores *bri1-5* signal deficiency, it can be suggested that BRS1 does not have a regulatory role and only provides a better condition for triggering the BR signaling by making the apoplast environment more acidic.

3.4 Conclusions

In this study, we performed expression, pathway, and network analysis to provide more insight into the BR signaling by taking advantage of the availability of mutants for keys genes in BR signaling. Our results suggest that ABA signaling plays a significant role in alleviating the *bri1-5* dwarf phenotype. The fact that also other phytohormone signaling pathways are restored to the wild-type expression level in all *bri1-5* suppressors confirms the crosstalk between BR and other phytohormone signalings. The negative feedback from BR signaling on BR biosynthesis was also confirmed by quantitative evidence. In addition, a new indirect role for *BRS1* in BR signaling was suggested.

3.5 Methods

3.5.1 Expression profiling experiment and differential expression analysis

The two activation tagging suppressors of *bri1-5, bri1-5/bak1-1D,* and *bri1-5/brs1-1D* were obtained from our previous study [13]. An additional activation tagging suppressor line *bri1-5*/*bri1-1D* was generated in this study as previously described [4]. Wild-type (WS2), the loss-of-function BR mutant (*bri1-5*), and its three suppressor mutants (*bri1-5*/*brs1-1D*, *bri1-5*/*bak1-1D*, *bri1-5*/*bri1-1D)* were grown at 22°C in a long-day condition (16 h of light and 8 h of dark) in a greenhouse for seven days. All mutants were generated from the WS2 ecotype background. Microarray analysis of all genotypes was performed with three biological replicates. Total RNA of 7-day whole seedlings was isolated, labeled, and hybridized with an Arabidopsis ATH1 genome array according to the Affymetrix instructions. Scanning of the array was performed using the Agilent GeneArray Scanner. The data are available in GEO (GSE70843). The CEL files were preprocessed using the AFFY package (background correction, quantile normalization, and probe value summarization (RMA normalization)) [53]. The arrayQualityMetrics package was used to check the quality of the normalized expression values [54]. All samples passed the quality check. Non-unique probe sets were removed, and the expression value of genes was calculated as the average of the expression measured by the probe sets that covered the gene. The consistency between replicate samples was assessed using PCA. Differential expression was calculated by comparing the samples of the mutated lines with those obtained from the wild-type (WS2) using the Limma package [55]. To define differentially expressed genes, the absolute fold change and false discovery rate (FDR) threshold were set at 1.5 and 0.05, respectively resulting in a total of 1413 differentially expressed genes. Further, eight differentially expressed genes were randomly selected and confirmed by RT-qPCR following standard protocol with three biological and three technical repeats [56]. The primers used for qPCR are presented in Table S3.3. Pathway enrichment analysis was carried out using the MapMan software [57] and GO enrichment was performed using TAIR GO enrichment [58].

3.5.2 Phenotypic analysis

The phenotypic impact of suppressor lines was evaluated through the measurement of the root, hypocotyl, and epidermal cell length on the 7-day old seedlings. The root and hypocotyl lengths were measured on 40 plants and hypocotyl epidermal cells were scanned using the electron microscope (SEM) on samples taken from 3 random plants for each line. The mean of epidermal cell length per image was determined using Fiji software [59]. Statistical analysis was performed using One-way ANOVA followed by Tukey's multiple comparisons in R [60].

3.5.3 Retrieving BR-responsive genes

We performed a literature study to compile a list of "high-confidence" BR-responsive genes, i.e., genes regulated by exogenous BR treatment or in a line containing a gain-of-function mutation in a BR signaling gene with consistent transcriptional response from at least five references [32, 33, 61-69]. A full list of marker genes together with the references in which they were found and an indication of their expression behavior upon addition of external BRs (or gain of mutation in BR signaling genes) is given in supplementary file (Additional file 2). The intersection of the genes that were differentially expressed in the *bri1-5* mutant line with the list of "high-confidence" BR-responsive genes was used to evaluate the consistency between the expression data of our study and literature and to determine which suppressor best restored the expression of those genes to WS2 level.

3.5.4 Network analysis

A high confidence Arabidopsis interaction network was compiled from the following sources: 64185 FunTFBS regulatory interactions were obtained from PlantRegMap [55], 96827 protein-protein interactions from AtPIN [70] and 34003 metabolic interactions from KEGG [71]. This resulted in a final number of 182748 interactions between 21263 unique genes. In this integrated network, nodes represented genes and edges the interactions between the genes.

To perform network analysis with PheNetic [25] network edges need to be weighted. The weight is derived from the log2 fold change (logFC) expression of a gene as indicated below. To each gene, we assigned as logFC, the highest value that was observed for this gene across the assessed mutant lines as compared to WS2. To assign a p-value to this logFC, we empirically estimated the distribution of the observed max logFC for all genes. As we expected that most of the genes would not change their expression this is an estimate of the null distribution. The mean (μ) and standard deviation (σ) of this distribution was estimated empirically using maximum likelihood implemented in the MASS package [72]. Using the values of the mean and the standard deviation each gene was assigned a significance score (gene-score) based on a two-tailed T-test which reflects the degree to which the gene has been affected as compared to other genes as follows [25]:

$$\text{gene\_score} = \text{abs}(1 - 2 * \phi_{(\mu,\sigma)}\max(\text{logFC}))$$

The edge weight between a source (S) and target (T) nodes was derived by the product of the S gene-score and T gene-score.
This weighted interaction network was used together with the seed gene list (list of differentially expressed genes) in PheNetic [24, 25]. PheNetic aims at connecting as many genes as possible from the seed list on the interaction network in the most parsimonious way (using the least number of edges). By enforcing such a parsimonious solution, PheNetic detects subnetworks in which genes from the seed list are closely connected. Such connected components can be viewed as proxies of pathways. The PheNetic *expression* subcommand was run in the downstream mode with the following parameters: min cost: 0.1; max cost: 5; step size: log scale between max and min cost with 28 steps; path-length=4; k-best paths: 50; for all other parameters the default values were used. For each edge cost, the highest-scoring subnetwork was selected. Furthermore, the Jaccard index of all the subnetworks with the same edge cost was computed. For each cost, the subnetwork is rejected if it has a low stability score (i.e., Jaccard index smaller than 0.5) or if it is too large (more than 500 interactions). The final subnetworks are then the union of all the "best subnetworks" for each edge penalty that passed the stability and size requirements.

3.5.5 Gene groups used for network analysis

The gene groups used for network analysis were derived from the Venn diagram displayed in Figure 3.3. To perform network analysis, we defined respectively restoring genes (group A, genes with altered expression in the *bri1-5* mutant, but restored to WS2 level in at least 2 suppressors), compensatory genes (group B, (genes not differentially expressed in the *bri1-5* mutant, but differentially expressed in at least two suppressor strains), non-restored genes (group C, genes that differentially expressed in the *bri1-5* mutant and at least two of the suppressor strains) and genes that are uniquely affected in respectively the *bri1-5/brs1-1D* (group D), the *bri1-5/bak1-1D* (group F), and the *bri1-5/bri-1D* (group E) (Additional file 1).

To study the interaction between restored, compensatory, and non-restored genes, we combined the genes of groups A, B, and C (789 genes) to perform network analysis. Each group consists of its core genes. The core of group A (restored genes) consists of 118 genes that are restored to the wild-type state by all suppressors. The core of group B (compensatory genes) consists of 23 genes, i.e., genes that were significantly differentially expressed in all three *bri1-5* suppressors, but not in *bri1-5* mutant. The core of group C (non-restored genes) consists of 149 genes that are affected in *bri1-5* mutant but not restored by any suppressors. However, as the size of the gene sets is dependent on the choice of an arbitrary threshold, we assumed that some of the genes belonging to the processes represented by these core gene sets were found to be significantly differentially expressed in two of the three lines only (and slightly below the threshold in the third). That is why we extended the core gene sets with the genes that were differentially expressed in at least two lines, rather than in all three of them. Extended gene sets A, B and C were subsequently combined to study the interaction between the restored, compensatory, and non-restored genes.

Authors' contributions

Conceptualization: J.L, T.S., R.S.R., K.M.; Methodology: T.S., R.S.R.; Investigation: D.Z., X.G., J.Y., G.M; Writing – Original Draft: T.S., R.S.R, K.M; Writing – Review & Editing: J.L., K.M., G.M; Funding Acquisition: J.L., K.M, R.S.R., T.S.; Resources: J.L.; Supervision: K.M., J.L. All authors read and approved the final manuscript.

3.6 References

1. Seyed Rahmani R, Shi T, Zhang D, Gou X, Yi J, Miclotte G, Marchal K, Li J: **Genome-wide expression and network analyses of mutants in key brassinosteroid signaling genes.** *BMC genomics* 2021, **22:**1-17.
2. Clouse SD, Sasse JM: **BRASSINOSTEROIDS: Essential Regulators of Plant Growth and Development.** *Annu Rev Plant Physiol Plant Mol Biol* 1998, **49:**427-451.
3. Noguchi T, Fujioka S, Choe S, Takatsuto S, Yoshida S, Yuan H, Feldmann KA, Tax FE: **Brassinosteroid-insensitive dwarf mutants of Arabidopsis accumulate brassinosteroids.** *Plant physiology* 1999, **121:**743-752.
4. Li J, Lease KA, Tax FE, Walker JC: **BRS1, a serine carboxypeptidase, regulates BRI1 signaling in Arabidopsis thaliana.** *Proc Natl Acad Sci U S A* 2001, **98:**5916-5921.
5. Li J, Chory J: **A putative leucine-rich repeat receptor kinase involved in brassinosteroid signal transduction.** *Cell* 1997, **90:**929-938.
6. Kinoshita T, Cano-Delgado A, Seto H, Hiranuma S, Fujioka S, Yoshida S, Chory J: **Binding of brassinosteroids to the extracellular domain of plant receptor kinase BRI1.** *Nature* 2005, **433:**167-171.
7. Gruszka D: **The brassinosteroid signaling pathway-new key players and interconnections with other signaling networks crucial for plant development and stress tolerance.** *Int J Mol Sci* 2013, **14:**8740-8774.
8. Planas-Riverola A, Gupta A, Betegon-Putze I, Bosch N, Ibanes M, Cano-Delgado AI: **Brassinosteroid signaling in plant development and adaptation to stress.** *Development* 2019, **146**.
9. Nam KH, Li J: **BRI1/BAK1, a receptor kinase pair mediating brassinosteroid signaling.** *Cell* 2002, **110:**203-212.
10. Wang X, Kota U, He K, Blackburn K, Li J, Goshe MB, Huber SC, Clouse SD: **Sequential transphosphorylation of the BRI1/BAK1 receptor kinase complex impacts early events in brassinosteroid signaling.** *Dev Cell* 2008, **15:**220-235.
11. Chen W, Lv M, Wang Y, Wang PA, Cui Y, Li M, Wang R, Gou X, Li J: **BES1 is activated by EMS1-TPD1-SERK1/2-mediated signaling to control tapetum development in Arabidopsis thaliana.** *Nat Commun* 2019, **10:**4164.
12. Wang Z-Y, Nakano T, Gendron J, He J, Chen M, Vafeados D, Yang Y, Fujioka S, Yoshida S, Asami T: **Nuclear-localized BZR1 mediates brassinosteroid-induced growth and feedback suppression of brassinosteroid biosynthesis.** *Developmental cell* 2002, **2:**505-513.
13. Li J, Wen J, Lease KA, Doke JT, Tax FE, Walker JC: **BAK1, an Arabidopsis LRR receptor-like protein kinase, interacts with BRI1 and modulates brassinosteroid signaling.** *Cell* 2002, **110:**213-222.
14. Wen J, Li J, Walker JC: **Overexpression of a serine carboxypeptidase increases carpel number and seed production in A rabidopsis thaliana.** *Food and Energy Security* 2012, **1:**61-69.
15. Nolan TM, Brennan B, Yang M, Chen J, Zhang M, Li Z, Wang X, Bassham DC, Walley J, Yin Y: **Selective Autophagy of BES1 Mediated by DSK2 Balances Plant Growth and Survival.** *Dev Cell* 2017, **41:**33-46.e37.

16. Peres A, Soares JS, Tavares RG, Righetto G, Zullo MAT, Mandava NB, Menossi M: **Brassinosteroids, the Sixth Class of Phytohormones: A Molecular View from the Discovery to Hormonal Interactions in Plant Development and Stress Adaptation.** *Int J Mol Sci* 2019, **20**.
17. Chinchilla D, Zipfel C, Robatzek S, Kemmerling B, Nurnberger T, Jones JD, Felix G, Boller T: **A flagellin-induced complex of the receptor FLS2 and BAK1 initiates plant defence.** *Nature* 2007, **448:**497-500.
18. Kemmerling B, Schwedt A, Rodriguez P, Mazzotta S, Frank M, Qamar SA, Mengiste T, Betsuyaku S, Parker JE, Mussig C, et al: **The BRI1-associated kinase 1, BAK1, has a brassinolide-independent role in plant cell-death control.** *Curr Biol* 2007, **17:**1116-1122.
19. Li J: **Multi-tasking of somatic embryogenesis receptor-like protein kinases.** *Curr Opin Plant Biol* 2010, **13:**509-514.
20. Albrecht C, Boutrot F, Segonzac C, Schwessinger B, Gimenez-Ibanez S, Chinchilla D, Rathjen JP, de Vries SC, Zipfel C: **Brassinosteroids inhibit pathogen-associated molecular pattern-triggered immune signaling independent of the receptor kinase BAK1.** *Proc Natl Acad Sci U S A* 2012, **109:**303-308.
21. Kim BH, Kim SY, Nam KH: **Genes encoding plant-specific class III peroxidases are responsible for increased cold tolerance of the brassinosteroid-insensitive 1 mutant.** *Mol Cells* 2012, **34:**539-548.
22. Kim SY, Kim BH, Lim CJ, Lim CO, Nam KH: **Constitutive activation of stress-inducible genes in a brassinosteroid-insensitive 1 (bri1) mutant results in higher tolerance to cold.** *Physiol Plant* 2010, **138:**191-204.
23. Guo H, Li L, Aluru M, Aluru S, Yin Y: **Mechanisms and networks for brassinosteroid regulated gene expression.** *Current opinion in plant biology* 2013, **16:**545-553.
24. De Maeyer D, Renkens J, Cloots L, De Raedt L, Marchal K: **PheNetic: network-based interpretation of unstructured gene lists in E. coli.** *Mol Biosyst* 2013, **9:**1594-1603.
25. De Maeyer D, Weytjens B, Renkens J, De Raedt L, Marchal K: **PheNetic: network-based interpretation of molecular profiling data.** *Nucleic Acids Res* 2015, **43:**W244-250.
26. Cai Z, Liu J, Wang H, Yang C, Chen Y, Li Y, Pan S, Dong R, Tang G, de Dios Barajas-Lopez J: **GSK3-like kinases positively modulate abscisic acid signaling through phosphorylating subgroup III SnRK2s in Arabidopsis.** *Proceedings of the National Academy of Sciences* 2014, **111:**9651-9656.
27. Belkhadir Y, Jaillais Y: **The molecular circuitry of brassinosteroid signaling.** *New Phytol* 2015, **206:**522-540.
28. Takahashi Y, Zhang J, Hsu P-K, Ceciliato PH, Zhang L, Dubeaux G, Munemasa S, Ge C, Zhao Y, Hauser F: **MAP3Kinase-dependent SnRK2-kinase activation is required for abscisic acid signal transduction and rapid osmotic stress response.** *Nature communications* 2020, **11:**1-12.
29. Choudhary SP, Yu J-Q, Yamaguchi-Shinozaki K, Shinozaki K, Tran L-SP: **Benefits of brassinosteroid crosstalk.** *Trends in plant science* 2012, **17:**594-605.
30. Tian H, Lv B, Ding T, Bai M, Ding Z: **Auxin-BR interaction regulates plant growth and development.** *Frontiers in plant science* 2018, **8:**2256.
31. Fujioka S, Yokota T: **Biosynthesis and metabolism of brassinosteroids.** *Annual review of plant biology* 2003, **54:**137-164.
32. Goda H, Sawa S, Asami T, Fujioka S, Shimada Y, Yoshida S: **Comprehensive comparison of auxin-regulated and brassinosteroid-regulated genes in Arabidopsis.** *Plant physiology* 2004, **134:**1555-1573.
33. Yu X, Li L, Zola J, Aluru M, Ye H, Foudree A, Guo H, Anderson S, Aluru S, Liu P: **A brassinosteroid transcriptional network revealed by genome-wide identification of BESI target genes in Arabidopsis thaliana.** *The Plant Journal* 2011, **65:**634-646.
34. Kumar K, Askari F, Sahu MS, Kaur R: **Candida glabrata: a lot more than meets the eye.** *Microorganisms* 2019, **7:**39.
35. Michelet B, Boutry M: **The plasma membrane H+-ATPase (A highly regulated enzyme with multiple physiological functions).** *Plant Physiology* 1995, **108:**1.

36. Wang H, Tang J, Liu J, Hu J, Liu J, Chen Y, Cai Z, Wang X: **Abscisic acid signaling inhibits brassinosteroid signaling through dampening the dephosphorylation of BIN2 by ABI1 and ABI2.** *Molecular plant* 2018, **11:**315-325.
37. Kulik A, Wawer I, Krzywińska E, Bucholc M, Dobrowolska G: **SnRK2 protein kinases—key regulators of plant response to abiotic stresses.** *Omics: a journal of integrative biology* 2011, **15:**859-872.
38. Lin Z, Li Y, Zhang Z, Liu X, Hsu C-C, Du Y, Sang T, Zhu C, Wang Y, Satheesh V: **A RAF-SnRK2 kinase cascade mediates early osmotic stress signaling in higher plants.** *Nature communications* 2020, **11:**1-10.
39. Miyazono K-i, Miyakawa T, Sawano Y, Kubota K, Kang H-J, Asano A, Miyauchi Y, Takahashi M, Zhi Y, Fujita Y: **Structural basis of abscisic acid signalling.** *Nature* 2009, **462:**609-614.
40. Zhou X-Y, Song L, Xue H-W: **Brassinosteroids regulate the differential growth of Arabidopsis hypocotyls through auxin signaling components IAA19 and ARF7.** *Molecular plant* 2013, **6:**887-904.
41. Liu K, Li Y, Chen X, Li L, Liu K, Zhao H, Wang Y, Han S: **ERF72 interacts with ARF6 and BZR1 to regulate hypocotyl elongation in Arabidopsis.** *Journal of experimental botany* 2018, **69:**3933-3947.
42. Cho H, Ryu H, Rho S, Hill K, Smith S, Audenaert D, Park J, Han S, Beeckman T, Bennett MJ: **A secreted peptide acts on BIN2-mediated phosphorylation of ARFs to potentiate auxin response during lateral root development.** *Nature cell biology* 2014, **16:**66-76.
43. Fedina E, Yarin A, Mukhitova F, Blufard A, Chechetkin I: **Brassinosteroid-induced changes of lipid composition in leaves of Pisum sativum L. during senescence.** *Steroids* 2017, **117:**25-28.
44. Pokotylo I, Kretynin S, Khripach V, Ruelland E, Blume YB, Kravets V: **Influence of 24-epibrassinolide on lipid signalling and metabolism in Brassica napus.** *Plant growth regulation* 2014, **73:**9-17.
45. Zafari M, Ebadi A, Sedghi M, Jahanbakhsh S: **Alleviating effect of 24-epibrassinolide on seed oil content and fatty acid composition under drought stress in safflower.** *Journal of Food Composition and Analysis* 2020, **92:**103544.
46. Yin Y, Wang Z-Y, Mora-Garcia S, Li J, Yoshida S, Asami T, Chory J: **BES1 accumulates in the nucleus in response to brassinosteroids to regulate gene expression and promote stem elongation.** *Cell* 2002, **109:**181-191.
47. Ryu H, Kim K, Cho H, Hwang I: **Predominant actions of cytosolic BSU1 and nuclear BIN2 regulate subcellular localization of BES1 in brassinosteroid signaling.** *Molecules and cells* 2010, **29:**291-296.
48. González-García M-P, Vilarrasa-Blasi J, Zhiponova M, Divol F, Mora-García S, Russinova E, Caño-Delgado AI: **Brassinosteroids control meristem size by promoting cell cycle progression in Arabidopsis roots.** *Development* 2011, **138:**849-859.
49. She J, Han Z, Kim T-W, Wang J, Cheng W, Chang J, Shi S, Wang J, Yang M, Wang Z-Y: **Structural insight into brassinosteroid perception by BRI1.** *Nature* 2011, **474:**472-476.
50. Maxfield FR, McGraw TE: **Endocytic recycling.** *Nature reviews Molecular cell biology* 2004, **5:**121-132.
51. Hager A: **Role of the plasma membrane H+-ATPase in auxin-induced elongation growth: historical and new aspects.** *Journal of plant research* 2003, **116:**483-505.
52. Lv B, Tian H, Zhang F, Liu J, Lu S, Bai M, Li C, Ding Z: **Brassinosteroids regulate root growth by controlling reactive oxygen species homeostasis and dual effect on ethylene synthesis in Arabidopsis.** *PLoS Genetics* 2018, **14:**e1007144.
53. Gautier L, Cope L, Bolstad BM, Irizarry RA: **affy--analysis of Affymetrix GeneChip data at the probe level.** *Bioinformatics* 2004, **20:**307-315.
54. Kauffmann A, Gentleman R, Huber W: **arrayQualityMetrics--a bioconductor package for quality assessment of microarray data.** *Bioinformatics* 2009, **25:**415-416.
55. Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, Smyth GK: **limma powers differential expression analyses for RNA-sequencing and microarray studies.** *Nucleic Acids Res* 2015, **43:**e47.

56. Wong ML, Medrano JF: **Real-time PCR for mRNA quantitation.** *Biotechniques* 2005, **39:**75-85.
57. Thimm O, Bläsing O, Gibon Y, Nagel A, Meyer S, Krüger P, Selbig J, Müller LA, Rhee SY, Stitt M: **MAPMAN: a user-driven tool to display genomics data sets onto diagrams of metabolic pathways and other biological processes.** *The Plant Journal* 2004, **37:**914-939.
58. Lamesch P, Berardini TZ, Li D, Swarbreck D, Wilks C, Sasidharan R, Muller R, Dreher K, Alexander DL, Garcia-Hernandez M: **The Arabidopsis Information Resource (TAIR): improved gene annotation and new tools.** *Nucleic acids research* 2012, **40:**D1202-D1210.
59. Schindelin J, Arganda-Carreras I, Frise E, Kaynig V, Longair M, Pietzsch T, Preibisch S, Rueden C, Saalfeld S, Schmid B: **Fiji: an open-source platform for biological-image analysis.** *Nature methods* 2012, **9:**676-682.
60. Team RC: **R: A language and environment for statistical computing.** Vienna, Austria; 2013.
61. Goda H, Shimada Y, Asami T, Fujioka S, Yoshida S: **Microarray analysis of brassinosteroid-regulated genes in Arabidopsis.** *Plant Physiology* 2002, **130:**1319-1334.
62. Müssig C, Fischer S, Altmann T: **Brassinosteroid-regulated gene expression.** *Plant physiology* 2002, **129:**1241-1251.
63. Nemhauser JL, Mockler TC, Chory J: **Interdependency of brassinosteroid and auxin signaling in Arabidopsis.** *PLoS Biol* 2004, **2:**e258.
64. Vert G, Nemhauser JL, Geldner N, Hong F, Chory J: **Molecular mechanisms of steroid hormone signaling in plants.** *Annu Rev Cell Dev Biol* 2005, **21:**177-201.
65. Mouchel CF, Osmont KS, Hardtke CS: **BRX mediates feedback between brassinosteroid levels and auxin signalling in root growth.** *Nature* 2006, **443:**458-461.
66. Nemhauser JL, Hong F, Chory J: **Different plant hormones regulate similar processes through largely nonoverlapping transcriptional responses.** *Cell* 2006, **126:**467-475.
67. Goda H, Sasaki E, Akiyama K, Maruyama-Nakashita A, Nakabayashi K, Li W, Ogawa M, Yamauchi Y, Preston J, Aoki K: **The AtGenExpress hormone and chemical treatment data set: experimental design, data evaluation, model data analysis and data access.** *The Plant Journal* 2008, **55:**526-542.
68. Guo H, Li L, Ye H, Yu X, Algreen A, Yin Y: **Three related receptor-like kinases are required for optimal cell elongation in Arabidopsis thaliana.** *Proceedings of the National Academy of Sciences* 2009, **106:**7648-7653.
69. Sun Y, Fan X-Y, Cao D-M, Tang W, He K, Zhu J-Y, He J-X, Bai M-Y, Zhu S, Oh E: **Integration of brassinosteroid signal transduction with the transcription network for plant growth regulation in Arabidopsis.** *Developmental cell* 2010, **19:**765-777.
70. Brandao MM, Dantas LL, Silva-Filho MC: **AtPIN: Arabidopsis thaliana protein interaction network.** *BMC Bioinformatics* 2009, **10:**454.
71. Kanehisa M, Goto S: **KEGG: kyoto encyclopedia of genes and genomes.** *Nucleic acids research* 2000, **28:**27-30.
72. Ripley B, Venables B, Bates DM, Hornik K, Gebhardt A, Firth D, Ripley MB: **Package 'mass'.** *Cran r* 2013, **538:**113-120.

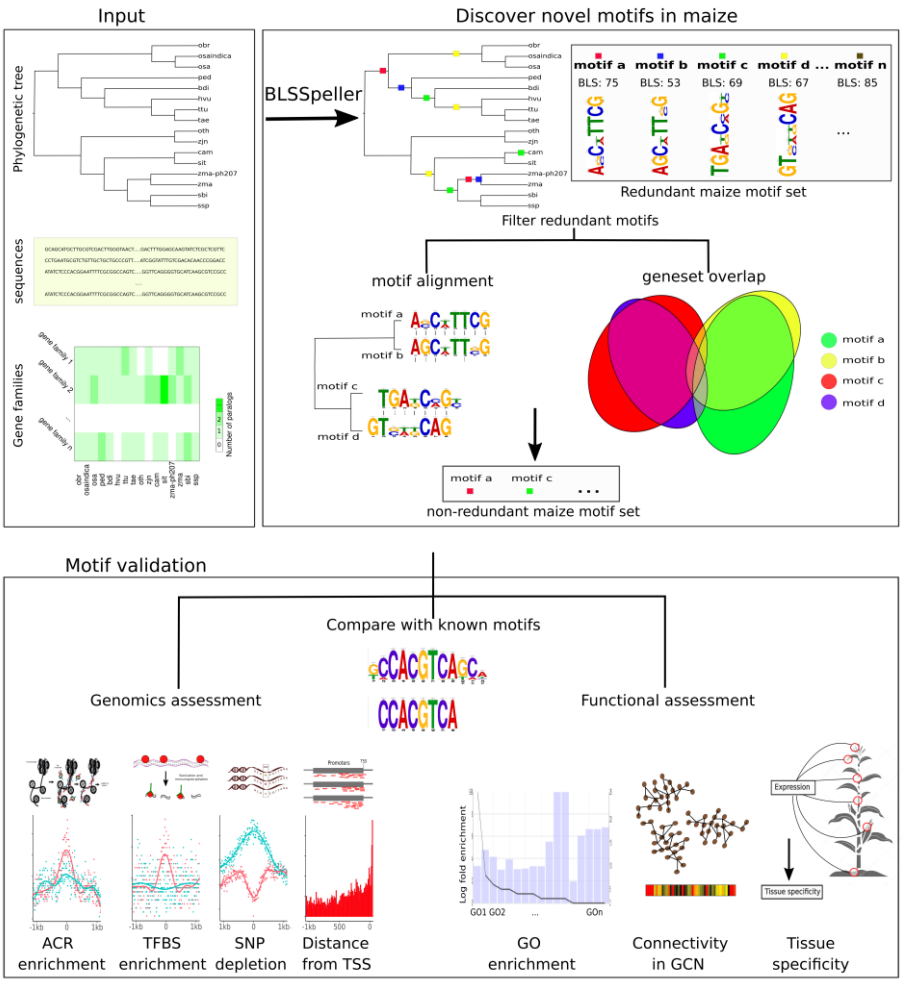
4 Chapter 4

BLSSpeller to discover novel regulatory motifs in maize

"Nothing in biology makes sense except in the light of evolution"

-Theodosius Dobzhansky

Plants need to quickly respond to internal and environmental changes. Understanding gene regulation is key to understanding the genotype-phenotype relation. The signals for gene expression regulation are clusters of DNA binding sites recognized by transcription factors (TFs) hidden in non-coding regions. Identifying the location of the functional TF binding sites, which are short and degenerate, can contribute to our understanding of gene regulation and are relevant for applications of e.g. genome editing in which mutations are introduced to obtain a desired phenotype. In the chapter below we describe a study in which we adapted BLSSpeller, a comparative approach to identify TF binding sites and applied it to identify motifs in *Zea mays*. The candidate preprocessed the data for BLSSpeller, implemented additional steps for redundancy filtering, performed the validation analyses, contributed to the visualization, and summarized the results in a coherent story. Page 4-22 includes an overview of the contributions of all authors.



BLSSpeller to discover novel regulatory motifs in maize

Razgar Seyed Rahmani, Dries Decap, Jan Fostier*, Kathleen Marchal*

Under review in *DNA Research*, 2022

Abstract

With the decreasing cost of sequencing and availability of larger numbers of sequenced genomes, comparative genomics methods are becoming increasingly attractive to complement experimental techniques for the task of transcription factor binding site identification. In this study, we redesigned BLSSpeller, a motif discovery algorithm, to cope with larger sequence datasets. BLSSpeller was used to identify novel motifs in *Zea mays* in a comparative genomics setting with 16 monocot species. We discovered 61 motifs of which 20 matched previously described motif models in Arabidopsis. In addition, novel, yet uncharacterized motifs were detected, several of which are supported by available sequence-based and/or functional data. Instances of the predicted motifs were enriched around transcription start sites and contained

signatures of positive selection. Moreover, the enrichment of the predicted motif instances in open chromatin and transcription factor binding sites indicates their functionality, supported by the fact that genes carrying instances of these motifs were often found to be coexpressed and/or enriched in similar GO functions. Overall, our study unveiled several novel candidate motifs that might help our understanding of the genotype to phenotype association in crops.

4.1 Introduction

Most of the genetic variation associated with phenotypic variation in plants is located in non-coding regions [1, 2]. Hence, uncovering the functional regulatory signals hidden in non-coding regions is crucial to improve our understanding of transcriptional regulation from both a fundamental and applied point of view [3-5]. Specific cis-regulatory elements (CRE) are recognized by specific transcription factors (TFs) and play an important role in regulating the rate and timing of gene expression. Maize is one of the world's most important crops, with a well-studied genome and a large body of experimental data on gene regulation already available, including the experimental identification of transcription factor binding sites [5, 6], and the characterization of active chromatin regions [7, 8]. However, like for most crop species, the availability of experimental information is, for technical [7, 9, 10] and budgetary reasons, limited. Comparing ChIP-seq TF binding site profiles obtained between a wild strain and a strain carrying a mutation in the TF has elucidated binding sites of the studied TF [11, 12]. However, those studies are restricted to the profiling of a handful of TFs in maize [6, 13-16]. Even one of the most comprehensive large-scale profiling DAP-seq experiments covered only 104 TFs of the ~2000 annotated TFs in maize, and focused on leaf tissue only [5]. In principle, open chromatin identification methods like ATAC-seq also allow identifying regions with putative regulatory function, but with low resolution (non-specific peaks covering hundreds to a thousand bp) and in a condition-dependent way [17].

