UNIVERSITEIT GENT

FACULTY OF SCIENCES

A Functional and Regulatory Perspective on *Arabidopsis thaliana*

Ken Heyndrickx

Promoter: Prof. Dr. Klaas Vandepoele

Ghent University
Faculty of Sciences
Department of Plant Biotechnology and Bioinformatics
VIB Department of Plant Systems Biology
Comparative and Integrative Genomics

Research funded by a PhD grant of the Institute for the Promotion of Innovation through Science and Technology in Flanders (IWT Vlaanderen).

Dissertation submitted in fulfilment of the requirements for the degree of Doctor in Sciences:Bioinformatics.

Academic year: 2014-2015

Examination Commitee

Prof. Dr. Geert De Jaeger (chair)

Faculty of Sciences, Department of Plant Biotechnology and Bioinformatics, Ghent University

Prof. Dr. Klaas Vandepoele (promoter)

Faculty of Sciences, Department of Plant Biotechnology and Bioinformatics, Ghent University

Prof. Dr. Detlef Weigel

Department of Molecular Biology, Max Planck Institute for Developmental Biology, 72076 Tübingen, Germany

Prof. Dr. Stein Aerts

Laboratory of Computational Biology, Center for Human Genetics, University of Leuven, Leuven 3000, Belgium

Prof. Dr. Kathleen Marchal

Department of Plant Biotechnology and Bioinformatics, Faculty of Sciences, Ghent University, B-9000 Ghent, Belgium

Prof. Dr. Tim De Meyer

Lab of Bioinformatics and Computational Genomics, Department of Mathematical Modeling, Statistics and Bioinformatics, Faculty of Bioscience Engineering, Ghent University, Ghent, Belgium

Prof. Dr. Pieter de Bleser

Inflammation Research Center, Flanders Institute of Biotechnology (VIB) and Department of Biomedical Molecular Biology, Ghent University, Ghent, Belgium

Thank You!

The process of a PhD is hardly a solitary process, and therefore there are a number of people I would like to thank.

First of all, I would like to thank Klaas Vandepoele, for having a great enthusiasm in guiding students, which made it a a self-evidence to continue my scientific career in his group after my Master's thesis under his supervision. His excellent guidance and scientific drive have taught me greatly. I would like to particularly thank him for having the generosity of letting his PhD student speak at the CSHL conference in New York, rather than wanting to speak himself.

This turned out to be an important point in my PhD, as this led to the wonderful collaboration I had with Detlef Weigel, whom I thank for being the kind of person that reaches out to young PhD students. My three month-stay in Tübingen was a wonderful experience, and one I will cherish. In particular, I would like to thank the bioinformatics group Stefan, Jonas and Euncheon for the many nice lunch conversations; Nacho for the interesting ChIP conversations; Rebecca, Effy and Cris for being fun; Damian and Jörg for all the World Championship fanatics (Belgium was only practising in Brasil, in France 2016 and Qatar 2018 we mean business); and all of X World for the great atmosphere.

Closer to home, I would like to thank everyone in the BEG group for four fun years, where the 3 o'clock breaks of the fellowship of the lounge have been a wonderful antidote against stress. Particular acknowledgements go to Lieven, for having to deal with my humour and making the best brownies in the world; Jan, my publishing partner in crime for being the perfect companion for fun, science and conferences; and Bram, with whom I shared the entire PhD experience from the start, too many fun moments to remember, stress-breaking walks, and last but not least: ribbekes from the Ribhouse.

Everyone knows you can't achieve anything in life if not given the opportunity, so I would like to thank my parents for supporting me throughout my life and setting me up with all these opportunities. I would also like to thank my brother Sven, who has driven my by not making it easy to prove that I am in fact, the smarter one. :-p

Finally, I would like to thank my best friend of almost 12 years without whom I could not picture having gone through this all: my wife Stefanie. Thank you for being the supportive and loving person you are, even when my mood isn't optimal.

Samenvatting

Genexpressie is een dynamisch proces, dat voortdurend beïnvloed wordt door signalen van binnen —en buitenaf. Het doel van dit project bestond uit twee delen. Als eerste trachtten we te bestuderen hoe regulatie en expressie georganiseerd zijn in het genoom en in de context van functioneel geassocieerde genmodulen. Als tweede doel hadden wij voor ogen deze data aan te wenden in de context van functiepredictie van genen in Arabidopsis met een tot op heden ongekende functie.

We begonnen met een systematische identificatie van functionele genmodulen (i.e. sets van genen die geassocieerd zijn met elkaar op basis van een biologische eigenschap) en het bestuderen van de onderliggende regulatorische en expressiecomponent. De modulen werden afgelijnd op basis van grootschalige expressiedata, functionele gen-annotatie, experimenteel geverifieerde eiwit-eiwit interacties en regulatorische TF-doelwitgeninteracties. De beperkte overlap tussen de verschillende geselecteerde experimentele inputdata bevestigde het voordeel van het combineren van verschillende datatypes. De systematische identificatie resulteerde in 1 562 modulen, die 13 142 genen omvatten. De meeste modulen toonden een significant niveau van co-expressie, maar de *cis*-regulatorische en functionele coherentie was beperkt. Hub genen werden significant geassocieerd bevonden met letaliteit in de embryofase en boden bewijs van *crosstalk* tussen verschillende biologische processen. Om na te gaan in welke mate de co-expressiecomponent van de modulen bewaard was over verschillende species heen werden de modulen overgezet naar andere species op basis van orthologie. Op basis van de expressie data van de verschillende species werd aangetoond dat 58% van de modulen een significante conservatie vertoonde van co-expressie. Naast het bestuderen van regulatie hebben we de modulen ook aangewend in de context van functiepredictie. Op basis van modulen konden 5 562 genen geannoteerd en geëvalueerd worden op basis van nieuw beschikbare experimentele functionele annotaties. Voor 197 genen die sinds de start van onze analyse een nieuwe functie hadden toegewezen gekregen kon 38,1% voorspeld worden in de modulecontext. Voorspelde functies vallen binnen de domeinen van celwand biogenese, xyleem en floëem, celcyclus, hormoonsignalisatie en circadiane ritmes. Globaal gezien werden hypotheses gegenereerd voor respectievelijk 1 701 en 43 624 functioneel ongekende genen in Arabidopsis en zes andere plantenspecies.

Vervolgens hadden we als doel om de genomische organisatie van transcriptionele regulatie te bestuderen in Arabidopsis. Publiek beschikbare data van ChIP experimenten voor 27 TF's werden geselecteerd. Alle experimenten werden gereanalyseerd op een uniforme wijze om zo onderlingen compatibiliteit te garanderen. Dit resulteerde in een netwerk van 15 188 potentiële doelwitgenen, verbonden door 46 619 potentiële regulatorische interacties. Op basis van de gecombineerde bindingsprofielen werden hub genen en *highly occupied target* regions geïdentificeerd. Binnen het onderzochte netwerk waren deze genen significant geassocieerd met ontwikkeling, stimulussignalisatie en gen regulatorische processen. Met de controverse omtrent de functionaliteit van HOT regio's in het biomedische veld in het achterhoofd hebben we meerdere analyses uitgevoerd om aan te tonen dat zij in planten weldegelijk functionele binding bevatten. HOT regio's bevatten aangerijkte DNA motieven, zijn aangerijkt voor differentieel geëxpresseerde genen, en zijn vaak geconserveerd binnen de cruciferen en de dicotyle planten. Een andere set van atypische gebonden regio's zijn deze die op een grote afstand ($<$ 4kb) liggen van hun dichtstbijzijnde gen. Net als de andere gebonden regio's zijn zij onder de invloed van negatieve selectie. Daarenboven zijn zij aangerijkt voor een chromatinestaat die geassocieerd is met het Polycomb repressieve complex. Het aantal bindingen in de nabijheid van een gen bleek gelinkt te zijn aan de breedte van expressie van

het gen in kwestie. Hypothesen betreffende co-binding en tethering tussen TF's betrokken bij bloemontwikkeling en lichtregulatie werden geformuleerd aan de hand van de overlap in bindingsprofielen en motiefaanwezigheid na integratie van *non-canonical* en *canonical* motiefinformatie.

In parallel met de genoomwijde analyse van de experimentele ChIP data werd een exploratieve zoektocht naar TF bindingsplaatsen uitgevoerd op basis van sequentieconservatie in 12 species. We hebben een fylogenetisch *footprinting framework* opgezet gebruik makende van zowel alignerings- als nietaligneringsgebaseerde methoden. De *footprinting* aanpak werd ingesloten in een uitgebreid achtergrondmodel om de significantie te verzekeren van de resulterende geconserveerde niet-coderende sequenties op een FDR cut off van 5%. In totaal werden 69 361 *footprints* geëxtraheerd, die geassocieerd konden worden met 17 895 genen. Door de *footprints* te linken aan TF's waarvoor het bindingsmotief reeds gekend was (uit literatuur en experimentele studies) kon een genregulatorisch netwerk opgesteld worden bestaande uit 40 758 interacties. De relevantie van deze bindingsplaatsen werd aangetoond door middel van hun lokalisatie (2/3 ligt in een open chromatine regio). Daarenboven waren de genen nabijgelegen genen aangerijkt voor experimenteel geverifieerde *in vivo* doelwitgenen van de gekende TF's. Door middel van een geïntegreerde aanpak van vijf verschillende biologische validatiescores konden we de kwaliteit van het netwerk verder aantonen. In een laatste *proof-of-concept* experiment slaagden we erin om de statische interacties om te zetten naar een dynamisch netwerk door te doelwitgenen en TF's te linken aan hun expressieprofiel in specifieke condities.

Summary

Gene expression is a dynamic process, responding to various internal and external cues. The aims of this project consisted of two parts: one is the study of how transcriptional regulation and expression is organised across the genome and across functional gene modules. The other consists of using this data to assign function to unknown Arabidopsis genes.

We started by performing a systemic identification of functional gene modules (i.e. sets of genes that are associated based on a biological property) and to study the underlying levels of coexpression and coregulation. The modules were delineated based on large-scale expression data, functional gene annotations, experimental protein-protein interactions, and transcription factor-target interactions. The little overlap between different selected experimental input data sets corroborates the advantage of combining multiple data types. The resulting set of 1,563 modules covered 13,142 genes. Most modules displayed a significant level of expression coherence (i.e. the degree to which genes in a module coexpress), but functional and *cis*-regulatory coherence (using DNA motif presence as a proxy) was less prevalent. Hub genes were significantly associated with embryo lethality and provided evidence for crosstalk between different biological processes. To test the conservation of the coexpression component underlying functional modules, the modules were translated into other plant species using orthology. Based on expression data in those species, it was established that 58% of the modules showed conserved coexpression across multiple plants. Apart from studying regulation, the modules were explored in the context of function prediction. Based on the modules, 5,562 genes were annotated and evaluated using newly acquired experimental gene-GO associations. Out of 197 recently experimentally characterized genes, we found that 38.1% of newly associated gene functions could be inferred through the module context. New confirmed functions included cell wall biogenesis, xylem and phloem pattern formation, cell cycle, hormone stimulus, and circadian rhythm. Overall, biological hypotheses were generated for 1,701 unknown genes in Arabidopsis and six other plant species (43,621 genes).

Next, we aimed at studying the genomic organisation of transcriptional regulation in Arabidopsis. Publicly available genome-wide ChIP experiments were selected for a total of 27 TFs. All experiments were re-analysed in a uniform manner to ensure comparability between the experiments. This resulted in a experimental network containing 15,188 potential target genes connected by 46,619 potential regulatory interactions. Based on the integration of all these binding profiles, we identified hub targets and highly occupied target (HOT) regions. In the context of the currently profiled network, genes with many binding events in their regulatory regions are enriched development, stimulus responses, signalling and gene regulatory processes. Taking the controversy concerning HOT regions and their functionality in the biomedical field in consideration, we collected several lines of evidence that TF binding at plant HOT regions is functional. HOT regions harbour specific DNA motifs, are enriched for differentially expressed genes, and are often conserved across crucifers and dicots, even though they are not under higher levels of purifying selection than non-HOT regions. Another set of atypical bound regions was the set of distal regions, lying further than 4kb from their closest genes. Similar to all bound regions, distal bound regions were found to be under purifying selection. In addition, they are enriched for a chromatin state associated with regulation by the Polycomb repressive complex. The number of binding events in the vicinity of a gene is linked to their expression breadth. Hypotheses on co-binding and tethering between specific TFs involved in flowering and light regulation were formulated to explain part of the low correspondence between binding and DNA motif presence through integration of non-canonical

and canonical DNA motif information.

In parallel with the genome-wide analysis based on experimental ChIP data, we performed an exploratory search of TF binding sites based on sequence conservation across 12 species. We have developed a phylogenetic footprinting approach based on alignment and non-alignment-based techniques in concert. The footprinting approach was embedded in an elaborate background framework to ensure significance at an false discovery rate of 5%. In total, 69,361 footprints were extracted, located in the regulatory regions of 17,895 genes. By associating the footprints with TFBS of which the binding TF was known obtained from literature and experimental studies, we built gene regulatory network composed of 40,758 interactions. Relevance of these binding sites was shown through their localisation (2/3 of all CNSs) in DNase I hypersensitive sites. The resulting network shows significant enrichment towards experimentally verified *in vivo* targets of the known TFs in the network. Using an integrated approach of five different biological validation metrics, we substantiated the quality of the network. In a final proof-of-concept experiment, we studied the regulatory events in the context of detailed expression data. This allowed us to convert the static CNSs into condition-dependent regulatory networks.

Table of Contents

Preface

This manuscript is aimed at providing a scientific overview of the research I performed over the past four years. It consists of a general introduction, followed by three research chapters, and a general conclusion of the results together with my perspectives on the future of regulatory genomics.

Given the complexity of the matter to non-experts, the introduction is aimed at providing necessary knowledge to understand the research chapters. Therefore, it provides low-level information on the different processes and techniques that form the basis of the performed experiments. The general introduction is by no means a complete review of the field in question. The field of regulatory genomics is absolutely booming and a lot of exciting research is being done, far beyond the scope of this introduction. Therefore, I have aimed a providing the original publications, in combination with good reviews as entry-points for further study for the different techniques and concepts. More specific introductions, tackling the specific matter of the research chapters are provided embedded within the chapters. Relevant advances in the field towards the future are explored in the general conclusion and perspectives.

This being said, I wish you an interesting read.

List of Abbreviations

3C	Chromosome conformation capture
4C	Chromosome conformation capture-on-chip
5C	Chromosome conformation capture carbon copy
BP	Biological process
CC	Cellular component
CDS	Coding sequence
ChIP	Chromatin Immunoprecipitation
CMM	Conserved motif mapping
CNS	Conserved non-coding sequence
CRE	*Cis*-regulatory element
DE	Differential expression
DNA	deoxyribonucleic acid
EC	Expression coherence
EMSA	Electrophoretic Mobility Shift Assay
ENCODE	ENCyclopedia Of DNA Elements
eQTL	Expression quantitative trait loci
FDR	False discovery rate
FE	Fold enrichment
GEO	Gene expression omnibus
GO	Gene Ontology
GRN	Gene regulatory network
HC network	High confidence network
HOT	Highly occupied target
incRNA	Intronic non-coding RNA
INDEL	Small insertion or deletion
JA	Jasmonic acid
KS test	Kolmogorov-Smirnov test
lincRNA	Long intergenic non-coding RNA
lncRNA	Long non-coding RNA
ME network	Multiple-evidences network
MF	Molecular function
miRNA	miRNA
MQSE	Multi-query seed expansion
mRNA	Messenger RNA
ncRNA	Non-coding RNA
PBM	Protein binding microarray
PCC	Peason correlation coefficient
PCR	Polymerase chain reaction
PMRD	Plant miRNA database
Pol I, II and III	DNA polymerase I, II and III
PPI	Protein-protein interaction
QC	Quiescent center
RNA	Ribonucleic acid
rRNA	Ribosomal RNA
SA	Salicylic acid
Scmm	Comparative motif mapping score
SELEX	Systematic Evolution of Ligands by EXponential enrichment
siRNA	Short interfering RNA
SMSP	Multi species phylogeny footprinting score
SNP	Single nucleotide polymorphism
snRNA	Small nuclear RNA
SRA	Short read archive
TAIR	The Arabidopsis information resource
TF	Transcription factor
TFBS	Transcription factor binding site
TPR	True positive rate
tRNA	Transfer RNA
UTR	Untranslated region
VIGS	Virus-induced gene silencing
WGD	Whole genome duplication

Chapter 1

Introduction

As many hikers will be able to tell, you can never make the same hike twice. Surroundings are always changing, always adapting to seasons passing. Around the year, different flowers color the fields depending on the time of year, fruits color the trees in summer; and in fall, the pinnacle of colour is achieved by the colouring tree leaves while nature prepares for winter. All these changes are the result of complex systems in which plants perceive signals from their surroundings, and subsequently adapt their internal systems to constantly optimise towards their ultimate goal: grow and procreate. Apart from following robust seasons, sudden changes such as predator attacks need to be dealt with, shifting resources from growth to defence. Suffice it to say that the system regulating the plant needs to be complex to be dealing with constant perturbation.

Internally, the system traces back to the cell's components: proteins, microRNAs (miRNAs), metabolites, etc. Many of these interact to form large, dynamic networks and regulate each other's activity. As such, they form interacting complexes and signalling cascades that transfer exogenous signals down to the core of the complex, which can then provide an answer to the perturbation at hand. Underlying it all are the genes from which proteins and ribonucleic acid (RNA) components are formed. This PhD thesis handles with the manner in which genes are regulated to feed the system of all its components.

1.1 Levels of regulation

Regulation occurs at two main levels: the gene expression level (i.e. the creation of the gene products based on their deoxyribonucleic acid [DNA] sequence) and the protein level (e.g. post-translational modifications, conformational changes, and protein-protein interactions). In theory, all steps are potential points of regulation (Figure 1.1). Because of the numerous steps between activation of transcription and the functional protein, a change in gene expression level does not necessarily indicate a change in protein level/activity and vice versa. Genes expression consists of two major steps: transcription (DNA is transcribed to messenger RNA [mRNA]) and translation (mRNA is translated into peptides).

Regulation on the protein level is widely implemented in signalling pathways of plant hormones, where proteins undergo modifications upon stimulus and are poised for degradation (e.g. AUX/IAA upon auxin stimulus and DELLA upon giberellic acid stimulus[1]). These cascades will often ultimately result in a stable transcriptomic response, and exhibit the interconnection between the different regulatory levels. This thesis will focus on transcriptional regulation, but we refer to Walton et al.[2] for an elaborate review of studies on protein regulation, in plant hormone signalling cascades specifically. Given the scope of the thesis, the following sections will focus on transcription.

1.2 Gene expression

Transcription

In a gene, the two strands of DNA are called the coding strand, and the template strand. These arbitrary names reflect the fact that the resulting mRNA will have the same sequence as the coding strand, while the template (opposite) strand is the one actually being used to guide the mRNA synthesis (Figure 1.2).

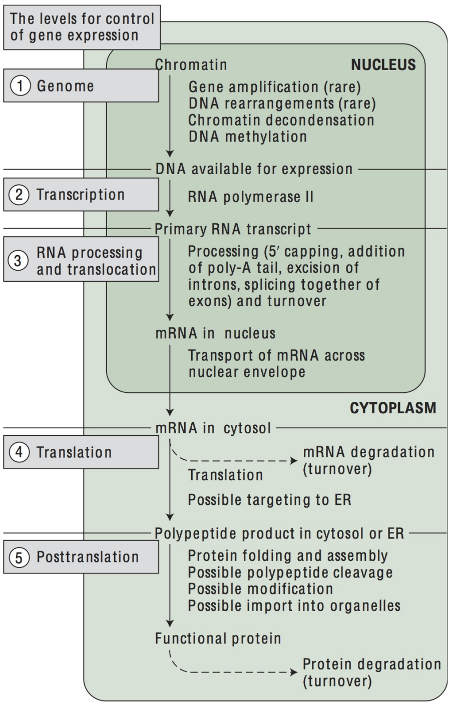

The levels for control of gene expression
NUCLEUS
Chromatin
① Genome
Gene amplification (rare)
DNA rearrangements (rare)
Chromatin decondensation
DNA methylation
DNA available for expression
② Transcription
RNA polymerase II
Primary RNA transcript
③ RNA processing and translocation
Processing (5′ capping, addition of poly-A tail, excision of introns, splicing together of exons) and turnover
mRNA in nucleus
Transport of mRNA across nuclear envelope
CYTOPLASM
mRNA in cytosol
④ Translation
mRNA degradation (turnover)
Translation
Possible targeting to ER
Polypeptide product in cytosol or ER
⑤ Posttranslation
Protein folding and assembly
Possible polypeptide cleavage
Possible modification
Possible import into organelles
Functional protein
Protein degradation (turnover)

Figure 1.1: **From gene to protein: levels of regulation.** *Eukaryotic gene expression can be regulated at multiple levels: (1) genomic regulation, by gene amplification, DNA rearrangements, chromatin (the collection of DNA and its associated proteins) decondensation or condensation, or DNA methylation; (2) transcriptional regulation; (3) RNA processing, and RNA turnover in the nucleus and translocation out of the nucleus; (4) translational control (including binding to endoplasmatic reticulum (ER) in some cases); (5) post-translational control (including mRNA turnover in the cytosol, and the folding, assembly, modification, and import of proteins into organelles).* Source: Taiz and Zeiger[3]

The DNA is transcribed to RNA by a DNA-dependent RNA polymerase. The type of polymerase differs depending on the type of gene: RNA polymerase II (Pol II) is responsible for transcription of pre-mRNAs, microRNAs (miRNAs), and a class of small nuclear RNAs (snRNAs).[4–6] Because Pol II is involved in the expression of protein coding genes, it is by far the most elaborately studied. RNA polymerase I (Pol I) transcribes most of the ribosomal RNAs (rRNAs). The best known targets of RNA polymerase III (Pol III) are the different tRNAs.

In the case of RNA genes, the process of transcription directly synthesizes the gene product. In case of protein-coding genes, the mRNA undergoes translation to a protein. The transcription of protein-coding and miRNA genes consists of four stages: promoter recruitment, initiation, elongation and termination. Although all four stages are complexly regulated, the study of gene expression regulation is aimed primarily towards promoter recruitment and initiation.[7]

The initiation requires the assembly of the basal transcription apparatus at the core promoter region: the pre-initiation complex.[8] Historically, the promoter is the sequence located upstream of the transcription start site (the first nucleotide that is copied into the mRNA molecule). Nevertheless, TF binding can also occur in the untranslated regions (UTRs), downstream, or in one of its introns[9], as evidenced both from genome-wide chromatin immunoprecipitation (ChIP) studies (e.g. ERF115[10]; see also section 1.6), and expression quantitative trait loci (eQTL) analyses.[11] eQTL analyses are large scale studies that assign causal relations between genomic variations and an observed expression differences. The promoter can roughly be divided into a proximal part (or core promoter) and a distal part (located 5' of the core promoter). The distal part contains the *cis*-regulatory elements – also called transcription factor binding sites (TFBS) or sequence motifs – required for spatio-temporal expression (Figure 1.3). The TFBSs are recognised by transcription factors (TFs), which function in the assembly of the pre-initiation complex.There are two classes of TFs: general TFs and spatio-temporal TFs. General TFs are those always required in the formation of the pre-initiation complex.[12] In addition, spatio-temporal TFs are responsible for the spatio-temporal expression of the target gene.[8] Interactions mediated by components of the basal machinery and both types of TFs ensure efficient and regulated transcription.[3,8] Transcription

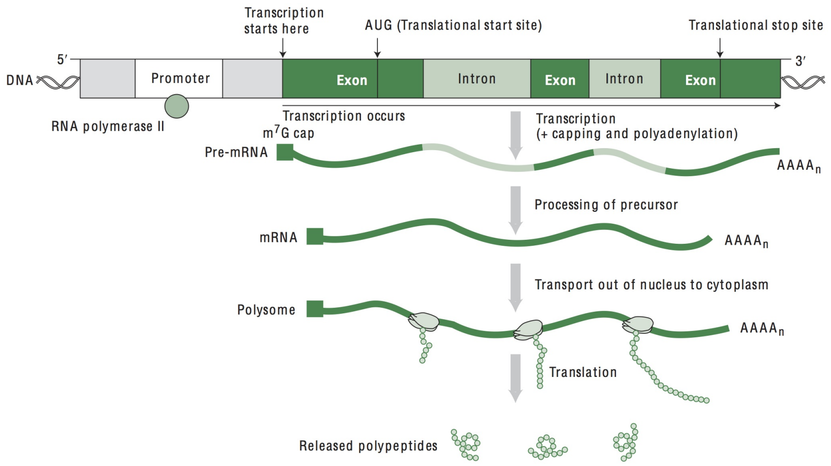


Figure 1.2: **Gene expression in eukaryotes.** *RNA polymerase II binds to the promoter of genes that encode proteins. Genes are divided into introns and exons. Transcription from the template strand proceeds in the 3' to 5' direction at the transcription start site. The pre-mRNA is processed into a mature mRNA (removal of introns) after which this leaves the nucleus to be translated by the ribosomes.* Source: Taiz and Zeiger[3]

initiation ends with the incorporation of the first few nucleotides of the mRNA.

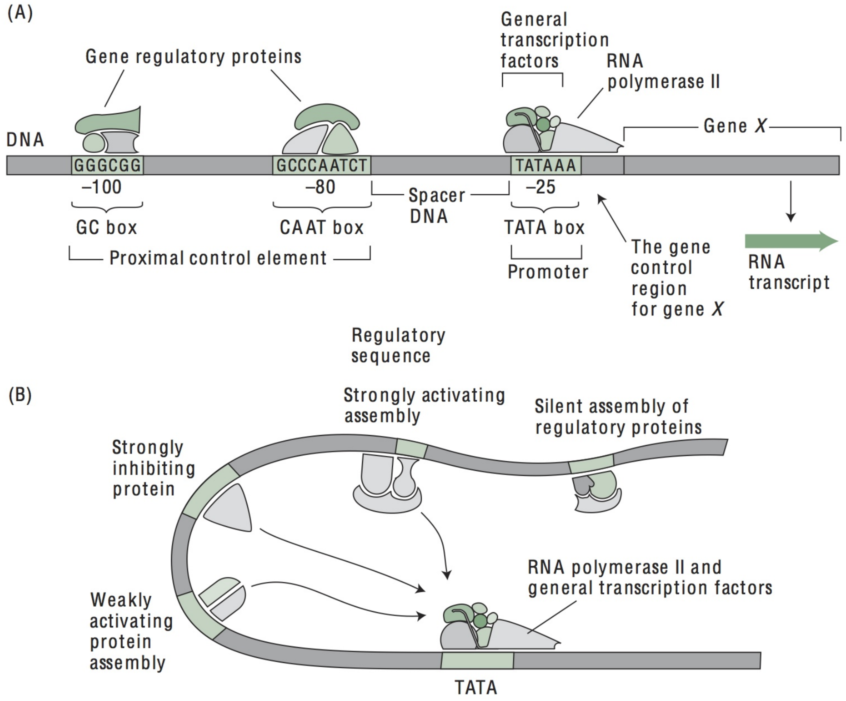

(A)
Gene regulatory proteins
General transcription factors
RNA polymerase II
DNA
Gene X
GGGCGG
GCCCAATCT
TATAAA
–100
–80
–25
Spacer DNA
GC box
CAAT box
TATA box
Proximal control element
Promoter
The gene control region for gene X
RNA transcript
Regulatory sequence
(B)
Strongly activating assembly
Silent assembly of regulatory proteins
Strongly inhibiting protein
RNA polymerase II and general transcription factors
Weakly activating protein assembly
TATA

Figure 1.3: **Organization and regulation of a typical eukaryotic gene.** *(A) Features of a typical eukaryotic RNA polymerase II minimum promoter and proteins that regulate gene expression. (B) Regulation of transcription by distal regulatory sequences and trans-acting factors. The trans-acting factors bound to distal regulatory sequences can act in concert to activate transcription by making direct physical contact with the transcription initiation complex.* Source: Taiz and Zeiger[3]

When the transcriptional machinery is in place, the transcript is elongated to its full length. However, the (near complete) assembly of the initiation complex can also be part of a strategy to poise the promoter.[13] This ensures a fast response to subsequent additional signals that either provides a last additional final TF, or remove a repressing TF (e.g. the removal of AUX/IAA from the ARF TFs after upon auxin stimulus[1]).

Elongation is a complex and highly regulated phase of the transcription cycle.[7] A lot of factors contribute to its dynamic control: some modulate activity of RNA polymerase II, others facilitate the transcription by influencing chromatin (the ensemble of DNA and its associated proteins; see section 1.3). Elongation plays a central role in coordinating transcription and various co-transcriptional RNA processing steps such as 5'-capping, splicing and polyadenylation.

Finally, when the full-length mRNA is formed, transcription is terminated by polyadenylation, the 5' end is capped and the the polymerase complex and all its co-factors are disassembled.

Translation

Transcription results in the formation of a pre-mRNA. The pre-mRNA is spliced into a mature mRNA by removal of the introns (Figure 1.2). The mature mRNA is read out by the ribosomes to assemble the correct series of amino acids from which peptides and proteins are formed.

1.3 Regulation of Transcription

Studying transcriptional regulation is often simplified to studying the gene regulatory network of TFs and target genes. In reality, whether or not a TF can bind the promoter of a target gene is influenced by the state of the chromatin, which is defined by the combination of different modifications present on the chromatin.[14] Lack or presence of cooperating TFs also influence either the binding of the TF, or whether the binding will result in an expression signal.

The Transcriptional Gene Regulatory Network

The transcriptional network is the complete collection of interactions between TFs and their target genes. Differential gene expression is accomplished by the presence/absence of regulating TFs. In *Arabidopsis thaliana* (hereafter Arabidopsis), around 1,700 genes are predicted to be transcription factors, representing ± 5% of the total gene count. The fact that this percentage is twice that of *C. elegans* possibly reflects a higher regulatory complexity.[15,16]

TFs typically consist of at least two domains: a DNA binding domain and a transcription activating domain or a transcription repressing domain. The DNA binding domain defines the specificity towards DNA, while the activating or repressing domain influences transcription of the target gene. TFs are divided into TF-families according to their DNA binding domain, of which 50-58 exist in Arabidopsis depending on the source.[17,18]

Conceptualising the interactions between TFs and their target genes as a directed graph allows for network analysis.[16,19] This mathematical approach has been shown to be able to retrieve biological information. A transcriptional network can be broken down into four levels of detail (Figure 1.4). The basic unit of the network is the interaction between TF and TFBS. The higher levels of the network all hold biological information. For example, the network motif level —not to be confused with the sequence motif —can be used to explain or model oscillations. The module level gives insight into dense subgroups of higher connectivity and those are linked to co-regulatory modules. And finally, the entire transcriptional network holds information about the global organisation of the transcriptional network and helps to find its central components, called hubs. The latter have been shown to coincide with vital components in the biological network.[20]

The Importance of the Chromatin State

Taking into account merely the presence or absence of the transcription factor to determine transcription initiation is a simplification. Whether or not a transcription factor can bind the DNA is not solely dependent on the presence/absence of the TF itself. It is also determined by the chromatin state of the region.

Chromatin is the whole of DNA and all of its associated proteins and consists of the paired DNA strands that are wrapped around nucleosomes (protein cores of different histone proteins). Chromatin is a dynamic structure, with two main possible states: heterochromatin and euchromatin. There is also a third —intermediate —state called bivalent chromatin, which contains both activating and repressing marks and is the chromatin variant of poised promoters.[21] The structure depends on the types of protein modifications that are present on the histon tails. Heterochromatin is tightly packed (⌀30 nm), and forms a closed conformation in which the genes residing in the region are silenced. Euchromatin is a loosely

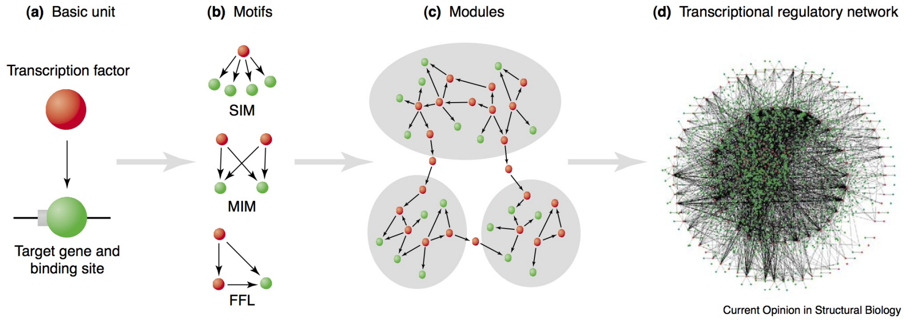

(a) Basic unit
(b) Motifs
(c) Modules
(d) Transcriptional regulatory network
Transcription factor
Target gene and binding site
SIM
MIM
FFL
Current Opinion in Structural Biology

Figure 1.4: **Transcriptional network analysis.** *(a) The basic unit of a transcriptional network is the interaction between a TF and its target gene. (b) Network motifs are combinations of basic units that provide the structure of the network and are linked to dynamics of the nework. Some motifs such as the feed-forward loop (FFL), single input module (SIM) and multiple input modules (MIM) are found commonly. (c) Modules are genes in the network that are associated more densely with each other than with the rest of the network. These often represent genes involved in the same process. (d) The complete transcriptional network —although holding all information —is the least usefull for biolocal analysis.* Source: Babu et al.[16]

packed (⌀11 nm), open configuration that harbours expressed genes. Historically, chromatin was mainly thought of as a DNA packaging mechanism. Now chromatin is seen as a highly dynamic structure affecting all DNA transactions: replication, repair, recombination, transposition, chromosome segregation and transcription.[22] In addition, chromatin states can potentially be inherited in cell lineages, meaning that the expression state of the genes they influence is stable across generations, even when the original stimulus responsible for has disappeared. When this is the case, these modifications are epigenetic: they represent a heritable trait that is not linked to changes in the DNA sequence. Unfortunately, the term epigenetics is often used in a loose fashion for any kind of chromatin modification, which is erroneous by definition and the subject of debate.[23]

Nucleosomes in DNA form a barrier for proteins that need to bind the DNA, including those that regulate gene expression (i.e. TFs). The restriction of the chromatin state on the DNA accessibility is dynamic and changes during development and in response to exogenous cues: e.g. stress, pathogen attack, temperature and light.[24] Chromatin states — and as a consequence the accessibility of the DNA — are modified by a variety of mechanisms and factors: covalent modifications of the histone core, the incorporation of histone variants, DNA methylation, chromatin-remodelling enzymes and small non-coding RNAs (transcriptional co-suppression; Figure 1.5).[22] Among the many mechanisms, histone modifications play a major role. How the histone modifiers are recruited to the DNA is largely unanswered, but one mechanism is recruitment by transcription factors.[24,25] So there seems to be an interplay between the epigenetic level and the transcription factors: transcription factors help modify the chromatin state and the chromatin state defines the accessibility of the DNA to transcription factors. Some examples of chromatin modifications and their link to transcription are shown in Figure 1.6.

Keeping in mind that plants need to be able to respond to a broad range of environmental factors, it is also interesting to mention that repressive histone modifications in Arabidopsis occupy smaller domains compared to those in metazoans, possibly making them more readily reversible in Arabidopsis. This may reflect the higher developmental plasticity so well known in plants.[24]

Post-transcriptional regulation

After transcription and splicing, the mRNA is ready to be translated. Post-transcriptional regulation interferes at this point. It is achieved through a sequence similarity dependent mechanism based on miRNAs or short interfering RNAs (siRNAs). The RNA molecules bind transcripts based on sequence similarity which allows it to target entire gene families at once (post-transcriptional co-suppression). A well-known example mechanism of miRNA regulation is its role in virus-induced gene silencing[27].

1.4 The *Arabidopsis thaliana* genome does not exist

While the "*Arabidopsis thaliana* genome" has been sequenced in 2010[28], this is in essence a simplification of the truth. As for humans, any plant is an individual, and as such has its own 'personalised'

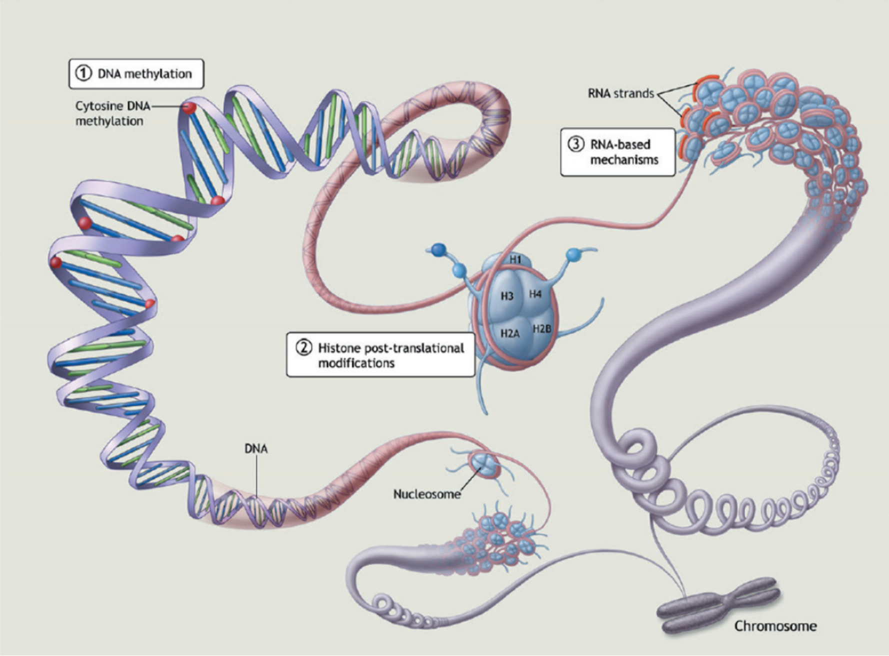


Figure 1.5: **Different mechanisms of epigenetic regulation.** *Different mechanism of epigenetic transcriptional regulation are shown. DNA methylation is the methylation of cytosine residues in a CG, CHG or CHH context (H is either A, T or C). Histone post-translational modification involves the addition of groups to the histone tails by histone modifying enzymes. RNA based mechanisms make use of ncRNA's. MiRNA or siRNA induce heterochromatin formation in regions with sequence similarity to the RNA.* Source: Allis et al.[26]

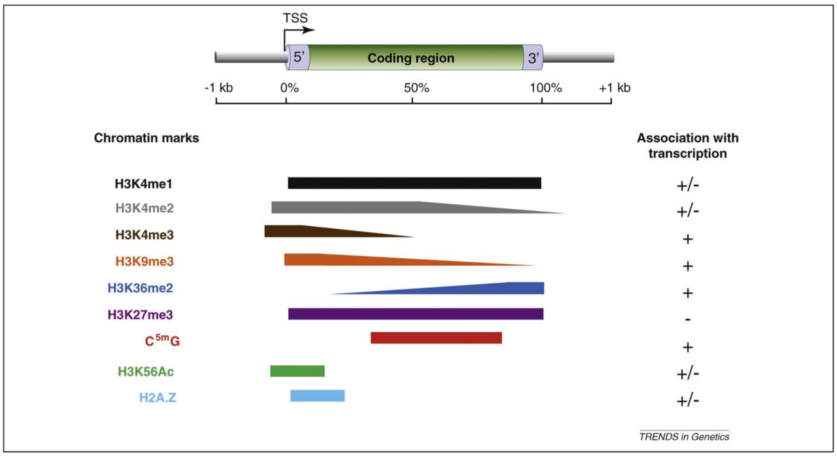


Figure 1.6: **Distribution of chromatin modifications over genes and their relationship with expression.** *Chromatin marks analyzed using genome-scale approaches are represented along a schematic Arabidopsis gene. 1 kb of 5' and 3' flanking regions, the transcription start site, the 5' and 3' UTRs and the coding region are indicated. No assumption is made about the co-deposition of the different marks at a single locus. Note that as in other eukaryotes, the region around the transcription start site of actively transcribed genes typically shows an apparent nucleosome depletion (not shown). (+) and (-) denote a tight association with active or repressed transcription, respectively, whereas (+/-) indicates no particular association. H3K4me1 = addition of 1 methyl group to histone 3 at the 4th residue which is a lysine; $C^{5m}G$ = methylation of a cytosine residue in CG context; H2A.Z = a histon variant.* Source: Roudier et al.[22]

genome. Because Arabidopsis is highly selfing, its population was historically considered to be a collection of asexual lineages (ecotypes). This would have meant that the variation in each lineage was fixed and that variation of different ecotypes would never transfer into another lineage. Any new difference of a plant compared to its parent would then have arisen from mutation. The genome sequence of Arabidopsis is in fact the sequence of a variant called *Columbia-0*, which is also the most widely used for experimentation.

This ecotype view has been shown to be false, as there is clear evidence of recombination between different accessions.[29] As a consequence, there is no phylogeny of variants. A phylogeny would implicate that each lineage had remained independent after it 'speciated' from another lineage. This does not signify that a complete population cannot exhibit a tree-like population structure, which simply indicates that individuals within the same population are much more related to each other than to a different population. The latter can be observed in the isolation by distance when relating the genomic distances to the geographical distances between pairs of accessions.

The study of the variation in genomes between populations and individuals is called population genomics. Possible variation are single nucleotide polymorphisms (SNPs), small insertions and deletions (INDELS), and reversions. The first studies on population genomics were based on PCR-based sequencing[29] or array-based profiling of a limited set of sequence regions.[30] With the appearance of next-generation sequencing, complete genomes are now resequenced, with as prime example the complete resequencing of over 1200 accessions from across the world in the 1001 Genomes Project.[31]

Variation can be non-synonymous or synonymous, meaning the change in the DNA has or has no effect respectively. Non-synonymous changes are easily explained in the context of protein-coding genes, where the incorporated amino acid is altered due to a change in the codon. In contrast, synonymous changes do not alter the amino acid and thus have no influence on the final gene product.

However, non-synonymous and synonymous variation is of equal —if not more —importance in the non-coding genome, where a non-synonymous SNP would affect the regulation of a gene. Around 60% of the genomic variation lies in the non-coding genome Cao et al.[31], which is of importance because the non-coding genome harbours a plethora of regulatory elements. Any variation in TFBSs can influence the wiring of the transcriptional network, either due to loss of the binding site, or due to a changed interaction specificity between TF and TFBS. At this point, it is impossible to assign a non-synonymous/synonymous label on all the non-coding SNPs, simply because the non-coding genome is far from profiled.

Using eQTL analyses, it is possible to link variations in the genomes to differences in expression.[32] Historically, this was done with a limited set of markers[33], but the methodology has now evolved to using the complete set of known SNPs in a genome thanks to whole-genome resequencing.[34] Many non-coding SNPs have since been associated with differences in gene expression. Their importance is stressed by their involvement in many disease-associated variants in humans where they systematically perturb transcription factor recognition sequences, frequently alter allelic chromatin states, and form regulatory networks[35] and phenotypic variation in plants. The latter has been extensively reviewed by Cubillos et al.[36].

Therefore, while we are still nowhere near having unravelled the transcriptional network in *Columbia-0*, the next challenge of mapping variations in the transcriptional network in a population is already known.

1.5 Evolution of the transcriptional network

In addition to being variable in a single species, the transcriptional network has also evolved differently from the ancestral state in different species. Similarly to the variants, evolution of the wiring of the transcriptional network is believed to have been one of the major driving forces in adaptation of different species to different conditions. It is known that genome duplication is a major driving force for evolution[37] because this allows for rewiring in a in a duplicated network, thus lowering the potential detrimental influence (Figure 1.7). Explanations on how rewiring of the interactions would evolve on the molecular scale are depicted in Figure 1.8. The latter can of course also happen without a prior duplication, as was the case for the LEAFY TF.[38]

On the level of cooperating genes, analysis of the yeast transcriptional network has indicated that most transcriptional modules (sets of genes that are coexpressed and share at least one motif and are thus presumed to be coregulated and cooperate functionally) in a network are conserved[39–41], yet some are lineage specific. The mechanism by which they arise is transcriptional network expansion. The transcriptional network can expand by either duplication of the TF, duplication of the target gene, or both (Figure 1.7). In most cases, a duplicated gene (i.c. TF) will loose its function, but in some cases there is neofunctionalisation (i.c. loss and gain of interactions). If this happens, this could be the 'seed' for a new transcriptional module. The conservation of transcriptional modules can also be used to identify TFBSs through a methodology called phylogenetic footprinting[42], which is explained more thoroughly in section 1.6.

Linked to the evolution of the sequence motifs, one can investigate the conservation of TF-target interactions. While a TF can co-evolve with its TFBS, it can also loose its binding. By comparing the binding events of two TFs across human, mouse, dog, rat, and chicken, Schmidt et al.[43] showed that the motifs are rarely conserved, but that some regulation is conserved due to the emergence of a new sequence motif in the vicinity. This process is called turn-over. It was shown that the binding sites that are organised in *cis*-regulatory modules (multiple *cis*-regulatory elements located together) are more often conserved. When conserved, they function in critical pathways.[44] Changes in *cis*-regulatory modules were also shown to correlate better with changes in transcript levels compared to changes in single binding events.[45]

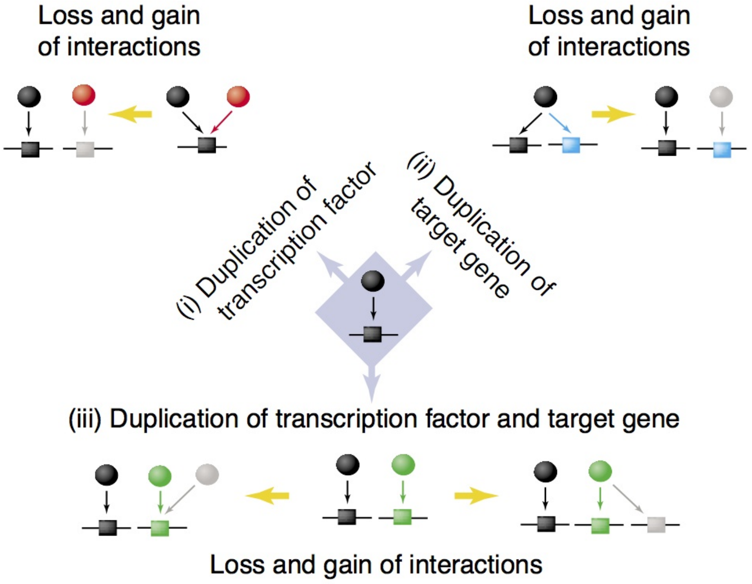

Loss and gain of interactions
Loss and gain of interactions
(i) Duplication of transcription factor
(ii) Duplication of target gene
(iii) Duplication of transcription factor and target gene
Loss and gain of interactions

Figure 1.7: **Mechanisms of transcriptional network expansion.** *Possible scenarios for the evolution of the basic unit are duplication of (i) the transcription factor, (ii) the target gene and (iii) both. Transcription factor duplication results in both copies regulating the same target. During divergence, new or existing regulatory interactions may be gained or lost. Similarly, target gene duplication results in both copies being regulated by the same transcription factor. Divergence may result in gain or loss of regulators.* Source: Babu et al.[16]

1.6 METHODOLOGY: Profiling the gene regulatory network

Chromatin Immunoprecipitation (ChIP)

The most widely used method to detect the binding sites of a known TF is ChIP (Figure 1.9). The TF (or any other DNA binding protein) is cross-linked to the DNA. The DNA is then sonicated into small pieces and the bound DNA region is extracted using either antibodies directly against the TF (e.g. Thibaud-Nissen et al.[46]) or antibodies against a protein TAG that has been genetically fused to the TF (e.g. Verkest et al.[47]). The cross-linking is reversed, releasing the DNA from the TF. Strategies exist that perform the pull-down on native DNA (no cross-linking), but these are more suited for histone modifications.[48] Initially, the technique was used to verify target gene binding using PCR (requires prior knowledge about the binding site). An alternative methodology With development of the tiling array (ChIP-chip;[49]), and later next-generation sequencing (ChIP-seq;[50]), the technique could be used in a exploratory manner. In the ChIP-chip methodology, the bound DNA is profiled by hybridisation on a tiling array. Different tiling arrays exist, but the Affymetrix one has been most widely used in Arabidopsis. On the array, the Arabidopsis genome is probed by stretches of 25bp, spaced by 10bp, which results in an overall resolution of 35bp. With the emergence of next-generation sequencing, the tiling array was replaced by the complete sequencing of the immunoprecipitation sample. The sequencing theoretically provides

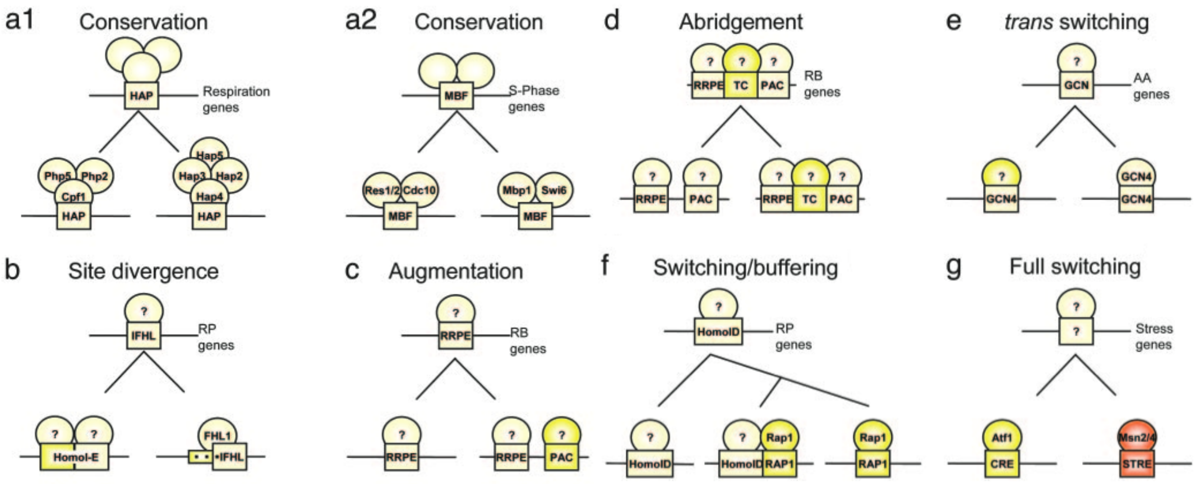


Figure 1.8: **Alternative mechanisms for the evolution of the regulation of transcriptional modules.** *Each panel shows a distinct scenario of the inferred evolution of an ancestral regulatory program (Upper) into programs observed in 2 or more extant species (Lower). For each module, a schematic representative promoter is shown (black line) along with cis-elements (boxes) and transcription factors (ovals). Ancestral conserved sites and proteins are in light yellow, and innovations and divergences are in bright yellow or red.* ***(a)*** *Conservation of both the cis-element and trans-factors* ***(b)*** *A gradual divergence of binding site sequence.* ***(c)*** *Augmentation of an existing program by the emergence of a new site along an ancestral one.* ***(d)*** *Abridgement of an augmented program by binding site loss.* ***(e)*** *Switching of the transcription factor while maintaining the same cis-element.* ***(f,g)*** *Full switching of a program from one cis-element to another. In some cases (f), this can occur by a combination of augmentation and abridgement.* Source: Tanay et al.[41]

access to binding/non-binding information for every single nucleotide in the genome.

Performing a ChIP experiment results in a snapshot of the regulatory state, since all interactions are 'frozen' into place by cross-linking. While this ensures detection of both active interactions, and interactions that are part of a poised promoter complex, interactions that are not of relevance in the profiled condition and transient interactions will be missed.[13] In the light of condition-specific conditions, methodologies have now been developed to perform differential ChIP-Seq analysis.[52,53] These allow the detection of binding events that are different across conditions, or even across different natural accessions.

Computationally, a ChIP experiment requires the identification of local enrichments of signal compared to the control. A tiling array essentially contains DNA polymers (probes) that represent the entire genome. *Sensu strictu*, each DNA polymer on the array should overlap with the next, to truly cover the complete genome. Instead however, the Affymetrix Arabidopsis Tiling array consists of 25bp probes, with 10bp gaps in between. Since DNA sequences extracted by ChIP are usually around 200bp, this does not interfere with the signal-capture.

The data needed to analyse a tiling array hybridisation experiment consists of a CEL file, and a bpmap file. The former contains the signal intensities for each position on the array, while the latter contains the coordinates of each probe's position on the genome. Given the premise of the ChIP biology, the analysis consists of identifying regions in the genome that have higher signals in the ChIP sample compared to a control sample (complete input sample). The combination of the signal information and the genomic coordinates of the different probes leads to information such as visualised in Figure 1.10.

The field of microarray data analysis is very mature, and different procedures have been devised to optimally compare samples with their controls. Software tools that can be used to detect peaks include HMM (hidden markov model)[55], TileMap[56], MAT (model-based analysis of tiling Array)[50], and BAC (Bayesian analysis of ChIP-chip)[57].

In a ChIP-Seq experiment, the IP sample is profiled by using next-generation sequencing technology instead of hybridising it on a tiling array. The sequencing results in millions of reads for which their genomic region of origin needs to be determined. This step is achieved by mapping the sequences back the the reference genome.

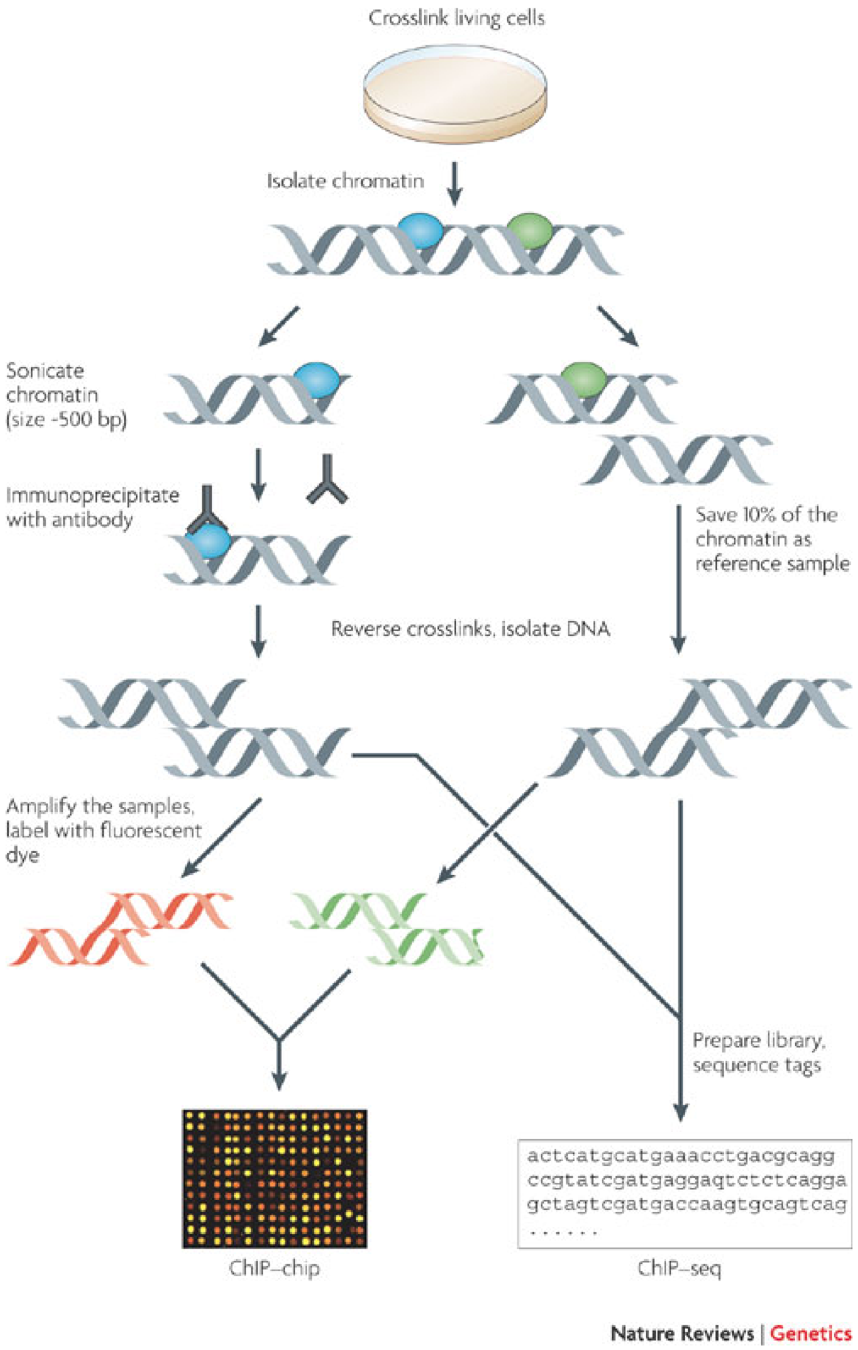

Crosslink living cells
Isolate chromatin
Sonicate chromatin (size ~500 bp)
Immunoprecipitate with antibody
Save 10% of the chromatin as reference sample
Reverse crosslinks, isolate DNA
Amplify the samples, label with fluorescent dye
Prepare library, sequence tags
actcatgcatgaaacctgacgcagg
ccgtatcgatgaggagtctctcagga
gctagtcgatgaccaagtgcagtcag
......
ChIP–chip
ChIP–seq
Nature Reviews | Genetics

Figure 1.9: **ChIP Methodology.** Source: Farnham [51]

Since the emergence of next-generation sequencing, the diversity of available mapping tools has equally boomed. It has been a rapidly developing field, constantly pushed by the evolutions of the sequencers themselves towards longer reads and greater output. Although the read length has gone up in ChIP-Seq experiments from 32bp to 75bp, long reads are of less importance in ChIP experiments as the whole protocol is based on finding local stacks of reads.

Different mappers have been developed using different algorithmic strategies and implementations. There are two major schools: hash-table based and Burrows-Wheeler Transform (BWT) based. The most commonly used tools for each category are Bowtie / BWA (BW) and MAQ / GSNAP. Over the years, evaluating and benchmarking has been a daunting task, and very dependent on the definition of correctness. An overview of benchmarking papers, together with a recent benchmark is given by [58]. In general, one can say that the choice of mapper is dependent on the question at hand. BWA is most often used for ChIP analyses, while GSNAP, which is focused on dissecting complicated splicing patters is a

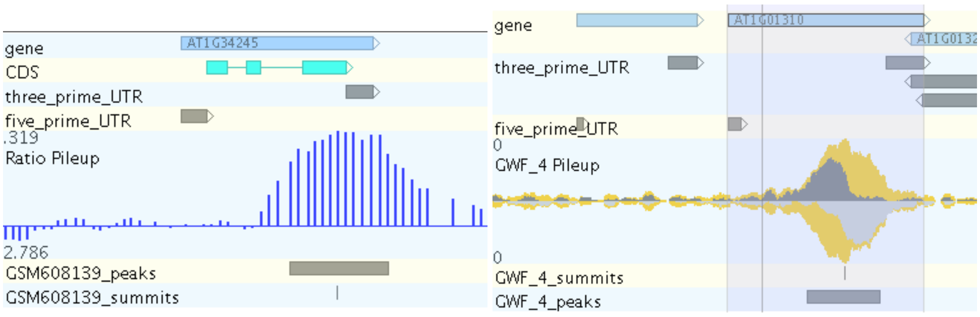


Figure 1.10: **ChIP-chip (left) and ChIP-Seq (right) signal visualisation in GenomeView.**[54] *(Left) Visualisation of the 25bp probe signal ('Ratio Pileup') mapped onto its correct genomic coordinates leading to the identification of a TF-bound region. (Right) Visualisation the pileup of the DNA reads mapped at their correct genomic location, leading to the identification of a TF-bound region.*

good choice for RNA-Seq.

Finally, the local enrichment of reads is called by a peak caller, of which many have been developed over the years. Each of them have slightly different approaches, ranging from different manners to compare to sample to control, different underlying distributions and different manners to calculate FDR. An overview of a number of them is given in Fig. 1.11.

Yeast-1-Hybrid

Whereas ChIP techniques are TF-centred, the yeast-1-hybrid methodology allows detection of protein-DNA interactions of any sort.[60] The technology is based on the genetic fusing of a library of proteins to a strong transcriptional activation domain. The sequence of interest is cloned in front of a reporter gene, which allowed screening of binding events based on the expression of the reporter gene.

At first, the yeast-1-hybrid system was mainly used with short DNA elements (30 bp) as DNA baits. The technique was drastically improved to facilitate the high-throughput and unbiased identification of protein-DNA interactions.[61–63] The improved system can be used with both small (e.g., *cis*-regulatory elements), and large DNA fragments (e.g., gene promoters). The use of promoters circumvents the need to identify functional sequence motifs for a gene of interest *a priori*. The system was used to, among others, construct a gene regulatory network in root.[64,65] Because of the complementary perspectives, the combination of ChIP and yeast-1-hybrid would allow to traverse and experimentally map the transcriptional network.

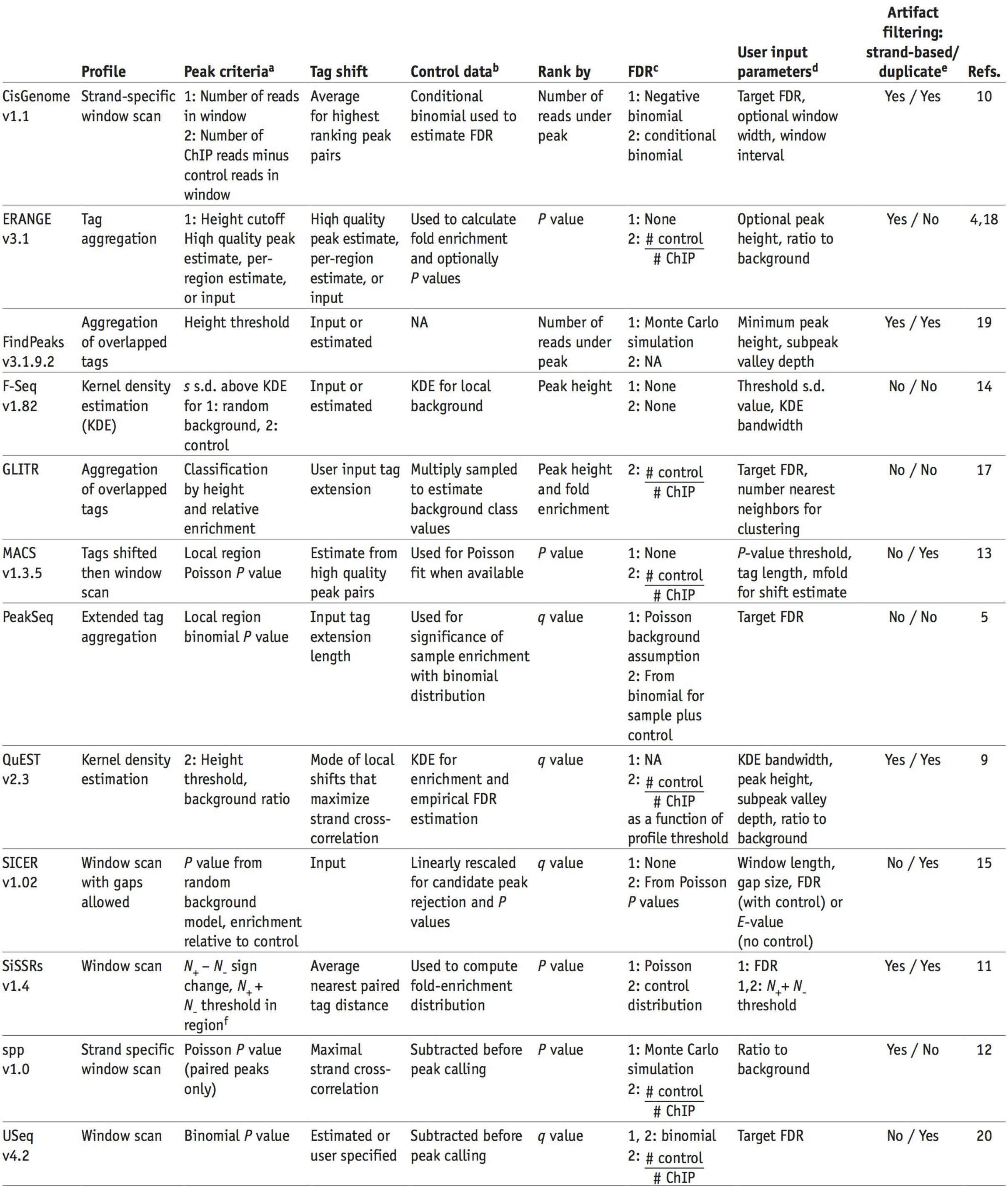
Indirect Methods

Finally, there are a number of experimental methods that determine the sequence motif that is bound by a TF *in vitro*, but do not directly determine the interaction between a TF and its target gene(s). To determine the actual genomic locations where the TF binds, the sequence motif has to be computationally mapped on the genomic sequence, which is prone to false positives. Integration of mapping with other data types such as functional and expression coherence of nearby target genes, or data on the chromatin state of the mapping location is often used to enrich the list towards true positive mappings.[66]

Electrophoretic Mobility Shift Assay (EMSA) is a gel-based method to separate protein-bound DNA molecules from unbound DNA molecules and identify the motif for a given TF. By introducing one point mutation per DNA molecule, it is possible to accurately determine the nucleotides that are necessary for the TF to bind and as such detect the motif.[67]

Systematic Evolution of Ligands by EXponential enrichment (SELEX) is based on the iterative selection and amplification of the DNA sequence with the highest affinity for a given TF. Using a PCR, a batch of random DNA sequences is generated. Multiple rounds of ligand selection and amplification exponentially enrich the population for the highest affinity species that can be clonally isolated and characterized.[68–70]

Protein binding microarrays (PBMs) have been gaining a lot of ground as it is a high-throughput



	Profile	Peak criteria[a]	Tag shift	Control data[b]	Rank by	FDR[c]	User input parameters[d]	Artifact filtering: strand-based/ duplicate[e]	Refs.
CisGenome v1.1	Strand-specific window scan	1: Number of reads in window 2: Number of ChIP reads minus control reads in window	Average for highest ranking peak pairs	Conditional binomial used to estimate FDR	Number of reads under peak	1: Negative binomial 2: conditional binomial	Target FDR, optional window width, window interval	Yes / Yes	10
ERANGE v3.1	Tag aggregation	1: Height cutoff Hiqh quality peak estimate, per-region estimate, or input	Hiqh quality peak estimate, per-region estimate, or input	Used to calculate fold enrichment and optionally *P* values	*P* value	1: None 2: # control / # ChIP	Optional peak height, ratio to background	Yes / No	4,18
FindPeaks v3.1.9.2	Aggregation of overlapped tags	Height threshold	Input or estimated	NA	Number of reads under peak	1: Monte Carlo simulation 2: NA	Minimum peak height, subpeak valley depth	Yes / Yes	19
F-Seq v1.82	Kernel density estimation (KDE)	*s* s.d. above KDE for 1: random background, 2: control	Input or estimated	KDE for local background	Peak height	1: None 2: None	Threshold s.d. value, KDE bandwidth	No / No	14
GLITR	Aggregation of overlapped tags	Classification by height and relative enrichment	User input tag extension	Multiply sampled to estimate background class values	Peak height and fold enrichment	2: # control / # ChIP	Target FDR, number nearest neighbors for clustering	No / No	17
MACS v1.3.5	Tags shifted then window scan	Local region Poisson *P* value	Estimate from high quality peak pairs	Used for Poisson fit when available	*P* value	1: None 2: # control / # ChIP	*P*-value threshold, tag length, mfold for shift estimate	No / Yes	13
PeakSeq	Extended tag aggregation	Local region binomial *P* value	Input tag extension length	Used for significance of sample enrichment with binomial distribution	*q* value	1: Poisson background assumption 2: From binomial for sample plus control	Target FDR	No / No	5
QuEST v2.3	Kernel density estimation	2: Height threshold, background ratio	Mode of local shifts that maximize strand cross-correlation	KDE for enrichment and empirical FDR estimation	*q* value	1: NA 2: # control / # ChIP as a function of profile threshold	KDE bandwidth, peak height, subpeak valley depth, ratio to background	Yes / Yes	9
SICER v1.02	Window scan with gaps allowed	*P* value from random background model, enrichment relative to control	Input	Linearly rescaled for candidate peak rejection and *P* values	*q* value	1: None 2: From Poisson *P* values	Window length, gap size, FDR (with control) or *E*-value (no control)	No / Yes	15
SiSSRs v1.4	Window scan	$N_+ - N_-$ sign change, $N_+ + N_-$ threshold in region[f]	Average nearest paired tag distance	Used to compute fold-enrichment distribution	*P* value	1: Poisson 2: control distribution	1: FDR 1,2: $N_+ + N_-$ threshold	Yes / Yes	11
spp v1.0	Strand specific window scan	Poisson *P* value (paired peaks only)	Maximal strand cross-correlation	Subtracted before peak calling	*P* value	1: Monte Carlo simulation 2: # control / # ChIP	Ratio to background	Yes / No	12
USeq v4.2	Window scan	Binomial *P* value	Estimated or user specified	Subtracted before peak calling	*q* value	1, 2: binomial 2: # control / # ChIP	Target FDR	No / Yes	20

Figure 1.11: **ChIP-Seq Peak Calling Algorithms.** [a]*The labels 1: and 2: refer to one-sample and two-sample experiments, respectively.* [b]*These descriptions are intended to give a rough idea of how control data is used by the software. 'NA' means that control data are not handled.* [c]*Description of how FDR is or optionally may be computed. 'None' indicates an FDR is not computed, but the experimental data may still be analyzed; 'NA' indicates the experimental setup (1 sample or 2) is not yet handled by the software. # control / # ChIP, number of peaks called with control (or some portion thereof) and sample reversed.* [d]*The lists of 'user input parameters' for each program are not exhaustive but rather comprise a subset of greatest interest to new users.* [e]*'Strand-based' artifiact filtering rejects peaks if the strand-specific distributions of reads do not conform to expectation, for example by exhibiting extreme bias of tag populations for one strand or the other in a region. 'Duplicate' filtering refers to either removal of reads that occur in excess of expectation at a location or filtering of called peaks to eliminate those due to low complexity read pileups that may be associated with, for example, microsatellite DNA.* [f]*N+ and N- are the numbers of positive and negative strand reads, respectively.* Source: Pepke et al.[59]

methodology for determining the sequence specificity of TFs.[71] A TF of interested is purified, either in its native state, or genetically fused with a tag. The PBM contains naked double-stranded DNA and the binding site of the TF is determined by investigating the hybridisation signal of the TF with the DNA.

In silico strategies for motif discovery

Apart from experimental methods of determining binding locations and sequence motifs of TFs, a number of computational strategies exist as well. The difference in the strategies lies in the method to maximise the signal-to-noise ratio necessary to retrieve the correct sequence motif. There are two common approaches: integration of coexpression information and integration of phylogenetic information.

When using coexpression information, the underlying assumption is the fact that coexpressed genes are more likely to share a common motif. But although it is true that genes with a shared motif show, on average, a higher degree of coexpression, is has been concluded in *Drosophila melanogaster* that the reverse is not true.[72] Nevertheless, this assumption has proven to be valuable when trying to detect motifs in coexpression datasets. The second assumption is that the coexpression is the result of coregulation at the level of transcriptional initiation. Because of the assumptions, motif discovery on these kinds of datasets involves mining for enriched motifs in the coexpression clusters. This means that not all genes in the module are required to have the motif because of the incomplete correctness of the assumptions.[25,73,74]

A second strategy is the incorporation of phylogenetic information. The underlying idea is that functional motifs will be conserved to a higher degree than non-functional regions of the promoter.[75–77] As such, functional motifs can be found overrepresented in the different regulatory regions of homologous genes. Phylogenetic information can be used as an integrative method to improve the motif discovery with coexpression information[78] or as a stand-alone strategy where all non-coding regions of interest of entire genomes are scanned for stretches of sequence that are conserved in the orthologs.[77]

Two approaches exist based on this idea: phylogenetic footprinting and phylogenetic shadowing. The difference lies in the number of species and the evolutionary distance involved. In phylogenetic footprinting, typically two species are used that are distantly related, e.g. human and mouse. This method retrieves ancient motifs that were already present in the common ancestor of the compared species. In phylogenetic shadowing, multiple closely related species are compared, e.g. multiple primates, and this allows detection motifs that arose more recent in evolution.[74,79] These multiple species are necessary because of the close evolutionary distance between them. All the evolutionary distances are 'summed' to have enough divergence, allowing to distinguish between conserved and non-conserved sequences. A more elaborate introduction of this topic can be found in section 5.1.

Similar to the indirect experimental methods, the disadvantage of these methods is the fact that they do not establish a regulatory link between a TF and its targets. The result of these predictions is a sequence motif, but no knowledge of which TF binds the motif. Based on known TF binding motifs and/or expression information on the possible upstream TFs, potential regulatory links can be inferred but this is prone to error. Network inference in itself is an entire field on its own, with many algorithmic approaches[80].

1.7 METHODOLOGY: Profiling the chromatin

Apart from knowing which TF can bind which promoter, we need to account for the chromatin state before we can expand the binding event to a generalised interaction. Ideally, ChIP experiments for TF binding should be accompanied by chromatin profiling experiments in the same condition. Only when we have such a complete view on a binding event, will we be able to exact rules for TF binding.

DNA Methylation

DNA methylation is the process where cytosine residues receive a methyl group on their structure and functions in transposon silencing and gene regulation (Figure 1.5). Loss of methylation leads to developmental aberrations due to wrongful gene activation, and loss of transposon silencing.[81–83] Methylation patterns are often inherited, and can thus be an epigenetic modification. The robust inheritance of DNA methylation does not lie in single site methylations, but rather in larger regions of contiguous methylation.[84,85] The mutation rate for these larger regions are similar to that of classic DNA mutations. Because

DNA methylation exhibits a clock-like accumulation in a geographically dispersed Arabidopsis population, it also reflects genetic distance.[86]

Chromatin Modification

Apart from direct methylation of the DNA, a number of additional modifications can be applied to the histones of the chromatin (Figure 1.5). The different modifications known are methylation (not to be confused with direct DNA methylation), acetylation, ubiquitinylation, sumoylation, phosphorylation, ADP ribosylation, deimination and proline isomerisation.[87] Note that only when these marks are stable and inherited after cell division, the modifications can be called epigenetic. The study of chromatin modifications is a field on its own, which is clear from the elaborate reviewing in Arabidopsis in the context of —among others —seed performance and plant development[88,89], gene responsiveness[90], bud dormancy[91], and flower development.[92]

Apart from the study of different modifications, integrative studies aim at determining the different quantitative combinations in which they occur across the genome, as well as their genomic context. The latter has given rise to the notion of chromatin states. The first study to define four chromatin states was Roudier et al.[14], based on twelve modifications on chromosome 4 (including DNA methylation), marking active genes, repressed genes, silent repeat elements and intergenic regions. More recently, Sequeira-Mendes et al.[93] and Wang et al.[88] have defined nine and six states based on 11 histone modifications (combined with CG methylation, nucleosome occupancy, and three histone variants) and 13 histone modifications(combined with two histone variants and DNA methylation) respectively.

DNase I and MNase Hypersensitivity

DNase I hypersensitivity (i.e. DNA that is extensively cleaved upon addition of DNase I) provides a method to map changes in chromatin structure. The method is based on the difference between generalised sensitivity and hypersensitivity. Both are linked to the open chromatin state but the former is inherent in all actively expressed genes while the latter refers to regions showing extreme sensitivity in short stretches of DNA ranging from 100 to 400 bp in length. These are likely to harbour functional motifs. The regions can be determined at various resolutions ranging from a few hundred bases to a single nucleotide. Given a detected region, the motif can be determined using different follow-up experiments.[67,74] Note that this method does not require prior knowledge about the TF.

Complementary to DNase I, which maps the open regions, micrococcal nuclease (MNase) maps occluded regions. Although similar in the principle that both nucleases will cut accessible DNA, MNase will digest DNA until it is prohibited by a DNA-binding protein. As such, the regions that are profiled are the occluded ones rather than the open ones[94,95]. A review of the different methodologies on profiling nucleosomes is provided by Zentner and Henikoff[96].

Chromatin Folding

Finally, the spatial organisation of the chromatin is of importance for bringing TFBSs close to the promoter (Figure 1.3) on which they act. A nice example is the promoter looping required for the correct expression of FLC.[97] The latter exemplifies the interplay between chromatin organisation and chromatin modification since the loop is disrupted during vernalisation by polycomb dependent epigenetic silencing.

With the development of the of the chromosome conformation capture (3C) method[98], and its successors chromosome conformation capture-on-chip (4C)[99], chromosome conformation capture carbon copy (5C)[100], and Hi-C[101] techniques, it has become possible to identify chromatin regions that lie in adjacency of each other. The 3D packing of the chromatin is not random as it needs to be efficiently untangled when there is need for transcription. Overall, the Arabidopsis chromatin interacts in 3D following the linear strand: most interactions for a given region are with the adjacent regions on the strand.[88,102,103] Nevertheless, special structures such as the KNOT[104] and positive strips of long-range interaction have already been identified.[88]

1.8 Glossary of Terms

Throughout the Introduction and results, some databases, tools and measures are mentioned that could use further clarification. Rather than occluding the main text with these explanations, many of these terms are shortly explained in the glossary below.

- **GO (Slim):** Gene Ontology. A dynamic, controlled and structured vocabulary that is used to annotate genes. GO a slime is a trimmed version of the GO structure to get a broader overview of gene annotations.[105]
- **AraNet:** Bayesian network of gene-gene functional associations in Arabidopsis *thaliana*.[106]
- **CORNET:** Web tool containing protein-protein interactions and expression data for Arabidopsis *thaliana*.[107]
- **AtRegNet:** Database of regulatory interactions in Arabidopsis thaliana.[17]
- **TAIR:** The Arabidopsis information Resource. Web portal collecting, among other things, all functional data of Arabidopsis *thaliana*.
- **AGRIS:** Database containing DNA motifs in Arabidopsis *thaliana*.[17]
- **PLACE:** Database containing DNA motifs in Arabidopsis *thaliana*.[108]
- **Embryo-lethal genes:** Genes that when mutated lead to a non-viable individual.[109]
- **AthaMap:** A genome-wide map of potential transcription factor and small RNA binding sites in Arabidopsis thaliana
- **PhosPhAt:** The Arabidopsis Protein Phosphorylation Site Database containing interactions between kinases and targets.[110]
- **GEO:** Gene Expression Omnibus. Database containing metadata and results of publicly available gene expression and ChIP experiments.[111]
- **SRA:** Short Read Archive. Database containing the raw data of publicly available next-generation sequencing experiments.
- **MapMan:** A dynamic, controlled and structured vocabulary that is used to annotate genes.[112]
- **psRNAtarget:** a tool to predict miRNA target sites.[113]
- **PMRD:** Plant miRNA Database. Database containing miRNA genomic coordinates in plants.[114]
- **Expression Coherence:** the fraction of gene pairs in a module that exhibit significant coexpression out of all possible gene pairs.
- **CAST:** Cluster Affinity Search Technique. Network-based clustering algoritm that assigns genes to clusters based on their affinity towards the entire cluster iteratively.[115]
- **Weeder:** Exact word-based *de novo* motif finding tool.[116]
- **MotifSampler:** Position Weight Matrix-based *de novo* motif finding tool.[117]
- **MotifRanking:** Tool to collapse different motifs into a single core one (if possible).[117]
- **Peak-Motifs:** Tool determine enriched DNA motifs in a ChIP dataset.[118]
- **BWA:** Tool map next-generation sequence reads to the genome.[119]
- **STAMP:** DNA motif alignment webtool and comparison to databases of motifs.[120]
- **MACS:** Model-based Analysis of ChIP-Seq. Tool for calling significantly enriched regions in a sample versus control setting of a ChIP-experiment.[52]

- **rMAT:** R version of MAT, the Model-based Analysis of Tiling Array. Tool used for ChIP-chip.[121]
- **Starr:** R package for ChIP-chip analysis.[122]
- **VCF tools:** Tool package for parsing VCF files (Variant Calling Format) and performing population genomics statistics.[123]
- **BEDtools:** Tool package to perform operations on different BED-format files.[124]
- **LASTZ:** DNA Sequence alignment tool that can cope with repeat-rich sequences (e.g. genomes).
- **MULTIZ:** DNA alignment tool that can handle inversions and duplications.[125]
- **PhyloP:** Tool to determine the conservation in an alignment.[126]
- **PhastCons:** Tool to identify sequence constraint in an alignment.[127]
- **DialignTX:** Pairwise DNA alignment tool.[128]
- **Sigma:** Pairwise DNA alignment tool specifically designed for non-coding DNA.[129]
- **ACANA:** Pairwise DNA alignment tool.[130]
- **Seaweeds:** Pairwise DNA alignment tool based on a moving window.[131]
- **Polycomb Repressive complex:** A complex of polycomb family proteins that function in gene regulation by remodelling the chromatin nearby their target genes.[132]

Chapter 2

Research Aims and Scope

Changes in gene expression have been observed under a wide variety of conditions and changes in environment: cell differentiation, stress response, etc. Because of its importance, there has been great effort in trying to unravel the process (and its elements) of regulated gene expression. This PhD thesis focuses on plants, which exhibit the a great range of environmental responses.[133] Because plants are sessile, it is of vital importance that they be able to cope with different conditions each meteorological season brings by versatile gene expression.

The primary objective of this project was to study the organisation of transcriptional regulation based on experimental high-throughput data. The research will be subdivided in two levels: the module level and the gene level. On the module level we will study genes that are functionally associated, and investigate how different associations relate to the regulation of the module. As many genes in Arabidopsis lack functional information to date, an important secondary objective was to provide reliable functional annotation for as many genes as possible. On the gene level, we will investigate the organisation and conservation of gene expression networks.

Initially, we will study the organisation of gene regulation through the context of functional gene modules. Modules will be built based on an integration of protein-protein interactions, TF-targets, and Gene Ontology association, and coexpression. For each of the module types, we will assess whether there is evidence for the genes being regulated by the same regulators. In addition, we will investigate the conservation of the regulation of modules by building orthologous modules. The modules will also be used for function predicting using the guilt-by-association principle, where conserved regulation is an added evidence of the correctness of the prediction. The transfer of modules to other species will facilitate translational research from model species to crops.

Next, all available Arabidopsis ChIP-Seq data sets will be collected, and an automated pipeline will be developed to analyse all datasets in a homogeneous manner. Next, the general properties will be analysed per data set. More specifically, we will study (i) which regions of the genome are bound by TFs (intergenic, intronic, exonic, etc.); (ii) whether a bound region can be unambiguously assigned to a target gene; (iii) whether a DNA motif can be found within the bound region; (iv) how many regions a TF binds; (v) whether a correlation can be detected between the location of a bound region and (a combination of) chromatin signatures, and (vi) what the nucleotide variation is in bound regions across the Arabidopsis thaliana population. This will allow us to detect long-range enhancers, disclose indirect regulation and link the regulatory code to e.g. histon modifications.

Subsequently, we will study the conservation of bound regions between related species. For a given binding site, the regulatory sequences will be aligned with the orthologous regions of other species within the dicotyledonous species. Based on these alignments, the species-specificity of binding sites will be evaluated using a conservation score. Using this score, we will evaluate whether the genomic position of a TFBS is a determining factor in its conservation, and whether conservation scores are similar for different biological processes.

Chapter 3

Systematic identification of functional plant modules through the integration of complementary data sources[a]

Abstract

A major challenge is to unravel how genes interact and are regulated to exert specific biological functions. The integration of genome-wide functional genomics data, followed by the construction of gene networks, provides a powerful approach to identify functional gene modules. Large-scale expression data, functional gene annotations, experimental protein-protein interactions, and transcription factor-target interactions were integrated to delineate modules in Arabidopsis thaliana. The different experimental input data sets showed little overlap, demonstrating the advantage of combining multiple data types to study gene function and regulation. In the set of 1,563 modules covering 13,142 genes, most modules displayed strong coexpression, but functional and *cis*-regulatory coherence was less prevalent. Highly connected hub genes showed a significant enrichment towards embryo lethality and evidence for crosstalk between different biological processes. Comparative analysis revealed that 58% of the modules showed conserved coexpression across multiple plants. Using module-based functional predictions, 5,562 genes were annotated and an evaluation experiment disclosed that, based on 197 recently experimentally characterized genes, 38.1% of these functions could be inferred through the module context. Examples of confirmed genes of unknown function related to cell wall biogenesis, xylem and phloem pattern formation, cell cycle, hormone stimulus, and circadian rhythm, highlight the potential to identify new gene functions. The module-based predictions offer new biological hypotheses for functionally unknown genes in Arabidopsis (1,701 genes) and six other plant species (43,621 genes). Furthermore, the inferred modules provide new insights into the conservation of coexpression and coregulation, as well as a starting point for comparative functional annotation.

[a] This chapter is based on Heyndrickx and Vandepoele [134]. KSH and KV designed the study and wrote the manuscript. KSH performed the analyses and created all figures.

3.1 Introduction

The sequencing of *Arabidopsis thaliana* (hereafter Arabidopsis) and the emergence of high-throughput functional genomics techniques like microarrays, systematic T-DNA knock-out
screens, and protein-protein interaction (PPI) mapping have enabled the development of integrative approaches to study gene function and regulation. One of the major challenges of computational biology is the integration and exploitation of genome-wide data sets such as transcriptome and interactome data, metabolomics and other -omics data, and large-scale phenotyping.[135] Data integration is often performed through gene network analysis[106,136] and the resulting networks can increase our knowledge of functional gene relationships and the interplay of different types of interactions. However, to study specific biological processes, networks are frequently studied through gene modules.[19] From a practical point of view, modules are typically identified as highly connected subgraphs within the network.[137] Depending on the type of interaction data, different types of modules are defined and examples in Arabidopsis include coexpression modules[138–140], protein complexes[107,141–143], and modules grouping genes that are regulated by the same transcription factor (TF).[144] Genes can be part of different (sometimes overlapping) modules, while modules can be involved in different biological processes. As a consequence, gene networks are frequently highly connected, revealing the pleiotropic roles of different genes. Consequently, the module context can be explored to identify genes that are present in many different modules and that have a functional association with many other genes (hub genes[137]). These hub genes represent important components of biological systems and can provide crosstalk between different processes.

Modules based on expression data are typically inferred through clustering of genes with similar expression profiles. Most often, each gene pair receives an expression similarity measure and this coexpression information is used to detect highly connected sub-graphs in the coexpression network, representing modules. Although numerous expression network analyses have been performed in Arabidopsis, some studies focused on a specific process using guide-genes (genes known to function in the process) to draw new hypotheses about the functional interplay between functionally known and unknown genes based on guilt-by-association.[138,145,146] Other studies employed module delineation and guilt-by-association on a genome-wide scale to predict gene functions.[139,140,147–150] From a regulatory point of view, module genes are often regulated by multiple *cis*-regulatory elements (referred to as motifs), organized into *cis*-regulatory modules (not to be confused with the gene module).[151] Therefore, coexpression modules are often used to investigate the *cis*-regulatory elements controlling the genes within the modules using known DNA motifs or *de novo* motif finding.[149,150,152]

A disadvantage of coexpression analysis is the false assumption that coexpressed genes are *de facto* coregulated.[153] The emergence of chromatine immuno-precipitation (ChIP) allows the direct profiling of the regions bound by a TF and the detection of TF target genes. The technique can be applied in a genome-wide fashion when followed by a whole-genome tiling array (ChIP-chip) or deep sequencing (ChIP-Seq).[144] The ChIP technique provides a snapshot of the regulatory binding state of the genome by cross-linking all proteins to nearby bound DNA. In Arabidopsis, ChIP-chip/Seq has been applied to a range of TFs, primarily those active in flowering and development. Because of the static nature of a ChIP experiment (it is a snapshot of the biological state), the genome-wide profiling of TF binding sites is often combined with differential expression analysis in a knockout[154–157] or an inducible overexpression line.[46,158–160] By combining these two data types, TF target interactions can be viewed with respect to the expression of both the TF and the target, thus transforming the static ChIP image to a set of dynamic transcriptional modules. A third type of modules is based on PPI networks. Although there have been several PPI studies in Arabidopsis, their main focus lay on building the interactome, rather than on breaking down the network to the module level.[141,143,161–163] Studies that did explore the network module contexts, found modules recapitulating known biological functions and also suggesting new biological hypotheses for several plant-specific genes, often through the integration with expression data.[107,142,143,164]

Although several plant studies performed some kind of data integration when delineating gene modules, the number of data types is often limited. Recently, a few Arabidopsis studies have been published reporting large networks for function prediction based on multiple data types. These networks were built combining expression and PPI data with sequence data[136], genetic and physical interaction data[165], phylogenetic profiles and gene location[166], and the integration of functional genomics, proteomics and comparative genomics data sets.[106] Apart from studying gene modules in a one species, recent studies have applied comparisons across species to identify conserved gene coexpression in plants.[167–169] The analysis of coexpression networks between more distantly related species exploits the assumption that predicted gene function associations, occurring by chance within one organism, will not be conserved in a multi-species context. Consequently, the analysis of conserved modules with specific functions provides an invaluable approach for biological gene discovery in model species and for translation of new gene functions into species with agricultural or economical value.[170]

In this study, we investigated how Arabidopsis genes are organized into gene modules based on four different data types (Gene Ontology or GO, PPI, ChIP, and AraNet) and studied the functional and regulatory properties of these modules. Furthermore, module evolution was examined by integration of orthologous sequences and expression data of six related plant species. Overall, our results revealed that currently available experimental data sources are highly complementary, different functional categories show varying levels of regulatory complexity, a large fraction of Arabidopsis gene modules is conserved in other plant species, and conserved modules provide a valuable source to study gene functions.

3.2 Results

Construction of Arabidopsis gene modules using experimental and computational gene associations

Based on an ensemble of primary data sets covering TF target interactions from AtRegNet[17], probabilistic gene-gene associations from AraNet[106], non-electronic gene-GO annotations (see list of evidence codes in Material and Methods) from TAIR[171], and PPIs from CORNET[107], functional gene modules were delineated in Arabidopsis (Table 3.1). To assemble a set of high quality gene associations, the GO, PPI, and TF targets data were filtered to only contain experimental information (see Materials and Methods). In contrast, the AraNet data is an integration of 24 distinct types of gene associations (e.g. coexpression, PPI, shared protein domains, similarity in phylogenetic profile, orthology) including both experimental and computational observations. In total, the final input data set covered 22,492 unique genes and > 1 million interactions, with the largest fraction coming from the AraNet network. Nearly all gene associations were unique to one input data type, with the fraction of unique associations ranging from 75% for PPI to 99% for AraNet and TF targets (Table 3.1).

Table 3.1: **Overview of the primary data sets and delineated modules with their properties.**

Data Type	Primary Data Sets		Modules			
	No. of Genes	No. of Associations (% Unique)[a]	No. of Genes	No. of Modules (% Unique)[b]	Functional Enrichment[c]	Motif Enrichment[d]
PPI	3,194	7,210 (75%)	597	72 (95%)	51	43
AraNet	19,647	1,062,222 (99%)	6,377	419 (99%)	116	172
TF targets	9,422	13,037 (99%)	5,127	518 (96%)	51	224
GO	6,588	89,100 (n.a.)	7,750	1,105 (99%)	943	341
Total	22,492	1,089,661	13,428	2,114	1,161	
Nonredundant modules			13,142	1,563	676	772

[a] Percentage of associations unique for this data type. As GO does not consist of pairwise gene-gene associations, no unique fraction is reported. n.a., Not available. [b] Percentage of modules unique for this data type (based on the output of detecting redundant modules across different input data types). [c] Based on BP GO categories and experimentally annotated embryo-lethal genes. [d] Calculated for the nonredundant modules only.

To delineate gene modules from the different gene association data sets, two clustering strategies were applied (Figure 3.1). Firstly, for the TF targets and GO data, expression information was integrated to cluster genes into modules (expression-based clustering, see Materials and Methods). This was done because the TF target ChIP data provides a static image of genome-wide TF binding and as a consequence, TF target genes do not necessarily form functionally coherent modules. By integrating expression data, these static images are converted into spatial-temporal TF target maps. Similarly, GO categories do not represent functionally coherent gene modules.[150] Therefore, per GO category, genes with non-electronic GO annotations were used as prior information to guide the creation of coexpression clusters using different expression compendia from CORNET.[107] Genes used as guides are referred to as seed genes in the remainder of the manuscript. Different Arabidopsis expression compendia (see Materials and Methods) were used because the degree of coexpression can be influenced by the specific expression data used.[172] Therefore, genes from GO categories were clustered using the compendium in which the coexpression was the highest, measured by Expression Coherence (EC). EC is a measure for the amount of expression similarity within a set of genes for a given expression compendium (see Materials and Methods). All GO categories across the three GO hierarchies 'Biological Process', 'Molecular Function', and 'Cellular Component' (abbreviated as BP, MF, and CC, respectively) were used as sources for seed genes to build modules of different specificity (i.e. general versus very specific processes). As many genes in Arabidopsis have not yet been functionally annotated, many GO categories are incomplete. To overcome this problem, GO category-based seed sets were expanded with genes showing high coexpression with the seed genes prior to the clustering (Multi-Query Seed Expansion, MQSE; see Materials and Methods). Since different TFs can regulate the same gene and genes can be associated with multiple GO categories, genes can belong to more than one resulting module. Secondly, PPI and AraNet gene associations were clustered based on the connectivity of the genes in their respective input networks without linking to expression data (referred to as connectivity-based clustering, see Materials and Methods). As a consequence, highly connected sub-graphs were identified in both networks to delineate PPI and AraNet modules, respectively.

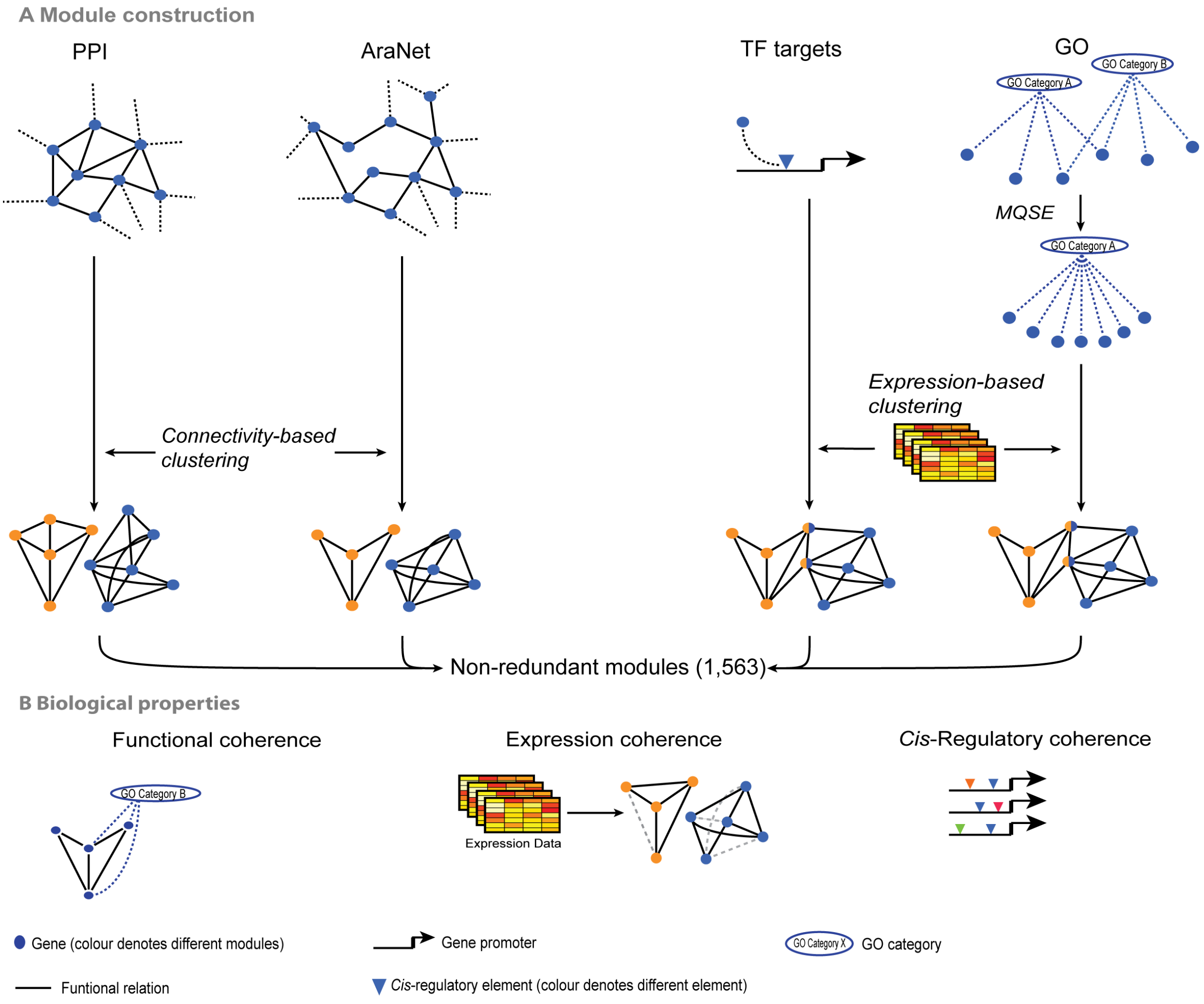


Figure 3.1: **Delineation of functional gene modules.** *A, Four different primary data sets were processed to extract functional gene modules, resulting in 1,563 non-redundant modules. Data types are in roman font, methods are in italic font. B, Biological properties (functional coherence, expression coherence, and cis-regulatory coherence) of gene modules were characterized. Dotted lines indicate gene-GO associations and non-significant PCCs for the functional coherence and the expression coherence panels, respectively. In the cis-regulatory coherence panel, the blue triangle represents an enriched motif.*

All modules from the different input data types (PPI: 72, AraNet: 419, TF targets: 518, GO: 1,105) were compiled into one final dataset covering 2,114 coexpression modules derived from GO, transcriptional modules derived from TF targets, PPI modules derived from the PPI network, and AraNet modules. To determine the extent to which the different datasets complement each other, the overlap between the different data types was assessed (see Materials and Methods). On the level of gene content, 40% of the genes in the modules is present in more than one input data type (Figure 3.2A). However, the overlap based on the gene-gene associations both in the input (Table 3.1) and the module associations was drastically smaller with only 3% of the gene pairs within a module having support by more than one primary data type (Figure 3.2B). After removing redundant modules based on the number of shared genes (see Materials and Methods), the final data set consisted of 1,563 modules comprising 13,142 genes (63% of all genes on the ATH1 microarray). Based on the redundant modules, the low overlap between different data types, was confirmed, as most modules (1,556 / 1,563) could only be found through a single data type (Figure 3.2C). Examples of modules confirmed by multiple data types (7) include genes related to amino acid metabolism and transport (see Table S1[a] for modules and gene sets discussed throughout the article). The majority of modules contained between five to ten genes (50%), while larger module sizes were increasingly less frequent (Figure 3.2D). These observations were in line with the notion of a hierarchical structure of biological networks, where smaller and more specific clusters reside within larger and more general clusters.[173]

Functional, expression and *cis*-regulatory coherence of plant modules

Based on the gene modules inferred through the different primary data types, we next sought to characterize different biological properties. The investigated properties describe the level of coexpression among the genes in a module, whether the module genes are potentially regulated by the same transcription factor, and whether a specific function or biological process can be linked to a module (Table 3.1; Figure 3.3). An additional website[b] is available to browse modules, genes, coexpression information, primary gene associations, functional annotations, and motifs.

For each module, the level of coexpression was determined using EC. To minimize the possible influence of the specific expression data set used to determine the level of coexpression, EC scores were initially calculated for each module based on a global compendium and other specific compendia, and only the maximum EC score was retained for further analysis. Note that for GO and TF targets, these compendia correspond with the expression data used to delineate the modules. Overall, for the non-redundant set of 1,563 modules, the median EC score was above 50%, indicating that coexpression is an important property of most modules (Figure 3.3A). Comparing the maximal EC scores for modules derived from different primary data types, revealed that coexpression levels were also high for PPI and AraNet modules (98.6% and 88.5% show significant EC), despite the fact that expression information was not directly integrated during the module delineation. At the 10% EC threshold, which corresponds with a p-value ≤ 0.02 (based on randomized gene modules, see Materials and Methods), the difference between the EC scores from the global and specific expression compendia was the largest for the TF target modules.

To assess the *cis*-regulatory module properties (*cis*-regulatory coherence), *de novo* motif finding was performed to identify putative transcription factor binding sites in the 1-kb promoters of the genes. The motif finding was performed with the complementary tools Weeder and MotifSampler.[152,174,175] To discard potentially false motifs, enrichment analysis was performed and only motifs showing significant enrichment within a module were retained (q-value ≤ 0.01). Redundant motifs within modules were removed based on sequence similarity and gene-motif occurrences (see Materials and Methods), resulting in 1,544 different motifs in the modules. MotifSampler and Weeder exclusively supported 1,190 (77.1%) and 285 (18,5%) motifs, respectively, while 69 (4.5%) motifs were supported by both tools, emphasizing their complementarity. To validate the reliability of motifs found by only one tool, the overlap of motifs found by MotifSampler or Weeder was compared with a set of 515 known motifs from PLACE[176] and AGRIS.[17] Of the 1,544 *de novo* motif instances in modules, 528 corresponded to a known motif.

[a] http://www.plantphysiol.org/content/suppl/2012/05/15/pp.112.196725.DC1/196725Table_S1.xls
[b] http://bioinformatics.psb.ugent.be/cig_data/plant_modules/

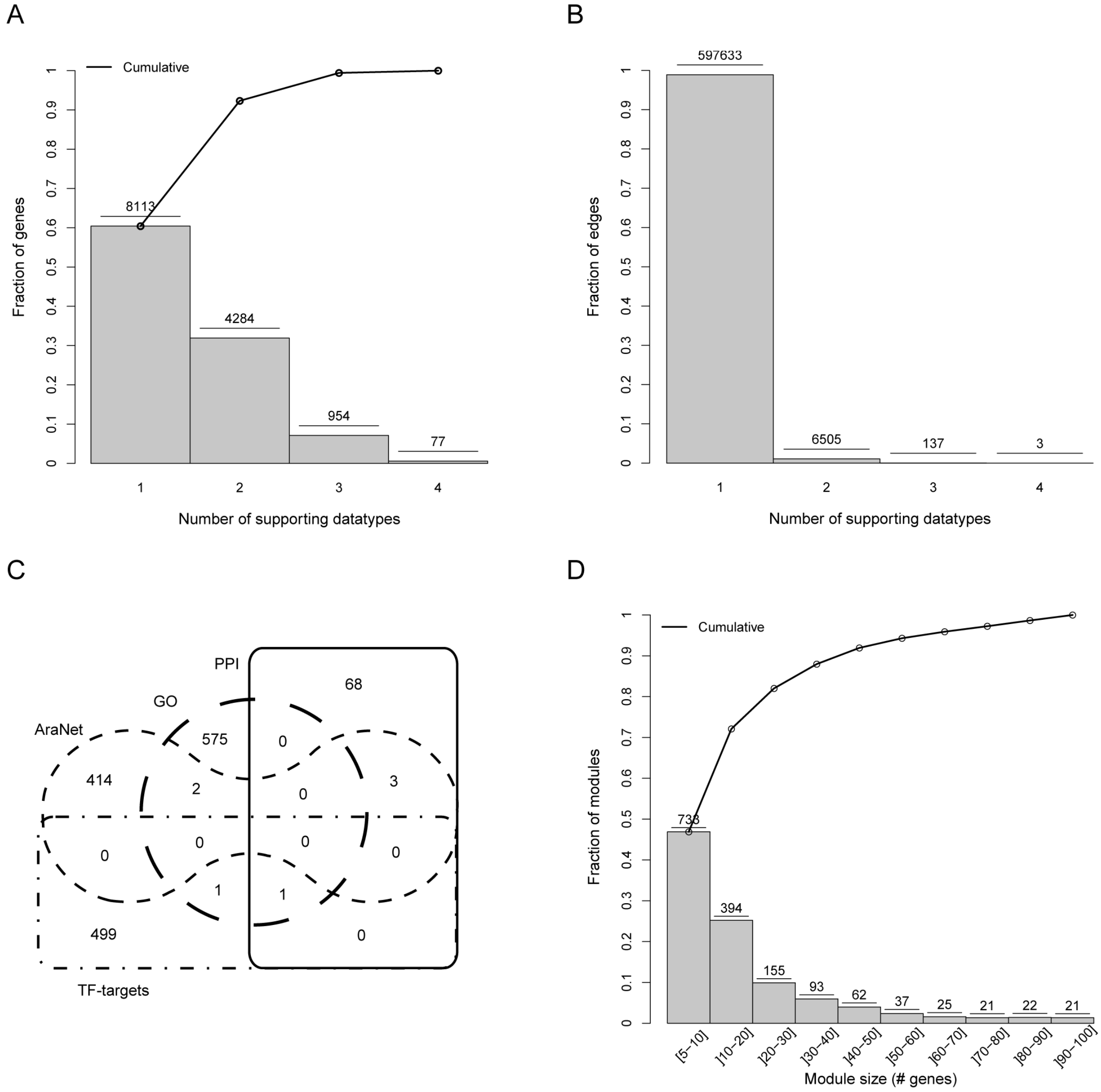


Figure 3.2: **Basic properties of the derived functional gene modules.** *A, Number of different module types per gene. B, Number of different input data types per module edge. C, Overlap between the different types of modules. D, Gene size distribution for the set of 1,563 non-redundant gene modules.*

For these 528 known motif instances, 408 (77.3%) and 71 (13.4%) were found uniquely by MotifSampler and Weeder, respectively, and 49 (9.3%) were retrieved by both tools. In addition, both methods reported a similar but complementary fraction of known motifs (MotifSampler 408/1,190 [34.3%] and Weeder 71/285 [24.9%]) among their total number of reported motifs. To facilitate downstream analysis, the combined set of *de novo* motifs and known motifs from PLACE and AGRIS was grouped into 813 motif families based on sequence similarity (see Materials and Methods). Within these *de novo* motif families, 65 contained a known motif, while 748 families contained purely *de novo* motifs. Finally, the *cis*-regulatory coherence was defined as the fraction of modules with at least one enriched motif (Figure 3.3B). *cis*-regulatory coherence scores ranged from 40% (AraNet: 172/419 and TF target: 224/502) to 60% (PPI: 43/72 and GO: 341/579). In total, 49.4% of the non-redundant set of modules contained at least one motif (772/1563). A weak but significant (R^2 = 0.03; p-value < 1.42e-11) relation was found for the number of different motif families in one module in function of EC. Apart from the *cis*-regulatory coherence analysis, these motifs provide an important resource to annotate and map specific TF target interactions at the module level.

The functional coherence was determined by GO enrichment analysis for non-electronic biological process annotations and enrichment for genes associated with embryo lethality. Information about genes

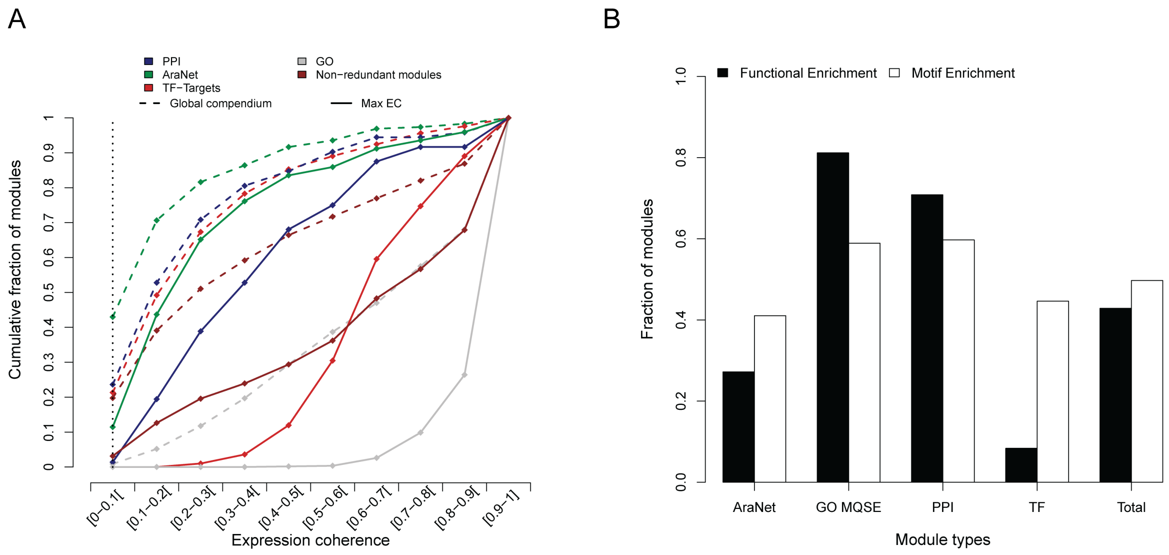


Figure 3.3: **Functional, expression and *cis*-regulatory coherence.** *A, Comparison of expression coherence scores between the modules from different input data types. The EC scores are shown for both the general compendium (dotted line) and the compendium showing the maximum EC (solid line). The vertical dotted line indicates the threshold for significant EC. B, GO-BP and motif enrichment statistics for the modules delineated using the different input data types.*

involved in embryo lethality was based on the SeedGenes database.[109] The functional coherence revealed large differences between modules from the different primary data types (Table 3.1; Figure 3.3B). As expected, the GO modules showed the highest functional coherence (80% of the modules). While for AraNet and PPI, respectively 27% and 72% of the modules showed functional coherence, the TF targets data had the lowest functional coherence (10% of the modules). Overall, 40% of the modules could be linked to a significantly enriched biological process or embryo lethality, while 98% of the modules contained one or more genes with a known experimental annotation. To obtain an overview of the different biological processes in which the modules were involved, the module predictions were categorized according to their GO Slim terms (Figure 3.4). Control experiments indicated that there were no significant enrichments towards any GO category in either the complete set of input genes, nor the complete set of resulting modules.

Hub genes and organization of transcriptional regulation in Arabidopsis

Genes can have pleiotropic roles and can thus be involved in multiple processes or modules. Because of the different input data types and the way different GO categories were used to guide module detection using MQSE, genes can occur in multiple though non-redundant modules. Hub genes[137] were identified as genes that are present in a large number of modules and are possibly providing crosstalk between the different biological processes they are involved in. The number of modules per gene ranged from 1 to 26, following a power law, making the gene-module associations a scale-free network (Figure 3.5A; Figure A.1). Genes present in more than ten modules (116 genes, top 5%) were extracted as hub genes, and a functional enrichment analysis revealed that these genes are involved in immune response, photosynthesis, cell cycle, and carbohydrate metabolism (Table S1[a] and Figure A.2A), which is in accordance with earlier studies.[177,178] Among the hub genes, we found MEK1 (MAPKK), MPK11 and MPK4 (MAPK), SNAP33 (SNARE), RABH1C (RAB GTPase), and XLG2 (GTP-binding protein), revealing that several hub genes are involved in signal transduction. Evidence for crosstalk mediated by hub genes was found for chromatin modification and development, through the genes CYP71, AT5G63960, and FUSED. Light response and photosynthesis were found to be coupled through the genes LIL3:1, FBA1, ISPF, and DXR. Finally, SYP121, SYP122, ATPAD4, NSL1, PBS3, WRKY70 (TF), JAZ1, ATNPR1, ATRCD1, ATCTL1, and AT1G15430 (based on module-based GO prediction) describe the crosslink between response to biotic/abiotic stimuli and hormone signaling through jasmonic acid (JA) and salicylic acid (SA). In addition, hub genes are also three-fold enriched for embryo lethal genes, confirming the

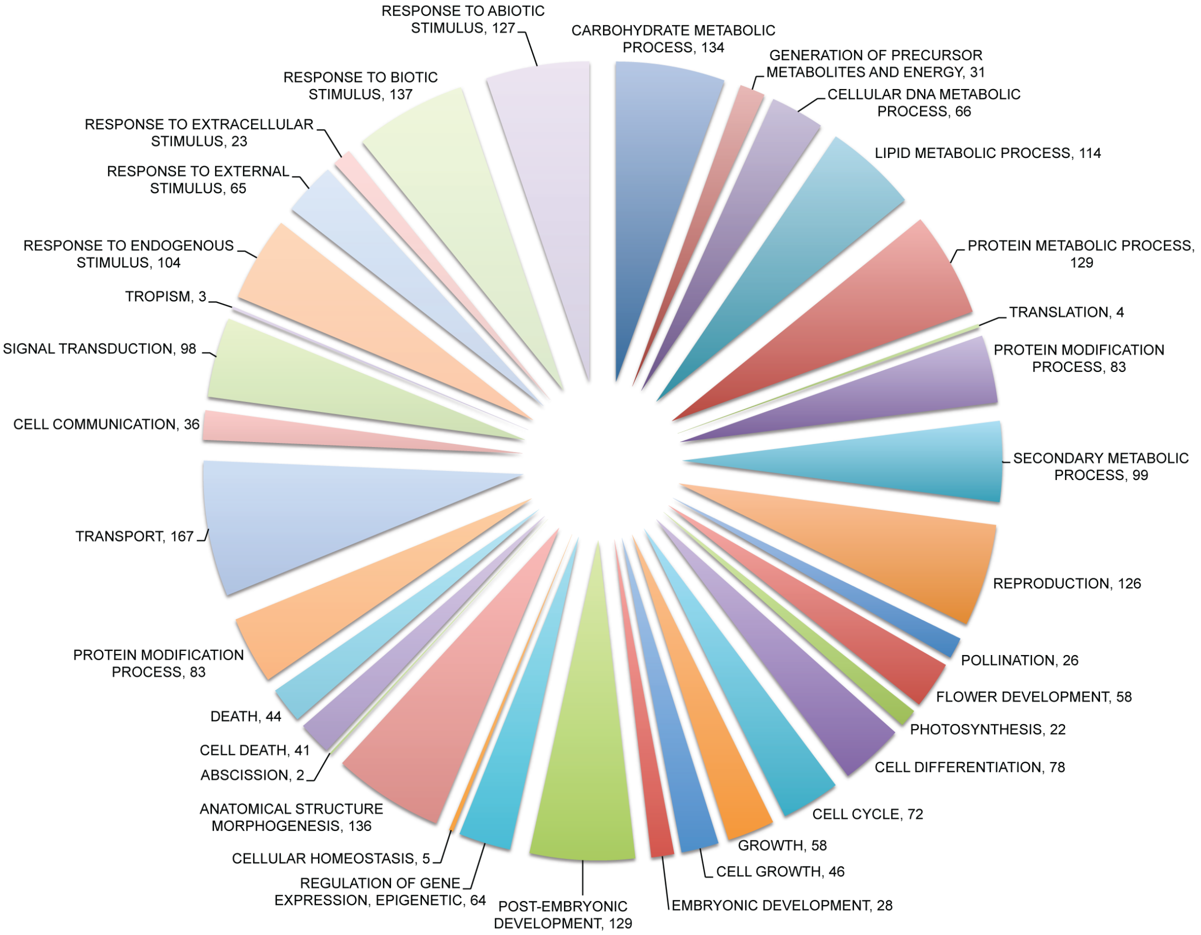


Figure 3.4: **Overview of GO-BP slim biological processes in which modules were predicted to be involved in.** *Modules with multiple GO-BP annotations can be present in different GO slim categories.*

relationship between network connectivity and essentiality.[173]

Besides investigating gene-module organization, the organization of motifs was also examined at the module and gene level. On the module level, the number of motifs ranged from zero to eight (Figure 3.5B) and modules regulated by five or more motifs (approximately 2%; Table S1[a]) are involved in processes associated with flower development, protein synthesis, and stimulus responses. On the gene level, the number of motifs per gene varied from zero to 26 (Figure 3.5C). Genes are mostly regulated by one to five motifs, but approximately 2% are regulated by more than ten motifs. These highly regulated genes are involved in cell cycle, systemic acquired response, and salicylic acid signaling (Table S1[a]).

To define the regulatory complexity of a gene, the number of modules and the number of motifs were combined in one plot (Figure 3.5D). A gene is considered complexly regulated when present in multiple modules and harboring multiple motifs. A significant positive correlation was found between the number of motifs and the number of modules (adjusted $R^2 = 0.40$; p-value $\leq 2.2e\text{-}16$). Whereas for the GO-BP slim main category 'Biological process' the linear fit followed the 1:1 line, not all genes follow this strict "one module - one motif" principle. Examining the module - motif relationships for different GO-BP slim subcategories revealed processes where genes were present in many modules, but without being regulated by many motifs. This indicates that, based on the number of motifs, hub genes are not necessarily regulated by many TFs (Figure A.2B). Carbohydrate metabolism, lipid metabolism, secondary metabolism, photosynthesis, cellular homeostasis, and generation of precursor metabolites and energy consistently showed a linear fit with a less steep slope, indicating more modules than motifs. Reversely, DNA metabolism and cell cycle showed a steeper slope than the main 'Biological process' category, indicating more motifs than modules and combinatorial regulation.

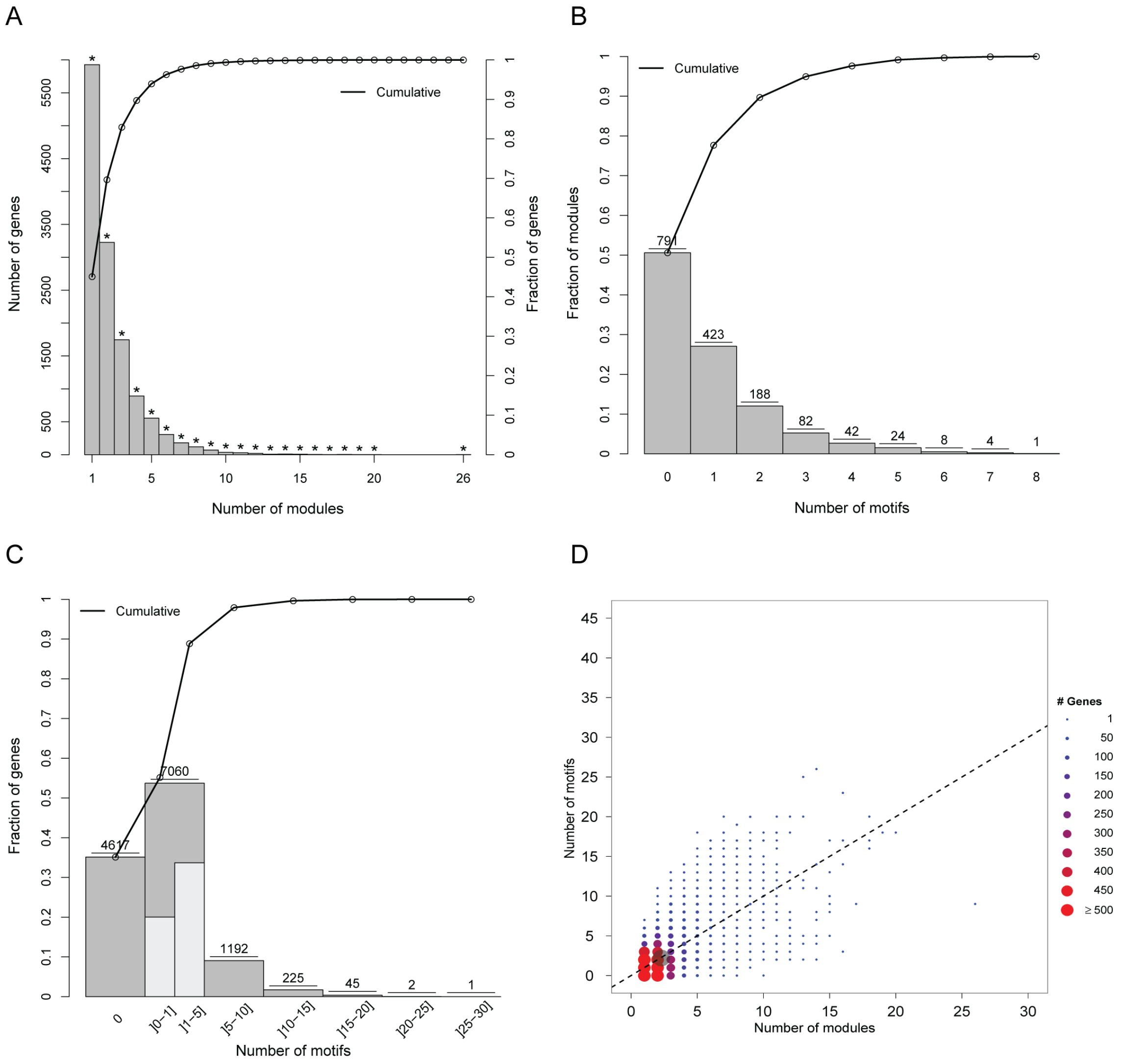


Figure 3.5: **Regulatory complexity of genes in modules.** *A, Number of modules in which a gene is present. Asterisks denote values higher than zero. B, Number of motifs per module. C, Number of motifs per gene promoter. D, Regulatory complexity, defined as a combination of the number of modules in which a gene is present, and the number of motifs in its promoter. All 13,142 genes are included and the number of genes at each coordinate is given as a colored size scale. The grey circle indicates the average regulatory complexity for all 13,142 genes. The dotted line is the function f(x) = x.*

When isolating the top 200 genes based on regulatory complexity (i.e. genes with a high number of modules and motifs), functional enrichments were found related to immune response, stress response, and cell cycle (Table S1[a]).

Conservation of gene modules in other plants

Based on the inferred Arabidopsis modules and their different biological properties, we next characterized if these modules are conserved in other plant species, since it has been shown that dynamic properties are primarily conserved at the module level.[179] The evolution of functional gene modules was examined using conservation of coexpression (EC) and conservation of regulatory DNA motifs (*cis*-regulatory coherence) based on orthologous genes in the dicots Glycine max (soybean), Medicago truncatula, Populus trichocarpa (poplar), Vitis vinifera (grapevine), and the monocots Zea mays (maize) and Oryza sativa ssp. japonica (rice). Orthologous modules were delineated using the PLAZA integrative orthology approach, which infers orthologous genes using complementary detection methods (i.e. phylogenetic trees, OrthoMCL families, and Best-Hits-and-Inparalogs families), which are considered as evidences.[180] For each Arabidopsis gene, the orthologous gene(s) with the highest number of evidences was (were) retained in each of the sampled species (Table S2[c]). Orthologous modules were subsequently constructed by grouping the orthologous genes based on with the Arabidopsis modules. Despite the potential problem of modules expanding significantly in size due to one-to-many orthology relationships, the applied ensemble approach retrieved one-to-one orthologs for on average (over the six species) 67% of the genes.

To study coexpression conservation, EC scores were calculated in the six species using publicly available microarray data (see Materials and Methods). For gene pairs with multiple orthologs, coexpression was considered present when at least one orthologous gene pair showing significant coexpression was found. Orthologous genes missing from the microarray were not taken into account. EC values of orthologous modules with less than five genes on the microarray of the respective species were marked as missing to distinguish them from zero values. The EC scores were compared to those of a set of random modules with the same gene size distribution (Figure A.3A) and based on these background scores, 910 (58%) modules with EC $\geq$ 10% in three or more species showed a significant conservation of coexpression (p-value $\leq$ 0.025; Table S1[a]). These conserved modules comprised a wide range of functions and biological processes, while modules with ultra-conserved coexpression (i.e. EC > 10% in 7 species, 92 modules; Table S1[a]) showed enrichments for processes linked with energy metabolism (e.g. NADPH metabolism, photosynthesis, starch biosynthesis).

For the set of modules with significant coexpression conservation in other plants, the conservation of *cis*-regulatory coherence was investigated, since conservation of both properties would strongly indicate conservation of regulation. To measure motif conservation, enrichment analysis for each of the motifs present in the original Arabidopsis modules was conducted in each of the species based on the promoter sequences of genes in the orthologous modules (Figure A.3B). Fifty-five percent of modules with conserved coexpression (500/910 modules) had at least one enriched motif in Arabidopsis, and based on the comparative motif analysis, we were able to confirm motif enrichment for 27.4% of these modules in at least one other species (137/500 modules; Table S1[a]). Four modules exhibited both expression and motif conservation in all seven species. These were involved in ribosome assembly, DNA modification, and response pathways and harbored motifs such as SORLIP2, SITEIIATCYTC, TELOBOX, UP1/2, BS1EGCCR, E2F, ABRE, and G-box. In contrast, 42% of modules without coexpression conservation had at least one motif in Arabidopsis (272/653 modules), but for only 5% of those modules, the motif enrichment was conserved (20/272 modules). This result showed that modules with conserved coexpression in other species are four-fold enriched in motif conservation compared to modules lacking conserved coexpression. The modules with conserved motif enrichment harbor 90 motif families (5% of all motif families), of which 67 represent new motifs and 23 were previously known. A detailed map associating motifs with specific functional categories is shown in Figure A.4.

Module-based functional annotation of unknown plant genes

Complementary to the cross-species analysis of different regulatory module properties, the conserved module contexts provide a promising resource for hypothesis-driven gene discovery in other plant species. The Arabidopsis sequencing project was succeeded by the Arabidopsis 2010 program, of which the goal was the annotation of all Arabidopsis genes by 2010[d]. Despite many efforts based on forward and reverse genetics, and computational predictions, functional annotation is still lacking for many genes. Although advanced computational gene function prediction tools have been developed[106,181], our main intention was to investigate how the integrated gene associations could lead to new functional hypotheses.

Since the initial download of the GO data for the module delineation (hereafter referred to as 'data freeze'), 2,940 genes belonging to the gene modules have received new experimental GO-BP annotations. Since these gene-GO associations were not available at the time of the module delineation, they form an ideal basis to evaluate the module-based gene function predictions inferred through the integration of the different primary gene associations. These new associations can be categorized in three groups: (i) genes that had no GO information from any hierarchy in the input data; (ii) genes that had no GO information

[c] http://www.plantphysiol.org/content/suppl/2012/05/15/pp.112.196725.DC1/196725Table_S2.xls

[d] http://www.arabidopsis.org/portals/masc/FG_projects.jsp

with non-electronic evidence tags in the BP hierarchy; and (iii) other experimental BP genes that had non-electronic BP information available, which was not linked to the new experimental association. To evaluate our module-based function predictions, very general categories were not taken into account to avoid an overestimation of the number of true positives (see Materials and Methods). Results showed that out of the 2,940 genes with a new experimental GO-BP, 1,460 genes were assigned to modules with GO-BP enrichment and 29.7% (434) of those had a correct GO-BP inferred through the modules (Table 3.2; Table S3[e]). For the 197 functionally unknown genes from category (i), this percentage was 38.1%. Conversely, from the perspective of the modules, 5,562 genes received a new module-based GO-BP prediction of which 434 genes had their prediction confirmed by a new experimental GO annotation (7.8%; Table 3.3). Based on the fraction of true positives for the functionally unknown genes from category (i), this would suggest that > 2,000 genes (38.1% of 5,562) can be correctly characterized based on the functional coherence of the modules. The results for the different categories are presented in more detail in Tables II and III. All new module-based Arabidopsis functional annotations were submitted to TAIR.

Table 3.2: **Comparison of 2,940 genes having new experimental GO-BP annotations (of which 1,460 are present in modules) with the module-based function predictions.**

Genes	Unknown[a]		Unknown Experimental BP[b]		Other Experimental BP[c]		Total	
	No. Predicted	No. Confirmed[d]	No. Predicted	No. Confirmed	No. Predicted	No. Confirmed	No. Predicted	No. Confirmed
All Genes[e]	197	75 (38.1)	255	108 (42.4)	1,008	251 (24.9)	1,460	434 (29.7)
Conserved	166	65 (39.2)	195	80 (41)	871	215 (24.7)	1,232	360 (29.2)
Not conserved	48	10 (20.8)	83	31 (37.3)	315	52 (16.5)	446	93 (20.9)

[a]No GO information from any hierarchy in the input data. [b]No GO information with nonelectronic evidence tags in the BP hierarchy. [c]Nonelectronic GO information is available in the BP hierarchy, which is not linked to the new experimental association. [d]Numbers in parentheses represent percentages of confirmed genes (number confirmed/number predicted). [e]Genes that were present in both conserved and nonconserved modules could gain a prediction by both. The total of genes in conserved and nonconserved modules is the set of unique genes from these two sets.

Despite the increasing number of genes receiving experimental GO-BP annotations during the last decades, still 7,233 Arabidopsis genes exist for which no GO-BP information is available (neither experimental nor electronic information in any GO hierarchy; Table S1[a]). From these functionally unknown genes, 3,553 genes were assigned to a module of which 68% (2,419/3,553) were part of a module that showed expression conservation (Table 3.4). Based on a functional enrichment analysis using GO or embryo lethal genes, a functional annotation could be associated to 1,701 genes. The fraction of modules containing genes of unknown function and having enrichment-based functional predictions was roughly two times higher for conserved modules compared to modules lacking expression conservation (1,435/2,419 and 266/1,134, respectively). The newly annotated genes in the coexpression conserved modules represented a wide range of biological processes, as can be seen in Figure A.5. Based on gene orthology in the significantly coexpression conserved modules, 43,621 genes with unknown experimental GO-BP in other plants could be assigned a function.

Table 3.3: **Comparison of 5,562 module-based function predictions with new experimental GO-BP annotations.**

Genes	Unknown[a]		Unknown Experimental BP[b]		Other Experimental BP[c]		Total	
	No. Predicted	No. Confirmed[d]	No. Predicted	No. Confirmed	No. Predicted	No. Confirmed	No. Predicted	No. Confirmed
All genes[e]	2,241	75 (3.3)	2,386	108 (4.5)	935	251 (26.8)	5,562	434 (7.8)
Conserved	1,826	65 (3.6)	1,926	80 (4.2)	818	215 (26.3)	4,570	360 (7.9)
Not conserved	565	10 (1.8)	645	31 (4.8)	260	52 (20)	1,470	93 (6.3)

[a]No GO information from any hierarchy in the input data. [b]No GO information with nonelectronic evidence tags in the BP hierarchy. [c]Nonelectronic GO information is available in the BP hierarchy, which is not linked to the new experimental association. [d]Numbers in parentheses represent percentages of confirmed genes (number confirmed/number predicted). [e]Genes that were present in both conserved and nonconserved modules could gain a prediction by both. The total of genes in conserved and nonconserved modules is the set of unique genes from these two sets.

The following paragraphs report a number of examples of module-based gene function predictions that correspond with recent experimental work, which can be explored using the additional data web-

[e] `http://www.plantphysiol.org/content/suppl/2012/05/15/pp.112.196725.DC1/196725Table_S3.xls`

site[f]. The first module, MQSE_BP_GO:0006030_3 (Figure 3.6A), is derived from the GO term ’chitin metabolic process’, and also includes some PPI, TF targets, and AraNet edges. The module contains five true positive genes: MYB63[182], IRX15 and IRX15-L[183], ANAC073[184], and RWA1[185], all of which were correctly predicted to be involved in cell wall biogenesis. MYB63 and ANAC073 are a MYB and a NAC TF, respectively, and whereas MYB63 was known to be involved in JA/SA response pathways[186], no BP information was known for ANAC073. In contrast, RWA1, IRX15, and IRX15-L were completely unknown (no GO in any hierarchy). Additionally, eight currently functionally unknown genes (AT2G41610, AT2G31930, AT1G09610, IQD10, AT1G72220, AT1G33800, IQD13, and AT4G27435) are present in the module. Furthermore, the genes reported in the module correspond with those found by Persson and co-workers in their study of cell wall biogenesis.[145] Out of the four genes that were tested by mutant analysis in their investigation, IRX8 was present as seed gene in the input data, but CTL2 and AT4G27435 were added by the MQSE methodology (AT5G03170 was not present in the module). In addition, looking at the 25 highest ranked genes with CESA4, CESA7, and CESA8 (including the four tested genes), we observed four genes that were seed genes, and ten genes that were added to our module by MQSE.

Table 3.4: **Module-based annotation of genes for which the GO-BP is currently unknown using experimental GO and embryo lethality data.**

Genes	No. of Genes of Unknown Function[a]	Module-Based Annotation		
		No. of Genes with GO Enrichment	No. of Genes Predicted with Embryo Lethality	Total No. of Genes with Functional Prediction (Unique[b])
Modules	3,553	1,680	281	1,701
Conserved	2,419	1,418	275	1,435
Not conserved	1,134	262	6	266
Not in modules	3,680			

[a]No GO-BP information (of any evidence type) is available in the current gene-GO association file. [b]Genes can be predicted by both GO and embryo lethality.

The second module originated from the GO category ’meristem initiation’ (MQSE_BP_GO: 0010014_1; Figure 3.6B) . The true positive gene in this module is PXY, which had only a computational BP annotation related to protein amino acid phosphorylation. Based on the module, the gene was predicted to be involved in xylem and phloem pattern formation, which has recently been annotated by an experimental GO annotation.[187] The module contains multiple genes known to be involved in xylem and phloem pattern formation, including AtPIN1, IFL1, and ATHB15. All genes in the module have experimental associations with meristem-related processes, which refers to the formation of phloem and xylem out of cambium cells (meristematic tissue).

The third module PPI_14 (Figure 3.6C) is based on the experimental PPI network, but many edges are supported by AraNet as well. This PPI module contains 14 genes and is predicted to be involved in DNA endoreduplication, the process of continued DNA replication without mitosis in order to support cell growth. Genes AT1G32310, AT1G06590, and OSD1 were unknown, but AT1G06590 has recently been experimentally validated.[188] Experiments have shown that a hemizygous mutant line of this gene has an endoreduplication index (the mean number of endoreduplication cycles) significantly different compared to wild-type plants. Genes in the module with a known link to endoreduplication were APC8, APC6, FZR2, CDC27B, and APC10.