However, experimental information on regulation can be complemented with computational analyses. Comparative genomics methods to accurately pinpoint the location of functional TF binding sites were already popular in the pre-genomics era [18-21] and have regained scientific interest with the decreasing cost of sequencing and availability of many sequenced genomes [22-24]. This is particularly true for plants, where TF binding sites are known to be well-conserved across closely related species and often located in the close neighbourhood of the coding genes [5, 7, 25]. Comparative approaches have been successfully applied to identify transcription factor binding sites in many organisms [25-28].

Even though many tools for phylogenetic footprinting have been developed, they are not designed to cope with the large volumes of the data that are currently available. BLSSpeller is a unique approach that exhaustively explores the full sequence space [29]. It allows, like many state-of-the-art comparative approaches [19, 27, 30] to account for the phylogenetic relatedness between the orthologs during its search for conserved motifs. In addition, by allowing for an alignment-free search of motifs conserved across species [29], it can discover binding sites that

were relocated during evolution. In order to include a larger number of sequences in the comparative analysis, BLSSpeller was redesigned. This ability increases its statistical power and reduces spurious predictions.
BLSSpeller was used to discover novel motifs in *Zea mays*, several of which were supported by complementary sequence-based and/or functional information. Overall, we discovered 20 motifs that perfectly matched previously described motif models in Arabidopsis, together with several yet undescribed motifs generating a useful resource of novel predictions that can complement results from functional genomics studies.

4.2 Results

We used BLSSpeller [29] to identify conserved motifs in a comparative genomics setting and validated our predictions by means of publicly available genomics and functional data. To identify motifs in a comparative setting, we obtained from PLAZA [31] gene sequences of 16 monocot species, including maize (Table S4-1). Genes were grouped into gene families where each gene family can contain both paralogs and orthologs. BLSSpeller was used to search in the promoter sequences of gene families with at least 3 species for conserved motifs.

Details on BLSSpeller can be found in the materials and methods. Briefly, BLSSpeller scores, for all possible motifs of a prespecified length (e.g., 8 bp), their degree of conservation within each gene family, denoted by the Branch Length Score (BLS) [32]. Conceptually, the BLS of a motif expresses the fraction of promoter sequences in a gene family that contain the motif, weighted by their relative phylogenetic distance. When the BLS of a motif within a gene family exceeds a prespecified BLS threshold, the motif is said to be *conserved* within that gene family. We used different thresholds on the BLS to also enable the identification of motifs that were conserved in only a subset of the species of a gene family. Conservation is soft constrained by allowing for some degeneracy in a motif. The BLS assigns more relevance to conservation in different species (orthologs) than conservation within a species (paralogs).

To reduce false-positive predictions, an additional selection criterion is imposed: given that biological processes tend to be regulated by multiple genes, a true motif is more likely to be conserved in multiple gene families. Therefore, we identified as the more reliable motifs, those predictions that were conserved within more gene families than random motifs of the same length and nucleotide composition. The degree to which a predicted motif is conserved within more gene families than a random motif is reflected by its recurrence score (see Materials and Methods) and is computed for multiple BLS thresholds. Motifs with a recurrence score of 0.9 or higher for any of the considered BLS thresholds were retained. A recurrence score of 0.9 for a given BLS threshold means that the observed motif is conserved (at that BLS threshold) within ten times as many gene families than expected. This filtering step largely reduced the number of motif candidates and resulted in a final list of 1295 motifs.

Figure S4-1-a shows the effect of the filtering based on the recurrence threshold. It shows that the fraction of retained motifs after applying the recurrence-based filtering increases with the

threshold on the BLS, indicating that motifs with a lower BLS are less likely to occur significantly more often across families than random motifs and hence that these motifs are indeed more likely to be spurious.

To investigate the impact of the number of paralogs in a gene family on the BLS of the motifs in that family, we plotted per gene family the number of motifs with a high BLS (> 0.95) as a function of the average number of paralogs over species within a gene family, before and after applying the filtering based on the recurrence score threshold (Figure S4-1, panels b and c). Figure S4-1-b shows that even though the BLS scheme downweighs the impact of paralogs, it is not entirely independent of the average number of paralogs present in gene families as the number of predicted motifs increases with an increase in the average number of paralogs. This is to be expected as a higher number of paralogs implies a larger sequence space and hence a higher probability of detecting by chance a conserved motif with a high BLS. As shown in Figure S41-c, recurrence filtering removes many of these likely spurious motifs detected in motif families with a high average number of paralogs.

4.2.1 Identifying motifs and instances relevant to *zea mays*

To select from the motifs predicted by BLSSpeller those that are relevant in maize, we assessed for each motif whether gene families exist that contained a motif instance in the corresponding maize sequences. This resulted in a final selection of 1292 motifs and 2,320,402 instances. However, many of these motifs are redundant as for instance the same motif can be recovered with a different level of degeneracy. Therefore, redundant motifs were removed based on the degree of similarity between the motifs and the degree to which the motifs covered similar genes (see Materials and Methods). This resulted in a final list of 61 non-redundant motifs, each of which with at least 20 instances in different maize genes (Appendix S1). The average GC content of these motif instances was 62% as compared to an average GC content of 45% for the maize promoter regions. For integration with complementary genomics data (see below), also redundant motif instances were removed, and a set of random motifs and instances was generated that has the same nucleotide content and distribution of the number of instances in maize as the predicted motifs (see Materials and Methods).

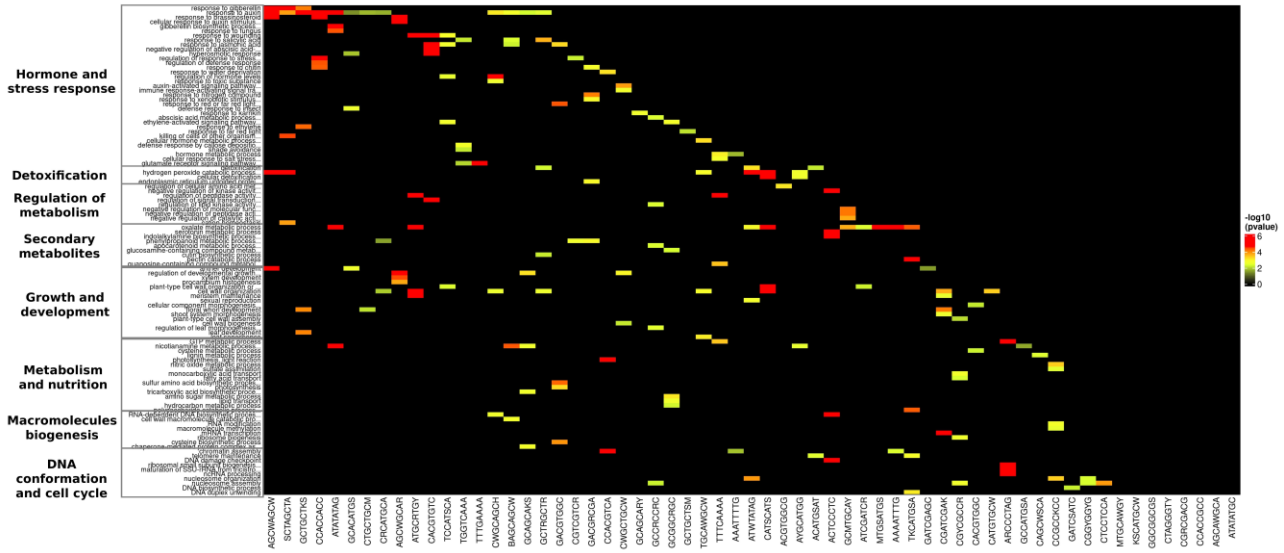


Figure 4.1 Enriched GO terms (Biological Process) for the gene sets corresponding to the predicted maize motifs. Each column represents a gene set in maize sharing the indicated motif and each row indicates a biological process that is found enriched amongst the genes in the gene set. The entries indicate the -log10 p-values of the GO enrichment. Only the five most significantly enriched GO terms are shown for each gene set. The GO terms were grouped into eight groups based on their biological similarity.

4.2.2 Predicted motifs are associated with processes known to be conserved across species

Genes with the same motifs are expected to be coregulated and hence involved in the same biological processes. To assess whether the gene sets containing the same motif are indeed functionally similar, we performed GO enrichment. For each motif, a representative gene set was compiled by taking for each gene family in which the motif occurred, the genes of maize that contained an instance of the considered motif. The gene sets corresponding to 53 out of the 61 motifs were enriched for at least one biological function (adj. p-value < 0.01) (Figure 4-1). For the genes sets corresponding to the random motifs this was only true for 15 out of the 61 random motifs (Figure S4-2). Enriched GO terms of the representative gene sets related to hormone and stress response, detoxification, regulation of metabolism, secondary metabolites, growth and development, metabolism and nutrition, macromolecules biogenesis, and DNA conformation and cell cycle (Figure 4-1). Although overall, the gene sets of the different motifs covered a variety of biological processes, processes related to “hormone and stress response” were most frequently found enriched, indicating that, as expected, the genes carrying the predicted motifs are involved in key processes known to be conserved across species [33-35].

4.2.3 Validation of the predicted maize motifs by comparing with Arabidopsis motifs

To validate the motifs, we compared the 61 maize motifs with experimentally verified motifs in *A. thaliana* obtained from a DAP-seq experiment [36]. The study of O’Malley et al. provides

one of the most comprehensive compendia of experimentally generated TF binding sites in plants, covering binding sites for 529 TFs. Our predicted maize motifs match significantly better to known Arabidopsis motifs than random motifs. Of the 61 maize motifs, 20 motifs perfectly match (q-value < 0.05) known Arabidopsis motifs, whereas the same was true for 3 random motifs only (Figure S4-3). Based on these similarities, the retrieved motifs cover Arabidopsis binding sites for a diverse set of transcription factors, including members of C2H2, AP2, bHLH, NLP, ERF, ARE, HMG, ABI (Table S4-2).

Subsequently, we assessed the functional similarity between our predicted motifs and their corresponding matching motifs in Arabidopsis. Hereto, we assumed the function of the Arabidopsis motif was the same as the function of the TF associated with the motif (TAIR website). For the maize motifs, the function was inferred by performing GO enrichment **on the representative gene set** (set of genes that share an instance of the same motif) for each of the motifs**.** Table S4-2 shows that at least for some motifs, a clear similarity can be detected between the inferred functions of the predicted motifs in maize and their matching motifs in Arabidopsis. The large overlap between our predicted motifs and experimentally validated motifs supports the relevance of our predictions.

Although for most of the predicted motifs a unique one to one match with an Arabidopsis motif was found, the Arabidopsis motif, NLP_tnt.AtNLP4, shows a high similarity with 4 different maize motifs (GCAGCARY, GCAGCAKS, AGCWGCAR, BAGCAGCW) (Table S4-2), suggesting these motifs might still be redundant or might represent binding sites for different TFs with similar functions in maize. Indeed the maize genes having an instance of each of the 4 aforementioned motifs are enriched in “metabolism and development”, consistent with the AtNLP4 function in Arabidopsis which involved in cell elongation and response to nitrate.

4.2.4 Genes sharing instances of the same predicted motif are coexpressed

Assuming that sharing a motif implies coregulation and hence coexpression, we assessed whether maize genes that have instances of a similar motif also exhibit a similar expression profile. Hereto, we used a previously published expression compendium of maize [37], comprising 8 developmental stages of the same genotype, 28 tissue-specific datasets sampled from the same tissue from multiple inbred lines, 5 tissue-genotype datasets originating from multiple tissues of specific inbred lines and 4 datasets from recombinant inbred populations (see Table S4-3 for details).

To measure similarities in coexpression, we built coexpression networks. The different studies in the compendium each capture gene expression associations in different biological contexts. The compendium is therefore highly unbalanced in the number of matching tissues/conditions. To avoid that the coexpression analysis would be biased by this unbalance, we built a coexpression network for the data of each study separately, resulting in 45 coexpression networks. To assess whether sets of genes that share instances of the same motif were also coexpressed, we measured the degree to which they were connected in each of the 45 coexpression networks (see Materials and Methods). The results show that for the majority of

the gene sets sharing a motif, the genes are significantly more connected than random gene sets of the same size (Figure 4-2, Figure S4-4). This observation further corroborates the functionality of our predicted motifs.

The degree to which the coregulated gene sets displayed a connectivity in the coexpression network increased with the number of samples in the dataset for which the coexpression network was built and with the type of samples that were profiled: In contrast to networks derived from tissue profiling experiments, recombinant inbred line (RIL) networks did not capture strong connectivity between the genes sharing a motif, despite the fact that they were derived from experiments with relatively many samples (Figure 4-2).

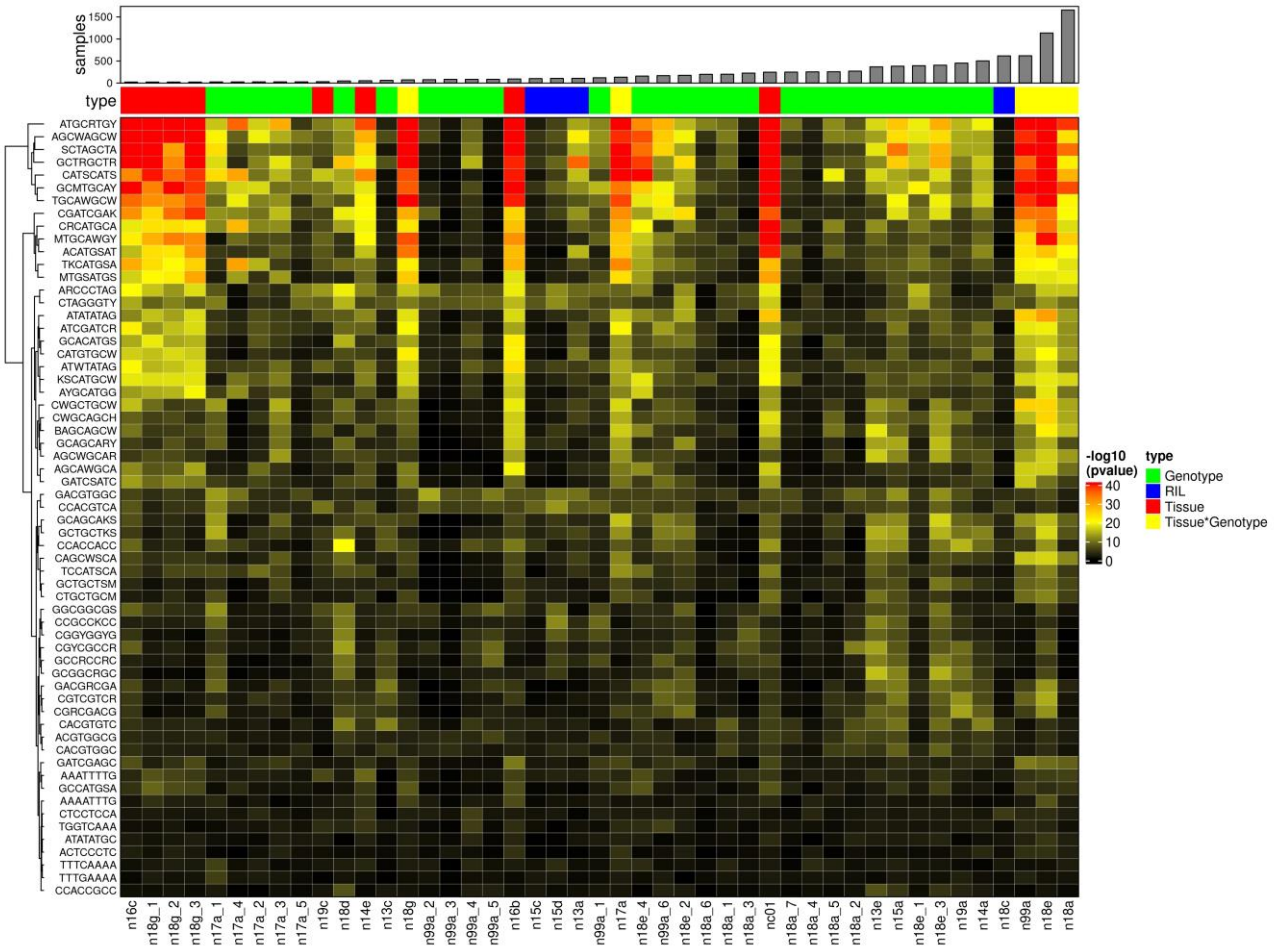


Figure 4.2 Connectivity of gene sets sharing the same predicted motif in the coexpression networks. Rows display the different motifs for which a gene set is considered. Columns display the different experimental conditions for which a coexpression network was constructed (sorted by sample size). Entries indicate the -log10 p-values of observing the same connectivity in a coexpression network by chance as was observed for the gene sets that share the same predicted motif. The colored annotation bar indicates the type of experimental dataset, and the top annotation bar indicates the sample size of each experimental dataset.

Like comparing with known motifs, the coexpression analysis showed that some motifs might either be variants of one true binding site or represent binding sites of closely related TFs. For example, gene sets with instances of respectively the motif ATWTATAG or ATATATAG in their promoter regions are connected in the same coexpression networks (Figure 4-2). This high connectivity in the same coexpression networks indicates that those motifs probably represent the same binding site and should have been merged into a single motif. Similar observations were made for the motifs [AGCWAGCW, TGCAWGCA], [GCTRGCTR, SCTAGCTA], [CRCATGCA, TKCATGSA, ACATGSAT], [GCTGCTKS, GCAGCAKS].

4.2.5 Tissue-specific genes have a larger number of predicted motifs

The timing, level, and tissue-specificity of gene expression depends on the presence of cis regulatory elements that recruit specific transcription factors in response to tissue demands [38]. We investigated whether an association exists between the degree to which a maize gene displays a tissue specific expression pattern and the number of motifs that were predicted to have an instance in the promoter of the gene. To assess the level of tissue-specific expression of a gene, we used RNA-seq expression data from 8 different tissues profiled in maize B73 line [39]. Tissue specific expression was measured using the Tau index [40] (Materials and Methods). Figure 4-3-a shows that in general, the more a gene has instances of different motifs, the higher the gene's degree of tissue-specific expression (higher Tau index). This observation is in line with literature indicating that tissue specific genes have evolved numerous unique binding sites during evolution and that the combinatorial effects of several binding sites may be critical to provide the proper response to developmental, condition, and tissues specific demands [41, 42]. Unlike the predicted motifs, for random motifs no relationship between the number of instances of different motifs and the degree of tissue specific expression (Tau index) was observed (Figure 4-3-c).

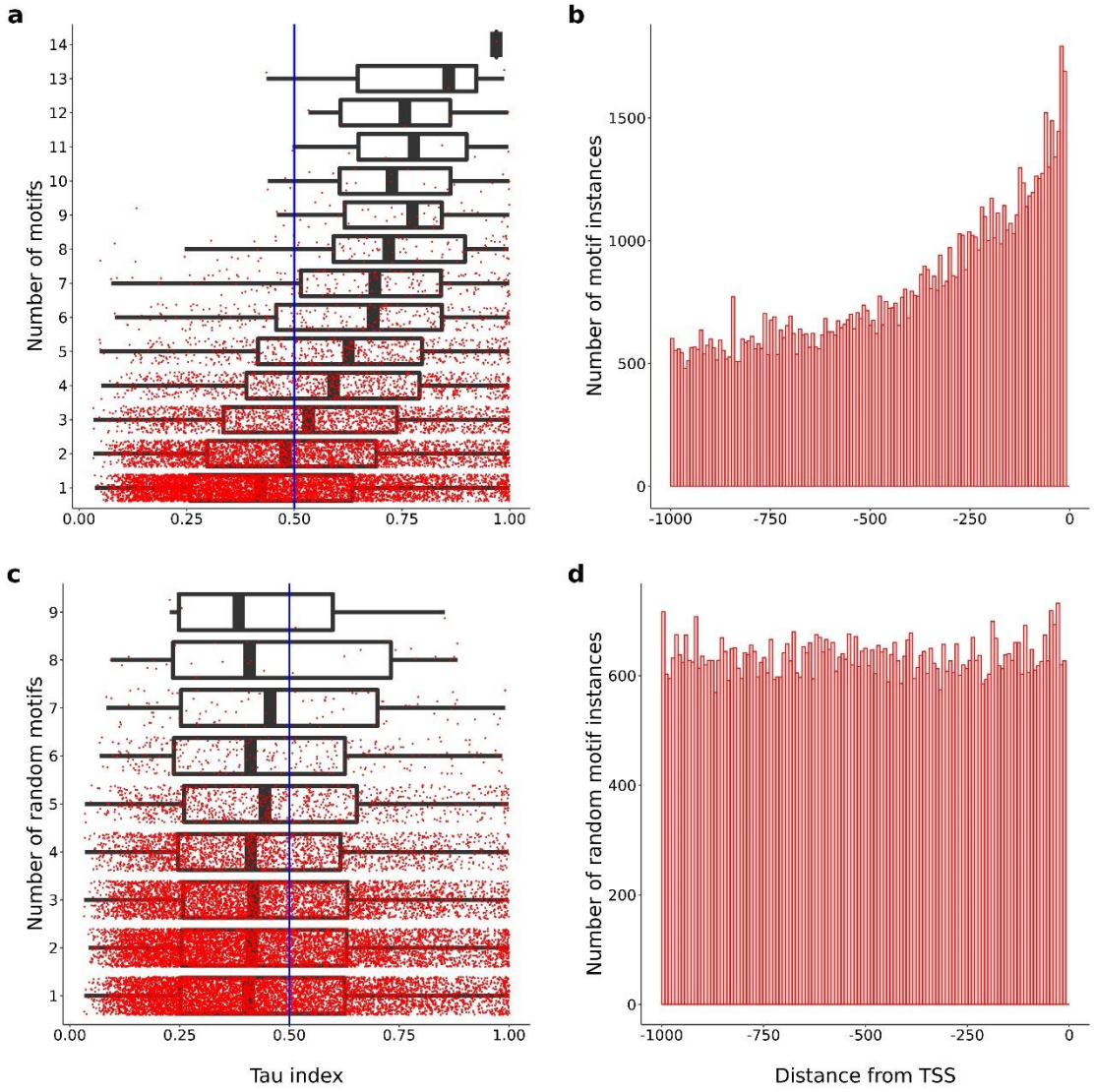


Figure 4.3 Position of the predicted motifs relative to the TSS and tissue specificity of the expression of gene sets carrying instances of the predicted motifs in maize. Panels a and c show how the tissue specificity of gene expression (Tau index) is associated with the number of motifs that have at least one instance in the gene. X-axis: Tua index, y-axis: number of unique (non-redundant) motifs. Panels b and d show the association between the positions of the motif instances in maize and the location of the TSS for respectively the predicted and the random motifs. x-axis: distance from the TSS; y-axis: number of motif instances (61 non-redundant motifs) per 8-bp bins across 1000 bp upstream of TSS.

4.2.6 Location of the predicted motif instances is biased towards the TSS

The region in the close vicinity of the TSS is known to play a central role as binding site for TFs [43]. Therefore, the specific positional relationship between transcription factor binding sites and the TSS has widely been used to predict the validity of transcription factor binding sites [44-46]. To validate the motifs predicted by BLSSpeller, we assessed whether the instances of our predicted motifs in maize were more frequently located around the TSS site than

instances of random motifs. Figure 4-3-b and d, show that this is indeed the case, further supporting the biological validity of our motif predictions in maize.

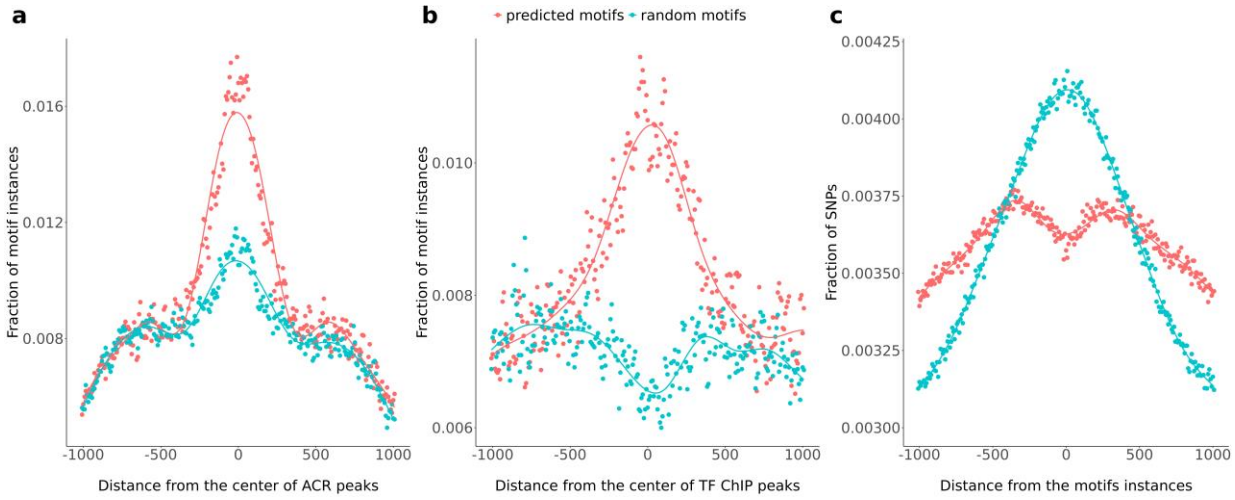


Figure 4.4 Overlap between the location of the predicted motif instances and respectively ACR, ChIP-seq binding locations and SNPs in maize. (a): Fraction of respectively the predicted and random motif instances that overlap with accessible chromatin regions and their flanking regions. To include flanking regions, the sequences were extended 1 kb up- and 1 kb downstream of the center of the open chromatin region. (b): Fraction of respectively the predicted and random motif instances and the ChIP-seq corresponding binding sites of 104 TFs. The ChIP-seq binding sites are extended 1 kb up and downstream of the center of the binding peak to include flanking regions. (c): Fraction of SNPs that overlap with the location of a motif instance and their surroundings. To include flanking regions, the sequences were extended 1 kb up and 1 kb downstream of the center of the location of motif instances of respectively the predicted and random motifs.

4.2.7 Predicted motif instances overlap with open chromatin regions and TF binding sites

To further validate our predicted motifs, we assessed to what extent instances of the predicted motifs were occurring more frequently than expected by chance in nucleosome-depleted regions and/or in regions known to functionally bind TFs. Hereto, we used the data from a study that profiled accessible chromatin regions (ACR) in maize at single-cell resolution using more than 50,000 cells from 6 organs using scATAC-seq [7], and a comprehensive ChIP-seq study in maize that profiled the putative DNA binding sites for 104 TFs in leaf tissue [5].
Figure 4-4 a-b shows that a significant overlap exists between our predicted motif instances and respectively the nucleosome-depleted regions and ChIP-bound regions. Both panels show that our predicted motif instances occur more frequently in the peaks of the ACRs and ChIP-seq data than random motif instances. For ACRs, also the random motif instances are enriched near the centers of the experimentally identified ACRs, but not as extreme as what we observe for the predicted motif instances. The random motifs were generated to have sequence properties

similar to the ones of the predicted motifs. Because ACRs have a high GC content [7], the enrichment of random motif instances near the center of the ACR peaks indicates that the enrichment of the predicted motif instances in the ACRs is partially the result of their GC content. In contrast to what is observed for the ACR peaks, random motif instances are depleted in the proximity of the ChIP-seq binding sites. These observations further confirm the relevance of our predicted motifs.
The same overlap between the predicted motif instances and active chromatin regions and TF binding regions (ACRs and ChIP-seq binding sites), but then for each motif separately, can be seen in Figures S4-5-9. The results show that for the majority of the predicted motifs, a trend similar to the one displayed in Figure 4-4 is observed. The trend observed in Figure 4-4 a-b is therefore not driven by the pattern of only a few predicted motifs.

4.2.8 Predicted motifs are under selection

True binding sites tend not to tolerate mutations that interfere with their functionality. To further validate whether this was also true for the predicted motifs we compared, the frequency with which these predicted motifs accumulated SNPs in their binding sites and flanking regions (1-kb up- and downstream of the locations of the predicted motif instances). Figure 4-4-c shows that the regions centered around the predicted motif instances are more depleted in SNPs than the regions centered around random motif instances. Figures S4-5-9 show the same pattern at the level of the individual motifs. The results presented in Figures S4-5-9 clearly show that most motifs are under selective constraints and therefore likely functionally relevant.
The degree of polymorphism in the regions flanking the random motif instances was lower than in the regions centered around the random motif instances (Figure 4-4-c) and drops faster than what was observed for the regions flanking the predicted motif instances. Repeating the analysis using different sets of random motifs showed that this pattern was robust. This decrease in polymorphism for regions flanking random motif instances can be explained by the fact that moving away from a random motif instance (intergenic region) increases the chance to reach a genic region for which a higher level of selective constraint and hence higher conservation is expected [47, 48]. The fact that this rate of decrease in polymorphism is lower for the predicted motifs might be the result of linkage of the flanking regions with beneficial SNPs within the motifs. In other words, it suggests the existence of a fitness trade-off for an individual between gaining a deleterious SNP in the flanking region versus a deleterious SNP in the predicted binding site: the potential deleterious effect of SNPs in regions flanking a true binding site can be offset by the beneficial effect of not having a SNP at the position of a true functional motif. Such compensatory effect cannot be achieved for random motifs and hence they tolerate fewer SNPs in their flanking regions, which are more likely to be genic regions. Overall, the strong depletion of SNP close to the predicted motif instances suggests that both natural and artificial selection has resulted in the fixation of variants in the predicted motifs in individuals with favorable phenotypes.

4.2.9 Prioritization of potential novel motifs in *zea mays*

Supplementary Figures S4-5-9 show the results of the aforementioned genomic and functional assessments at the level of the individual motifs. Overall, most motifs seem to be supported by at least one genomic and/or functional line of evidence or feature. For several predicted motifs, the inferred function based on GO enrichment of the genes sets representative for the motif in maize corresponds to the function of the matching motif-TF pair in Arabidopsis: predicted maize motifs corresponding to gene sets enriched in hormone and stress response, matched the Arabidopsis motifs of known master regulators of response to hormones and stress in plants i.e., C2C2, bHLH34, NLP4, TGA4, ERF13, ABI3, MYB74 (Figures S4-5-9). Likewise, motifs of gene sets enriched in growth and development showed high similarity with the motifs of several main players in plant development ERF6, AREB3 bZIP42, GATA20, ZIM. One predicted motif of which the target gene set was enriched in DNA conformation and cell cycle showed a high similar to the Arabidopsis motif of TBP3 (ARCCCTAG) (Figure S4-5, motif 1). TBP3 encodes a telomeric DNA binding protein and belongs to Single Myb Histone (SMH) gene family, of which the members preferentially bind to double-stranded telomeric repeats [49].

However, not all predicted motifs could be clearly associated with known TF binding sites in Arabidopsis. To prioritize the most reliable motif predictions, we used additional genomics and functional data to support our findings.

According to their shared supporting features, 5 groups of predicted motifs were distinguished. The first group of motifs show a high similarity to known Arabidopsis motifs and are particularly strongly supported by all levels of genomic evidence (Figure S4-5): compared to their flanking regions and random motif instances, the location of the predicted motif instances in this group is strongly centered around the TSS, is depleted for polymorphisms, and shows overlap with active chromatin and ChIP-seq bound regions. These motifs occur in genes that tend to be broadly expressed in many tissues (lower Tau index) and that are enriched in processes that are widely conserved across species such as “DNA conformation and cell cycle”, “Metabolism and nutrition”, and “Regulation of metabolism” (Figure S4-5). The more uniform expression behavior across conditions of these genes might explain why their mutual correlation and hence also connectivity in the coexpression network is more difficult to capture (as their expression might not sufficiently vary across conditions). Figure 4-5-a shows a representative motif from this group of motifs.