The fourth module (MQSE_BP_GO:0051726_1; Figure 3.6D) was identified based on the GO term ’regulation of cell cycle’ and includes the functionally unknown genes AT5G48310, AT3G56870, AT3G14190, AT1G10780, AT2G32590, AT3G42660, AT3G56870, AT4G14200, AT3G58650, AT5G01910, and AT4G39630. Given the strong coexpression in the entire module (EC = 0.97) and the conservation of the coexpression (in all six species but Medicago), there is strong evidence that these genes are involved in cell cycle regulation as well. A large fraction of the genes are co-regulated by the E2FA-DPA TF complex. An essential role in cell division coincides with the observed embryo lethality of the module genes HTR12, EMB2795, POLA2, SMC2, AESP, and ATSMC3. The prediction for AT5G55820, which was only known to be functionally involved in embryo sac and seed development,

[f] `http://bioinformatics.psb.ugent.be/cig_data/plant_modules/`

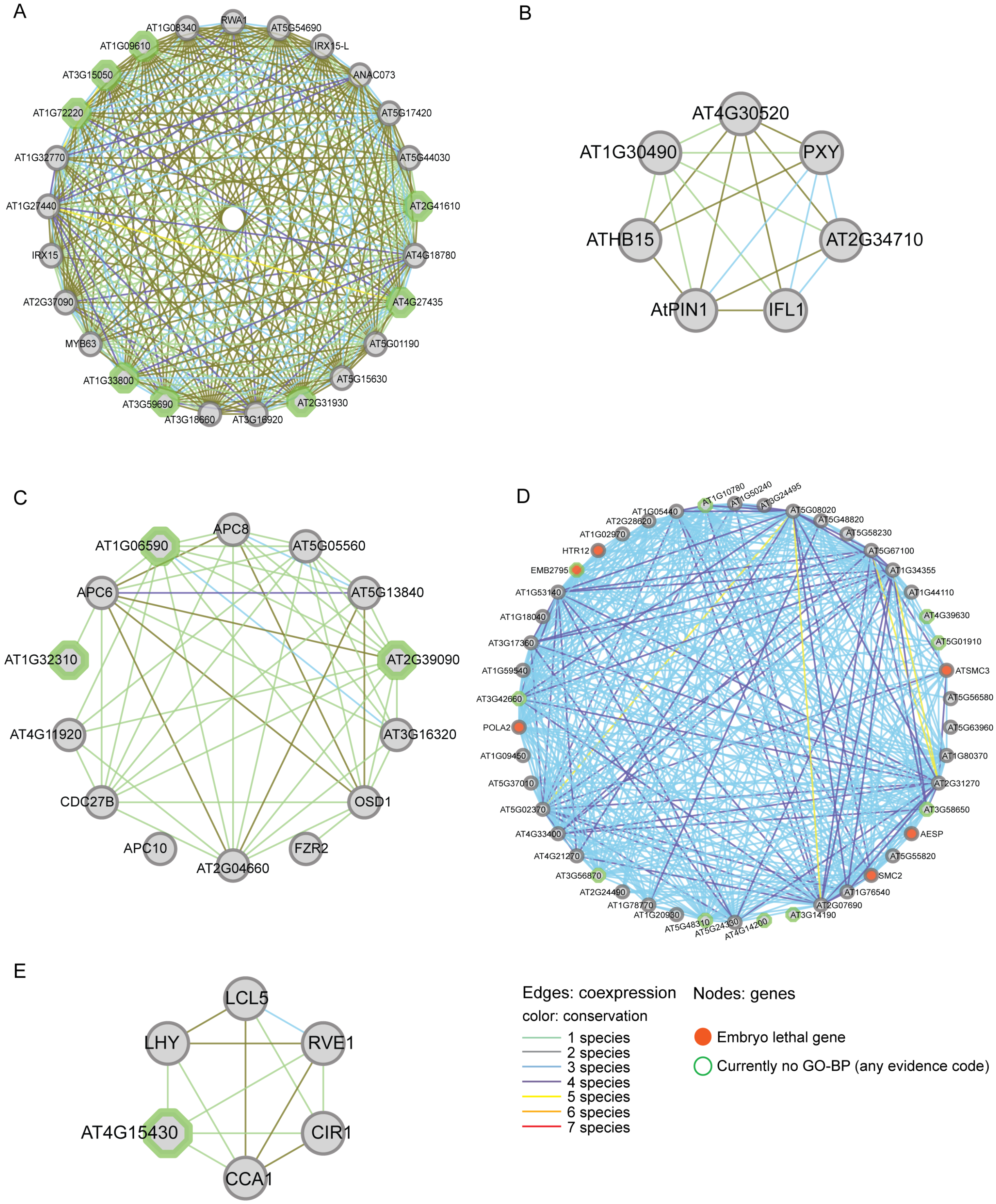


Figure 3.6: **Example of a delineated module with true positive genes.** *A, Cell wall biogenesis. B, Xylem and phloem pattern formation. C, DNA endoreduplication. D, Cell cycle regulation. Edges with EC conservation in less than three species are hidden. E, Response to GA. Modules can be explored in detail using the additional data website.*

is supported by additional InterPro domain evidence as it contains the 'inner centromere protein, ARK-binding domain'. This domain is involved in the coordination of chromosome segregation during cell division in yeast[189], thus linking it to the cell cycle. Furthermore, the *de novo* motif discovery retrieved enriched motifs with an E2F core (TCCCGC).

The last module MQSE_BP_GO:0009739_3 is delineated from the GO category 'response to gibberellin (GA) stimulus' (Figure 3.6E), and has some AraNet and PPI edges as well. The functional prediction of the module yielded the GO terms 'response to GA', as well as 'response to salt stress and hormones (auxin, JA, SA, and abscisic acid)'. However, the module also showed enrichment towards 'circadian rhythm' and 'long-day photoperiodism, flowering' (GO:0007623 and GO:0048574, respectively). These two predictions are particularly interesting as the module contains LCL5, encoding a MYB family TF that was only known to be involved in response to hormone stimuli, but has recently been experimentally assigned to both 'photoperiodism, flowering' and 'circadian rhythm'.[190] Next to the newly assigned MYB LCL5, the module contains two more MYB TF genes (RVE1 and CIR1). Although the MYB TF gene RVE1 had a GO-BP association based on a traceable author statement, the annotation 'regulation of cellular transcription' (GO:0045449) was far from specific. Together with the unknown gene AT4G15430, the module thus provides a strong prediction for two functionally unknown genes. CIR1, LHY, and CCA1 were the known circadian regulators on which the module prediction was based. The module is enriched for the motif sTsAGCCACwAn, which contains the SORLIP1 (Sequences Over-Represented in Light-Induced Promoter 1) core (CCAC) described in PLACE, which is a phytochrome A-induced motif. Finally, given the enrichments for genes responsive to a GA stimulus and circadian clock genes, this module reaffirms the crosstalk between both processes reported by Arana and coworkers.[191]

3.3 Discussion

To delineate a wide range of gene modules, an ensemble of input data types was assembled, based on experimental gene associations (GO, PPI, and TF targets) and AraNet. Although the different combined data sets comprised more than 1 million gene associations, the overlap between individual data sets was surprisingly low. This observation was confirmed by the large fraction of unique associations per primary data type and the low overlap in gene content between the modules before redundancy removal, indicating the advantage of combining different experimental data sources. Based on a set of 2,355 Arabidopsis proteins, Lysenko and co-workers also reported that the integration of multiple data sets, apart from sequence-based gene functions, was beneficial for the functional annotation of modules inferred using graph-based clustering.[192] In addition, their data revealed that, despite the integration of experimental data sources, only a limited number of all Arabidopsis genes could be embedded into an integrative network. Complementary to network construction methods that start from a limited number of experimentally characterized genes, other studies have applied clustering tools on large expression compendia to identify gene modules at a genome-wide scale[139,149];.[140] Although including more genes, these approaches typically yield a limited number of functional modules, as functional gene information is mostly incorporated during post-processing to link modules to specific biological processes.[140,173] To circumvent this problem, we developed the MQSE method to use genes with non-electronic GO annotations as guide genes to define coexpression modules. While guide gene approaches are typically applied for the analysis of a specific process[138,145,146], the integration of all GO categories resulted in a set of modules covering a wide range of processes in Arabidopsis (Figure 3.4). Although Cho et al. also integrated different GO annotations during the delineation of yeast modules[193], as far as we are aware, this approach has not been applied to plants. GO-based clustering without any modification to the gene sets would result in many missing genes due to the incomplete functional annotation of the Arabidopsis genome and the low expression coherences in some categories. To overcome this problem, the guide gene MQSE strategy allowed to fine-tune the GO seed sets prior to expression clustering by identifying strongly coexpressed seeds and by adding more than thousand genes with highly similar expression profiles. Whereas MQSE is related to the Multi-Experiment Matrix (MEM) method of Adler and coworkers[194], MEM uses one gene as seed, while our approach can integrate multiple seed genes. This is a significant improvement since this allows the analysis of coregulation within a process of interest. Secondly, whereas the output of MEM is a ranked list of genes that are coexpressed with the query gene, there is no determination of an optimal set of coexpressed genes. In contrast, MQSE returns the optimal set of coexpressed genes using a rank-based enrichment score.

Based on the EC scores and the percentage of modules for which a regulatory DNA motif could be found (50%), it is clear that coexpression and coregulation are two important factors to ensure the proper functioning of genes. Remarkably, PPI is the second best data type when considering expression and *cis*-regulatory coherence, indicating that interacting genes are also frequently coregulated. Conversely, the *cis*-regulatory coherence of the TF target data was not higher than in other data sets, supporting the concerns about the specificity of ChIP data sets, as many reported TF targets do not correlate with each other at the expression level.[144] However, the EC of TF target data set was influenced most by different expression compendia, suggesting differences in the condition specificity for the different target genes (Figure 3.3A). The analyzed module properties indicate that GO combined with coexpression and PPI data is the most suited to delineate functionally and regulatory coherent modules. The same trend was observed when determining true positive module-based GO predictions per input type, as true positives were found in 214 (37%) GO, 22 (31%) PPI, 47 (11%) AraNet, and 15 (3%) TF modules. In addition, we observed that highly integrative approaches, such as AraNet, yielded many modules lacking functional coherence and that more than thousand conserved gene modules were found, based on one of the other primary data types.

On the organizational level, it is clear that, as for other biological networks, most genes are present in few modules, while a limited number of hub genes exists. On the regulatory level, a similar pattern was observed with most modules and genes containing a limited number of motifs. The maximum number of 26 motifs per genes is high, but in line with a recent estimation of the number of binding sites per gene, being, based on available Arabidopsis Chip-Seq studies, up to 75 binding events per gene.[144] Although it is currently unclear whether this pattern holds for all genes, this estimate provides an experimental indication that complex regulation, as indicated by our modules, will be true for some genes. The variation in regulatory complexity for different GO-BP slim categories confirms that function, apart from other factors, is an important element contributing to a gene's regulation.[168,195]

Genome-wide modular approaches have often been used to infer functions for functionally unknown genes. However, to our knowledge, this study is the first one to integrate different functional data types as well as conserved coexpression in seven species (soybean, Medicago, poplar, grapevine, rice, and maize) to characterize new plant gene functions. Whereas integrative approaches have been performed combining heterogeneous data in Arabidopsis[136,166], Mutwil and co-workers included cross-species expression information to study gene functions in seven plant species.[169] An important advantage of the module-based approach with respect to function prediction is that homologs are not required for a gene to receive a prediction. In agreement with a recent comparative transcriptomics study reporting conserved modules between maize and rice[167], we observed that modules showing ultra-conserved coexpression, primarily cover genes that are related to energy and housekeeping functions, such as photosynthesis, ribosome biogenesis, and translation. However, the 910 modules showing significant coexpression in other angiosperms, cover a broad range of biological processes and provide a valuable resource to identify new gene functions and translate biological information from model species to crops. Based on our module-based functional predictions, 5,562 Arabidopsis genes received a functional annotation and an evaluation experiment showed that, based on a set of previously functionally unknown genes that were recently experimentally characterized, 38.1% of these gene functions could be inferred through the modules. Clearly, the annotation of genes of unknown function seems to benefit from the integration of coexpression conservation, as modules showing conserved coexpression, recover almost two times more experimental GO-BP annotations compared to non-conserved modules. However, true positive annotations could be found in non-conserved modules as well, thus not only providing support for these annotations, but also suggesting that high-quality experimental data sets are important to study species-specific or adaptive gene functions. Overall, as a result of the integration of sequence and expression data for six plant species, the module-based predictions offer new biological hypotheses for currently functionally unknown genes in Arabidopsis (1,701 genes) and six other plant species (43,621 genes).

3.4 Material and Methods

Data sets

Twelve expression data sets (abiotic stress conditions, biotic stress conditions, developmental stages, flowering tissue, genetic modification, hormone treatment, leaf tissue, root tissue, seed tissue, all stress conditions, whole plants, AtGenExpress, and a general compendium) for Arabidopsis were retrieved from the CORNET database in November 2010.[107] The expression data for soybean (15,753 genes), Medicago (17,614 genes), poplar (28,969), grapevine (8,255 genes), rice (34,153 genes), and maize (10,068 genes) were assembled from NCBI Gene Expression Omnibus.[111] CEL files were analyzed using a custom-made CDF (at least five probes per probe set) and normalized using the RMA method.[196] A list of experiments for the different species is reported in Table S4[g]. Redundant experiments were removed by clustering experiments over genes, and experiments with Pearson correlation coefficient (PCC) âĽ'ě 0.99 were considered redundant. The number of retained experiments was 1,153, 43, 108, 39, 258, and 85 for soybean, Medicago, poplar, grapevine, rice, and maize, respectively. AraNet gene associations were retrieved in November 2010.[106] GO associations[105] for Arabidopsis genes were retrieved from the PLAZA2.0 database in November 2010.[197] Genes assigned to a GO term were recursively assigned to all of the GO terms' parental terms. Only gene-GO associations with non-electronic evidence codes were taken into account for module delineation: EXP, IDA, IPI, IMP, IGI, IEP, IC, and TAS. The PPI data were downloaded from the CORNET database in November 2010[107] and only experimentally identified PPIs were retained. Interaction data of TFs and their targets were retrieved from the AtRegNet database in November 2010.[17] The targets of each TF were divided based on the effect on their expression: activation, repression, and all (this group also contains the genes with unknown effect). Orthologous genes were identified using the integrative orthology method available from PLAZA 2.0 only retaining orthologs with the highest number of evidences.[180] The embryo lethal genes were obtained from the SeedGenes database by selecting for confirmed embryo defective genes.[109]

Module delineation using expression- and connectivity-based clustering

Both connectivity-based clustering and expression-based clustering were performed with a Perl implementation of the graph-based Cluster Affinity Search Technique (CAST) algorithm.[115] Connectivity-based clustering was directly applied to the PPI and AraNet input gene associations, and was optimized by selecting the threshold that maximized the largest number of genes assigned to modules, and the number of modules with GO functional enrichment (PPI: 0.5 and AraNet: 0.33).

Expression-based clustering was performed using a relative PCC threshold (95th percentile) based on a set of 10,000 random gene pairs, specific to each expression compendium. Clustering was optimized for each set of genes (either a set of TF target genes or a set of genes with a common GO annotation) by prior selection of the CORNET expression compendium with the best EC for the given set of genes. The minimum and maximum clustering size was set at 5 and 100, respectively.

The GO seed genes were submitted to the MQSE approach prior to clustering. The MQSE approach adds new genes that show significant coexpression, while also removing seed genes that do not coexpress coherently with the other seed genes. The decision of which genes to add and which genes to remove is based on a rank statistic that incorporates the number of coexpressed seed genes, the standard deviation of the expression profile of the coexpressed seed genes, and the median rank towards all seed genes (see Protocol S1). The final expanded gene set is defined by selecting the top set of ranked genes yielding the highest significant enrichment towards seed genes. Subsequently, these expanded gene sets are clustered using CAST after which only clusters with enrichment towards the initial seed genes are retained, to ensure retention of the initial functional category (hypergeometric distribution, p-value âĽ'd' 0.05).

To identify and remove redundant modules within and across the different data types, the gene overlap between all modules was assessed using the Jaccard score. In cases where one module was completely embedded in the other, the overlap score was set at 1. Based on all pairwise overlap scores, modules were clustered by CAST using a score cut-off of 0.85. As a result, overlapping modules were grouped in a cluster of similar modules and the most highly connected module in each cluster was assigned as being the representative (i.e. the module with the highest average overlap in the cluster of similar modules).

In order to make the module information publicly available, an additional data website was developed[h]. From the start page, all genes, modules, and GO categories from the module data set can be queried. Results include the modules and their genes, regulatory DNA motifs, comparative coexpression

[g] http://www.plantphysiol.org/content/suppl/2012/05/15/pp.112.196725.DC1/196725Table_S4.xls

[h] http://bioinformatics.psb.ugent.be/cig_data/plant_modules/

results, and visualizations of the modules based on either the comparative coexpression links or the input data gene associations. Bulk downloads are also available.

Expression Coherence

The EC for a set of N genes was calculated as the fraction of all possible N*(N-1) / 2 gene pairs with a PCC higher than or equal to the threshold value defined for that compendium.[198] The p-value for an EC threshold of 10% in Arabidopsis modules was estimated at $\leq$ 0.02 based on 960,000 random modules with a size distribution identical to the real data set.

Gene functional annotation

GO enrichment analysis was based on the same GO dataset as for the module delineation (described under 'Data sets'). Enrichment of a GO category in a module was calculated as the ratio of the module frequency over the genome-wide frequency. The enrichment values were validated statistically using the hypergeometric distribution and adjusted using FDR correction for multiple hypotheses testing.[199] The significance level was set at 0.01 and at least two genes in the cluster had to be associated with the GO label before a GO was assigned to a module. Due to this stringent threshold, some GO modules determined by MQSE lack enrichment in the final set of non-redundant modules. Enrichment towards embryo lethal genes was performed similarly.

Motif finding

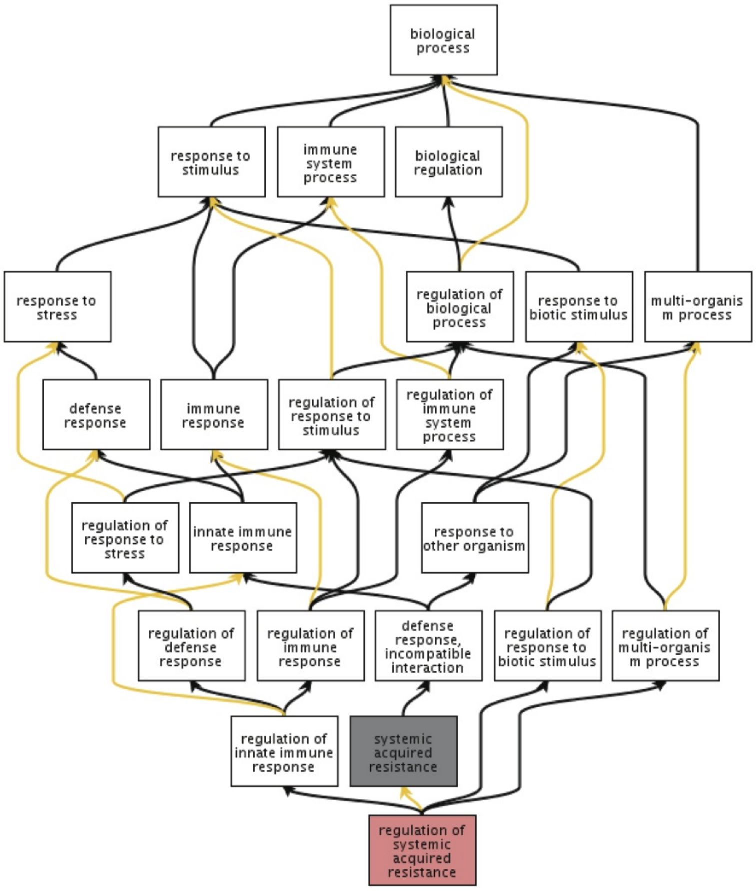
De novo motifs were identified using MotifSampler (default settings) followed by MotifRanking (default settings)[175] and Weeder (default settings)[174] for word sizes ranging from 6 to 12, on the 1-kb promoter (sequence upstream of start codon, based on TAIR9) taking both strands into account. MotifSampler was run with a third order background model based on all Arabidopsis promoters from PLAZA2.0. Weeder motifs were transformed to position weight matrices (PWMs) based on their reported frequency matrix. Motif enrichment was determined for each motif based on genome-wide promoter mapping of their PWMs using MotifLocator (default settings).[175] Enrichment was defined as the ratio of the module frequency over the genome-wide frequency and enrichment values were statistically evaluated using the hypergeometric distribution, adjusted by the FDR correction for multiple hypothesis testing.[199] Only significantly enriched motifs with a corrected p-value $\leq$ 0.01 were retained. To determine motif representatives (and remove redundancy) within each module, motifs were clustered based on sequence similarity and gene-motif occurrences. To compare sequence similarity, motif PWMs were transformed into vectors and for each pair of motifs, the PCC between the vectors was determined using a sliding window while retaining a minimum overlap of six nucleotides. Subsequently, the motifs were clustered using a PCC threshold of 0.75. The results of the sequence-based clustering were submitted to the occurrence-based clustering, based on the method described by Xie and co-workers.[77] Based on these results, a set of non-redundant motifs was defined for each module and motifs with a similar sequence, but residing in a distinct set of genes, were considered as distinct motifs. Known motifs were extracted from AGRIS[17] and PLACE[176], and the redundancy was removed similarly as for the modules.

Motif conservation was determined by mapping the PWMs on the 1-kb promoters (both strands) of the different species with MotifLocator. For each species, backgrounds of the third order were built based on all 1-kb promoters (PLAZA2.0). For each module, the enrichment was determined in each species based on the occurrences in the orthologous module and the genome-wide occurrences. P-values for enrichment were calculated based on the hypergeometric distribution and corrected by FDR.

Motif annotation was performed by integrating the module functional annotation and the coexpression conservation. For each motif family, the motif instances across different modules were used to translate the functional annotation of the module to the motif family. Furthermore, each motif family annotation obtained in this manner was weighted by the expression conservation of the module. When multiple modules supported the association between GO and motif family, the expression conservation was averaged over the different modules. The motif - GO network was created using Cytoscape[200] and reduced by retaining the most specific GO nodes (and discarding related but less significant nodes).

Functional prediction of genes of unknown function

To validate module-based GO predictions, an updated Gene Ontology gene association file was downloaded from TAIR on 20/01/2012. All associations with non-electronic evidence tags that were created after the input data freeze, were compared with the module-based predictions. Note that some new experimental gene associations were derived from publications prior to the data freeze. A prediction was called as true positive if and only if the most specific common parent between the prediction and the new experimental association was more specific than any existing experimental GO-BP term. If the most specific common parent was a general term (GO:0008150, GO:0051704, GO:0009987, GO:0008152, GO:0044237, GO:0044238, GO:0050794, GO:0044260, GO:0043170, GO:0044249, GO:0050789, GO:0034645, GO:0010468, GO:0031326, GO:0010556, GO:0051171, GO:0009889, GO:0080090, GO:0019222, GO:0060255, GO:0065007, GO:0031323, GO:0009058, GO:0006139, GO:0009059, GO:0034641, GO:0044267), it was not considered a true positive. The different categories for a true positive prediction listed in Table 3.2 are visualized in Figures 3.7-3.8. The categories 'unknown' and 'unknown experimental BP' were the same from the perspective of the true positive determination, as in both cases there were no existing GO-BP categories in the input data (only non-electronic evidence GO-BPs were selected for input data). These scenarios are depicted by Figure 3.7. The third category, 'other experimental BP', describes genes that had GO-BP annotations with experimental evidence codes, but of which the true positive prediction was not linked to the existing annotations (Figure 3.8). As such, the predictions were not a consequence of the existing non-electronic annotations.


biological process
response to stimulus
immune system process
biological regulation
response to stress
regulation of biological process
response to biotic stimulus
multi-organis m process
defense response
immune response
regulation of response to stimulus
regulation of immune system process
regulation of response to stress
innate immune response
response to other organism
regulation of defense response
regulation of immune response
defense response, incompatible interaction
regulation of response to biotic stimulus
regulation of multi-organis m process
regulation of innate immune response
systemic acquired resistance
regulation of systemic acquired resistance

QuickGO - http://www.ebi.ac.uk/QuickGO

Figure 3.7: **True Positive Gene Annotation Prediction for AT1G73805.** *In the input data set, the gene AT1G73805 had no annotations with a GO-BP with any type of evidence. The module-based GO prediction 'systemic acquired resistance' (GO:0009627; grey) has been experimentally confirmed through the new GO-BP association 'regulation of systemic acquired resistance' (GO:0010112; red). Black lines represent 'is a' relationships and yellow lines indicate regulatory relationships.*

Genes that did not have GO-BP associations with non-electronic evidence types in the updated GO association file were selected as currently unknown. The functional prediction was based on the en-

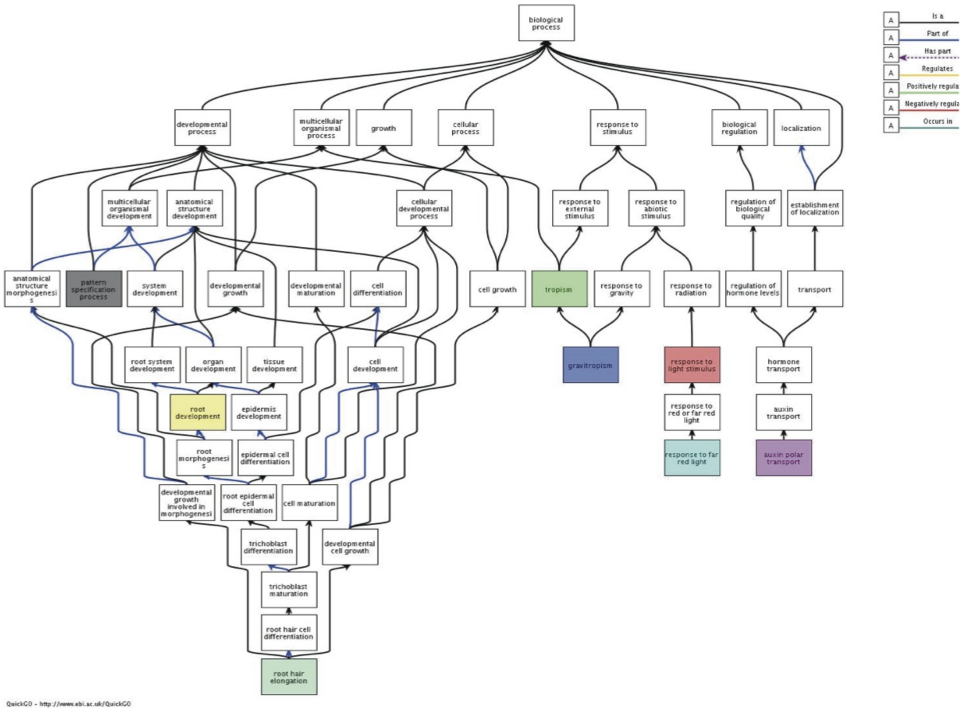

biological process
Is a
Part of
Has part
Regulates
Positively regula
Negatively regula
Occurs in
developmental process
multicellular organismal process
growth
cellular process
response to stimulus
biological regulation
localization
multicellular organismal development
anatomical structure development
cellular developmental process
response to external stimulus
response to abiotic stimulus
regulation of biological quality
establishment of localization
anatomical structure morphogenesis
pattern specification process
system development
developmental growth
developmental maturation
cell differentiation
cell growth
tropism
response to gravity
response to radiation
regulation of hormone levels
transport
root system development
organ development
tissue development
cell development
gravitropism
response to light stimulus
hormone transport
root development
epidermis development
response to red or far red light
auxin transport
root morphogenesis
epidermal cell differentiation
response to far red light
auxin polar transport
developmental growth involved in morphogenesi
root epidermal cell differentiation
cell maturation
trichoblast differentiation
developmental cell growth
trichoblast maturation
root hair cell differentiation
root hair elongation
QuickGO - http://www.ebi.ac.uk/QuickGO

Figure 3.8: **True positive gene annotation prediction for AT1G70940.** *In the input data set, the gene AT1G70940 had several GO-BP annotations a non-electronic evidence code: pattern specification process (GO:0007389), root development (GO:0048364), root hair elongation (GO:0048767), tropism and gravitropism (GO:0009606 and GO:0009630) and auxin polar transport (GO:0009926). However, the module-based prediction provided a new annotation 'response to far red light' (GO:0010218; cyan). Since the data freeze, this gene has been experimentally associated with the response to light stimuli (GO:0009416). As the most specific overlap between these two terms is response to light stimulus, this association is a true positive. Although the gene had existing GO-BP with experimental support, these annotations had influence on the predicted annotation. Black lines represent 'is a' relationships and blue lines indicate 'part of' relationships.*

richments for GO-BP categories and embryo lethal genes. Orthologous genes without non-electronic GO-BP associations were assigned the functional prediction of the Arabidopsis module if and only if these modules had a significant EC conservation, as well as a significant EC in the respective species.

3.5 Acknowledgements

We thank Sara Movahedi for microarray normalization, Michiel Van Bel for technical assistance with the additional Web site, Stefanie de Bodt for critical reading of the manuscript and providing the CORNET data, Annick Bleys for help in preparing the manuscript, and Yves Van de Peer for general support.

CHAPTER **4**

A functional and evolutionary perspective on transcription factor binding in *Arabidopsis thaliana*[a]

Abstract

Understanding the mechanisms underlying gene regulation is paramount to comprehend the translation from genotype to phenotype. The two are connected by gene expression, and it is generally thought that variation in transcription factor (TF) function is an important determinant of phenotypic evolution. We analysed publicly available genome-wide ChIP experiments for 27 transcription factors (TFs) in Arabidopsis thaliana and constructed an experimental network containing 46,619 regulatory interactions and 15,188 target genes. We identified hub targets and Highly Occupied Target (HOT) regions, which are enriched for genes involved in development, stimulus responses, signaling and gene regulatory processes in the currently profiled network. We provide several lines of evidence that TF binding at plant HOT regions is functional, in contrast to that in animals, and not merely the result of accessible chromatin. HOT regions harbor specific DNA motifs, are enriched for differentially expressed genes, and are often conserved across crucifers and dicots, even though they are not under higher levels of purifying selection than non-HOT regions. Distal bound regions are under purifying selection as well, and are enriched for a chromatin state showing regulation by the Polycomb repressive complex. Gene expression complexity is positively correlated with the total number of bound TFs, revealing insights in the regulatory code for genes with different expression breadths. The integration of non-canonical and canonical DNA motif information yields new hypotheses on co-binding and tethering between specific TFs involved in flowering and light regulation.

[a] This chapter is based on Heyndrickx et al. [201]. K.S.H., K.V., C.W., and D.W. designed the research methodology. K.S.H. (ChIP pipeline, binding properties, population genomics, expression and motif analysis) J.V.d.V. (DH I enrichment analysis, chromatin state analysis, CNS pipeline and analysisFigures 1B, 2E), and C.W. (1001 genomes assembly and SNP calling) performed data analyses. K.S.H., J.V.d.V, D.W., and K.V. wrote the manuscript.

4.1 Introduction

Unravelling the mechanisms underlying gene regulation is an important premise to understand how the genotype is translated into a functional organism. Transcriptional regulation by transcription factors (TFs) is one of the most investigated mechanisms, as it can be considered the primary level of regulation.[202] The emergence of Chromatin Immunoprecipitation (ChIP) followed by genome-wide readout through microarray (ChIP-chip) or deep sequencing (ChIP-Seq) has stimulated the experimental identification and comprehensive characterization of target genes bound by a specific TF.[49,50] Studying a single TF using ChIP (henceforth referring to both ChIP-chip and ChIP-Seq) is already valuable to examine its DNA binding motif, identify putative target genes and unravel its biological role through the functional analysis of its targets. Going further, the integration of complementary functional genomics data sets has the potential to provide insights regarding the bound DNA and the mechanisms underlying co-regulation by multiple TFs.

While these genome-wide approaches can open many interesting avenues for subsequent studies, the biological interpretation of ChIP studies involves a number of important challenges. Firstly, ChIP data have revealed only weak correlation between TF binding and transcriptional regulation of the potential target genes.[154] Possible explanations are the dependency on other condition-specific factors, such as cofactors or chromatin remodelling, for the correct regulation of the target gene, or that many of the observed binding events are non-functional. In the latter case, such binding events are suggested to be the result of passive thermodynamics instead of active recruitment[203] and non-functional binding events have been linked with highly bound genes (hub target genes) and Highly Occupied Target (HOT) regions (bound by many TFs) in the worm Caenorhabditis elegans and in the yeast Saccharomyces cerevisiae.[204,205] Secondly, some TF-bound regions show enrichment for multiple different DNA sequence motifs, complicating the identification of directly regulated targets. In regions of the genome of Arabidopsis thaliana (hereafter Arabidopsis) bound by SEPALLATA3 (SEP3), a TF involved in flower development, enrichment was found for five known TF sequence motifs.[158] Multiple enriched DNA binding motifs in a ChIP data set can be a sign of cooperative binding by multiple TFs, or of tethering, where the profiled TF associates with the chromatin through a protein-protein interaction with a second TF. Some of the first integrative regulatory studies were in the context of the ModENCODE and ENCODE projects in C. elegans[205–207], Drosophila melanogaster[208,209], and Homo sapiens.[210–212] Information on protein-protein interactions, miRNA-target interactions and gene expression profiles has been harnessed for the identification of master regulators and network motifs[207,211], and for inferring gene regulatory networks and predictive models of gene expression levels of target genes.[205,213] Ferrier et al.[144] and Mejia-Guerra et al.[214] have already generated an overview of the available TF profiling studies in Arabidopsis. They also listed several challenges related to unravelling TF binding complexity in plants; however, an integrated experimental gene regulatory network describing cooperative TF binding events in plants is currently missing.[144,214]

Here, an integrative study of 27 genome-wide TF profiling experiments containing 15,188 potential target genes in Arabidopsis is presented, in combination with complementary TF perturbation information, chromatin states, population genomic data and various functional data sets. We study the organisation and mechanisms underlying TF regulation and uncover the following insights in transcriptional regulation in plants:

- Grouping potential target genes into modules of functionally related genes offers, complementary to filtering potential target genes using DNA motifs, a valuable approach to identify TF-regulated genes, and provides a computational alternative to differentially expressed genes obtained through TF perturbation experiments.
- TF binding is organised in distinct islands across the genome that correlate well with DNase I hypersensitive (DH) sites. TF bound regions have different levels of complexity, ranging from being bound by a single TF to up to more than half of the profiled TFs.
- Hub potential target genes are enriched for functions related to signalling and regulation, responses to stimuli and development, and are examples of complexly bound genes. Furthermore, through

the integration of miRNA and kinase networks, we confirmed that TFs themselves are complexly targeted through several mechanisms.

- Broad expression and high gene expression levels are correlated with complex regulation by many TFs, offering insights into how transcriptional control for genes expressed under numerous conditions is encoded in the genome.

- Cross-species sequence conservation, population sequence diversity, and chromatin states of the bound regions together with functional analysis of the potential target genes indicate that HOT regions are functional and do not reflect spurious binding events due to open chromatin. This pattern is different from results in animals, where it has been reported that HOT-associated genes are less likely to be regulated than other genes.

- Overlap with chromatin states links a subset of distal upstream bound regions to binding events under regulation by the Polycomb complex, an important repressor complex in plant development.

- For several TFs, a large number of DNA binding events are associated with non-canonical motifs, generating new testable hypotheses of co-binding TFs and TFs associating with chromatin through tethering.

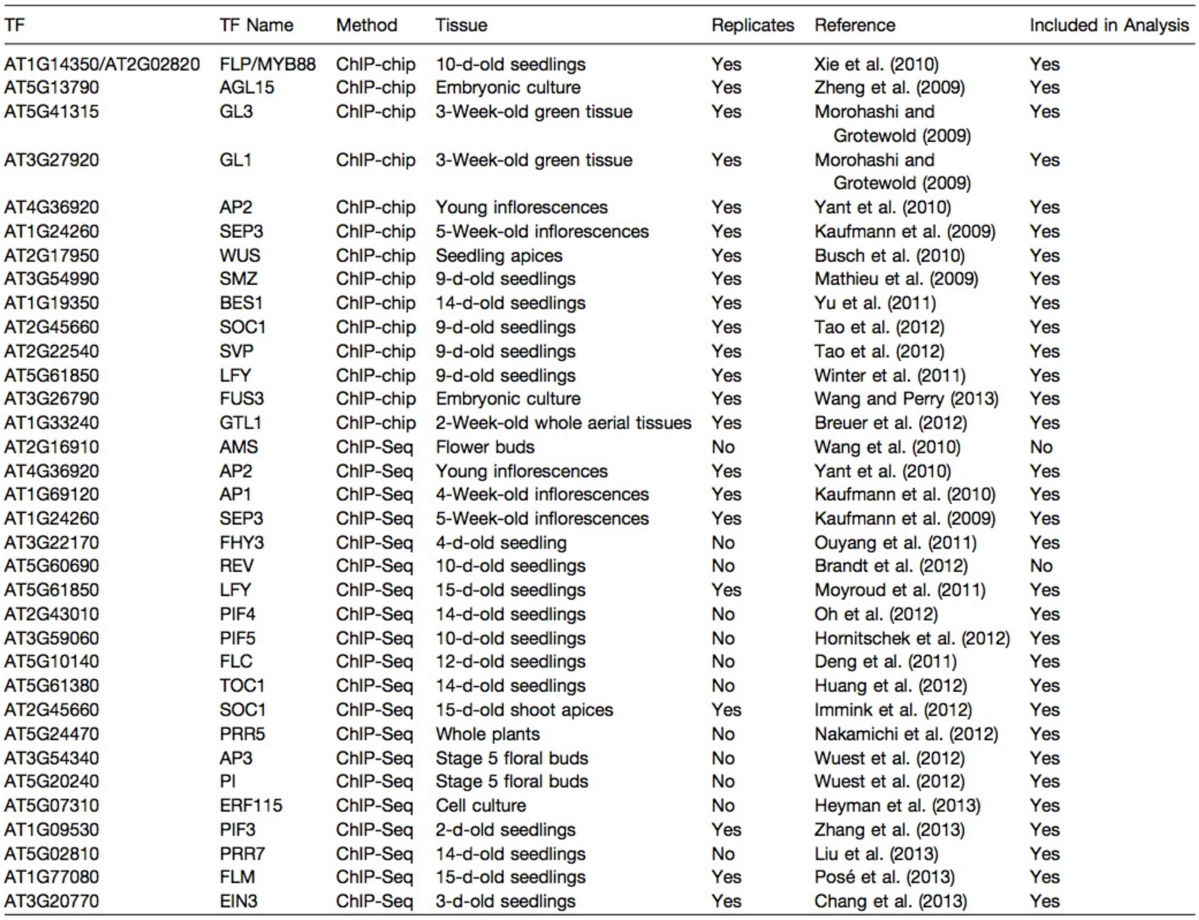
4.2 Results

Construction of an experimental Arabidopsis gene regulatory network through the integration of TF ChIP experiments

At the start of our study, 34 ChIP experiments had been performed in Arabidopsis using the Affymetrix Tiling Array or short read sequencing, profiling 30 different TFs (Table 4.1). These factors are primarily involved in flowering (AGAMOUS-LIKE 15 [AGL15], APETALA1 [AP1], APETALA2 [AP2], APETALA3 [AP3], SEPALLATA3 [SEP3], SCHLAFMUTZE [SMZ], SUPPRESSOR OF OVEREXPRESSION OF CO 1 [SOC1], SHORT VEGETATIVE PHASE [SVP], PISTILLATA [PI], LEAFY [LFY], FLOWERING LOCUS C [FLC], WUSCHEL [WUS], FOUR LIPS/MYB88 [FLP/MYB88], and FLOWERING LOCUS M [FLM]), circadian rhythm and light response (PSEUDO RESPONSE REGULATOR 5 [PRR5], PSUEDO RESPONSE REGULATOR 7 [PRR7], SOC1, TIMING OF CAB EXPRESSION 1 [TOC1], PHYTOCHROME INTERACTING FACTOR 3 [PIF3], PHYTOCHROME INTERACTING FACTOR 4 [PIF4], PHYTOCHROME INTERACTING FACTOR 5 [ÃŔF5], REVOLUTA [REV], and FAR-RED ELONGATED HYPOCOTYLS 3 [FHY3]), cell cycle (WUS), hormone signalling (BRI1-EMS-SUPPRESSOR 1 [BES1] and ETHYLENE-INSENSITIVE 3 [EIN3]), and other aspects of development (GLABRA 1 [GL1], GLABRA 3 [GL3], GT2-LIKE 1 [GTL1], FUSCA 3 [FUS3], ABORTED MICROSPORES [AMS] and ETHYLENE RESPONSE FACTOR 115 [ERF115]). To create comparable data sets, we developed an analysis pipeline consisting of quality control, platform-specific signal processing, and peak calling to re-process all raw data in a standardised and uniform manner (see Methods, Figure 4.1). Thus, the integrated network comprised 27 unique TFs binding near 15,188 potential target genes, covering 46,619 unique TF-target interactions (Figure 4.2). For the remainder of this manuscript, we use the terms potential target genes or bound genes for genes that were associated with a TF binding event. Genes that are bound and display differential expression (DE) upon perturbation of the TF will be referred to as TF-regulated genes. The TFs for which DE data was available are listed in B.1.

Genome-wide ChIP experiments can lead to the identification of many potential target genes, some of which have no known functional association with the TF. The integration of DE data, which results in a set of high confidence, directly regulated target genes is often used to filter ChIP data. However, TF binding can also be part of a strategy to poise the promoter for fast response to subsequent other signals that lead to a transcriptional response of the target gene. In the latter case there would be no DE response of the potential target gene in the perturbation experiment.[13] Therefore, as an alternative to using TF perturbation in a single condition, we sought potential target genes that show functional coherence, a sign of bona fide regulated genes. False positive potential target genes will not show functional coherence with

Table 4.1: **Arabidopsis TF ChIP Data Sets Used.**



TF	TF Name	Method	Tissue	Replicates	Reference	Included in Analysis
AT1G14350/AT2G02820	FLP/MYB88	ChIP-chip	10-d-old seedlings	Yes	Xie et al. (2010)	Yes
AT5G13790	AGL15	ChIP-chip	Embryonic culture	Yes	Zheng et al. (2009)	Yes
AT5G41315	GL3	ChIP-chip	3-Week-old green tissue	Yes	Morohashi and Grotewold (2009)	Yes
AT3G27920	GL1	ChIP-chip	3-Week-old green tissue	Yes	Morohashi and Grotewold (2009)	Yes
AT4G36920	AP2	ChIP-chip	Young inflorescences	Yes	Yant et al. (2010)	Yes
AT1G24260	SEP3	ChIP-chip	5-Week-old inflorescences	Yes	Kaufmann et al. (2009)	Yes
AT2G17950	WUS	ChIP-chip	Seedling apices	Yes	Busch et al. (2010)	Yes
AT3G54990	SMZ	ChIP-chip	9-d-old seedlings	Yes	Mathieu et al. (2009)	Yes
AT1G19350	BES1	ChIP-chip	14-d-old seedlings	Yes	Yu et al. (2011)	Yes
AT2G45660	SOC1	ChIP-chip	9-d-old seedlings	Yes	Tao et al. (2012)	Yes
AT2G22540	SVP	ChIP-chip	9-d-old seedlings	Yes	Tao et al. (2012)	Yes
AT5G61850	LFY	ChIP-chip	9-d-old seedlings	Yes	Winter et al. (2011)	Yes
AT3G26790	FUS3	ChIP-chip	Embryonic culture	Yes	Wang and Perry (2013)	Yes
AT1G33240	GTL1	ChIP-chip	2-Week-old whole aerial tissues	Yes	Breuer et al. (2012)	Yes
AT2G16910	AMS	ChIP-Seq	Flower buds	No	Wang et al. (2010)	No
AT4G36920	AP2	ChIP-Seq	Young inflorescences	Yes	Yant et al. (2010)	Yes
AT1G69120	AP1	ChIP-Seq	4-Week-old inflorescences	Yes	Kaufmann et al. (2010)	Yes
AT1G24260	SEP3	ChIP-Seq	5-Week-old inflorescences	Yes	Kaufmann et al. (2009)	Yes
AT3G22170	FHY3	ChIP-Seq	4-d-old seedling	No	Ouyang et al. (2011)	Yes
AT5G60690	REV	ChIP-Seq	10-d-old seedlings	No	Brandt et al. (2012)	No
AT5G61850	LFY	ChIP-Seq	15-d-old seedlings	Yes	Moyroud et al. (2011)	Yes
AT2G43010	PIF4	ChIP-Seq	14-d-old seedlings	No	Oh et al. (2012)	Yes
AT3G59060	PIF5	ChIP-Seq	10-d-old seedlings	No	Hornitschek et al. (2012)	Yes
AT5G10140	FLC	ChIP-Seq	12-d-old seedlings	No	Deng et al. (2011)	Yes
AT5G61380	TOC1	ChIP-Seq	14-d-old seedlings	No	Huang et al. (2012)	Yes
AT2G45660	SOC1	ChIP-Seq	15-d-old shoot apices	Yes	Immink et al. (2012)	Yes
AT5G24470	PRR5	ChIP-Seq	Whole plants	No	Nakamichi et al. (2012)	Yes
AT3G54340	AP3	ChIP-Seq	Stage 5 floral buds	No	Wuest et al. (2012)	Yes
AT5G20240	PI	ChIP-Seq	Stage 5 floral buds	No	Wuest et al. (2012)	Yes
AT5G07310	ERF115	ChIP-Seq	Cell culture	No	Heyman et al. (2013)	Yes
AT1G09530	PIF3	ChIP-Seq	2-d-old seedlings	Yes	Zhang et al. (2013)	Yes
AT5G02810	PRR7	ChIP-Seq	14-d-old seedlings	No	Liu et al. (2013)	Yes
AT1G77080	FLM	ChIP-Seq	15-d-old seedlings	Yes	Posé et al. (2013)	Yes
AT3G20770	EIN3	ChIP-Seq	3-d-old seedlings	Yes	Chang et al. (2013)	Yes

other potential target genes, in contrast to genuine regulated genes.[66] To delineate functionally coherent subsets of bound genes per TF, the enrichment of potential targets was determined in 1,563 functional gene modules.[134] The latter comprise 13,142 genes annotated with specific functional descriptions based on co-expression, experimental Gene Ontology (GO) information, experimental protein-protein interaction data, protein-DNA interactions described in AtRegNet[17] or AraNet gene function predictions.[106]

The benefits of this strategy are illustrated by the finding that potential target genes are greatly enriched for DE genes in 10 out of 15 ChIP experiments for which DE data is available and for which $> 20\%$ of the potential target genes are in modules (Figure B.1). For an additional 4 experiments (LFY, FHY3, PI, and AP1) the effect was marginal. The applicability of this approach is by definition dependent on the presence of the potential target genes in the functional gene modules. For GLT1, GL1, BES1, GL3, PIF3 and GL3, there was support for less than 20% of the potential target genes and these were concentrated in very few modules, leading to ineffective sub-selection.

In addition to the functional module enrichment, de novo motif finding using Peak-Motifs[118] was performed on the sequences underneath the bound regions identified after peak calling. Selecting for potential target genes that are associated with a peak containing a significant DNA motif is based on the fact that most TFs are thought to bind at specific DNA sequences, although some bind through protein-protein interactions with other DNA-binding factors.[215] The motif-based subset improved the enrichment for DE genes, albeit less consistently than the enrichment in functional modules. The combination of both criteria led to an additional gain in enrichment for some experiments (SOC1 ChIP-Seq, FUS3, PIF5, GL3, both LFY experiments, FHY3, AP3, PI, PRR5, and both AP2 experiments; Figure B.1). We conclude that the selection of potential target genes based on enrichment in functional modules, and to a lesser extent DNA motif enrichment, complements TF perturbation data to filter genome-wide ChIP data sets towards TF-regulated genes.

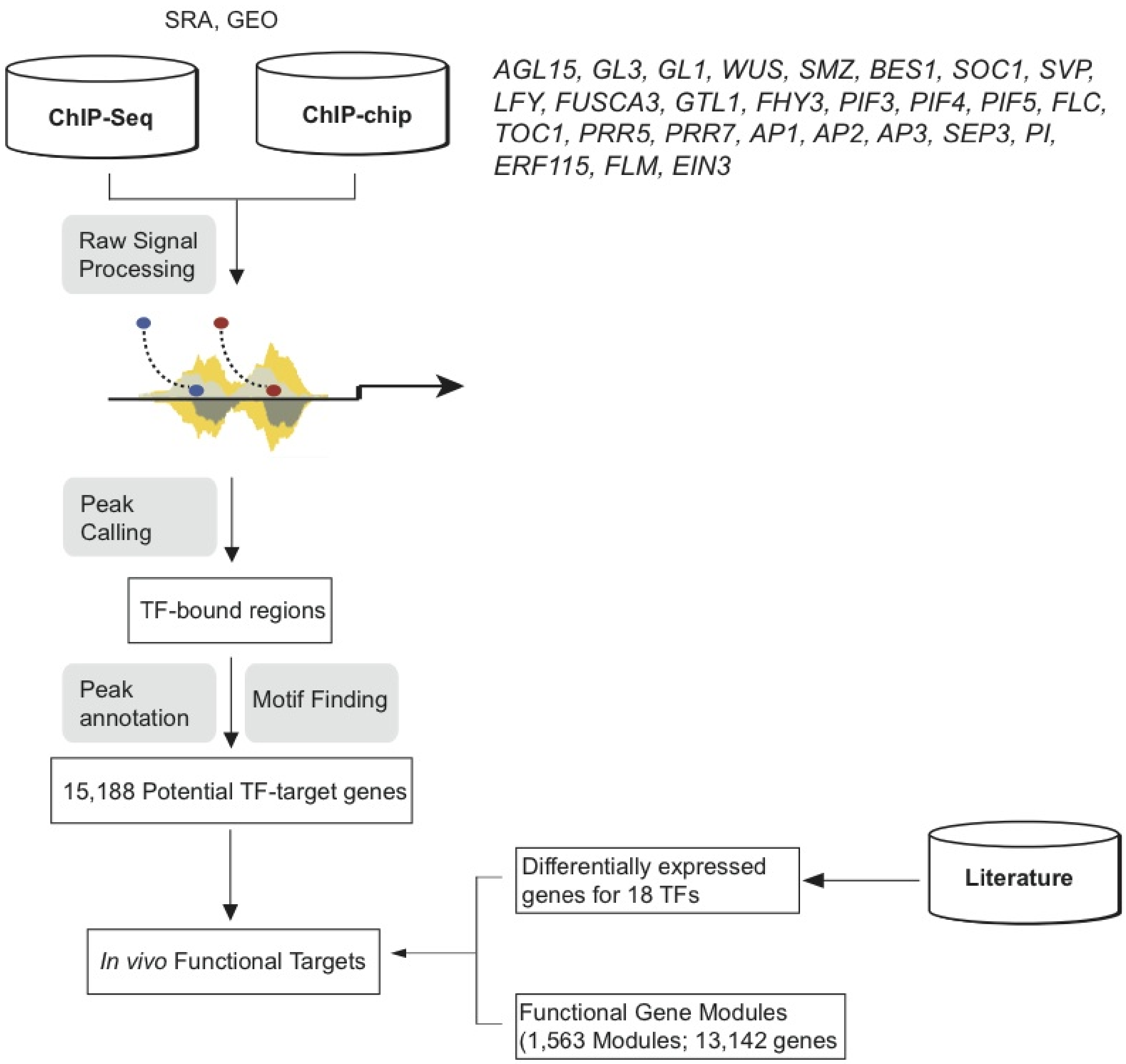

SRA, GEO
ChIP-Seq
ChIP-chip
AGL15, GL3, GL1, WUS, SMZ, BES1, SOC1, SVP, LFY, FUSCA3, GTL1, FHY3, PIF3, PIF4, PIF5, FLC, TOC1, PRR5, PRR7, AP1, AP2, AP3, SEP3, PI, ERF115, FLM, EIN3
Raw Signal Processing
Peak Calling
TF-bound regions
Peak annotation
Motif Finding
15,188 Potential TF-target genes
In vivo Functional Targets
Differentially expressed genes for 18 TFs
Literature
Functional Gene Modules (1,563 Modules; 13,142 genes

Figure 4.1: **Overview of the data and methodology used in this study.**

We made use of the results described above to extract high-confidence subnetworks. In the multiple-evidences (ME) network, a TF-target gene interaction is kept only when it has additional support of (i) DE or the complementary approach of the functional modules or (ii) a significantly enriched DNA motif. The High-Confidence (HC) network is filtered for TF-target gene interactions that are supported by both (i) and (ii). Whereas the ME network contains all 27 TFs and 10,990 potential target genes (30,072 interactions; Figure B.2A), the HC network is reduced to 25 TFs and 3,957 potential target genes (8,872 interactions; Figure B.2B online). The experiments described in this manuscript were performed on these networks in addition to the complete network and unless mentioned otherwise, results were found to be robust in the subnetworks. The entire set of peak-called regions can be accessed and downloaded[b] (see Methods). The GenomeView[54] visualisation also includes the DH sites[216] discussed below.

TF-binding properties

There are large differences in the number of potential target genes for different TFs, ranging from 56 (WUS) to 6790 (AGL15) (Figure 4.2A). While some of this variability might arise from the different experimental conditions, the similarity in the number of potential target genes for TFs that have been profiled using both ChIP-chip and ChIP-Seq indicates that those effects are minor. More important

[b] http://bioinformatics.psb.ugent.be/cig_data/RegNet/

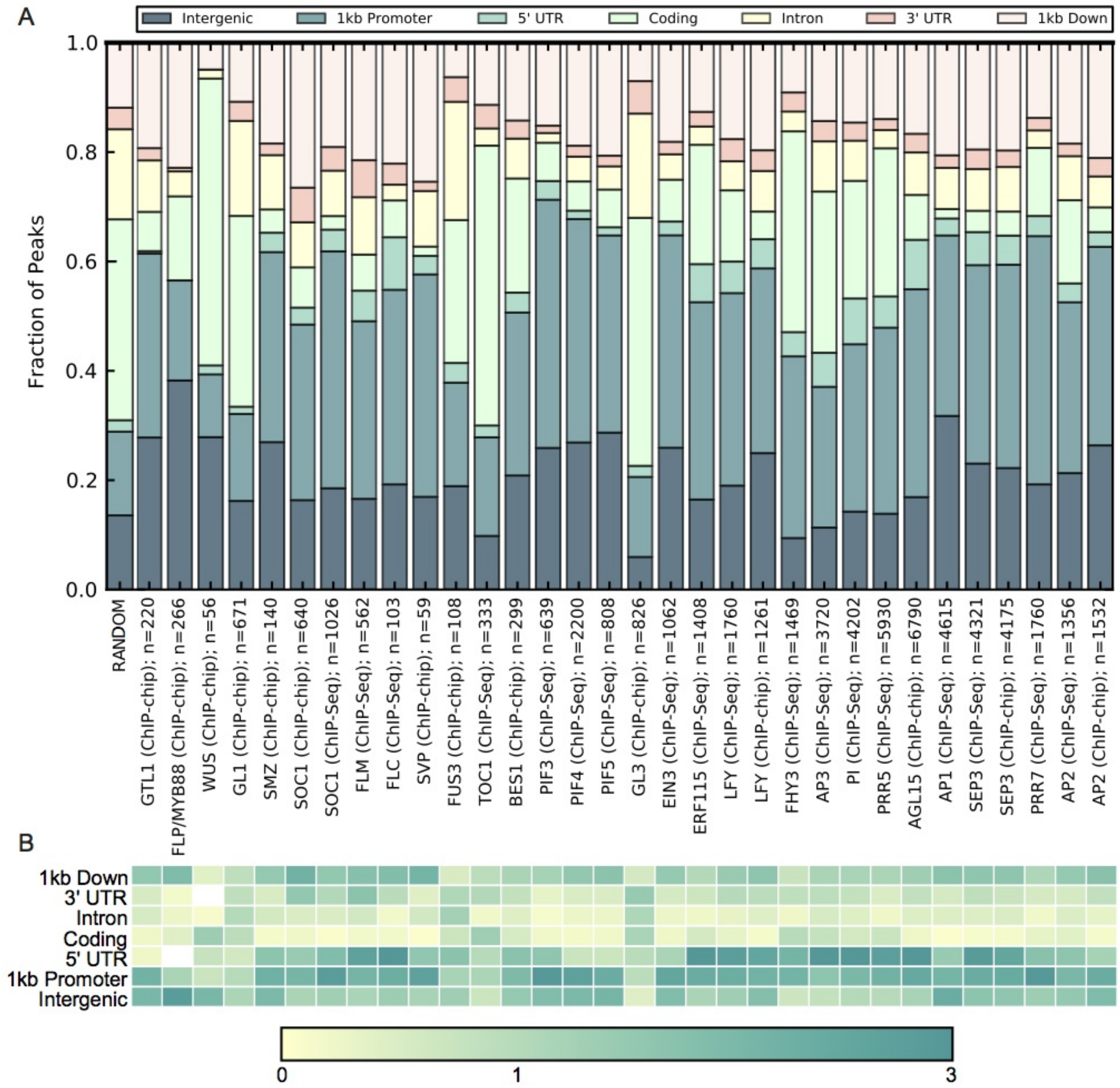


Figure 4.2: **Number of potential target genes per TF and the distribution of ChIP peaks across different types of genomic regions.** *The coloured bars represent the fractions of peaks in each of the genomic regions (left y-axis). The exact number of potential target genes is given in the labels at the bottom as n. TFs are ordered following the hierarchical clustering based on potential target gene overlap.*

than the overall number of potential target genes, is the type of genes that are bound (Figure B.3A online), and more specifically, the number of potential target genes that are gene expression regulators (TFs or miRNAs; Figure B.3B online). The highest fraction of regulators among potential target genes is 18% (for FLC). The fraction gradually lowers to 6% (for GL1), but given the sigmoidal shape of the distribution, the majority of TFs have around 12-14% potential target genes that are regulators (compared to the expected 6%). With regard to transcriptional regulation of miRNAs, the fraction of bound miRNAs ranges from 0 to 1.8% (for FUS3). Among the miRNAs that are found as potential target genes of TFs, we find known flowering regulators such as miR172 and miR156.[217] Thus, this network will also be a valuable resource to investigate transcriptional regulation of miRNAs in flowering.

A second important difference between TFs is the distribution of the types of bound genomic regions and how this compares against a random experiment (Figure 4.2A-B). Based on the function of TFs in transcriptional regulation, we would expect to see the majority of binding sites in close proximity of the potential target genes. Although most TFs exhibit depletion of exonic binding (Figure 4.2B), there are TFs with a substantial amount of intragenic binding in exons (WUS, GL1, FUS3, TOC1, GL3, ERF115, BES1, FHY3, AP3, PI, PRR5, AP2, PRR7). To ensure that the differences in binding distribution between TFs were not an effect of assigning a bound region based on its 1-bp-peak summit, the observed distributions were confirmed based on the overlap using the entire peak regions (Figure B.4). The robustness of TF binding sites in codons in the ME and HC subnetworks (Figure B.2) confirms their relevance. They might be instances of what has been termed dual-use codons in plants.[218]

Concerning the position of the binding sites with respect to the gene, we observed that 57% and 28% of the binding events are upstream and downstream of the potential target gene, respectively. Overall, 89% (23,891 / 26,717) of all upstream binding sites are within 2 kb of the transcription start site (73% in 1kb promoter). At the 3' end of the gene, 91% (11,687 / 12,828) of all binding sites are within 2 kb and 72% within 1 kb from the transcription stop. The highest fraction of binding for all TFs is close to the transcription start site (Figure B.5). To group TFs having similar binding profiles within a locus context, we clustered binding information for the different TFs. Whereas for some TFs binding is restricted to a small region around the gene body (see clusters 1, 6 and 7 in Figure B.5), the binding landscape of clusters 2, 4 and 5 is more diffuse across the 2 kb upstream region (e.g. AP1). SVP (cluster 3) is unique based on the fact that it is the only TF in the data set with substantial binding at 300-400 bp downstream of the transcription termination site.

Detection of hub targets and HOT regions

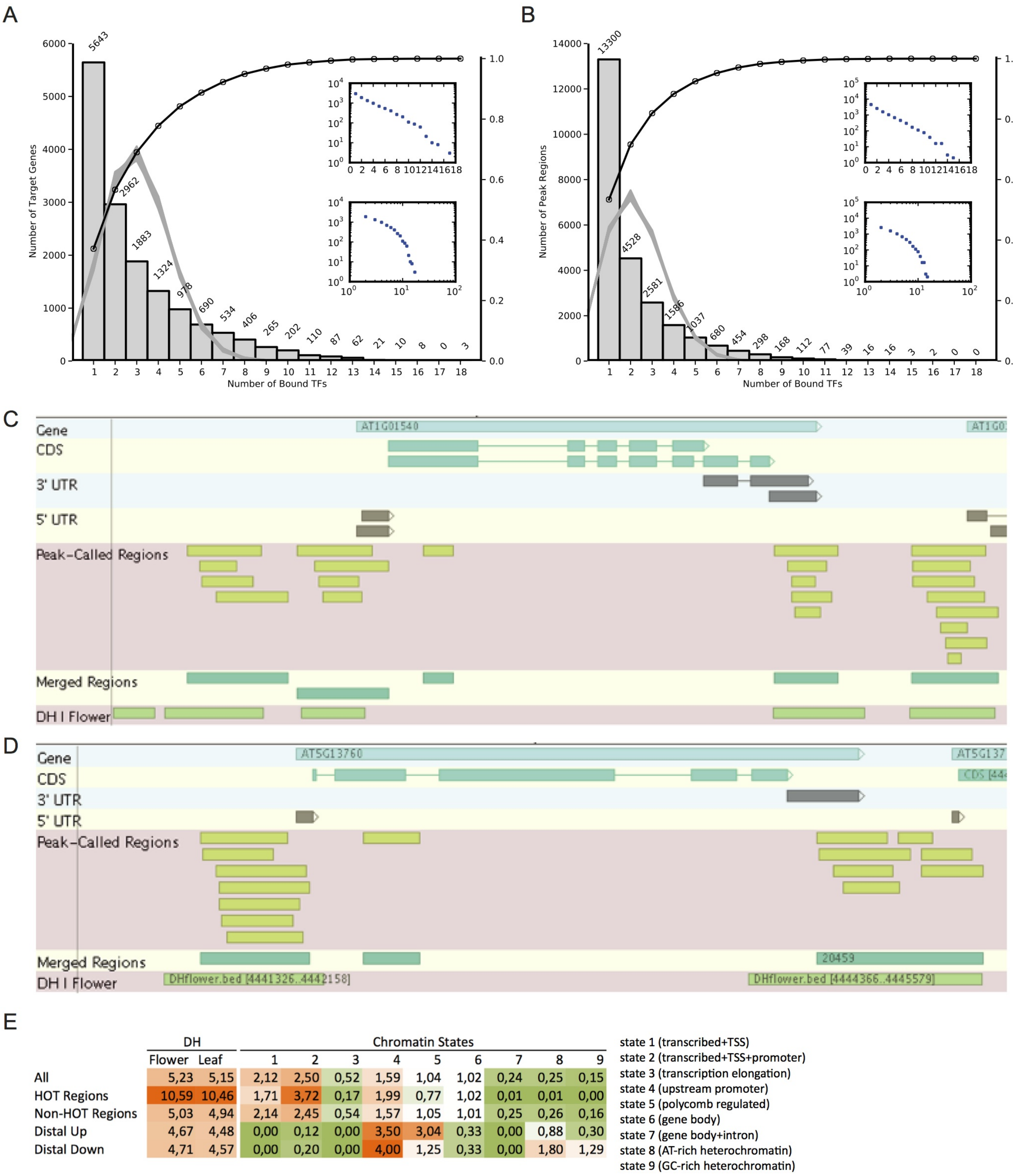
To estimate the complexity of gene regulation in the network, all TF-target gene interactions were integrated for the 27 unique TFs. The majority (63%) of the potential target genes are bound by more than one TF (Figure 4.3A), but the number of genes decreases rapidly for an increasing number of bound TFs, reaching a maximum of 18 bound TFs per potential target gene. The distribution itself best fits an exponential seen as a linear relation in a log-y scale (top insets Figure 4.3A), instead of the more commonly described power-law (which would be linear in a log-log scale, bottom inset). In a network context, hub genes are attributed the important function of providing crosstalk between different processes.[137] To delineate the hub genes in the ChIP gene regulatory network, a random TF-gene target distribution was built (Figure 4.3A) by randomising the relationships between TFs and potential target genes while preserving the number of potential target genes per TF.[213] Based on the 99th percentile values of the randomised distributions, we defined the 1,174 potential target genes that are bound by eight TFs or more as target hubs. Non-hub genes include all other genes.

In complement to the hub target genes, we delineated HOT regions in the genome as regions in which many TFs bind. HOT regions differ from hub genes as the hub genes can be bound by many TFs each binding at a different position (Figure 4.3C-D). To delineate HOT regions, all peak-called regions from all 27 TFs were merged (see Methods) and collapsed. To avoid chaining of multiple single-bound regions into long stretches based on limited overlap, all peaks were trimmed to regions of 235 bp at each side of the summit (unless original regions were shorter), which is the average length of all peaks (Figure B.6A). This resulted in conservative 'merged regions' with a median length of 349 bp that were used to identify HOT regions (Figure 4.3C; Figure B.6B online). The region occupancy followed an exponential curve, where approximately 44% of the regions are bound by more than 1 TF (Figure 4.3B). A total of 1,185 HOT regions were defined as those being bound by seven or more TFs. Non-HOT regions include all other merged regions.

Whereas hub genes measure TF complexity at the level of the target gene, HOT regions define how many TFs bind to the same region at such close proximity that the ChIP peaks could not be discerned from each other. Similar to peak annotation of individual binding events, each HOT region is assigned to the closest gene to obtain the potential target genes associated with HOT regions. Based on the two gene lists, we observe that of the 1,174 hub genes, 355 (30%) are not associated with HOT regions, because of the TFs binding at different regions (Figure 4.3D). The distributions are robust in the ME and HC subnetworks (Figure B.7).

Target hubs are enriched for regulatory genes

Through the integration of different datasets, the regulatory complexity was also functionally investigated. Hub genes are significantly enriched for genes involved in stimulus responses, development,



	DH		Chromatin States								
	Flower	Leaf	1	2	3	4	5	6	7	8	9
All	5,23	5,15	2,12	2,50	0,52	1,59	1,04	1,02	0,24	0,25	0,15
HOT Regions	10,59	10,46	1,71	3,72	0,17	1,99	0,77	1,02	0,01	0,01	0,00
Non-HOT Regions	5,03	4,94	2,14	2,45	0,54	1,57	1,05	1,01	0,25	0,26	0,16
Distal Up	4,67	4,48	0,00	0,12	0,00	3,50	3,04	0,33	0,00	0,88	0,30
Distal Down	4,71	4,57	0,00	0,20	0,00	4,00	1,25	0,33	0,00	1,80	1,29

state 1 (transcribed+TSS)
state 2 (transcribed+TSS+promoter)
state 3 (transcription elongation)
state 4 (upstream promoter)
state 5 (polycomb regulated)
state 6 (gene body)
state 7 (gene body+intron)
state 8 (AT-rich heterochromatin)
state 9 (GC-rich heterochromatin)

Figure 4.3: **Organisation of Hub genes and HOT regions.** *(A) Histogram of the number of regulating TFs per potential target gene and (B) per peak region. The black line is the cumulative number of targets, the grey band is the ensemble of 1,000 random distributions. The insets are log, or double log transformed representations of the same data. (C) Four examples of peak region merging. (D) Example of a hub gene, which is not classified as HOT-associated because the set of the regulating TFs bind at two distinct regions (5' and 3'). The 'Peak-Called Regions' track contains all regions called in any of the single TF experiments used for the Hub and HOT analysis. Zooming in shows the name of the TF binding at each region based on the name of the region. The âĂŸMerged RegionsâĂŹ track shows the result of our merging procedure of separate binding regions into genomic binding regions. The DH I Flower track shows the results of the study by Zhang et al.*[216]*. (E) Enrichment of bound regions for DH sites in flower and leaf and in the chromatin states delineated by Sequeira-Mendes et al.*[93]*.*

signalling, and process regulation. No enrichment for these GO terms was found in the non-hub genes, nor in a more specific set of low-complexity genes, defined as potential target genes bound by one or two TFs. While these processes are enriched in hub genes in the currently profiled network, it will be important to see whether this pattern is confirmed in other subsets of the complete Arabidopsis transcriptional network.

To further explore the functional properties of hub genes, other gene function information was collected, including all TFs from AGRIS, miRNAs from ’AthaMap MicroRNA targets’[219], embryo-lethal genes[109], and the set of kinases described by PhosPhAt.[110] Although there is a significant enrichment for TFs in the entire set of potential target genes, the enrichment is dependent on the level of target complexity: there is a significant 3-fold enrichment of TFs in hubs while they are significantly underrepresented among genes bound by less than three TFs (Fold enrichment [FE] = 0.87). Similarly, there is a significant enrichment for kinases in the hub genes (FE = 3.15). No enrichment could be found for miRNAs or embryo-lethal genes among the hub genes.

In addition to evaluating the enrichment of miRNAs and kinases in the set of TF hubs, we determined hub target genes of the miRNAs and kinases in their respective networks in the same manner as in the TF network (Figure B.8). Kinase hub targets are defined as being phosphorylated by ≥ 5 kinases, whereas miRNA hub targets are regulated ≥ 6 miRNAs. Interestingly, both the miRNA and kinase hubs are significantly enriched for DNA-dependent nucleic acid binding and TF activity. Three kinase hubs (ATBZIP12, BIN2, and ABI5) are also TF target hubs, all of which are involved in brassinosteroid signalling. The enrichment for TF activity in hubs of different network types reveals that genes related to transcriptional regulation are also complexly regulated through other regulatory mechanisms.

Expression levels are correlated with the total number of bound TFs

Apart from function, we evaluated expression of the potential target genes in the context of regulatory complexity (see Methods). Because our TF set involved a large number of known flowering regulators (Table 4.1), we focused on potential target genes associated with flowering based on the functional modules (n=406 genes). They were divided into low-complexity genes (bound by $\leq$3 TFs), intermediate-complexity (bound by 3-7 TFs), and hub or high-complexity genes, and compared using the Kolmogorov-Smirnov (KS) test.

Expression breadth, defined as the number of conditions in which a gene is expressed, is positively correlated with the number of regulating TFs of the potential target genes (Figure 4.6; p-value $\leq$ 0.05). Although high-complexity genes also display a U-shaped distribution with some genes being expressed in only a few conditions, genes expressed in only a single condition are most frequently bound by only one or a few TFs. To determine whether the observed correlation was due to the presence of HOT regions or the added complexity of all nearby bound regions, we compared the distributions for the hub genes (Figure 4.6) and those of the HOT-associated genes, and found the shift was not significant when comparing HOT- and non-HOT-associated genes (Figure B.9). Therefore, we conclude that the total regulatory TF complexity of the potential target genes is the main responsible factor. This is supported by the same analyses performed on the subnetworks, where the shift is consistently larger for hub target genes than for HOT-associated target genes (data not shown). Similarly, using median gene expression levels instead of expression breadth confirms this bias (Figure B.10).

To assess whether the signal was due to a difference in CG content of HOT regions, we calculated the %GC of the bound regions in relation to its complexity (i.e. number of bound TFs). Figure 4.4 Shows the %CG histograms for the different types of bound regions (low, intermediate, and HOT). Whereas there indeed seems to be a difference in the shape of the distributions (confirmed by a Kolmogorov-Smirnov test), the difference is due to the reduced variation with higher binding scores (Figure 4.5). It is not the case that the regions with the lowest number of bound TFs exhibit the highest %CG overall. We do not observe a correlation between complexity and %GC.

HOT regions are enriched for DNase I hypersensitive (DH) sites

A common characteristic of all genomic regions associated with regulatory proteins is a pronounced sensitivity to DNase I digestion.[216] We evaluated the overlap between DH sites from flower and leaf[216] with our merged regions describing TF binding (Figure 4.3E). All bound regions (non-HOT and HOT) are significantly enriched for flower DH sites (p-value $\leq$ 0.001), with the enrichment in HOT regions being twice as high as in non-HOT regions. The fraction of HOT regions that overlap with DH sites is 87%, compared to 55% for non-HOT regions. The same patterns were observed when using the DH sites

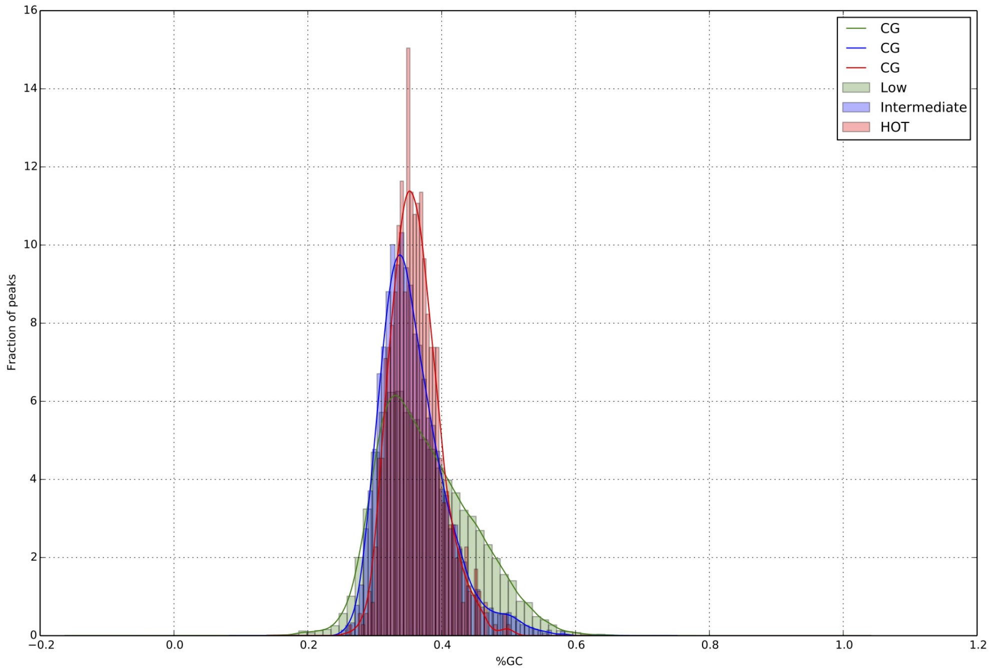


Figure 4.4: **Normed histograms of the %CG values for low, intermediate and high-complexity (HOT) bound regions**

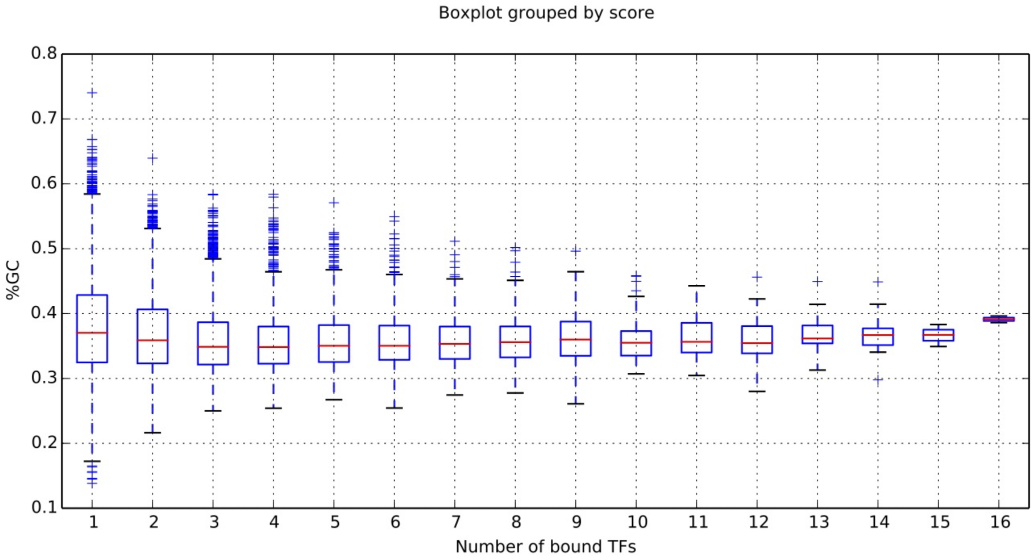


Figure 4.5: **Boxplots of %CG in merged region in function of the number of bound TFs (score).**

determined in leaf tissue. The significant overlap of DH sites with bound regions in general confirms their susceptibility to transcriptional regulation while the higher enrichment for HOT regions suggest a more steady open chromatin state, possibly because of the high number of TF binding events.

Hub and HOT-associated genes respond to TF perturbation

Next, we investigated how TF perturbation affected potential target genes, and how this was reflected by regulatory complexity. Van Nostrand and Kim (2013) reported that HOT-associated potential target genes in C. elegans are less responsive to TF perturbation in C. elegans. For each of the 18 TFs with

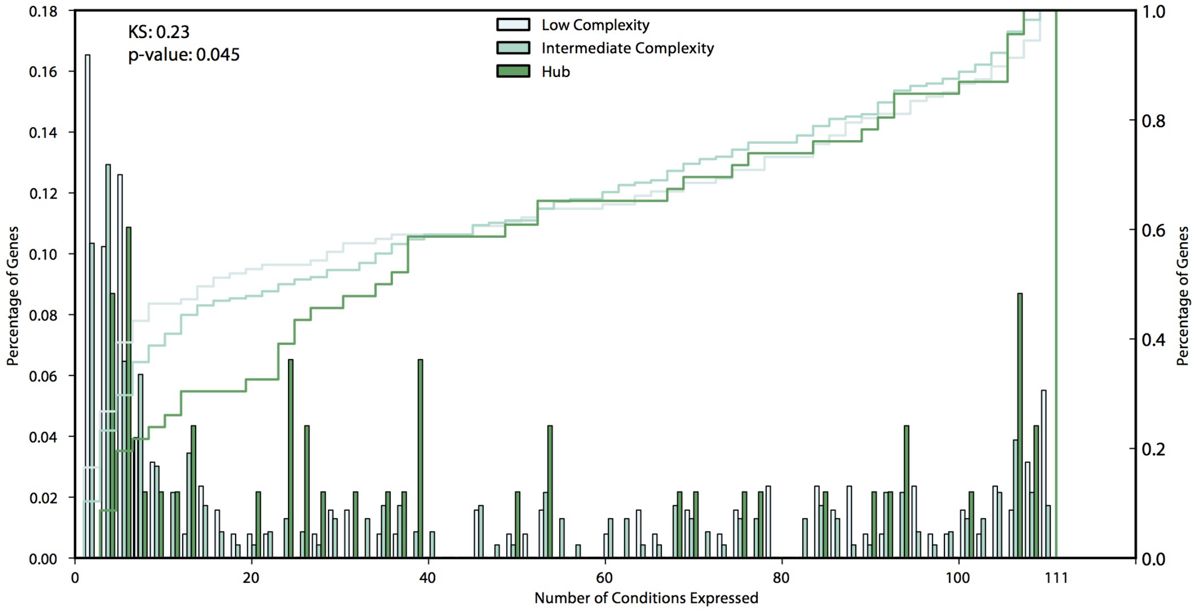


Figure 4.6: **Expression breadth as a function of regulatory complexity.** *Expression breadth distributions based on a non-redundant expression compendium of 111 conditions for three series of complexity: Low: ≤ 3 TFs; Intermediate: ≥ 3 TFs and $\leq$ 8TFs; hub: ≥ 8 TFs (n=406). The lines indicate the cumulative histograms. The Kolmogorov-Smirnov (KS) statistic and p-value are calculated between the low complexity and the hub series.*

perturbation data in our data set, we compared the enrichment for DE genes, defined as genes that respond to perturbation of the profiled TF among non-hub - non-HOT genes (low-complexity binding) and hub - HOT genes (high-complexity binding). Overall both low- and high-complexity bound genes are significantly enriched for DE genes, and in most (13/18) data sets, there is no significant reduction in expression responsiveness in hub genes or HOT-associated genes (Figure B.11). Also, TFs display higher DE enrichment in the high-complexity bound gene sets. Deviating patterns are found for some specific TFs: PIF3 potential target genes show higher DE enrichment in non-hub genes and non-HOT-associated genes, while FUS3-, PIF4-, LFY-, and PI-bound genes exhibit almost no difference in enrichment. Only FLC potential target genes have different patterns for hub and HOT-associated genes.

Chromatin states of bound regions

The Arabidopsis genome can be divided into nine chromatin states[93] based on nine genome-wide histone modification marks, three histone variants, nucleosome density, genomic G+C content, and CG methylated residues. The combination of these marks into signatures or states holds more power for functional association than different marks in isolation. With regard to our set of TF-bound regions (Figure 4.3E), we observed significant enrichment for state 1 (associated with transcribed regions and transcription start sites), 2 (similar to 1, but lower nucleosome density and located outside the gene body but in the promoter) and 4 (similar to state 2, but with fewer active marks, mostly overlapping with non-coding intergenic regions and upstream promoter). By contrast, the bound regions were significantly depleted for states 3 (transcription elongation), 7 (gene body and intron), 8 (AT-rich heterochromatin), and 9 (GC-rich heterochromatin). The association with states 1 and 2, and the depletion for 3 and 7 appears to be a direct consequence of the location of most bound regions near genes, and the enrichment for state 4 and depletion for states 8 and 9 confirm the functionality of the intergenic bound regions.

Based on the ChIP peak-gene distance distribution, we defined a set of 195 distal bound regions as those further than 4 kb away from the closest gene. Although a small fraction (11%) of the distal upstream bound regions lies in heterochromatic regions (state 8 and 9), they are significantly depleted for these heterochromatin-typical states. Interestingly, the remainder of the distal upstream bound regions can be split into enrichment towards states 4 and 5 (Polycomb chromatin). The Polycomb pathway is an important repressive pathway in development, including flowering, which is known to act by regulating chromatin accessibility to binding sites. When the repression is overcome, TF binding leads to target gene regulation.[132] The enrichment for state 5 suggests that the distal upstream bound regions are candidate distal elements where the chromatin is under regulation by the Polycomb complex similar to the distal element of FLOWERING LOCUS T[220], which is brought to close association with the proximal promoter through a chromatin loop.[221] While downstream distal elements appear to show similar enrichment patterns for state 4 and 5, the sample size is too small to obtain significant results.

Population sequence diversity and conservation of bound DNA

If bound regions are of functional importance for transcriptional regulation, we expect them to be under purifying selection. Based on complete re-sequencing data of 369 Arabidopsis strains from the 1001 Genomes project[222], we assessed the nucleotide diversity within the bound regions using the average number of nucleotide differences per site, π.[223] We compared the TF-bound regions with fourfold degenerate (4D) sites and other sets of genomic regions (Figure 4.7). 4D sites are thought to be the most neutrally evolving sites in the genome, as such mutations do not affect the encoded amino acid, and coding sequences are less likely to have other regulatory functions. 4D sites are indeed less constrained than either intergenic regions or 1 kb up- and downstream regions of genes (π of 0.0052, 0.0050, and 0.0034 respectively, versus π of 0.0070 for 4D sites), but bound regions have the lowest diversity (p-value $\leq$ 0.001 based on reshuffling; see Methods). The diversity of bound regions is similar to that of 5' and 3' UTRs, and almost as low as coding sequences. Importantly, the ME and HC subnetworks show only little additional constraint for bound regions (Figure 4.7).

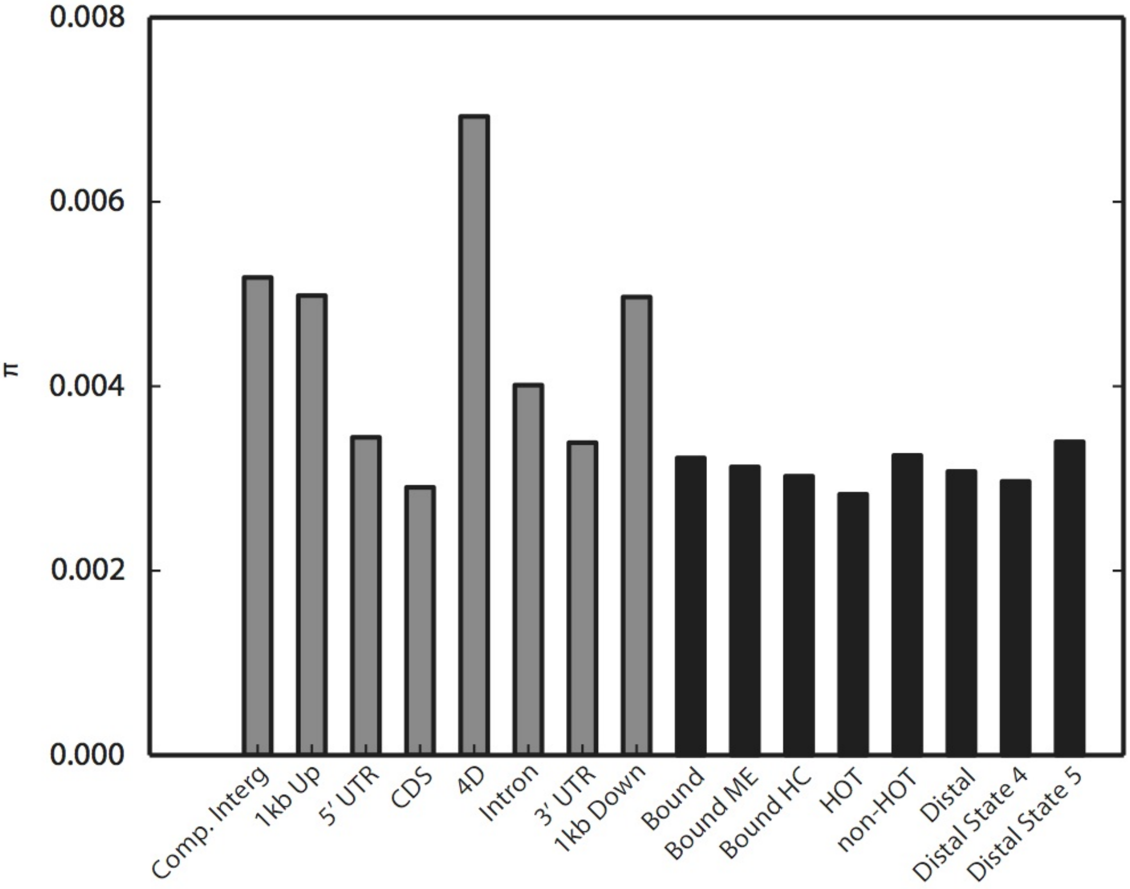


Figure 4.7: **Nucleotide diversity (π) in different sets of genomic DNA.** *Nucleotide diversity values based on 369 Arabidopsis strains for different genomic regions, including bound regions from the complete network (Bound), the subnetworks (Bound ME and Bound HC) and distal bound regions (all, and the subsets lying in the chromatin states 4 and 5). Comp. Interg. is the complete intergenic space and 4D are fourfold degenerate sites in coding sequences (CDS). Bound DNA as black bars.*

In addition, we examined HOT regions and distal bound regions in comparison to the non-HOT regions and proximal bound regions, respectively. HOT regions show reduced π values compared to the non-HOT regions, which can be explained by the necessity to retain binding sites for more TFs than non-HOT regions, and further corroborating the functionality of HOT regions. Similarly, distal bound regions show similar π values compared to regions acting proximally, providing evidence for their functionality.

Because of their function, bound regions are also often conserved across species, which is the premise of genome-wide studies of conserved non-coding sequences (CNSs). We determined the fraction of bound regions exhibiting conservation within the crucifers[224] and within the dicot lineage[225] based on overlap with CNSs. Overall, CNSs supported 35% and 29% of the 24,898 bound regions in the crucifer and dicot data, and 15% are supported in both sets. Bound regions are significantly enriched for overlap with CNSs in crucifers (3.2 fold) and dicots (1.6 fold). For the set of 1185 HOT regions, we observe that 72% and 52% overlap with a conserved region, which results in a slightly higher enrichment of HOT regions in CNSs of crucifer (3.8 fold) and dicot (1.6 fold) datasets compared to non-HOT regions (3.2 and 1.5 fold, respectively). This result complements the findings of the population sequence diversity analysis regarding the higher constraint on HOT regions.

Hypotheses to explain the diversity of motifs in bound regions

Combinatorial control, where different TFs cooperate in a context-dependent manner, is an important principle in transcriptional regulation (Singh, 1998;.[148] For all 27 TFs, we determined the overlap in potential target genes (Figure 4.8A), and clustered them accordingly. Importantly, when all experiments, including ChIP-chip and ChIP-Seq experiments for the same TF, were taken into account, all experiments of a single TF clustered together, rather than clustering based on the ChIP method used. We observed significant overlaps for 255 out of the 351 TF pairs in our data set, showing that there is high degree of overlap in the genes that are targeted by TFs involved in flowering, circadian rhythm and light response. Among the profiled TFs, there are two major protein-protein interaction clusters: light response (marked in orange), and a flowering cluster (marked in green; Figure 4.8B). Interacting TFs can be retrieved from the overlap analysis (Figure 4.8A), albeit the flowering cluster is split up in three smaller clusters, potentially revealing the more common interactions. Since HOT-associated genes have a large influence on co-binding statistics[209], the same matrix was constructed using only the non-HOT-associated genes. Although fewer significant TF pairs were found (208 / 351), the cluster structure of the matrix is robust, also when using the subnetworks (Figure B.12).

Whereas co-targeting of potential target genes reveals possible co-regulation, co-binding of TFs in close proximity of each other, i.e., in the same bound region, can identify co-binding complexes. Therefore, we integrated de novo motif finding for each of the profiled TFs (see Methods). An overview of all enriched motif logos per TF, together with their frequency and location within the peak regions is given in B.2. Importantly, motif definitions were determined stringently, meaning that differences in flanking nucleotides were considered as different motifs, as can be seen for PIF5. Flanking nucleotides have been shown to add important specificity in motif recognition, and are therefore not collapsed into a single degenerate consensus binding site.[226–228] Motifs were ranked by occurrence, with the most frequent motif denoted as the primary motif.

For each factor, we evaluated whether any of the *de novo* motifs corresponded to the canonical motif (the motif that is known to be bound by the TF, as opposed to non-canonical motifs), based on motif alignments against the AGRIS database and comparison with motifs from literature (B.3). Notably, we observed for several TFs that the primary motif is not the canonical motif. For most TFs, such as TOC1, only a single motif fitted the canonical motif description, whereas for others, such as PIF5, multiple motifs matched the canonical motif as motif differences resided in the flanking nucleotides.

In the traditional view, a DNA motif is expected to explain the binding site of the TF in the peak region. A single motif however rarely covers more than 40% of the ChIP peaks (Figure B.13A), raising the question of how the TF might be associated with the chromatin in the remainder of the peaks. Interestingly, when taking all significantly enriched motifs into account, the fraction of peaks with at least one motif increased to 45-80%. This large increase indicates different motifs are rarely present in the

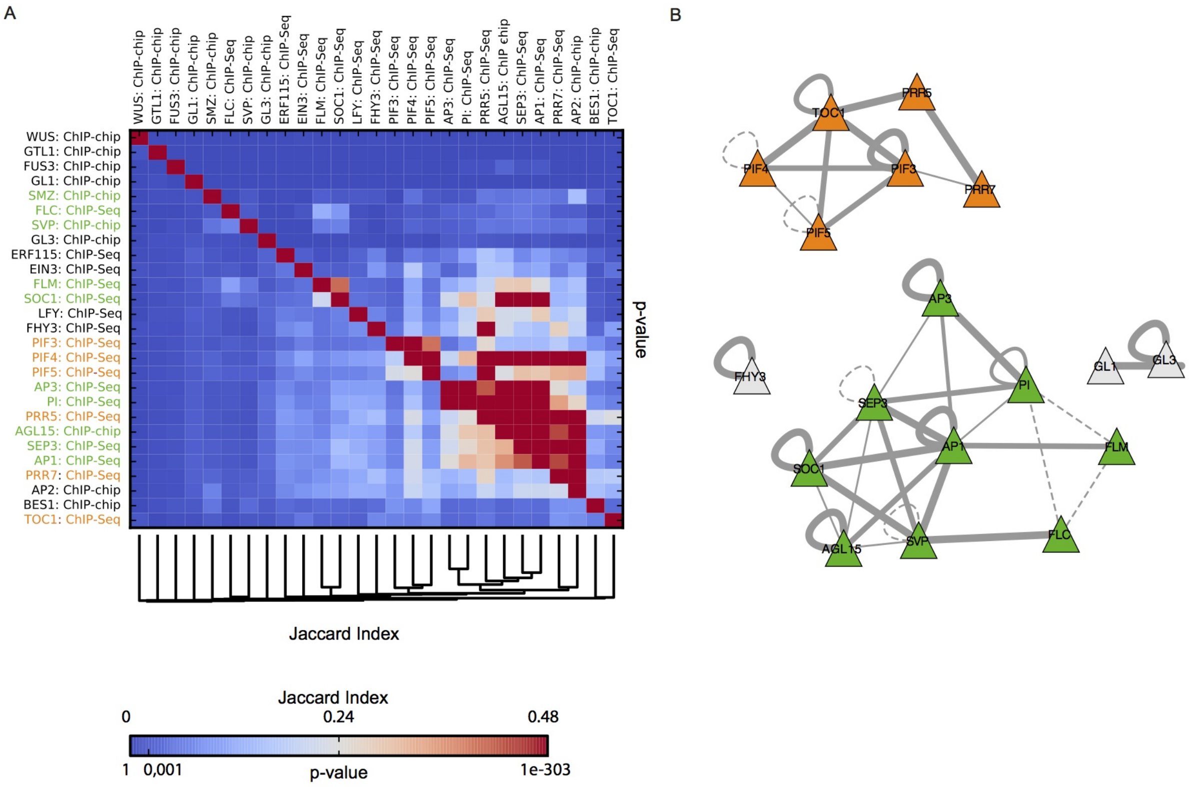


Figure 4.8: **Co-regulation and protein complexes of TFs.** *(A) TF co-binding matrix based on common target genes and average-linkage hierarchical clustering based on Jaccard Index. The lower left half displays the Jaccard index and the upper right half hypergeometric p-values of overlap between the two sets of regulated genes, corrected using the Bonferroni method. (B) Experimental and predicted PPIs between the TFs. Solid lines indicate experimentally determined PPIs and dotted lines indicate predicted interactions. Line thickness indicates number of supporting experiments.*

same subset of peaks. Enrichment for DE genes shows that the sets of genes uniquely associated with the non-primary motifs likely represent regulated target genes (with the exception of some motifs of GL1, FLC, BES1, PIF4, and GL3; Figure B.13B). Notably, we did not observe a reduction in the fraction of peaks with motif instances between non-HOT regions and HOT regions (Figure B.14).

Co-binding TFs, where one TF binds through association with another TF with a different DNA binding specificity, or where TFs modify each others' DNA binding specificity, provide a possible explanation for the widespread occurrence of different DNA motifs for the same TFs. This can only be the case if TFs bind in the same bound region. Considering all identified motifs per TF and the complete set of potential target genes, we systematically categorised binding events via canonical and non-canonical motifs. TFs can bind (i) peaks where only a canonical motif instance is present; (ii) peaks where both canonical and non-canonical motif instances are present; or (iii) peaks where only non-canonical motif instances are present. Peaks of type I fulfil the traditional view of TF binding, where a TF binds its target directly. Type II represents co-binding, where a second TF binds in cooperation with the profiled TF. The peaks of type III represent tethering, where the profiled TF associates with the chromatin through a partnering TF, an example being TCP binding via a protein-protein interaction with AS2.[212,215] Based on the fraction of these three peak types for a given TF, we observe that most TFs bind a mixture of these peak types (Figure 4.9). Only a few TFs such as SEP3 and FLM tend to bind peaks that almost always include a canonical motif.

To find explanations for the different DNA motifs in a single ChIP experiment, we assessed the co-occurrence of pairs of TFs in bound regions (Figure 4.10). From the perspective of each TF, its entire peak set was divided into the different categories of peaks and the number of co-occurrences was statistically evaluated. Starting from this matrix, we tested whether known co-binding regulators could be recovered, and derived new testable hypotheses for several TFs. Firstly, in the -ChIP data for the MADS domain protein SEP3, multiple different CArG motifs, typical for MADS domain proteins, are enriched

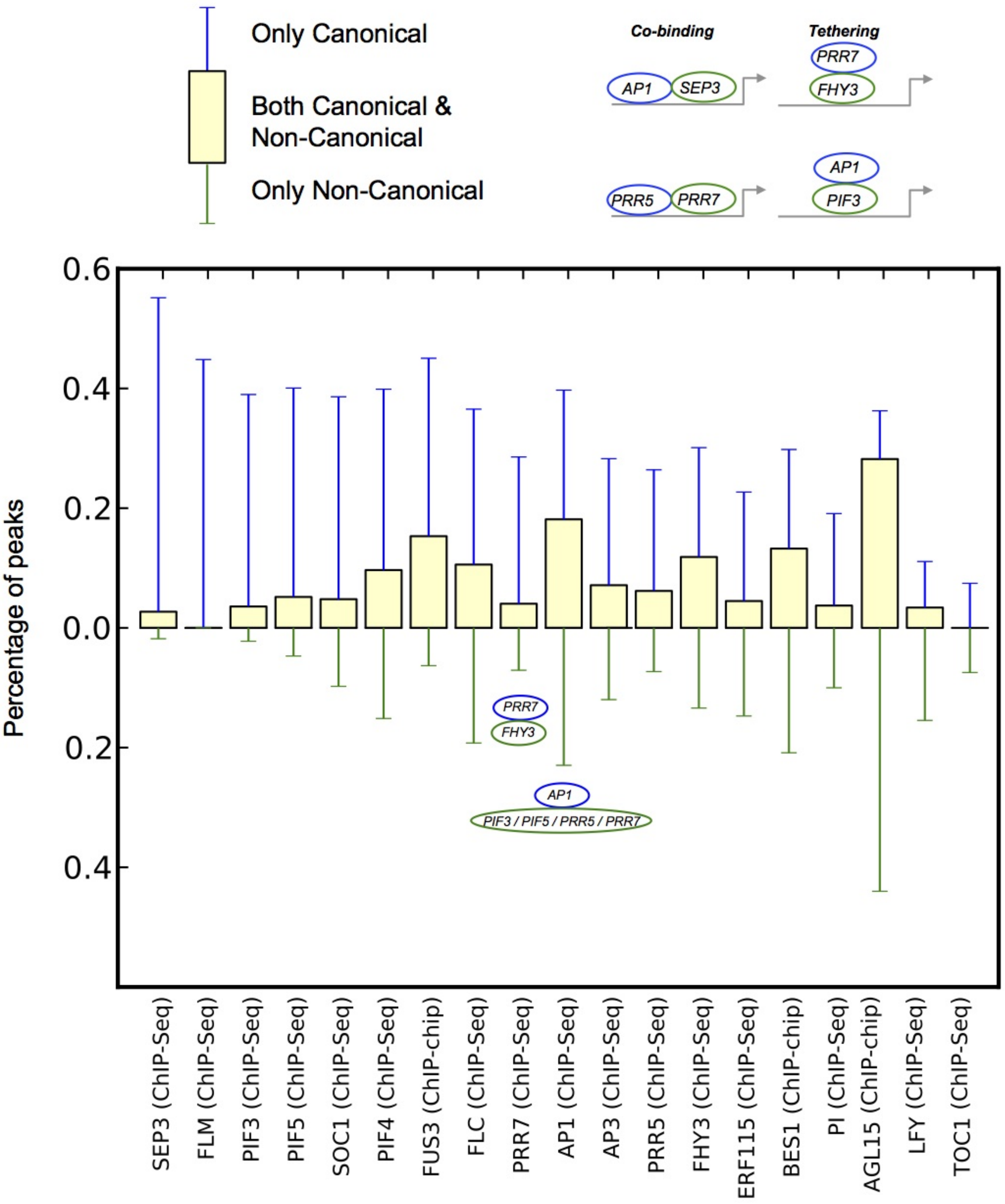


Figure 4.9: **Co-regulation and protein complexes of TFs.** *(A) TF co-binding matrix based on common target genes and average-linkage hierarchical clustering based on Jaccard Index. The lower left half displays the Jaccard index and the upper right half hypergeometric p-values of overlap between the two sets of regulated genes, corrected using the Bonferroni method. (B) Experimental and predicted PPIs between the TFs. Solid lines indicate experimentally determined PPIs and dotted lines indicate predicted interactions. Line thickness indicates number of supporting experiments.*

(B.2). Binding of many other MADS box TFs is significantly enriched in the SEP3-bound regions. All TFs that form a protein-protein interaction with SEP3[229] have high co-binding scores: AGL15, AP1, SOC1, PI, and AP3. Although there is no protein-protein interaction known or predicted between SEP3 and FLM, we observe a highly significant co-binding pattern in the same regions for these TFs as well. Overall, the co-binding of the different MADS TFs is a likely explanation for the different CArG motifs (different flanking nucleotides) found in the peaks of SEP3. Similarly, PIF3-4-5 and PRR5-7 show highly significant co-binding scores within their respective TF family members, fitting with the protein-

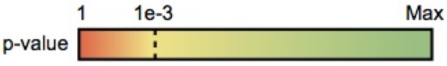
protein interactions between them. Overall, for TF pairs that have a known protein-protein interaction, the co-occurrence scores are higher compared to pairs without interactions.

Apart from canonical CArG motifs, many MADS domain TFs have non-canonical G-boxes as secondary motifs. Based on the co-binding of other TFs with MADS TFs, we attempted to identify new cooperative TF interactions. For instance, AP1-bound regions that harbour non-canonical motifs also often bind PIF5, PIF3, PRR5, and PRR7. This suggests a link between the presence of the G-box, and the co-binding of these TFs. In the PRR7 peaks, there is a relationship between the presence of the FHY3-FAR1 binding site (FBS) motif (CACGCG; Lin et al. (2007)) and FHY3 found in the PRR7 peaks. FHY3, which has an FBS motif as canonical motif, shows very high co-binding scores in the peaks with both motif types, and in those with only non-canonical motifs. The fact that PRR7 has high co-binding with FHY3 in its type II and III peaks, but low co-binding in its type I peaks with only canonical motifs, corroborates the hypothesis that the non-canonical FBS in PRR7 is explained by FHY3. A similar signal can be seen for AP1 and PRR7, and LFY and PRR7, where there is only significant co-binding in AP1 peaks where the G-box (PRR7 canonical motif) is found. In both cases, we hypothesise a tethering event.

4.3 Discussion

Large-scale analysis of TF binding can provide insights into the organisation and complexity underlying transcriptional regulation. To investigate gene regulatory networks in Arabidopsis, we have compiled an experimental network comprising 46,619 unique TF-target regulatory interactions based on 27 TF ChIP profiling experiments. Given the different data analysis methodologies of the different source studies, we reprocessed the raw data following a uniform pipeline to obtain an unbiased view on potential target genes for different TFs. Prior to our study, the AtRegNet platform has made great efforts to collect and store all Arabidopsis regulatory information from both small- and large-scale studies.[17] However, given the rapid increase in genome-wide ChIP studies in Arabidopsis, the AtRegNet database as of the writing this manuscript is lacking 21 of the experiments included in this study. In contrast to AtRegNet, we did not include data from small scale-studies, as we were primarily interested in discerning binding patterns and properties of TFs for which global genomic binding information is required. Through the integration of different functional datasets including Gene Ontology, functional modules, embryo-lethal genes, miRNAs and kinases, as well as DNA motif finding information, our gene regulatory network provides a functional view of TF regulation in Arabidopsis as well as an entry point to predict functions for unknown genes in the set of potential target genes.

To investigate the organisation of regulation and binding sites among the potential target genes, all ChIP data sets were merged, and the distributions of the number of regulators per potential target gene and number of binding events per region were quantified. In both cases, an exponential distribution was observed, which is distinct from the commonly described power-law in biological networks.[137] However, the exponential distribution was also reported in the C. elegans gene regulatory network by Cheng et al. (2011). We delineated hub genes and HOT regions, two proxies for complex gene regulation. In contrast to the modENCODE study[206] where HOT regions had to be bound by more than 65% of the profiled TFs, our definition of HOT regions is based on a percentile score inferred through network randomizations, as was done by Shalgi and co-workers[230], avoiding a static ad-hoc threshold. Functional analysis of the potential target genes revealed that the genes bound by few TFs are depleted for TFs, while potential target genes with high TF complexity were enriched for TFs. In addition, TFs were also enriched in the hubs of kinase and miRNA networks, showing that regulatory genes in plants, such as those involved in hormone signalling, are complexly targeted in different types of regulatory networks. Through overlap analysis with DH sites, all bound regions showed significant enrichment for open chromatin regions. HOT regions consistently exhibited higher enrichments, likely caused by a constraint on the chromatin to maintain an open conformation because of the high number of binding TFs. This open conformation raises concerns about whether the binding in HOT regions truly affects the regulation of the associated target gene, or merely represents a state of massive TF binding, due to increased local accessibility of DNA, without any regulatory consequences. There is evidence in non-plant species that (i) at H. sapiens HOT regions, TF occupancy is strongly predictive of transcription preinitiation complex recruitment



ID	Type	AGL15	BES1	FUS3	ERF115	SEP3	AP1	LFY	FLC	FHY3	PIF5	PIF4	TOC1	PRR5	PI	AP3	SOC1	PIF3	PRR7	FLM
AGL15	nCan	▨	4,48	0,06	0,16	50,72	50,82	0,00	7,95	0,37	8,05	10,25	0,72	5,70	41,22	10,01	46,58	4,13	24,92	28,98
	Both	▨	7,04	0,68	0,12	93,68	127,76	0,13	8,15	0,07	9,75	29,42	0,10	11,43	67,04	28,28	64,71	8,99	26,76	31,78
	Can	▨	0,01	1,15	0,01	8,34	9,60	0,17	1,26	0,00	0,62	0,19	0,04	0,00	1,79	0,35	10,42	0,01	0,17	5,39
	PPI	▨	-	-	-	+	+	-	-	-	-	-	-	-	-	-	+	-	-	-
BES1	nCan	3,95	▨	0,00	1,09	10,22	5,56	0,43	0,00	5,12	18,60	15,35	4,71	22,23	2,09	2,66	1,27	6,10	20,86	1,81
	Both	4,66	▨	0,00	2,06	4,39	3,97	0,42	0,00	10,00	10,52	15,00	11,34	18,47	8,56	7,68	1,03	10,71	26,29	1,18
	Can	2,17	▨	0,70	2,59	3,72	1,34	1,84	1,73	0,20	18,74	17,15	2,66	18,76	3,62	2,77	0,77	7,18	20,31	3,83
	PPI	-	▨	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
FUS3	nCan	0,39	0,00	▨	0,46	0,35	0,05	0,36	0,00	0,43	0,66	0,00	0,00	0,43	0,00	0,00	0,00	0,00	0,00	0,82
	Both	0,53	1,75	▨	0,57	0,31	0,32	0,41	0,00	0,00	0,35	0,28	0,00	0,09	0,04	0,00	0,00	0,00	0,85	0,00
	Can	0,10	0,00	▨	0,00	0,15	0,10	0,59	0,00	0,00	0,16	0,18	0,00	0,07	0,03	0,00	0,38	0,00	0,32	0,00
	PPI	-	-	▨	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
ERF115	nCan	0,13	0,88	1,12	▨	0,15	0,06	0,05	0,22	0,26	1,38	0,42	1,29	0,45	0,10	0,05	0,45	0,87	1,42	0,05
	Both	0,39	2,05	0,00	▨	0,02	0,03	0,01	0,00	0,37	1,56	0,28	0,15	0,29	0,03	0,18	0,10	0,98	0,57	0,00
	Can	0,45	0,94	0,00	▨	0,00	0,00	0,01	0,00	1,36	0,33	0,10	0,13	0,88	0,17	0,03	0,29	0,07	2,35	0,01
	PPI	-	-	-	▨	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
SEP3	nCan	1,48	0,49	0,00	0,03	▨	12,37	0,11	0,00	0,23	0,63	5,17	1,00	2,02	1,40	0,92	2,55	0,14	2,52	0,73
	Both	8,38	2,09	0,33	0,83	▨	27,83	1,52	1,65	0,25	4,45	6,26	0,57	8,89	9,19	5,00	7,29	0,47	6,28	4,51
	Can	132,18	11,06	0,06	0,00	▨	300,00	0,21	13,99	0,00	27,83	53,43	0,31	29,95	60,04	22,88	159,02	6,45	39,14	72,18
	PPI	+	-	-	-	▨	+	-	-	-	-	-	-	-	+	+	+	-	-	-
AP1	nCan	44,39	3,81	0,09	0,02	109,55	▨	0,29	3,33	0,00	14,64	25,87	0,57	4,40	35,76	11,78	31,27	8,97	13,76	8,42
	Both	77,01	5,65	0,34	0,01	154,55	▨	1,82	8,29	0,00	9,28	34,20	0,87	15,81	74,10	32,10	21,64	4,50	25,43	20,15
	Can	32,81	0,87	0,12	0,00	88,29	▨	1,58	0,43	0,00	1,53	14,50	0,00	0,52	45,39	18,16	16,65	0,36	2,24	6,59
	PPI	+	-	-	-	+	▨	-	-	-	-	-	-	-	+	+	+	-	-	+
LFY	nCan	0,97	1,12	0,82	0,01	2,12	1,61	▨	0,44	0,18	2,71	3,49	0,71	6,55	2,81	0,91	1,65	1,79	20,25	2,69
	Both	1,12	0,00	0,00	0,96	1,47	1,20	▨	0,00	0,36	2,64	3,48	0,89	2,43	2,64	0,66	2,60	0,96	4,87	0,33
	Can	0,02	0,56	0,00	0,05	0,10	0,24	▨	0,33	0,12	0,58	0,25	0,11	0,82	1,50	0,36	0,48	0,04	1,32	0,16
	PPI	-	-	-	-	-	-	▨	-	-	-	-	-	-	-	-	-	-	-	-
FLC	nCan	3,22	0,66	0,00	0,00	5,51	4,27	0,69	▨	0,13	3,32	3,36	0,51	2,17	1,84	2,34	9,96	2,76	5,09	8,10
	Both	4,02	0,00	0,00	0,31	3,50	2,77	0,00	▨	0,00	2,26	1,04	0,00	1,05	4,51	3,02	8,76	1,48	2,11	15,34
	Can	6,71	1,37	0,00	0,09	3,44	1,50	0,20	▨	0,28	1,19	0,88	0,00	1,73	2,25	2,35	21,55	0,79	0,45	31,25
	PPI	-	-	-	-	-	-	-	▨	-	-	-	-	-	+	-	-	-	-	+
FHY3	nCan	1,42	7,26	0,21	2,38	0,39	0,04	0,02	0,23	▨	4,82	5,55	9,34	19,86	3,62	1,89	1,97	4,77	16,95	0,14
	Both	1,95	5,39	0,00	0,77	1,86	0,46	0,96	0,00	▨	8,61	6,87	4,88	41,37	8,05	5,47	1,72	6,01	45,68	2,28
	Can	0,95	1,58	0,14	0,05	0,89	0,01	0,47	0,91	▨	1,61	2,36	1,36	18,35	0,89	0,47	0,31	0,43	10,06	0,62
	PPI	-	-	-	-	-	-	-	-	▨	-	-	-	-	-	-	-	-	-	-
PIF5	nCan	1,38	7,54	0,00	0,64	2,22	1,96	0,67	0,80	16,71	▨	26,90	0,76	12,15	1,77	1,65	0,27	7,04	13,65	1,83
	Both	0,31	9,89	0,00	0,55	4,02	1,22	0,56	0,00	13,20	▨	24,42	6,96	17,77	1,40	1,67	0,06	30,67	22,27	0,18
	Can	6,33	24,31	0,41	1,60	16,65	11,40	3,98	2,76	0,18	▨	167,94	3,76	53,31	2,24	1,00	3,86	108,39	41,22	2,68
	PPI	-	-	-	-	-	-	-	-	-	▨	+	+	-	-	-	-	+	-	-
PIF4	nCan	4,50	4,15	0,00	0,17	6,52	12,59	0,24	1,12	4,61	11,05	▨	1,30	9,49	2,18	1,36	4,27	22,34	15,69	3,60
	Both	5,65	24,75	0,51	2,59	11,48	7,04	1,83	1,05	34,60	81,04	▨	7,96	51,32	6,44	6,21	3,56	68,96	54,86	5,65
	Can	4,33	19,49	0,18	0,03	41,10	20,30	1,07	1,84	0,00	261,76	▨	4,86	59,96	2,25	0,67	4,40	136,04	55,55	5,02
	PPI	-	-	-	-	-	-	-	-	-	+	▨	+	-	-	-	-	+	-	-
TOC1	nCan	1,56	6,57	0,00	0,42	1,50	2,11	0,25	0,00	7,56	4,51	5,15	▨	11,47	3,18	3,54	1,69	1,78	14,94	0,00
	Both	0,49	1,92	0,00	1,22	0,00	0,00	0,00	0,00	1,19	1,46	1,00	▨	0,51	0,67	0,75	0,00	1,57	1,11	0,00
	Can	0,74	5,29	0,00	0,18	2,76	0,43	0,48	0,00	2,09	5,46	14,96	▨	16,97	6,34	3,00	0,09	9,92	31,96	0,00
	PPI	-	-	-	-	-	-	-	-	-	+	+	▨	+	-	-	-	+	-	-
PRR5	nCan	4,74	3,66	0,33	1,04	9,18	3,22	0,51	0,37	0,00	5,11	9,18	0,98	▨	0,99	0,18	8,47	0,26	30,04	6,69
	Both	13,61	19,88	0,75	7,27	14,01	6,61	3,40	1,26	27,14	36,31	52,72	10,33	▨	11,19	9,07	3,09	25,99	174,11	2,22
	Can	16,46	53,71	0,03	3,62	23,48	10,23	3,73	4,05	121,89	102,81	136,39	20,88	▨	27,87	16,36	8,76	69,83	286,26	5,84
	PPI	-	-	-	-	-	-	-	-	-	-	-	+	▨	-	-	-	-	+	-
PI	nCan	33,73	23,22	0,16	6,45	34,82	34,00	6,38	4,11	10,48	31,38	44,04	14,65	76,14	▨	132,18	16,09	23,36	90,38	6,43
	Both	19,36	10,12	0,00	1,42	9,93	20,04	1,09	1,26	9,12	5,04	21,52	6,55	31,90	▨	82,66	9,54	3,87	29,92	5,17
	Can	30,23	1,37	0,30	0,08	21,43	69,46	1,53	1,26	0,93	0,60	2,65	0,19	0,69	▨	210,87	22,56	0,57	1,93	6,58
	PPI	-	-	-	-	+	+	-	+	-	-	-	-	-	▨	+	-	-	-	+
AP3	nCan	4,73	0,93	0,04	0,02	1,16	5,53	0,04	0,73	0,33	0,65	1,83	0,40	0,25	182,88	▨	4,44	0,58	4,57	3,80
	Both	21,06	1,36	1,25	0,34	15,87	22,60	1,34	1,94	0,05	1,03	1,00	1,12	0,57	145,36	▨	15,08	0,61	3,69	7,95
	Can	37,66	4,46	0,00	0,57	32,69	59,44	1,14	5,75	2,91	0,99	4,78	1,17	2,44	300,00	▨	23,50	0,11	9,40	12,25
	PPI	-	-	-	-	+	+	-	-	-	-	-	-	-	+	▨	-	-	-	-
SOC1	nCan	17,54	1,39	0,00	0,52	28,52	19,29	0,91	7,82	0,85	2,44	7,50	1,33	13,01	14,10	6,98	▨	0,50	18,14	17,72
	Both	11,19	2,44	0,68	1,05	18,26	15,50	0,23	5,59	1,78	8,80	8,01	0,21	3,77	12,89	6,26	▨	4,18	8,28	21,96
	Can	76,57	1,10	0,09	0,02	75,68	52,22	0,02	13,85	0,19	0,30	0,78	0,29	0,85	33,36	21,35	▨	0,01	1,52	78,84
	PPI	+	-	-	-	+	+	-	-	-	-	-	-	-	-	-	▨	-	-	-
PIF3	nCan	1,32	3,23	0,00	0,24	0,23	0,87	0,00	0,00	5,00	1,95	2,81	1,58	3,48	0,47	0,64	0,34	▨	3,41	0,55
	Both	0,38	4,96	0,00	0,73	1,64	2,27	0,23	0,00	13,07	22,71	18,71	4,09	11,00	1,19	1,20	0,53	▨	13,08	2,60
	Can	4,67	10,87	0,00	0,07	7,19	3,60	0,64	2,49	0,74	141,01	137,75	10,00	36,19	2,44	0,64	0,65	▨	38,34	3,33
	PPI	-	-	-	-	-	-	-	-	-	+	+	+	-	-	-	-	▨	+	-
PRR7	nCan	2,13	5,96	0,33	0,89	4,91	3,38	2,94	0,00	77,73	10,15	17,10	2,95	62,65	7,04	4,28	0,83	9,85	▨	1,79
	Both	2,75	10,63	0,51	3,42	3,02	4,40	2,32	0,00	49,34	11,33	19,85	11,65	42,46	11,06	5,67	1,57	9,26	▨	2,49
	Can	15,07	17,34	0,05	2,15	9,42	8,06	4,50	3,68	0,27	34,15	60,75	30,15	199,20	15,48	10,23	6,56	25,41	▨	6,77
	PPI	-	-	-	-	-	-	-	-	-	-	-	-	+	-	-	-	+	▨	-
FLM	nCan	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	▨
	Both	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	▨
	Can	54,30	1,38	0,16	0,00	38,22	15,35	0,06	35,95	0,93	3,50	4,78	0,06	2,65	12,44	11,69	122,68	1,95	5,56	▨
	PPI	-	-	-	-	-	+	-	+	-	-	-	-	-	+	-	-	-	-	▨

Figure 4.10: **TF cobinding matrix based on merged regions.** *For each TF comparison, -10log(p-value) is given for the genomewide co-occurrence in the merged regions, split per type of region based on the presence of non-canonical and canonical motif instances. nCan = peaks with only non-canonical motif instances; Both = peaks with both canonical and non- canonical motif instances; Can = peaks with only canonical motif instances; PPI: + means a PPI is known between the TFs.*

and moderately predictive of initiating Pol II recruitment, but not of transcript abundance[231]; (ii) highly expressed loci are very amenable to ChIP in yeast, leading to HOT regions[204], and (iii) DNA motifs appear to be of less importance for TF binding in human HOT regions.[232] To assess whether HOT represent functional regulatory elements in plants, we investigated the expression of HOT-associated genes, together with purifying selection patterns, chromatin states, and DNA motifs in HOT regions.

Firstly, we found that for most TFs, there is no indication that genes associated with HOT regions are less prone to be responsive upon perturbation of the profiled TF than non-HOT-associated genes. These results differ from those in C. elegans modENCODE, where it has been suggested that HOT-associated genes are less prone to be regulated by the binding TFs. Instead, HOT-associated genes tend to be ubiquitously expressed[205], which is not the case for the plant HOT-associated genes delineated here. However, it should be noted that Van Nostrand and Kim (2013) inferred this pattern for only two TFs, raising the question whether this finding represents a global trend that is valid for other TFs as well. Secondly, the percentage of peaks, as well as the distribution of canonical and non-canonical motifs, harbouring a motif instance is similar in HOT regions and non-HOT regions, revealing that sequence-specific TF binding is prevalent in HOT regions as well. This is again in contrast with results found in humans, where the ENCODE project concluded that open chromatin facilitated TF binding in HOT regions even in the absence of specific binding motifs for the particular TF examined.[232] Through the integration of genome-wide chromatin states, we explored whether different types of bound regions are enriched for specific states, which could indicate functional differences. Overall, we observed that both HOT and non-HOT regions are strongly enriched for states describing proximal and distal promoters, as well as transcription start sites, and are depleted for heterochromatin. Furthermore, based on nucleotide diversity data from 369 re-sequenced Arabidopsis strains, we found that bound regions, both HOT and non-HOT, show strong signatures of purifying selection. Combining these different results, we therefore concluded that the binding events occurring in Arabidopsis HOT regions are functional and are mediated by specific DNA binding motifs, and are not merely the result of increased accessibility due to an open chromatin configuration.

While we have shown that HOT regions are indicative of functional binding, one of the consistent observations in genome-wide ChIP experiments is poor correlation between binding, DNA motif presence, and transcriptional response for candidate target genes. Possible explanations are the incorrect assignment of a binding site to a potential target gene, functional redundancy among related TFs, conditional differences between ChIP and transcript profiling (different cell-type, developmental stage, or physiological condition), or an incompatible chromatin state.[144] Additional hypotheses are that there is a transcriptional response following the binding event, but the mRNA is immediately degraded, or that the binding merely facilitates binding of co-factors essential for activation or repression of the targets.[13] A last explanation is that transcript-profiling studies in part capture indirect regulation. With respect to the DNA motif presence in ChIP peak sequences, we have shown that when taking into account significantly enriched non-canonical or non-primary motifs, the fraction of peaks with a motif instance substantially increased. Furthermore, we observed for some TFs that the most frequent motif does not match the canonical motif, which is consistent with the ENCODE results.[212] Importantly, the potential target gene sets associated with canonical and non-canonical motifs are similarly enriched for DE genes, implying that both types of motifs mediate TF regulation.

Based on the non-canonical motifs and the TF co-occupancy at merged regions, we have inferred co-binding events that are significantly more frequent compared to what would be expected by chance. For example, the different motifs matching CArG boxes in one MADS domain TF ChIP profiling study can be explained by the extensive co-binding among MADS domain family members.[229] Furthermore, the G-boxes found enriched in regions bound the AP1 MADS domain TF can be explained by co-binding of PIF3, PIF5, PRR5, and PRR7. Similarly, we could correlate the significant enrichment of a non-canonical FBS motif in the peaks of PRR7 to the co-binding with FHY3. Because these motifs and co-binding is most strongly enriched in peaks with only non-canonical motifs, we hypothesise that these binding events occur through tethering.[212] Whereas it has recently been shown, based on in vitro in protein binding microarrays[233], that some plant TFs can bind different DNA sequences, based on our co-binding observations, we conclude that the non-canonical DNA motifs can for the most part be explained as the result of cooperative TFs binding the same region.

In conclusion, the integration of different experimental ChIP datasets has revealed a number of insights regarding the organisation of binding events on a genome-wide scale. In addition, we showed that bound regions show a clear signal of purifying selection based on a population diversity, as well as conservation analysis. Finally, we provide testable hypotheses for the cooperative regulation of TFs through tethering based on the integration of DNA motif information for the different binding events.

4.4 Material and Methods

ChIP-Seq processing

Raw reads were downloaded from the NCBI Sequence Read Archive (SRA, Wheeler et al.[234]; accession IDs listed in). The quality of the raw data was evaluated with FASTQC (v0.10.0; `http://www.bioinformatics.bbsrc.ac.uk/projects/fastqc`), and adaptors and other overrepresented sequences were removed using the fastx-toolkit (v0.0.13; `http://hannonlab.cshl.edu/fastx_toolkit/`). The reads were mapped to the unmasked TAIR10 reference genome of Arabidopsis (TAIR10_chr_all.fas; `ftp.arabidopsis.org`) using BWA with default settings for all parameters (v0.5.9; Li et al.[119]). Reads that could not be assigned to a unique position in the genome were removed using samtools (v0.1.18; Li et al.[119]) by setting the mapping quality threshold (-q) to 1. Redundant reads were removed, retaining only one read per start position, using Picard tools (v1.56; `http://picard.sourceforge.net`). Peak calling was performed using MACS (v2.0.10, Zhang et al.[52]; default parameters except -g 1.0e8 and FDR $\leq$ 0.05). When replicates were available, the Pearson correlation coefficient (PCC) between the peak FPKM (fragment per kilobase per million) values was calculated for all peak-called regions across the different replicates (Figure B.15). Since most ChIP-Seq studies were performed without biological replication, the analysis was continued with the better replicate, with the choice of replicate being based on the results of the motif enrichment under the peaks (see Methods on Peak Calling). A few of the older experiments (SRP002328, SRP003928, and SRP000783) had lower PCC values between replicates than recent studies because of lower consistency in quality. Both for experiments with high and low PCC values between replicates, the replicate with better motif enrichment was retained (see Methods on Motif Finding further). An overview of which replicates were used for the samples is provided in B.4. For EIN3, the time point at which the maximal number of binding events occurred (4h) was processed.[235] REV, AMS, and FLP/MYB88 were removed from the data set due to a very low number of peaks in the results, the lack of paired-end read processing in the computational pipeline, and an abnormally high fraction of peak regions near transposable elements (Figure B.1A), respectively. All experiments were visually inspected with GenomeView.[54]

ChIP-chip processing

Raw CEL files were downloaded from Gene Expression Omnibus (GEO, Barrett et al.[111]; accession IDs are listed in section 4.4). The Affymetrix Tiling array bpmap files were updated to the current TAIR10 annotation with Starr.[122] Normalisation and peak Calling was performed with the Bioconductor[236] package rMAT[121] in R (R Core Team, 2012). The PairBinned method was used to normalise the arrays and peaks were called using a FDR cutoff of 0.05 except for the data sets GSE13090, GSE24684, GSE43291, and GSE40519, in which the p-value was set of 10^{-3} (in analogy to the original study, and necessary to obtain peak calling results). The minimum requirement of consecutive enriched probes was set at of eight. Other parameters were left at their default setting. All replicates were taken into account by the rMAT algorithm.

Peak annotation

Peak regions were annotated based on the location of their summits. A peak was assigned to the closest gene as annotated in the TAIR10 release represented in the PLAZA2.5 database[180]; peaks can be assigned both 5' and 3' of a gene. Each assignment is considered as a potential TF-target interaction. The peak locations were categorized by assigning a peak to one of the following genomic regions: intergenic, 1 kb promoter (1 kb upstream of Transcription Start Site), 5' UTR, coding, intron, 3' UTR, and 1 kb down of the Transcription stop site. For Figure B.2, the assignment based on the entire peak regions shows the average fraction of the peak lengths assigned to each genomic region.

Motif finding

The sequences of the complete peak regions were masked for coding sequence and submitted to the Peak-Motifs algorithm using default settings.[118] Motifs that could be aligned with a correlation score $\geq 75\%$ were collapsed. For each returned DNA motif, enrichment was defined as the ratio of the peak set frequency over the frequency in 1,000 random sets of peaks of the same size and length distribution sampled without replacement from the complete non-coding genome space (intergenic + UTR). The motifs from Peak-Motifs were mapped using matrix-scan[237] using the same parameters as used by Peak-Motifs. To determine whether a motif corresponded with a TF's canonical DNA motif, de novo motifs were compared with known motifs from the AGRIS database[17] using the STAMP web tool with default settings.[120]

Population genomic analyses

Single nucleotide polymorphism (SNP) data was downloaded from the 1001 Genomes project[c] on April 10 2014. Positions were only taken into account when they were sequenced in 70% of the strains. π values[223] were calculated per site using VCFtools[123] and recalculated into region π values for the different genomic data sets used. For the large intergenic regions (complete, 1kb up, and 1 kb down), the regions with information in less than 70% of the accessions were discarded. For the other (smaller) genomic elements, it was required that they were covered completely by regions with 70% information. The significance of the difference in π for different regions was determined by shuffling the bound regions across the Arabidopsis intergenic space 1000 times using BEDTools[124] and its python extension Pybedtools.[238] The p-value was empirically determined by counting the number of iterations in which the overlap was larger in the reshuffled than in the real data set.

Integrated functional data sets

Protein-protein interaction data was taken from the CORNET database[239], excluding the EVEX and AraNet relations. The functional modules were taken from our previous study.[134] Phosphorylation data was downloaded from PhosPhAt on March 24 2013.[110] Only those interactions were taken into account that describe a verified relationship between the kinase and the target protein itself: protein regulation, activation/inactivation, phosphorylation, dephosphorylation, and autophosphorylation. The miRNA target data was extracted from the Supplemental table 1 of Bulow et al.[219]. MicroRNA-target relations were filtered for psRNATarget[113] expectation scores lower or equal to 3. DH sites (flowering and leaf tissue) were from Zhang et al.[216]. DE data was obtained from the publications as listed in Supplemental Table 1. Genes were removed when they were present as being up —and downregulated upon perturbation of a TF, because of different time points and conditions. The GO and MapMan gene annotations were downloaded on May 15 2013. Enrichment of a functional category in a set of genes was calculated as the ratio of the set frequency over the genome-wide frequency. All functional enrichment values (GO[105], MapMan[112], functional modules[134], and DE) were validated statistically using the hypergeometric distribution and adjusted using FDR correction for multiple hypotheses testing.[199] The significance level was set at 0.05. For DE enrichment, the potential target genes were filtered for those present on the ATH1 microarray.

Hub targets and HOT regions

Target Hub genes were identified as described by Shalgi et al.[230]. For TFs that were profiled by both ChIP-chip and ChIP-Seq, only one of the experiments was taken into account. Hub genes are targeted by more TFs than the 99th percentile of the maximal value in 1,000 randomizations of the columns in the TF

[c] `http://1001genomes.org/projects/MPICWang2013/`

to gene matrix. The TF-target randomisation preserved the number of targets for each TF but reassigned each link. Following this procedure, target hubs are genes that are targeted by $\geq$ 8 TFs. For the ME and HC networks, the cut off values for hub genes were $\geq$ 7 and $\geq$ 6, respectively. For the determination of the HOT regions, all peak regions of all 27 TF data sets were merged after pruning long peak regions to the median length of all peak regions (470 bp, Figure B.2). Gene regulatory complexity was defined as the numbers of TFs that bind to peak regions assigned to a specific gene through peak annotation. The HOT regions were determined using the same strategy as the target hubs, being bound by $\geq$ 7 TFs. For the ME and HC networks, the cut off values was $\geq$ 6.

Enrichment analysis of bound regions in different genomic regions

The DH sites in flower and leaf were downloaded from NCBI SRA database (accession ID SRP009678.[216] The chromatin states were downloaded from supplemental dataset 2 from Sequeira-Mendes et al.[93]. The CNS data in dicots and crucifers was taken from Van de Velde et al.[225] and Haudry et al.[224], respectively. The HOT and non-HOT bound region files of each TF were formatted as BED files. Overlap analysis was performed using the BEDTools function intersectBed. For DH sites and chromatin states, the observed presence was determined with -u parameter and the -f parameter set to 0.5.[124] Because of the very long CNS regions in the crucifer data set, the overlap requirement was set to 50bp. In contrast, the dicot CNSs are very short since they resemble actual binding sites and here, CNSs were required to be completely embedded in bound regions. The expected presence in bound regions was determined by shuffling the DH sites data set 1000 times using shuffleBed, excluding the actual positions of the real instances. The overlap was determined using the same parameters for each shuffled file and the median number of shared elements present over 1,000 shuffled files was used as a measure for the expected presence. This was used to calculate enrichment as the ratio between observed presence and expected presence.

TF co-regulation and co-binding

For the co-regulatory matrix, the TFs were clustered based on the Jaccard distance (1 - Jaccard Index) between their target sets using average linkage hierarchical clustering. The overlap was validated statistically using the hypergeometric p-value, with Bonferroni correction for multiple hypothesis testing. The cut-off for significance was set at 0.001.

The co-binding statistics per type of peak (based on the presence of canonical and non-canonical motifs) were generated per query TF. For each query TF, the entire peak set was divided into the different categories of peaks (only canonical, both canonical and non-canonical, and non-canonical). Based on the merged regions to which each peak is associated, the number of times each other TF binds in the same merged region was counted. The p-value for this overlap (number of merged regions in which they co-bind) given the total set of merged regions, the set of merged regions associated with the query TF (split per type), and the set of merged regions associated with the co-binding TF was calculated with the hypergeometric distribution.

Expression values and condition specificity

Expression values were determined based on the filtered microarray compendium 2 from the CORNET database.[239] For condition-specificity, a gene was considered expressed if the log2 expression value was above 7.5. The Kolmogorov-Smirnov test was executed using Scipy.[240]

Accession numbers

NCBI SRA and Gene Expression Omnibus accession IDs are as follows: FLPMYB88, GSE19763; AGL15, GSE17717; GL3, GSE13090; GL1, GSE13090; AP2, E_MEXP_2653, SRP002328; SEP3, GSE14635, SRP000783; WUS, E_MEXP_2499; SMZ, E_MEXP_2068; BES1, GSE24684; SOC1, GSE33297, SRP020612; SVP, GSE33297; LFY, GSE28063, SRP003928; FUS, GSE43291; GTL1, GSE40519; AMS, SRP002566; AP1, SRP002174; FHY3, SRP007485; REV, SRP006211; PIF4, SRP010570; PIF5, SRP010315; FLC, SRP005412; TOC1, SRP010999; PRR5, SRP011389; AP3, SRP013458; PI, SRP013458; ERF115, GSE48793; PIF3, SRP014179; PRR7, SRP028272; FLM, SRP026163; EIN3, SRP017902; DH sites, SRP009678.

Acknowledgements

This work was supported by the Multidisciplinary Research Partnership "Bioinformatics: from nucleotides to networks" Project (no 01MR0310W) of Ghent University (K.V.), the Agency for Innovation by Science and Technology (IWT) in Flanders (K.S.H. and J.V.D.V.), an FWO travel grant (K.S.H), and the Max Planck Society (C.W. and D.W.).

CHAPTER 5

Inference of transcriptional networks in *Arabidopsis thaliana* through conserved non-coding sequence analysis[a]

Abstract

Transcriptional regulation plays an important role in establishing gene expression profiles during development or in response to (a)biotic stimuli. Transcription factor binding sites (TFBS) are the functional elements that determine transcriptional activity and the identification of individual TFBS in genome sequences is a major goal to inferring regulatory networks. We have developed a phylogenetic footprinting approach for the identification of conserved non-coding sequences (CNSs) across 12 dicot plants. Whereas both alignment and non-alignment-based techniques were applied to identify functional motifs in a multi-species context, our method accounts for incomplete motif conservation as well as high sequence divergence between related species. We identified 69,361 footprints associated with 17,895 genes. Through the integration of known TFBS obtained from literature and experimental studies, we used the CNSs to compile a gene regulatory network containing 40,758 interactions, of which two-thirds act through binding events located in DNase I hypersensitive sites. This network shows significant enrichment towards in vivo targets of known regulators and its overall quality was confirmed using five different biological validation metrics. Finally, through the integration of detailed expression and function information, we demonstrate how static CNSs can be converted into condition-dependent regulatory networks, offering new opportunities for regulatory gene annotation.

[a] This chapter is based on Van de Velde et al. [225]. J.V.D.V, K.S.H. and K.V. designed the research methodology and wrote the manuscript. J.V.D.V (CNS pipeline implementation, evaluation, condition-specific networks, and all figures,), K.S.H. (CNS pipeline design and implementation, ChIP pipeline and data collection), and K.V. (condition-specific networks) performed data analysis.

5.1 Introduction

Transcriptional regulation is a complex and dynamic process in which transcription factors (TFs) play a fundamental role. Although being subject to many potentially overlapping control mechanisms, such as miRNA regulation and chromatin accessibility coordinated by histone modifications and DNA methylation, the binding of TFs on specific genomic locations modulating gene expression levels is pivotal for the proper control of different biological processes. TF binding events can have a direct or indirect effect on the activation or repression of gene transcription. More complex regulation of gene expression is achieved through cooperative binding of different TFs adding an extra combinatorial level of control.[15] These regulatory mechanisms allow organisms to process different endogenous signals related to growth and development and to respond to changing environmental conditions including different types of (a)biotic stresses.