Motifs belonging to group 2 also show a high similarity to Arabidopsis motifs and are particularly well supported by the functional evidence and by the enrichment of their motif instances upstream of the TSS (Figure S4-6). Because their target genes show high variability in expression across conditions (high Tau index), they also show relatively high connectivity in the coexpression networks, representative of several different conditions. Their target genes are highly enriched for “secondary metabolites”, “hormone and stress response” and “growth and development”, processes for which a high level of polymorphism between individuals has been described [50, 51]. The overlap between motif instances and SNPs confirms that indeed the

motifs in this group tend to be associated with a rather high level of polymorphism (an increase rather than a depletion of SNPs at the motif position). Such a high level of polymorphism could be associated with natural and artificial selection in maize which benefits from heterosis and is associated with increased fitness. If functionally divergent alleles enable adaptation to different environments, spatially heterogeneous natural selection (balancing selection) might maintain locus-specific polymorphism [52]. This is in line with the fact that genes in this group are enriched for "secondary metabolites", and "hormone and stress response" processes of which are responsible for adaptation and are under more relaxed purifying selection or under stronger diversifying selection [50, 53-55]. In addition, the less clear support provided by the overlap between motif instances and active chromatin regions (ACR and ChIP-seq) can be due to the mismatch between the conditions in which the chromatin profiling experiments were performed and the conditions under which the genes carrying the motifs of this group are expressed, which are as shown in Figure S4-6 rather tissue-specific. A representative motif of this group is shown in Figure 4-5-b.

The third group represents potentially novel motifs that show no match with a corresponding Arabidopsis motif (insignificant p-value and presence of strong mismatch in at least one position) (Figure S4-7), but that are well supported by genomic and/or functional assessments. Motifs in this group were sorted based on their Tau-index (from least to most tissue-specific, Figure S4-7). Also here, we observe -similar to what we noticed for motifs of groups 1 and 2- that the higher the tissue specificity of the expression of the genes carrying the motif, the more the genes show connectivity in the corresponding tissue-specific coexpression networks and the less they were supported by ACR, ChIP-seq, SNP evidence. A representative motif of this group is shown in Figure 4-5-c.

Motifs belonging to group 4 (Figure S4-8) and group 5 (Figure S4-9) are less clearly supported by additional evidence. Even though the genes carrying these motifs show enrichment in a certain GO function (groups 4 and 5) and that motifs of group 4 (Figure S4-8) in addition to this show similarity to known Arabidopsis motifs, we could not find any support for these predictions through the genomics or expression-based evidence (low connectivity in the coexpression network).

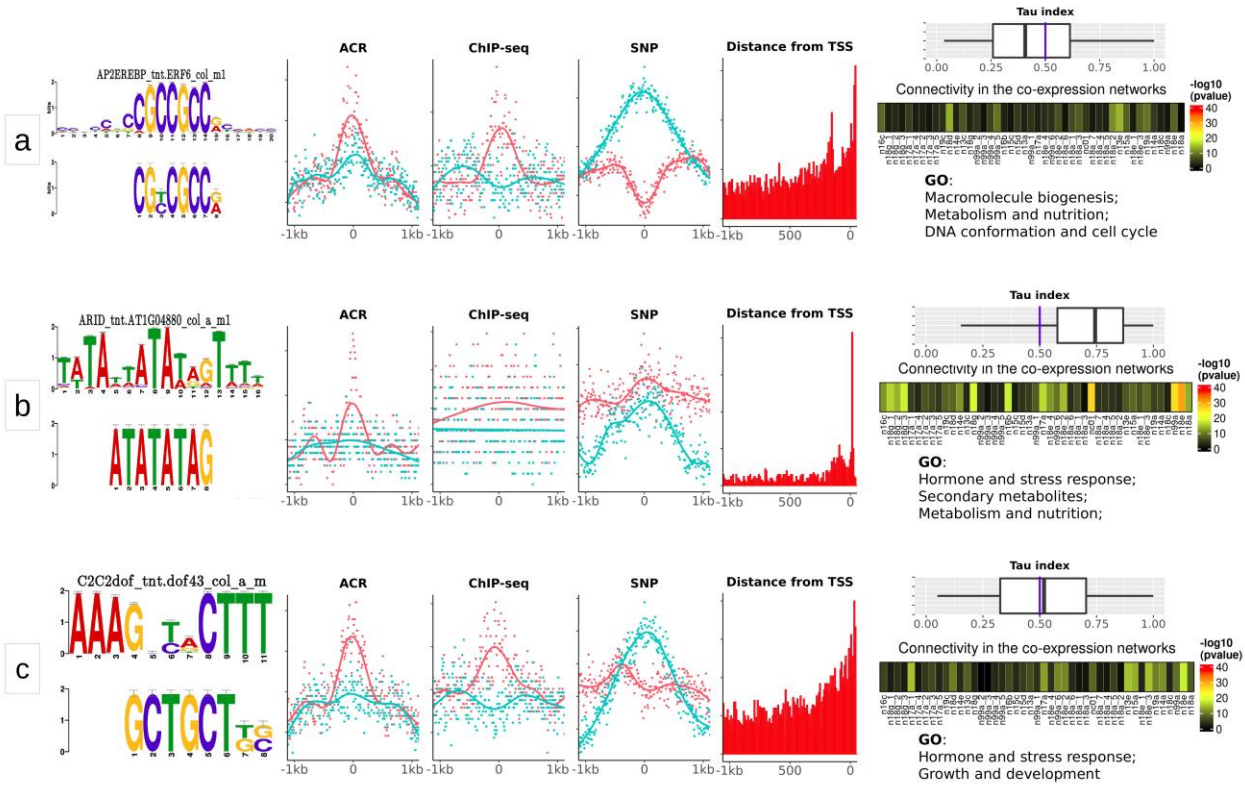


Figure 4.5 Representative motifs for each group. a) group 1: motifs that show a high similarity to known Arabidopsis motifs, with instances on non-tissue specific genes in maize, and that are strongly supported by all levels of genomic evidence; b) group 2: group of motifs similar to Arabidopsis motifs, with instances in tissue-specific genes in maize, and strongly supported by functional evidence; c) group 3: novel motifs that do not show any similarity with Arabidopsis motifs, but supported by genomics and/or functional evidence. For each row: left panel: most similar Arabidopsis motif (top) and predicted motif (bottom); middle panels: overlap between the motif instances (predicted: red and random: blue) and respectively accessible chromatin regions (ACRs), ChIP-seq binding sites, the occurrence of SNPs, and distance from TSS; Right panel: from top to bottom tissue specificity distribution (Tau index), connectivity in coexpression networks, and GO enrichment for genes having at least one instance of the predicted motif under consideration.

4.3 Discussion

With the increasing availability of sequenced genomes, comparative analyses to identify transcription factor binding sites are an attractive alternative to complement tedious experimental approaches. In this work, we present BLSSpeller, an exhaustive alignment-free search for motifs conserved in orthologous sequences of related species. We applied BLSSpeller to identify motifs in *Zea mays* by using a comparative genomics approach using sixteen monocot species.

Overall, we observed that despite the high "AT" content of plant promoters, the motifs identified by BLSSpeller are enriched in "GC". This is in line with previous studies in plants which also

observed that regions in which transcription factor binding sites occur are GC enriched i.e., open chromatin regions [7, 56] or experimentally identified binding sites [5].
Complementing the predictions with publicly available complementary data sources allowed pinpointing several promising motif candidates in *Zea mays*. Overall, we identified several motifs, both motifs with a match to known motifs in Arabidopsis, but also novel motif candidates that were supported by additional genomic assessments: the location of the instances of these predicted motifs are occurring more frequently than expected by chance in the regions upstream of the TSS, in active chromatin regions or ChIP-seq bound regions. In addition, the instances of these motifs tend to be depleted of SNPs, consistent with the action of purifying selection. Also, functional assessments could support the motifs: genes that share instances of the same predicted motifs are enriched in similar GO functions and tend to be more often coexpressed than random gene sets (displaying connectivity in the coexpression networks). We observed that there was an inverse relationship between the level to which predicted motifs were supported by the functional versus the genomic assessments. Genes that were functionally well supported by expression analysis tended to display a more tissue-specific expression and were therefore less supported by the large, but condition-dependent chromatin binding assays. In contrast, processes that were ubiquitously expressed and not condition dependent were less supported by expression connectivity but well supported by the available chromatin accessibility data. This indicates that the chromatin accessibility landscape varies largely between conditions and is likely a major factor in contributing to tissue-specific expression and that expression connectivity is easier to capture for processes that are variably expressed across conditions.
So, summarizing we showed that BLSSpeller, a high performant alignment for motif detection procedure, can successfully be used to predict novel motifs in a comparative way. Combining such predictions with available genomics and functional data allowed further elucidating transcriptional regulation in *Zea mays*. Although their impact requires further characterization, the provided motifs offer a large and valuable source for further investigation.

4.4 Materials and Methods

4.4.1 Datasets used for motif detection

The reference genomes, the structural annotations for 16 species, and orthologous gene families (inferred using the PLAZA integrative method) were downloaded from the PLAZA monocots 4.5 [31]. The details about the size of the genome, number of chromosomes, and annotated genes of the species used in this study are provided in Table S4-1. Each gene family consists of all homologous (orthologous and paralogous) genes from all 16 species (41,970 gene families). Gene families with orthologs in less than 3 species were removed, resulting in a total of 21,727 gene families used for motif detection.
Motif detection was performed on 1 kb upstream of the transcription start site (TSS). For the genes located on the negative strand, the reverse complement of the extracted sequence was

considered. Low-complexity and homopolymeric sequences were masked using RepeatMasker [57] prior to performing motif detection.

4.4.2 BLSSpeller to perform Phylogenetic Footprinting

The original implementation of BLSSpeller [29] was limited to only few species. To improve its computational performance and enable the analysis of larger datasets, BLSSpeller was reimplemented in Apache Spark in order to take advantage of parallel, distributed-memory compute platforms. Compute-intensive parts, e.g., enumerating conserved motifs in a gene family, were implemented in the C++ programming language for efficiency reasons. Additional functionality was implemented to pinpoint the location of the conserved motifs in each species. Below and in Figure S4-10, we summarize the key steps of BLSSpeller. For every gene family (Figure S4-10-a), a generalized suffix tree, truncated at depth *k*, is constructed of the promotor sequences of genes within that gene family. The suffix tree is traversed in a depth-first manner to exhaustively enumerate all motifs of a prespecified length *k*. The motifs can contain up to a prespecified number of degenerate characters from the IUPAC alphabet. To this end, information on multiple children of a node in the suffix tree is aggregated. For example, if a degenerate character R (representing A or G) is introduced, child nodes "A" and "G" are explored and motif occurrences in both branches are aggregated.

For each length-*k* motif, the suffix tree reveals in which promoter sequences (and hence: species) an instance of that motif appears (Figure S4-10-b). The degree of conservation of the motif within the gene family is expressed by the Branch Length Score (BLS). The BLS takes a value between 0 (motif occurs only in a single species) and 1 (motif occurs in all species of the gene family). It is calculated by finding, in the phylogenetic species tree, the minimum spanning tree that connects the relevant subset of species and summing the weights of the horizontal branches in that tree. If the motif appears in multiple paralogs of the same species, the branch length to that species is only accounted for once. If a gene family does not contain a representative ortholog in each of the considered species, we delete the branches in the species tree corresponding to the missing species and rescale the branch lengths so that the sum of the weights on the branches of the tree again amounts to 1. Within the tree of a gene family, a higher BLS corresponds to a motif that appears in relatively more species and/or more distantly related species. A motif for which the BLS exceeds a predefined BLS threshold in a gene family is said to be *conserved* within that gene family. We use multiple BLS thresholds (i.e., 0.07, 0.13, 0.41, 0.54, 0.65, 0.75, 0.85 and 0.95) (Figure S4-10-b) to also allow the discovery of motifs that are conserved only in a subset of the species and might have a relatively lower BLS score.

Subsequently, we calculate the *recurrence score* (referred to as the 'confidence score' in the original publication [32]) for every motif and BLS threshold (Figure S4-10-d). The recurrence score is calculated as 1 – (*expected recurrence* of the motif at the considered BLS threshold) / (*recurrence* of the observed motif at the considered BLS threshold). Here, the recurrence of a motif equals the total number of gene families in which the motif is conserved. The expected recurrence of a motif is estimated as the median number of gene families in which

motifs with the same nucleotide content as the observed motif are conserved. A recurrence score of 0.90 for a given BLS threshold means that the observed motif is conserved (at that BLS threshold) in ten times as many gene families than expected.

The recurrence and expected recurrence of a motif are computed as follows. As described before, we use the suffix tree to exhaustively enumerate all motifs and we create a binary matrix indicating for each motif (rows of the matrix) whether it meets a certain BLS threshold (columns of the matrix). This procedure is repeated for all gene families (Figure S4-10-c). The matrices of all gene families are aggregated into a single matrix, where each matrix element now corresponds to the recurrence of a certain motif at a certain BLS threshold. To calculate the expected recurrence of a motif, we group all motifs with the same nucleotide content and refer to these as a nucleotide content group. We extend this list by adding a count of 0 for all motif permutations that are not present. Next, per nucleotide content group and per BLS threshold, the expected recurrence is computed as the median value among all motifs in the nucleotide content group (Figure S4-10-c). This median value can be 0.

The software can be obtained at https://bitbucket.org/dries_decap/bls-speller-spark. BLSSpeller was used to discover 8-bp conserved motifs with at most 3 degenerate sites in the promoter sequences of 16 monocot species. Considering the fact that the median length of experimentally identified cis-elements is 8-bp [58], we opted for motifs of length 8, as this gives the best balance between detecting biologically relevant motifs while maintaining computational tractability. To obtain a set of reliable motifs we selected motifs with a recurrence score of at least 0.9 in at least one of the considered BLS thresholds. This resulted in 1295 different motifs with 2,320,402 instances.

4.4.3 Identifying motif instances in *Zea mays*

To identify the instances of these motifs in maize, we consider for each motif the lowest BLS threshold at which the corresponding recurrence score is 0.9 or higher (Figure S4-10-d). We then identify all gene families in which this motif has a BLS that meets the said threshold. The motif instances (genomic locations) in maize are selected. As such, we obtained 1292 motifs with at least one instance in maize.

Because BLSSpeller is an exhaustive approach, it will output highly similar motifs that only differ from each other in a limited number of (degenerate) characters or that have largely overlapping instances. To reduce the level of redundancy, we removed motifs with fewer than 20 or more than 3000 instances (25 motifs were removed) as motifs with instances in too few or too many unique genes are either hard to validate or are likely correspond to general TF binding sites. To remove redundancy among the remaining motifs, they were compared in a pairwise manner. The criteria to decide whether two motifs were sufficiently similar to be considered redundant are i) the pairwise alignment distance between the two motifs and ii) the degree of overlap among the genes that contain instances of these motifs.

To identify the pairwise alignment distance, the motifs were sorted by their number of degenerate sites (higher to lower) and subjected to an all to all mutual pairwise alignment. The

distance between motifs is defined using the following cost model in the pairwise alignment: a match score of 0, a mismatch penalty of 1 and an indel penalty of 1.

The pairwise degree of overlap in genes that contain an instance of two considered motifs is determined as follows:

$$\text{Degree of overlap in genes} = \frac{G_i \cap \mathrm{G_j}}{\min(N_i, N_j)}$$

where the numerator indicates the number of genes in maize containing an instance of both motifs i and j, and the denominator indicates the minimum of the number of genes in maize that contain instances of motifs i and j, respectively.

Motif "i" is considered redundant if its alignment distance to motif "j" is less than 5 and the degree of overlap in genes is larger than 0.5. This indicates that the smallest gene set of a motif is contained for at least 50% in the larger gene set of another motif. The redundant motifs are removed from the sorted list of motifs in a greedy manner, hereby keeping for each similar set of motifs the most degenerate one (as this one contains most if not all the instances of the other motif). This resulted in 61 non-redundant motifs (referred to as non-redundant motifs hereafter, Appendix 1).

For downstream analyses that were performed at the level of the motif instances (overlap of motif instances with chromatin regions and ChIP-seq peaks or when assessing the degree of polymorphism), overlapping instances (minimum overlap in bp = 5) were removed. This is done in order to prevent that the same instance would contribute multiple times to the analysis. Overlapping instances were removed as follows: if two motifs contain overlapping instances, we retain the instance for the motif that covers the highest number of genes and remove the instances of the other motifs that overlapped with the selected motif. This resulted in 50,354 non-overlapping instances for 61 non-redundant motifs.

4.4.4 Generating random motifs and random instances in *Zea mays*

Here we generated a set of random motifs and their instances to be used in the comparative downstream motif validation analysis. First, the GC content was inferred from the instances of 61 non-redundant motifs in maize. Then, 1000 random 8-bp *k*-mers were generated using IUPAC DNA codes with the same GC content with at most 3 degenerate sites. Simulating motifs like this ensures that the random motifs share properties comparable to those of our predicted motifs: i.e., having the same GC content and at most three degenerate sites.

To identify random maize motif instances, we identified all gene families in which a given random motif occurs. Like for the predicted motifs (see Datasets used for motif detection**)**, we only considered gene families that have sequences of at least 3 different species (at least sequences of three species should be included). From the maize sequences, the random motif instance was extracted irrespective of the BLS score of that family. Subsequently, on the obtained motif instances, the same redundancy filtering criteria mentioned above were applied: random motifs with instances in more than 3000 or fewer than 20 different maize genes were removed, and overlapping instances were filtered as we did for the predicted motifs. Of the remaining random motifs, 61 random motifs with a pairwise distance of higher than 4 to any of the conserved motifs were randomly selected. In this way, the random motifs underwent the

same filtering criteria as the predicted motifs while being sufficiently different from any of the predicted motifs.

4.4.5 Comparing motifs with the Arabidopsis motif compendium

The Arabidopsis DAP-seq motif compendium was obtained from O'Malley [36] and predicted consensus motifs were compared using the Tomtom tool from the MEME suite [59]. An adjusted p-value smaller than 0.05 was used to determine significant similarity between the compared motifs.

4.4.6 Tissue specificity (Tau Index)

To calculate the tissue specificity index of a gene, we used expression datasets profiled in maize from 8 different tissues [39]. RNA-seq raw sequencing reads were downloaded from SRA [60]. Read quality was assessed using FastQC [61]. Remaining adapters, low-quality bases (Phred quality score < 20) and reads shorter than 50 bp were filtered using Trimmomatic [62]. The cleaned reads were aligned to the *Zea mays* B73 AGPv4 genome assembly [63] using the STAR software [64]. FeatureCount [65] was used to quantify expression as count values using the annotation (Zea_mays.B73_RefGen_v4.49.gtf) provided by EnsemblPlants. Only uniquely mapped reads were considered for expression quantification. The count data were normalized for library size and gene length differences using TMM normalization implemented in the edgeR package [66]. Tissue-specific expression of genes was assessed by the Tau index [40]. The Tau index was calculated as follows:

$$Tau = \frac{\sum_{i=1}^{n}(1-\overline{x}_i)}{n-1}$$

$$\overline{x}_i = \frac{x_i}{max_{1\le i\le n}(x_i)}$$

where x_i is the expression of a gene in tissue *i* and *n* is the number of tissues in which expression was profiled. The higher the Tau index, the more tissue-specific the expression of the gene is.

4.4.7 GO enrichment

Functional gene annotation for *Zea mays* B73 AGPv4 was downloaded from www.maizegdb.org. GO enrichment, limited to the "Biological Process" ontology, was performed using the topGO package [67]. P-values were adjusted by the BH method using the p.adjust function in R. From the result, the GO terms with more than 1500 annotated genes or less than 3 significant genes, and adjusted p-value > 0.01 were removed. After filtering redundant GO terms, the five most significantly enriched GO terms for each cluster are visualized using the pheatamap function in R [68].

4.4.8 Coexpression analysis

To perform coexpression analysis, we used the datasets that have been collected and processed by [37] (Table S4-3). In this study, a comprehensive compilation of RNAseq data from studies in maize with more than 20 samples spanning different genotypes, tissues, and conditions was made. Lowly expressed (FPKM < 1 in more than 90% of samples), low variance genes, tRNA, and ribosomal genes were filtered out from each dataset. For each dataset, the rank matrix was derived from the gene-gene Pearson correlation matrix calculated on the log2-transformed FPKM values. The rank matrix was transformed to the Mutual Rank (MR) matrix using the following formula:

$$MR(AB) = \sqrt{rank_{(A\rightarrow B)} * rank_{(B\rightarrow A)}}$$

where rank (A →B) is the rank of the correlation of gene B with gene A as compared to its correlation with all other genes [69]. Then the MR matrix was converted to the probability matrix (P) using:

$$P_{(AB)} = e^{-(MR_{(AB)}-1)/10}$$

The P matrix was converted to the coexpression graph after removing p-values smaller than 0.05. The connectivity score between nodes (genes) in each network was calculated using the Katz index [70]. The Katz index assesses connectivity of two nodes in a graph by exploiting the neighborhood of the nodes. When calculating the connectivity between two nodes, it considers all paths in the graph that connect the two nodes, but favors shorter paths by assigning them a higher weight. Here we considered all the paths connecting two nodes with a maximum length of three, weighted according to their path length to calculate the similarity between any two nodes in each coexpression network. The Katz index can be calculated and normalized for the *degree* of the connecting genes (normalized Katz index) as follows:

$$Katz\_index = \alpha A + \alpha^2 A + \alpha^3 A$$

$$nomalized\_Katz_{ij} = \frac{Katz\_index_{ij}}{\sqrt{d_i * d_j}}$$

Where the "A" is the adjacency matrix, alpha is a parameter to weight the order of the neighbors (set to 0.3), and "d" indicates the degree for a gene "i" and "j". The normalized Katz index ranges between 0 and 1 (highly skewed toward zero). A high value indicates high connectivity. For the normalized Katz index ideally a distribution with an exponential decay is expected for a random gene set. However, distributions of normalized Katz-indexes for gene sets sharing a random motif are indicative of some residual connectivity in the coexpression networks which is to be expected, given the large connectivity in the coexpression network (Figure S4-11). However, we observe that the majority of pairwise connectivity scores for random gene sets fall below 0.05. Therefore, we considered 0.05 as the threshold to distinguish between random and non-random connectivity (see below).

To assess whether genes that share the same motif in maize were more connected in each of the coexpression networks than expected by chance, we first extracted the sets of genes that shared the same motif. A gene set is here defined as a set of genes that share an instance of the same motif. Subsequently, all pairwise Katz scores between genes in a set were assessed and the

number of times the Katz score was above a predefined threshold (0.05) was counted, where 0.05 was defined based on the distribution of the Katz score of random gene sets. The obtained count was referred to as the gene set score. For each gene set, 1000 random gene sets of the same size were obtained by randomly selecting genes from the coexpression network. For each of the random gene sets, the gene set score was calculated as described above. The distribution of these gene set scores obtained for the 1000 random sets was used to construct the null distribution, which followed a normal distribution after log2 transformation. The parameters of this null distribution were estimated using Maximum likelihood. The ccdf (complementary cumulative distribution function) was used to obtain the p-value of the observed gene set score given null distribution. A small p-value indicates that connectivity between the genes of the gene set corresponding to a certain predicted motif is significantly higher than expected by chance. This analysis was performed for each coexpression network separately.

4.4.9 Overlap with active chromatin regions and degree of polymorphism

The ACRs, ChIP-seq peaks, and SNPs were downloaded from the following sources: 165,913 non-overlapping 500 bp accessible chromatin regions, integrated over more than 50000 single cells were obtained from [7]; 144,890 non-overlapping TF binding integrated from a ChIP-seq study covering 104 TFs obtained from [5]; SNPs for maize were obtained from Imputed HapMap 3.2.1 (uplifted to B73 AGPv4) [71]; For the active chromatin and ChIP-seq regions (ACRs and ChIP-seq), sequences were selected that covered up to 1kb up and downstream of the center of the experimental ACR or ChIP seq peak to include the flanking regions. We subsequently assessed which fraction of the total instances of the 61 motifs overlapped with the sequences contained in a window centered at a prespecified position up or downstream of the active chromatin and ChIP-seq regions. For the plots of the individual motifs (Figure S4-5-9) we performed the same analysis, but only focusing on the instances of one particular motif.
To assess the fraction of SNPs located at the location of the motif instances and in their flanking regions, we counted the number of SNPs in a sliding window of 8-bp that starts at the location of the motif instance and that moves up to 1kb up and downstream of motif instance locations and divided this by the total number of SNPs occurring in the entire sequences considered for the above-mentioned analysis. The GenomicRange package [72] was used to count the overlap between the windows that contained SNPs, ChIP-seq or ACR regions and the windows containing motifs and the result was visualized using ggplot2 [73]. The same analyses were performed for both predicted and random motifs.

Author contributions

Supervision, K.M. and J.F.; Method development D.D., J.F. Formal analyses and visualization, R.S.R.; Validation, application, R.S.R.; K.M; Writing - Original Draft, R.S.R, D.D, J.F and K.M; Writing – Review and Editing, R.S.R, D.D, J.F. and K.M; Funding Acquisition, K.M. and J.F.

References

1. Miculan M, Nelissen H, Ben MH, Marroni F, Inze D, Pe ME, Dell'Acqua M: **A forward genetics approach integrating genome-wide association study and expression quantitative trait locus mapping to dissect leaf development in maize (Zea mays).** *The Plant journal: for cell and molecular biology* 2021, **107:**1056-1071.
2. Wallace JG, Bradbury PJ, Zhang N, Gibon Y, Stitt M, Buckler ES: **Association mapping across numerous traits reveals patterns of functional variation in maize.** *PLoS genetics* 2014, **10:**e1004845.
3. Cherry TJ, Yang MG, Harmin DA, Tao P, Timms AE, Bauwens M, Allikmets R, Jones EM, Chen R, De Baere E: **Mapping the cis-regulatory architecture of the human retina reveals noncoding genetic variation in disease.** *Proceedings of the National Academy of Sciences* 2020, **117:**9001-9012.
4. Salvi S, Sponza G, Morgante M, Tomes D, Niu X, Fengler KA, Meeley R, Ananiev EV, Svitashev S, Bruggemann E: **Conserved noncoding genomic sequences associated with a flowering-time quantitative trait locus in maize.** *Proceedings of the National Academy of Sciences* 2007, **104:**11376-11381.
5. Tu X, Mejía-Guerra MK, Franco JAV, Tzeng D, Chu P-Y, Shen W, Wei Y, Dai X, Li P, Buckler ES: **Reconstructing the maize leaf regulatory network using ChIP-seq data of 104 transcription factors.** *Nature communications* 2020, **11:**1-13.
6. Eveland AL, Goldshmidt A, Pautler M, Morohashi K, Liseron-Monfils C, Lewis MW, Kumari S, Hiraga S, Yang F, Unger-Wallace E: **Regulatory modules controlling maize inflorescence architecture.** *Genome research* 2014, **24:**431-443.
7. Marand AP, Chen Z, Gallavotti A, Schmitz RJ: **A cis-regulatory atlas in maize at single-cell resolution.** *Cell* 2021, **184:**3041-3055. e3021.
8. Ricci WA, Lu Z, Ji L, Marand AP, Ethridge CL, Murphy NG, Noshay JM, Galli M, Mejía-Guerra MK, Colomé-Tatché M: **Widespread long-range cis-regulatory elements in the maize genome.** *Nature plants* 2019, **5:**1237-1249.
9. Bartlett A, O'Malley RC, Huang S-sC, Galli M, Nery JR, Gallavotti A, Ecker JR: **Mapping genome-wide transcription-factor binding sites using DAP-seq.** *Nature protocols* 2017, **12:**1659-1672.
10. Gerstein MB, Kundaje A, Hariharan M, Landt SG, Yan K-K, Cheng C, Mu XJ, Khurana E, Rozowsky J, Alexander R: **Architecture of the human regulatory network derived from ENCODE data.** *Nature* 2012, **489:**91-100.
11. Johnson DS, Mortazavi A, Myers RM, Wold B: **Genome-wide mapping of in vivo protein-DNA interactions.** *Science* 2007, **316:**1497-1502.
12. Kheradpour P, Kellis M: **Systematic discovery and characterization of regulatory motifs in ENCODE TF binding experiments.** *Nucleic acids research* 2014, **42:**2976-2987.
13. Bolduc N, Yilmaz A, Mejia-Guerra MK, Morohashi K, O'Connor D, Grotewold E, Hake S: **Unraveling the KNOTTED1 regulatory network in maize meristems.** *Genes & development* 2012, **26:**1685-1690.
14. Li C, Yue Y, Chen H, Qi W, Song R: **The ZmbZIP22 transcription factor regulates 27-kD γ-zein gene transcription during maize endosperm development.** *The Plant Cell* 2018, **30:**2402-2424.
15. Pautler M, Eveland AL, LaRue T, Yang F, Weeks R, Lunde C, Je BI, Meeley R, Komatsu M, Vollbrecht E: **FASCIATED EAR4 encodes a bZIP transcription factor that regulates shoot meristem size in maize.** *The Plant Cell* 2015, **27:**104-120.
16. Yang H, Liu X, Xin M, Du J, Hu Z, Peng H, Rossi V, Sun Q, Ni Z, Yao Y: **Genome-wide mapping of targets of maize histone deacetylase HDA101 reveals its function and regulatory mechanism during seed development.** *The Plant Cell* 2016, **28:**629-645.
17. Yocca AE, Edger PP: **Current status and future perspectives on the evolution of cis-regulatory elements in plants.** *Current opinion in plant biology* 2022, **65:**102139.