Despite the functional importance of transcriptional regulation and the fact that 1500-1700 TFs have been identified in Arabidopsis thaliana[15,18], knowledge about the genes controlled by different TFs is still very limited. AtRegNet, which is a part of the AGRIS database[241], summarizes regulatory interactions collected from small and large-scale experiments and contains 728 interactions when filtering on direct and confirmed targets. This paucity of experimentally validated regulatory interactions can be partially explained by the fact that previously used methods like electrophoretic mobility shift assay[242], systematic evolution of ligands by exponential enrichment[243] and Yeast-one-hybrid[244] are labour-intensive and only yield a small number of interactions.[214] More recent techniques such as protein binding microarrays, chromatin immunoprecipitation (ChIP) with readout through microarray (ChIP-chip) or next-generation sequencing (ChIP-Seq), allow TF protein-DNA binding to be analyzed in a high-throughput manner. However published binding results using these methods have revealed a weak correlation between the binding of a TF and transcriptional regulation of the potential target genes.[245]

Dozens of software tools have been developed to delineate regulatory regions based on experimental features, such as co-regulation, or using advanced computational methods.[73] Although the naÃŕve mapping of known DNA sequence motifs to promoter regions is frequently used to explore cis-regulatory elements, this approach yields many false positives because TF binding sites are often short and typically contain some level of degeneracy in the binding motif.[152] Although experimentally characterized open chromatin regions, profiled through DNase I hypersensitive (DH) sites, offer a global picture of accessible regions throughout the genome and can aid in reducing the motif search space[216], determining individual TF binding events remains a major challenge. A promising solution for the computational detection of functional elements is phylogenetic footprinting, which identifies conservation in orthologous genomic sequences.[75,246] Orthologs are homologous genes derived from a speciation event in the last common ancestor of the compared species. Regions of non-coding DNA in the genome that are conserved across related species are likely to be under purifying selection and this signature can be seen as evidence for functionality.[42,148,150,247–251] Overall, it is not trivial to make the distinction between conserved non-coding sequences (CNSs) that have arisen due to neutral sequence carry-over and functionally constrained CNSs in closely related species. With the advent of methods such as PhastCons[127], which make use of aligned genomes and statistical models of sequence evolution, it has become possible to determine CNSs in closely related species. These methods have shown greater power in the detection of functional elements and lineage-specific conservation than detection methods based on comparing more distantly related genomes in vertebrates, insects, worm and yeast.[127] However, these approaches require aligned genomes and the fraction of the genome that can be aligned drops drastically ($\leq 40\%$) when comparing species from different genera in flowering plants.[252] This is due to large-scale genome rearrangements and high sequence divergence. Furthermore, taxon sampling is still limited for flowering plants with the exception of the Brassicaceae lineage. These factors make global alignment strategies for the detection of CNSs impractical for many of the currently available plant genomes.[253] An additional difficulty for phylogenetic footprinting in plants lays in the fact that it is not trivial to identify one-to-one orthology in plants,, due to a wealth of paralogs (homologous genes created through a duplication event) in almost all plant lineages.[180] Besides continuous duplication events, for instance via tandem duplication, many plant paralogs are remnants of whole genome duplications (WGDs). In flowering plants, the frequent WGDs in several lineages result in the establishment of one-to-many and many-to-many orthologs (or co-orthologs). As a consequence, methods for identifying CNSs that were successfully applied in yeast or vertebrates don't work well in plants, as these methods cannot cope with complex orthology relationships.[148,254]

Recently three approaches to identify genome-wide CNSs using multiple plant genomes have been published. Baxter and co-workers used a local pairwise alignment approach, implemented in the Seaweed alignment plot tool[131], to search for CNSs in the 2kb upstream of the transcription start site in Arabidopsis.[251] Pairwise alignments were generated between orthologous genes of Arabidopsis and three highly diverged dicots: Papaya, Poplar and Grapevine (Carica papaya, Populus trichocarpa and Vitis vinifera). The conservation scores associated with each pairwise alignment were aggregated while orthologs were delineated using a combination of synteny and reciprocal best BLAST hits. Haudry et al.[224] generated a whole genome alignment approach using a combination of the LASTZ (Harris, 2007) and MULTIZ[125] tools across nine closely related Brassicaceae species. In this study a genomic region was aligned with one or multiple regions in another species as a means to cope with polyploidy. Conservation in the aligned regions was determined using PhyloP[126] yielding a set of 95,142 Arabidopsis CNSs. Similarly, Hupalo and Kern (2013) created a whole genome alignment between 20 closely and distantly related angiosperm genomes by making use of the LASTZ tool, and used PhastCons[127] to identify sequence constraint.

To generate a comprehensive overview of cis-regulatory elements in the Arabidopsis genome, we developed a phylogenetic footprinting framework that identifies CNSs between 12 distantly related genomes. Through the integration of information about known transcription factor binding sites (TFBS), gene expression profiles, open chromatin states and different gene function annotations, the static CNSs were annotated and translated into a gene regulatory network capturing known and condition-specific regulatory interactions. In addition, we confirm using different experimental datasets and biological validation metrics the quality of the inferred network.

5.2 Results

Detection of CNSs using a multi-species footprinting approach

We used a comparative genomics approach across 12 dicot plants to discover CNSs in Arabidopsis. A computational framework was developed that uses the mapping of known motifs as well as de novo local alignments to identify regulatory motifs conserved in multiple species. A local alignment-based approach between orthologous regions was applied because global alignment strategies are impractical for many of the currently available plant genomes due to massive loss of synteny conservation (Figure C.1). The selected comparator dicot species used in this study are reported in Figure C.1. The first method, called Comparative Motif Mapping (CMM), requires a candidate motif (e.g. a transcription factor binding site represented as a consensus sequence or position count matrix) as input, and assesses the motif conservation on, for example, the 2kb promoter of an Arabidopsis gene. Conservation is scored based on the occurrence of the motif in the promoter regions of the orthologs from the query gene in 11 other species, allowing for incomplete motif conservation. The statistical significance of a motif conserved in a set of orthologous genes is determined by comparing the observed conservation score to a background model that is built from conservation scores generated by processing the same motif on a large number of randomly assembled non-orthologous families, containing the same species composition and having the same sequence length distribution as in the real set of orthologs (see Methods). Based on the phylogenetic footprinting principle, the assumption behind this statistical model is that conservation of functional motifs will be higher between orthologous genes than between randomly chosen non-orthologous genes. As orthologous genes between Arabidopsis and all other comparator species show saturated substitution patterns (the fraction of synonymous substitutions per synonymous site, Ks ≥ 1, see Methods), the identified CNSs show selective constraint indicating biological functionality.

The second method is alignment-based and uses a multi-species scoring approach to detect CNSs, without requiring prior motif information. All footprints extracted from pairwise local alignments between the query gene and its orthologs are collapsed onto the corresponding region of the query gene. As such, the number of species that supports each nucleotide through a pairwise alignment is determined. In the next step, conserved footprints are extracted and scored based on the number of species in which they are conserved. Significant footprints are determined using a pre-computed background model built with scores of footprints derived from non-orthologous families to which each real footprint is compared. The same assumption regarding higher functional sequence conservation between orthologous genes than between randomly chosen genes is made. For the alignment-based approach four alignment tools were implemented in the framework and their performance was compared. These tools were DIALIGN-TX[128], Sigma[129], ACANA[130] and the Seaweeds alignment plot tool.[131] The proposed methods are able to cope with high sequence divergence when aligning non-coding sequences between related species. As many motif and alignment comparisons are being made for thousands of genes, the false discovery rate (FDR) was estimated by comparing the significant results of the real runs with those of control runs. The FDR is defined as the ratio between the number of false positives estimated by the control run and the number of rejected null hypotheses in the real run, and provides a better measure for controlling false positives compared to the false positive rate, as the latter does not correct for the multiple tests performed per query gene. Control runs are identical to real runs with the exception that the orthologous families are randomly generated, maintaining the species constitution and gene size as observed in the real families (see Methods). Unless mentioned otherwise, all presented results have an FDR below 10%.

After updating the TAIR10 genome annotation with 791 new miRNA loci obtained from the plant microRNA database (PMRD)[114], three different genomic sequence types were defined to identify CNSs (2kb upstream, 1kb downstream and intron). In this analysis upstream and downstream are used relative to the translation start site and translation stop site, respectively, because it has been shown, both through promoter deletion experiments as well as using genome-wide ChIP analyses, that regulatory elements can be found in 5' and 3' untranslated region (UTR).[255–257] Another reason to include UTRs is that not all genes have information about their UTR available. In total, the different genomic sequences cover 83% of the non-coding Arabidopsis genome and 84% of all complete intergenics. Gene orthology information was retrieved from the PLAZA 2.5 integrative orthology method[180], which uses a combination of different detection methods to infer consensus orthology predictions, both for simple one-to-one as well as for more complex many-to-may gene relationships. Here, two different orthology definitions were used to delineate orthologs. The first definition uses a simple 'best BLAST hit'-derived method that includes inparalogs, called best-hit and in-paralogous families (BHIF), while the second definition, called consensus orthology, requires that at least two PLAZA detection methods confirm an orthologous gene relationship (see Methods). Orthologs could be obtained for 24,241 Arabidopsis genes using BHIF and for 21,300 genes using the consensus definition. For Arabidopsis genes with orthology information, 70% and 90% have orthologs in at least 10 species for the consensus and BHIF definition, respectively (Figure C.2).

Combining phylogenetic footprinting experiments from the alignment-based and CMM runs, we identified in total 69,361 significant CNSs associated with 17,895 genes. These conserved regions cover 1070 kb of the Arabidopsis genome and all CNSs are available through a genome browser (see Methods). The median length of a CNS was 11bp, while the largest and smallest CNS were 514bp and 5bp, respectively (Figure 5.1A). All of the significant CNSs were conserved in at least two comparator species while the median number of supporting species was six (Figure 5.1B). This result illustrates the strong multi-species nature and potential functionality of the identified CNSs. Analyzing the contribution of comparator species to footprints conserved in only two species showed no bias towards the most closely related comparator species. Half of the CNSs are located in the 1 kb promoter region of annotated genes and a large number of conserved regions were associated with introns (10,872) and downstream sequences (6953) (Figure 5.1C). The alignment-based and CMM detection methods detect 30% and 60% of all CNSs uniquely, respectively, while 10% is shared by both methods. CMM covers 473 kb and the alignment-based-approach covers 686 kb. The complementarity of the two different orthology definitions was evaluated by determining the uniquely detected CNSs and revealed that 70% of detected CNSs were found using both definitions. The consensus and BHIF definition detected 19% and 11% unique CNSs, respectively.

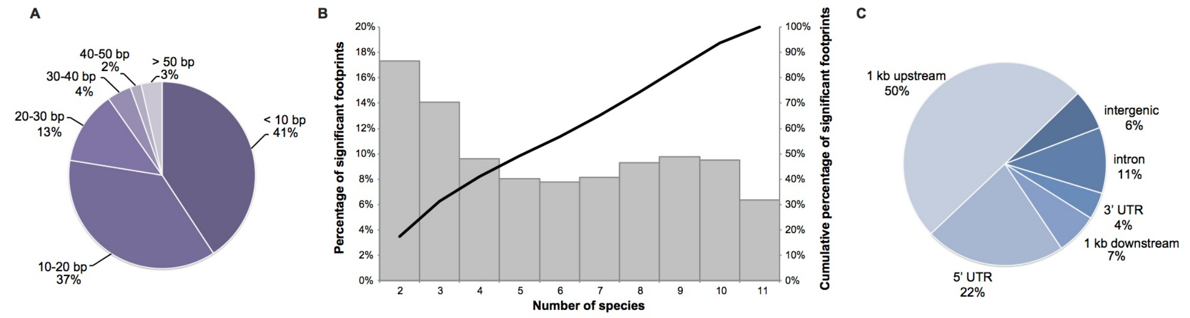


Figure 5.1: **Overview of CNS properties.** *A) Length distribution of significantly conserved footprints. All footprints are grouped in bins of size 10bp. B) Overview of significantly conserved footprints in relation to the number of species in which the footprint was conserved. For all conservation scores the relative percentage of significant footprints is shown (grey boxes) as well as a cumulative distribution (black line). C) Breakdown of CNS over different genomic regions.*

Besides regulatory elements, other structural features such as incorrectly annotated exons or missing genes may show significant conservation across related genomes. To determine whether any of the identified footprints represent coding features, we performed a sequence similarity search of all CNSs against a large set of known plant proteins (see Methods). Only 499 CNSs (0.01% of all footprints) showed a significant hit against the plant protein database and were discarded for downstream analysis.

Evaluation of different phylogenetic footprinting approaches using an experimental gold standard

In order to evaluate whether our footprints correspond with known regulatory sequences, we compared our CNSs against the AtProbe dataset, which contains 144 experimentally determined cis-regulatory elements (see Methods and Supplemental Online Data set 1[b]). Overall, our CNSs recovered 26% of the experimental binding sites. This global true positive rate (TPR) was analyzed in more detail per detection method (Figure C.3). Sigma, the best performing alignment tool, scores equally well compared to CMM as both methods have a TPR of 19%. This result indicates that Sigma, which finds conserved regions without any prior information, has sensitivity comparable to CMM, for which prior motif information is required. Additionally, these methods are complementary as they uniquely detected 22% and 16% of the recovered AtProbe elements, respectively. Whereas ACANA and Seaweeds-60 recovered experimental instances (TPR of 5% and 3%, respectively), DIALIGN-TX and Seaweeds-30 did not, which is due to the generation of spurious alignments yielding many false positives in the control runs.

To further validate our set of CNSs, we compared our results with three other CNS datasets from published genome-wide phylogenetic footprinting approaches (Figure 5.2).[224,251,252] Apart from evaluating the sensitivity of the different studies, which relates to finding true positive AtProbe results, we also assessed the specificity, which relates to identifying negative results. The latter is important, as a method that would assign each non-coding nucleotide to a CNS would yield a high sensitivity but a low specificity, due to many false positives. Although it is not trivial to assemble a negative dataset of genomic regions free from any regulatory sequence, we estimated false positives by reshuffling the AtProbe genomic locations 1000 times and determining the overlap with CNSs detected per footprinting study. The estimated number of false positives was used to determine enrichment for known regulatory elements (observed number of elements over expected number of elements, see Methods). This approach does not guarantee that the reshuffled dataset, which covers in essence randomly selected non-coding genomic regions that have no overlap with real AtProbe instances, contains only true negatives. However the reshuffled dataset can be used as a proxy to estimate the specificity of different footprinting studies as the same biases are present in the negative dataset for all methods.

Comparing the CNSs from the different studies showed that Haudry et al. (2013) has the highest recovery of experimental binding sites (35% TPR), followed by our results (26% TPR) and Baxter et al.

[b] http://www.plantcell.org/content/suppl/2014/06/16/tpc.114.127001.DC1/tpc127001_Supplemental_Datasets.xls

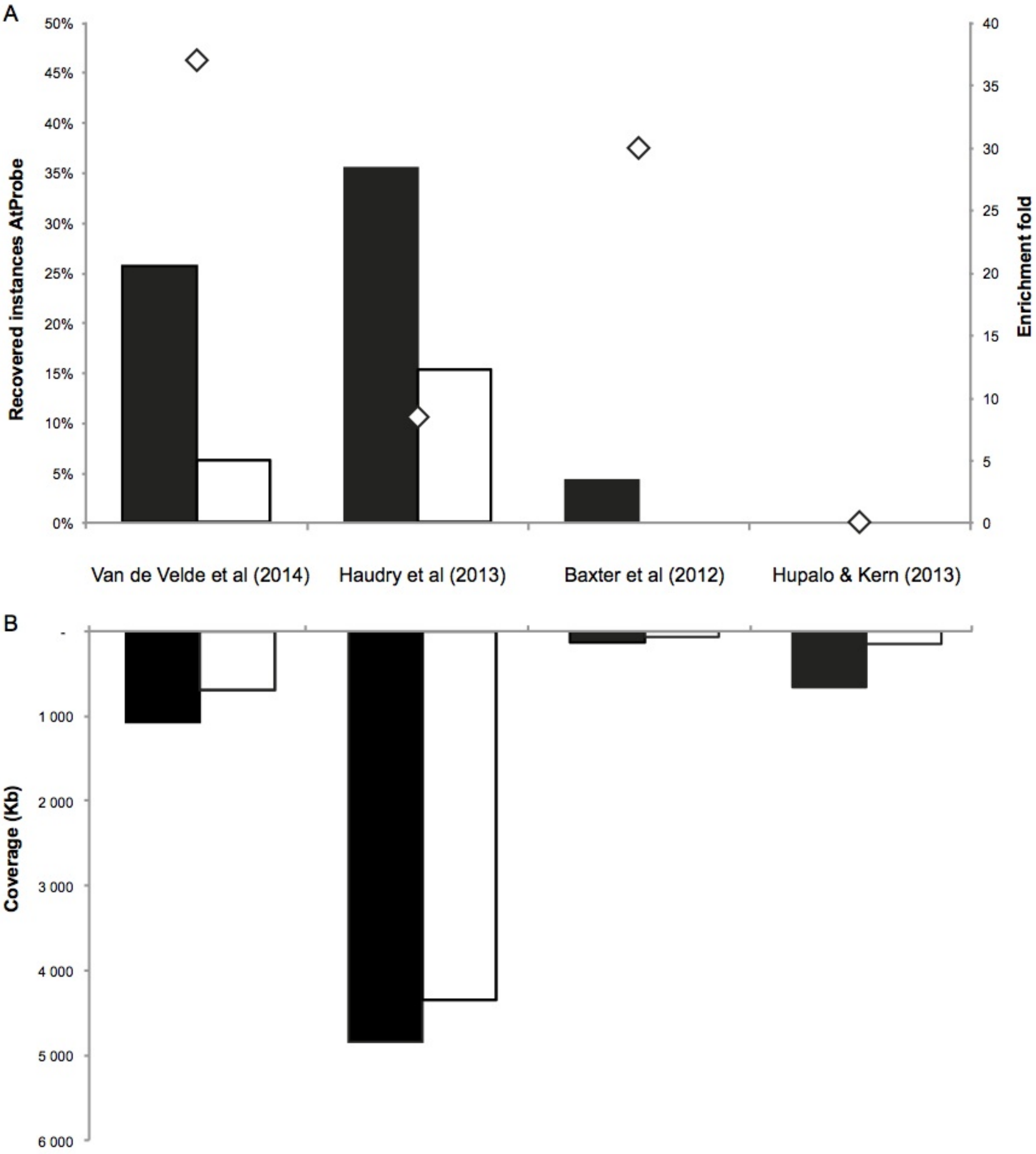


Figure 5.2: **Recovery of AtProbe elements and comparison of CNSs from different phylogenetic footprinting studies.** *(A) Overview of the recovery of experimental AtProbe elements in four different CNS studies. Black boxes show the percentage of recovered elements and white boxes shows the percentage of uniquely recovered elements. Diamonds depict fold enrichments, which are defined as the ratio of the observed overlap over the expected overlap by chance. (B) Genome-wide coverage of CNSs. Black boxes show the total number of nucleotides assigned to CNSs per study while white boxes show the number of nucleotides in CNSs that are unique to a single study.*

(2012) (4% TPR). An overview of retrieved CNSs for the AtProbe genes for this study and Haudry et al. (2013) can be found in Figure C.4. However, comparing the specificity using the shuffled AtProbe datasets reveals that Haudry et al. (2013) has a lower enrichment towards experimentally determined elements (8.5 fold enriched) than our approach (37 fold enriched) (Figure 5.2). Determining the genome-wide coverage for the different CNS datasets revealed that Haudry et al. (2013), identified constraint for 4,834 kb of non-coding DNA. This coverage is substantially larger than our dataset (1,070 kb) and those of Baxter et al. (2012) and Hupalo and Kern (2013), which cover 137 kb and 658 kb, respectively (Figure 5.2). Overall, our method, which we have shown to be accurate based on the analysis of known regulatory sites, identifies 64% of the nucleotides covered by our CNSs as evolutionary constrained which were not identified by other methods, indicating that our phylogenetic footprinting approach covers a large fraction of unique CNSs.

Conserved motif instances identify *in vivo* functional regions

To evaluate the functionality of the identified CNSs and to verify whether these conserved footprints can provide a template to computationally map TF-target interactions, detailed comparisons of the CNSs were made against different experimentally determined datasets. DH sites are associated with regions of open chromatin where the DNA is accessible and as such provide a global perspective on possible protein binding to the genome. Overall, 48% and 47% of our CNSs overlapped with a recently published set of DH sites in flower and leaf tissue, respectively.[216] This overlap is significant (p-value $\leq$ 0.001) and shows high fold enrichment (4.0 for both DH sets, see Methods), revealing that a large part of the CNSs can be accessed by TFs and as such can act as a functional TFBS. Our set of CNSs also exhibited a significant overlap with H3K4me3, H3K9ac and H3K4me2 marks (2.6, 2.2 and 1.7 fold enriched, respectively; Figure C.5). These histone modifications are indicative of active promoters and enhancer elements.[22,258] Interestingly, our regions showed an even higher enrichment for regions where DH sites, H3K4me3, H3K9ac and H3K4me2 coincide (6.3 fold enriched, p-value $\leq$ 0.001), corroborating that several of the conserved regions are associated with actively transcribed genes.

Whereas the experimental datasets profiling different chromatin states act as a proxy for functionality, more detailed regulatory information can be obtained by comparing the CNSs with experimental datasets comprising functional TFBS. To delineate a high-quality dataset of *in vivo* functional TF-targets covering directly regulated genes, publicly available ChIP-Seq data was combined with enriched motifs in ChIP-Seq peaks and TF-perturbation expression profiles (see Methods). This was done for 15 TFs (AGAMOUS-LIKE 15 (AGL15), APETALA1 (AP1), APETALA2 (AP2), APETALA3 (AP3), SUPPRESSOR OF OVEREXPRESSION OF CO 1 (SOC1), PISTILLATA (PI), LEAFY (LFY), FLOWERING LOCUS C (FLC), PSEUDO RESPONSE REGULATOR 5 (PRR5), PHYTOCHROME INTERACTING FACTOR 3 (PIF3), PHYTOCHROME INTERACTING FACTOR 4 (PIF4), PHYTOCHROME INTERACTING FACTOR 5 (PIF5), FAR-RED ELONGATED HYPOCOTYLS 3 (FHY3), BRI1-EMS-SUPPRESSOR 1 (BES1) and FUSCA 3 (FUS3)) yielding a dataset of 2807 regulatory interactions (Supplemental Online Data set 2[b]). Importantly, these *in vivo* functional targets were determined independently of any comparative information and thus provide an independent dataset to evaluate our footprints. Overlap analysis revealed that in total 787 functional binding sites (28%) were successfully recovered by our CNSs. Although the recovery rate for individual TF varies from 8% for AP3 to 57% for PRR5 (median recovery 36%), the number of recovered genes for all 15 TFs was significantly higher compared to the number of recovered target genes expected by chance ($p \leq 0.001$, see Supplemental Dataset 2[b] and Figure 5.3).

To compare the specificity by which our CNSs identified functional TFBS with other computational methods, two other protocols were evaluated. Whereas the first approach is based on the simple mapping of all positional count matrices of all 15 TFs on the non-coding genomic DNA, the second approach comprises motif mapping in open non-coding chromatin regions that were identified through DH sites.[216] Enrichment analysis using shuffled datasets of the *in vivo* functional regions (see Methods) revealed that our CNSs yielded higher specificity for functional regulatory elements than either of these alternative protocols (median fold enrichment of 41.2 for CNSs versus 2.6 and 12.8 fold enrichment for the simple and DH site-based mapping methods, respectively) (Figure 5.3, Supplemental Online Data set 3[b] and Figure C.6).

Construction and biological evaluation of an Arabidopsis gene regulatory network

To get an overview of how transcriptional regulation is organized on a genome-wide level, motif information was combined with our CNSs to construct a gene regulatory network (GRN) containing 40,758 interactions (see Methods). This GRN includes 157 TFs that, based on conserved binding sites, have

[b] `http://www.plantcell.org/content/suppl/2014/06/16/tpc.114.127001.DC1/tpc127001_Supplemental_Datasets.xls`

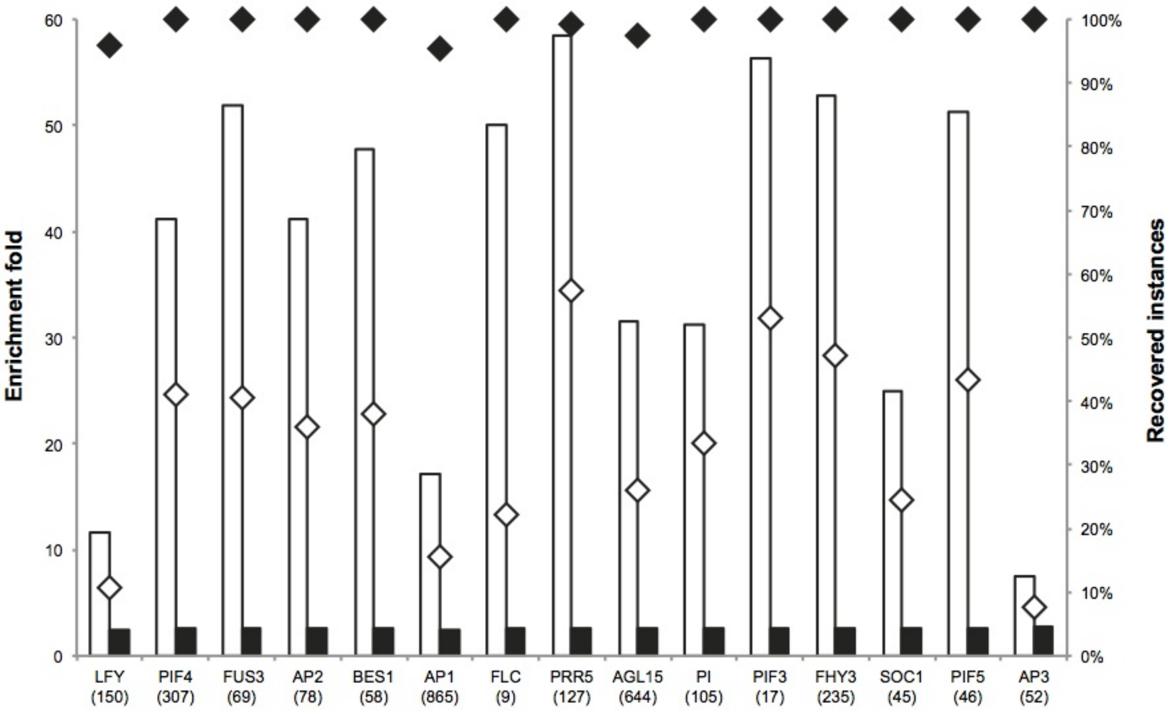


Figure 5.3: **Recovery of *in vivo* functional targets using CNS information.** *White and black boxes show fold enrichments for CNSs and naÃŕve motif mapping, respectively. White and black diamonds show the fraction of recovered elements for CNSs and a simple motif mapping approach, respectively.*

one or more target genes and covers 11,354 genes in total (Supplemental Online Data set 4[b]). On average, a TF in the predicted network has 259 target genes while each target gene is regulated by 4 TFs. The number of target genes per TF and their associated GO enrichment can be seen in Figure C.7. For these interactions, 64.6% of the conserved binding sites are overlapping with a leaf or flower DH site. To evaluate our network we used an experimental GRN of 1092 confirmed interactions derived from AtRegNet[259] and a collection of regulatory interactions obtained from small-scale studies concerning secondary cell wall metabolism.[260] Overlap analysis between the predicted network and the experimental network revealed that edges present in the predicted network are significantly more likely to also be present in the experimental network than would be expected by chance (4.65 fold enrichment, p-value $\leq$ 0.001; see Methods). Apart from comparing the global overlap between both networks, we also assessed the overlap between the predicted and experimental TF-target interactions for individual TFs for which motif information was available. For a sub-set of TFs with ten or more known target genes, a significant overlap was found for nine out of 13 TFs (p-value $\leq$ 0.001), which covers 99 out of 385 (26%) experimentally determined gene regulatory interactions.

To evaluate which role intronic regions have in transcriptional gene regulation through TF binding, an intron-specific GRN was generated. This network consists of 2821 interactions between 123 TFs and 1552 target genes. Six out of the 99 experimentally confirmed interactions that were retrieved were unique to this network (See Supplemental Online Data set 5). Examples of correctly inferred intron interactions are binding events of AP2 and LFY to the intron of AGAMOUS (AG).[261] Similarly, TF-miRNA regulation was studied by constructing a small sub-network containing 24 TF-miRNA targets for 14 TFs and 10 target miRNAs (Supplemental Online Data set 6). One of the retrieved interactions is the known binding of the ABRE binding factor (ABF1) to the promoter of mir168a.[215] Another interesting, however unconfirmed, interaction is that between AP2 and mir167a, the latter which is known to play a role in flowering maturation.[262]

In addition to the recovery of known regulatory interactions, the biological relevance of the predicted target genes was studied using five independent biological datasets. Gene Ontology (GO)[105], Mapman[112] and functional gene modules[134] describe functional annotations and were used to assess if target genes of the same TF participate in similar biological processes or have similar functions. The functional modules comprise a set of 13,142 genes (1562 modules) annotated with specific functional descriptions based on experimental GO information, protein-protein interaction data, protein-DNA interactions or AraNet gene function predictions. The evaluation of our GRN is made based on the assumption that a set of true target genes of a TF will have a higher enrichment for functional annotations than randomized networks.[213] For each TF, the enriched functional annotations were determined and compared against that of randomized networks (see Methods). Next to the three functional datasets, two general gene expression compendia were used, stress and development[239], to investigate if genes targeted by the same TFs (called co-regulated targets) are more likely to be expressed at similar developmental stages or under similar stress conditions. Following Marbach et al. (2012), co-regulated gene pairs are defined as genes having 50% or more shared regulators. The average level of co-expression was calculated using correlation analysis for all co-regulated gene pairs and compared to that of randomized networks (see Methods). All five biological metrics were performed on the CNS-based GRN as well as on the experimental GRN and we observed that both networks were significantly enriched for all five biological datasets (p-value $\leq$ 0.05, Figure 5.4). A detailed comparison revealed that GO fold enrichment was higher in the predicted network. Although the opposite is true for both Mapman and the functional modules, there is still a significant enrichment in our predicted GRN, illustrating the functional coherence of the predicted target genes. The discrepancy between different functional annotation datasets can largely be explained by the fact that for GO annotations a filtering step using GO slim terms was performed in order to have sufficient annotations for all genes in the network. These terms are very broad and as such enrichment will be lower compared to the two other functional classification datasets. Based on the stress and development expression datasets, a higher level of co-expression was observed for co-regulated genes in the predicted and experimental GRN, compared to random GRNs (Figure 5.4). The CNS-based network outperformed the experimental network, as the fold enrichments were higher for the predicted GRN in both expression datasets. A similar evaluation was performed on two sub-sets of the predicted network, which were defined based on the number of species in which a regulatory interaction is conserved. The predicted network was divided into a highly (conservation CNS $\geq$6 species) and a moderately conserved (conservation CNS 2-6 species) sub-network. Both the highly and the moderately conserved sub-networks showed significant enrichment for co-expression and functional coherence, indicating that CNSs with support from a lower number of species are also biologically meaningful (Figure C.8).

Combining the CNS-based network with expression information to identify condition-specific gene regulatory interactions

To investigate the biological role of the predicted GRN, the static gene regulatory interactions were converted into condition-specific interactions through the integration of expression information. Co-expression was determined between a TF and each predicted target gene based on 11 expression compendia from the CORNET database[239], comprising gene expression profiles from microarray experiments performed for different organs (flower, leaf, root, seed), during development, under different treatments and stresses (hormone, biotic and abiotic stress) (sdee Methods). Co-expression between a TF and a predicted target gene can act as a proxy for regulation as both are frequently expressed in the same conditions.[263] 6957 Interactions between a TF and its predicted target genes showed significant co-expression in one or maximum three expression compendia (Supplemental Online Data set 7[b]). Examples of specific co-expression patterns of predicted TF-target interactions that are confirmed by experimentally confirmed target genes include interactions for MYB DOMAIN PROTEIN 58 (MYB58) under biotic stress, MYB DOMAIN PROTEIN 83 (MYB83) in leaf and for AP2 and ELONGATED HYPOCOTYL 5 (HY5) under abiotic and biotic stress. MYB DOMAIN PROTEIN 63 (MYB63) shows co-expression of target genes in five different compendia, including (a)biotic stress and hormone (Figure C.9). The following paragraphs highlight examples of condition-dependent GRNs.

Five secondary wall NAM-ATAF1/2-CUC2 (NAC) TFs were selected to illustrate how integrating co-expression information into the predicted GRN can be used for modelling of the transcriptional network in different conditions and plant organs. SECONDARY WALL-ASSOCIATED NAC DOMAIN 1 (SND1) is a master transcriptional regulator activating the developmental program of secondary cell

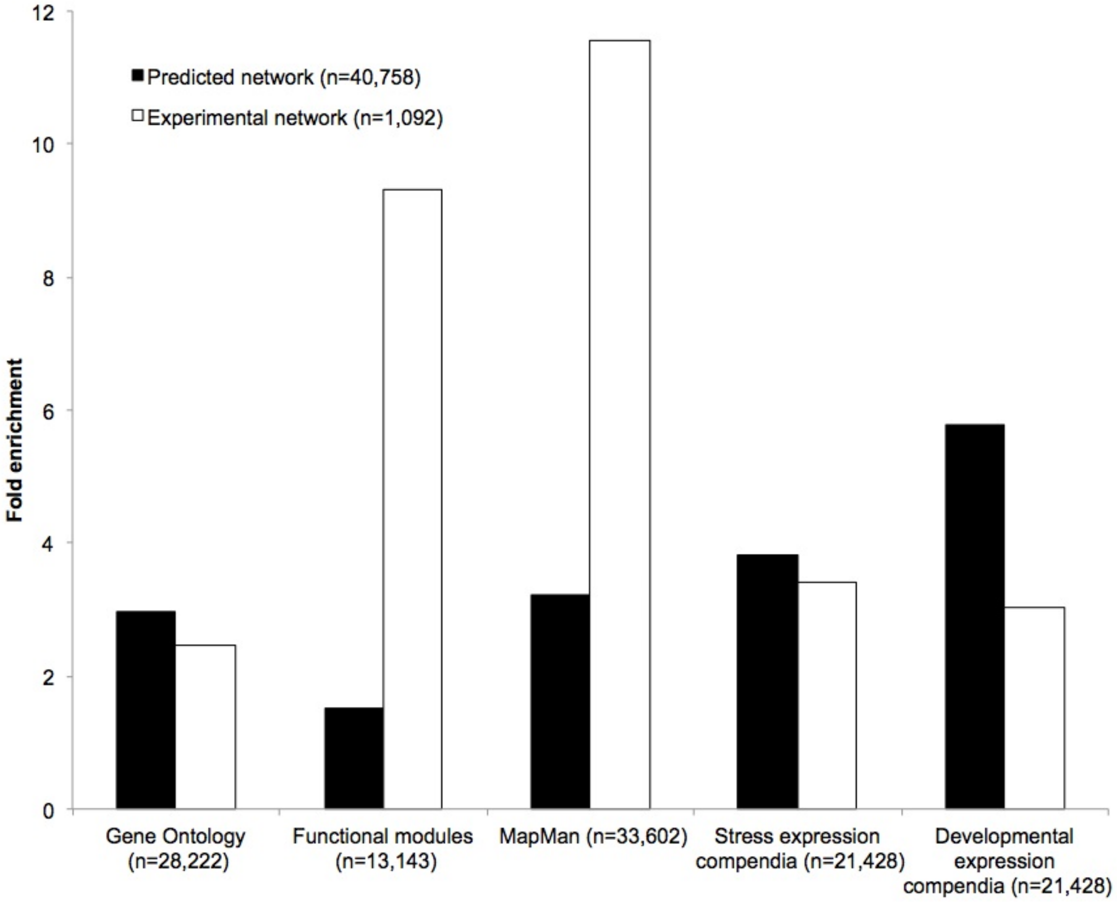


Figure 5.4: **Evaluation of the biological relevance of the predicted network using different biological metrics assessing functional and expression coherence.***Gene Ontology annotations, Mapman annotations and functional modules together with a stress and developmental expression compendium were used to evaluate the biological relevance of the predicted GRN. A comparison of fold enrichment is depicted between the predicted network (black bars) and the experimental network (white bars). All reported fold enrichments are significant (p-value $\leq$ 0.05). Numbers in parentheses report the number of regulatory interactions in the two networks and the number of genes having functional or expression information, respectively.*

wall (SCW) biosynthesis. SND1 and its functionally related homologs NAC SECONDARY WALL THICKENING PROMOTING FACTOR1 (NST1), NAC SECONDARY WALL THICKENING PROMOTING FACTOR2 (NST2), VASCULAR-RELATED NAC-DOMAIN 6 (VND6) and VASCULAR-RELATED NAC-DOMAIN 7 (VND7) regulate the same downstream targets in different cell types.[184] While SND1 and NST1 activate the SCW biosynthetic program in fibers, VND6 and VND7 specifically regulate SCW biosynthesis in vessels, and NST1 and NST2 act together in regulating SCW biosynthesis in endothecium of anthers (Mitsuda and Ohme-Takagi, 2008;.[184] These five TFs bind to an imperfect palindromic 19-bp consensus sequence designated as secondary cell wall NAC binding element, (T/A)NN(C/T)(T/C/G)TNNNNNNNA(A/C)GN(A/C/T)(A/T), in the promoters of their direct targets.[264] For VND6 an additional binding site has been described (CTTNAAAGCNA).[265] Based on the predicted targets of these 5 TFs, we used the co-expression information to introduce specificity through condition-dependent regulation. For SND1, NST1 and NST2 we studied target genes co-expressed in a flower and a seed expression compendium, because of their role in SCW biosynthesis in flower and reproductive organs (Mitsuda and Ohme-Takagi, 2008;[184] (Figure 5.5). Auxin, cytokinin, and brassinosteroids play pivotal roles in xylem vessel formation (Fukuda, 2004) and VND6 and VND7 show elevated expression levels in presence of these three hormones.[266] Both TFs reside in the same functional module, which is annotated with the GO term 'response to brassinosteroid stimulus'.[134] Therefore, VND6 and VND7 targets co-expressing in a hormone compendium were selected. For all TFs, predicted target genes were only selected if they were part of a functional module grouping two or more predicted target genes. This network groups 5 TFs showing 69 condition-specific interactions with 24 target genes (Figure 5.5). The SCW network contains a large number of experimentally confirmed interactions (14/69) and nearly all genes in the network are involved in SCW metabolism based on GO annotations (21/24). In this network, two TFs, namely MYB DOMAIN PROTEIN 46 (MYB46) and SECONDARY WALL-ASSOCIATED NAC DOMAIN PROTEIN 3 (SND3), which are known direct targets involved in the SCW pathway, are present. Interestingly, these genes do not have a co-expression link with SND1 in flower or seed, but a co-expression link is present with NST1, a TF that cooperates with SND1 in SCW biosynthesis in fibers.[184] Overexpression of MYB46 leads to activation of the entire SCW biosynthetic program and its co-expressing targets in seed, flower and hormone expression compendia show a large number of shared targets with the five master regulators as well as a large set of genes involved in SCW biosynthesis.[184]

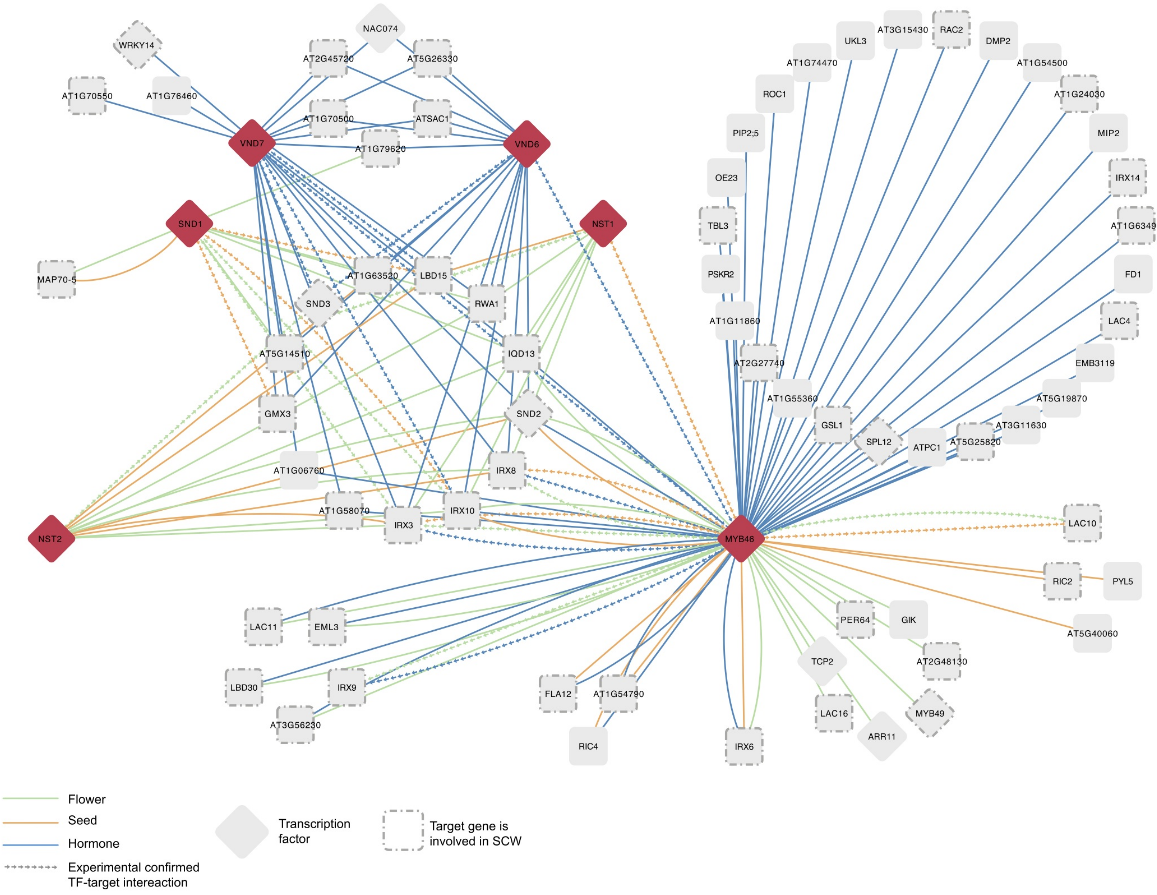


Figure 5.5: **A condition-specific secondary cell wall gene regulatory network.***Gene Ontology annotations, Mapman annotations and functional modules together with a stress and developmental expression compendium were used to evaluate the biological relevance of the predicted GRN. A comparison of fold enrichment is depicted between the predicted network (black bars) and the experimental network (white bars). All reported fold enrichments are significant (p-value $\leq$ 0.05). Numbers in parentheses report the number of regulatory interactions in the two networks and the number of genes having functional or expression information, respectively.*

A similar approach was applied to delineate condition-specific targets for AP3 and PI, two TFs that have been shown to act as bifunctional transcription factors in flower development.[267] AP3 and PI are necessary for the proper development of the petals and stamens.[268,269] Plant hormones such as jasmonic acid have been shown to play a role in both stamen and petal development.[270,271] The expression data for these two TFs shows induction in jasmonic acid treatment conditions. Therefore co-expressed target genes in the hormone expression compendium were selected. This approach resulted in a hormone-specific GRN with 223 target genes and 237 interactions. The network shows a strong enrichment for genes involved in flower development (53/223)(Figure C.10. Additional evidence for the relevance of this network was generated through integrating ChIP-Seq and differential gene expression data. The ChIP and differential expression experiments were performed at the early-intermediate floral stage (stage 4-5 flowers)[267]. In this network, we observe 11 interactions that are confirmed through binding of the TF in the ChIP-Seq data and also 6 interactions that are confirmed through differential expression of the gene after TF perturbation. Interestingly, AG is a predicted co-expressed target gene of AP3 in the hormone-specific network and AG has been shown to be involved in stamen development through regulation of jasmonic acid biosynthesis genes.[272]

5.3 Discussion

In this study we developed a new phylogenetic footprinting approach to identify conserved non-coding sequences in Arabidopsis through the comparison with 11 dicot genomes. Distantly related species were used because of the premise that, in comparison to one another, all non-coding regions that are not under functional constrained will have undergone one or more mutations. A set of 69,361 CNSs associated with 17,895 genes was delineated through the combination of an alignment-based and a non-alignment-based approach. Twenty-eight percent of the CNSs were found downstream of genes, in introns or more than 1kb upstream of a gene, indicating that regulatory elements are not restricted to the first hundreds of base pairs upstream of a gene.[253,273]

A previous evaluation study reported that phylogenetic footprinting in plants works best by comparing genomes that have diverged less than 100mya or have non-saturated substitution patterns.[253] Phylogenetic footprinting methods that use genome synteny inferred through genome alignments as primary source of orthology information indeed have difficulties integrating distantly related genomes.[252] This is due to the frequent nature of polyploidy and genome rearrangements in dicot plants (Figure C.1) causing problems for global genome alignment methods. Here, a combination of different gene orthology prediction methods was used that do not rely on synteny information. As such, our approach is well-suited to incorporate more distantly related species including many-to-many gene orthology relationships. Our alignment-based approach is best summarized as a multiple local alignment strategy, since first local pairwise alignments are identified which are subsequently aggregated on the Arabidopsis reference genome in order to obtain multi-species footprints. We demonstrated that this approach is very suitable for detecting CNSs over large phylogenetic distances, as half of our CNS are conserved in six or more species, spanning $\geq$100 million years of evolution (Figure 5.1B). Furthermore, approaches based exclusively on pairwise alignments lack the power to detect a large set of our CNSs over a similar evolutionary distance.[251,253]

Comparing our CNSs with the experimental AtProbe benchmark dataset showed that both alignment and non-alignment-based approaches have a similar performance, recovering 19% of the experimental regulatory elements. Both approaches are complementary as they together recovered 26% of the AtProbe elements. This is largely explained by the fact that the alignment-based approach identifies large conserved regions, typically covering clusters of individual TFBS, whereas the non-alignment-based approach will also identify short conserved motifs. Based on a comparison of our footprints with three recently published studies[224,251,252], 64% of our CNSs represent newly discovered constrained sequences. This finding is in agreement with Haudry et al. (2013) who found that their CNSs show limited conservation outside the Brassicaceae lineage. Compared to Baxter et al. (2012) and Hupalo and Kern (2013), both the number of comparator species as well as the different alignment strategy contribute to the difference in identified CNSs. Comparison with the three previously published CNS datasets revealed that our CNSs have the highest enrichment for experimentally determined regulatory elements. Haudry et al. (2013) recovered a larger number of bases covered by CNSs with a lower enrichment towards the AtProbe elements. Although these results could indicate that their higher coverage is associated with a reduced specificity, additional explanations can be formulated. As demonstrated by Haudry et al. (2013), their CNSs also contain other types of functional non-coding sequences, such as RNA genes, which are not accounted for in our benchmark. CNSs could also cover long-range enhancers. Also, the conservation of functional non-coding sequences is likely greater within the Brassicaceae lineage due to more specialized developmental processes and adaptation to environmental conditions, whereas our set of CNSs covers the regulation of processes that are highly conserved across a wide range of dicot plants. A subset of the AtProbe regulatory elements recovered was unique to this analysis, corroborating the complementarity of our CNSs with these previous studies.

The biological relevance of our CNSs was further evaluated by overlap analysis with a number of different chromatin modification marks. Enrichment analysis showed that our CNSs are highly enriched for DH sites as well as for histone marks promoting transcription indicating that our CNSs are located within open chromatin regions or nearby actively transcribed regions. Processing of 15 TF ChIP-chip/seq experiments together with the corresponding transcriptome profiling studies after TF perturbation generated a high-quality dataset of 2807 *in vivo* functional binding sites. In total 28% of these regions were successfully recovered. Mapping the position count matrices for all 15 TFs genome-wide and retaining only instances overlapping with a CNS, showed to be more specific to recover functional binding sites compared to filtering using DH sites. In contrast to simple motif mapping approaches which are associated with high false positive rates, computationally identified CNSs as well as experimental DH sites offer two complementary data sources to start performing systematic regulatory genome annotation in plants. The largest bottleneck for identifying all functional regions through conservation analysis is caused by the highly degenerative nature of certain binding sites, such as CArG boxes for AP1 and AP3 (CC(A/T)6GG).[274] The newly developed algorithm will not detect these binding sites as significantly conserved because these sites will have high conservation scores in both the real and control run. Another explanation for the low recovery of functional binding sites for some TFs is the fact that the position count matrices that are used to evaluate conservation in the orthologous regions of distantly related organisms might be too specific for Arabidopsis, making it more difficult to identify conserved instances. Finally, in some cases a regulatory interaction might be species or clade-specific, making comparative methods impractical. Overlap analysis of the recovered *in vivo* binding sites elements with CNSs from the three other studies showed that 52.3% of the 787 recovered functional regions were uniquely discovered by our approach. This further supports our conclusion that this study captures a unique fraction of regulatory elements in Arabidopsis.

Whereas several studies reporting plant CNSs have suggested different lines of evidence to indicate that sequence conservation implies functional conservation and a role for CNSs in transcriptional regulation[224,247–249,251,252], their success in inferring regulatory networks has been hampered by the difficulty to convert CNSs into TF-target interactions. Based on different publicly available databases and ChIP studies, TFs for which motif information was available were integrated with the CNSs to generate a gene regulatory network containing 40,758 TF-target interactions. Overlap analysis with an experimental GRN containing 1092 confirmed regulatory interactions showed that the predicted network is highly enriched for experimental edges. In addition, the functional and expression coherence of the target genes in the different GRNs was evaluated by integrating five different biological datasets. Application of these different validation metrics on the experimental and predicted network were used to assess the functional and co-regulatory properties of the different TF-target interactions. Whereas both GRNs showed significant enrichment for all biological datasets, the predicted network outperformed the experimental network for the stress and developmental expression compendia and also for GO functional annotations. Application of the co-expression metric on two sub-networks with edges supported by CNSs showing conservation in a different number of species revealed that also regulatory interactions with lower species support are biologically relevant. Although the predicted GRN, like the experimental network, lacks many true regulatory relationships, comparison with experimentally validated targets as well as validation through the different biological datasets showed that the predicted network is of high overall quality. Compared to the experimental network, where each TF regulates on average 12 target genes, our GRN predicts on average 20 times more target genes for 157 TFs. As our GRN likely identifies many true interactions, which have not been detected and validated experimentally, it provides an important step forward towards the systematic regulatory annotation of individual genes.

A sub-network containing unique regulatory interactions based on intronic CNSs recovered a small subset of experimental interactions, confirming that intronic regions also play an important role in transcriptional regulation in plants. The TF-miRNA network contained only 24 TF-miRNA interactions, for which one previously described interaction between ABF1 and mir168a could be confirmed. A major challenge for phylogenetic footprinting of miRNA genes and the construction of miRNA GRNs is the lack of miRNA orthology information across a number of related species, which is a prerequisite for most phylogenetic footprinting methods.

Although the predicted GRN offers additional information on the transcriptional regulators controlling individual target genes, the static nature of these CNS-based interactions offers few insights about the biological context of these regulatory events. We demonstrated how integrating expression data for different organs and conditions with the predicted interactions through co-expression analysis provides an effective approach to obtain condition-specific networks. Based on 11 compendia containing gene expression profiles in different biological contexts, we identified 6597 regulatory interactions where a TF specifically co-expressed with its target gene in one or a few conditions. As shown for the secondary cell wall and AP3/PI networks, this co-expression information can be used to filter the set of predicted interactions and to identify previously unknown target genes as well as new regulators acting downstream of the TF under investigation. Furthermore, for different TFs and signaling cascades, it also becomes possible to investigate how the transcriptional control of some direct target genes changes in different conditions while other targets show constitutive co-expression.

Apart from integrating sequence conservation and expression information, other approaches combining complementary functional datasets may improve the power to correctly identify regulatory interactions. For example, the incorporation of additional regulatory information such as differentially expressed genes from TF perturbation experiments or genomic regions marked with transcription-promoting chromatin modifications can offer new ways to identify functional target genes. With the advent of TF binding data from protein binding microarray experiments for an increasing number of TFs[66,233] our CMM approach combined with co-expression analysis offers a practical means to convert in vitro TF binding information from protein binding microarrays into functional and condition-specific GRNs.

5.4 Material and Methods

Sequence and orthology information

The 12 dicotyledonous genomes used in this paper were Arabidopsis thaliana (TAIR10), Carica papaya (Hawaii Agriculture Research Center), Glycine max (JGI 1.0), Malus domestica (IASMA), Populus trichocarpa (JGI 2.0), Fragaria vesca (Strawberry Genome 1.0), Medicago truncatula (Mt 3.5) Lotus japonicus (Kazusa 1.0), Theobroma cacao (CocoaGen v1.0), Ricinus communis (JCVI 1.0), Manihot esculenta (Cassava4) and Vitis vinifera (Genoscope_v1) and were obtained from the PLAZA 2.5 database.[180]The structural annotation of the genomes in PLAZA 2.5 was updated by adding all known miRNAs obtained from the plant microRNA database.[114] miRNA sequences were downloaded from PMRD and mapped to the genomes using BLASTN[275] and GenomeThreader (-mincoverage 0.89 -minalignmentscore 0.95)[276] and only unique mappings were retained. The overlap with existing RNA gene annotations in PLAZA 2.5 and the database was determined by using BLASTN (e-value $\leq$ 1e-10) against all transcripts, and only RNA genes lacking overlap with already annotated loci were added. In total, 791 new miRNA loci were added in Arabidopsis and 20% of all miRNAs have orthologs in one or more related dicot genome.

Three sequence types, upstream, downstream and intronic, were used to identify CNSs. Upstream sequences were restricted to the first 1000/2000 bp upstream from the translation start site or to a shorter region if the adjacent upstream gene is located within a distance smaller than 1000/2000 bp (n = 33,703). 1000 and 2000bp upstream sequences were processed as two independent runs. Downstream sequences were restricted to the first 1000 bp downstream from the stop codon or to a shorter region if the adjacent downstream gene was within 1000bp (n = 33,809). The intronic sequence type is defined as the complete gene locus with exons masked (n = 20,608).

Orthologs for each Arabidopsis gene were determined in 11 comparator dicot species using the PLAZA Integrative Orthology method.[180] The included orthology detection methods are OrthoMCL[277], phylogenetic tree-based orthologs and BHIF. Through Ks graphs in the PLAZA 2.5 platform, we confirmed that all included dicot species have saturated substitution patterns (mean Ks$\geq$1) when comparing orthologous gene pairs with Arabidopsis.[180]

Synteny conservation

Orthologs were determined for each Arabidopsis protein-coding gene using the PLAZA Integrative Orthology method demanding that the orthology prediction is supported by at least two detection methods. The conservation of the orthologous relationship for the flanking gene upstream and downstream of each ortholog was determined for each of the comparator species.

Comparative Motif Mapping

Known motifs were mapped on the regions covered for the three sequence types for all included species using dna-pattern allowing no mismatches.[278] 692 cis-regulatory elements were obtained from AGRIS[259], PLACE[176] and Athamap.[279] In addition, 44 positional count matrices were obtained from Athamap and for 15 TFs positional count matrices were obtained from ChIP-Seq data (see section 'ChIP-Seq *in vivo* targets'). Positional count matrices were mapped genome-wide using MatrixScan using a p-value cut-off $\leq$ 1e-05.[278]

For each Arabidopsis gene and per sequence type, a conservation score SCMM is determined per motif. The SCMM is calculated as the number of species in which this motif was conserved in an orthologous family context. The statistical significance of each motif with SCMM was tested through a comparison with the SCMM derived from 1000 random gene families that have the same number of orthologs and species but are lacking an orthologous relationship to the query gene. Evaluation of the statistical significance using larger sets of random families (1000-100,000) confirmed that the p-values obtained using 1000 non-orthologous families are robust.

The FDR was calculated through a control experiment in which the entire analysis, including all Arabidopsis genes, was performed using non-orthologous genes. For each query gene a family was randomly assembled sampling non-orthologous genes, but maintaining the number of genes and the species composition of the real orthologous family. The real and control run were compared and footprints in the real run with a p-value that corresponds to a FDR $\leq 10\%$ were retained.

Alignment-based phylogenetic footprinting

Pairwise alignments were generated between all Arabidopsis query genes and their orthologous genes for all three sequence types and two orthology definitions. ACANA and DIALIGN-TX were run with standard parameters. Seaweeds was run with the step size parameter set to 1 and window size to 60 bp and 30bp (referred to as Seaweeds-60 and Seaweeds-30, respectively) and only alignments with an alignment score higher than 40 and 20, respectively, were retained. Sigma was run with the -x parameter set to 0.5.

All pairwise alignments were aggregated on the query sequence generating a multi-species conservation plot that shows for each position of the investigated region how many species support this nucleotide through pairwise footprints. All footprints for each level of conservation are extracted from the multi-species conservation plot and each footprint is defined by its length and a multi-species level conservation score SMSP, which denotes the number of comparator species supporting that footprint.

For each alignment tool and sequence type, a pre-computed pairwise background library, including $\geq$25 million alignments, was used to determine significant conservation of footprints. The background model was created by binning all investigated regions of all species on length, selecting 150 genes from each bin and making pairwise alignments for all possible length bin combinations. The reasoning behind this binning approach is that we wanted to compare the investigated region of the query gene with a background model consisting of genes that have regions of similar size. For each Arabidopsis gene, 1000 non-orthologous (random) gene families with the same species and ortholog composition as the query gene were generated and their pairwise alignments were obtained from the background library. Multi-species conservation is calculated for each family and the footprints obtained from all random families are binned on length. Each bin needs to contain at least 1000 multi-species footprints together with their associated scores, otherwise one or more subsequent bins (with greater lengths) were added. Finally, the statistical significance of each real footprint was then evaluated by counting the number of footprints in random families that have an equally good or better SMSP in the associated background length bin. Comparison of results between using a background library and generating these random families on-the-fly for each gene has pointed out that the results are not altered but processing time is greatly improved. Again, the real and control run were compared and footprints in the real run with a p-value that corresponds to a FDR $\leq 10\%$ were retained.

Browsing results in GenomeView

The complete set of CNSs, overlapping known motifs and DH sites can be browsed through the link `http://bioinformatics.psb.ugent.be/cig_data/Ath_CNS/Ath_CNS.php`. While loading, when asked, the file format needs to be specified to BED format.

Protein-coding potential of CNSs

The coding potential of a CNS was determined using BLASTX[275] against the PLAZA 2.5 protein database (780,667 proteins from 25 Viridiplantae species) and all significant hits were removed. To establish an appropriate e-value cutoff for a significant hit, we randomly permuted each sequence in our CNS dataset set and performed the BLASTX search using this set of sequences to obtain the distribution of e-values for random sequences with the same length distribution.[251] We then performed the same BLASTX search on the real sequences, using the minimum e-value from the random set (e-value $\leq$ 0.001) as the cutoff for a significant hit.

Overlap of CNSs with benchmarks

Our CNS dataset was compared with different functional datasets. The first one was the Arabidopsis thaliana promoter binding element database (AtProbe) (`http://exon.cshl.org/cgi-bin/atprobe/instance.pl`), which contains 172 experimentally determined regulatory sequences in 76 Arabidopsis genes. This dataset was curated by removing results from promoter deletion experiments and CREs for which mapping data was not correct with the coordinates in the dataset, resulting in a dataset of 144 CREs present in 63 genes (Supplemental Online Data set 1^{b}). The benchmark dataset was formatted as a BED file and the overlap (recovery of elements) was determined using the BEDTools function intersectBed with -u parameter and the -f parameter on 0.5.[124] This means that an experimental CRE was considered 'correctly identified' if more than half of the region was overlapping with a CNS. CNS datasets from three recent studies were obtained through the UCSC genome browser at `http://genome.genetics.rutgers.edu/` (table top10conserved) for Hupalo and Kern (2013), the authors for the CNS data of Arabidopsis from Haudry et al. (2013) or were assembled from supplementary data.[251] These files were also formatted as BED files and compared with the AtProbe benchmark. False positives were determined by shuffling the AtProbe dataset 1000 times using shuffleBed, excluding coding sequences and the actual AtProbe instances. The overlap with CNS files was determined for each shuffled file and the median number of recovered elements over 1000 shuffled files was used as a measure for false positives. This estimation of false positives was used to calculate a fold enrichment, defined as the ratio between observed overlap and expected overlap by chance.

A list of 2807 in *in vivo* functional targets was assembled from genes that were annotated to a TF ChIP-Seq peak in non-coding DNA in which a DNA motif was significantly enriched, and that show regulatory response in the corresponding TF perturbation experiment (see Supplemental Online Data set 2^{b}). Overlap and enrichment for *in vivo* functional targets was determined in the same way as for the AtProbe benchmark. For DH sites and histone modifications datasets the number of overlapping CNSs was also determined using BEDTools. Enrichment of our CNS dataset for these marked chromatin regions was determined as described above.

Detection of DNase I hypersensitive sites and histone modifications

The BED files with the flower and leaf DH sites were downloaded from the SRA database, SRA accession number SRP009678.[216] The histone modification datasets (H3K4me3, H3K4me2, H3K9ac) were downloaded from the SRA database, GEO accession number GSE28398.[280] The reads were mapped to the unmasked TAIR10 reference genome of Arabidopsis thaliana (TAIR10_chr_all.fas; ftp.arabidopsis.org) using CLC assembly cell 4.2.0 with -c parameter for colorspace reads and -r to ignore redundant reads. Peak calling was performed using DFilter 1.0 with -std 2.[281]

ChIP-Seq *in vivo* targets

For the ChIP-Seq datasets (PHYTOCHROME INTERACTING FACTOR 4 [PIF4][282], PHYTOCHROME INTERACTING FACTOR 5 [PIF5][283], APETALA1 [AP1][160], APETALA2 [AP2][157], FLOWERING LOCUS C [FLC][284], FAR-RED ELONGATED HYPOCOTYLS 3 [FHY3][285], PSEUDO RESPONSE REGULATOR 5 [PRR5][286], APETALA3 [AP3][267], PISTILLATA [PI][267] and PHYTOCHROME INTERACTING FACTOR 3 [PIF3][287]), raw reads were downloaded from the SRA database (SRA accession numbers SRP010570, SRP010315, SRP002174, SRP002328, SRP005412, SRP007485, SRP011389, SRP013458, SRP014179). The quality of the raw data was checked with FASTQC (v0.10.0; `http://www.bioinformatics.bbsrc.ac.uk/projects/fastqc/`). Adaptors and other overrepresented sequences were removed using fastx (v0.0.13; `http://hannonlab.cshl.edu/fastx_toolkit/`). The reads were mapped to the unmasked TAIR10 reference genome of Arabidopsis thaliana (TAIR10_chr_all.fas; `ftp.arabidopsis.org`) using BWA with default settings (v0.5.9[119]). Reads that could not be assigned to a unique position in the genome were removed using samtools (v0.1.18[119]) by setting the mapping quality threshold (-q) at 1. Redundant reads were removed, retaining only one read per start position, using Picard tools (v1.56; `http://picard.sourceforge.net`). Peak calling was performed using MACS (v2.0.10;[52]). The genome size (-g) was set at 1.0e8, and the FDR cut-off was set at 0.05. Other parameters were set at their default values.

For the ChIP-chip data (BRI1-EMS-SUPPRESSOR 1 (BES1)[288], SUPPRESSOR OF OVEREXPRESSION OF CO 1 (SOC1)[289], AGAMOUS-LIKE 15 (AGL15)[290], LEAFY (LFY)[291] and FUSCA 3 (FUS3)[292], raw CEL files were downloaded from GEO (GEO accession numbers GSE24684, GSE33297, GSE17717, GSE28063, GSE43291). The Affymetrix Tiling array bpmap files were updated to the current TAIR10 annotation with Starr.[122] Peak Calling was performed with rMAT.[121] The PairBinned method was used to normalise the arrays. Peaks were called using a FDR cutoff of 0.05 except for the datasets GSE13090, GSE24684, GSE43291, and GSE40519, in which the p-value was set of 1 x 10-3 (in analogy to the original study, and necessary to obtain peak calling results). The minimum requirement of consecutive enriched probes was set at of eight. Other parameters were left at their default setting.

Peak regions were annotated based on the location of their summits as determined by MACS. A peak was assigned to the closest gene as annotated in the TAIR10 release present in the PLAZA2.5 database.[180] Both upstream, intron and downstream regions of the peak were taken into account. The complete (exon-masked) peak regions were submitted to the Peak-Motifs algorithm using default settings.[118] The p-value for motif enrichment in the peak set compared with the genomic background was calculated by mapping the motifs using matrix-scan[237] (using the same default parameters of Peak-Motifs) in 1,000 random sets of peaks of the same size and length distribution sampled without replacement from the complete intergenic genome space. Only motifs with significant enrichment (p-value $\leq$ 0.05) towards peak regions for a specific TF were retained. Lists of differentially expressed genes following perturbation of the TF were gathered from their respective publications (for SOC1, the original study describing the data was[293]).

Construction and analysis of a CNS-based gene regulatory network

Based on the known motifs compiled from the different databases and literature (see section Comparative Motif Mapping), we retained 157 TFs for which specific motif information was available. A conserved gene regulatory network was created with intersectBed (-f parameter was set to 1 demanding complete motif presence in the conserved region, -u parameter was also used), which determined the overlap between a BED file containing all CNSs, together with their associated genes, and BED files with genome-wide occurrences of the motifs of all 157 TFs. Although in most cases experiments have confirmed the specificity of the association between a TF and its binding site, we cannot exclude that predicted target genes identified through a CNS are regulated by a member of the same TF family. Overlap between the predicted GRN and the experimental network (n=1092) was evaluated by counting how may TF-target interactions from the experimental network were present in the predicted network and enrichment between two networks was defined as the number of interactions that are present in both networks divided by the number of interactions expected by chance. The number of common interactions expected by chance is given by the mean of the hypergeometric distribution: N1*N2/T, where N1 and N2 are the number of interactions in the two networks, and T is the total number of possible interactions. Statistical significance of the observed number of overlapping edges was evaluated using the hypergeometric distribution.[213] Overlap was also determined per TF, demanding that a TF had at least ten target genes.

Functional enrichment was determined for each network by using five biological datasets. Three functional datasets, Gene Ontologies[105], Mapman[112], functional modules[134] and two expression datasets, a stress expression compendia (336 microarray experiments) and a developmental expression compendia (135 microarray experiments).[239]

For the functional annotation datasets the enrichment of functional terms was determined within the set of target genes for each TF through the hypergeometric distribution with Bonferroni correction. A enrichment score (-log(p-value)*fold enrichment) was created for each significantly enriched term and the average of all enrichment scores within the network was determined. For Gene Ontology only GO slim terms were taken into account. For the expression datasets a gene pair was considered to be co-regulated in the given network if the two genes had $\geq$50% of their regulators in common. These gene pairs were identified by computing the Jaccard similarity coefficient between the set of regulators of the first gene and the second gene. For each co-regulated gene pair, we then measured the similarity of the expression profile between both genes using the Pearson correlation coefficient. Finally, the biological similarity was summarized by taking the average over all co-regulated gene pairs. For both functional annotation and expression datasets the same procedure was repeated for 100 randomized versions of the network, and fold enrichment was computed as the ratio of the average functional enrichment score, or average Pearson correlation coefficient, of the original network to the average of the randomized networks. Network randomization was done by permuting the labels of all TFs and permuting the labels of all genes, which preserves the network structure. This assures that the observed enrichment is not due to potential biases arising from structural properties of the network. Statistical significance was assessed at a level of 0.05 using a one-sided Wilcoxon rank-sum test to compare the functional enrichment scores or Pearson correlation coefficient from the original network with a random sample from the randomized networks that has the same size as the real set of scores.[213] P-values obtained using 100 randomizations were identical to those from obtained through 1000 randomizations.

Construction and analysis of condition-specific GRNs

Co-expression was determined between all TFs and target genes using the Pearson correlation coefficient based on 11 CORNET expression compendia: Abiotic stress TAIR10 (256 exp), Biotic stress TAIR10 (69 exp), Microarray compendium 2 TAIR10 (111 exp), Development TAIR10 (135 exp), Flower TAIR10 (72 exp), Hormone treatment TAIR10 (140 exp), Leaf TAIR10 (212 exp), Root TAIR10 (258 exp), Seed TAIR10 (83 exp), Stress (abiotic+biotic) TAIR10 (336 exp), Whole plant TAIR10 (85 exp) from.[239] A Z-score transformation of correlation coefficients was performed in order to determine significant co-expression. A TF-target interaction was deemed significantly co-expressing if the Z-score was bigger or smaller than 2. Only TF-target interactions that showed significant co-expression in less than four compendia was used as an additional filter to obtain specificity. This threshold was selected because of the presence of three stress-related compendia.

Acknowledgements

We thank Bram Verhelst for technical assistance during the processing of the miRNA annotations, M. Blanchette for sending us the A. thaliana CNS dataset and Edward Himelblau for proofreading. This work was supported by the Multidisciplinary Research Partnership "Bioinformatics: from nucleotides to networks" Project (no 01MR0310W) of Ghent University. K.S.H. and J.V.D.V are indebted to the Agency for Innovation by Science and Technology (IWT) in Flanders for a predoctoral fellowship.