18. Monsieurs P, Thijs G, Fadda AA, De Keersmaecker SC, Vanderleyden J, De Moor B, Marchal K: **More robust detection of motifs in coexpressed genes by using phylogenetic information.** *BMC bioinformatics* 2006, **7:**1-15.
19. Blanchette M, Tompa M: **FootPrinter: a program designed for phylogenetic footprinting.** *Nucleic acids research* 2003, **31:**3840-3842.
20. Cliften P, Sudarsanam P, Desikan A, Fulton L, Fulton B, Majors J, Waterston R, Cohen BA, Johnston M: **Finding functional features in Saccharomyces genomes by phylogenetic footprinting.** *science* 2003, **301:**71-76.
21. Blanchette M, Tompa M: **Discovery of regulatory elements by a computational method for phylogenetic footprinting.** *Genome research* 2002, **12:**739-748.
22. Wei H, Liu J, Guo Q, Pan L, Chai S, Cheng Y, Ruan M, Ye Q, Wang R, Yao Z: **Genomic organization and comparative phylogenic analysis of NBS-LRR resistance gene family in Solanum pimpinellifolium and Arabidopsis thaliana.** *Evolutionary Bioinformatics* 2020, **16:**1176934320911055.
23. Hou L, Xie J, Wu Y, Wang J, Duan A, Ao Y, Liu X, Yu X, Yan H, Perreault J: **Identification of 11 candidate structured noncoding RNA motifs in humans by comparative genomics.** *BMC genomics* 2021, **22:**1-14.
24. Taboada-Castro H, Castro-Mondragón JA, Aguilar-Vera A, Hernández-Álvarez AJ, van Helden J, Encarnación-Guevara S: **RhizoBindingSites, a Database of DNA-Binding Motifs in Nitrogen-Fixing Bacteria Inferred Using a Footprint Discovery Approach.** *Frontiers in Microbiology* 2020, **11**.
25. Rombauts S, Florquin K, Lescot M, Marchal K, Rouzé P, Van de Peer Y: **Computational approaches to identify promoters and cis-regulatory elements in plant genomes.** *Plant physiology* 2003, **132:**1162-1176.
26. Blanchette M: **Computation and analysis of genomic multi-sequence alignments.** *Annu Rev Genomics Hum Genet* 2007, **8:**193-213.
27. Aerts S: **Computational strategies for the genome-wide identification of cis-regulatory elements and transcriptional targets.** *Current topics in developmental biology* 2012, **98:**121-145.
28. Fickett JW, Wasserman WW: **Discovery and modeling of transcriptional regulatory regions.** *Current Opinion in Biotechnology* 2000, **11:**19-24.
29. De Witte D, Van de Velde J, Decap D, Van Bel M, Audenaert P, Demeester P, Dhoedt B, Vandepoele K, Fostier J: **BLSSpeller: exhaustive comparative discovery of conserved cis-regulatory elements.** *Bioinformatics* 2015, **31:**3758-3766.
30. Carmack CS, McCue LA, Newberg LA, Lawrence CE: **PhyloScan: identification of transcription factor binding sites using cross-species evidence.** *Algorithms for Molecular Biology* 2007, **2:**1-17.
31. Van Bel M, Diels T, Vancaester E, Kreft L, Botzki A, Van de Peer Y, Coppens F, Vandepoele K: **PLAZA 4.0: an integrative resource for functional, evolutionary and comparative plant genomics.** *Nucleic acids research* 2018, **46:**D1190-D1196.
32. Stark A, Lin MF, Kheradpour P, Pedersen JS, Parts L, Carlson JW, Crosby MA, Rasmussen MD, Roy S, Deoras AN: **Discovery of functional elements in 12 Drosophila genomes using evolutionary signatures.** *Nature* 2007, **450:**219-232.
33. Berens ML, Berry HM, Mine A, Argueso CT, Tsuda K: **Evolution of hormone signaling networks in plant defense.** *Annual Review of Phytopathology* 2017, **55:**401-425.
34. Katsir L, Chung HS, Koo AJ, Howe GA: **Jasmonate signaling: a conserved mechanism of hormone sensing.** *Current opinion in plant biology* 2008, **11:**428-435.
35. Chater C, Kamisugi Y, Movahedi M, Fleming A, Cuming AC, Gray JE, Beerling DJ: **Regulatory mechanism controlling stomatal behavior conserved across 400 million years of land plant evolution.** *Current Biology* 2011, **21:**1025-1029.
36. O'Malley RC, Huang S-sC, Song L, Lewsey MG, Bartlett A, Nery JR, Galli M, Gallavotti A, Ecker JR: **Cistrome and epicistrome features shape the regulatory DNA landscape.** *Cell* 2016, **165:**1280-1292.

37. Zhou P, Li Z, Magnusson E, Gomez Cano F, Crisp PA, Noshay JM, Grotewold E, Hirsch CN, Briggs SP, Springer NM: **Meta gene regulatory networks in maize highlight functionally relevant regulatory interactions.** *The Plant Cell* 2020, **32:**1377-1396.
38. Kaufmann K, Pajoro A, Angenent GC: **Regulation of transcription in plants: mechanisms controlling developmental switches.** *Nature Reviews Genetics* 2010, **11:**830-842.
39. Kremling KA, Chen S-Y, Su M-H, Lepak NK, Romay MC, Swarts KL, Lu F, Lorant A, Bradbury PJ, Buckler ES: **Dysregulation of expression correlates with rare-allele burden and fitness loss in maize.** *Nature* 2018, **555:**520-523.
40. Yanai I, Benjamin H, Shmoish M, Chalifa-Caspi V, Shklar M, Ophir R, Bar-Even A, Horn-Saban S, Safran M, Domany E: **Genome-wide midrange transcription profiles reveal expression level relationships in human tissue specification.** *Bioinformatics* 2005, **21:**650-659.
41. Vanneste S, Coppens F, Lee E, Donner TJ, Xie Z, Van Isterdael G, Dhondt S, De Winter F, De Rybel B, Vuylsteke M: **Developmental regulation of CYCA2s contributes to tissue-specific proliferation in Arabidopsis.** *The EMBO journal* 2011, **30:**3430-3441.
42. Siefers N, Dang KK, Kumimoto RW, Bynum IV WE, Tayrose G, Holt III BF: **Tissue-specific expression patterns of Arabidopsis NF-Y transcription factors suggest potential for extensive combinatorial complexity.** *Plant physiology* 2009, **149:**625-641.
43. Quandt K, Frech K, Karas H, Wingender E, Werner T: **Matlnd and Matlnspector: new fast and versatile tools for detection of consensus matches in nucleotide sequence data.** *Nucleic acids research* 1995, **23:**4878-4884.
44. Worsley-Hunt R, Bernard V, Wasserman WW: **Identification of cis-regulatory sequence variations in individual genome sequences.** *Genome medicine* 2011, **3:**1-14.
45. Ma S, Shah S, Bohnert HJ, Snyder M, Dinesh-Kumar SP: **Incorporating motif analysis into gene co-expression networks reveals novel modular expression pattern and new signaling pathways.** *PLoS genetics* 2013, **9:**e1003840.
46. Tabach Y, Brosh R, Buganim Y, Reiner A, Zuk O, Yitzhaky A, Koudritsky M, Rotter V, Domany E: **Wide-scale analysis of human functional transcription factor binding reveals a strong bias towards the transcription start site.** *PloS one* 2007, **2:**e807.
47. Clark RM, Schweikert G, Toomajian C, Ossowski S, Zeller G, Shinn P, Warthmann N, Hu TT, Fu G, Hinds DA: **Common sequence polymorphisms shaping genetic diversity in Arabidopsis thaliana.** *science* 2007, **317:**338-342.
48. Tatarinova TV, Chekalin E, Nikolsky Y, Bruskin S, Chebotarov D, McNally KL, Alexandrov N: **Nucleotide diversity analysis highlights functionally important genomic regions.** *Scientific reports* 2016, **6:**1-12.
49. Procházková Schrumpfová P, Vychodilová I, Dvořáčková M, Majerská J, Dokládal L, Schořová Š, Fajkus J: **Telomere repeat binding proteins are functional components of A rabidopsis telomeres and interact with telomerase.** *The Plant Journal* 2014, **77:**770-781.
50. Warren AS, Anandakrishnan R, Zhang L: **Functional bias in molecular evolution rate of Arabidopsis thaliana.** *BMC evolutionary biology* 2010, **10:**1-10.
51. Andolfatto P, Wong KM, Bachtrog D: **Effective population size and the efficacy of selection on the X chromosomes of two closely related Drosophila species.** *Genome biology and evolution* 2011, **3:**114-128.
52. Lee C-R, Mitchell-Olds T: **Environmental adaptation contributes to gene polymorphism across the Arabidopsis thaliana genome.** *Molecular biology and evolution* 2012, **29:**3721-3728.
53. Zhou J, Lemos B, Dopman EB, Hartl DL: **Copy-number variation: the balance between gene dosage and expression in Drosophila melanogaster.** *Genome Biology and Evolution* 2011, **3:**1014-1024.
54. Shi T, Rahmani RS, Gugger PF, Wang M, Li H, Zhang Y, Li Z, Wang Q, Van de Peer Y, Marchal K: **Distinct expression and methylation patterns for genes with different fates following a single whole-genome duplication in flowering plants.** *Molecular biology and evolution* 2020, **37:**2394-2413.
55. Schuster-Böckler B, Conrad D, Bateman A: **Dosage sensitivity shapes the evolution of copy-number varied regions.** *PloS one* 2010, **5:**e9474.

56. Morton T, Petricka J, Corcoran DL, Li S, Winter CM, Carda A, Benfey PN, Ohler U, Megraw M: **Paired-end analysis of transcription start sites in Arabidopsis reveals plant-specific promoter signatures.** *The Plant Cell* 2014, **26:**2746-2760.
57. Chen N: **Using Repeat Masker to identify repetitive elements in genomic sequences.** *Current protocols in bioinformatics* 2004, **5:**4.10. 11-14.10. 14.
58. Reineke AR, Bornberg-Bauer E, Gu J: **Evolutionary divergence and limits of conserved non-coding sequence detection in plant genomes.** *Nucleic acids research* 2011, **39:**6029-6043.
59. Gupta S, Stamatoyannopoulos JA, Bailey TL, Noble WS: **Quantifying similarity between motifs.** *Genome biology* 2007, **8:**1-9.
60. Leinonen R, Sugawara H, Shumway M, Collaboration INSD: **The sequence read archive.** *Nucleic acids research* 2010, **39:**D19-D21.
61. Andrews S: **FastQC: a quality control tool for high throughput sequence data. 2010.** 2017.
62. Bolger AM, Lohse M, Usadel B: **Trimmomatic: a flexible trimmer for Illumina sequence data.** *Bioinformatics* 2014, **30:**2114-2120.
63. Jiao Y, Peluso P, Shi J, Liang T, Stitzer MC, Wang B, Campbell MS, Stein JC, Wei X, Chin C-S: **Improved maize reference genome with single-molecule technologies.** *Nature* 2017, **546:**524-527.
64. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR: **STAR: ultrafast universal RNA-seq aligner.** *Bioinformatics* 2013, **29:**15-21.
65. Liao Y, Smyth GK, Shi W: **featureCounts: an efficient general purpose program for assigning sequence reads to genomic features.** *Bioinformatics* 2014, **30:**923-930.
66. Robinson MD, McCarthy DJ, Smyth GK: **edgeR: a Bioconductor package for differential expression analysis of digital gene expression data.** *Bioinformatics* 2010, **26:**139-140.
67. Alexa A, Rahnenfuhrer J: **topGO: enrichment analysis for gene ontology.** *R package version* 2010, **2:**2010.
68. Team RC: **R: A language and environment for statistical computing.** 2013.
69. Obayashi T, Hayashi S, Saeki M, Ohta H, Kinoshita K: **ATTED-II provides coexpressed gene networks for Arabidopsis.** *Nucleic acids research* 2009, **37:**D987-D991.
70. Bonchi F, Esfandiar P, Gleich DF, Greif C, Lakshmanan LV: **Fast matrix computations for pairwise and columnwise commute times and Katz scores.** *Internet Mathematics* 2012, **8:**73-112.
71. Bukowski R, Guo X, Lu Y, Zou C, He B, Rong Z, Wang B, Xu D, Yang B, Xie C: **Construction of the third-generation Zea mays haplotype map.** *Gigascience* 2018, **7:**gix134.
72. Lawrence M, Huber W, Pages H, Aboyoun P, Carlson M, Gentleman R, Morgan MT, Carey VJ: **Software for computing and annotating genomic ranges.** *PLoS computational biology* 2013, **9:**e1003118.
73. Wickham H: **ggplot2.** *Wiley Interdisciplinary Reviews: Computational Statistics* 2011, **3:**180-185.

5 Chapter 5

"One of the main concerns of philosophy is to warn people against jumping to conclusions"
- Jostein Gaarder

5.1 General Conclusions and Future Perspectives

Advances in omics technologies (i.e., (epi)genomics, transcriptomics, proteomics, and metabolomics) have shifted the main bottleneck in biology from data collection to data analysis and interpretation. New biological questions can be answered by mining thousands of genomic datasets that span different technological platforms, genotypes, tissues, and developmental stages. However, due to their inherent data differences, analyzing and integrating multiple-omics datasets will remain an ongoing challenge for years to come. Continuous efforts to develop new methods in order to combine available datasets with novel datasets are required. The ultimate goal of integrative studies is to associate the genetic code to phenotypes such as gene expression, or gene function in general, using information contained in different molecular layers. In this dissertation, we have contributed in the following way to reach this goal:

- **Associate gene function with evolutionary patterns using multi-omics data**: In the first study, we collaborated on a project that aimed at systematically exploring the association between functional divergence and copy number state using different omics datasets. We performed integrative methylation and expression analyses combined with relevant genomic analyses to study how functional constraints and dosage balance determine the fate of genes after a single round of WGD. This helps uncovering genetic mechanisms that create functional divergence and adaptation after a WGD.
- **Inferring gene function by integrating expression data with prior information**: In the second collaborative study, an integrative network-based approach was used to shed more light on the BR signaling pathway and its crosstalk with other hormone signaling pathways. To this end, available information at different molecular levels (protein-protein, regulatory, and metabolic interactions) was integrated into an interaction network in Arabidopsis. Subsequently, we mapped the gene list resulting from the

expression analysis on the interaction network by applying a network-based method. This allowed showing how biological entities are interacting in the BR related pathways to drive the observed mutant phenotype. Furthermore, a new function for less characterized genes was proposed to be investigated by future studies.

- **Identifying functional regulatory elements using a comparative approach and omics datasets**: In the final part, we turned our attention to integrating omics datasets to the cis-regulatory element identification problem. In a comparative genomic setting using sequence information of related species as input to BLSSpeller, the upstream sequence of genes was explored to identify candidate cis-regulatory elements in maize. We took advantage of the availability of large-scale omics datasets to validate our predictions. In addition, a multi-step filtering approach was applied to classify the predictions, based on the genomic and functional behavior of their corresponding genomic regions and nearby genes. The predicted motifs will be a valuable resource of novel cis-regulatory elements that can complement results from functional genomics studies. In particular, the improved version of the BLSSpeller and our integrative guidelines can be applied to extend gene regulatory networks in other plant species.

5.2 Pitfalls of integrative approaches

The advantages and benefits of using integrative methods have been described in the introduction and the presented papers. In the following section, I highlight based on my personal experience some limitations and room for future improvement of the use of integrative approaches in plant system biology.

5.2.1 Lack of diverse and relevant prior information for non-model species

A critical step in network-based approaches is conceptualization and capturing relevant prior knowledge by reviewing the available literature and searching for omics-based resources. The most important factors to account for, when including a dataset to build an omics derived prior network, are whether relevant sample types are available, whether biological and/or technical replicates are available that proof data reproducibility. As a result, only a subset of the available datasets passes the minimal quality control criteria one envisages in an integrative network inference effort. Moreover, because of financial constraints, omics datasets (with both good quality and large sample size) for non-model crops species are much sparser than for model plants. For example, comprehensive datasets from (epi)genomic, transcriptomic, proteomic, and metabolomic profiling experiments are available for Arabidopsis, but lack for non-model plants. This lack of sufficient omics and meta-data for non-model plant species is one of the main barriers to use integrative, and then particularly network-based approaches in these organisms. In addition, often the most relevant prior information for a question at hand is not available. For instance, the most relevant prior information to explain the underlying regulatory mechanisms behind the differentially expressed genes is the protein abundance of TFs under

relevant conditions. When available, these data often come at a limited time resolution (limited number of sampled time points), and often with a poor correlation ($r \sim 0.3$) between transcript-protein level [1-3], As such, interactions between signaling proteins and targets are difficult to infer from transcriptome and proteome profiling data, and one has to rely on experimentally identified protein interaction and/or regulatory interactions. Such data are available for well-defined model organisms only and limited to profiling a restricted set of molecular entities under limited conditions. Furthermore, with the reduced cost of sequencing, reference genomes are continuously improved, but much less attention is paid to the updating of the associated molecular interaction networks, as this depends on much more tedious experimental approaches.

5.2.2 Limitation of biological networks for inferences

Ideally, a network that represents biological truth would contain all edges relevant to a particular condition of interest, but in reality, the interaction networks inferred from omics data contain both false positives, condition irrelevant, and missing edges. Therefore, the application of networks in genomics is more challenging than in other fields such as social or engineering fields. The most straightforward application of networks can be assigning labels to unlabeled nodes based on knowledge of other nodes in the network (guilt by association). For example, Facebook can correctly predict and suggest friends for a new user, based on connections in the networks of friends. However, in biological networks uncertainty on the edges (mostly predicted or inferred from model organisms) and an overabundance of unlabeled nodes make this inference challenging. For example, one may try to infer the function of a gene without annotation based on information on its neighbors. However, if only for a small portion of genes the function has been established experimentally, the predicted outcomes from a guilt by association approach might be unreliable.

The use of graph-based methods for interpreting in-house generated gene lists from omics data depends even more on the reliability of the underlying interaction scaffold used to drive the analysis. Because most graph-based algorithms perform poorly on over-connected networks, one needs to find, when designing a prior interaction scaffold, a good trade-off between being comprehensive, yet not including too many false positive interactions. Also, when considering interactions derived from different molecular omics data, one needs to ensure that the inferred edges are used in an independent way, e.g., interactions derived from one layer might have been used to adjust and correct other layers [4]. For instance, GO annotation has been widely used to correct other biological networks (e.g., coexpression networks) leading to an artificially high correlation between biological networks [5].

Propagation-based methods [6, 7] are popular graph-based analysis methods that can be used to perform a topology-based weighting of a graph or to prioritize genes. The similarity scores between any two nodes is used to assess the relative impact of one node on the other node in the graph [8]. To measure the similarity, it is desired to consider the number of paths connecting two nodes and the length of the path without getting lost in the large space of the graph.

Therefore, often local propagation is performed (e.g., random walk with restart). The advantage of such an approach is that the impact of hubs in the network is automatically downweighed as hubs might affect many neighboring nodes but only with a relatively low impact [7]. However, when dealing with a heterogeneous graph, the use of propagation-based methods might not always capture correctly the biological flow of information. For instance, the connectivity in a protein interaction (PI) network is different from the connectivity in a TF-target network. PI often comes in connected components (particularly when measured with pull down experiments), whereas TF-target interactions show a directed and sparser connectivity. However, the biological signal through this TF-target relation is not weaker than the signal in the PI network, but might not be captured by a propagation-based method. This might explain why applying network-propagation methods to prioritize candidate genes for stress response on plant biological data (Arabidopsis and maize) did not provide satisfying results.

5.2.3 Statistical validation of the network-based approaches

Classical statistical or machine-learning methods have their own underlying assumptions that need to be verified. Network-based methods are no exception as they are based on statistical or machine-learning methods. However, a large number of variables or sample space make the validation of these assumptions and correction for multiple testing almost impossible [9, 10]. As a result, network-based approaches are prone to high rates of false positives (spurious predictions). In addition, the complex integrative approaches are often not well understood by users and the output resulting from these approaches is not sufficiently critically evaluated by the users. This may result in misinterpretation and in the generation of false negative and positive predictions, which significantly impact reproducibility, and adversely affect scientific progress in the field. On top of that, the lack of well-defined ‘gold standard’ pipelines for integrative approaches allows investigators to play with parameters and models (subjective model and parameter selection) until non-significant results become significant (p-hacking) [10]. More interpretable algorithms with a solid statistical basis and a gold-standard guideline are desirable to extract unbiased results from network-based approaches and multiple testing issues should ideally be addressed. However, the lack of gold standard data often obviates the benchmarking and parameter optimization.

5.2.4 Pitfalls when constructing coexpression networks

How a coexpression network is constructed considerably affects the success of the downstream analysis. Choosing the right combination of the biological datasets used to construct a network is crucial. For example, constructing a coexpression network from heterogeneous expression profiling experiments that differ largely in sample size is challenging. When considering each dataset separately for constructing condition (or study) specific networks, it is often difficult to determine a cutoff on edge reliability that is common to all datasets, as for instance, the degree of correlation is not comparable between datasets of different sizes. A correlation value of 0.2 for instance, has more chance to represent a true association in a network derived from a study

with a large sample size than from a smaller sized study. The approach to construct coexpression networks used in this thesis (mutual rank) can deal at least partially with an imbalanced sample size. In practice, we observed that the used approach (mutual rank) resulted in networks with a comparable number of edges, irrespective of the size of the datasets from which they were constructed (ranging from 20 to > 1000 samples). Therefore, using z-score based or mutual rank-based approaches is highly recommended when constructing coexpression networks from datasets that are heterogeneous in size. This forces the two nodes to be at the nearest neighbors of each other in contrast to using a simple correlation method that requires one of them belongs to the nearest neighbors of the other. The limitation of those approaches, however, is that less likely to produce high-degree (hub) nodes which is a common property of the biological network and reflects the TF and target relationship. Using very strict cutoff results in a network with many connected components which makes the network useless as a connected network is required for most network analysis methods.

5.2.5 Measuring similarity between groups of genes in a biological network

We adopted a method to explore the neighborhood (up to third-order neighbor) of nodes in a network in order to assign a high topological similarity score to nodes that are connected by many neighbors. The approach, based on the Katz index (described in chapter 4) turned out to be very useful in providing a trade-off between only considering direct links in a network or considering the entire local network topology (e.g., with random walks with restart) which are more sensitive to hubs and not suitable for large networks. The Katz index-based approach used in chapter 4 (motif project) is an extension of the employed method to cluster duplicate pairs on the coexpression network described in chapter 2, with a parameter (alpha: the likelihood of effectiveness of a single edge) to weigh neighboring nodes based on their order (first to higher orders neighbors). A length “t” path, then, has a likelihood α^t of being effective. Although this method is powerful for undirected networks with few hub nodes such as coexpression and protein interaction networks, it may not return reasonable results for sparse (metabolic) or directed networks with many hub nodes (regulatory networks). Therefore, the type and structure of the network must be considered in selecting the most appropriate method for measuring the similarity between groups of genes in the biological networks. For instance, in a metabolic network, two genes that are involved in the same biological process (biologically close to each other) might be connected by several intermediate genes (structurally far from each other), indicating the structure has a different meaning in different biological networks.

5.3 Future prospective for each investigated biological question and conclusion

5.3.1 What can be improved to improve studying the fate of genes after gene duplication?

When studying the fate of genes after WGD we observed that a higher expression level of the genes retained after WGD is associated with lower methylation in their flanking regions, pointing towards the role of the gene flanking methylation in regulating expression levels. However, the higher expression level of single copy genes could not be explained by the methylation of the gene flanking regions. Rather it was associated with a high gene body methylation. Increasing evidence supports that in contrast to the induced repression of gene expression by the methylation of flanking regions, gene body methylation would induce gene expression [11]. We also noticed that the higher gene body methylation of single copy genes was associated with the presence of TE and repeats elements. A high methylation of TE and repeat elements is required to silence the TE movement and to maintain the integrity of the gene structure and function and makes the hypothesis on the ability of gene body methylation in inducing gene expression less likely, mostly because the molecular mechanisms of this phenomenon is less clear. Studying other epigenetic modifications, such as histone modification marks will be essential to understanding the role of epigenetic regulation in shaping the evolutionary fate of genes following gene duplication. In addition, moving beyond transcriptomic data and exploring the abundance of proteins and metabolites is key to studying the dosage-balance constraints of duplicate genes. This is particularly true because of the limitation that comes with the use of the short read sequencing data. Differences in gene length and level of sequence divergence among groups of genes with different fates (single-copy, WGD, etc.), might influence the expression results. Although the effect of differences in gene length on the observed number of mapped reads is removed by normalizing the read counts by the gene length, aks FPKM, this normalization is not perfect: observing one read for a gene with 1kb length is not necessarily the same as ten reads for a gene with 10 kb length [12]. In addition, the higher the sequence conservation between paralogs (duplicated genes), the more ambiguous the correct mapping of the reads to these duplicates gene becomes, which may also affect the conclusions related to expression divergence. Repeating those analyses at the transcript level using long-read sequencing data can remove those biases in future studies.

In addition, the role of expression divergence during subfunctionalization of duplicate copies can be further studied at the cis-regulatory divergence level by using information from closely related species. Incorporating sequence, expression, and methylation information from closely related species will enhance the power and accuracy to detect the cis-regulatory divergences that have shaped the expression divergence (subfunctionalization) between duplicate copies. BLSSpeller which has been introduced in the fourth chapter can in principle be redesigned to identify subfunctionalization (motif divergence) during evolution. This potential is reflected in the findings of the second and the fourth chapters which show that tissue-specific genes are associated with a higher copy number (chapter 2) and a higher number of unique motifs in their

upstream regions (Chapter 4). This is consistent with the fact that a higher copy number state relaxes the selection pressure and allows for regulatory divergence and eventually a higher tissue specificity [13].

5.3.2 Future prospective and follow-up study to validate the hypothesis made for BR signaling

Although some interesting results have been revealed by our integrative analysis, our study is limited to transcriptome analysis, whereas BR signaling is likely regulated to a large extent at the post-transcriptional level (i.e., phosphorylation). By using network analysis, we can partially deal with this missing information, but this approach is limited by the incompleteness of the interaction network. Hence, more extensive validation studies are required to confirm our hypotheses. Besides, this study mainly focuses on the suppressor lines of *bri1-5*. Future work by adding enhancer lines such as *bri1-5/bri1-1*, *bri1-5/brs1-1*, *and bri1-5/bak1-1* or by studying the effect of suppressors, not only in the *BRI1* as a loss-of-function background, but also using double or triple mutants of multiple genes of the *BRI1* gene family and/or *BRI1*-like genes [14] could be interesting to further elucidate the mechanism of *BRI* signaling. In addition, future research by scRNA-seq of *bri1-5* suppressor mutants across different time points of hypocotyl growth would be key to elucidating the detailed mechanisms of BR signaling at the level of cell types [15, 16]. The limitations and pitfalls of network-based methods that have been described are also relevant to our BR signaling study.

5.3.3 Limitations and future extensions of BLSSpeller

The performance of BLSSpeller is determined by the number of related species included in the study and the quality of the orthologous gene family membership. All genes are grouped in gene families where each species can contribute more than one sequence (paralogs) to a gene family. All the paralogs are represented by a single node in the phylogenetic tree, meaning the presence of a motif in only one paralog is enough to call the motif conserved in that species regardless of the number of paralogs that gene has for that species. In other words, more paralogs imply more sequence space on which a motif can be detected. Although the BLS and recurrence score filtering down-weight the impact of the number of paralogs, there is room for improvement by incorporating all paralogs in the phylogenetic tree. This can be achieved by considering gene family-specific phylogenetic trees, and taking into account the duplications and speciation events in such trees. The latter might be particularly important to account for the difference between duplicates that occurred before or after speciation. When assuming that subfunctionalization of the promoter sequences is going hand in hand with subfunctionalization at protein level, BLSSpeller could be used with a gene tree to identify subfunctionalization. This would ideally require that the score function is adapted as such that motif sequences of genes that are close on the tree contribute relatively less to the motif model than motif hits in

more distantly related sequences, but at the same time that closely related sequences are penalized more if do not contain the motif than distantly related sequences on the gene tree. Adapting the scoring function as such is not trivial however as a motif is absence or presence is not clearly defined. Therefore the same could be achieved pragmatically by simply postprocessing the results of BLsspeller, run with a gene tree.

Another limitation is related to the rescaling of the phylogenetic tree in order to detect motifs that occur in smaller gene families with missing species. This step comes at the expense of finding potentially more spurious motifs as it is relatively easier to find a random motif being conserved in a smaller number of species. Dealing with the imbalance in the number of orthologs in gene families can be investigated in future studies. Relaxing the predefined motif length can also improve the results and capture motifs with different lengths as this is a characteristic of the known motifs. It should also be noted that the motifs are only predictive of TF binding sites and inferring the corresponding TF to each motif is difficult as multiple TFs from similar families can bind to similar motifs. Those TFs may additively or synergistically regulate target genes or may act in a redundant manner to mitigate the effects of perturbation caused by genetic variations. Therefore, inferring transcriptional interactions from motifs, even for experimentally identified CRE, is still a challenging task.

5.3.4 Extend and modify the input for motif discovery

Although the considered upstream region (1 kb) is the region with the strongest signals for cis-regulatory elements, a substantial portion of regulatory elements does not reside in this region. A larger upstream region up to a few kb and/or different non-coding regions relevant to gene regulation (e.g., 3' and introns) could be considered for motif identification. However, in the current version of BLSSpeller, the scoring function (BLS) relies merely on the presence/absence of the motif in the sequence. Therefore, the length of the sequence under consideration could influence the probability that a motif hit is found by chance. Using more sequence space will therefore give rise to more spurious hits as the statistics are no longer correct. The second imposed filtering criterium (recurrence score) partially compensates for this issue but inan arbitrary and statistically sound way. Considering the sequence length in the scoring system would therefore further improve the performance of the algorithm and allows to include larger regulatory regions, including downstream of 3' regions.
It has been shown that introns, intragenic regions, enhance the expression of their respective genes [17]. Therefore, considering introns allows us to discover the regions of regulation that reside in those non-coding regions. However, introns play an essential role in alternative splicing, mRNA-stability and therefore can contain RNA genes, such as snoRNAs, long non-coding RNAs (lncRNAs), miRNAs, and small-interfering RNAs (siRNAs) [18]. These can be conserved relatively more and over longer distances (higher signal to noise region) than regulatory motifs (of TFs) and therefore might obviate finding the smaller and less conserved regulatory motifs in the same intergenic regions. Therefore, regions containing RNA genes

should be masked prior to perform a motif search and it should be taken care of that no flanking coding regions are still present in the input set.

If the chromatin accessible regions are available for considered species, considering only those accessible regions can avoid finding spurious hits, but might come at the expense of missing tissue-specific motifs (as chromatin accessibility data are condition specific and only available for particular conditions), as has been discussed in chapter 4.

References

1. Aizat WM, Ibrahim S, Rahnamaie-Tajadod R, Loke K-K, Goh H-H, Noor NM: **Proteomics (SWATH-MS) informed by transcriptomics approach of tropical herb Persicaria minor leaves upon methyl jasmonate elicitation.** *PeerJ* 2018, **6:**e5525.
2. Mata CI, Fabre B, Parsons HT, Hertog ML, Van Raemdonck G, Baggerman G, Van de Poel B, Lilley KS, Nicolaï BM: **Ethylene receptors, CTRs and EIN2 target protein identification and quantification through parallel reaction monitoring during tomato fruit ripening.** *Frontiers in plant science* 2018, **9:**1626.
3. Peng Z, He S, Gong W, Xu F, Pan Z, Jia Y, Geng X, Du X: **Integration of proteomic and transcriptomic profiles reveals multiple levels of genetic regulation of salt tolerance in cotton.** *BMC plant biology* 2018, **18:**1-19.
4. Voit EO: **The best models of metabolism.** *Wiley Interdisciplinary Reviews: Systems Biology and Medicine* 2017, **9:**e1391.
5. Lee I, Li Z, Marcotte EM: **An improved, bias-reduced probabilistic functional gene network of baker's yeast, Saccharomyces cerevisiae.** *PloS one* 2007, **2:**e988.
6. Lei X, Yang X, Fujita H: **Random walk based method to identify essential proteins by integrating network topology and biological characteristics.** *Knowledge-Based Systems* 2019, **167:**53-67.
7. Valdeolivas A, Tichit L, Navarro C, Perrin S, Odelin G, Levy N, Cau P, Remy E, Baudot A: **Random walk with restart on multiplex and heterogeneous biological networks.** *Bioinformatics* 2019, **35:**497-505.
8. Cowen L, Ideker T, Raphael BJ, Sharan R: **Network propagation: a universal amplifier of genetic associations.** *Nature Reviews Genetics* 2017, **18:**551-562.
9. De Smet R, Marchal K: **Advantages and limitations of current network inference methods.** *Nature Reviews Microbiology* 2010, **8:**717-729.
10. Misra BB, Langefeld C, Olivier M, Cox LA: **Integrated omics: tools, advances and future approaches.** *Journal of molecular endocrinology* 2019, **62:**R21-R45.
11. Gent JI, Ellis NA, Guo L, Harkess AE, Yao Y, Zhang X, Dawe RK: **CHH islands: de novo DNA methylation in near-gene chromatin regulation in maize.** *Genome research* 2013, **23:**628-637.
12. Dillies M-A, Rau A, Aubert J, Hennequet-Antier C, Jeanmougin M, Servant N, Keime C, Marot G, Castel D, Estelle J: **A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis.** *Briefings in bioinformatics* 2013, **14:**671-683.
13. Cheng F, Wu J, Cai X, Liang J, Freeling M, Wang X: **Gene retention, fractionation and subgenome differences in polyploid plants.** *Nature plants* 2018, **4:**258-268.
14. Lozano-Elena F, Caño-Delgado AI: **Emerging roles of vascular brassinosteroid receptors of the BRI1-like family.** *Current opinion in plant biology* 2019, **51:**105-113.
15. Zhang T-Q, Xu Z-G, Shang G-D, Wang J-W: **A single-cell RNA sequencing profiles the developmental landscape of Arabidopsis root.** *Molecular plant* 2019, **12:**648-660.
16. Liu Q, Liang Z, Feng D, Jiang S, Wang Y, Du Z, Li R, Hu G, Zhang P, Ma Y: **Transcriptional landscape of rice roots at the single-cell resolution.** *Molecular Plant* 2021, **14:**384-394.

17. Back G, Walther D: **Identification of cis-regulatory motifs in first introns and the prediction of intron-mediated enhancement of gene expression in Arabidopsis thaliana.** *BMC genomics* 2021, **22:**1-24.
18. Bush SJ, Chen L, Tovar-Corona JM, Urrutia AO: **Alternative splicing and the evolution of phenotypic novelty.** *Philosophical Transactions of the Royal Society B: Biological Sciences* 2017, **372:**20150474.

6 Supplementary figures and tables

Supplementary data for chapter 2 are attached but are available online at:

https://academic.oup.com/mbe/article/37/8/2394/5826357?login=true#supplementary-data

6.1 Genome-wide expression and network analyses of mutants in key brassinosteroid signaling genes

Complete supplementary data for chapter 3 are available online at:
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8220701/

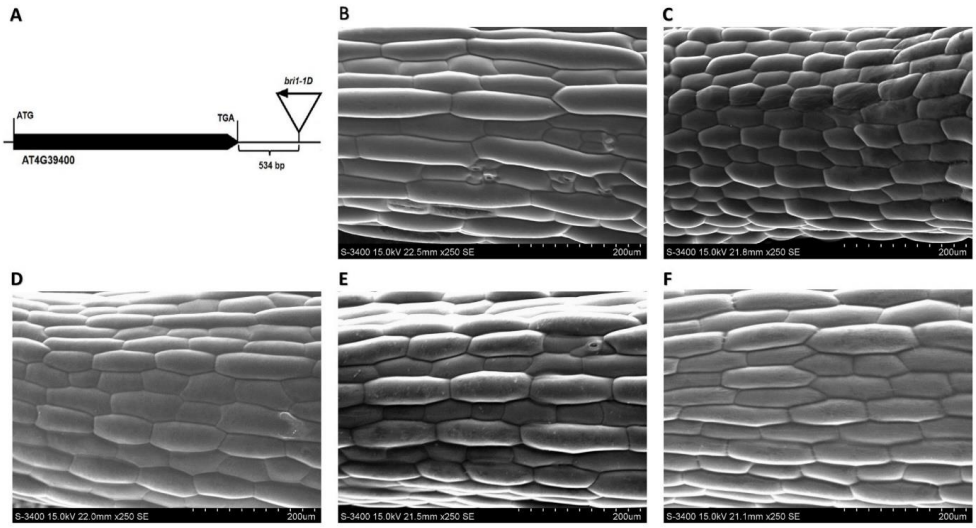


Figure S3.1: The T-DNA insertion site for *bri1-1D* (A), and microscopic images of 7-day old hypocotyl cells for WS2 (B), *bri1-5* (C), *bri1-5/bak1-1D (*D*), bri1-5/bri1-1D* (E), *bri1-5/brs1-1D* (F).

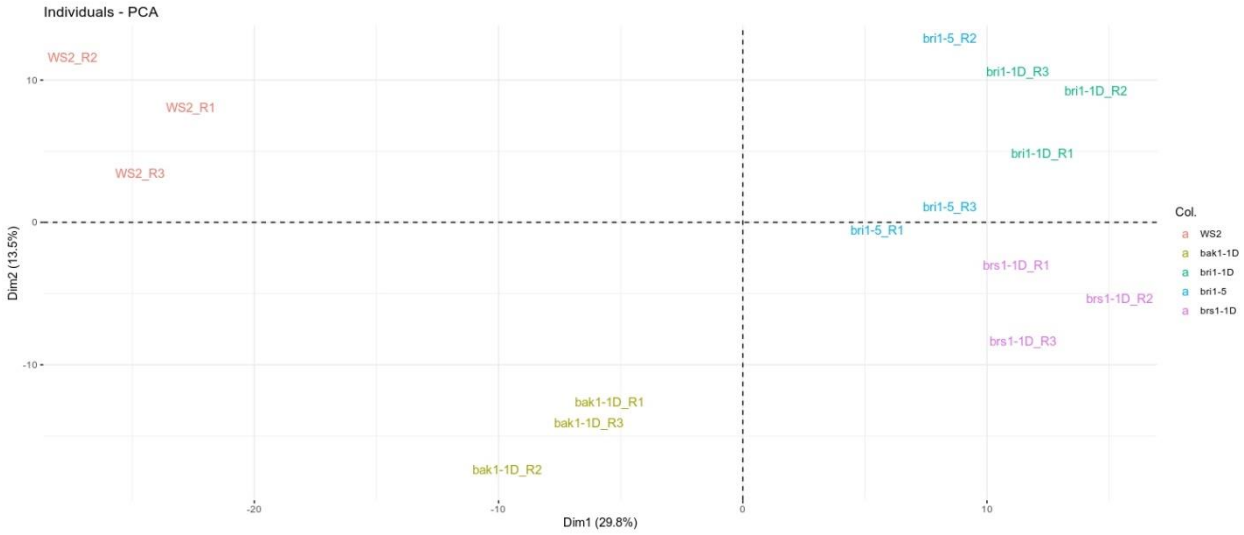


Figure S3.2: PCA plot for assessing the reproducibility of the gene expression dataset. Samples taken from the same genotype are represented in the same color. The plot indicates high consistency between replicate samples as they are located close to each other when plotted on the first and second principal components.

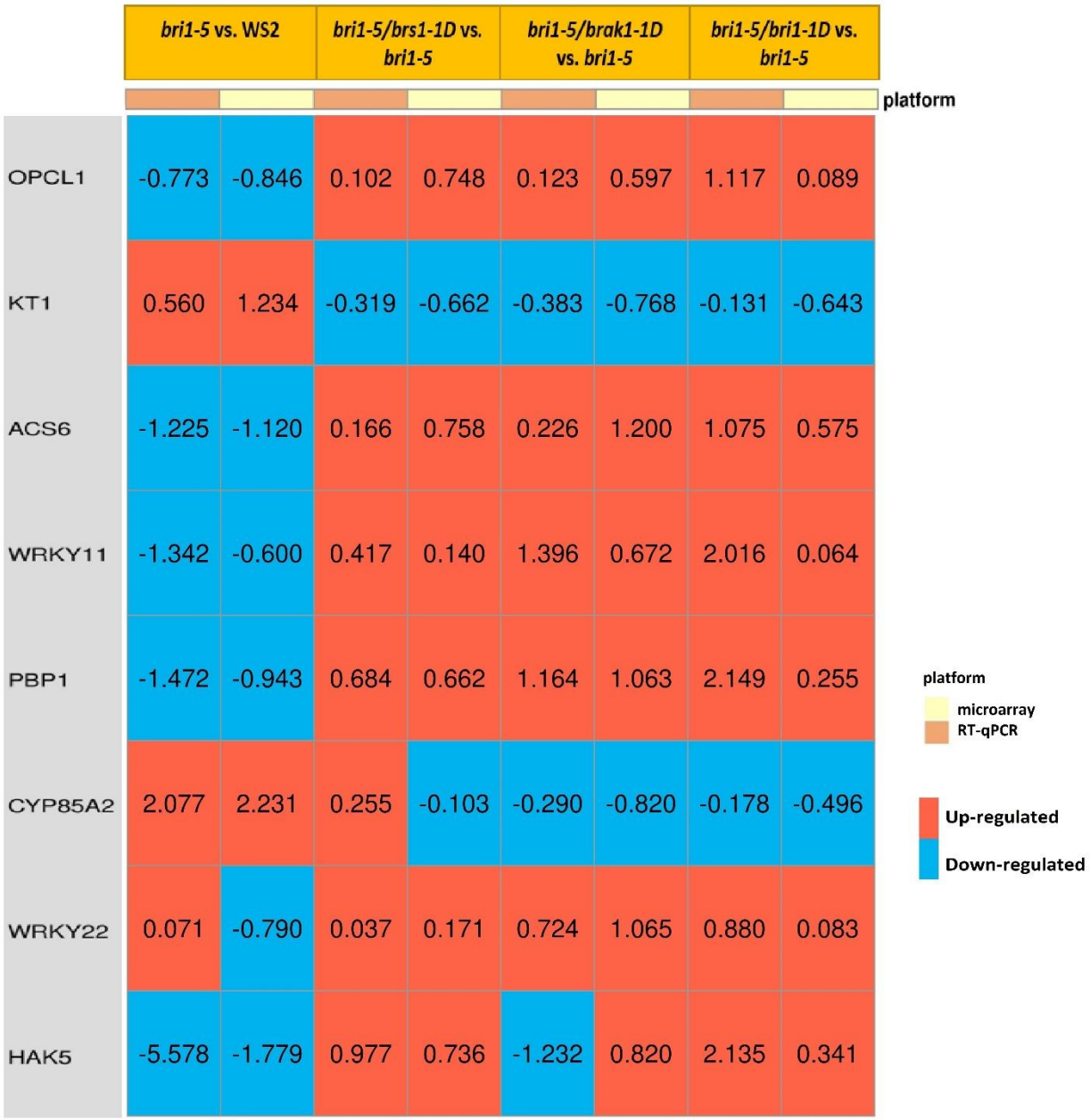


Figure S3.3: RT-qPCR results for relative expression of selected genes and their corresponding values from microarray analysis. The values represent the log2 of relative expression (sample1/sample2). Rows indicate gene names and columns show the comparison between the indicated lines. Columns with pink header represent the RT-qPCR values, and columns with yellow header are microarray measurements. The red color on the heatmap indicates that the gene has been up-regulated in sample1 as compared to sample2, while blue indicates down-regulation.

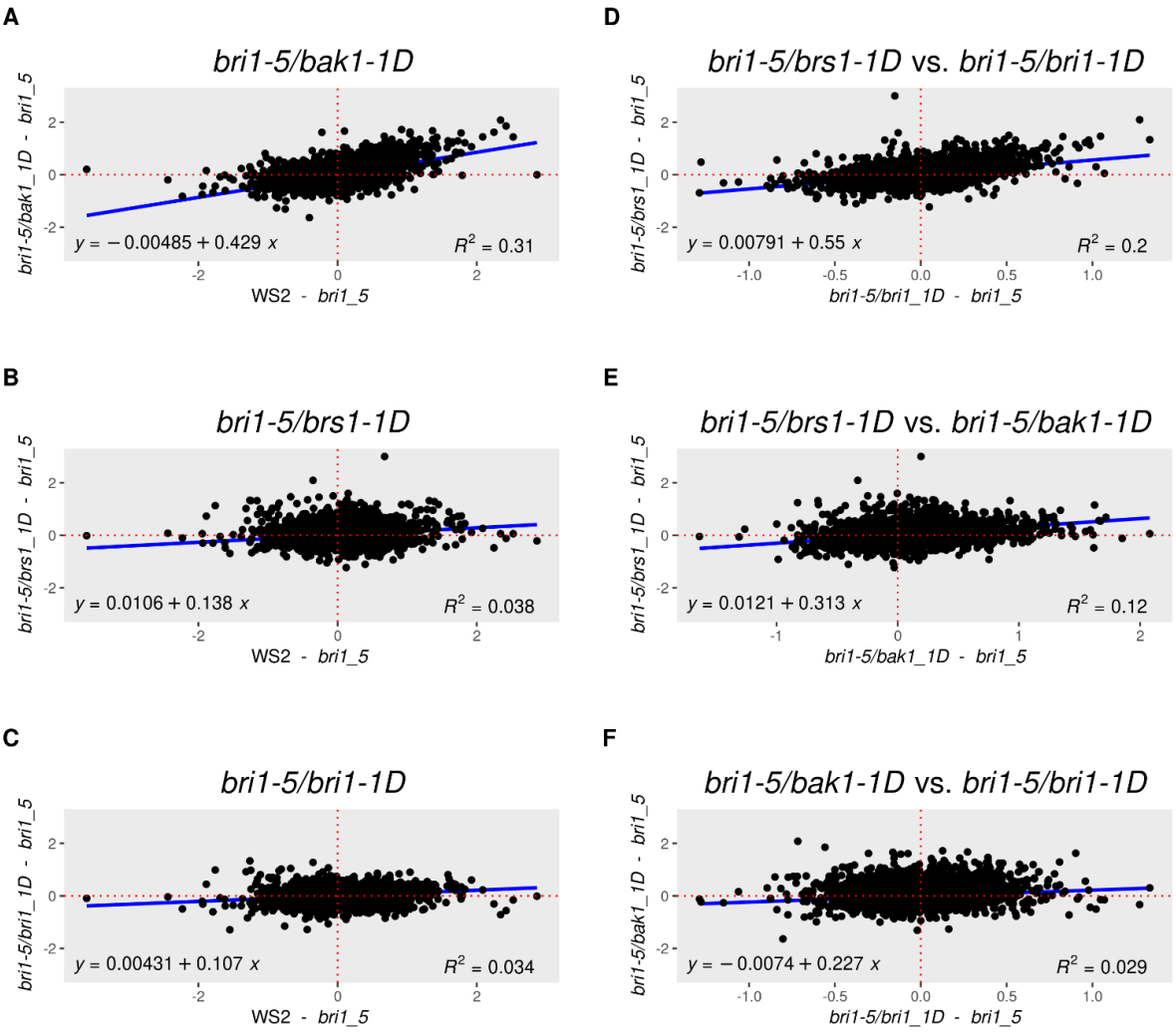


Figure S3.4. Comparing genome-wide expression impact between *bri1-5* suppressor lines. Panel A-C: Scatter plots display the log2 mean expression difference between each suppressor and *bri1-5* (Y-axis) versus the log2 mean expression difference of WS2 and *bri1-5* (X-axis). A value around zero on the Y-axis and a non-zero value on the X-axis means that a gene that was affected in the *bri1-5* mutant is not restored by the suppressor line displayed in the subplot title. A positive regression slope indicates that genes that are affected in the *bri1-5* versus WS2, have been restored to WS2 level by the suppressor line displayed on the subplot title. So it indicates that the aberrant expression observed in the *bri1-5* with the WS2 as a reference is restored for in the suppressor mutant. Hence, a stronger positive regression indicates that the suppressor mutant better approximates the WS2 and hence better compensates for the *bri1-5* phenotype. Panels D-F indicate to what extent the same genes were affected in the *bri1-5* suppressors lines (*bri1-5*/*bri1-1D*, *bri1-5*/*brs1-1D,* and *bri1-5*/*bak1-1D*) using the *bri1-5* as a common reference. A higher positive correlation (or regression slope) indicates that similar genes are affected by each pair of compared mutant lines.

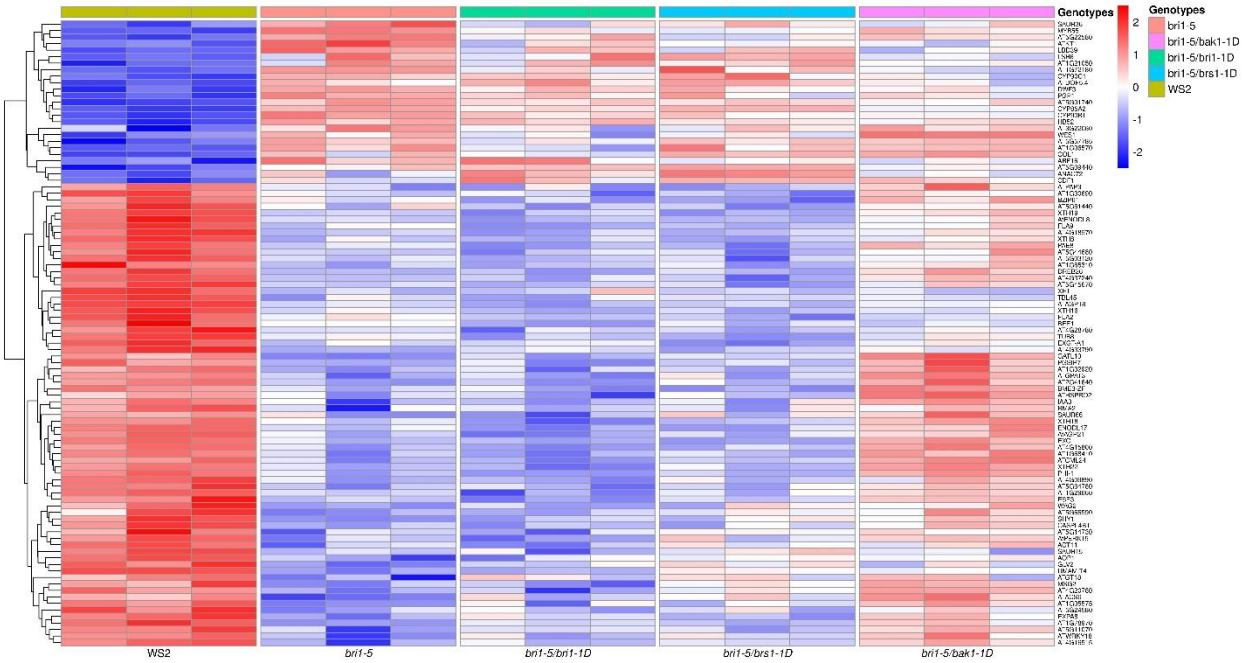


Figure S3.5: Heatmap of expression of the marker genes that up/down regulation of their expression was confirmed by at least 5 independent references and also affected in the *bri1-5* line of our study. For each line, the row-scaled normalized expression data of the 3 biological replicates are shown as adjacent columns. In each row the gradient red color indicates the higher expression for the gene compared to other samples while blue indicates the lower expression.

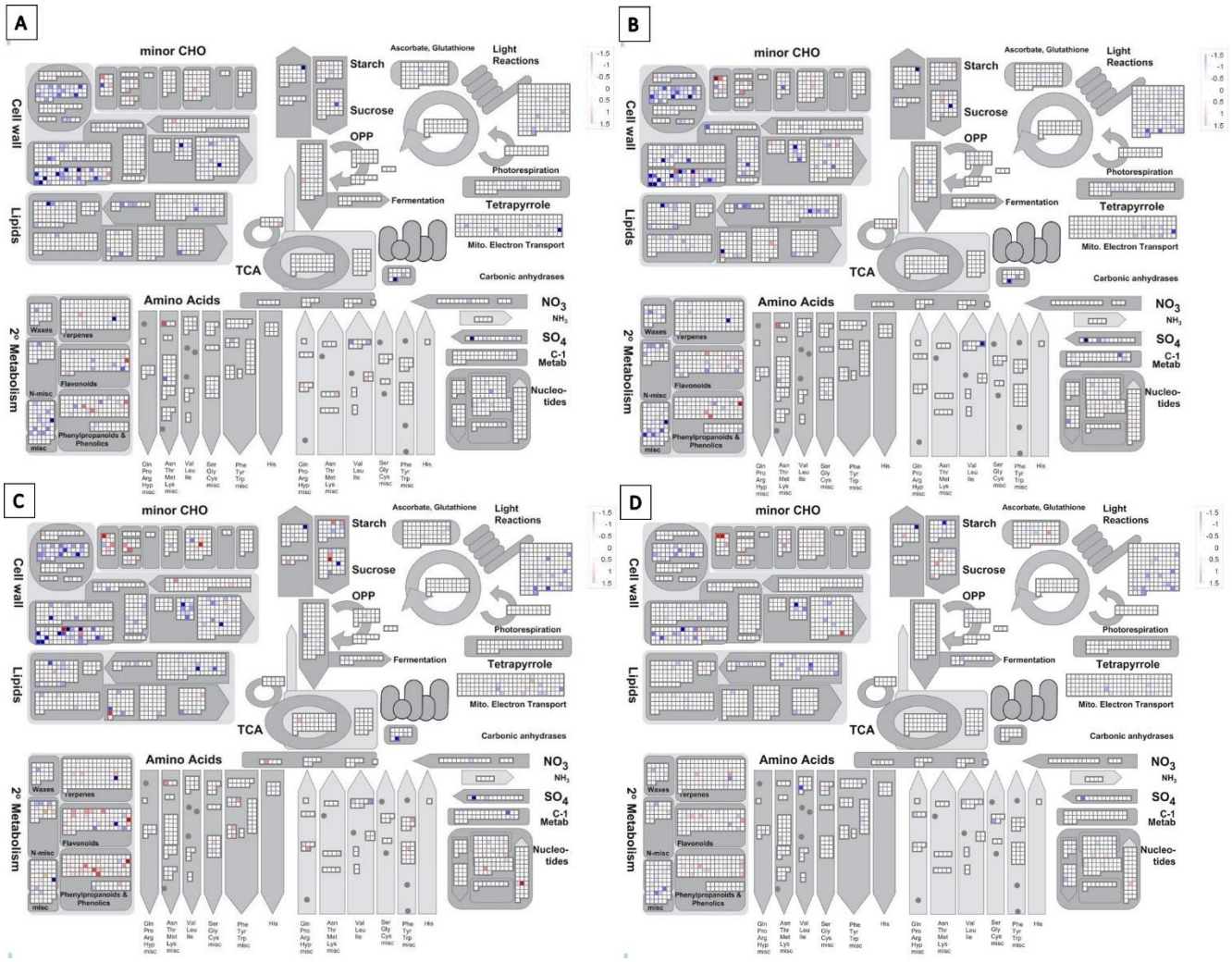


Figure S3.6: Pathway analysis (MapMan metabolism) showing for each mutant line the expression changes compared to WS2. Panel A: *bri1-5,* Panel B*: bri1-5/bri1-1D*, Panel C: *bri1-5/brs1-1D*, Panel D: *bri1-5/bak1-1D*.

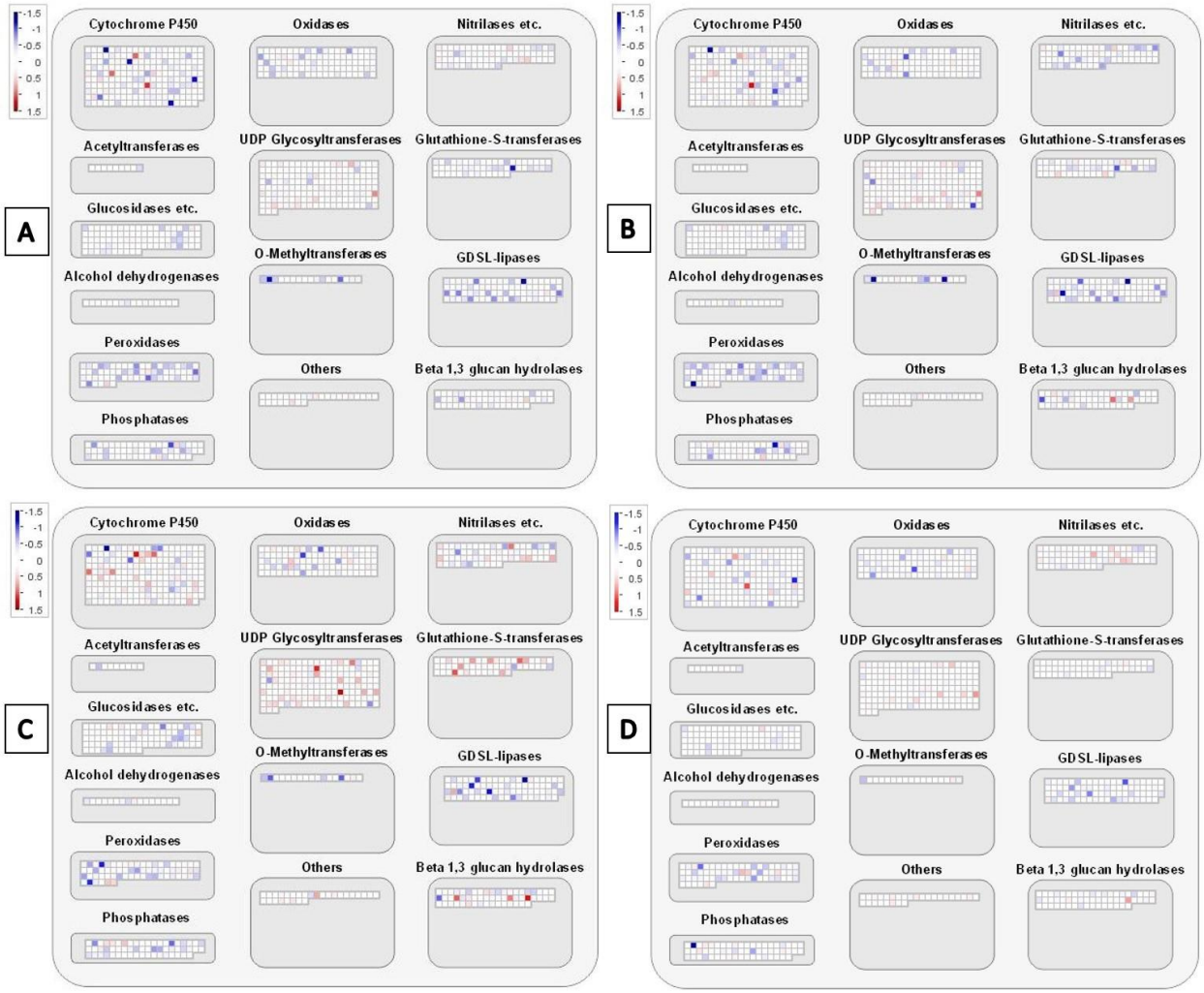


Figure S3.7: Pathway analysis (MapMan: large enzyme families) showing for each mutant line the expression changes compared to WS2. Panel A: *bri1-5,* Panel B*: bri1-5/bri1-1D*, Panel C: *bri1-5/brs1-1D*, Panel D: *bri1-5/bak1-1D*.

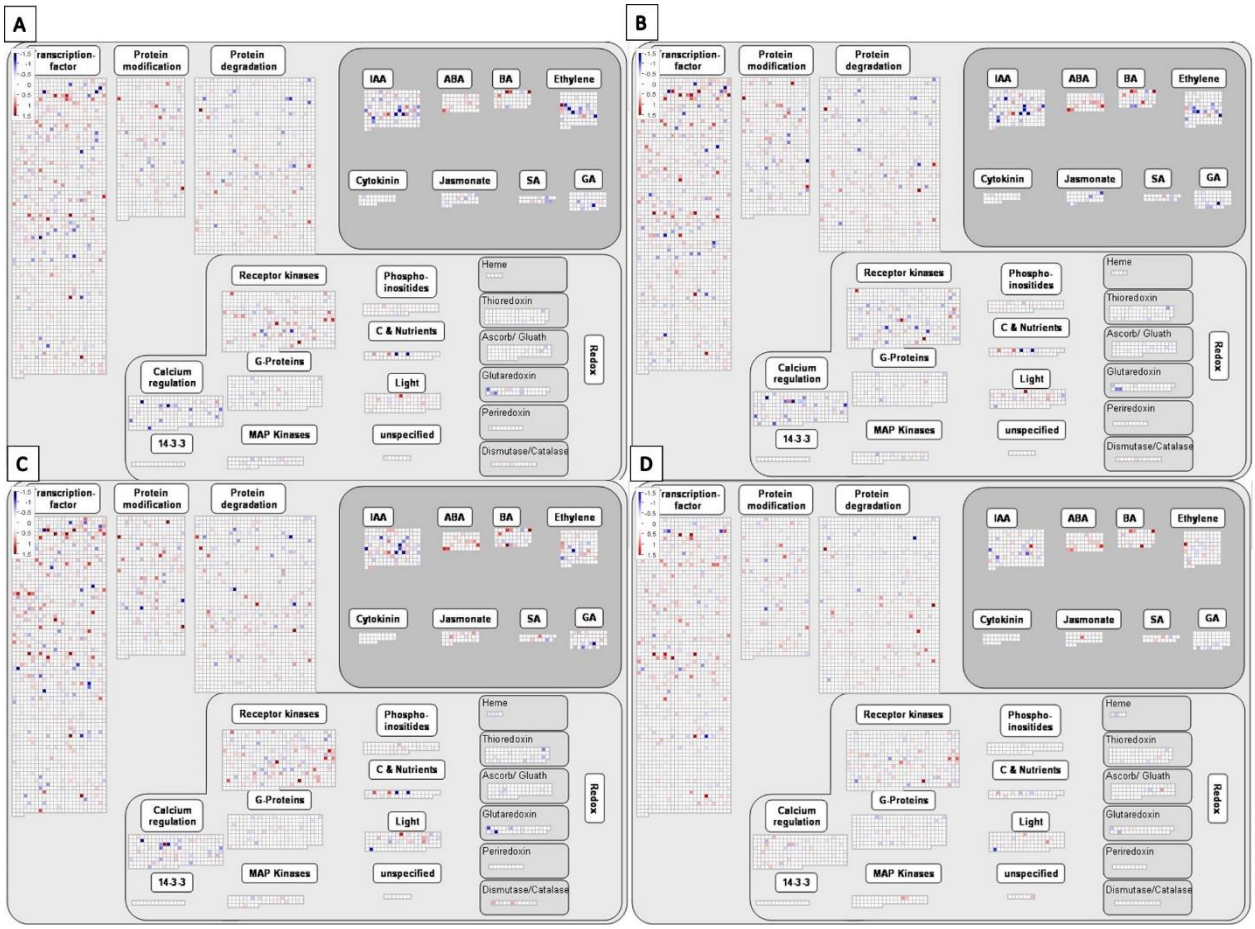


Figure S3.8: Pathway analysis (MapMan: gene regulation) showing for each mutant line the expression changes compared to WS2. Panel A: *bri1-5,* Panel B*: bri1-5/bri1-1D*, Panel C: *bri1-5/brs1-1D*, Panel D: *bri1-5/bak1-1D*.