Chapter **6**

General Conclusions and Perspectives

Before ending this PhD with a more general discussion of my perspectives on the field of regulatory genomics, I would like to follow up on the objectives that were set out at the start of this project. If you recall, we defined a primary objective and secondary objective, which were to study the organisation of transcriptional regulation, and providing functional annotation predictions to guide molecular biologists when studying unknown genes. I will start with the secondary objective, to end with general conclusions and future perspectives on the primary objective of regulatory genomics.

6.1 Unravelling gene function in Arabidopsis

Integrative Modules Follow-up

To provide leads on the function of unknown Arabidopsis genes, we applied a guilt-by-association approach on the integrated modules that were delineated based on different data types, that exhibited strong complementarity in their gene-gene associations.[134] These annotations were validated by cross-checking our predicitions with experimental results that had become available in the GO database since the start of our analysis (i.e. the data 'freeze'). It included completely unknown genes and genes for which a unrelated function had already been known. The validation showed that out of 1,460 genes that had gained a new annotation since the start of our analyses, we predicted 29.7% correctly. Based on these numbers, one could extrapolate that we could correctly predict 30% of all gene functions that were unknown at that point in time.

Reversely, among all of our 5,562 module-based function predictions, 7.8% of had been validated by new experimental annotations. Whereas the this confirmation rate seemed low, we claimed this to be a lower boundary since many of our predictions could still be validated in the future. As another two years have passed —also roughly the interval between the data freeze and the first validation cross-check —the validation was repeated in October 2014 using the latest Gene Ontology Data for Arabidopsis[a]. The methodology followed was the same as described in section 3.4, with the slight adjustment that I removed additional general GO terms of low information value (making the evaluation more stringent).

The comparison of the newest annotation file with the data freeze, 1,368 genes have received new experimental GO-BP annotations. Remarkably, this number is lower than the 1,460 genes in 2012. Upon further investigation, this was found to be due to clean-ups in the gene ontology annotation, as 891 genes were new compared to 2012, but 1,013 had been removed. Two randomly chosen examples are AT5G54390 (AHL) and AT4G34580 (COW1). For AHL, there was an annotation in 2012 with GO:0016311 (dephosphorylation) which is now gone. COW1 was associated with GO:0015914 (phospholipid transport). This association has now been changed to the 'Molecular Function' branch of the GO under the name 'phosphatidylinositol transporter activity'. Thus, due to added genes, and changes in existing annotations, the fraction of correctly predicted genes has risen to 54.3% (Table 6.1).

Complementary, based on the 5,397 predicted gene annotations, 13.8% has now been validated (Table 6.2). The rise in validation rate of our complete set of predicted annotations is of prime importance, as

[a] `http://purl.obolibrary.org/obo/go/go-basic.obo`

Table 6.1: **Comparison of 1,368 module genes having new experimental GO-BP annotations with the module-based function predictions.**

Genes	Unknown[a]		Unknown Experimental BP[b]		Other Experimental BP[c]		Total	
	No. Predicted	No. Confirmed[d]	No. Predicted	No. Confirmed	No. Predicted	No. Confirmed	No. Predicted	No. Confirmed
All Genes[e]	259	168 (64.9)	283	191 (67.5)	826	384 (46.5)	1,368	743 (54.3)
Conserved	222	147 (66.2)	222	154 (69.4)	723	336 (46.5)	1,167	637 (54.6)
Not conserved	56	27 (48.2)	93	49 (52.7)	231	64 (27.7)	380	140 (36.8)

failure to do so would have abolished our claim of it being a lower-bound.

Table 6.2: **Comparison of 5,397 module-based function predictions with new experimental GO-BP annotations.**

Genes	Unknown[a]		Unknown Experimental BP[b]		Other Experimental BP[c]		Total	
	No. Predicted	No. Confirmed[d]	No. Predicted	No. Confirmed	No. Predicted	No. Confirmed	No. Predicted	No. Confirmed
All genes[e]	2,299	168 (7.3)	2,319	191 (8.2)	779	384 (49.3)	5,397	743 (13.8)
Conserved	1,876	147 (7.8)	1,873	154 (8.2)	692	336 (48.6)	4,432	637 (14.4)
Not conserved	583	27 (4.6)	625	49 (7.8)	200	64 (31.3)	1,408	140 (10.0)

On a general note regarding function prediction, and the use of predictions in research, it is important to stress that, when utilising computational predictions, false positive predictions are always a risk. Therefore, any researcher should always use predictions with care, trace back the root of the underlying data if possible, and align the predictions with additional experimental results before accepting them as truth. Additionally, predictions should only carefully be used as a basis for further predictions, as this might propagate errors made in the initial predictions. However, based on the latest validation data presented above, we feel reassured that our predictions —when used properly —can make a contribution in guiding researchers when diving into unravelling the function of an unknown gene.

In addition, we note that our supplementary data website containing all of our predictions and information to trace back the sources has been consistently used, with almost 600 users and 7,000 pageviews in since publication (Figure 6.1). It has been used in numerous studies. The study of UGT72E1 found a possible explanation of its mutant phenotype[294] because of the prediction in sugar giberillic acid processes. Similarly, *hac1* has provides resistance against Paclobutrazol, which inhibits seed germination by inhibiting biosynthesis of gibberellins.[295] The resistance is thought to act through EIL4, which is predicted to be involved in gibberellic biosynthesis. These are just two quick examples of how the predictions have been used so far to complete experimental results with hypotheses based on the function predictions.

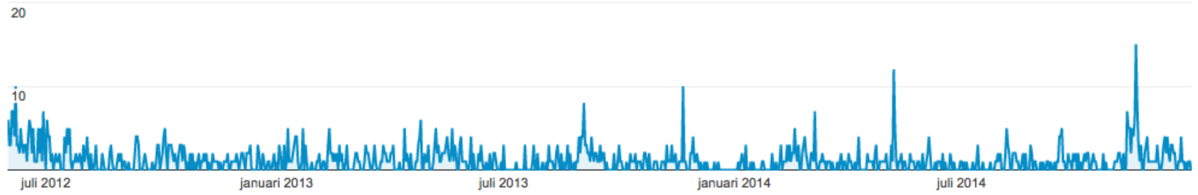


Figure 6.1: **Plant Modules Website Analytics.** *Unique users over the course of the past two years since publication.* Source: Google Analytics

ChIP-Seq as an aid to elucidate molecular pathways

The result chapters on the ChIP and the CNS network were focused on studying transcriptional regulation from a holistic viewpoint. But apart from studying regulation in itself, the binding pattern of a TF is a powerful method to gain knowledge on its biology. After all, the function of a TF is determined by the processess in which its targets are involved. Over the past four years, I have been involved in the biological elucidation of the processes around ETHYLENE RESPONSE FACTOR 115 (ERF115), ANGUSTIFOLIA3 (AN3), PEAPOD (PPD), and FAR1 RELATED SEQUENCE 12 (FSR12). The following paragraphs are meant as illustrations on how regulatory experiments can help in unravelling molecular pathways on a detailed scale by determining genome-wide leads. In combination with other data, these leads can be filtered down to a small set of genes that can be investigated further with classical methods.

ERF115[b]

Root development is a pivotal process in plant development, since it provides the plant with nutrients during its lifespan. Root growth is based on the continous division of the stem cell niche, located in the in the proximal part of the root. The mechanism that is in place to make sure that stem cells retain their stem cell identity is controlled by the quiescent center (QC), a group of organizing cells neighbouring the stem cells. The QC itself is restrained from dividing by CELL CYCLE SWITCH 52 A2 (CCS52A2), an E3 ubiquitin ligase[296]. Based on copurifucation experiments with CCS52A2, ERF115, which is a TF, was identified and shown to induce a QC cell division phenotype upon ectopic expression.

To indentify the mechanism underlying the ERF115 controlled QC divisions, it was analysed by transcriptomic and ChIP-Seq analysis. The ChIP-Seq analsis revealed 608 potential target genes. The GCC-box sequence motif was found enriched, which corresponds to the canonical of the ERF1 TF. Twenty potential target genes exhibited an expression shift upon overexpresssion of ERF155 (out of 259 upregulated genes, p-value ≤ 0.05).

One of the potential targets was PHYTOSULFOKINE PRECURSOR5 (PSK5), which was 8-fold upregulated in the ERF115^{OE} line. It has an ERF115 binding location at -780nt in relation to its promoter. Through additional analysis, the lead was confirmed that ERF115 is a rate-limiting factor of QC cell division by acting as a transcriptional activator of PSK5.

AN3[c]

AN3 is a known activator of cell proliferation in Arabidopsis, functioning as a transcriptional coactivator in leaf development. A developing leaf consists of a proliferation phase (cell division at the leaf primordia), followed by a transition phase, and an expansion phase (cell differentiation).[298,299] Being a transcriptional co-activator, it was known that AN3 was not a specific TF itself, but rather aids other TFs in recruiting the promoter initiation complex. The molecular mechanism through which AN3 acted however, was largely unknown.

Based on the protein-protein interaction formed by AN3 with subunits of the SWI/SNF complex, it was hypothesised that AN3 was involved in chromatin remodelling by recruiting the SWI/SNF complex to the promoters of its target genes. Simply put, chromatin remodelling prepares the chromatin for binding of TFs. To identify lead genes to test the hypothesis, transcriptomics analyses were combined with ChIP-Seq. The AN3 ChIP-Seq experiment resulted in 2,702 potential target genes, of which 20 were also found as upregulated in an AN3-GR line upon activation by DEX. There was an enrichment for TFs among the potential targets, pointing towards the role of AN3 as a regulator of an extended downstream transcriptional network. Based on these results, the hypothesis was drafted that AN3 binds SWI/SNF chromatin remodelling complex to recruit the complex to the target genes of AN3.

ChIP-qPCR on the promoters of AN3, GRF3, GRF5, CRF2, COL5, HEC1, and ARR4 with an antibody against a tagged SWP73B subunit in was performed in *Col-0*, and an *an3* background. The significant reduction of SWP73B localisation at the promoters of GRF5, GRF3, HEC1, COL5, and ARR4 in the *an3* background confirmed the role of AN3 in the optimal recruitment of the complex.

PPD[d]

During development, cell number and cell size are two important determinants of the final organ size. The size of plant organs is of interest in the context of increasing the yield of crops. Since the plant organ size in itself is highly heritable, there must be an underlying molecular mechanism of regulation. In addition to the different stages in a developing leaf (proliferation phase, transition phase, expansion phase), meristimoid cells that lie dispersed across the the leaf epidermis lead to the stomatal lineage.[301–303]

[b] This section is based on Heyman et al. [10]. K.S.H analysed the ChIP data and wrote the relevant sections.

[c] This section is based on Vercruyssen et al. [297]. K.S.H analysed the ChIP data and wrote the relevant sections.

[d] This section is based on unpublished workGonzalez et al. [300]. KSH performed the ChIP-Seq analysis.

Thus, leaf size in Arabidopsis is governed by (i) the number of cells incorporated in the leaf primordia, (ii) the rate of cell division, (iii) the developmental window of cell proliferation, (iv) the extent of cell expansion, and (v) the timing of meristemoid division. A positive effect on leaf growth was found in a *ppd* genomic deletion mutant, due to prolonged division activity of the epidermal meristemoids. The Arabidopsis contains two PPD genes —PPD1, and PPD2 —, due to a tandem repeat. The PPD genes are TFs belonging to the TIFY protein family. However, the mechanism through which the PPD genes act on the meristomoid division was unknown.

Transciptomic analyses identified 49 genes that were differentially expressed in the first leaf pair of *ami-ppd* (a line with overexpression of an artificial miRNA targeting the PPD genes) compared with wild type. ChIP-Seq analysis identified 1,919 potential target genes of PPD2, among which the two PPD genes themselves. The target genes were associated with regulation of transcription and hormone metabolism. The sequence of the bound regions led to the identification of two sequence motifs. The first motif, GmCACGTGkC, containing an ABF (abscisic acid-responsive elements binding factor) binding site sequence (CACGTGGC) or less specific a G-box sequence (CACGTG), preferentially located near the peak summit, is present in 726 peak sequences. The second motif, yctCACGCGCyt, is also related to a G-box sequence and found in 275 peak sequences. Out of the 49 differentially expressed genes, 13 (including PPD2) were found in the list of genes with a PPD2 binding site nearby.

Notable targets include the cell-cycle related CYCD3;2, CYCD3;3 and the chromatin organiser HMGA. This shows PPD might act by limiting the division of meristemoids during leaf development, more particularly the amplifying divisions as the activity of these D3-type cyclins proteins has been shown to be important for determining cell number in developing leaves.[304] A null mutation of HMGA has been shown to result in a decreased rate of cell proliferation in mice[305] thus hinting that PPD reduces cell proliferation following multiple paths.

FRS12[e]

During development, plants needs to control the fine balance between spending energy on growth or on defence strategies. Plants have adopted the pragmatic solution to invest in growth as long as no threat poses itself. FRS12 was picked up in a protein-protein interaction complex with NINJA, a repressor of jasmonate signalling. FRS12 is a TF of the FRS family, which also encompasses FHY3 and FAR1, well-known regulators of light response, one of the most important pathways for plant development. Thus, FRS12 and it close homologue FRS7 —also in the protein complex with NINJA —are good candidates to be central regulatory elements in the switch between growth and defence.

Genome-wide binding analysis was performed both at daytime and night-time, to asses the influence of light on the binding activity of FRS12. The analysis revealed 7,669 and 6,225 potential target genes at daytime and night-time respectively, of which 85% were shared between the two conditions. GO Slim analysis of loci binding sites highlighted diversified biological processes potentially regulated by FRS12 and included signalling, growth, flowering, development and metabolism. An inducible overexpression construct was used to perform transcriptomic analyses, resulting in 351 differentially expressed genes, of which 86 had a FRS12 binding event nearby.

We could not identify an influence of FRS12 on the JA-dependent wounding response (it was not impaired upon frs12 knock-down), and FRS12 could not be identified as a switch between growth and development. Nevertheless, we could gain insight in its role in light signalling. Among the target genes, we found genes such as the flowering time regulator GI and the photomorphogenesis regulators PIF4, PRE1 and PIL1 that showed strong peaks at their promoter regions thus partly unvealing the transcriptional cascade.

[e] This section is based on unpublished workRitter et al. [306]. KSH performed the ChIP-Seq analysis.

6.2 Organisation of transcriptional regulation

Number of binding sites per gene

Whereas the concept of *cis*-regulatory modules was well-known in transcriptional regulation in plants before the start of my PhD[148,307–309], most analyses were based on small-scale experiments. As a consequence, the degree to which different TFs can bind a same promoter has never been able to be studied.

Here, we explored the organisation of transcriptional regulation in the context of the sets of functionally associated genes (indirect evidence through *de novo* motif finding) and in the context of genomic binding sites. Based on the *de novo* motif finding in our set of modules we performed on the integrative modules, we found that the majority of genes harboured between 1 and five motifs, with just over 10% holding more than 5 DNA motifs. *Cis*-regulatory modules had only been described experimentally for 3-5 TFs, so we were cautious about over-interpretation of this figure, especially since it was based on a proxy rather than experimental evidence of binding. Nevertheless, in relation to the similar distribution for the number of modules a gene resides in, it the number of motifs seemed perfectly plausible. Shortly after our publication, Ferrier et al.[245] made the simple extrapolation of number of binding sites for a given gene based on the ChIP data available at the time for Arabidopsis. Based on his calculations, there would be around 75 binding events per regulated gene. The extrapolation was probably an overestimation because it only took into account protein-coding genes, while we know that miRNAs and lncRNAs have promoters as well, and it is extrapolates binding events, which are not all regulatory. Nevertheless, they supported our high number of hypothesised binding sites per gene.

Confirmation of the distribution of the number of binding sites came from our integrated analysis of ChIP data, which showed that up to 18 TFs can bind near a given gene in the limited network of 27 TFs. Although these conclusions are strictly speaking limited to the network that was profiled (which is primarily light and flowering related), the results from the systematic modules provide confidence that the pattern will hold across different subnetworks.

Function of hub genes in Arabidopsis

In both the systematic module analysis as well as the ChIP network, we extracted hub genes and evaluated their function. It is important to stress that different types analyses identify different types of hubs. In the module analysis, hubs are genes that are functionally associated with many modules, and thus exert a function in the context of many small subsets of genes. functional enrichment analysis revealed that these genes are involved in immune response, photosynthesis, cell cycle, and carbohydrate metabolism. In addition, hub genes are also three-fold enriched for embryo lethal genes. In the transcriptional network, hubs are genes that are bound (potentially regulated) by many TFs. Similarly, hub genes from in the ChIP network are significantly enriched for genes involved in stimulus responses, development, signalling, and process regulation. But whereas both sets of hub genes were enriched for response pathways, only the module hubs were enriched for embryolethal genes. Thus embryo-lethality is linked to genes that have broad functionality in many different modules but are not necessarily regulated by many TFs.

6.3 Future Data in Regulatory Genomics in Plants

Chromatin Conformation

DNA conformation methodologies have been proven useful in determining the global chromatin structure, but are less suited to aid in unravelling the regulatory genomic architecture. Another method, called ChIA-PET holds a far greater potential in the field of regulatory genomics. Whereas 3C and its successors[310] determine the chromatin interactions between regions of interest, ChIA-PET allows the selection of genomic regions based on its association with a protein of interest (Figure 6.2). It works by pulling down the ligated junctions associated with the protein of interest by use of an antibody. In other words, it allows the detection of chromatin loops bringing a TF in proximity of the promoter of a target gene while binding a sequence motif located elsewhere. A downside of the employment of ChIA-PET entails that random ligation events occasionally could occur between highly enriched DNA fragments. Therefore, a suitable control experiment will be critical for the ChIA-PET method in the same way as for ChIP studies.

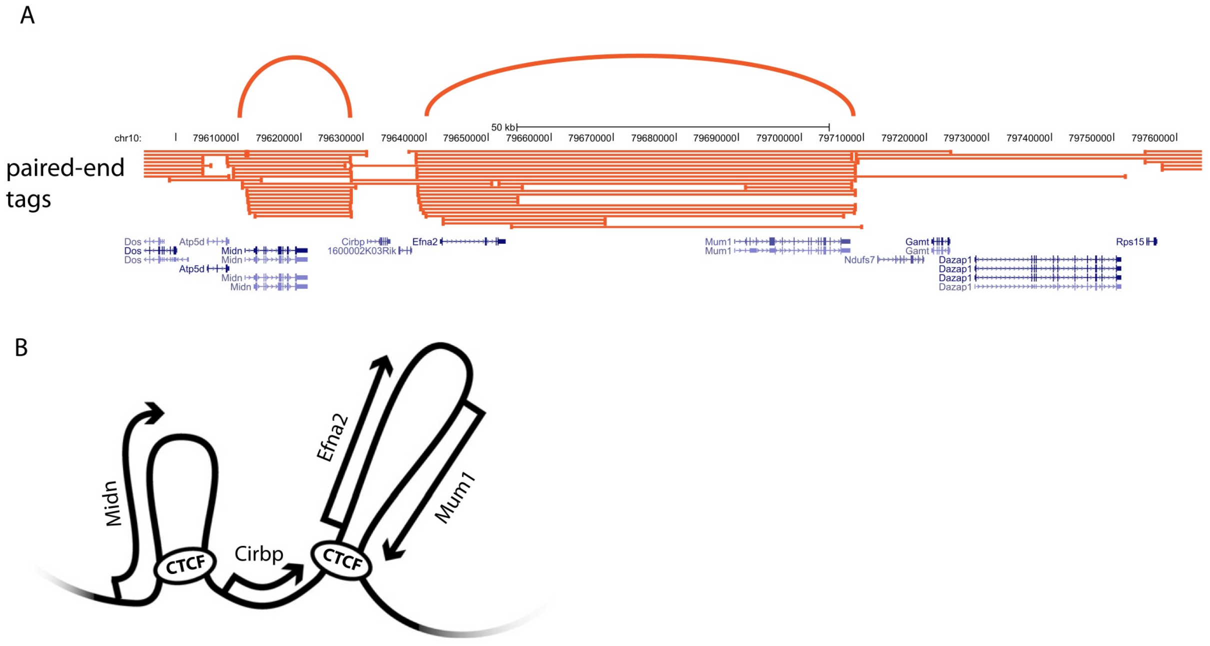


Figure 6.2: **ChIA-PET.** *ChIA-PET allows the detection of TF binding sites based on ChIP in addition to the chromatin loops formed between different binding sites. (A) ChIA-PET analysis results in paired-end reads that identify the TF binding site through their mapping position, and the chromatin interaction through their insert span. Arcs depict loops formed between binding sites. (B) Schematic representation of a possible explanation of the data shown in (A).* Source: de Wit and de Laat[310]

The methodology has been applied in a number of studies in the human field, with some very important results. Fullwood et al.[311] demonstrated that ER-α-mediated chromatin associations are essentially local, because less than 1% of the data correspond to the association among genomic regions located more than 1 Mb apart. Li et al.[215] scanned the genome for distant interactions based on RNAPII in cancer cells. Apart from interactions between promoters and long-range *cis*-regulatory elements, they found that interactions between promoter regions were also prevalent. Importantly, for the interactions between promoters and non-promoter regions, more than 40% of the non-promoter regulatory elements that were found to interact with a distant promoter regions exhibited no interaction with the nearest promoters. This is of impact on the assignment of ChIP-Seq regions to their potential target genes. Heidari et al.[312] performed ChIA-PET analyses on chromatin marks (marks known to associate with enhancers, promoters, and active regulatory elements), POLR2A (a RNAPII subunit), and RAD21 (part of the cohesin complex known to be involved in chromatin interactions during replication) and found that —for the TFs profiled —interactions between enhancers and promoters are cell-type specific. While proximal binding events were enriched at house-keeping genes, distal binding events confer regulation of dynamic biological processes. Chen et al.[313] discovered that many miRNA and protein-coding loci are coördinately expressed and functionally compartmentalised, forming transcription factories. Although data in Arabidopsis suggests that transcription factories do not exists because there is no evidence of large topological domains in conformation experiments[88,103,314], similar ChIA-PET experiments with RNAPII in different plants will give the definitive answer. On a sidenote: given the particularly dense genome of Arabidopsis, it might not be the ideal model for these kinds conformational analysis as it does not seem impossible that this could have impacted structure-based regulation.

All these results show a clear role of structural chromatin features in genomic regulation, and underline the importance of profiling the chromatin in future regulatory genomics studies. Note that ChIA-PET inherently not only detects the interactions, but also the protein binding sites, and is in essence replace ChIP-Seq as an all-in-one method to this purpose.[312]

Long Non-Coding RNA (lncRNA)

LncRNAs are a family of non-coding RNAs, with a minimal length of 200bp and containing multiple exons. With respect to genomic features, they can lie antisense to coding transcripts[315] (natural antisense transcripts), be encompassed in an intron (incRNAs), or reside in the intergenic space (lincRNAs). In comparison to miRNAs and siRNAs, lncRNAs are fairly new, and certainly the least well-understood of the regulatory non-coding RNAs. In plants at least some of them are transcribed by RNA polymerases IV and V[316].

LncRNAs can be detected using genome-wide strategies[317], and more than 16,000 lncRNAs are currently in the plant long non-coding RNA database[318]. The functions of the different lncRNAs remain largely unexplored. Natural antisense products can lead to the formation of siRNAs after processing. Other mechanisms described in human include transcriptional inference, initiation of chromatin remodelling, promoter inactivation by (i) binding to basal transcription factors, (ii) activation of an accessory protein, (iii) activation of transcription factors, (iv) oligomerisation of an activator protein, (v) transport of transcription factor, (vi) epigenetic silencing of gene clusters (probably not in plants as nuclear topological domains are not formed in Arabidopsis), and (vii) epigenetic repression of genes[319].

In plants, a number of mechanisms of lncRNAs have been identified. FLC is an flowering repressing TF, that is repressed itself following vernalisation to allow flowering. COOLAIR is a long antisense lncRNA of FLC[320] while COLDAIR is an incRNA on the sense strand[321] (Figure 6.3). COLDAIR induces a epigenetic repression of FLC in coordination with polycomb, necessary for vernalisation through H3K27me3 modifications. COLDAIR itself is upregulated upon degradation of FRIGIDA (FRI), thus showing a nice example of how protein regulation, chromatin modification and transcriptional regulation are intertwined. Importantly, FLC antisense promoter sequences of COOLAIR to a reporter gene is sufficient to confer cold-induced silencing of the reporter which means that bound regions from ChIP experiments at the 3' end of genes might actually need to be assigned to the antisense transcripts. Similarly, bound regions in introns might be regulating incRNAs, rather than the coding gene in which they reside. The example of the FLC locus, harbouring two lncRNAs in addition to its protein coding gene shows the complexity of future regulatory genomics analysis.

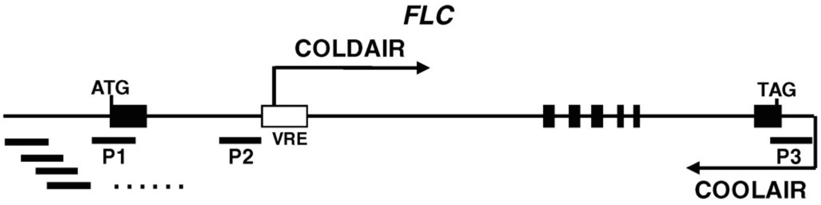


Figure 6.3: **FLC non-coding transcripts.** *Schematic representation of transcription start sites for COLDAIR and COOLAIR and the location of VRE at the FLC genomic region.* Source: Heo and Sung[321]

Plant lncRNAs can also act as the precursors of sRNAs, as 65,006 sRNAs found their loci on 5,891 lncRNAs.[322] They are found to regulate miRNAs by acting as a miRNA target mimic[323], and to modulate alternative splicing regulators by hijacking nuclear AS regulators and displacing their normal targets.[324]

Given the wide range of *modus operandi*, unravelling the function and mechanisms of all lncRNA will be a vast undertaking, but it is clear that this is yet another layer of transcriptional —and translational —regulation that will need to be solved before we can completely map the regulatory logic.

Non-Coding Variation and its Conservation

Finally, one of the most interesting directions of plant genomics with regard to evolution is the study of regulatory genomics in the context of population variation and conservation across species on the other hand. A simple pubmed search on 'Arabidopsis natural variation' returns a long list of publications released in 2014. It is an important question which variations are present, how they impact molecular mechanisms and how they have potentially lead to local adaptation.

One of the most studied traits in Arabidopsis in the context of natural variation —and thus functions nicely as an example —is flowering time. Flowering time is dependent on FRI and FLC, given that FLC is regulated by FRI. Different types of variations (affecting the coding region and affecting expression levels) have been found to influence the flowering time, for which two phenotypes exist: early-flowering and late-flowering. Johanson et al.[325] found two different deletions alleles that result in a the disruption of the open reading frame, thus rendering FRIGIDA useless. In addition, FLC has also been found to exist in five predominant haplotypes[326]. But whereas the FRI variants affected the protein completeness, the FLC variants affect FLC expression levels and rate of epigenetic silencing because of changes in non-coding *cis* variation. Rosas et al.[327] found that flowering time also varies in response to the variation in copy number of a small 7bp insertion in the promoter of CONSTANS (CO), effectively determining the number of CDF1 binding sites. To make matter even more complex, there exist flowering time determining pathways independent of FLC, of which the components also exhibit variation in their regulation thus forming a nice example of how complex natural variation is.[328] One of the future challenges of regulatory genomics will undoubtedly be to investigate how variation across the genome has affected the transcriptional network and subsequently, the manner in which these variations have led to adaptation.

One important note concerning the societal value of plant research is that whereas Arabidopsis has been a very useful model species and will continue to provide important insights, many studies are moving towards economically more interesting species. The analysis of regulatory variation in the context of the 3000 rice genomes project will therefore be of much greater societal value[329].

Experimentally studies of the conservation of transcriptional binding haven't been performed in plants, but have given interesting results in the animal field. The first study to experimentally profile the binding patterns of four orthologous TFs across species was performed in 2007[330], and revealed that 41% - 81% of the individual binding events were species specific. However, it was later shown that while many binding events were species specific, the ones associated with functional targets (i.e. genes that responded to TF knock-outs) were highly conserved.[331] A similar conclusion was reached by Schmidt et al.[43], who found that binding events are rarely conserved, but that genes that with expression levels that are dependent on the a TF were often bound by the TF in multiple species. A follow-up study by Ballester et al.[44] found that when organising human binding events into *cis*-regulatory modules, only half of those were found in a second species. However, the conserved ones were associated with liver pathways and disease loci identified by genome-wide association studies. Similarly, functional enhancers in Drosophila are more likely to be found in regions with conserved TF ChIP binding events[45].

The consistent conclusion that many single binding events are not conserved point towards the non-functionality and misannotation of many of them before anything else. Conclusions on the evolution of the transcriptional network should only be made based on binding events that exert an effect. The goal of being able to label binding sites functional or non-functional might potentially be reached by integration with other genomic data sets. After all, many important TFBSs are conserved, as shown by the proven track record of comparative genomics being able to identify conserved binding sites that can be linked to TFs and their functions. This was proven in a very recent publication[332], where binding profiles were created for 34 orthologous TFs in concert with chromatin modifications. Around half of the bound regions could be aligned between human and mouse. Within the set of bound regions that could be aligned, some factors exhibited no conserved binding events whereas for others up to 60% of binding events were conserved, showing that it is highly factor dependent, but in the same range as the results by Odom et al.[330]. Preference of binding location of TFs does appear well-conserved although promoter region binding is, as well as the chromatin states. Interestingly, while the primary motifs identified for a TF are well conserved, the secondary motifs (of associated factors) appear to be lineage-specific. TFs with conserved occupancy profiles are associated with pleiotropic functions, though to be due to increased selective pressure of having regulatory functions in multiple tissues.

Similar studies should be executed in the plant field in the future in order to assess to what extent these conclusions hold for transcriptional regulation in plants. Whereas most principles might be expected to be the same, the difference in functionality of HOT regions in animal compared with Arabidopsis shows differences might be present. As ChIP-studies in other plant species besides Arabidopsis are becoming more and more available, such experiments should be feasible in years to come.

6.4 The Interplay between Experiment and Computation

Finally, I would like to end my PhD thesis with my views on the future of being a bioinformatician/computational biologist/data scientist and what the field should strive for.

As genome-wide experimental approaches are becoming increasingly refined, expanded and utilised, the amount of genomic data that can be browsed in relation to transcriptional regulation increases. Whereas many of the data sets are generated in the context of specific research questions, integration of the different data sets is a biologically easily interpretable manner to learn what defines the context of an active binding event (e.g. mapping PWMs in open chromatin regions reduces false positive rate[225]).

The reason this learning is important is because we will never be able to understand the regulatory network unless we learn how different cues lead to binding of a TF, and potentially regulation following the binding event. No matter how complex the regulatory landscape looks at this point in time, we must never forget that nature is built logically, following the rules of physics and chemistry. In that aspect, it is of importance to generate (i) as complete data views as possible, in (ii) as many regulatory conditions as possible. The development of network inference tools and machine learning approaches will greatly profit of this boom in genomic data and will become increasingly important in detecting the regulatory logic and extracting rules of functional binding.

Importantly, one must also never forget that while studying transcriptional regulation in a integrated genomics perspective, is still a simplification of the complete picture. The link between binding and changes in transcript levels will still be obfuscated by mRNA stability and degradation. After the regulatory step of transcription, other points of regulation can influence the mRNA level, and even more the presence of the protein.

A Complete Data view

In order to gain insight into the logic that determines binding of a TF and regulation of the targeted gene, we need all the information on the current genomic state. Too many data sets are limited to one or two data types for an experiment, e.g. binding information (ChIP-Seq) and expression response (RNA-Seq following TF perturbation). When binding data is related to an expression response, the effect has consistently found to be low. But while any of those binding events may be true, it might indeed just be the case that there is no expression response. The promoter can be poised to activation, or the TF can simply bind a stretch of DNA because it's sequence motif while it lacks all other requirements for steering expression. The FRS12 binding pattern is a good example of this, as 85% is identical between day and night samples, while it is a light-signal steered TF.

Mapping the state of the chromatin, will ultimately be necessary to find which complete genomic state represents a binding event that confers regulation. Any ChIP binding experiment should include: mapping of the open chromatin using DNase I and MNase, mapping all known chromatin modifications using ChIP, determining the methylation state of the DNA, DNA conformation information.

Data with a High Information-to-Noise Ratio

So far, many methodologies have been applied to non-specific samples such as whole plant or seedlings. Even leaves are a collection of different cell types. While the information of the true binding is present (the binding in the cells where the TF is regulating its target genes), it is diluted by binding events (and lack thereof) in other cells. In the future, significant advances to learn the logic of transcriptional regulation will only be made when the profiling is performed in single-cells, such as single cell transcriptomics.[333]

The Future of Data Generation

Whereas many labs studying regulation are aware of these needs, at this point in time, a coordinated effort is missing. Ideas on how these goals could be reached, and what these goals should specifically encompass are excellently discussed by Lane et al.[334], which reaches the same conclusion as us with regards to what kinds of data are needed. The overall idea is that a coordinated effort should be set in place, following the example of the ENCODE projects in the animal model systems.

While the core principle of this idea is exactly what is necessary for data scientists to extract hypotheses, I would like to add my personal thoughts on the feasibility. I would like to believe that such an effort is possible in an academic setting, but this will ultimately depend on either joint funding, which stimulates labs to honour deadlines and put dedicated people on them. Secondly, any project needs project management, and the pENCODE effort would need someone with the necessary influence on individual group leaders to steer the project. Many labs have different adaptations and tweaks of methods, which could affect comparability of the results. However, the latter is a minor issue, as this would not diminish the vast improvement of data quality compared to what is available now. Personally, I see the greatest challenge in steering the project and managing the different group leaders.

Lastly, I see a need for a paradigm shift in the approach of molecular biology in the future. To put it radically, data scientists should not be analysing data generated by molecular biologists to answer the latter's specific question as is often the case now. Instead, molecular biologists should be generating the data required by data scientists. I deliberately is overly radicalised this statement, because ideally, scientists themselves should be trained with knowledge of both molecular biologists and data science. In a framework such as the envisioned pENCODE, the current paradigm will critically damage the progress.

6.5 Final Thoughts

Even if we could bring about all the ideal conditions and data years to come, it is perfectly possible that we are still aware of all layers of genomic information, and thus will not be able to fully relate binding to expression. But with the advances in regulatory genomics of the past 10 years in hindsight, I think I can be optimistic that unravelling the transcriptional regulation is within a lifetime's grasp.

Supplemental Data

SUPPLEMENTAL DATA FILE **A**

Integrative Plant Modules

A.1 Supplemental Figures

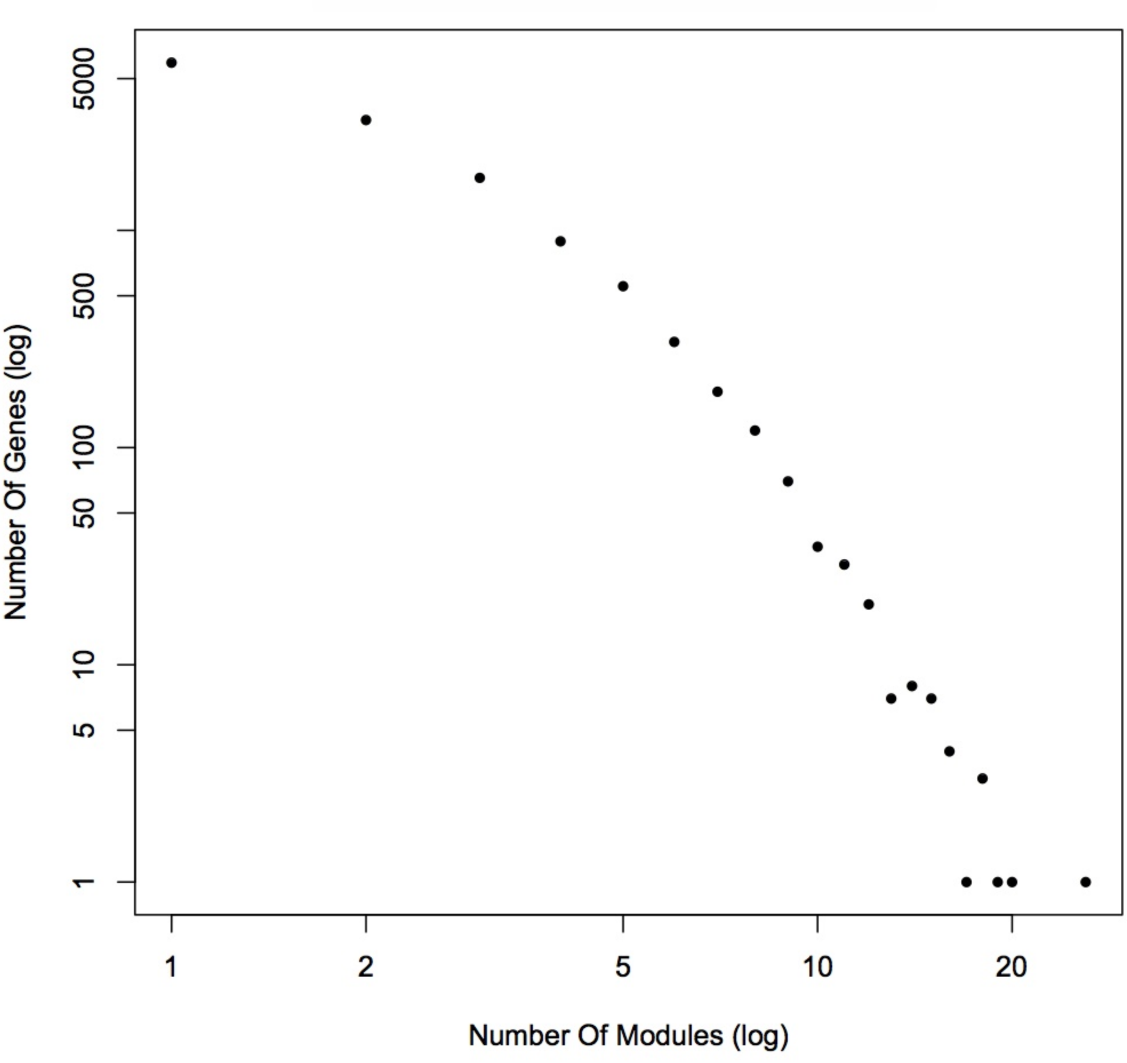


Figure A.1: **Number of modules per gene.** *Degree distribution of gene nodes in log scale.*

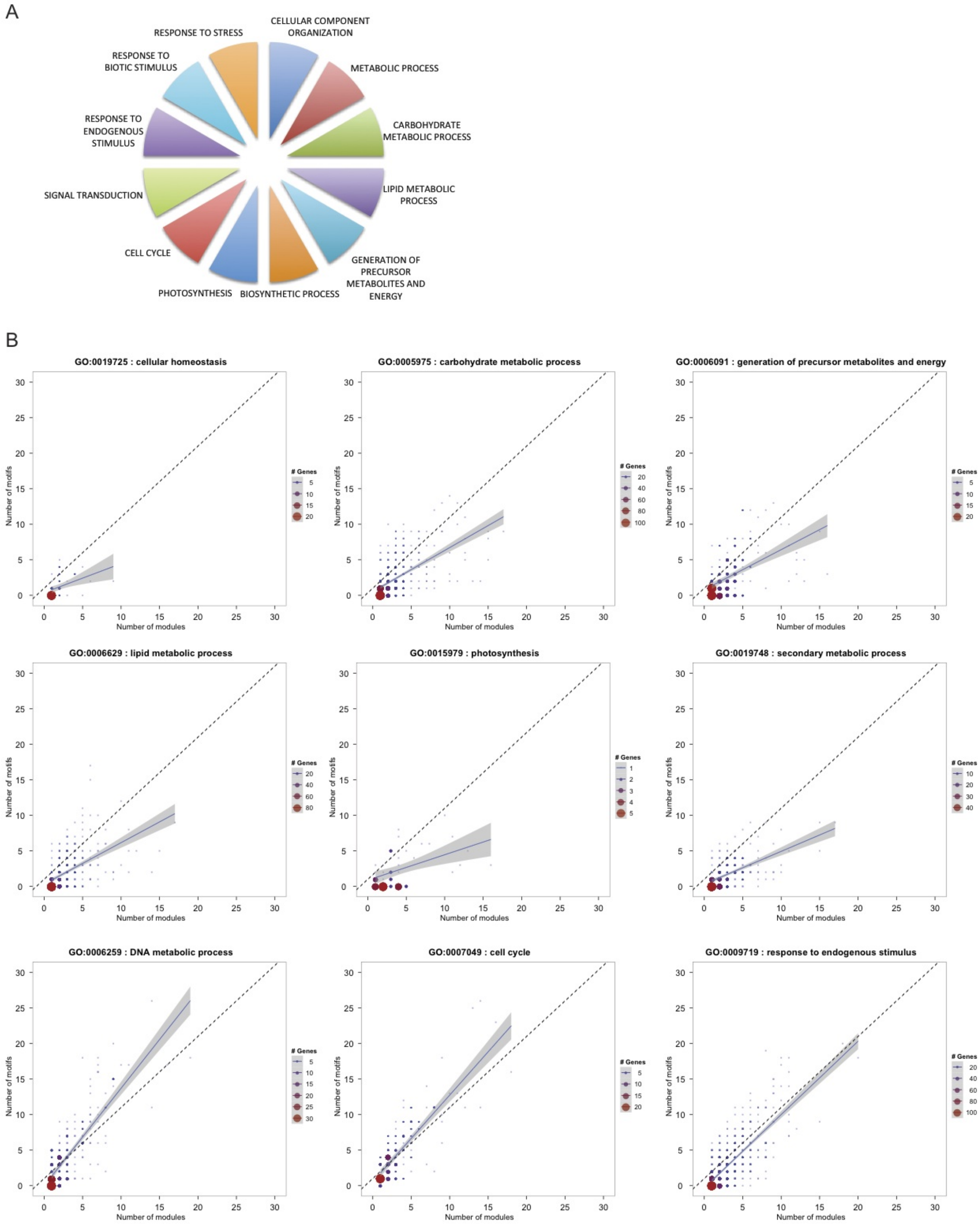


Figure A.2: **Functional enrichment of hub genes and regulatory complexity of different biological processes.** *A, Functional enrichment among hub genes represented by GO-BP slim categories. B, Regulatory complexity of different GO-BP processes. The number of genes at each coordinate is given as a colored size scale. The grey circle indicates the average regulatory complexity for all 13,142 genes. The dashed line is the function f(x) = x.*

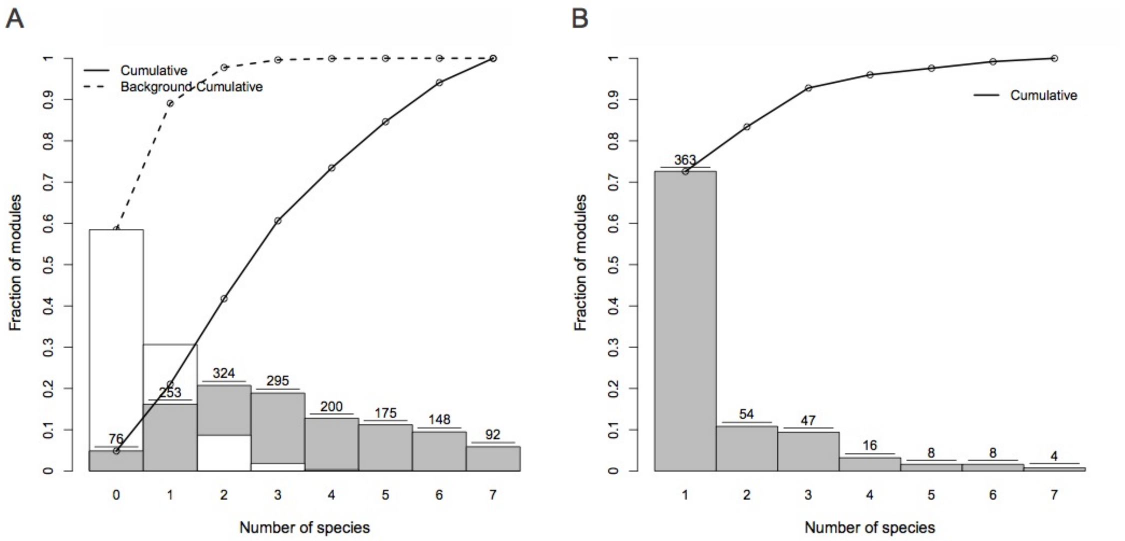


Figure A.3: **Conservation of EC and motif enrichment across the green plant lineage.** *A, EC conservation across seven species for real data (grey + black cumulative) and random data (white + black cumulative) respectively. B, Motif enrichment conservation for expression- conserved modules across seven species.*

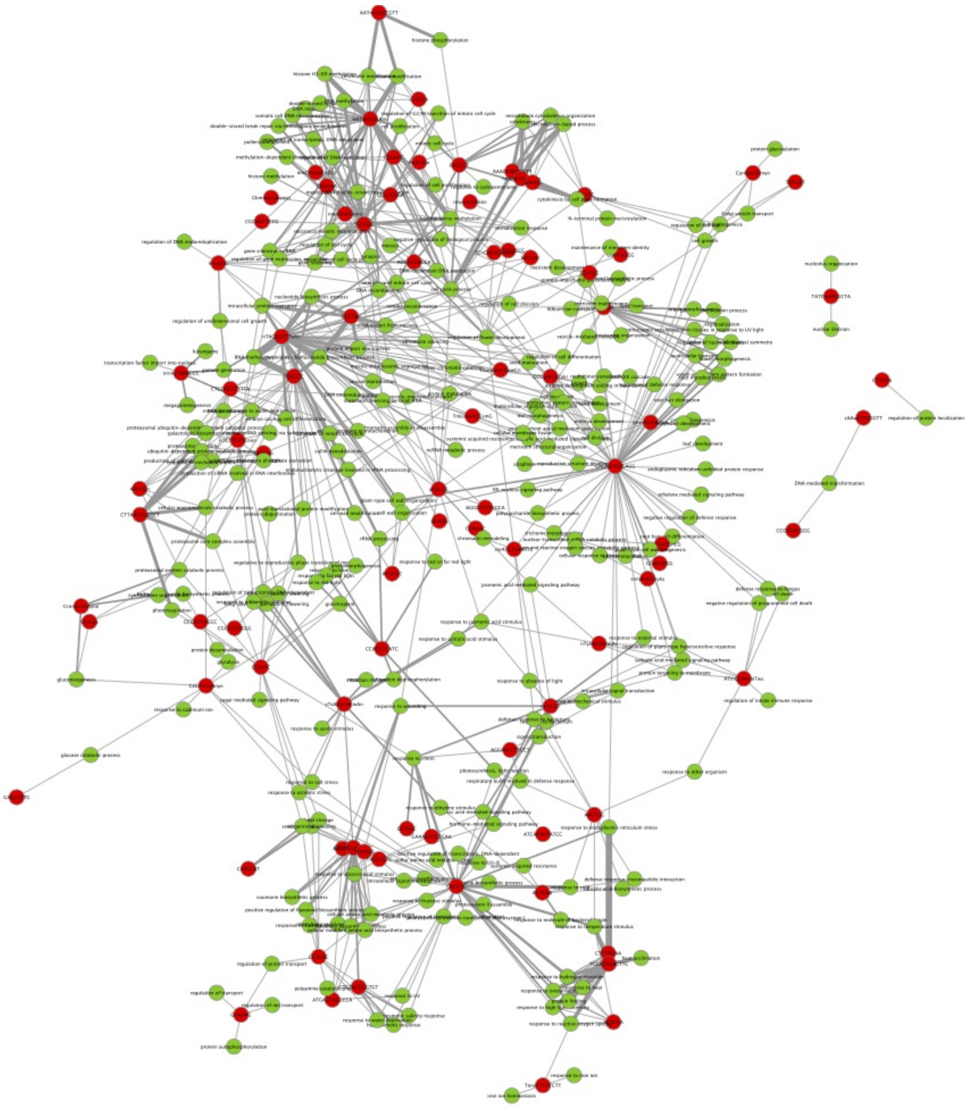

histone phosphorylation
histone H3-K9 methylation
regulation of G2/M transition of mitotic cell cycle
microtubule cytoskeleton organization
protein glycosylation
Golgi vesicle transport
cytokinesis by cell plate formation
N-terminal protein myristoylation
regulation of DNA endoreduplication
maintenance of meristem identity
nucleolus organization
nuclear division
regulation of unidimensional cell growth
transcription factor import into nucleus
regulation of protein localization
leaf development
ethylene mediated signaling pathway
negative regulation of defense response
DNA mediated transformation
defense response to fungus
negative regulation of programmed cell death
ER-nucleus signaling pathway
jasmonic acid mediated signaling pathway
response to jasmonic acid stimulus
response to salicylic acid stimulus
response to absence of light
response to wounding
regulation of innate immune response
response to other organism
glucose catabolic process
response to cadmium ion
sugar mediated signaling pathway
response to salt stress
response to osmotic stress
response to endoplasmic reticulum stress
defense response, incompatible interaction
regulation of protein transport
regulation of transport
protein autophosphorylation
response to reactive oxygen species
response to heat
iron ion homeostasis
response to iron ion
coumarin biosynthetic process
positive regulation of flavonoid biosynthetic process
photosynthesis, light reaction
proteasome core complex assembly
gluconeogenesis
tRNA processing
chromatin remodeling
systemic acquired resistance
protein deneddylation

Figure A.4: **Motif-GO map based on coexpression conservation.** *Motifs were associated to GO categories based on the modules in which they were retrieved. Edges were weighted by the average number of species with conserved coex- pression for these modules. Red nodes: motifs; Green nodes: GO categories.*

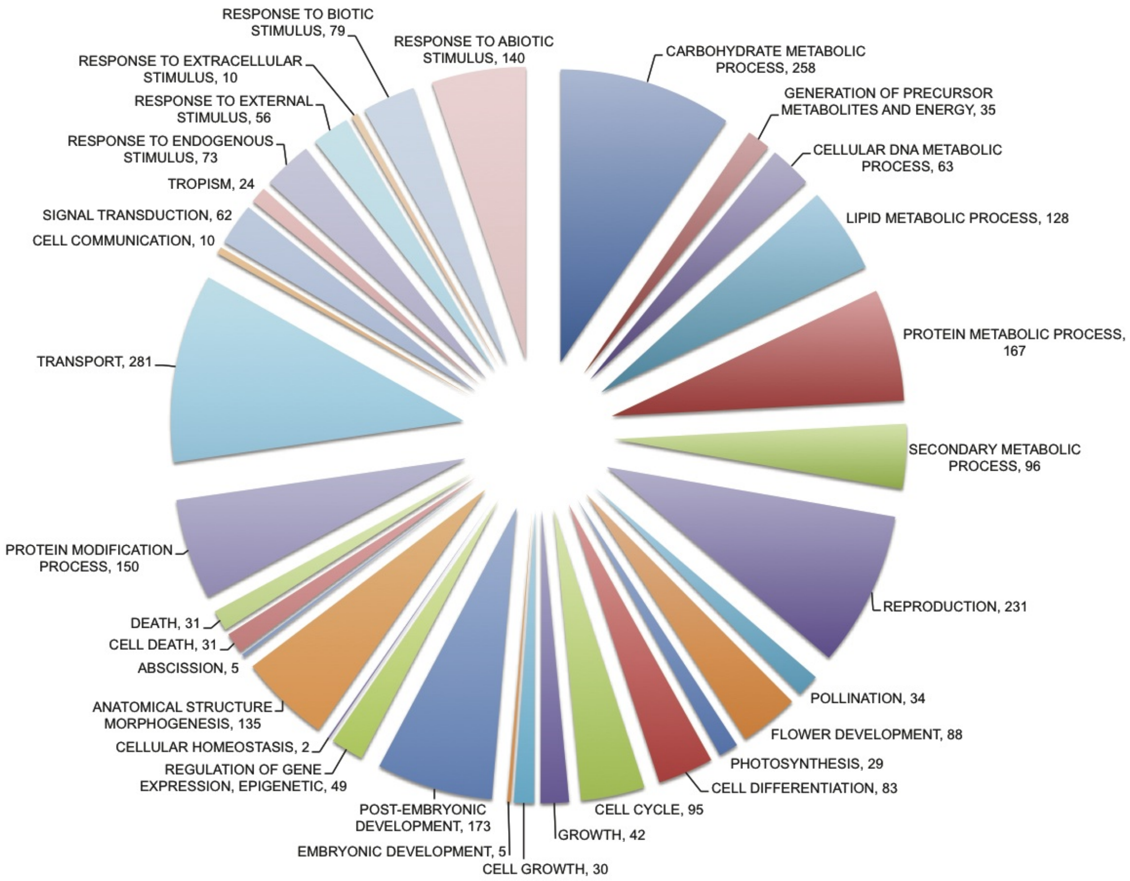


Figure A.5: **Overview of GO-BP slim predictions for 1,435 genes currently without GO-BP annotation.** *Modules with multiple GO-BP annotations can be present in different GO slim categories.*

A.2 MQSE Protocol

Multi-Query Seed Expansion: optimizing a set of seed genes prior to clustering Standard clustering techniques only utilize the genes from the input data. MQSE is a semi-supervised strategy for co-regulatory module detection that expands the gene sets (from here on referred to as âĂŸseed setâĂŹ) prior to the clustering with genes that are coexpressed with the seed set and removes seeds that do not show expression coherence with the other seeds.

1. For each seed, all genes that coexpress significantly (relative Pearson correlation coefficient or PCC threshold at the 95th percentile of a random distribution) with the seed are read (including the other seeds). The genes are ordered descending by the PCC with the seed. The order gives each gene a rank relative to the seed. The seed itself is removed from the ranking (rank 1 before removal). After iterating over all seeds, every gene has a list of ranks (a rank for each seed).
2. Calculate for each gene:
    a) The fraction of coexpressed seeds.
    b) The standard deviation of the expression profile of the coexpressed seeds.
    c) The median rank.
3. Calculate the ranking score for each gene (Scoreg in pseudocode).
4. Sort all genes based on their ranking score. Iterate over the ordered list taking into account one extra gene in every iteration. Each time the gene is a seed, calculate the enrichment score in the gene set towards the seeds. The gene set for which the enrichment score is highest, is returned as expanded gene set.

MQSE: Multi-Query Seed Expansion

```
for all $S$ in $Seeds$ do
    $i \leftarrow 0$
    for all $g$ in $Genes$ do
        if $g \neq S$ and $PCC_{S,g} > c$ then
            $G_{S,i} \leftarrow g$
            $i \leftarrow i+1$
        end if
    end for
    Sort $G$ on $PCC_{S,g}$ descending
    $k \leftarrow 0$
    for all $g$ in $Genes$ do
        $Ranks_{g,S} \leftarrow k$
        $k \leftarrow k+1$
    end for
end for
for all $g$ in $Genes$ do
    $Median\_Rank_g \leftarrow median(Ranks_g)$
    $Max\_SD \leftarrow \sum_{S=Seeds_{[0]}}^{Seeds_{[n]}} SD_S$
    $Sum\_SD \leftarrow \sum_{S=Ranks_{g,0}}^{Ranks_{g,n}} SD_S$
    $Score_g = \frac{\frac{Sum\_SD}{Max\_SD}}{\sqrt{Median\_Rank_g} * 100}$
end for
Sort $Score$ descending
for $i$ to $n$ do
    $List_{[i]} \leftarrow Score_i$
    Calculate enrichment towards seeds
    if $List$ is enriched then
        $Last\_enriched\_index \leftarrow i$
    end if
end for

$\Longrightarrow$ The extended list is $List_{[0-Last\_enriched\_index]}$
```

The ranking score

Ranking all the genes is a crucial step in the process. If the ranking does not place those genes that are associated with the seed genes on top, there will be no enrichment towards the seeds in the top of the ranked list (the seeds themselves are expected to be ranked on top, together with other coexpressed genes).

To calculate the rank score, several features were taken into account that are related to coregulation with the seeds: number of coexpressed seeds, standard deviation of the coexpressed seeds and the median rank of each gene towards the different seeds. Firstly, the number of coexpressed seeds represents the fact that we wish to find genes that are coexpressed with the entire seed set (or at least with as many seeds possible). Secondly, the number of coexpressed seeds is weighted by the standard deviations of each seed. This approach has also been used previously in a somewhat different form in MEM[194]. The assumption is that genes that have a low standard deviation are less likely to be regulated and the difference in expression levels between the different datasets is merely noise. Therefore, genes that are coexpressed with the most variable seeds are rewarded a greater weight. The third and final element of the score is the median rank. Every gene has been ranked to each of the seeds individually (if they coexpressed significantly). Based on these ranks, each gene has a median rank towards the seed set. The higher this rank (rank 1 being the highest), the higher the resulting ranking score was. All these elements were combined into a final score that is believed to place coexpressed genes that are likely to be coregulated with the seed set on top of the ranking.

For each gene g, Ranks is the list of ranks the gene has for each of the coexpressed seeds in Seeds. Max_SD is the sum of standard deviations of the different seeds (i.e. the maximum any gene can reach). Sum_SD is the sum of the standard deviations of the seeds with which the gene coexpresses significantly. Score is the ranking score calculated based on the different features. Note that the number of coexpressed seeds is represented in the sum of the standard deviations.

Supplemental Data File B

Function and Evolution of TF-bound DNA

B.1 Supplemental Figures

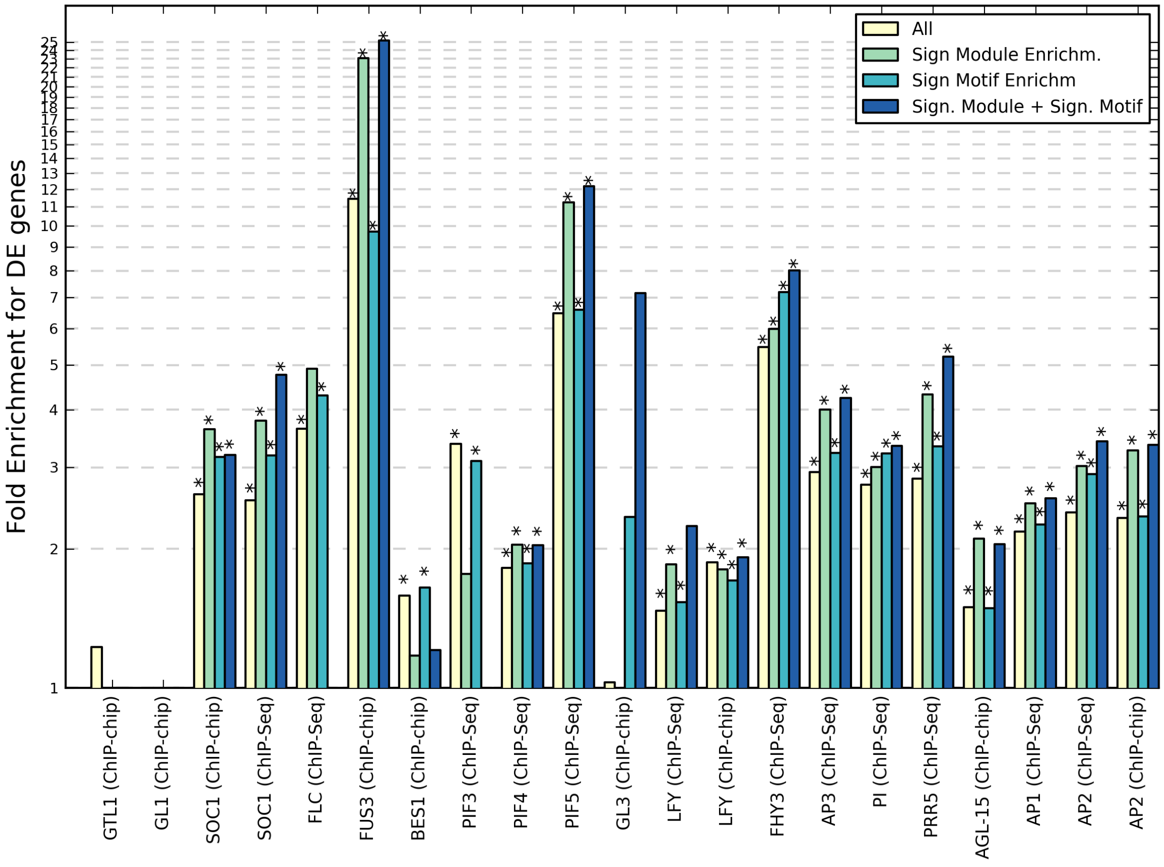


Figure B.1: **Enrichment for differentially expressed genes in network subcategories.** *Enrichment for differentially expressed genes in all potential target genes for subsets based on significant DNA motif enrichment and/or enrichment of the target genes in functional modules. Note that any value drawn indicates enrichment, as the lower bound is of the y-axis is 1. The gray dotted line marks twofold enrichment. * p-value ≤ 0.01.*

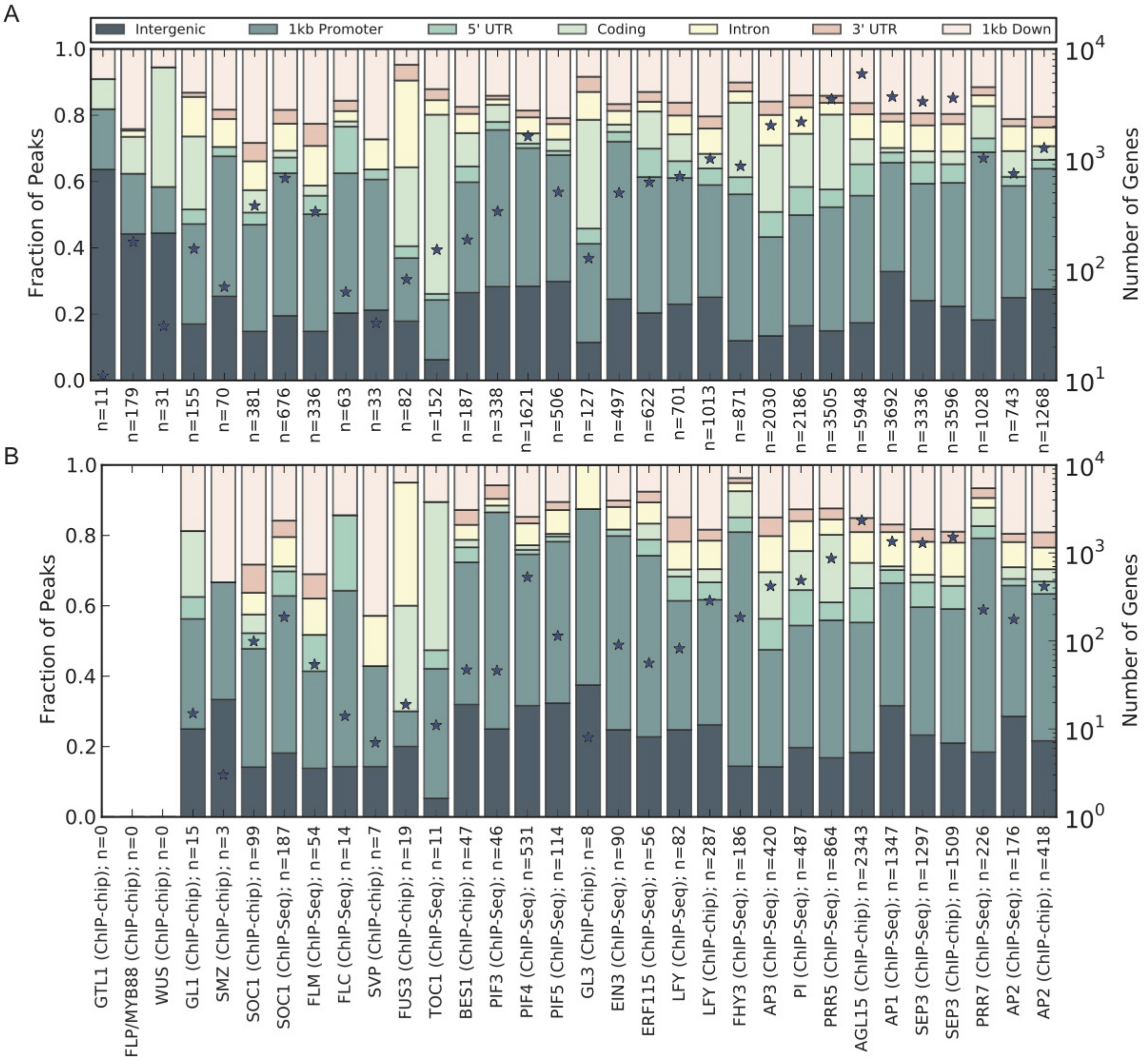


Figure B.2: **Number of potential target genes per TF and their distribution across different genomic regions for the (A) Multiple-Evidence subnetwork and (B) High-Confidence subnetwork.** *The coloured bars represent the fractions of genomic regions (left y-axis). Star signs represent the number of genes (right y-axis). The exact number of potential target genes is given in the labels (n=).*

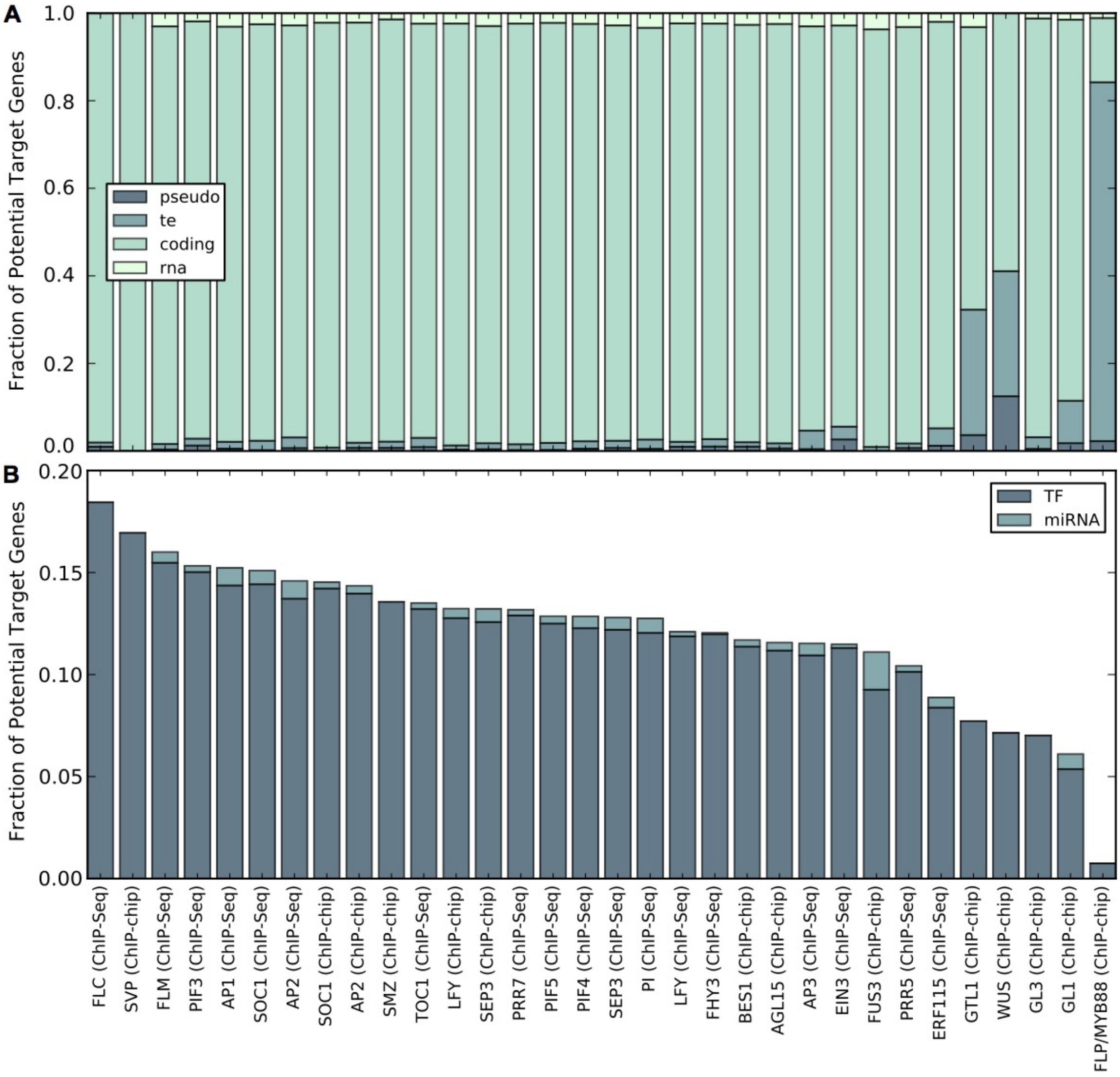


Figure B.3: **Fraction of different gene types bound by each TF.** *(A) The distribution of peak-associated gene types split up in coding, pseudogenes (pseudo), transposable elements (te), and all types of RNA genes (rna). (B) The fraction of peaks associated with regulator genes (TF and miRNAs).*