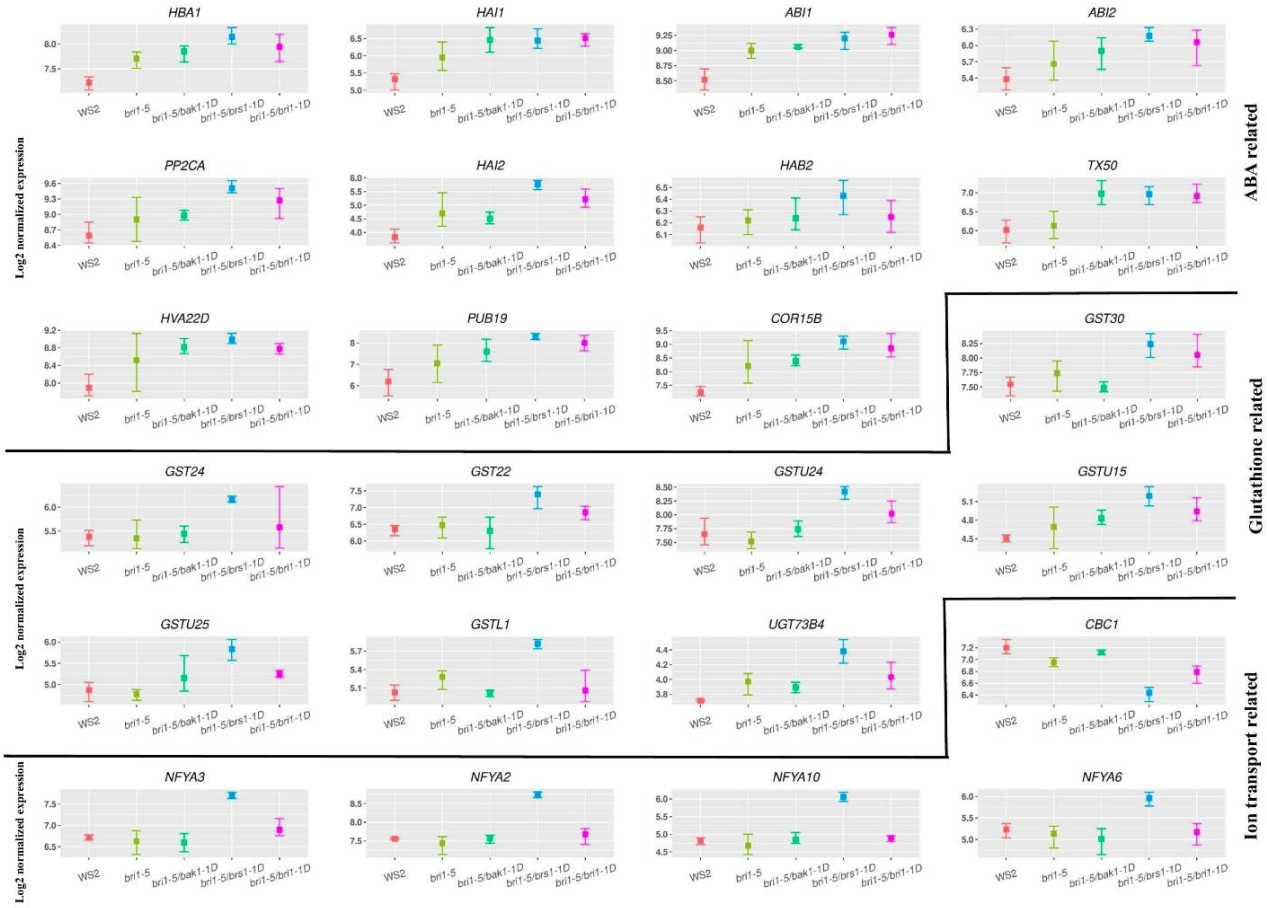


Figure S3.9: Expression pattern in each mutant line of genes related to ABA signaling, Glutathione metabolism, and ion related hemostasis as discussed in the main text. Mutant lines are represented in the x-axis. The y-axis indicates the log2 normalized expression value of the gene.

Table S3.1. RT-qPCR test of log-fold change (log-FC) of the genes that are overexpressed by activation-tagging in the suppressors at the 7 days seedling stage.

	Log-FC (RT-qPCR)	Log-FC (microarray)
BAK1 (*bri1-5/bak1-1D* vs. *bri1-5*)	3.576	1.611
BRI1 (*bri1-5/bri1-1D* vs. *bri1-5*)	1.847	0.911
BRS1 (*bri1-5/brs1-1D* vs. *bri1-5*)	2.388	5.386

Table S3.2: Summary of the most significant results obtained by MapMan pathway analysis (metabolism, regulation and, large-enzyme families overview). Left column: enriched pathways; entries provide for each line the degree to which the pathway is enriched. P-values are FDR corrected using Benjamini-Hochberg).

	bri1-5	***bri1-5/bri1-1D***	***bri1-5/brs1-1D***	***bri1-5/brak1-D***
RNA regulation of transcription	1e-20	1e-20	9.10e-14	1e-20
Protein degradation	6.77e-7	1.45e-9	7.46-3	6.17e-10
Protein post-ranslational modification	1.02e-4	5.46e-3	7.06e-2	3.13e-3
Amino acid metabolism	2.13e-2	2.71e-3	1.34e-3	1.39-5
Cell wall	2.01e-12	1.25e-11	5.97e-7	2.25e-2
Mitochondrial electron transport- / ATP synthesis	1.48e-8	3.63e-6	1.14e-5	2.69e-8
Lipid metabolism- FA synthesis and FA elongation	1.423-3	3.69e-3	2.04e-3	7.79e-7
Secondary metabolism- sulfur-containing	4.30e-3	1.17e-5	7.17e-2	2.63e-5
Cell wall- cell wall proteins (AGPs)	7.26e-9	7.16e-9	8.64e-9	9.83e-4
PS- light reaction	1.07e-8	1e-20	1.36e-11	1.56e-2
Cell wall- cellulose synthesis	9.30e-2	1.67e-3	4.63e-4	9.01e-2
Abscisic acid (ABA) metabolism	0.28	3.25e-2	7.11e-2	4.53e-5
Cytochrome P450	3.35e-4	0.87	0.72	0.34
Glutathione S transferases	8.87e-4 (down-regulated genes)	0.12	2.48e-6 (up-regulated genes)	0.41

Table S3.3. Designed primers for RT-qPCR.

Gene name	Gene symbol	Forward primer	Reverse primer
Actin	*ACT*	TCAGATGCCCAGAAGTCTTGTTCC	CCGTACAGATCCTTCCTGATATCC
AT1G20510	OPCL1	TCAAATCCAGTCCTCGCCTT	TCGTTCCTCGTCCATAAGCA
AT1G80840	WRKY40	AGGGATCAAATCAGCCCTCC	GAGCTACTCTCCGACACTCC
AT2G30070	KT1	TGGAACCCGCGTATAGTCTC	ACACCTCCAAGTGAAACCCA
AT3G30180	BR6OX2	GAAAGGACTTGTTGCCGGTT	CAACGAGCCTCTCATTAGCC
AT3G49580	LSU1	TTAAGTTGTGGCAGCGAACG	AGAGCATGCGATCGTGATAGT
AT4G01250	WRKY22	TCTTTGTCGGATACGGTGGT	AGAGATCAAAACTCGCCGGA
AT4G11280	ACS6	TTTCGTCACAAACGCAGCAT	GGTTATCTCAGCGTGCCTTG
AT4G13420	HAK5	GACCGAGATTACGGCAGAGA	CGCAAGTGCTTTGTCTCCTT
AT4G31550	WRKY11	CCACCGTCTAGTGTAACACTCGAT	TGCAACGGAGCAGAAGCAAGGAA
AT5G54490	PMP1	TTGCAAAGGGTTCGAGCTTC	TATCGAACATCGTCGTCCGT
AT4G33430	BAK1	AGACTGGGTGAAAGGGTTGT	TAGCTGCTCCACTTCTTCGT
AT4G39400	BRI1	TTTGCACGACCCCAAGAAAG	TGTGGTGAAGGAAAGCAAGC
AT4G30610	BRS1	TCACGGTGGAGAGAGTTTCC	CGTTGGGGTGTAAATGCTGT

6.2 BLSSpeller to discover novel regulatory motifs in maize

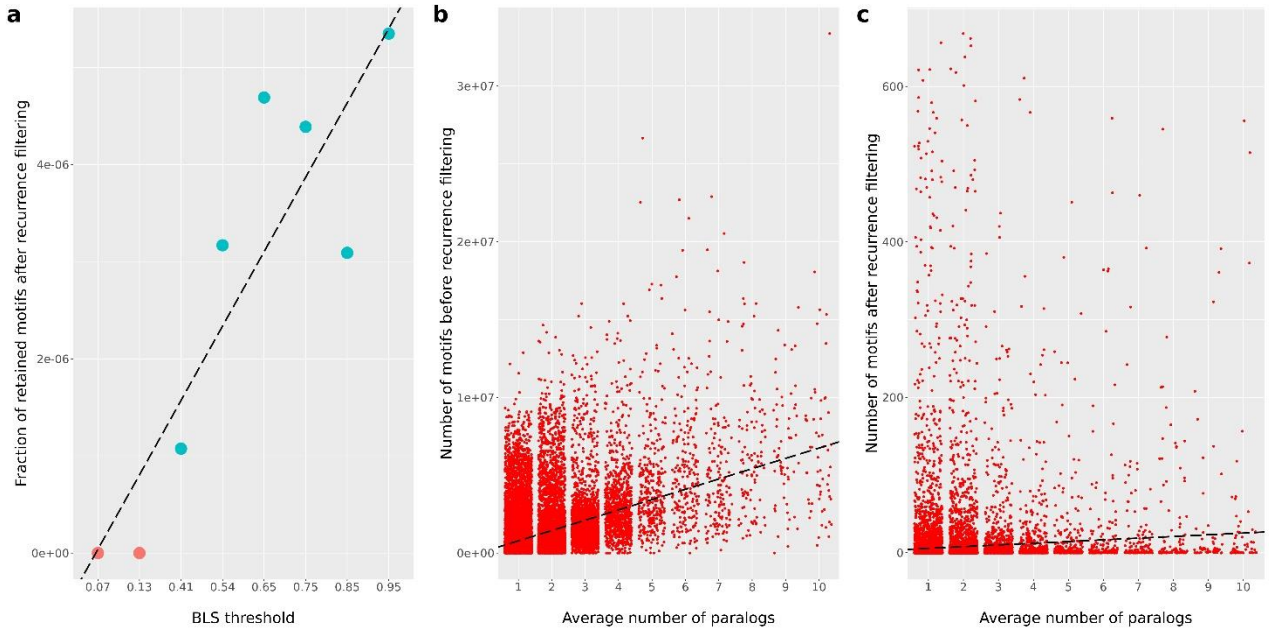


Figure S4.1: Distribution of fraction of motifs retained after applying recurrence filtering for different BLS thresholds (a); the number of motifs per gene family obtained at the highest BLS threshold (0.95) as a function of the number of paralogs in the gene family before (b) and after (c) applying the recurrence threshold (0.9). Panel (a) shows that after applying the recurrence filtering, no motifs were retained with a low BLS (low BLS thresholds 0.07 and 0.13) and that the fraction of motifs retained after applying the recurrence threshold increases with the BLS. Panels (b) and (c) show how the recurrence filtering penalizes the motifs detected in gene families with a high number of paralogs. The number of motifs with a high BLS increases with the number of paralogs in a gene family (b). This is to be expected as a higher number of paralogs implies a larger search space and hence also a higher probability of finding a highly conserved motif by chance. However, many of these motifs are filtered because of a low recurrence score, indicating that they were indeed likely spurious (c). The dashed line in each panel shows the regression line fitted to predict values on the y-axis based on values on the x-axis.

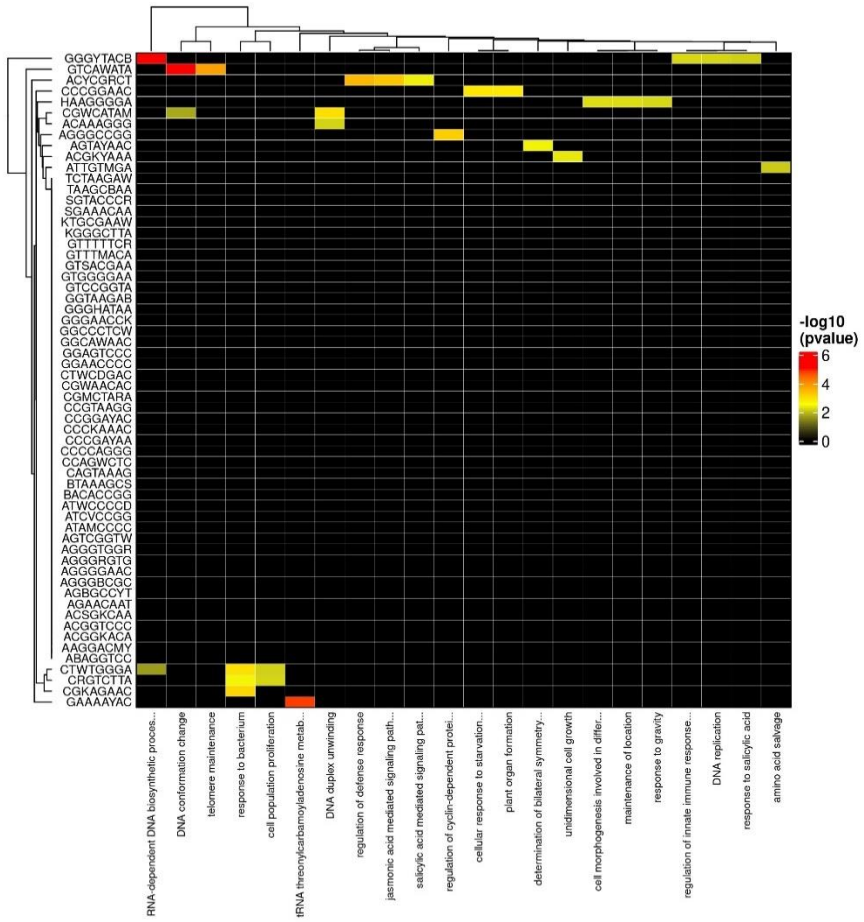


Figure S4.2: Enriched GO terms (Biological Process) for the gene sets corresponding to the random maize motifs. Each column represents a gene set sharing the corresponding random motif in maize and each column indicates a specific biological process. The entries indicate the -log10 p-values of the GO enrichment. Only the five most significantly enriched GO terms are shown. Unlike predicted motifs, only 15 gene sets sharing a random motif were enriched for at least one biological process

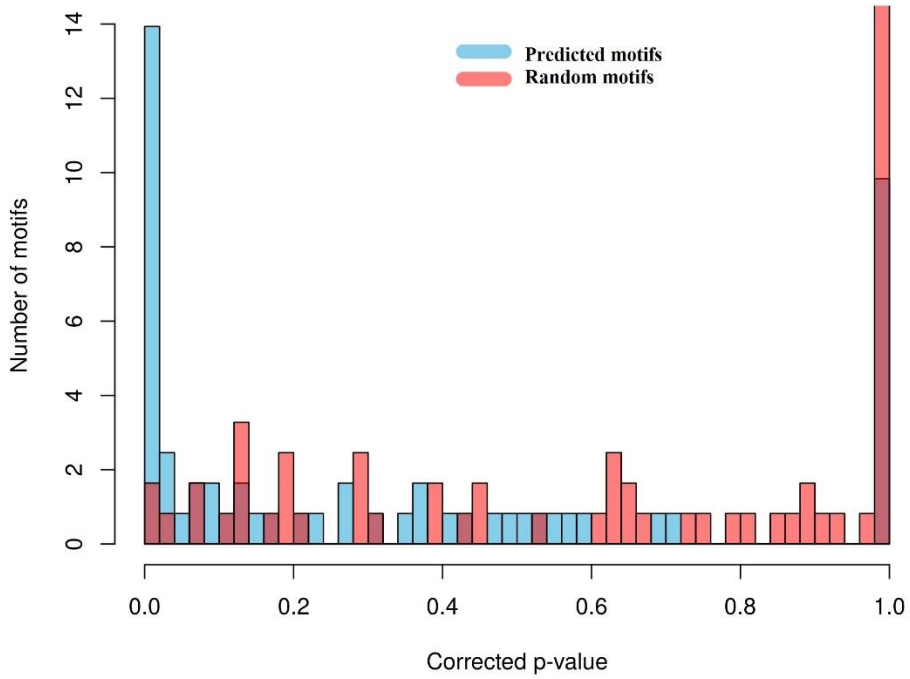


Figure S4.3. Comparison of the predicted and random motifs with DAP-seq motifs in Arabidopsis. Similarities were calculated with Tomtom [1]. The corrected similarity p-value (q-value) with the closest DAP-seq motifs for predicted (blue) and random motifs (red) are shown.

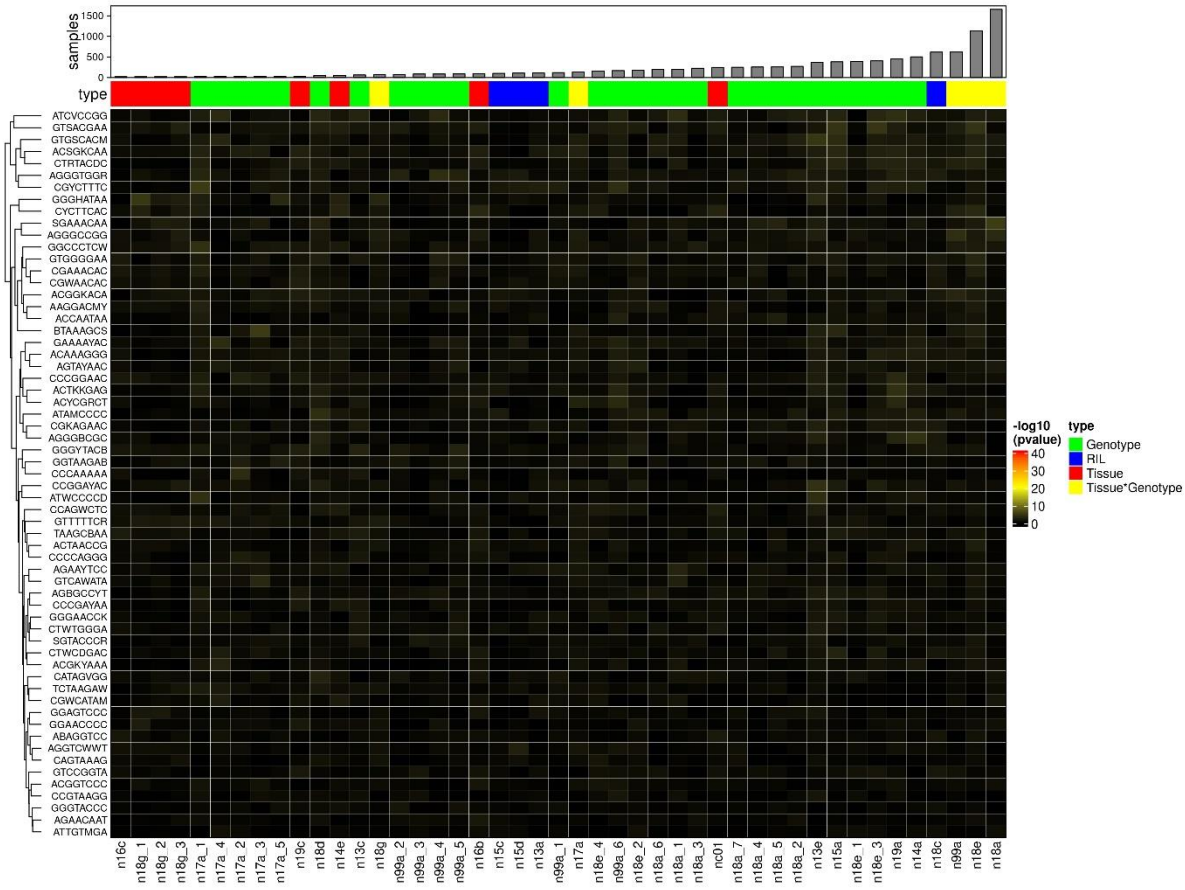
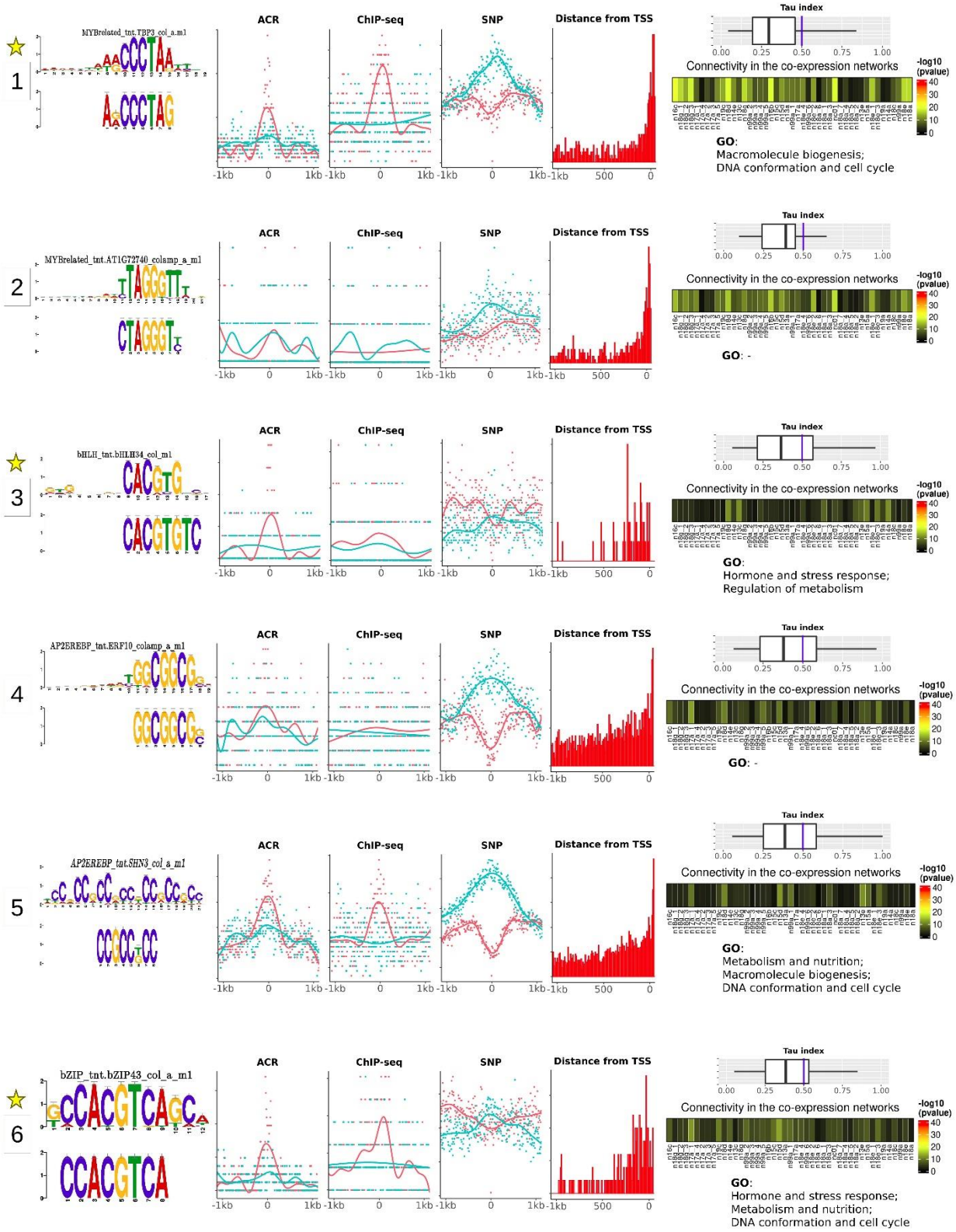
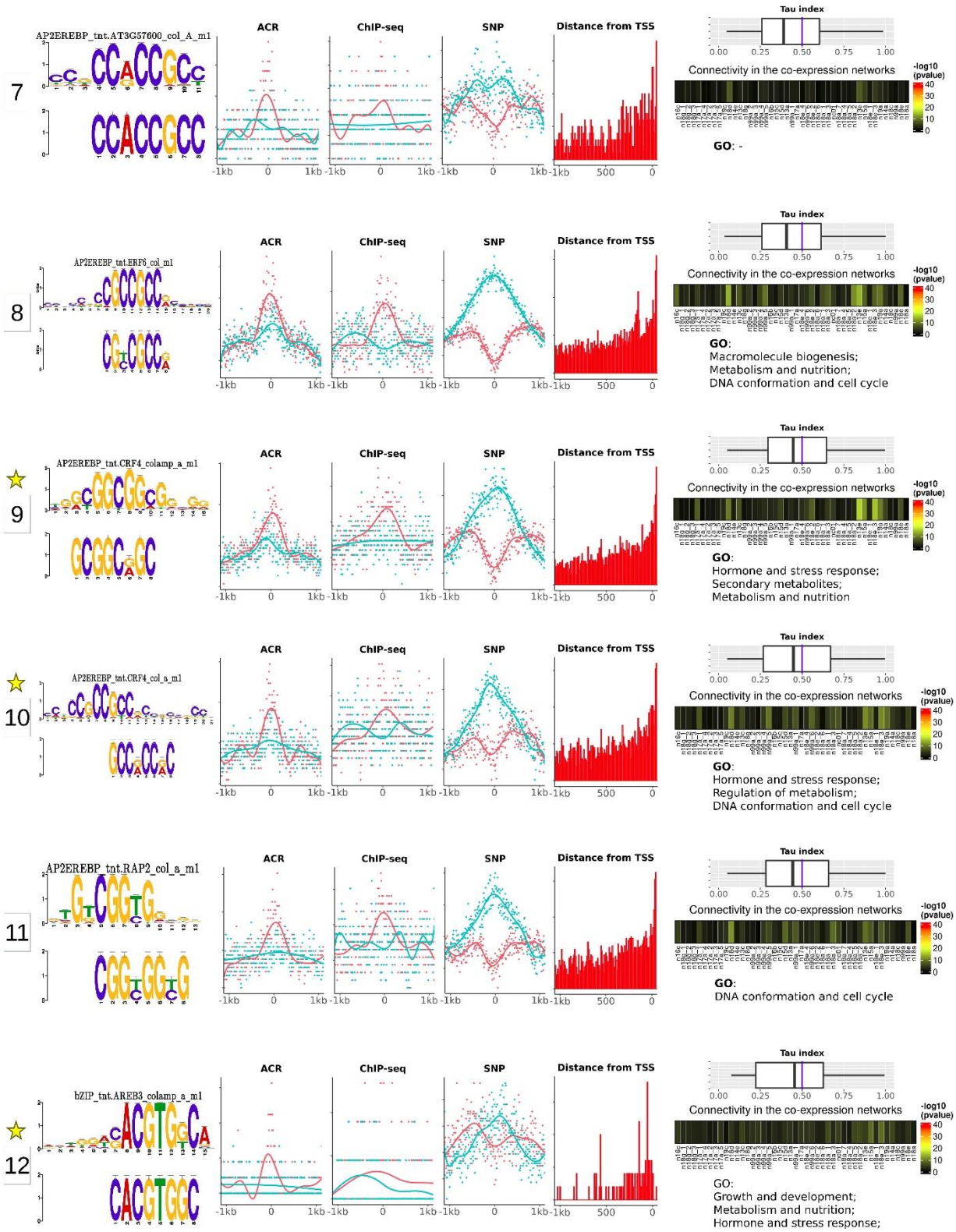
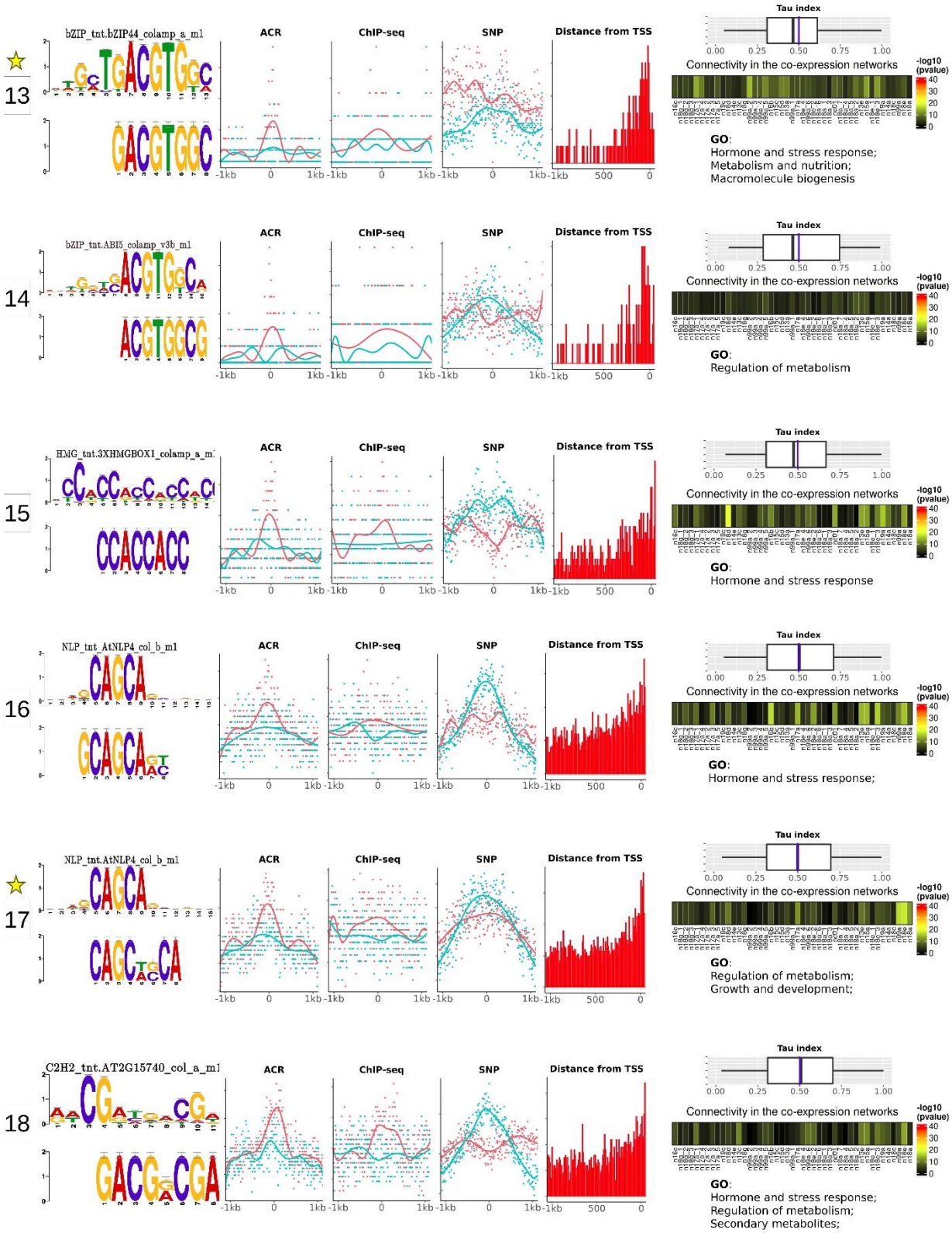
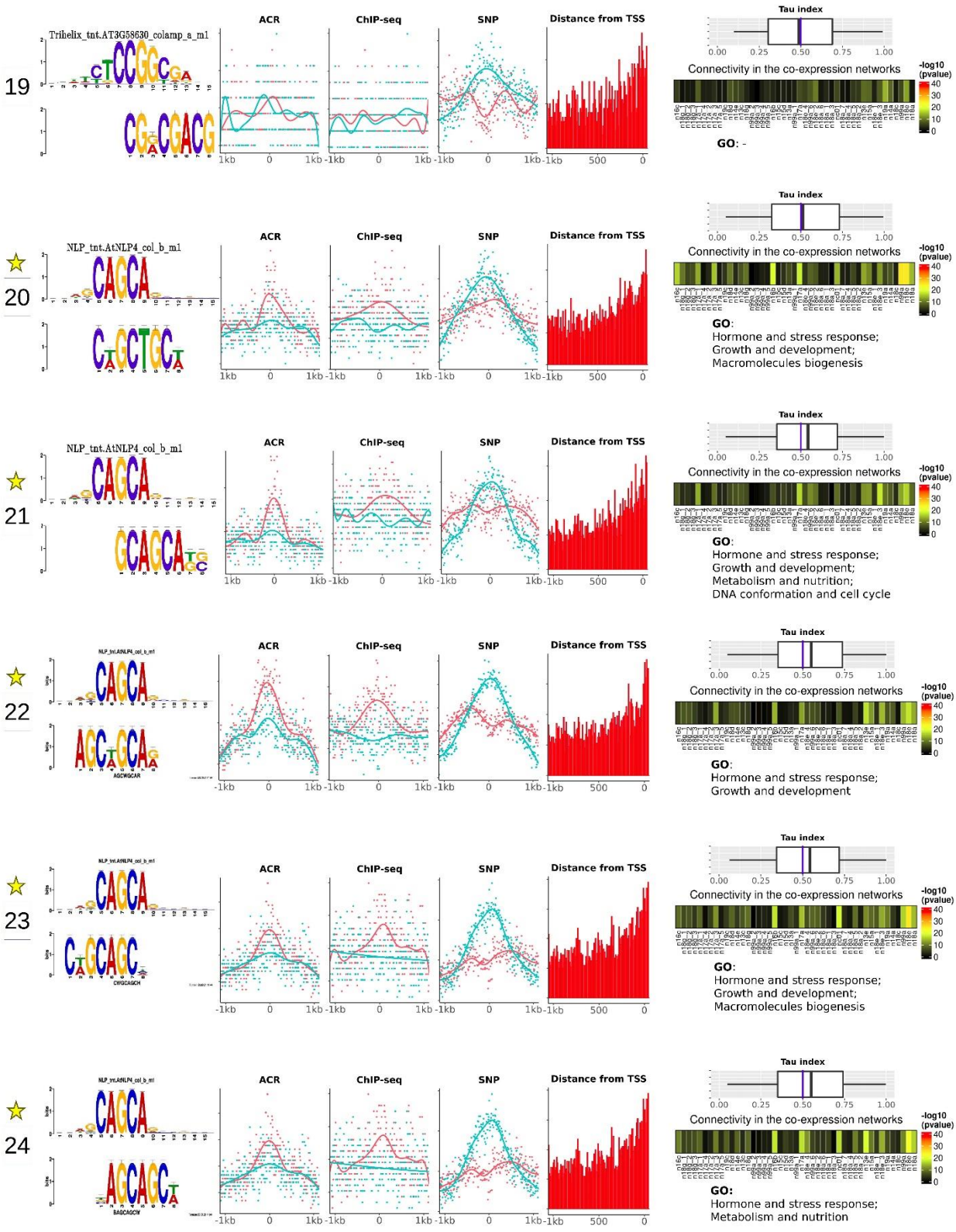


Figure S4.4: Connectivity of gene sets sharing the same random motifs in the coexpression networks. Rows display the different motifs for which a gene set is considered. Columns display the different experimental conditions for which a coexpression network was constructed (sorted by sample size). Entries indicate the -log10 p-values of observing the same connectivity in a coexpression network by chance as was observed for the gene sets that share the same random motif. The colored annotation bar indicates the type of experimental dataset, and the top annotation bar indicates the sample size in the log scale of each experimental dataset.