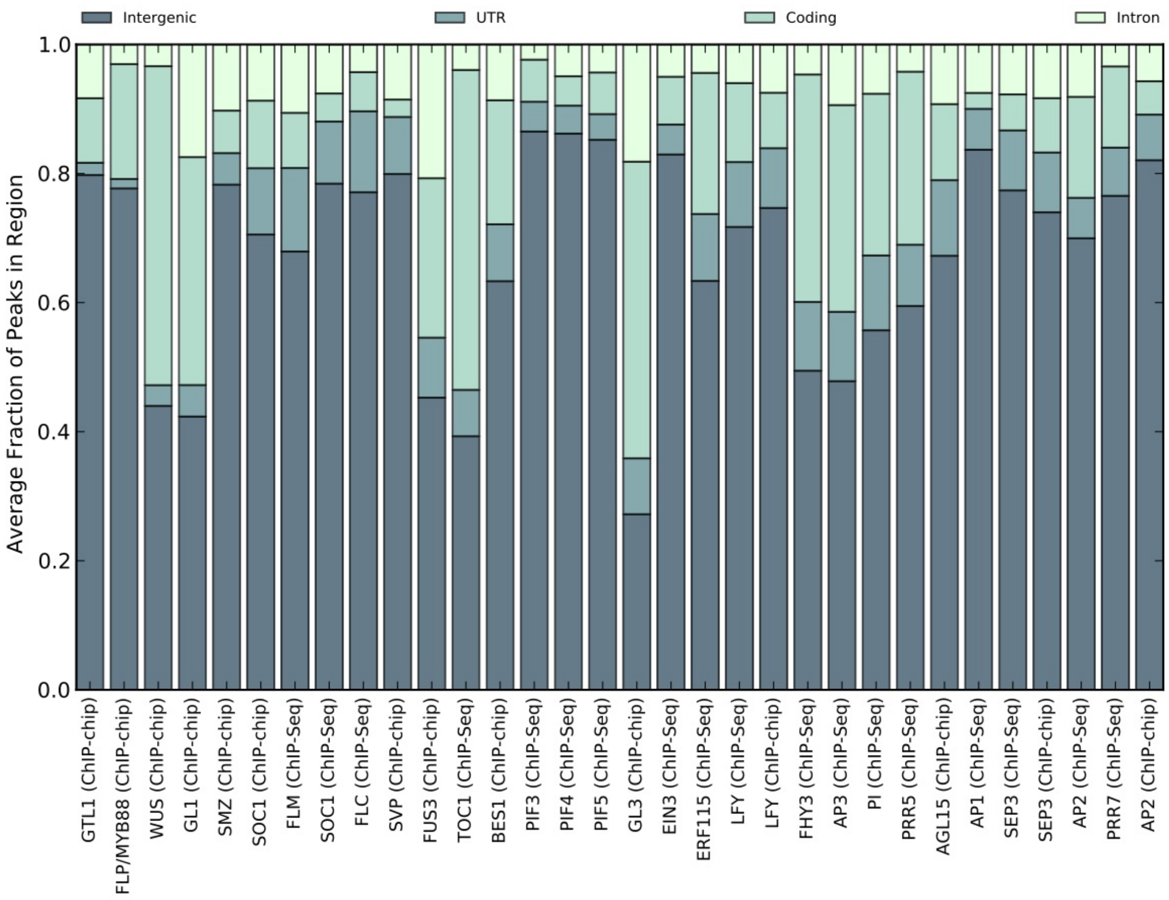


Figure B.4: **Peak region annotation based on the fraction of overlap of the entire peak region.** *The y-axis describes the average fraction of a peak that is assigned to each genomic region, only considering peaks uniquely covering a single region.*

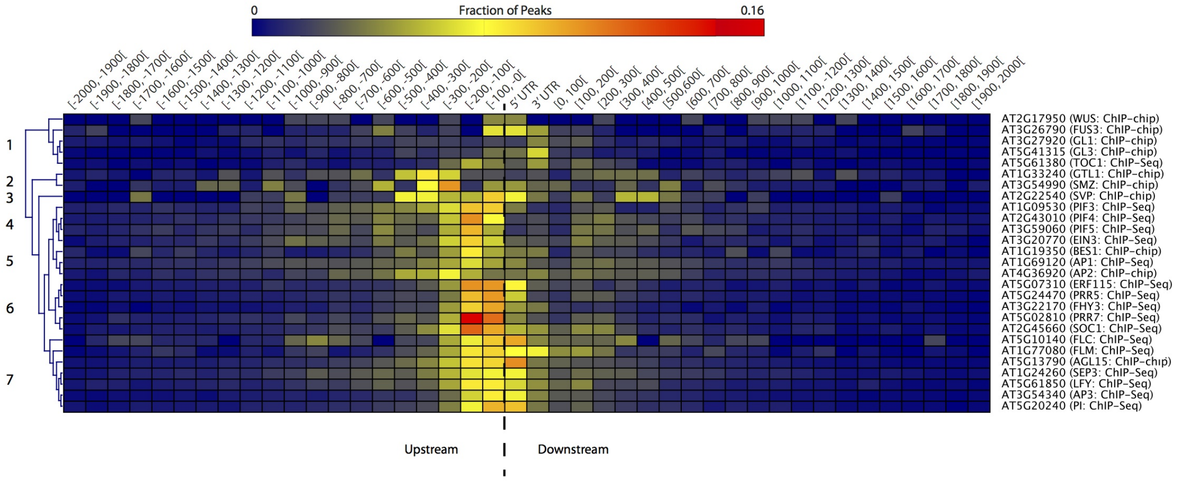


Figure B.5: **eak location binding preference for the different TFs.** *The values represent the fraction of all bound regions binding in each bin. Bins are size 100. TFs are hierarchically clustered based on the Pearson Correlation Coefficient between their location vectors. Numbers on the left indicate clusters used to discuss different subtypes in the results.*

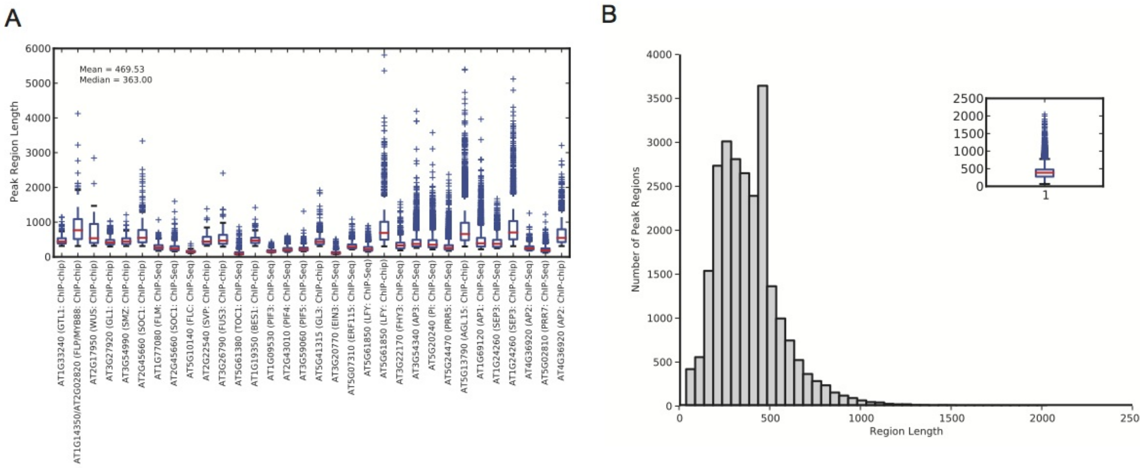


Figure B.6: **Length distributions of (A) all peak regions for each TF and (B) the merged regions based on all TF peak regions.** *The inset in (B) shows the boxplot for the merged region lengths.*

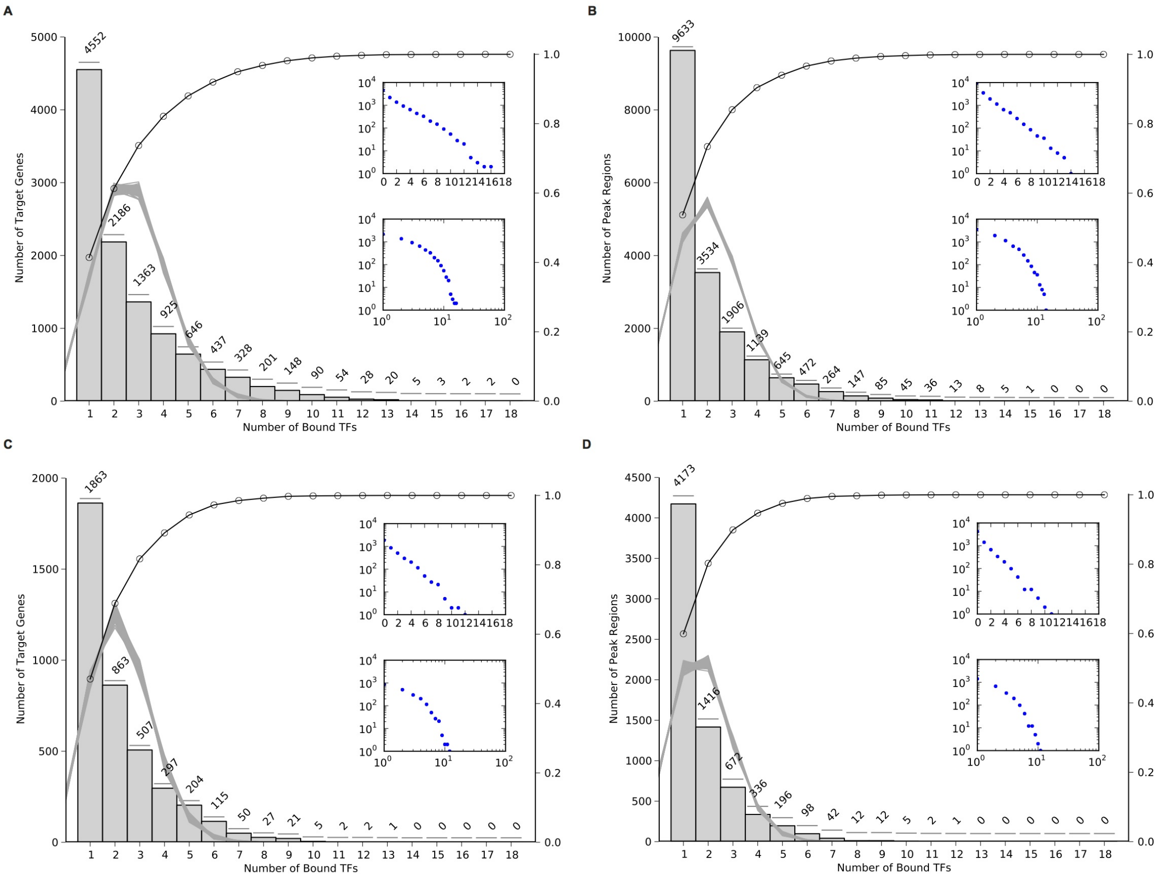


Figure B.7: **Histogram of the number of TFs per (A) potential target gene and (B) per peak region for the Multiple-Evidence (top) and High-Confidence subnetworks (bottom).** *The black line is the cumulative histogram, the grey band is the collection of 1000 random distributions. The upper inset represents the same data on log-y scale; the lower inset represents the same data on a log-log scale.*

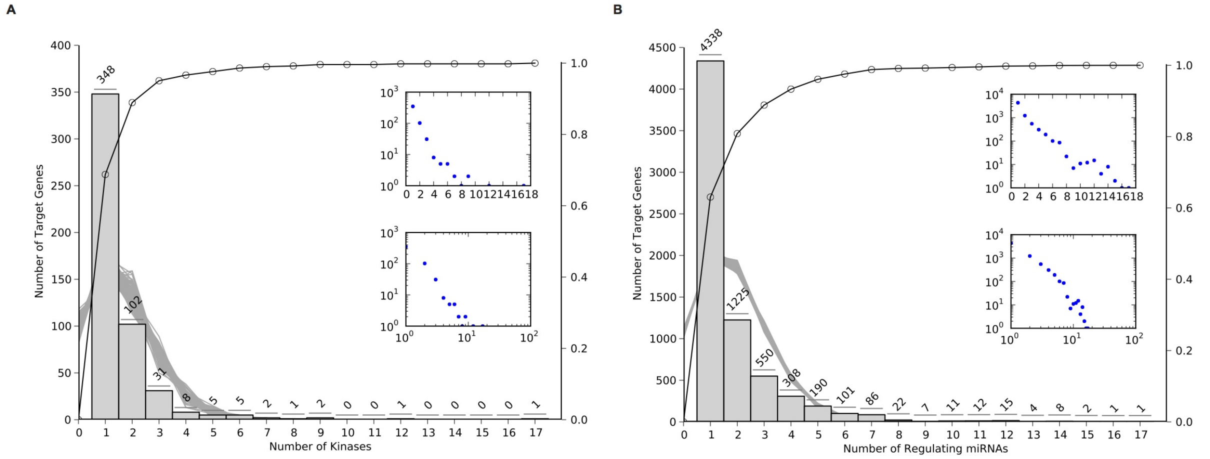


Figure B.8: **Histogram of the number of (A) regulating Kinases and (B) miRNAs per target gene.** *The black line is the cumulative histogram, the grey band is the collection of 1000 random distributions. The upper inset represents the same data on log-y scale (exponential distribution); the lower inset represents the same data on a log-log scale (power-law).*

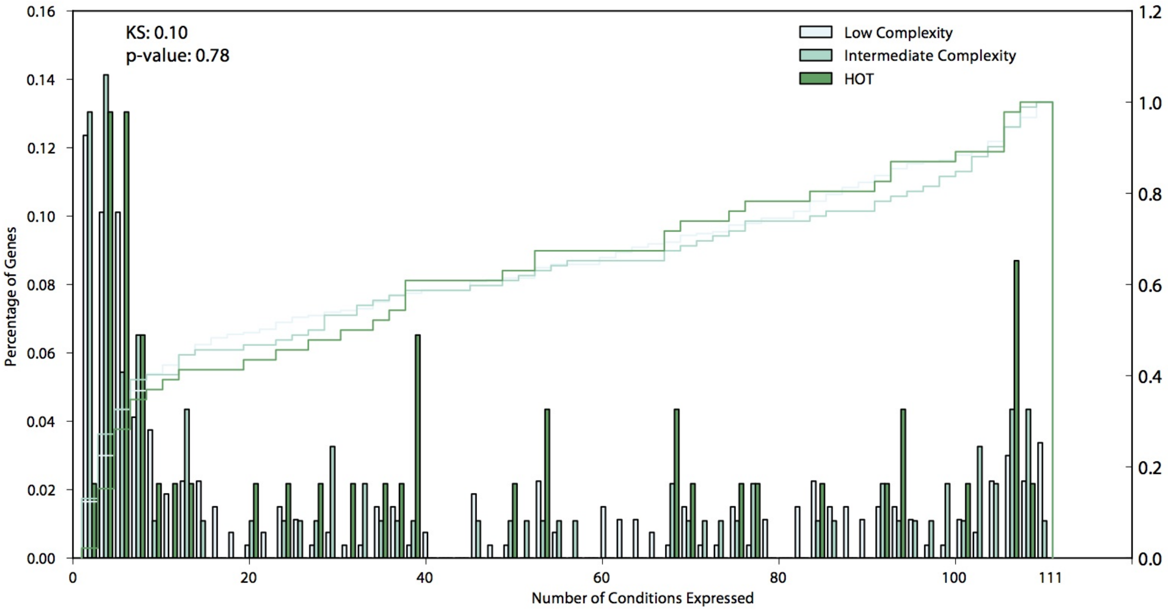


Figure B.9: **Expression breadth in function of regulatory complexity.** *Expression breadth histograms based on a non-redundant expression compendium of 111 conditions for three series of complexity: Low: < 3 TFs; Intermediate: >= 3 TFs and < 8TFs; hub: >= 8 TFs) for HOT associated genes involved in flowering versus genes associated with low complexity merged regions. The lines are the cumulative histograms. The Kolmogorov-Smirnov (KS) statistic and p-value are calculated between the low complexity regions and the HOT regions.*

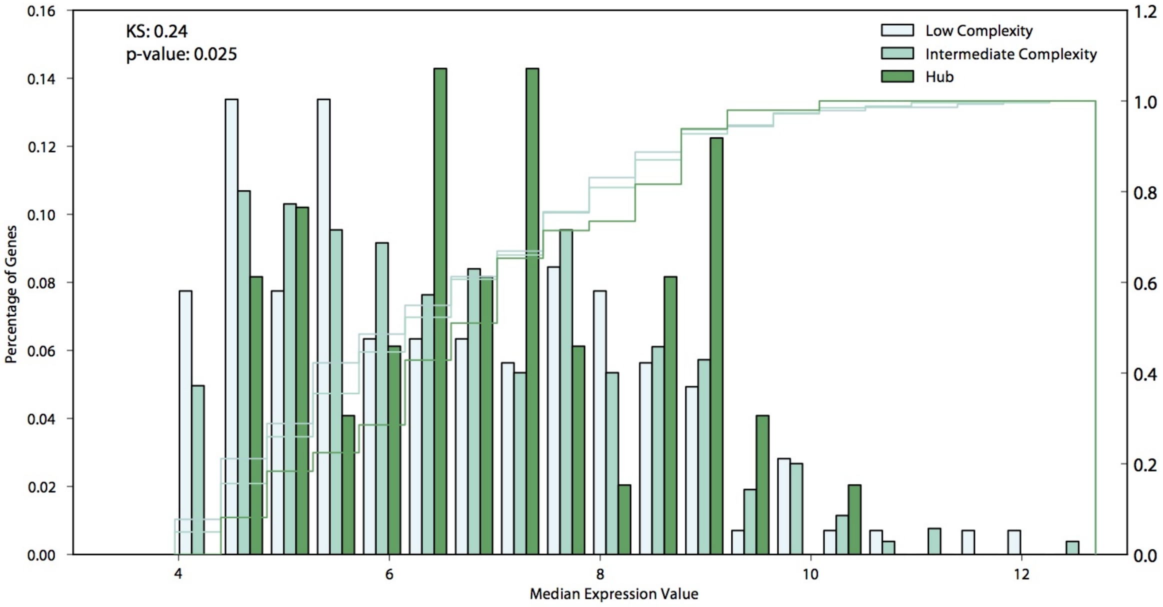


Figure B.10: **Histogram of the median expression values for flowering-associated genes for different series of regulatory complexity.** *KS: Kolmogorov-Smirnov.*

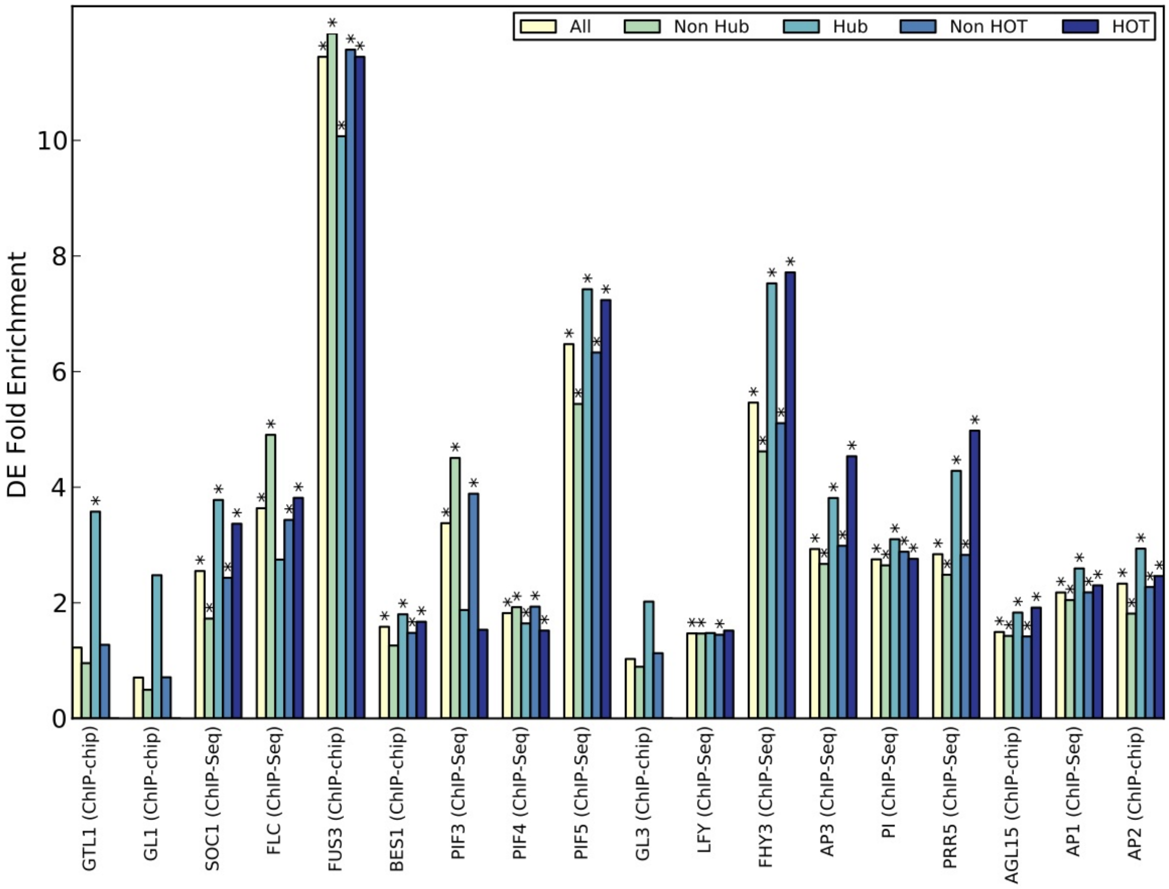


Figure B.11: **Enrichment for differentially expressed (DE) genes in non-HOT- associated and non-hub genes versus HOT-associated and hub genes.**

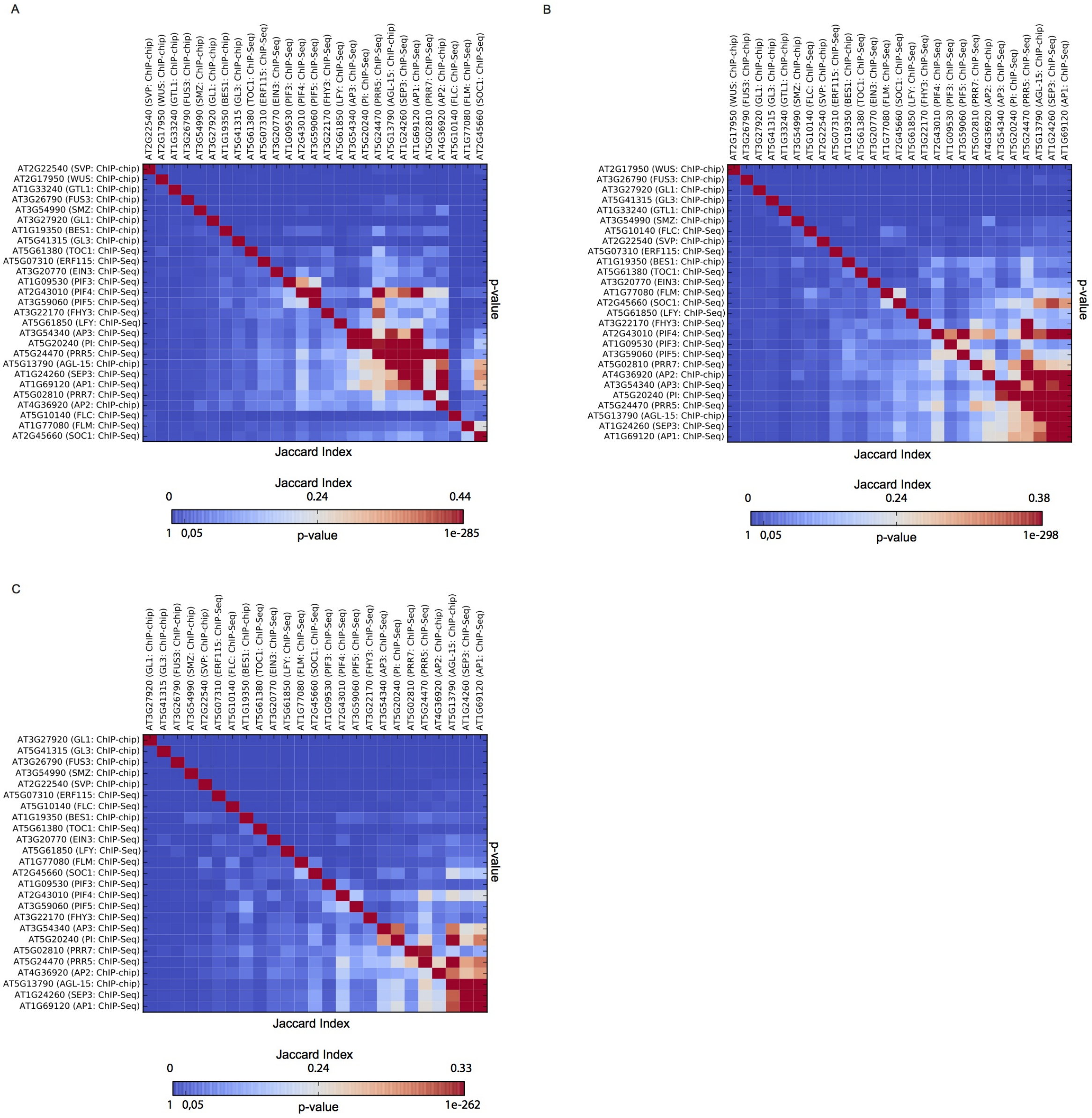


Figure B.12: **Robustness of the clusters TF co-regulation.** *TF cobinding matrix based on common potential target genes, and average-linkage hierarchical clustering based on Jaccard Index for (A) HOT-associated genes (B) ME subnetwork and (C) HC subnetwork. The lower left half displays the Jaccard Index while the upper right displays hypergeometric p-values of overlap between the two sets of bound genes, corrected using the Bonferonni method.*

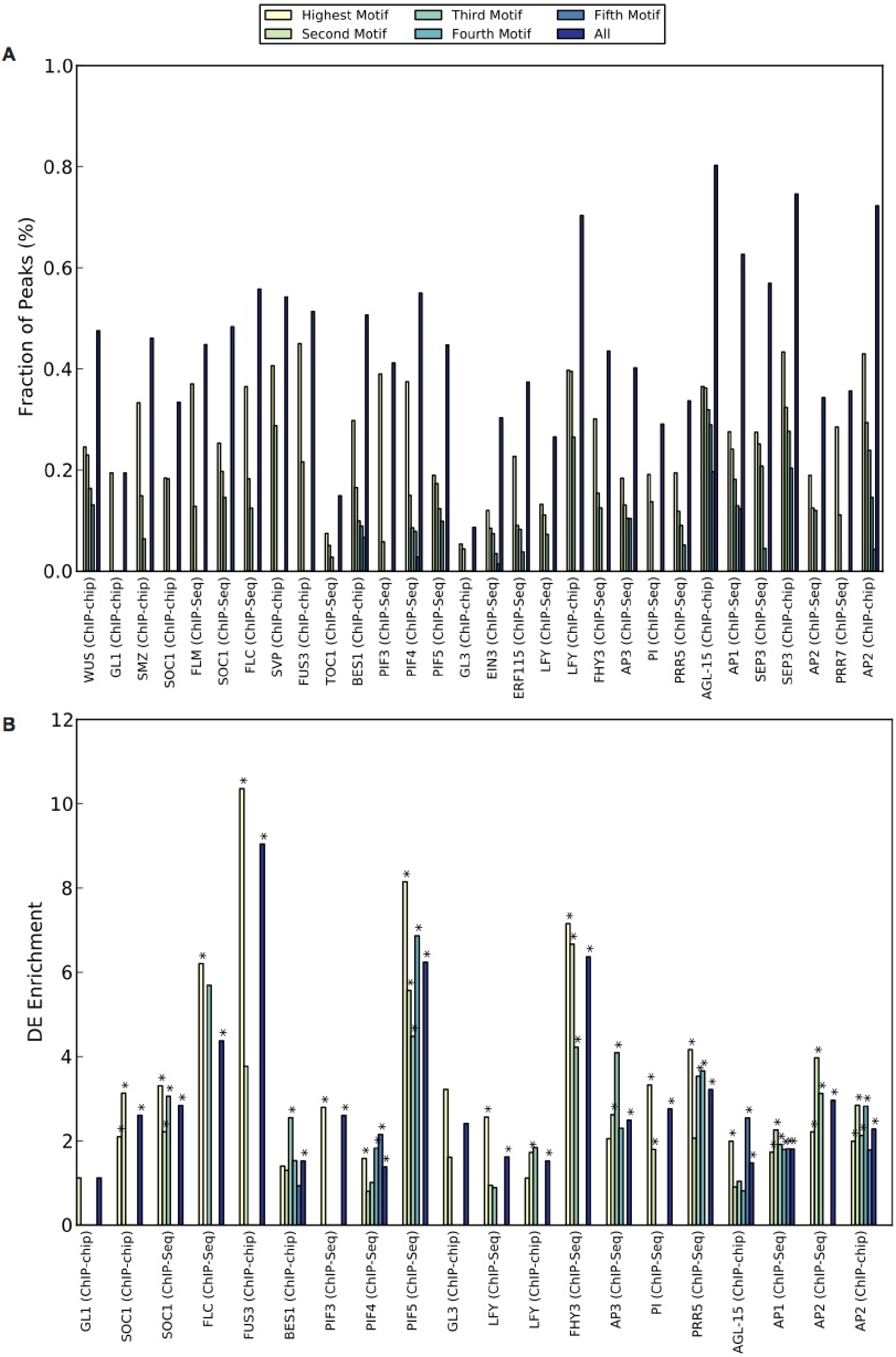


Figure B.13: **DNA motif statistics.** *(A) Fraction of peaks containing each DNA motif per TF. (B) DE enrichment based on the subset of peak-gene annotations uniquely associated with each motif. The term 'Highest Motif' refers to the primary motif while all other series refer to non-primary motifs.*

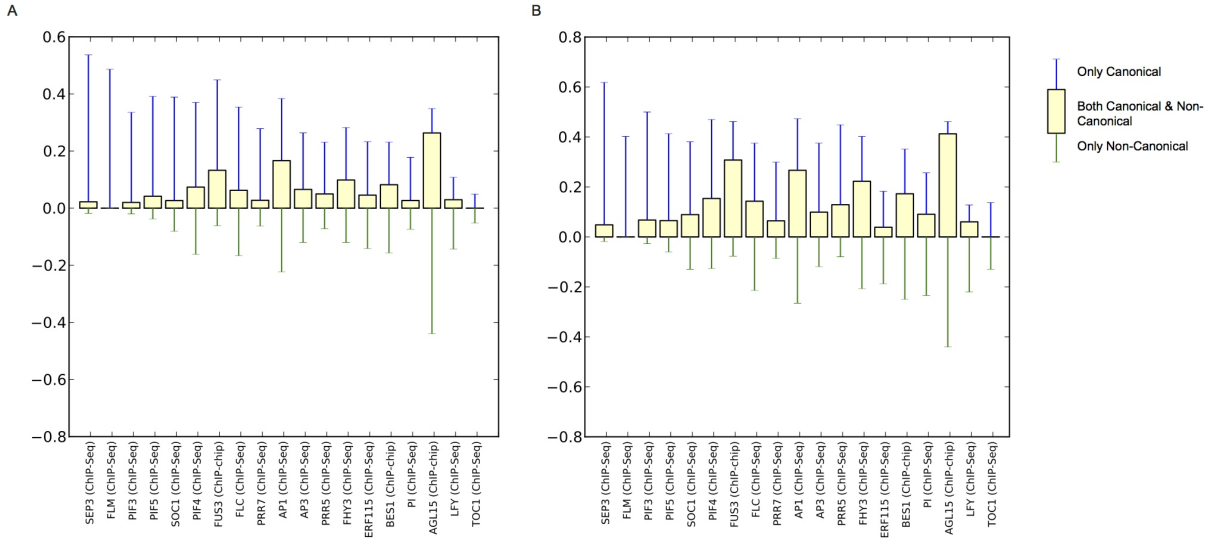


Figure B.14: **Canonical versus non-canonical motifs in (A) non-HOT and (B) HOT regions.** *For each TF blue error flags represent the fraction of peaks with solely a canonical motif, the yellow bar denotes the fraction of peaks with both a canonical and a non-canonical motif instance, and the green error flag indicates the fraction of peaks with exclusively non-canonical motifs.*

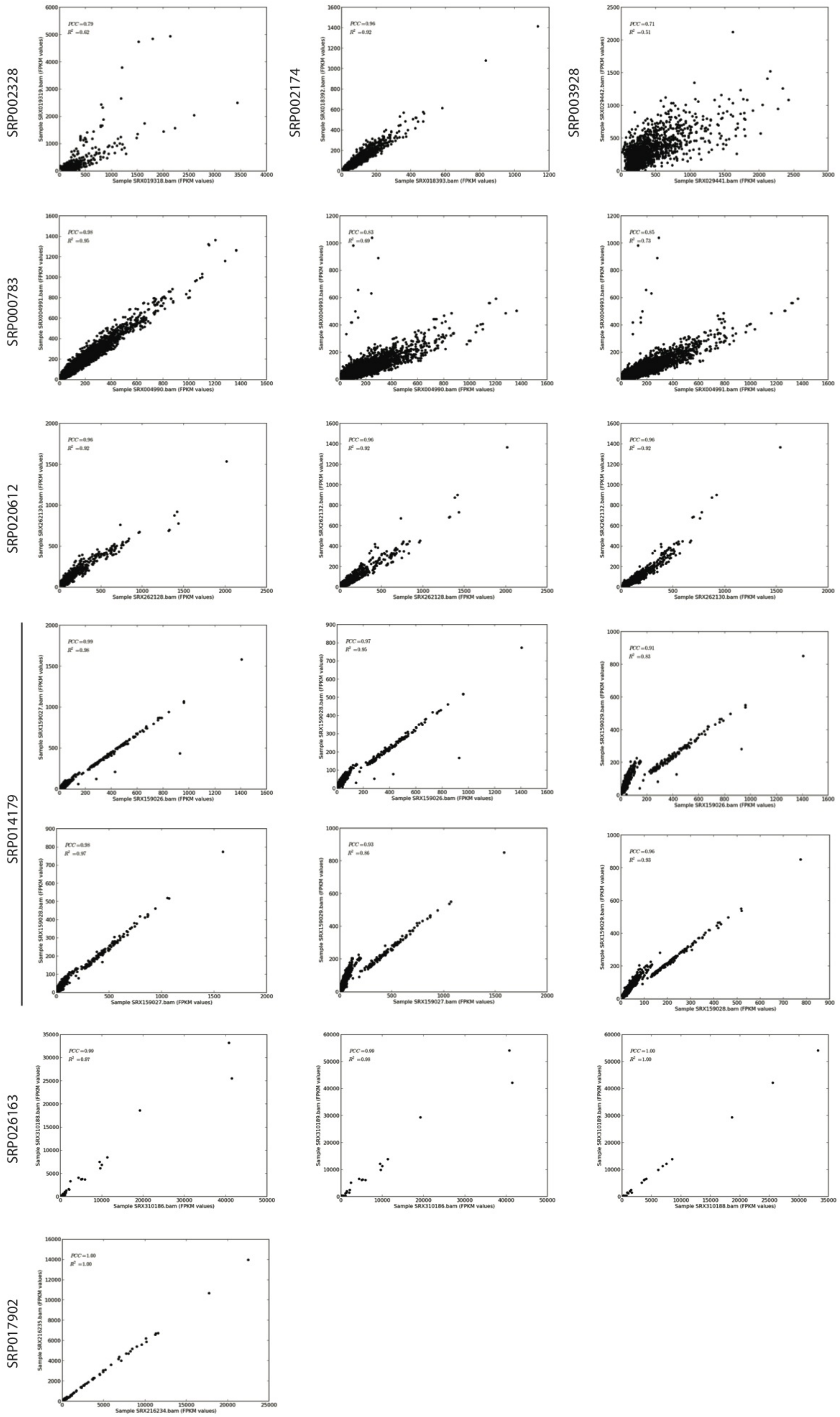
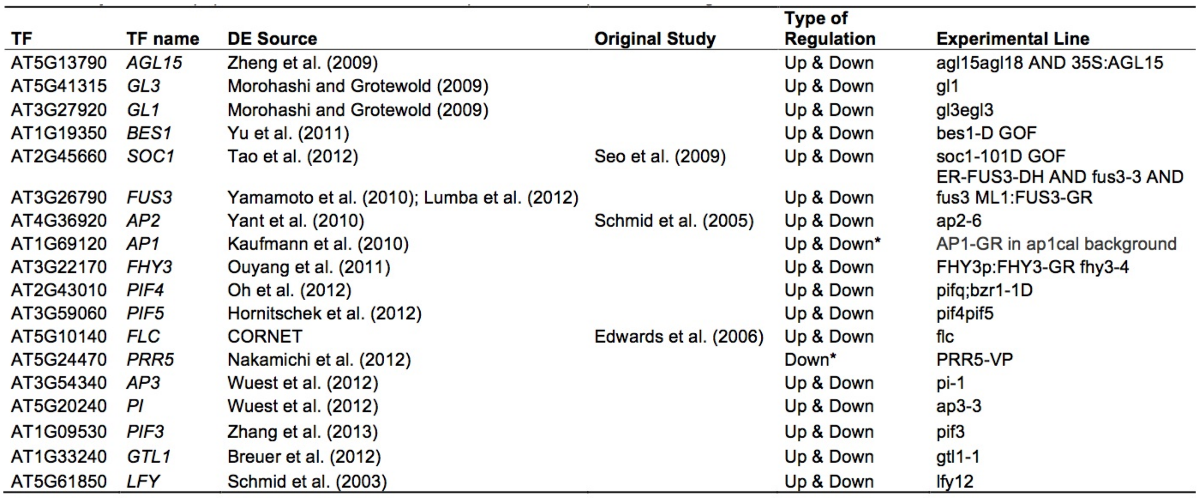


Figure B.15: **Number of modules per gene.** *For each ChIP-Seq data set with replicates, scatter plots are shown of the FPKM values of all peak regions for all pairwise combinations of replicates.*

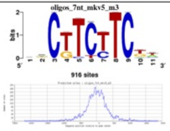
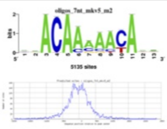
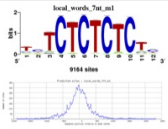
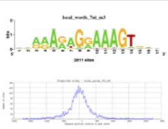
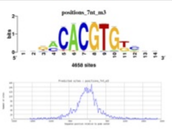
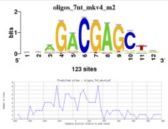
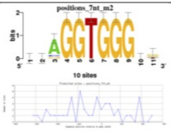
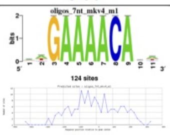
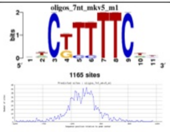
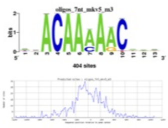
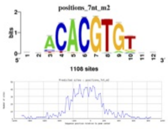
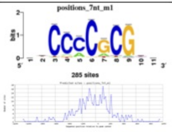
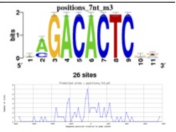
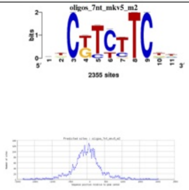
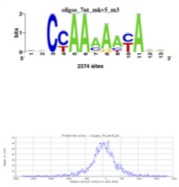
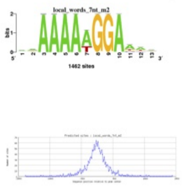
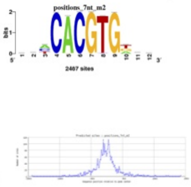
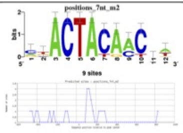
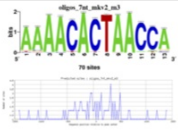
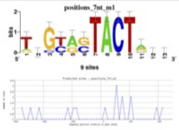
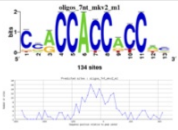
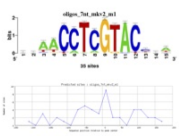
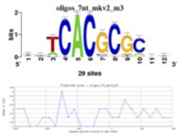
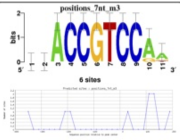
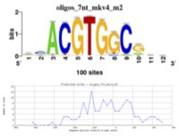
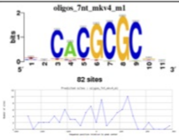
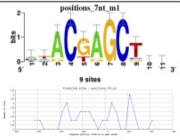
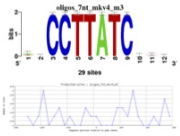
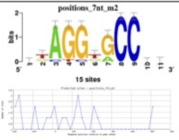
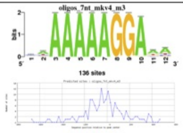
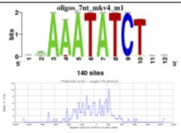
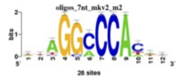
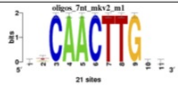
B.2 Supplemental Tables

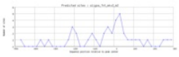
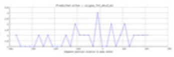
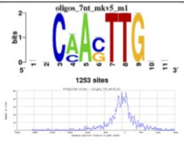
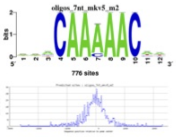
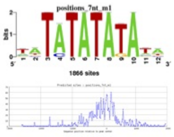
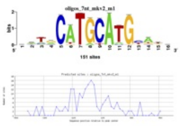
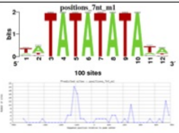
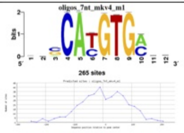
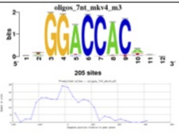
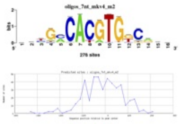
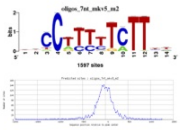
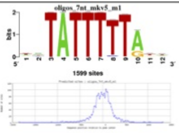
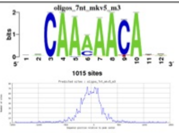
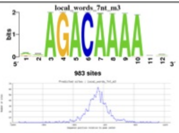
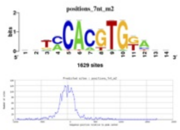
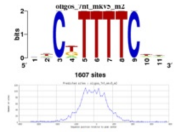
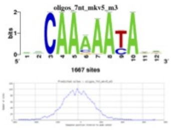
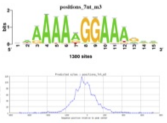
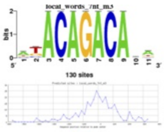
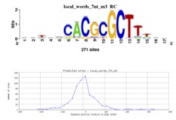
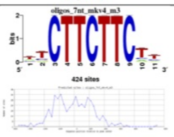
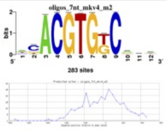
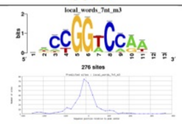
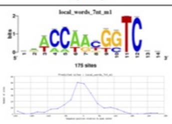
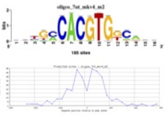
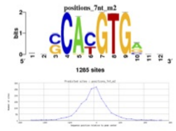
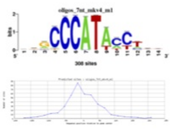
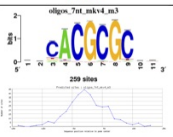
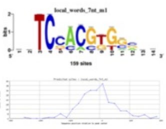
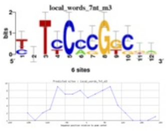
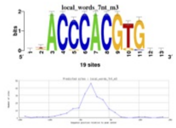
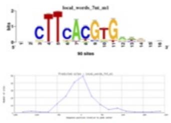
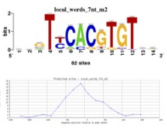
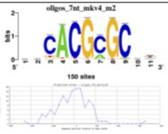
Table B.1: **Arabidopsis TF ChIP Data Sets Used.**



TF	TF name	DE Source	Original Study	Type of Regulation	Experimental Line
AT5G13790	*AGL15*	Zheng et al. (2009)		Up & Down	agl15agl18 AND 35S:AGL15
AT5G41315	*GL3*	Morohashi and Grotewold (2009)		Up & Down	gl1
AT3G27920	*GL1*	Morohashi and Grotewold (2009)		Up & Down	gl3egl3
AT1G19350	*BES1*	Yu et al. (2011)		Up & Down	bes1-D GOF
AT2G45660	*SOC1*	Tao et al. (2012)	Seo et al. (2009)	Up & Down	soc1-101D GOF
AT3G26790	*FUS3*	Yamamoto et al. (2010); Lumba et al. (2012)		Up & Down	ER-FUS3-DH AND fus3-3 AND fus3 ML1:FUS3-GR
AT4G36920	*AP2*	Yant et al. (2010)	Schmid et al. (2005)	Up & Down	ap2-6
AT1G69120	*AP1*	Kaufmann et al. (2010)		Up & Down*	AP1-GR in ap1cal background
AT3G22170	*FHY3*	Ouyang et al. (2011)		Up & Down	FHY3p:FHY3-GR fhy3-4
AT2G43010	*PIF4*	Oh et al. (2012)		Up & Down	pifq;bzr1-1D
AT3G59060	*PIF5*	Hornitschek et al. (2012)		Up & Down	pif4pif5
AT5G10140	*FLC*	CORNET	Edwards et al. (2006)	Up & Down	flc
AT5G24470	*PRR5*	Nakamichi et al. (2012)		Down*	PRR5-VP
AT3G54340	*AP3*	Wuest et al. (2012)		Up & Down	pi-1
AT5G20240	*PI*	Wuest et al. (2012)		Up & Down	ap3-3
AT1G09530	*PIF3*	Zhang et al. (2013)		Up & Down	pif3
AT1G33240	*GTL1*	Breuer et al. (2012)		Up & Down	gtl1-1
AT5G61850	*LFY*	Schmid et al. (2003)		Up & Down	lfy12

Table B.2: **Different significant DNA motifs per TF in order of prevalence.** *Upper row: motif logo; middle row: motif distribution in peak regions; lower row: percentage of peaks with the motif. In the TF column, the information provided are the TF AGI ID and the TF family.*

TF	TF name	Motif 1	Motif2	Motif3	Motif4	Motif 5
AT5G13790 MADS	*AGL15*	 36.55	 36.27	 31.91	 28.96	 19.66
AT5G41315 bHLH	*GL3*	 5.36	 4.40			
AT3G27920 MYB	*GL1*	 19.42				
AT4G36920 AP2-EREBP	*AP2* (chip)	 42.98	 29.40	 23.93	 14.58	 4.35
AT1G24260 MADS	*SEP3* (chip)	 43.34	 32.41	 27.64	 20.40	
AT2G17950 Homeobox	*WUS*	 24.59	 22.95	 16.39	 13.11	
AT3G54990 AP2-EREBP	*SMZ*	 33.33	 14.89	 6.38		
AT1G19350 BZR	*BES1*	 29.8	 16.56	 9.93	 8.94	 6.62
AT2G45660 MADS	*SOC1* (chip)	 18.41	 18.26			
AT2G22540 MADS	*SVP*					

		 40.68	 28.81			
AT5G61850 Orphan	*LFY* (chip)	 39.72	 39.50	 26.50		
AT3G26790 ABI3VP1	*FUS3*	 45.05	 21.62			
AT1G33240 Trihelix	*GTL1*	/ / /				
AT4G36920 AP2-EREBP	*AP2* (Seq)	 18.93	 12.51	 12.04		
AT1G69120 MADS	*AP1*	 27.54	 24.16	 18.15	 12.94	 12.35
AT1G24260 MADS	*SEP3* (Seq)	 27.46	 25.17	 20.80	 4.52	
AT3G22170	*FHY3*	 30.11	 15.40	 12.53		
AT5G61850 Orphan	*LFY* (Seq)	 13.22	 11.09	 7.24		
AT2G43010 bHLH	*PIF4*	 37.49	 14.97	 8.60	 7.90	 2.78
AT3G59060 bHLH	*PIF5*	 18.94	 17.34	 12.40	 9.87	

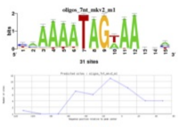
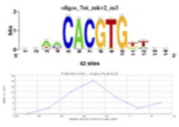
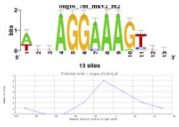
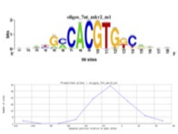
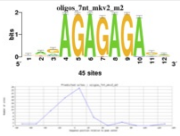
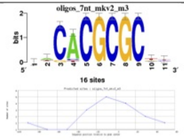
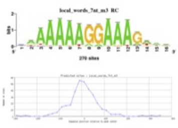
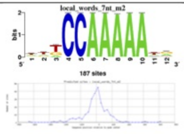
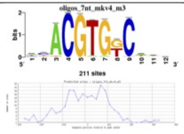
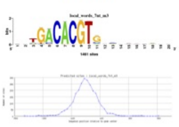
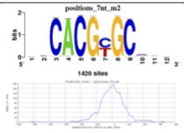
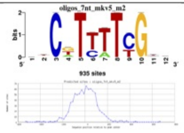
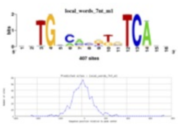
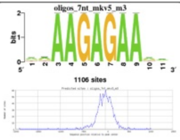
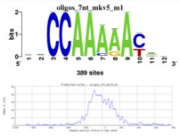
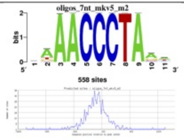
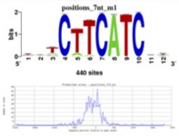
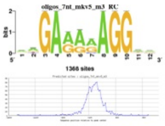
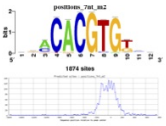
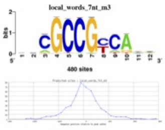
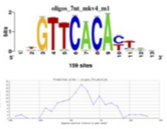
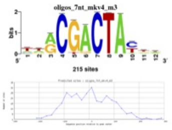
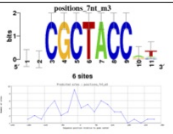
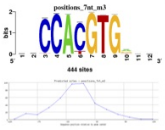
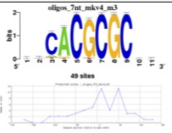
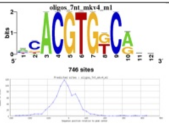
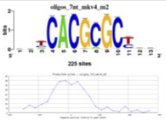
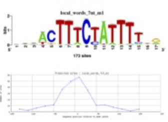
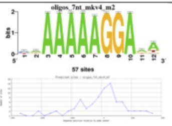
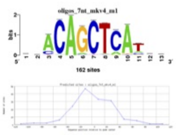
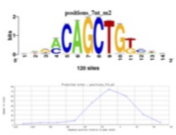
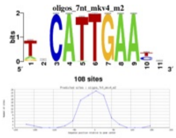
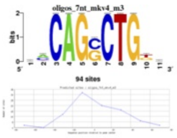
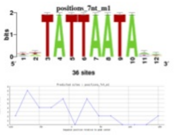
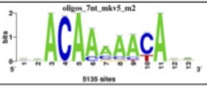
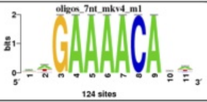
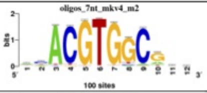
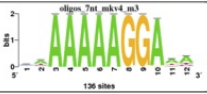
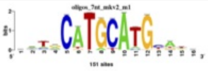
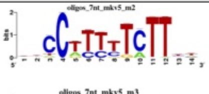
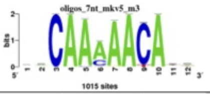
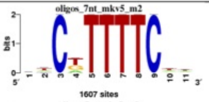
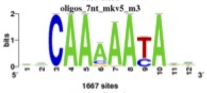
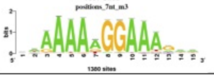
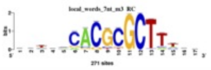
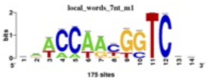
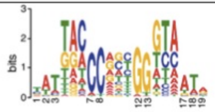
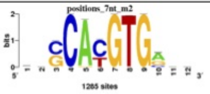
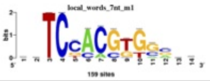
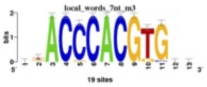
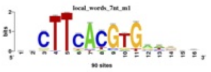
AT5G10140 MADS	*FLC*	 36.54	 18.27	 12.50		
AT5G61380 PRR family	*TOC1*	 7.45	 5.10	 2.75		
AT2G45660 MADS	*SOC1* (Seq)	 25.35	 19.72	 14.56		
AT5G24470 PRR family	*PRR5*	 19.42	 11.87	 9.07	 5.17	
AT3G54340 MADS	*AP3*	 18.38	 13.10	 10.45	 10.40	
AT5G20240 MADS	*PI*	 19.10	 13.73			
AT5G07310 AP2-EREBP	*ERF115*	 22.69	 9.04	 8.30	 3.82	
AT1G09530 bHLH	*PIF3*	 38.99	 5.8			
AT5G02810 PRR	*PRR7*	 28.55	 11.11			
AT1G77080 MADS	*FLM*	 37.06	 12.86			
AT3G20770 EIL	EIN3	 12.06	 8.51	 7.41	 3.47	 1.42

Table B.3: **Motifs from supplemental Table 3 that fit the TF's canonical motif.***Alignments were created based on the motifs on the AGRIS motif database, but references point to evidence of the motif being the canonical motif for the TF, not the necessarily to the reference of the motif specified in the database. / means that no canonical motif could be defined, or was found in our data set.*[a]

TF	TF name	Canonical Motif	Canonical Motif Sequence (Alignment)	Motif Name	Source
AT5G13790 MADS	*AGL15*		ACAAMAACA- CCAAAAATGG	CarG	Tang and Perry (2003)
AT5G41315 bHLH	*GL3*			E-box	Toledo-Ortiz et al. (2003)
AT3G27920 MYB	*GL1*		-----TGTTTTC TAGATTGTTT--	CCA1	Wang et al. (1997)
AT4G36920 AP2-EREBP	*AP2* (chip)	/			
AT2G17950 Homeobox	*WUS*	/			
AT3G54990 AP2-EREBP	*SMZ*	/			
AT1G19350 BZR	*BES1*		-ACGTG CACGTG	E-box	Yu et al. (2011)
AT2G45660 MADS	*SOC1* (chip)		-------AAAAAGGANA- GTTACTAAAAATGGAAAG	CarG	Tilly et al. (1998)
AT2G22540 MADS	*SVP*	/			
AT3G26790 ABI3VP1	*FUS3*		TNCATGCATGNA --CATGCATG--	RY	Monke et al. (2004)
AT1G33240 Trihelix	*GTL1*	/			
AT1G69120 MADS	*AP1*		AAGAAAAAGG CTAAAAATGG	CarG	Tilly et al. (1998)
			-CAAAAACA- CCAAAAATGG	CarG	Riechmann et al. (1996)
AT1G24260 MADS	*SEP3* (Seq)		--GAAAAMG- CCAAAAATGG	CarG	Tilly et al. (1998)
			-CAAAAATA- CCAAAAATGG	CarG	Tilly et al. (1998)
			CTTTCCYTTTT------- CTTTCCATTTTTAGTAAC	CarG	Tilly et al. (1998)
AT3G22170	*FHY3*		CACGCGC	FHY3-FAR1 binding site (FBS)	Lin et al. (2007)
AT5G61850* Orphan	*LFY* (Seq)		 *Image from SELEX exp. from reference*		Moyroud et al. (2011)
AT2G43010 bHLH	*PIF4*		SCACGTGR -CACGTG-	G-box	Toledo-Ortiz et al. (2003)
			TCCACGTGSN --CACGTG--	G-box	Toledo-Ortiz et al. (2003)
AT3G59060 bHLH	*PIF5*		CACGTGGGT CACGTG---	G-box	Toledo-Ortiz et al. (2003)
			CACGTGAAG CACGTG---	G-box	Toledo-Ortiz et al. (2003)

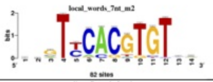
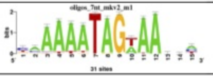
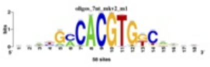
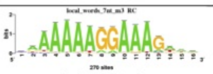
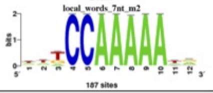
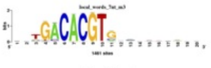
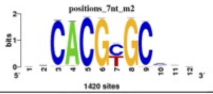
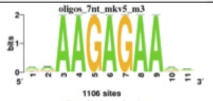
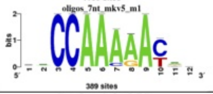
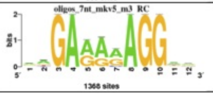
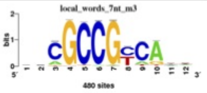
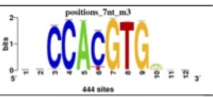
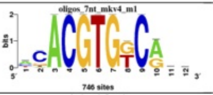
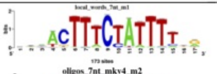
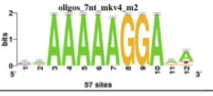
			NTYCACGTGT ---CACGTG-	G-box	Toledo-Ortiz et al. (2003)
AT5G10140 MADS	*FLC*		-------AAAATAGWAANNN GTTACTAAAAATGGAAAG--	CarG	Tilly et al. (1998)
AT5G61380 PRR	*TOC1*		GMCACGTGKC --CACGTG--	E-box	Nakamichi et al. (2010)
AT2G45660 MADS	*SOC1* (Seq)		-------AAAAAGGAAAGW GTTACTAAAAATGGAAAG-	CarG	Tilly et al. (1998)
			KCCAAAAA--- -CCAAAAATGG	CarG	Riechmann et al. (1996)
AT5G24470 PRR	*PRR5*		TGACACGTG ---CACGTG	E-box	Nakamichi et al. (2010)
			CACGYGC CACGTG-	E-box	Nakamichi et al. (2010)
AT3G54340 MADS	*AP3*		--AAGAGAA------- AAAACAGAATAGGAAA	CarG	Ito et al. (2004)
			--GTTTTTGG CCATTTTTGG	CarG	Riechmann et al. (1996)
AT5G20240 MADS	*PI*		CCTYTYTC-- CCATTTTTGG	CarG	Riechmann et al. (1996)
AT5G07310 AP2-EREBP	*ERF115*		TGRCGGCG -GGCGGC-	GCC	Cheng et al. (2013)
AT1G09530 bHLH	*PIF3*			G-box	Toledo-Ortiz et al. (2003)
AT5G02810 PRR	*PRR7*		CACGTGKCA CACGTG---	G-box	Nakamichi et al. (2010)
AT1G77080 MADS	*FLM*		-----YNAAAATAGAAAGT GTTACTAAAAATGGAAAG-	CarG	Tilly et al. (1998)
			-------AAAAAGGANA- GTTACTAAAAATGGAAAG	CarG	Tilly et al. (1998)

Table B.4: **Replicates used for each ChIP-Seq study with replicates.**

TF	TF name	# Replicates	Chosen Replicate	Reference
AT4G36920	AP2	2	SRX019318	Yant et al. (2010)
AT1G69120	AP1	2	SRX018393	Kaufmann et al. (2010)
AT1G24260	SEP3	3	SRX004990	Kaufmann et al. (2009)
AT5G61850	LFY	2	SRX029441	Moyroud et al. (2011)
AT2G45660	SOC1	3	SRX262130	Immink et al. (2012)
AT1G09530	PIF3	4	SRX159027	Zhang et al. (2013)
AT1G77080	FLM	3	SRX310188	Pose et al. (2013)
AT3G20770	EIN3	2	SRP017902	Chang et al. (2013)

Supplemental Data File C

Detection of Conserved Noncoding Sequences in Arabidopsis

C.1 Supplemental Figures

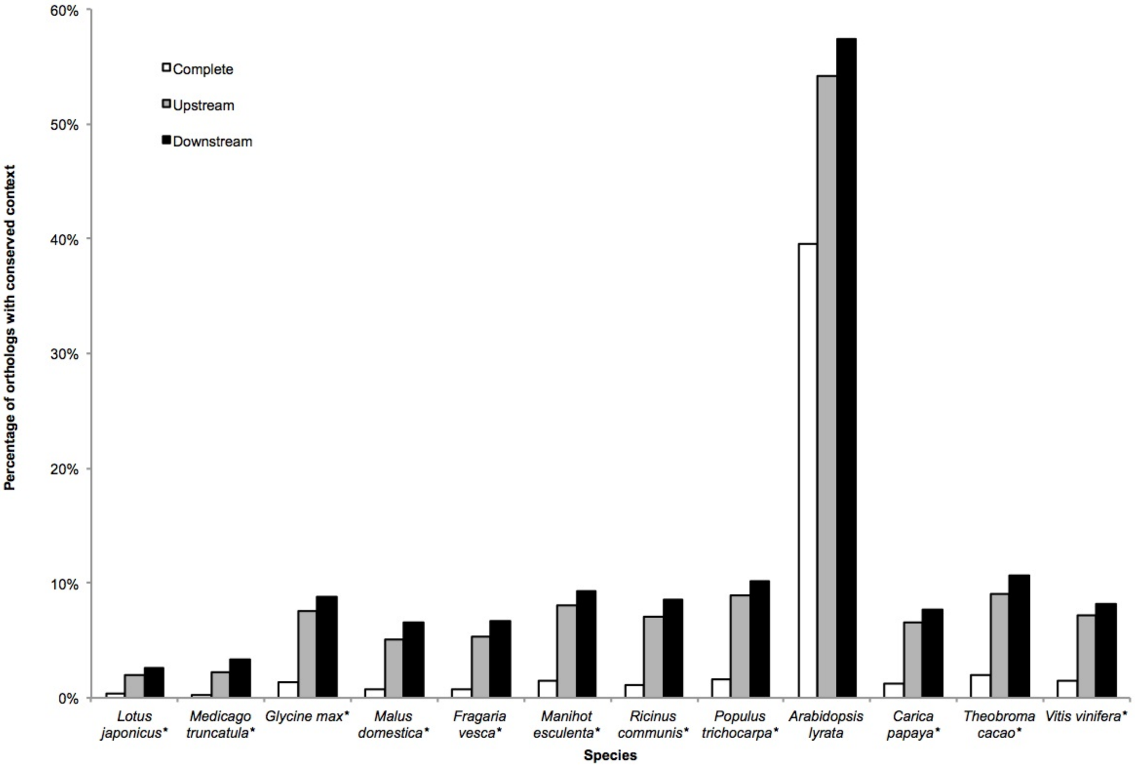


Figure C.1: **Overview of synteny conservation between Arabidopsis and other dicot species.** *This figure shows the percentage of orthologous genes for each Arabidopsis gene for which the flanking genes were conserved by collinearity. Criteria to score collinearity conservation were: 1) whether the genes upstream and/or downstream of the ortholog in the comparator species were orthologous to the genes upstream and/or downstream of the Arabidopsis test gene and 2) whether these orthologs maintained the same relative orientation. In the figure complete (both upstream and downstream)(white box), upstream (grey box) and downstream (black box) conservation is shown. Asterisks indicate species included for phylogenetic footprinting (Arabidopsis lyrata was excluded due to a non-saturated substitution pattern).*

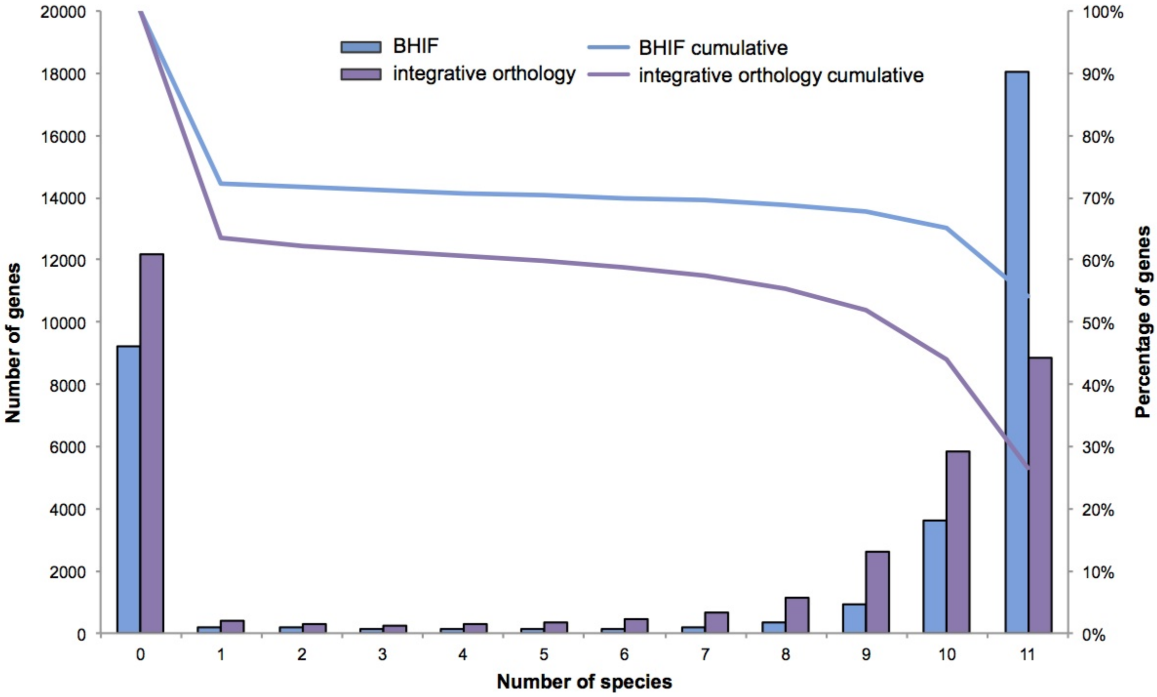


Figure C.2: **Distribution of genes that have orthologs in the dicot comparator species for each orthology detection method.** *The number of Arabidopsis genes with orthologs in different comparator dicot species is depicted for the integrative orthology (purple boxes) and BHIF method (blue boxes), respectively (left y-axis). A cumulative overview is also shown for both methods (purple and blue line, respectively) showing the total percentage of genes for which orthologs could be delineated (right y-axis).*