1
MYBrelated_tnt.TBP3_col_a.m1
ACR
ChIP-seq
SNP
Distance from TSS
Tau index
Connectivity in the co-expression networks
-log10 (pvalue)
GO:
Macromolecule biogenesis;
DNA conformation and cell cycle
2
MYBrelated_tnt.AT1G72740_colamp_a_m1
GO: -
3
bHLH_tnt.bHLH34_col_m1
GO:
Hormone and stress response;
Regulation of metabolism
4
AP2EREBP_tnt.ERF10_colamp_a_m1
GO: -
5
AP2EREBP_tnt.SHN3_col_a_m1
GO:
Metabolism and nutrition;
Macromolecule biogenesis;
DNA conformation and cell cycle
6
bZIP_tnt.bZIP43_col_a_m1
GO:
Hormone and stress response;
Metabolism and nutrition;
DNA conformation and cell cycle


7
AP2EREBP_tnt.AT3G57600_col_A_m1
ACR
ChIP-seq
SNP
Distance from TSS
Tau index
Connectivity in the co-expression networks
-log10 (pvalue)
GO: -
8
AP2EREBP_tnt.ERF6_col_m1
GO:
Macromolecule biogenesis;
Metabolism and nutrition;
DNA conformation and cell cycle
9
AP2EREBP_tnt.CRF4_colamp_a_m1
GO:
Hormone and stress response;
Secondary metabolites;
Metabolism and nutrition
10
AP2EREBP_tnt.CRF4_col_a_m1
GO:
Hormone and stress response;
Regulation of metabolism;
DNA conformation and cell cycle
11
AP2EREBP_tnt.RAP2_col_a_m1
GO:
DNA conformation and cell cycle
12
bZIP_tnt.AREB3_colamp_a_m1
GO:
Growth and development;
Metabolism and nutrition;
Hormone and stress response;


13
bZIP_tnt.bZIP44_colamp_a_m1
14
bZIP_tnt.ABI5_colamp_v3b_m1
15
HMG_tnt.3XHMGBOX1_colamp_a_m1
16
NLP_tnt.AtNLP4_col_b_m1
17
NLP_tnt.AtNLP4_col_b_m1
18
C2H2_tnt.AT2G15740_col_a_m1
ACR
ChIP-seq
SNP
Distance from TSS
-1kb
0
1kb
500
Tau index
0.00
0.25
0.50
0.75
1.00
Connectivity in the co-expression networks
-log10 (pvalue)
40
30
20
10
0
GO:
Hormone and stress response;
Metabolism and nutrition;
Macromolecule biogenesis
GO:
Regulation of metabolism
GO:
Hormone and stress response
GO:
Hormone and stress response;
GO:
Regulation of metabolism;
Growth and development;
GO:
Hormone and stress response;
Regulation of metabolism;
Secondary metabolites;


19
Trihelix_tnt.AT3G58630_colamp_a_m1
ACR
ChIP-seq
SNP
Distance from TSS
Tau index
Connectivity in the co-expression networks
-log10 (pvalue)
GO: -
20
NLP_tnt.AtNLP4_col_b_m1
GO:
Hormone and stress response;
Growth and development;
Macromolecules biogenesis
21
NLP_tnt.AtNLP4_col_b_m1
GO:
Hormone and stress response;
Growth and development;
Metabolism and nutrition;
DNA conformation and cell cycle
22
GO:
Hormone and stress response;
Growth and development
23
GO:
Hormone and stress response;
Growth and development;
Macromolecules biogenesis
24
GO:
Hormone and stress response;
Metabolism and nutrition

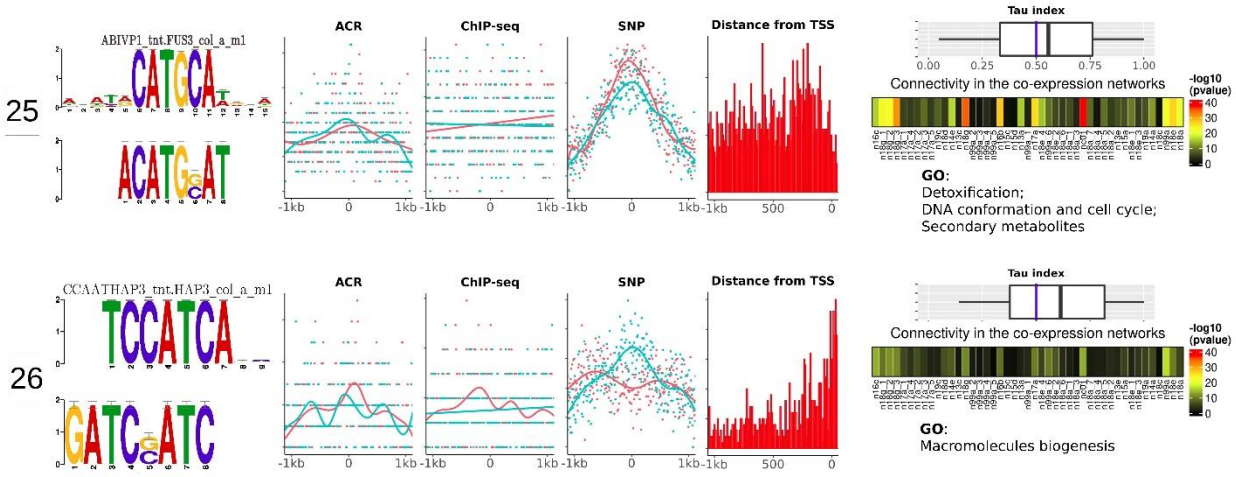
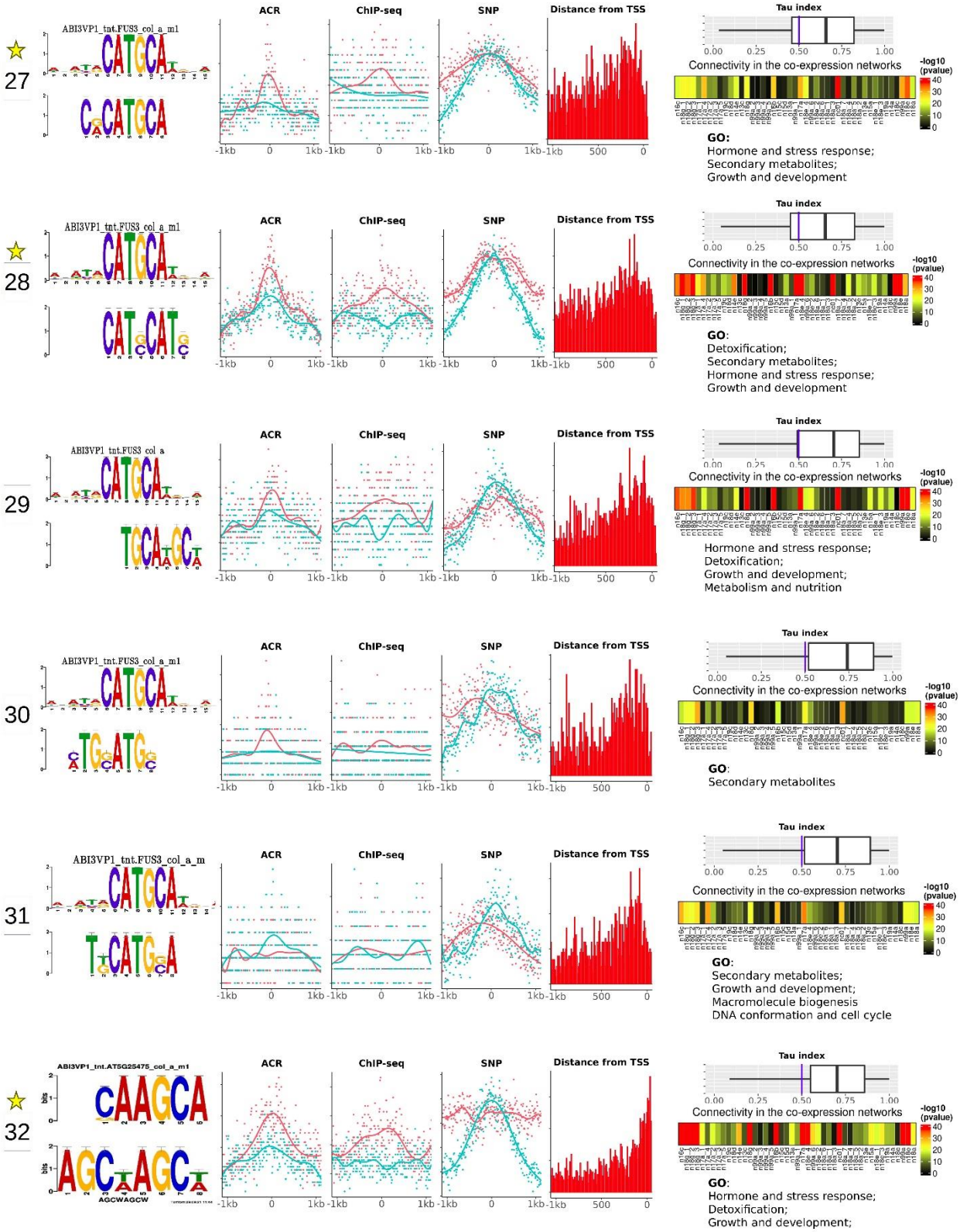


Figure S4.5: Overview of the genomic and functional assessments for predicted motifs that show a high similarity with a corresponding Arabidopsis motif and that are well supported by genomic assessments. Each row represents the results for a motif: left panel: the predicted motif is the bottom motif, and the top motif is the Arabidopsis motif most similar to the predicted motif. The yellow star indicates that the inferred function of the predicted motif is similar to the known function of the TF binding the Arabidopsis motif matching the predicted motif (no visual mismatch). Middle panels: show for the predicted motifs (indicated in red) and the random motifs (blue) the genomic assessments: i) the overlap between the motif instances with open chromatin regions (ACRs) and their surrounding sequences ii) the overlap between motif instances and ChIP-seq TF binding sites and their up-downstream regions, iii) fraction of SNPs occurring in the motif instances and in their up-downstream regions and iv) distance of the location of the motif instances from the TSS. Right panel: shows for the predicted motifs the functional assessments: shows for the genes in *Zea mays* that contain at least one instance of the considered motif i) the distribution of the tissue specificity of their expression (Tau index), ii) their connectivity in coexpression networks representative of different tissues, and iii) their GO enrichment. Motifs are sorted based on the Tau index of their gene sets (showing the genes with the lowest tissue-specific expression first).


27
ABI3VP1_tnt.FUS3_col_a_m1
ACR
ChIP-seq
SNP
Distance from TSS
Tau index
Connectivity in the co-expression networks
-log10 (pvalue)
GO:
Hormone and stress response;
Secondary metabolites;
Growth and development
28
ABI3VP1_tnt.FUS3_col_a_m1
GO:
Detoxification;
Secondary metabolites;
Hormone and stress response;
Growth and development
29
ABI3VP1_tnt.FUS3_col_a
GO:
Hormone and stress response;
Detoxification;
Growth and development;
Metabolism and nutrition
30
ABI3VP1_tnt.FUS3_col_a_m1
GO:
Secondary metabolites
31
ABI3VP1_tnt.FUS3_col_a_m
GO:
Secondary metabolites;
Growth and development;
Macromolecule biogenesis
DNA conformation and cell cycle
32
ABI3VP1_tnt.AT5G25475_col_a_m1
GO:
Hormone and stress response;
Detoxification;
Growth and development;

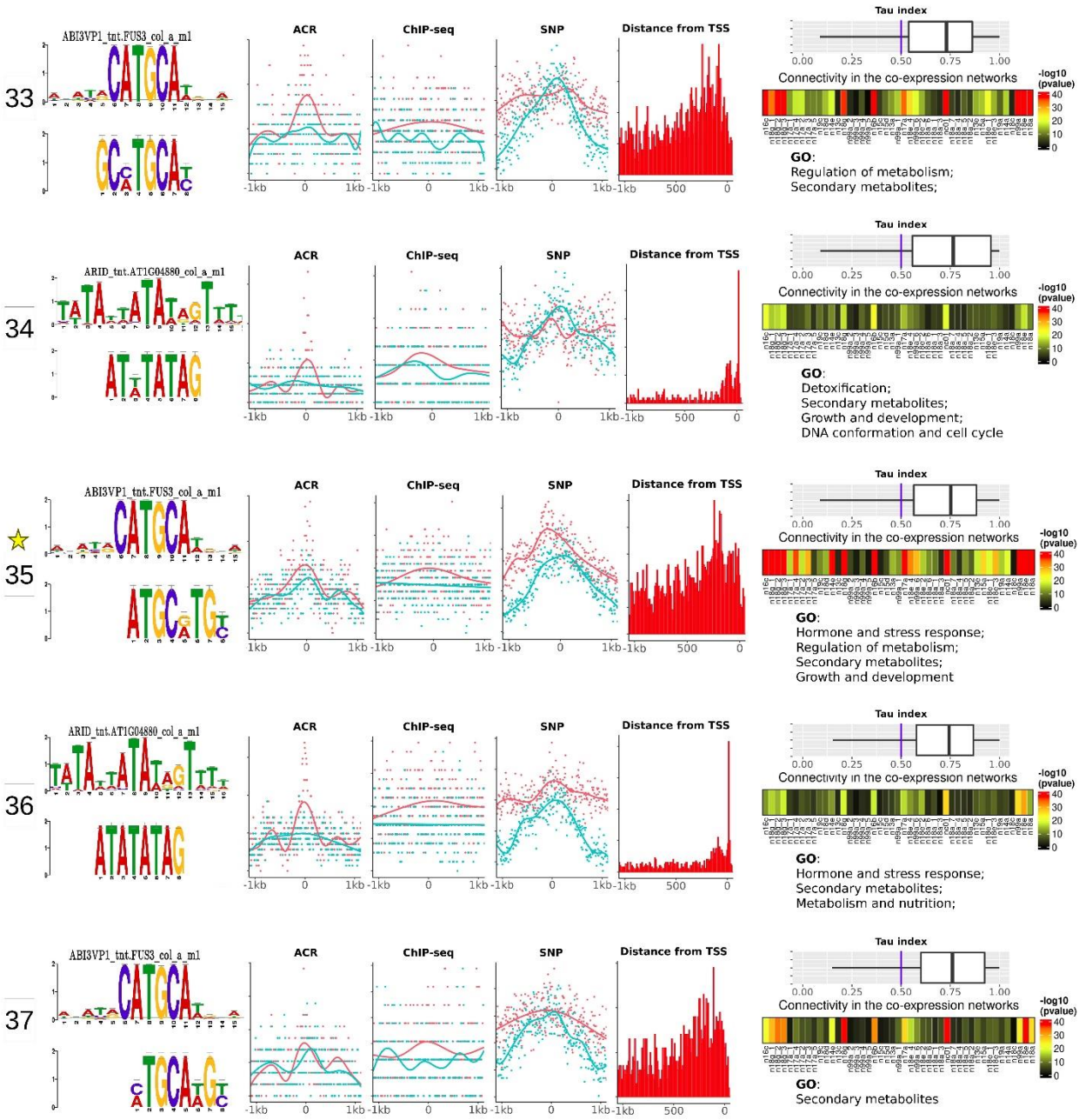
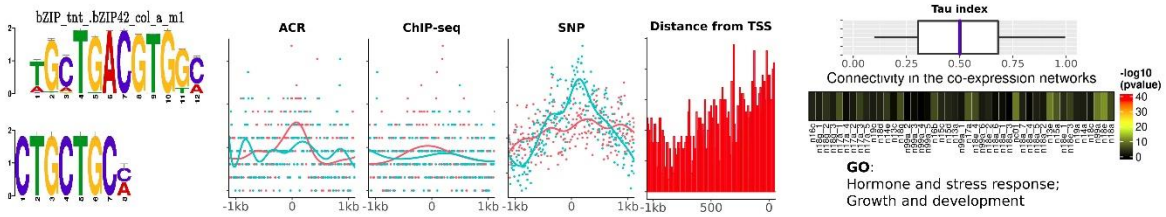
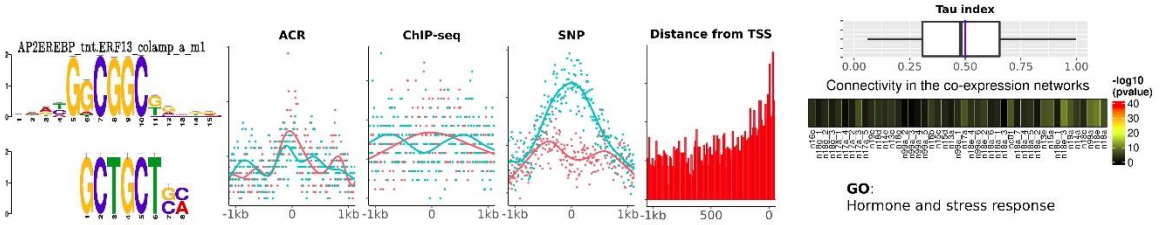


Figure S4.6: Overview of the results for predicted motifs that show a high similarity with a corresponding Arabidopsis motif and that are well supported by functional evidence. Legend is the same as Fig S5.

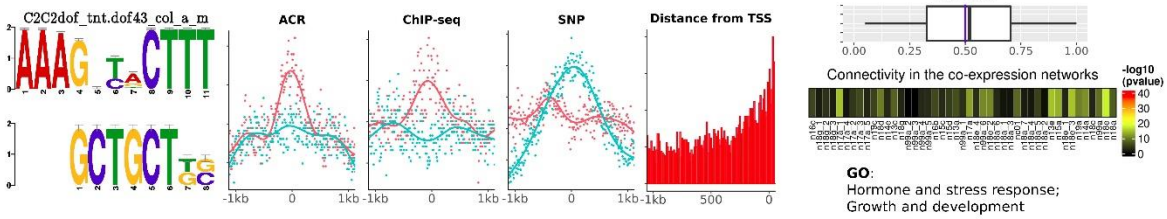
38



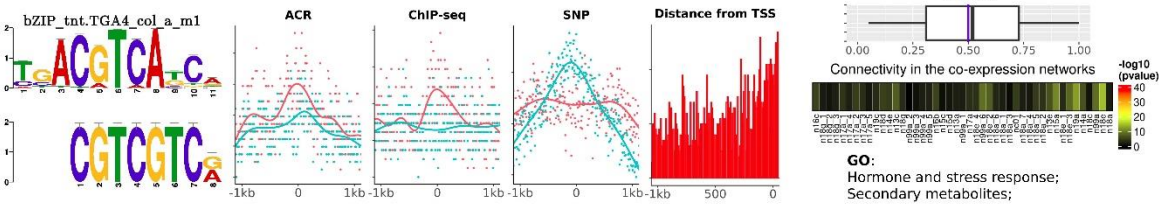
39



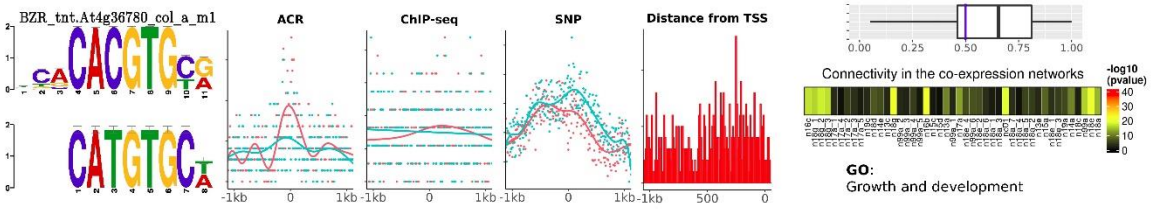
40



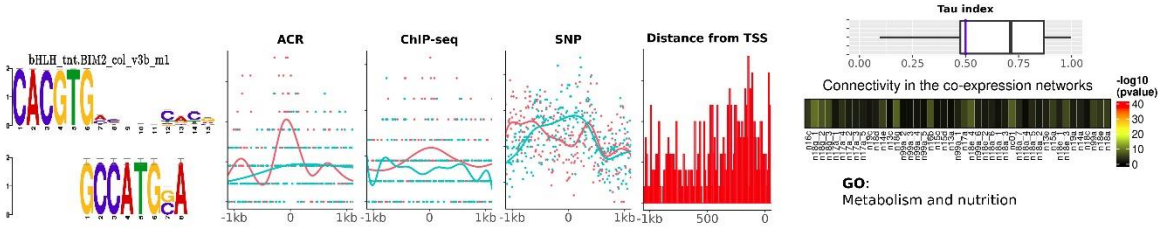
41

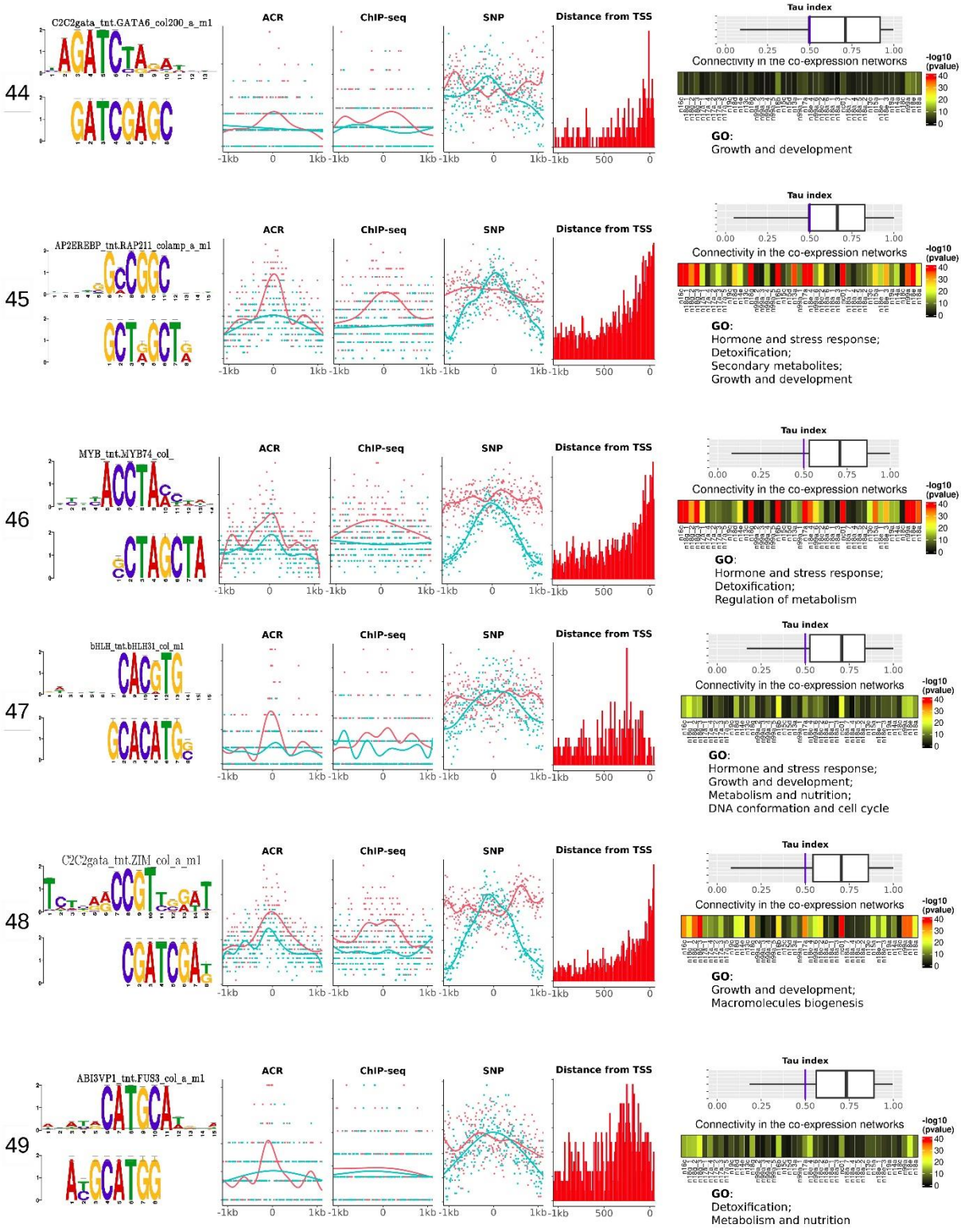


42



43




44
C2C2gata_tnt.GATA6_col200_a_m1
ACR
ChIP-seq
SNP
Distance from TSS
Tau index
Connectivity in the co-expression networks
-log10 (pvalue)
GO:
Growth and development
45
AP2EREBP_tnt.RAP211_colamp_a_m1
GO:
Hormone and stress response;
Detoxification;
Secondary metabolites;
Growth and development
46
MYB_tnt.MYB74_col_
GO:
Hormone and stress response;
Detoxification;
Regulation of metabolism
47
bHLH_tnt.bHLH31_col_m1
GO:
Hormone and stress response;
Growth and development;
Metabolism and nutrition;
DNA conformation and cell cycle
48
C2C2gata_tnt.ZIM_col_a_m1
GO:
Growth and development;
Macromolecules biogenesis
49
ABI3VP1_tnt.FUS3_col_a_m1
GO:
Detoxification;
Metabolism and nutrition

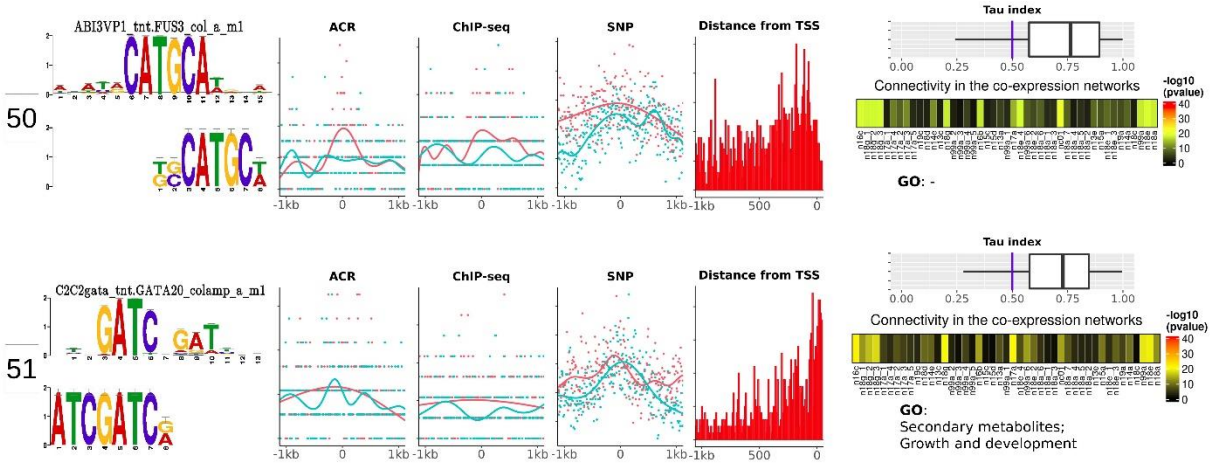


Figure S4.7: Overview of the results for predicted motifs that show no similarity with any of the Arabidopsis motifs described in the study of O'Malley et al 2016, but which are supported by genomic and/or functional assessments. These are promising novel motif predictions. Legend is the same as Fig S5.

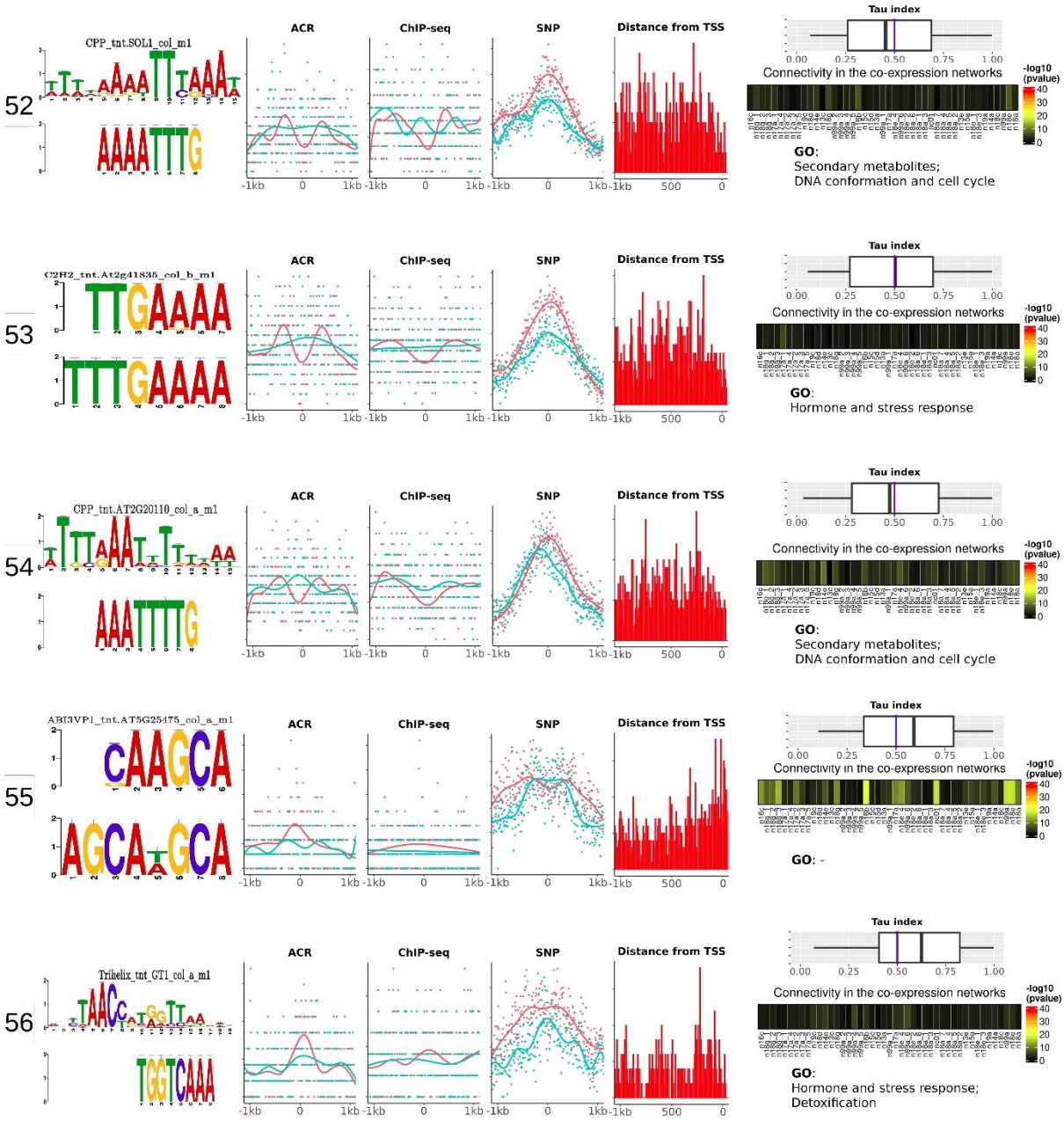


Figure Figure S4.8: Overview of the results for predicted motifs that show a high similarity with Arabidopsis motifs, but of which the instances are not supported by any of the genomics nor by the expression support. These likely represent spurious motifs despite the match in Arabidopsis. Legend as in Fig S5.

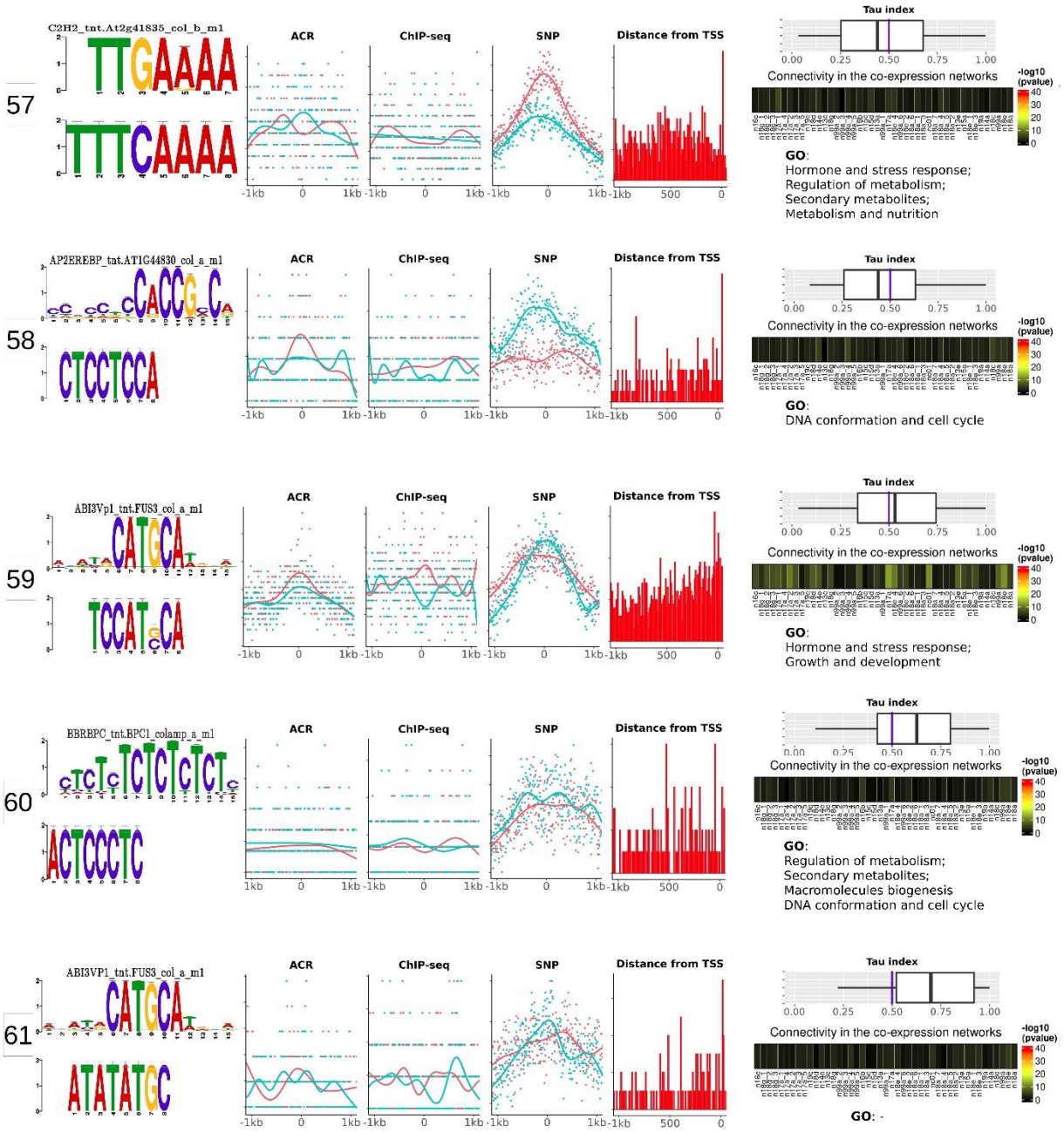


Figure Figure S4.9: Overview of the results for predicted motifs that no significant similarity with any of the Arabidopsis motifs and which are not supported by any of the genomic assessments nor by the functional assessments. These might represent spurious motifs. Legend as in Fig S5.

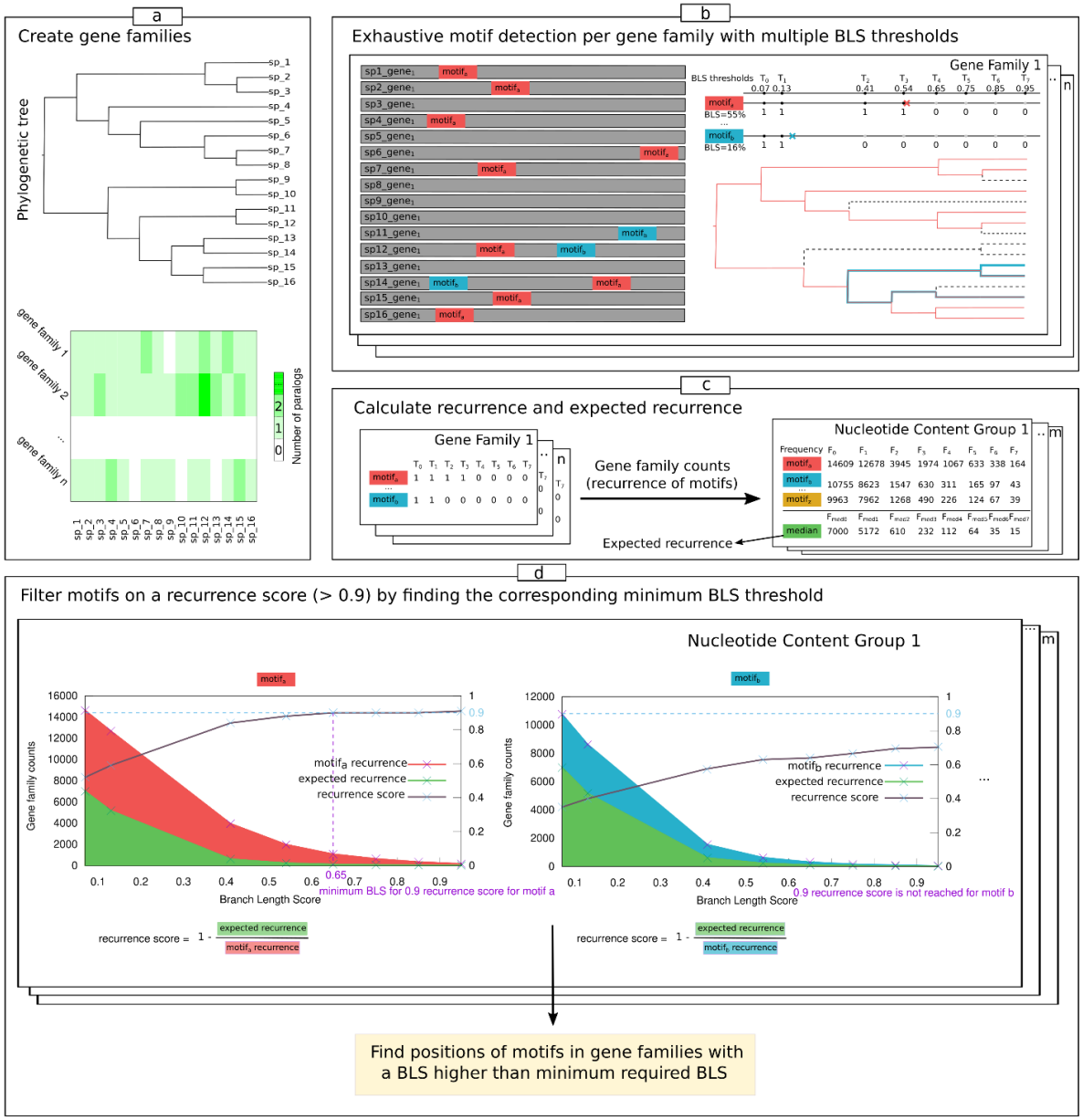


Figure S4-10: BLSSpeller workflow: Panel a: BLSSpeller requires as input gene families, consisting of related orthologs and paralogs obtained from a preselected set of species. The heatmap below the phylogenetic tree indicates that each species can contribute zero, one, or more genes to each gene family. **Panel b:** BLSSpeller will then search for motifs conserved in the upstream sequences of the genes belonging to a particular gene family. Hereto, for each length-*k* motif, it finds in which promoter sequences (and hence: species) an instance of that motif appears. Here, two motifs are highlighted in green and red. The degree of conservation of the motif within the gene family is expressed by the Branch Length Score (BLS). It is calculated by finding, in the phylogenetic species tree, the minimum spanning tree that connects the relevant subset of species and summing the weights of the horizontal branches in that tree. If the motif appears in multiple paralogs of the same species, the branch length to that species is accounted for only once. If a gene family does not contain a representative ortholog in each of the considered species, we delete the branches in the species tree corresponding to the missing species and

rescale the branch lengths so that the sum of the weights on the branches of the tree again amounts to 1 The branch lengths relevant to respectively motif *a* and *b* are indicated in red and green. Dashed lines indicate omitted branch lengths. A motif for which the BLS exceeds a predefined BLS threshold in a gene family is said to be *conserved* in that gene family. We use multiple BLS thresholds (i.e., T = 0.07, 0.13, 0.41, 0.54, 0.65, 0.75, 0.85 and 0.95). **Panel c:** calculation of the recurrence and expected recurrence of a motif. For each gene family, a binary matrix is constructed where rows represent motifs and columns indicate whether the motif meets the indicated predefined BLS threshold. The matrices of all gene families are aggregated into a single matrix, where each matrix element now corresponds to the frequency (F) with which a certain motif was found conserved at a certain BLS threshold across gene families or said otherwise recurs across gene families. To calculate the expected recurrence of a motif, we group all motifs with the same nucleotide content and refer to these as a "nucleotide content group". We extend this list by adding a count of 0 for all motif permutations that are not present. Next, per nucleotide content group and per BLS threshold, the expected recurrence is computed as the median value among all motifs in the nucleotide content group. **Panel d:** filter motifs on a recurrence score. Motifs that meet the predefined recurrence score (0.9) are retained and their positions in gene families with a BLS higher than minimum required BLS are extracted for each species. The minimum required BLS for motif *a* is 0.65 while motif *b* does not meet the recurrence sore of 0.9 at any of the considered BLS thresholds.

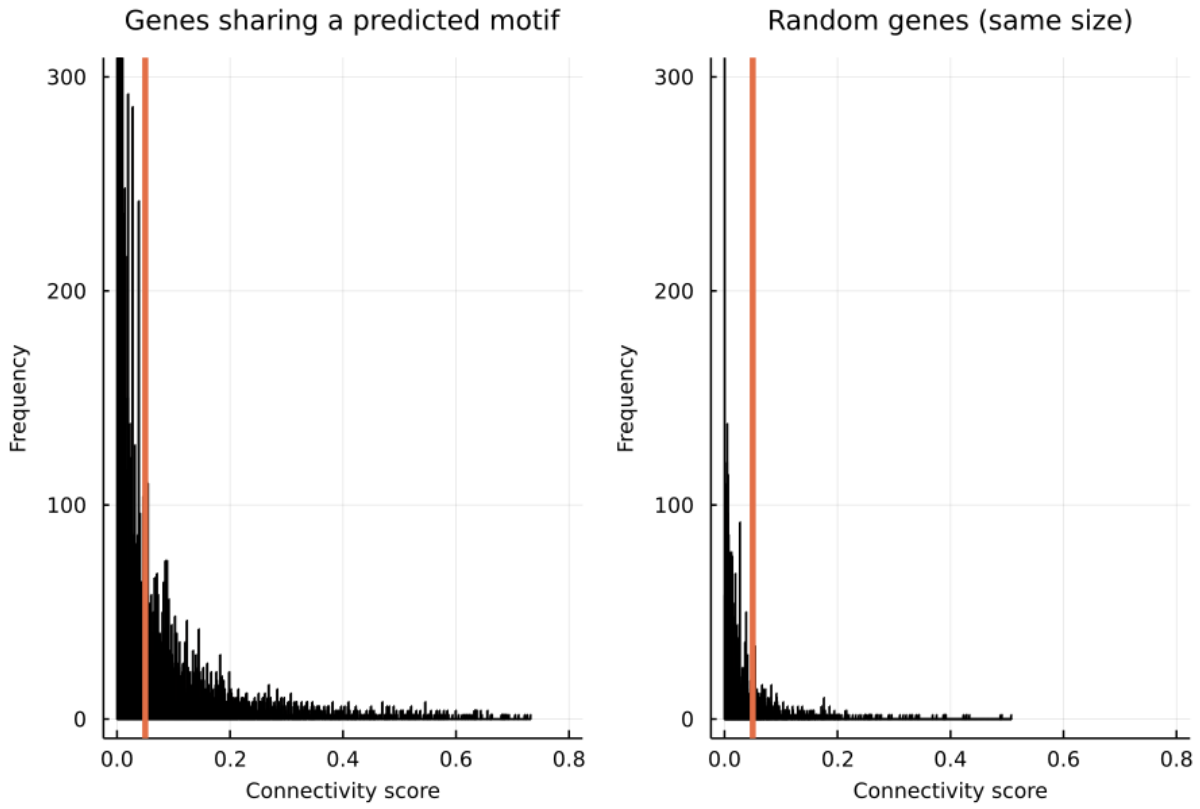


Figure S4.11: Distribution of the pairwise connectivity scores (normalized Katz index) derived for a representative coexpression network (n18a, see Table S3) for gene pairs that share a predicted motif (left) and gene pairs that share a random motif (right). For both plots, the same number of gene pairs was used to make the distributions comparable. The vertical red line indicates the threshold of 0.05 used in this study. The majority of the normalized Katz indices of gene pairs with random motifs fall below 0.05 with a strong peak at zero. The y axis is capped at 300.

Table S4.1: Details of monocot species used in this study. General characteristics of the genomes used in the comparative approach.

Id	Scientific name	Total number of genes	Number of Chr (2n)	Genome size	Reference to the genome sequence
bdi	Brachypodium distachyon	34,310	10	355 Mb	https://pubmed.ncbi.nlm.nih.gov/20148030/
cam	Cenchrus americanus	38,579	14	1.7 Gb	https://pubmed.ncbi.nlm.nih.gov/28922347/
hvu	Hordeum vulgare	43,050	14	4.98 Gb	https://pubmed.ncbi.nlm.nih.gov/23075845/
obr	Oryza brachyantha	34,155	24	260 Mb	https://pubmed.ncbi.nlm.nih.gov/23518283/
osa	Oryza sativa ssp. japonica	42,189	24	400 Mb	https://pubmed.ncbi.nlm.nih.gov/16100779/
osaindica	Oryza sativa ssp. indica	38,714	24	400 Mb	https://pubmed.ncbi.nlm.nih.gov/16100779/
oth	Oropetium thomaeum	28,446	18	244 Mb	https://pubmed.ncbi.nlm.nih.gov/26560029/
ped	Phyllostachys edulis	31,987	48	2.05 Gb	https://pubmed.ncbi.nlm.nih.gov/23435089/
sbi	Sorghum bicolor	34,211	20	730 Mb	https://pubmed.ncbi.nlm.nih.gov/19189423/
sit	Setaria italica	34,584	18	423 Mb	https://pubmed.ncbi.nlm.nih.gov/22580950/
ssp	Saccharum spontaneum	83,826	40	2.56-Gbp	https://pubmed.ncbi.nlm.nih.gov/30297971/
tae	Triticum aestivum	107,891	42	17 Gb	https://pubmed.ncbi.nlm.nih.gov/25035500/
ttu	Triticum turgidum	63,993	28	10.45 Gb	https://pubmed.ncbi.nlm.nih.gov/30962619/
zjn	Zoysia japonica ssp. nagirizaki	59,271	40	334 Mb	https://pubmed.ncbi.nlm.nih.gov/26975196/
zma	Zea mays B73	44,474	20	3.2 Gb	https://pubmed.ncbi.nlm.nih.gov/19965430/
Zma-ph207	Zea mays PH207	40,557	20	2.1 Gb	https://pubmed.ncbi.nlm.nih.gov/27803309/

Table S4.2: Functional similarity between the predicted motifs and their matching counterparts in Arabidopsis. The numbers on the left indicate the motif ID as assigned in Fig S5-9. Highlighted in yellow are the motifs for which a clear functional similarity could be detected between the maize and the Arabidopsis counterparts.

Num	Predicted motif	Arabidopsis motif	Function in Arabidopsis	GO enriched in *Zea mays*	adjusted p-values
53	TTTGAAAA	C2H2_tnt.At2g41835	stress response	glutamate receptor signaling pathway	4.3152e-06
5	CCGCCKCC	AP2EREBP_tnt.SHN3	ethylene responsive	protein-DNA complex assembly	1.4384e-07
3	CACGTGTC	bHLH_tnt.bHLH34_col_m1	stress response	regulation of abscisic acid signaling	5.394e-11
16	GCAGCARY	NLP_tnt.AtNLP4	nitrate-responsive	response to karrikin	0.0025
7	CCACCGCC	AP2EREBP_tnt.AT3G57600	response to drought	no enrichment	-
21	GCAGCAKS	NLP_tnt.AtNLP4	nitrate-responsive	regulation of developmental growth	0.00073718
4	GGCGGCGS	AP2EREBP_tnt.ERF10	ethylene response	no enrichment	-
13	CACGTGGC	bZIP_tnt.AREB3	ABA responsive	response to heat	0.0186459
13	GACGTGGC	bZIP_tnt.bZIP44	seed germination	response to light stimulus	4.36015e-05
6	CCACGTCA	bZIP_tnt.bZIP43	protein binding	protein-DNA complex assembly	3.596e-18
22	AGCWGCAR	NLP_tnt.AtNLP4	nitrate-responsive	regulation of developmental growth	2.9421818181 8182e-06
8	CGYCGCCR	AP2EREBP_tnt.ERF6	ethylene response	monocarboxylic acid transport	6.3e-07 0.002 0677
11	CGGYGGYG	AP2EREBP_tnt.RAP21	ethylene response	nucleosome organization	0.0052142
19	CGRCGACG	Trihelix_tnt.AT3G58630	protein binding	no enrichment	-
14	ACGTGGCG	bZIP_tnt.ABI5_colamp_v3b	ABA signaling	regulation of cellular amino acid metabolism	0.00085405
9	GCGGCRGC	AP2EREBP_tnt.CRF4	ethylene response	ethylene-activated signaling pathway	0.00157873
24	BAGCAGCW	NLP_tnt.AtNLP4	nitrate-responsive	nicotianamine metabolic process	4.5720571428 5714e-05
15	CCACCACC	HMG_tnt.3XHMGBOX1	(high mobility group)-box	response to brassinosteroid	3.596e-08
10	GCCRCCRC	AP2EREBP_tnt.CRF4	ethylene response, leaf development	regulation of leaf morphogenesis	0.0039556
37	MTGCAWGY	ABI3VP1_tnt.FUS3	Mitogen-activated protein kinase	plant-type cell wall organization	0.046748

Table S4.3: Details of the 45 expression datasets used for coexpression analysis.

Name	type	note	reference	samples
n16b	Tissue	B73	[2]	93
n16c	Tissue	B74	[3]	23
n14e	Tissue	B75	[4]	53
n19c	Tissue	seed dev	[5]	31
nc01	Tissue	combined	[2-6]	247
n13c	Genotype	seedling_leaf3	[7]	62
n13e	Genotype	kernel	[8]	368
n14a	Genotype	seedling	[9]	503
n15a	Genotype	SAM	[10]	383
n18d	Genotype	root_GCN	[11]	48
n19a	Genotype	seedling	[12]	453
n17a	Tissue*Genotype	5 tissues	[13]	133
n17a_1	Genotype	ear	[13]	26
n17a_2	Genotype	root	[13]	27
n17a_3	Genotype	shoot	[13]	27
n17a_4	Genotype	tassel	[13]	26
n17a_5	Genotype	SAM	[13]	27
n18a	Tissue*Genotype	7 tissues	[14]	1657
n18a_1	Genotype	Groot	[14]	201
n18a_2	Genotype	Gshoot	[14]	271
n18a_3	Genotype	kernel	[14]	226
n18a_4	Genotype	L3Base	[14]	254
n18a_5	Genotype	L3Tip	[14]	257
n18a_6	Genotype	LMAD	[14]	199
n18a_7	Genotype	LMAN	[14]	249
n18e	Tissue*Genotype	4 tissues	[15]	1136
n18e_1	Genotype	leaf	[15]	394
n18e_2	Genotype	root	[15]	176
n18e_3	Genotype	SAM	[15]	406
n18e_4	Genotype	seed	[15]	159
n99a	Tissue*Genotype	6 tissues	[16]	620
n99a_1	Genotype	endosperm	[16]	121
n99a_2	Genotype	internode	[16]	77
n99a_3	Genotype	leaf	[16]	84
n99a_4	Genotype	root	[16]	84
n99a_5	Genotype	shoot	[16]	85
n99a_6	Genotype	seedling	[16]	169
n18g	Tissue*Genotype	B+M+F1	[6]	73
n18g_1	Tissue	B73	[6]	23
n18g_2	Tissue	Mo17	[6]	23
n18g_3	Tissue	BxM	[6]	23
n13a	RIL	B73 x Mo17	[17]	107

n15c	RIL	MAGIC	[18]	102
n15d	RIL	B73 x H99	[19]	106
n18c	RIL	W22 x Teosinte	[20]	617

References

1. Gupta S, Stamatoyannopoulos J, Bailey T, Stafford W: **Quantifying similarity between motifs. Genome Biology.** vol; 2007.
2. Stelpflug SC, Sekhon RS, Vaillancourt B, Hirsch CN, Buell CR, de Leon N, Kaeppler SM: **An expanded maize gene expression atlas based on RNA sequencing and its use to explore root development.** *The plant genome* 2015, **9**.
3. Walley JW, Sartor RC, Shen Z, Schmitz RJ, Wu KJ, Urich MA, Nery JR, Smith LG, Schnable JC, Ecker JR: **Integration of omic networks in a developmental atlas of maize.** *Science* 2016, **353:**814-818.
4. Chen J, Zeng B, Zhang M, Xie S, Wang G, Hauck A, Lai J: **Dynamic transcriptome landscape of maize embryo and endosperm development.** *Plant physiology* 2014, **166:**252-264.
5. Yi F, Gu W, Chen J, Song N, Gao X, Zhang X, Zhou Y, Ma X, Song W, Zhao H: **High temporal-resolution transcriptome landscape of early maize seed development.** *The Plant Cell* 2019, **31:**974-992.
6. Zhou P, Hirsch CN, Briggs SP, Springer NM: **Dynamic patterns of gene expression additivity and regulatory variation throughout maize development.** *Molecular plant* 2019, **12:**410-425.
7. Eichten SR, Briskine R, Song J, Li Q, Swanson-Wagner R, Hermanson PJ, Waters AJ, Starr E, West PT, Tiffin P: **Epigenetic and genetic influences on DNA methylation variation in maize populations.** *The Plant Cell* 2013, **25:**2783-2797.
8. Fu J, Cheng Y, Linghu J, Yang X, Kang L, Zhang Z, Zhang J, He C, Du X, Peng Z: **RNA sequencing reveals the complex regulatory network in the maize kernel.** *Nature communications* 2013, **4:**1-12.
9. Hirsch CN, Foerster JM, Johnson JM, Sekhon RS, Muttoni G, Vaillancourt B, Peñagaricano F, Lindquist E, Pedraza MA, Barry K: **Insights into the maize pan-genome and pan-transcriptome.** *The Plant Cell* 2014, **26:**121-135.
10. Leiboff S, Li X, Hu H-C, Todt N, Yang J, Li X, Yu X, Muehlbauer GJ, Timmermans MC, Yu J: **Genetic control of morphometric diversity in the maize shoot apical meristem.** *Nature communications* 2015, **6:**1-10.
11. Schaefer RJ, Michno J-M, Jeffers J, Hoekenga O, Dilkes B, Baxter I, Myers CL: **Integrating coexpression networks with GWAS to prioritize causal genes in maize.** *The Plant Cell* 2018, **30:**2922-2942.
12. Mazaheri M, Heckwolf M, Vaillancourt B, Gage JL, Burdo B, Heckwolf S, Barry K, Lipzen A, Ribeiro CB, Kono TJ: **Genome-wide association analysis of stalk biomass and anatomical traits in maize.** *BMC plant biology* 2019, **19:**1-17.
13. Lin H-y, Liu Q, Li X, Yang J, Liu S, Huang Y, Scanlon MJ, Nettleton D, Schnable PS: **Substantial contribution of genetic variation in the expression of transcription factors to phenotypic variation revealed by eRD-GWAS.** *Genome biology* 2017, **18:**1-14.
14. Kremling KA, Chen S-Y, Su M-H, Lepak NK, Romay MC, Swarts KL, Lu F, Lorant A, Bradbury PJ, Buckler ES: **Dysregulation of expression correlates with rare-allele burden and fitness loss in maize.** *Nature* 2018, **555:**520-523.
15. Huang J, Zheng J, Yuan H, McGinnis K: **Distinct tissue-specific transcriptional regulation revealed by gene regulatory networks in maize.** *BMC plant biology* 2018, **18:**1-14.
16. Li Z, Zhou P, Della Coletta R, Zhang T, Brohammer AB, Vaillancourt B, Lipzen A, Daum C, Barry K, de Leon N: **Highly genotype-and tissue-specific single-parent expression drives dynamic gene expression complementation in maize hybrids.** *bioRxiv* 2019**:**668681.

17. Li L, Petsch K, Shimizu R, Liu S, Xu WW, Ying K, Yu J, Scanlon MJ, Schnable PS, Timmermans MC: **Mendelian and non-Mendelian regulation of gene expression in maize.** *PLoS genetics* 2013, **9:**e1003202.
18. Baute J, Herman D, Coppens F, De Block J, Slabbinck B, Dell'Acqua M, Pè ME, Maere S, Nelissen H, Inzé D: **Combined large-scale phenotyping and transcriptomics in maize reveals a robust growth regulatory network.** *Plant Physiology* 2016, **170:**1848-1867.
19. Baute J, Herman D, Coppens F, De Block J, Slabbinck B, Dell'Acqua M, Pè ME, Maere S, Nelissen H, Inzé D: **Correlation analysis of the transcriptome of growing leaves with mature leaf parameters in a maize RIL population.** *Genome biology* 2015, **16:**1-26.
20. Wang X, Chen Q, Wu Y, Lemmon ZH, Xu G, Huang C, Liang Y, Xu D, Li D, Doebley JF: **Genome-wide analysis of transcriptional variability in a large maize-teosinte population.** *Molecular Plant* 2018, **11:**443-459.

Curriculum vitae

Education:

P.h.D in Bioinformatics, Department of Plant Biotechnology and Bioinformatics, Ghent University, Belgium (2017-2022).

MS in Plant Biotechnology, Agronomy and Plant breeding Department, University of Tehran, Iran (2011-2014).

BSc in Plant Breeding, Department of Agronomy and Plant Breeding, Urmia University, Iran (2007-2011).

List of publications:

Published

Tao Shi, **Razgar Seyed Rahmani**, Paul F. Gugger, Muhua Wang, Hui Li, Yue Zhang, Zhizhong Li, Qingfeng Wang, Yves Van de Peer, Kathleen Marchal, and Jinming Chen. Distinct expression and methylation patterns for genes with different fates following a single whole-genome duplication in flowering plants. *Molecular biology and evolution*, 8 (2020): 2394-2413.

Razgar Seyed Rahmani†, Tao Shi†, Dongzhi Zhang, Xiaoping Gou, Jing Yi, Giles Miclotte, Kathleen Marchal, and Jia Li. "Genome-wide expression and network analyses of mutants in key brassinosteroid signaling genes." *BMC Genomics* 22, 1 (2021): 1-17.

Zhang Yue, **Razgar Seyed Rahmani**, Xingyu Yang, Jinming Chen, and Tao Shi. "Integrative expression network analysis of microRNA and gene isoforms in sacred lotus." *BMC Genomics* 21, 1 (2020): 1-13.

Under review

Razgar Seyed Rahmani†, Dries Decap†, Jan Fostier, Kathleen Marchal, BLSSpeller: discover novel regulatory motifs in maize, *DNA Research*, (2022).

Taher Mohasseli†, **Razgar Seyed Rahmani**†; Reza Darvishzadeh; Sara Dezhsetan; Kathleen Marchal, De novo transcriptome assembly of two maize genotypes identified pathways associated with differences in salt tolerance, Cereal Research Communications, (2022).

†: Same contribution.

Poster presented at international conferences

Razgar Seyed Rahmani, Tao Shi, Dongzhi Zhang, Xiaoping Gou, Jing Yi, Kathleen Marchal, Jia Li. Integrate expression and network analyses to unveil functional features of brassinosteroid signaling in Arabidopsis mutant lines. Intelligent System Biology for Molecular Biology ISMB (2021).

Razgar Seyed Rahmani, Dries Decap, Jan Fostier, Kathleen Marchal. BLSSpeller to discover novel regulatory motifs in maize. Plant Genomics and Gene Editing Congress: Europe (2022).

Courses followed during Ph.D.

Title	Code	Teacher	Organizers	Date/period	Credits
Predicting Modeling		Willem Waegeman	Ugent	2018-2019	16
Selected Topics in Mathematical Optimization		Michiel Stock	Ugent	2019-2020	16
Analysis of High Dimensional Data		Olivier Thas	Ugent	2020-2021	13
Applied High-throughput Analysis		De Meyer, Tim Jo Vandesompele	Ugent	2020-2021	15
Molecular Plant Breeding		Haesaert, Geert	Ugent	2020-2021	15