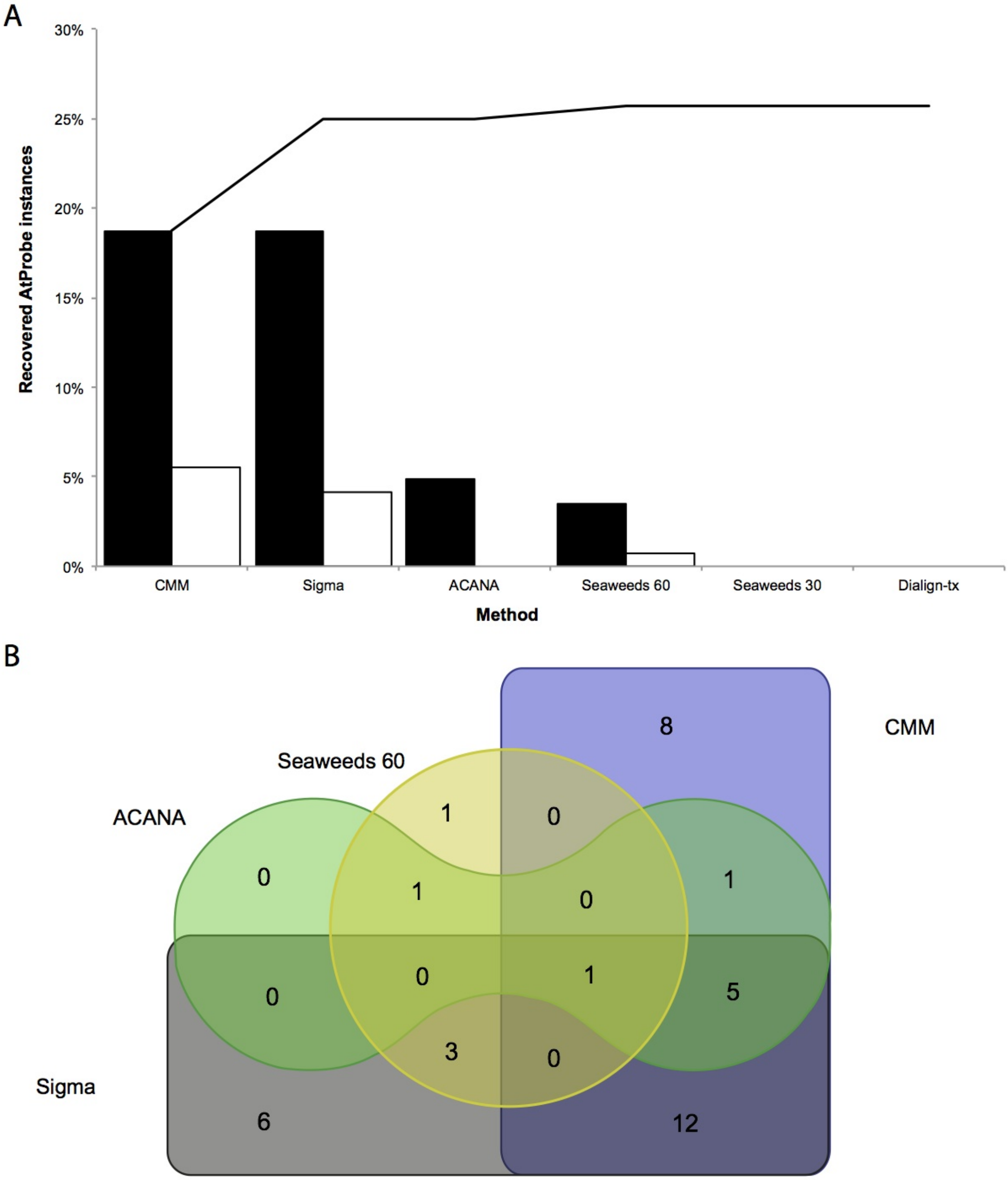


Figure C.3: **Recovery of experimental AtProbe elements using different phylogenetic footprinting approaches.** *A) For the different phylogenetic footprinting approaches developed in this study, the recovery of AtProbe elements was determined. Black boxes show the percentage of recovered elements while white boxes show the percentage of uniquely recovered elements. The black line shows the cumulative recovery over all methods. B) A venn diagram was constructed for the four methods that recovered AtProbe elements. The number of recovered elements for Sigma are displayed in black, for ACANA in green, for Seaweeds 60 in yellow and for CMM in purple.*

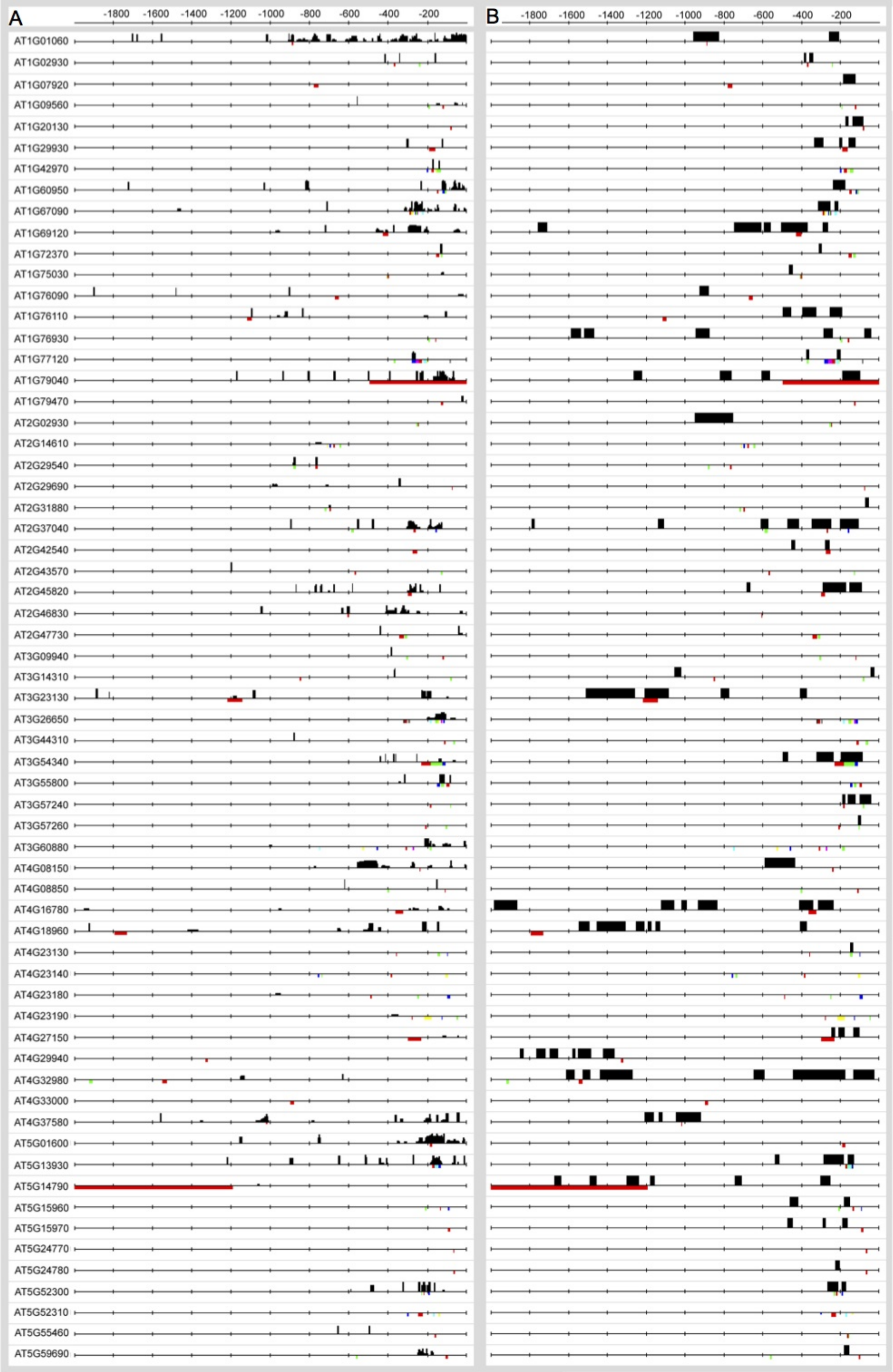


Figure C.4: **Recovery of AtProbe elements for the CNSs described in this paper (A) and by Haudry et al. (2013)(B).** *Black lines denote upstream sequences, colored boxes depict AtProbe elements, and black boxes show significant CNSs.*

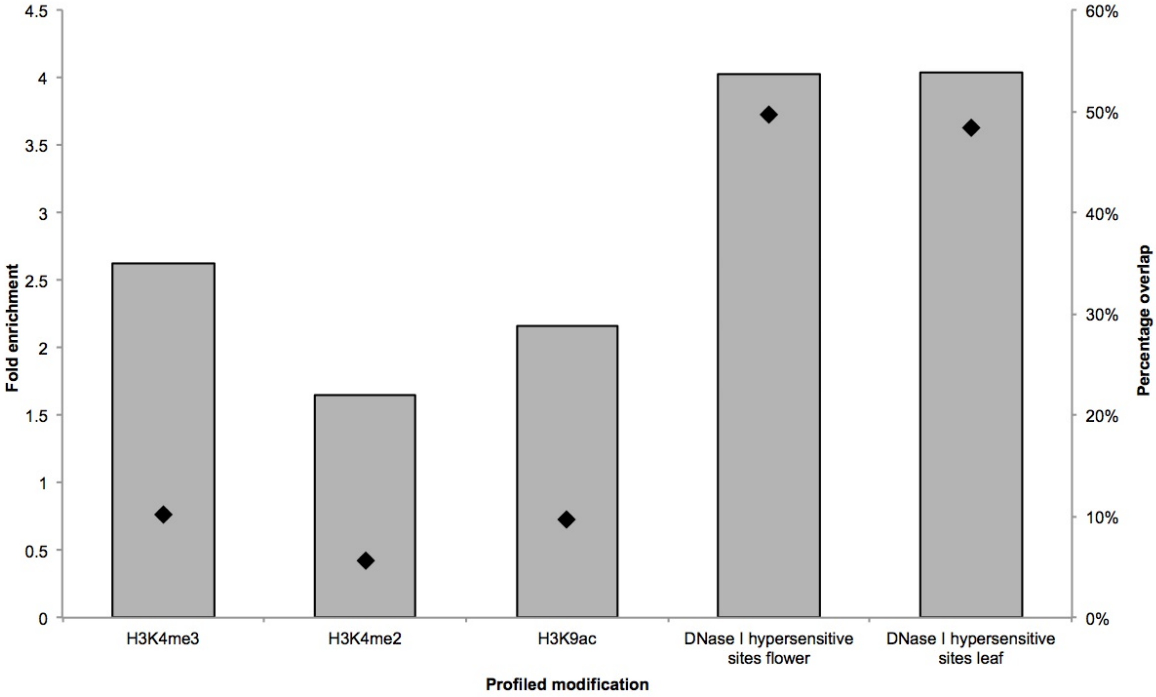


Figure C.5: **Enrichment and overlap of in vivo functional regions with CNSs.** *Grey boxes show the fold enrichment of different histone marks and DH sites. Black diamonds show the percentages of CNSs that overlap with each in vivo functional region dataset.*

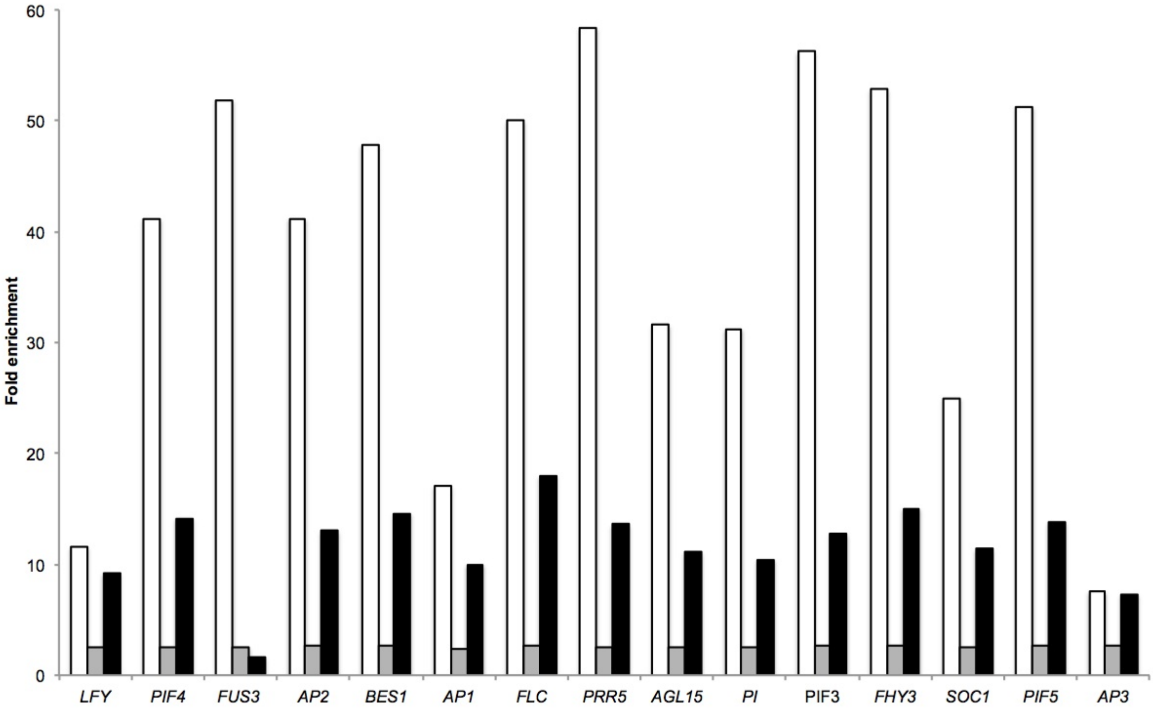


Figure C.6: **Comparison of fold enrichment for in vivo functional binding site regions.** *Fold enrichment for in vivo functional binding sites is shown for our CNSs dataset (white boxes),simple motif mapping (grey boxes) and motif mapping within DH sites (black boxes).*

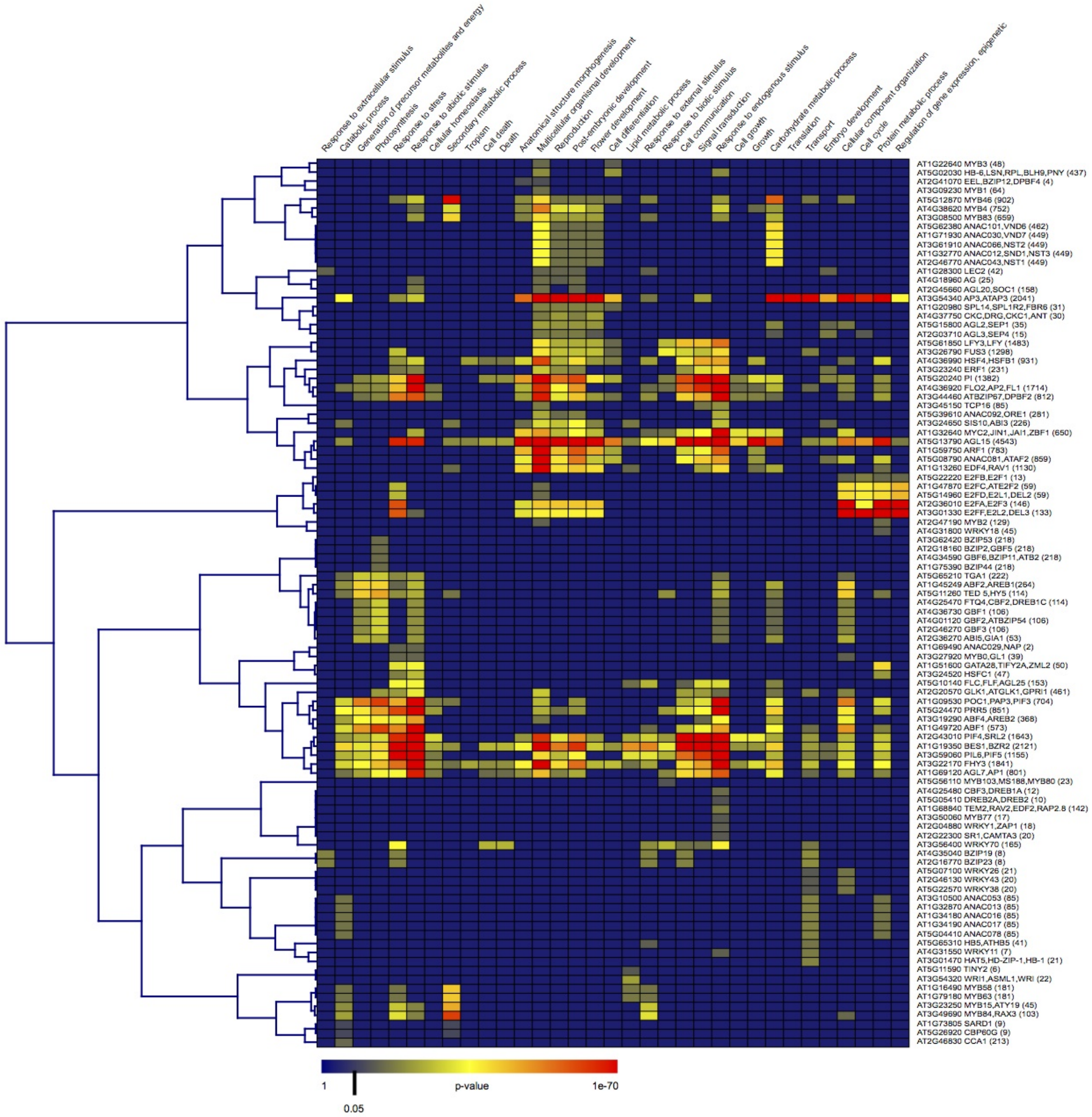


Figure C.7: **GO enrichment for all TF-targets in the predicted GRN.** *A heatmap was generated using Genesis that displays, per TF, the enrichment of target genes towards GO slim annotations (hypergeometric distribution + Bonferroni correction). The number of target genes for each TF is shown in parenthesis. The color gradient shows the p-values of the different enriched gene sets.*

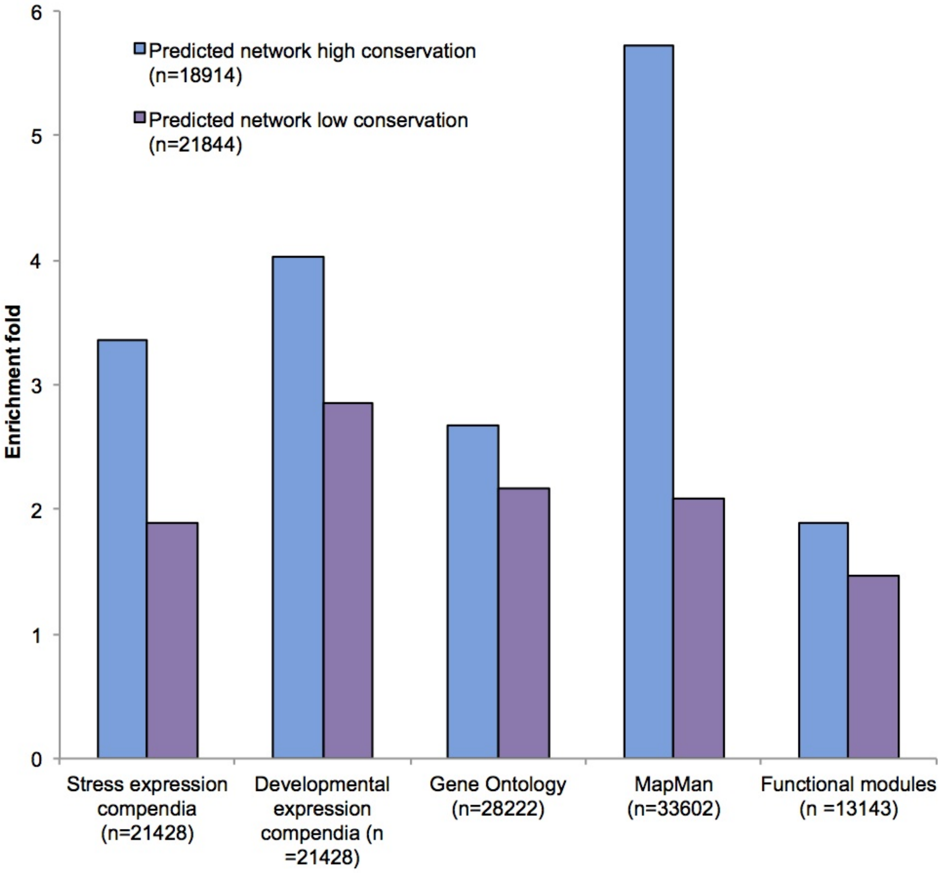
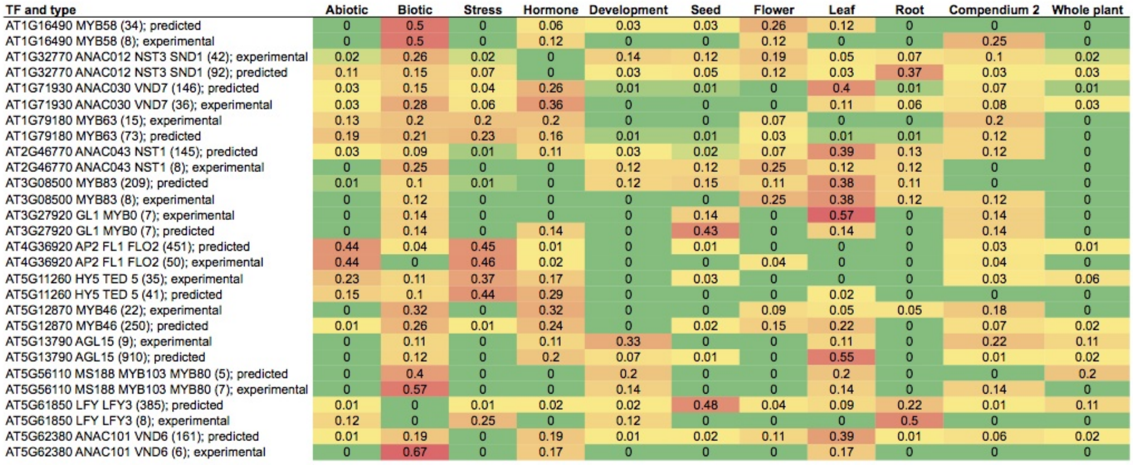


Figure C.8: **Evaluation of the biological relevance of highly and moderately conserved interactions using the biological validation metrics.** *Comparison of the five biological metrics for the predicted sub-networks with highly (blue boxes, >6 species) and moderately (purple boxes, 2-6 species) conserved interactions. Fold enrichments are shown for the CORNET stress and developmental expression compendia, Gene Ontology annotations, Mapman annotations and Functional modules. All reported fold enrichments are significant (p-value < 0.05).*



TF and type	Abiotic	Biotic	Stress	Hormone	Development	Seed	Flower	Leaf	Root	Compendium 2	Whole plant
AT1G16490 MYB58 (34); predicted	0	0.5	0	0.06	0.03	0.03	0.26	0.12	0	0	0
AT1G16490 MYB58 (8); experimental	0	0.5	0	0.12	0	0	0.12	0	0	0.25	0
AT1G32770 ANAC012 NST3 SND1 (42); experimental	0.02	0.26	0.02	0	0.14	0.12	0.19	0.05	0.07	0.1	0.02
AT1G32770 ANAC012 NST3 SND1 (92); predicted	0.11	0.15	0.07	0	0.03	0.05	0.12	0.03	0.37	0.03	0.03
AT1G71930 ANAC030 VND7 (146); predicted	0.03	0.15	0.04	0.26	0.01	0.01	0	0.4	0.01	0.07	0.01
AT1G71930 ANAC030 VND7 (36); experimental	0.03	0.28	0.06	0.36	0	0	0	0.11	0.06	0.08	0.03
AT1G79180 MYB63 (15); experimental	0.13	0.2	0.2	0.2	0	0	0.07	0	0	0.2	0
AT1G79180 MYB63 (73); predicted	0.19	0.21	0.23	0.16	0.01	0.01	0.03	0.01	0.01	0.12	0
AT2G46770 ANAC043 NST1 (145); predicted	0.03	0.09	0.01	0.11	0.03	0.02	0.07	0.39	0.13	0.12	0
AT2G46770 ANAC043 NST1 (8); experimental	0	0.25	0	0	0.12	0.12	0.25	0.12	0.12	0	0
AT3G08500 MYB83 (209); predicted	0.01	0.1	0.01	0	0.12	0.15	0.11	0.38	0.11	0	0
AT3G08500 MYB83 (8); experimental	0	0.12	0	0	0	0	0.25	0.38	0.12	0.12	0
AT3G27920 GL1 MYB0 (7); experimental	0	0.14	0	0	0	0.14	0	0.57	0	0.14	0
AT3G27920 GL1 MYB0 (7); predicted	0	0.14	0	0.14	0	0.43	0	0.14	0	0.14	0
AT4G36920 AP2 FL1 FLO2 (451); predicted	0.44	0.04	0.45	0.01	0	0.01	0	0	0	0.03	0.01
AT4G36920 AP2 FL1 FLO2 (50); experimental	0.44	0	0.46	0.02	0	0	0.04	0	0	0.04	0
AT5G11260 HY5 TED 5 (35); experimental	0.23	0.11	0.37	0.17	0	0.03	0	0	0	0.03	0.06
AT5G11260 HY5 TED 5 (41); predicted	0.15	0.1	0.44	0.29	0	0	0	0.02	0	0	0
AT5G12870 MYB46 (22); experimental	0	0.32	0	0.32	0	0	0.09	0.05	0.05	0.18	0
AT5G12870 MYB46 (250); predicted	0.01	0.26	0.01	0.24	0	0.02	0.15	0.22	0	0.07	0.02
AT5G13790 AGL15 (9); experimental	0	0.11	0	0.11	0.33	0	0	0.11	0	0.22	0.11
AT5G13790 AGL15 (910); predicted	0	0.12	0	0.2	0.07	0.01	0	0.55	0	0.01	0.02
AT5G56110 MS188 MYB103 MYB80 (5); predicted	0	0.4	0	0	0.2	0	0	0.2	0	0	0.2
AT5G56110 MS188 MYB103 MYB80 (7); experimental	0	0.57	0	0	0.14	0	0	0.14	0	0.14	0
AT5G61850 LFY LFY3 (385); predicted	0.01	0	0.01	0.02	0.02	0.48	0.04	0.09	0.22	0.01	0.11
AT5G61850 LFY LFY3 (8); experimental	0.12	0	0.25	0	0.12	0	0	0	0.5	0	0
AT5G62380 ANAC101 VND6 (161); predicted	0.01	0.19	0	0.19	0.01	0.02	0.11	0.39	0.01	0.06	0.02
AT5G62380 ANAC101 VND6 (6); experimental	0	0.67	0	0.17	0	0	0	0.17	0	0	0

Figure C.9: **Comparison between experimental and predicted GRN of co-expressed target genes in different conditions.** *The fraction of target genes showing specific co-expression in each condition is displayed. The color gradient shows the fractions of the target genes. The total number of target genes showing specific co-expression for each TF is shown in parenthesis.*

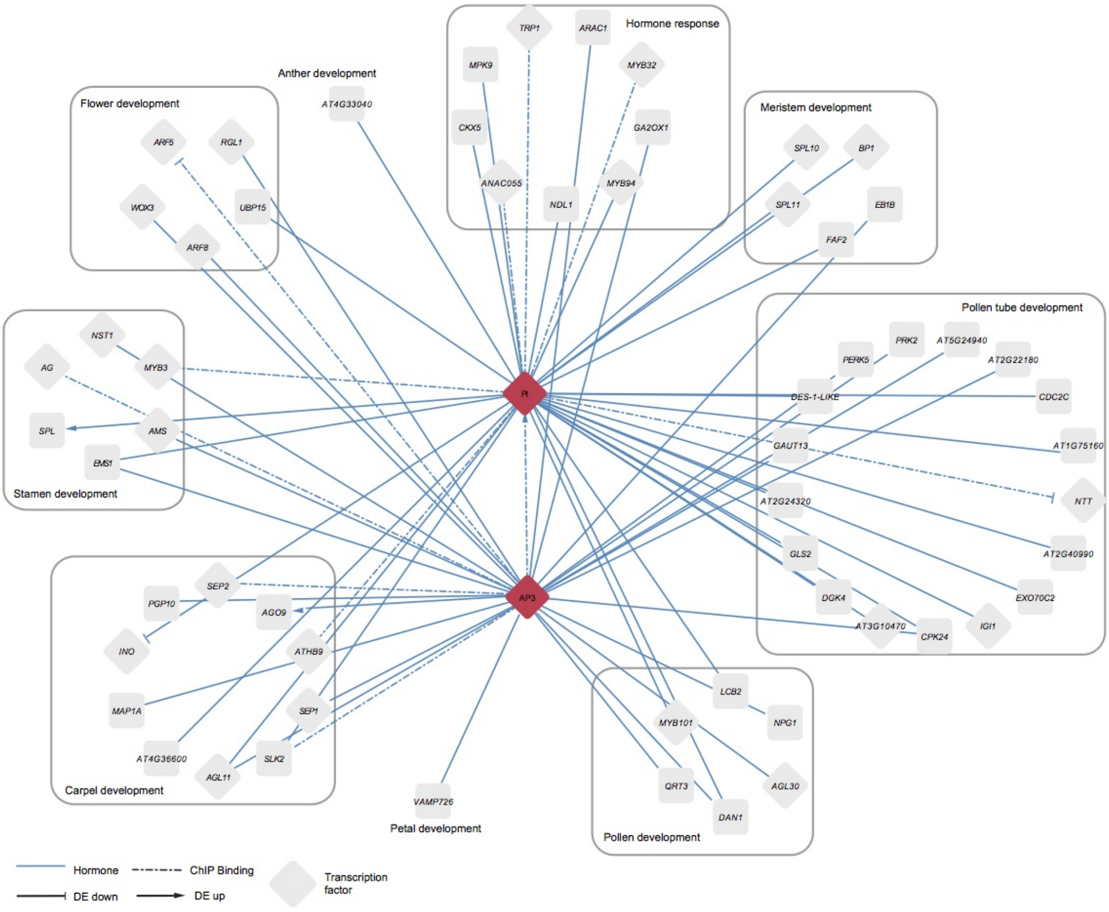


Figure C.10: **A condition-specific GRN for PI and AP3 based on hormone-specific TF-target co-expression edges.** *Genes that have GO annotations related to flower development are displayed. ChIP-bound regions associated with the target gene are shown as dashed lines while differentially expressed genes are shown by an arrowhead for up-regulation and by a vertical line for down-regulation, respectively. Red diamonds are the source TFs, grey diamonds are target genes that are TFs and rounded rectangles are other target genes. Rounded boxes depict different GO biological processes.*

Bibliography

[1] A. K. Spartz and W. M. Gray. Plant hormone receptors: new perceptions. *Genes Dev*, 22(16):2139–48, 2008. ISSN 0890-9369 (Print) 0890-9369 (Linking). doi: 10.1101/gad.1693208. URL http://www.ncbi.nlm.nih.gov/pubmed/18708574http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2735353/pdf/2139.pdf. Spartz, Angela K Gray, William M eng R01 GM067203/GM/NIGMS NIH HHS/ R01 GM067203-04/GM/NIGMS NIH HHS/ Research Support, N.I.H., Extramural Research Support, U.S. Gov't, Non-P.H.S. Review 2008/08/19 09:00 Genes Dev. 2008 Aug 15;22(16):2139-48. doi: 10.1101/gad.1693208.

[2] A. Walton, E. Stes, I. De Smet, S. Goormachtig, and K. Gevaert. Plant hormone signalling through the eye of the mass spectrometer. *Proteomics*, 2014. ISSN 1615-9861 (Electronic) 1615-9853 (Linking). doi: 10.1002/pmic.201400403. URL http://www.ncbi.nlm.nih.gov/pubmed/25404421http://onlinelibrary.wiley.com/doi/10.1002/pmic.201400403/abstract. Walton, Alan Stes, Elisabeth De Smet, Ive Goormachtig, Sofie Gevaert, Kris ENG 2014/11/19 06:00 Proteomics. 2014 Nov 18. doi: 10.1002/pmic.201400403.

[3] Lincoln Taiz and Eduardo Zeiger. *Plant Physiology*. Sinauer Associates, Inc., Sunderland, Massachusetts, USA, 2006. ISBN 0878938567.

[4] M. Sugiura. Plant in vitro transcription systems. *Annu. Rev. Plant Physiol. Plant Mol. Biol.*, 48:383–398, 1997.

[5] X. Zhou, J. Ruan, G. Wang, and W. Zhang. Characterization and identification of microrna core promoters in four model species. *PLoS Comput Biol*, 3(3):e37, 2007. ISSN 1553-7358 (Electronic) 1553-734X (Linking). doi: 10.1371/journal.pcbi.0030037. URL http://www.ncbi.nlm.nih.gov/pubmed/17352530. Zhou, Xuefeng Ruan, Jianhua Wang, Guandong Zhang, Weixiong eng Research Support, Non-U.S. Gov't Research Support, U.S. Gov't, Non-P.H.S. 2007/03/14 09:00 PLoS Comput Biol. 2007 Mar 9;3(3):e37. Epub 2007 Jan 9.

[6] Y. Lee, M. Kim, J. Han, K. H. Yeom, S. Lee, S. H. Baek, and V. N. Kim. Microrna genes are transcribed by rna polymerase ii. *EMBO J*, 23(20):4051–60, 2004. ISSN 0261-4189 (Print) 0261-4189 (Linking). doi: 10.1038/sj.emboj.7600385. URL http://www.ncbi.nlm.nih.gov/pubmed/15372072. Lee, Yoontae Kim, Minju Han, Jinju Yeom, Kyu-Hyun Lee, Sanghyuk Baek, Sung Hee Kim, V Narry eng Research Support, Non-U.S. Gov't England 2004/09/17 05:00 EMBO J. 2004 Oct 13;23(20):4051-60. Epub 2004 Sep 16.

[7] K. D. Grasser. Emerging role for transcript elongation in plant development. *Trends Plant Sci*, 10(10):484–90, 2005. ISSN 1360-1385 (Print) 1360-1385 (Linking). doi: S1360-1385(05)00198-6[pii]10.1016/j.tplants.2005.08.004.

[8] C. Molina and E. Grotewold. Genome wide analysis of arabidopsis core promoters. *BMC Genomics*, 6(1):25, 2005. ISSN 1471-2164 (Electronic) 1471-2164 (Linking). doi: 1471-2164-6-25[pii]10.1186/1471-2164-6-25.

[9] X. Gu, C. Le, Y. Wang, Z. Li, D. Jiang, Y. Wang, and Y. He. Arabidopsis flc clade members form flowering-repressor complexes coordinating responses to endogenous and environmental cues. *Nat Commun*, 4:1947, 2013. ISSN 2041-1723 (Electronic) 2041-1723 (Linking). doi: 10.1038/ncomms2947. URL http://www.ncbi.nlm.nih.gov/pubmed/23770815. Gu, Xiaofeng Le, Chau Wang, Yizhong Li, Zicong Jiang, Danhua Wang, Yuqi He, Yuehui eng Research Support, Non-U.S. Gov't England 2013/06/19 06:00 Nat Commun. 2013;4:1947. doi: 10.1038/ncomms2947.

[10] J. Heyman, T. Cools, F. Vandenbussche, K. S. Heyndrickx, J. Van Leene, I. Vercauteren, S. Vanderauwera, K. Vandepoele, G. De Jaeger, D. Van Der Straeten, and L. De Veylder. Erf115 controls root quiescent center cell division and stem cell replenishment. *Science*, 342(6160): 860–3, 2013. ISSN 1095-9203 (Electronic) 0036-8075 (Linking). doi: 10.1126/science.1240667. URL http://www.ncbi.nlm.nih.gov/pubmed/24158907http://www.sciencemag.org/content/342/6160/860http://www.sciencemag.org/content/342/6160/860.full.pdf. Heyman, Jefri Cools, Toon Vandenbussche, Filip Heyndrickx, Ken S Van Leene, Jelle Vercauteren, Ilse Vanderauwera, Sandy Vandepoele, Klaas De Jaeger, Geert Van Der Straeten, Dominique De Veylder, Lieven eng New York, N.Y. Science. 2013 Nov 15;342(6160):860-3. doi: 10.1126/science.1240667. Epub 2013 Oct 24.

[11] J. B. Veyrieras, S. Kudaravalli, S. Y. Kim, E. T. Dermitzakis, Y. Gilad, M. Stephens, and J. K. Pritchard. High-resolution mapping of expression-qtls yields insight into human gene regulation. *PLoS Genet*, 4(10):e1000214, 2008. ISSN 1553-7404 (Electronic) 1553-7390 (Linking). doi: 10.1371/journal.pgen.1000214. URL http://www.ncbi.nlm.nih.gov/pubmed/18846210. Veyrieras, Jean-Baptiste Kudaravalli, Sridhar Kim, Su Yeon Dermitzakis, Emmanouil T Gilad, Yoav Stephens, Matthew Pritchard, Jonathan K eng GM077959/GM/NIGMS NIH HHS/ HG002772/HG/NHGRI NIH HHS/ HG02585-01/HG/NHGRI NIH HHS/ Howard Hughes Medical Institute/ Research Support, N.I.H., Extramural Research Support, Non-U.S. Gov't 2008/10/11 09:00 PLoS Genet. 2008 Oct;4(10):e1000214. doi: 10.1371/journal.pgen.1000214. Epub 2008 Oct 10.

[12] G. Orphanides, T. Lagrange, and D. Reinberg. The general transcription factors of rna polymerase ii. *Genes Dev*, 10(21):2657–83, 1996. ISSN 0890-9369 (Print) 0890-9369 (Linking). URL http://www.ncbi.nlm.nih.gov/pubmed/8946909. Orphanides, G Lagrange, T Reinberg, D eng Review 1996/11/01 Genes Dev. 1996 Nov 1;10(21):2657-83.

[13] A. Para, Y. Li, A. Marshall-Colon, K. Varala, N. J. Francoeur, T. M. Moran, M. B. Edwards, C. Hackley, B. O. Bargmann, K. D. Birnbaum, W. R. McCombie, G. Krouk, and G. M. Coruzzi. Hit-and-run transcriptional control by bzip1 mediates rapid nutrient signaling in arabidopsis. *Proc Natl Acad Sci U S A*, 2014. ISSN 1091-6490 (Electronic) 0027-8424 (Linking). doi: 10.1073/pnas.1404657111.

[14] F. Roudier, I. Ahmed, C. Berard, A. Sarazin, T. Mary-Huard, S. Cortijo, D. Bouyer, E. Caillieux, E. Duvernois-Berthet, L. Al-Shikhley, L. Giraut, B. Despres, S. Drevensek, F. Barneche, S. Derozier, V. Brunaud, S. Aubourg, A. Schnittger, C. Bowler, M. L. Martin-Magniette, S. Robin, M. Caboche, and V. Colot. Integrative epigenomic mapping defines four main chromatin states in arabidopsis. *EMBO J*, 30 (10):1928–38, 2011. ISSN 1460-2075 (Electronic) 0261-4189 (Linking). doi: 10.1038/emboj.2011.103.

[15] J. L. Riechmann and O. J. Ratcliffe. A genomic perspective on plant transcription factors. *Curr Opin Plant Biol*, 3(5):423–34, 2000. ISSN 1369-5266 (Print) 1369-5266 (Linking). doi: S1369-5266(00)00107-2[pii].

[16] M. M. Babu, N. M. Luscombe, L. Aravind, M. Gerstein, and S. A. Teichmann. Structure and evolution of transcriptional

regulatory networks. *Curr Opin Struct Biol*, 14(3):283–91, 2004. ISSN 0959-440X (Print) 0959-440X (Linking). doi: 10.1016/j.sbi.2004.05.004S0959440X04000788[pii].

[17] S. K. Palaniswamy, S. James, H. Sun, R. S. Lamb, R. V. Davuluri, and E. Grotewold. Agris and atregnet. a platform to link cis-regulatory elements and transcription factors into regulatory networks. *Plant physiology*, 140(3):818–29, 2006. ISSN 0032-0889 (Print) 0032-0889 (Linking). doi: 140/3/818[pii] 10.1104/pp.105.072280.

[18] J. Jin, H. Zhang, L. Kong, G. Gao, and J. Luo. Planttfdb 3.0: a portal for the functional and evolutionary study of plant transcription factors. *Nucleic Acids Res*, 42(Database issue):D1182–7, 2014. ISSN 1362-4962 (Electronic) 0305-1048 (Linking). doi: 10.1093/nar/gkt1016. URL `http://www.ncbi.nlm.nih.gov/pubmed/24174544`. Jin, Jinpu Zhang, He Kong, Lei Gao, Ge Luo, Jingchu eng Research Support, Non-U.S. Gov't England 2013/11/01 06:00 Nucleic Acids Res. 2014 Jan;42(Database issue):D1182-7. doi: 10.1093/nar/gkt1016. Epub 2013 Oct 29.

[19] K. Aoki, Y. Ogata, and D. Shibata. Approaches for extracting practical information from gene co-expression networks in plant biology. *Plant Cell Physiol*, 48(3):381–90, 2007. ISSN 0032-0781 (Print) 0032-0781 (Linking). doi: pcm013[pii]10.1093/pcp/pcm013.

[20] J. D. Han, N. Bertin, T. Hao, D. S. Goldberg, G. F. Berriz, L. V. Zhang, D. Dupuy, A. J. Walhout, M. E. Cusick, F. P. Roth, and M. Vidal. Evidence for dynamically organized modularity in the yeast protein-protein interaction network. *Nature*, 430 (6995):88–93, 2004. ISSN 1476-4687 (Electronic) 0028-0836 (Linking). doi: 10.1038/nature02555nature02555[pii].

[21] B. E. Bernstein, T. S. Mikkelsen, X. Xie, M. Kamal, D. J. Huebert, J. Cuff, B. Fry, A. Meissner, M. Wernig, K. Plath, R. Jaenisch, A. Wagschal, R. Feil, S. L. Schreiber, and E. S. Lander. A bivalent chromatin structure marks key developmental genes in embryonic stem cells. *Cell*, 125(2): 315–26, 2006. ISSN 0092-8674 (Print) 0092-8674 (Linking). doi: 10.1016/j.cell.2006.02.041. URL `http://www.ncbi.nlm.nih.gov/pubmed/16630819http://www.cell.com/cell/pdf/S0092-8674(06)00380-1.pdf`. Bernstein, Bradley E Mikkelsen, Tarjei S Xie, Xiaohui Kamal, Michael Huebert, Dana J Cuff, James Fry, Ben Meissner, Alex Wernig, Marius Plath, Kathrin Jaenisch, Rudolf Wagschal, Alexandre Feil, Robert Schreiber, Stuart L Lander, Eric S eng CA84198/CA/NCI NIH HHS/ GM38627/GM/NIGMS NIH HHS/ HD045022/HD/NICHD NIH HHS/ Research Support, N.I.H., Extramural Research Support, Non-U.S. Gov't 2006/04/25 09:00 Cell. 2006 Apr 21;125(2):315-26.

[22] F. Roudier, F. K. Teixeira, and V. Colot. Chromatin indexing in arabidopsis: an epigenomic tale of tails and more. *Trends Genet*, 25(11):511–7, 2009. ISSN 0168-9525 (Print) 0168-9525 (Linking). doi: S0168-9525(09)00186-3[pii]10.1016/j.tig.2009.09.013.

[23] H. Ledford. Language: Disputed definitions. *Nature*, 455 (7216):1023–8, 2008. ISSN 1476-4687 (Electronic) 0028-0836 (Linking). doi: 10.1038/4551023a. URL `http://www.ncbi.nlm.nih.gov/pubmed/18948925`. Ledford, Heidi eng Historical Article News England 2008/10/25 09:00 Nature. 2008 Oct 23;455(7216):1023-8. doi: 10.1038/4551023a.

[24] J. Pfluger and D. Wagner. Histone modifications and dynamic regulation of genome accessibility in plants. *Curr Opin Plant Biol*, 10(6):645–652, 2007.

[25] S. Rombauts, K. Florquin, M. Lescot, K. Marchal, P. Rouze, and Y. van de Peer. Computational approaches to identify promoters and cis-regulatory elements in plant genomes. *Plant Physiol*, 132(3):1162–76, 2003. ISSN 0032-0889 (Print) 0032-0889 (Linking).

[26] C. David Allis, Thomas Jenuwein, Danny Reinberg, and Marie-Laure Caparros. *Epigenetics*. Cold Spring Harbor Laboratory Press, 2007. ISBN 0879697245.

[27] Qi Xie and Hui-Shan Guo. Systemic antiviral silencing in plants. *Virus Res*, 118(1-2):1–6, 2006.

[28] The Arabidopsis Genome Initiative. Analysis of the genome sequence of the flowering plant arabidopsis thaliana. *Nature*, 408(6814):796–815, 2000. ISSN 0028-0836 (Print) 0028-0836 (Linking). doi: 10.1038/35048692.

[29] M. Nordborg, T. T. Hu, Y. Ishino, J. Jhaveri, C. Toomajian, H. Zheng, E. Bakker, P. Calabrese, J. Gladstone, R. Goyal, M. Jakobsson, S. Kim, Y. Morozov, B. Padhukasahasram, V. Plagnol, N. A. Rosenberg, C. Shah, J. D. Wall, J. Wang, K. Zhao, T. Kalbfleisch, V. Schulz, M. Kreitman, and J. Bergelson. The pattern of polymorphism in arabidopsis thaliana. *PLoS Biol*, 3(7):e196, 2005. ISSN 1545-7885 (Electronic) 1544-9173 (Linking). doi: 10.1371/journal.pbio.0030196.

[30] A. Platt, M. Horton, Y. S. Huang, Y. Li, A. E. Anastasio, N. W. Mulyati, J. Agren, O. Bossdorf, D. Byers, K. Donohue, M. Dunning, E. B. Holub, A. Hudson, V. Le Corre, O. Loudet, F. Roux, N. Warthmann, D. Weigel, L. Rivero, R. Scholl, M. Nordborg, J. Bergelson, and J. O. Borevitz. The scale of population structure in arabidopsis thaliana. *PLoS Genet*, 6 (2):e1000843, 2010. ISSN 1553-7404 (Electronic) 1553-7390 (Linking). doi: 10.1371/journal.pgen.1000843.

[31] J. Cao, K. Schneeberger, S. Ossowski, T. Gunther, S. Bender, J. Fitz, D. Koenig, C. Lanz, O. Stegle, C. Lippert, X. Wang, F. Ott, J. Muller, C. Alonso-Blanco, K. Borgwardt, K. J. Schmid, and D. Weigel. Whole-genome sequencing of multiple arabidopsis thaliana populations. *Nat Genet*, 43(10):956–63, 2011. ISSN 1546-1718 (Electronic) 1061-4036 (Linking). doi: 10.1038/ng.911.

[32] R. C. Jansen and J. P. Nap. Genetical genomics: the added value from segregation. *Trends Genet*, 17(7):388–91, 2001. ISSN 0168-9525 (Print) 0168-9525 (Linking). URL `http://www.ncbi.nlm.nih.gov/pubmed/11418218`. Jansen, R C Nap, J P eng Research Support, Non-U.S. Gov't England 2001/06/22 10:00 Trends Genet. 2001 Jul;17(7):388-91.

[33] R. DeCook, S. Lall, D. Nettleton, and S. H. Howell. Genetic regulation of gene expression during shoot development in arabidopsis. *Genetics*, 172(2):1155–64, 2006. ISSN 0016-6731 (Print) 0016-6731 (Linking). doi: 10.1534/genetics.105.042275. URL `http://www.ncbi.nlm.nih.gov/pubmed/15956669http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1456214/pdf/GEN17221155.pdf`. DeCook, Rhonda Lall, Sonia Nettleton, Dan Howell, Stephen H eng Research Support, Non-U.S. Gov't Research Support, U.S. Gov't, Non-P.H.S. 2005/06/16 09:00 Genetics. 2006 Feb;172(2):1155-64. Epub 2005 Jun 14.

[34] F. A. Cubillos, O. Stegle, C. Grondin, M. Canut, S. Tisne, I. Gy, and O. Loudet. Extensive cis-regulatory variation robust to environmental perturbation in arabidopsis. *Plant Cell*, 2014. ISSN 1532-298X (Electronic) 1040-4651 (Linking). doi: 10.1105/tpc.114.130310. URL `http://www.ncbi.nlm.nih.gov/pubmed/25428981http://www.plantcell.org/content/early/2014/11/26/tpc.114.130310.full.pdf`. Cubillos, Francisco A Stegle, Oliver Grondin, Cecile Canut, Matthieu Tisne, Sebastien Gy, Isabelle Loudet, Olivier ENG 2014/11/28 06:00 Plant Cell. 2014 Nov 26. pii: tpc.114.130310.

[35] M. T. Maurano, R. Humbert, E. Rynes, R. E. Thurman, E. Haugen, H. Wang, A. P. Reynolds, R. Sandstrom, H. Qu, J. Brody, A. Shafer, F. Neri, K. Lee, T. Kutyavin, S. Stehling-Sun, A. K. Johnson, T. K. Canfield, E. Giste, M. Diegel, D. Bates, R. S. Hansen, S. Neph, P. J. Sabo, S. Heimfeld, A. Raubitschek, S. Ziegler, C. Cotsapas, N. Sotoodehnia, I. Glass, S. R. Sunyaev, R. Kaul, and J. A. Stamatoyannopoulos. Systematic localization of common disease-associated variation in regulatory dna. *Science*, 337(6099):1190–5, 2012. ISSN 1095-9203 (Electronic) 0036-8075 (Linking). doi: 10.1126/science.1222794. URL

http://www.ncbi.nlm.nih.gov/pubmed/22955828. Maurano, Matthew T Humbert, Richard Rynes, Eric Thurman, Robert E Haugen, Eric Wang, Hao Reynolds, Alex P Sandstrom, Richard Qu, Hongzhu Brody, Jennifer Shafer, Anthony Neri, Fidencio Lee, Kristen Kutyavin, Tanya Stehling-Sun, Sandra Johnson, Audra K Canfield, Theresa K Giste, Erika Diegel, Morgan Bates, Daniel Hansen, R Scott Neph, Shane Sabo, Peter J Heimfeld, Shelly Raubitschek, Antony Ziegler, Steven Cotsapas, Chris Sotoodehnia, Nona Glass, Ian Sunyaev, Shamil R Kaul, Rajinder Stamatoyannopoulos, John A eng F31 MH094073/MH/NIMH NIH HHS/ P30 DK056465/DK/NIDDK NIH HHS/ R01 HL088456/HL/NHLBI NIH HHS/ R01HL088456/HL/NHLBI NIH HHS/ R24 HD000836/HD/NICHD NIH HHS/ R24HD000836-47/HD/NICHD NIH HHS/ U01ES01156/ES/NIEHS NIH HHS/ U54 HG004592/HG/NHGRI NIH HHS/ U54HG004592/HG/NHGRI NIH HHS/ Research Support, N.I.H., Extramural New York, N.Y. 2012/09/08 06:00 Science. 2012 Sep 7;337(6099):1190-5. doi: 10.1126/science.1222794. Epub 2012 Sep 5.

[36] F. A. Cubillos, V. Coustham, and O. Loudet. Lessons from eqtl mapping studies: non-coding regions and their role behind natural phenotypic variation in plants. *Curr Opin Plant Biol*, 15(2):192–8, 2012. ISSN 1879-0356 (Electronic) 1369-5266 (Linking). doi: 10.1016/j.pbi.2012.01.005. URL http://www.ncbi.nlm.nih.gov/pubmed/22265229. Cubillos, Francisco A Coustham, Vincent Loudet, Olivier eng Research Support, Non-U.S. Gov't Review England 2012/01/24 06:00 Curr Opin Plant Biol. 2012 Apr;15(2):192-8. doi: 10.1016/j.pbi.2012.01.005. Epub 2012 Jan 20.

[37] S. De Bodt, S. Maere, and Y. Van de Peer. Genome duplication and the origin of angiosperms. *Trends Ecol Evol*, 20(11):591–7, 2005. ISSN 0169-5347 (Print) 0169-5347 (Linking). doi: S0169-5347(05)00249-1[pii]10.1016/j.tree.2005.07.008. URL http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=16701441http://www.cell.com/trends/ecology-evolution//retrieve/pii/S0169534705002491?_returnURL=http://linkinghub.elsevier.com/retrieve/pii/S0169534705002491?showall=true. De Bodt, Stefanie Maere, Steven Van de Peer, Yves England Trends in ecology & evolution Trends Ecol Evol. 2005 Nov;20(11):591-7. Epub 2005 Aug 9.

[38] C. Sayou, M. Monniaux, M. H. Nanao, E. Moyroud, S. F. Brockington, E. Thevenon, H. Chahtane, N. Warthmann, M. Melkonian, Y. Zhang, G. K. Wong, D. Weigel, F. Parcy, and R. Dumas. A promiscuous intermediate underlies the evolution of leafy dna binding specificity. *Science*, 343(6171):645–8, 2014. ISSN 1095-9203 (Electronic) 0036-8075 (Linking). doi: 10.1126/science.1248229. URL http://www.ncbi.nlm.nih.gov/pubmed/24436181. Sayou, Camille Monniaux, Marie Nanao, Max H Moyroud, Edwige Brockington, Samuel F Thevenon, Emmanuel Chahtane, Hicham Warthmann, Norman Melkonian, Michael Zhang, Yong Wong, Gane Ka-Shu Weigel, Detlef Parcy, Francois Dumas, Renaud eng Research Support, Non-U.S. Gov't New York, N.Y. 2014/01/18 06:00 Science. 2014 Feb 7;343(6171):645-8. doi: 10.1126/science.1248229. Epub 2014 Jan 16.

[39] A. P. Gasch, A. M. Moses, D. Y. Chiang, H. B. Fraser, M. Berardini, and M. B. Eisen. Conservation and evolution of cis-regulatory systems in ascomycete fungi. *PLoS Biol*, 2(12): e398, 2004. ISSN 1545-7885 (Electronic) 1544-9173 (Linking). doi: 10.1371/journal.pbio.0020398.

[40] J. Ihmels, S. Bergmann, J. Berman, and N. Barkai. Comparative gene expression analysis by differential clustering approach: application to the candida albicans transcription program. *PLoS Genet*, 1(3):e39, 2005. Ihmels, Jan Bergmann, Sven Berman, Judith Barkai, Naama AI 14666/AI/NIAID NIH HHS/United States AI50562/AI/NIAID NIH HHS/United States DE 14666/DE/NIDCR NIH HHS/United States Research Support, N.I.H., Extramural Research Support, Non-U.S. Gov't United States PLoS genetics PLoS Genet. 2005 Sep;1(3):e39.

[41] A. Tanay, A. Regev, and R. Shamir. Conservation and evolvability in regulatory networks: the evolution of ribosomal regulation in yeast. *Proc Natl Acad Sci U S A*, 102(20):7203–8, 2005. ISSN 0027-8424 (Print) 0027-8424 (Linking). doi: 0502521102[pii]10.1073/pnas.0502521102.

[42] M. Blanchette and M. Tompa. Discovery of regulatory elements by a computational method for phylogenetic footprinting. *Genome Res.*, 12(5):739–748, May 2002.

[43] D. Schmidt, M. D. Wilson, B. Ballester, P. C. Schwalie, G. D. Brown, A. Marshall, C. Kutter, S. Watt, C. P. Martinez-Jimenez, S. Mackay, I. Talianidis, P. Flicek, and D. T. Odom. Five-vertebrate chip-seq reveals the evolutionary dynamics of transcription factor binding. *Science*, 328(5981):1036–40, 2010. ISSN 1095-9203 (Electronic) 0036-8075 (Linking). doi: science.1186176[pii]10.1126/science.1186176.

[44] B. Ballester, A. Medina-Rivera, D. Schmidt, M. Gonzalez-Porta, M. Carlucci, X. Chen, K. Chessman, A. J. Faure, A. P. Funnell, A. Goncalves, C. Kutter, M. Lukk, S. Menon, W. M. McLaren, K. Stefflova, S. Watt, M. T. Weirauch, M. Crossley, J. C. Marioni, D. T. Odom, P. Flicek, and M. D. Wilson. Multi-species, multi-transcription factor binding highlights conserved control of tissue-specific biological pathways. *Elife*, 3:e02626, 2014. ISSN 2050-084X (Electronic) 2050-084X (Linking). doi: 10.7554/eLife.02626. URL http://www.ncbi.nlm.nih.gov/pubmed/25279814http://elifesciences.org/content/elife/3/e02626.full.pdf. Ballester, Benoit Medina-Rivera, Alejandra Schmidt, Dominic Gonzalez-Porta, Mar Carlucci, Matthew Chen, Xiaoting Chessman, Kyle Faure, Andre J Funnell, Alister P W Goncalves, Angela Kutter, Claudia Lukk, Margus Menon, Suraj McLaren, William M Stefflova, Klara Watt, Stephen Weirauch, Matthew T Crossley, Merlin Marioni, John C Odom, Duncan T Flicek, Paul Wilson, Michael D eng 15603/Cancer Research UK/United Kingdom WT095908/Wellcome Trust/United Kingdom WT098051/Wellcome Trust/United Kingdom Cancer Research UK/United Kingdom Research Support, Non-U.S. Gov't England 2014/10/04 06:00 Elife. 2014 Oct 3;3:e02626. doi: 10.7554/eLife.02626.

[45] M. Paris, T. Kaplan, X. Y. Li, J. E. Villalta, S. E. Lott, and M. B. Eisen. Extensive divergence of transcription factor binding in drosophila embryos with highly conserved gene expression. *PLoS Genet*, 9(9):e1003748, 2013. ISSN 1553-7404 (Electronic) 1553-7390 (Linking). doi: 10.1371/journal.pgen.1003748. URL http://www.ncbi.nlm.nih.gov/pubmed/24068946http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3772039/pdf/pgen.1003748.pdf. Paris, Mathilde Kaplan, Tommy Li, Xiao Yong Villalta, Jacqueline E Lott, Susan E Eisen, Michael B eng HG002779/HG/NHGRI NIH HHS/ K99/R00-GM098448/GM/NIGMS NIH HHS/ R00 GM098448/GM/NIGMS NIH HHS/ Howard Hughes Medical Institute/ Research Support, N.I.H., Extramural Research Support, Non-U.S. Gov't 2013/09/27 06:00 PLoS Genet. 2013;9(9):e1003748. doi: 10.1371/journal.pgen.1003748. Epub 2013 Sep 12.

[46] F. Thibaud-Nissen, H. Wu, T. Richmond, J. C. Redman, C. Johnson, R. Green, J. Arias, and C. D. Town. Development of arabidopsis whole-genome microarrays and their application to the discovery of binding sites for the tga2 transcription factor in salicylic acid-treated plants. *Plant J*, 47(1):152–62, 2006. ISSN 0960-7412 (Print) 0960-7412 (Linking). doi: TPJ2770[pii]10.1111/j.1365-313X.2006.02770.x.

[47] A. Verkest, T. Abeel, K. Heyndrickx, J. Van Leene, C. Lanz, E. Van De Slijke, N. De Winne, D. Eeckhout, G. Persiau, F. Van Breusegem, D. Inze, K. Vandepoele, and G. De Jaeger. A generic tool for transcription factor target gene discovery

in arabidopsis cell suspension cultures based on tandem chromatin affinity purification. *Plant Physiol*, 2014. ISSN 1532-2548 (Electronic) 0032-0889 (Linking). doi: 10.1104/pp.113.229617.

[48] J. Brind'Amour, S. Liu, M. Hudson, C. Chen, M. M. Karimi, and M. C. Lorincz. An ultra-low-input native ChIP-seq protocol for genome-wide profiling of rare cell populations. *Nat Commun*, 6:6033, 2015.

[49] B. Ren, F. Robert, J. J. Wyrick, O. Aparicio, E. G. Jennings, I. Simon, J. Zeitlinger, J. Schreiber, N. Hannett, E. Kanin, T. L. Volkert, C. J. Wilson, S. P. Bell, and R. A. Young. Genome-wide location and function of dna binding proteins. *Science*, 290(5500):2306–9, 2000. ISSN 0036-8075 (Print) 0036-8075 (Linking). doi: 10.1126/science.290.5500.2306.

[50] D. S. Johnson, A. Mortazavi, R. M. Myers, and B. Wold. Genome-wide mapping of in vivo protein-dna interactions. *Science*, 316(5830):1497–502, 2007. ISSN 1095-9203 (Electronic) 0036-8075 (Linking). doi: 10.1126/science.1141319.

[51] P. J. Farnham. Insights from genomic profiling of transcription factors. *Nat Rev Genet*, 10(9):605–16, 2009. ISSN 1471-0064 (Electronic) 1471-0056 (Linking). doi: nrg2636[pii]10.1038/nrg2636.

[52] Z. D. Zhang, J. Rozowsky, M. Snyder, J. Chang, and M. Gerstein. Modeling chip sequencing in silico with applications. *PLoS computational biology*, 4(8):e1000158, 2008. ISSN 1553-7358 (Electronic) 1553-734X (Linking). doi: 10.1371/journal.pcbi.1000158.

[53] M. Allhoff, K. Sere, H. Chauvistre, Q. Lin, M. Zenke, and I. G. Costa. Detecting differential peaks in chip-seq signals with odin. *Bioinformatics*, 30(24):3467–75, 2014. ISSN 1367-4811 (Electronic) 1367-4803 (Linking). doi: 10.1093/bioinformatics/btu722. URL `http://www.ncbi.nlm.nih.gov/pubmed/25371479`. Allhoff, Manuel Sere, Kristin Chauvistre, Heike Lin, Qiong Zenke, Martin Costa, Ivan G eng England Oxford, England 2014/11/06 06:00 Bioinformatics. 2014 Dec 15;30(24):3467-75. doi: 10.1093/bioinformatics/btu722. Epub 2014 Nov 3.

[54] T. Abeel, T. Van Parys, Y. Saeys, J. Galagan, and Y. Van de Peer. Genomeview: a next-generation genome browser. *Nucleic acids research*, 40(2):e12, 2012. ISSN 1362-4962 (Electronic) 0305-1048 (Linking). doi: 10.1093/nar/gkr995.

[55] H. Ji and W. H. Wong. TileMap: create chromosomal map of tiling array hybridizations. *Bioinformatics*, 21(18):3629–3636, jul 2005. doi: 10.1093/bioinformatics/bti593. URL `http://dx.doi.org/10.1093/bioinformatics/bti593`.

[56] W. Li, C. A. Meyer, and X. S. Liu. A hidden markov model for analyzing ChIP-chip experiments on genome tiling arrays and its application to p53 binding sequences. *Bioinformatics*, 21(Suppl 1):i274–i282, jun 2005. doi: 10.1093/bioinformatics/bti1046. URL `http://dx.doi.org/10.1093/bioinformatics/bti1046`.

[57] Raphael Gottardo, Wei Li, W. Evan Johnson, and X. Shirley Liu. A flexible and powerful bayesian hierarchical model for ChIP-chip experiments. *Biometrics*, 64(2):468–478, jun 2008. doi: 10.1111/j.1541-0420.2007.00899.x. URL `http://dx.doi.org/10.1111/j.1541-0420.2007.00899.x`.

[58] Ayat Hatem, Doruk Bozdağ, Amanda E Toland, and Ümit V Çatalyürek. Benchmarking short sequence mapping tools. *BMC Bioinformatics*, 14(1):184, 2013. doi: 10.1186/1471-2105-14-184. URL `http://dx.doi.org/10.1186/1471-2105-14-184`.

[59] S. Pepke, B. Wold, and A. Mortazavi. Computation for chip-seq and rna-seq studies. *Nat Methods*, 6(11 Suppl):S22–32, 2009. ISSN 1548-7105 (Electronic) 1548-7091 (Linking). doi: nmeth.1371[pii]10.1038/nmeth.1371.

[60] J. J. Li and I. Herskowitz. Isolation of orc6, a component of the yeast origin recognition complex by a one-hybrid system. *Science*, 262(5141):1870–4, 1993. ISSN 0036-8075 (Print) 0036-8075 (Linking). URL `http://www.ncbi.nlm.nih.gov/pubmed/8266075`. Li, J J Herskowitz, I eng AI18738/AI/NIAID NIH HHS/ Research Support, Non-U.S. Gov't Research Support, U.S. Gov't, P.H.S. New York, N.Y. 1993/12/17 Science. 1993 Dec 17;262(5141):1870-4.

[61] B. Deplancke, D. Dupuy, M. Vidal, and A. J. Walhout. A gateway-compatible yeast one-hybrid system. *Genome Res*, 14(10B):2093–101, 2004. ISSN 1088-9051 (Print) 1088-9051 (Linking). doi: 10.1101/gr.2445504. URL `http://www.ncbi.nlm.nih.gov/pubmed/15489331http://www.ncbi.nlm.nih.gov/pmc/articles/PMC528925/pdf/0142093.pdf`. Deplancke, Bart Dupuy, Denis Vidal, Marc Walhout, Albertha J M eng 4 R33 CA097516-02/CA/NCI NIH HHS/ Research Support, Non-U.S. Gov't Research Support, U.S. Gov't, P.H.S. 2004/10/19 09:00 Genome Res. 2004 Oct;14(10B):2093-101.

[62] J. S. Reece-Hoyes, A. Diallo, B. Lajoie, A. Kent, S. Shrestha, S. Kadreppa, C. Pesyna, J. Dekker, C. L. Myers, and A. J. Walhout. Enhanced yeast one-hybrid assays for high-throughput gene-centered regulatory network mapping. *Nat Methods*, 8(12):1059–64, 2011. ISSN 1548-7105 (Electronic) 1548-7091 (Linking). doi: 10.1038/nmeth.1748. URL `http://www.ncbi.nlm.nih.gov/pubmed/22037705http://www.nature.com/nmeth/journal/v8/n12/pdf/nmeth.1748.pdf`. Reece-Hoyes, John S Diallo, Alos Lajoie, Bryan Kent, Amanda Shrestha, Shaleen Kadreppa, Sreenath Pesyna, Colin Dekker, Job Myers, Chad L Walhout, Albertha J M eng GM082971/GM/NIGMS NIH HHS/ HG003143/HG/NHGRI NIH HHS/ HG005084/HG/NHGRI NIH HHS/ R01 HG003143/HG/NHGRI NIH HHS/ R01 HG005084/HG/NHGRI NIH HHS/ R01 HG005084-01A1/HG/NHGRI NIH HHS/ Research Support, N.I.H., Extramural Research Support, Non-U.S. Gov't Research Support, U.S. Gov't, Non-P.H.S. 2011/11/01 06:00 Nat Methods. 2011 Oct 30;8(12):1059-64. doi: 10.1038/nmeth.1748.

[63] V. Vermeirssen, B. Deplancke, M. I. Barrasa, J. S. Reece-Hoyes, H. E. Arda, C. A. Grove, N. J. Martinez, R. Sequerra, L. Doucette-Stamm, M. R. Brent, and A. J. Walhout. Matrix and steiner-triple-system smart pooling assays for high-performance transcription regulatory network mapping. *Nat Methods*, 4(8):659–64, 2007. ISSN 1548-7091 (Print) 1548-7091 (Linking). doi: 10.1038/nmeth1063. URL `http://www.ncbi.nlm.nih.gov/pubmed/17589517`. Vermeirssen, Vanessa Deplancke, Bart Barrasa, M Inmaculada Reece-Hoyes, John S Arda, H Efsun Grove, Christian A Martinez, Natalia J Sequerra, Reynaldo Doucette-Stamm, Lynn Brent, Michael R Walhout, Albertha J M eng DK068429/DK/NIDDK NIH HHS/ DK071713/DK/NIDDK NIH HHS/ Research Support, N.I.H., Extramural Research Support, Non-U.S. Gov't Validation Studies 2007/06/26 09:00 Nat Methods. 2007 Aug;4(8):659-64. Epub 2007 Jun 24.

[64] S. M. Brady, L. Zhang, M. Megraw, N. J. Martinez, E. Jiang, C. S. Yi, W. Liu, A. Zeng, M. Taylor-Teeples, D. Kim, S. Ahnert, U. Ohler, D. Ware, A. J. Walhout, and P. N. Benfey. A stele-enriched gene regulatory network in the arabidopsis root. *Mol Syst Biol*, 7:459, 2011. ISSN 1744-4292 (Electronic) 1744-4292 (Linking). doi: msb2010114[pii]10.1038/msb.2010.114.

[65] A. Gaudinier, L. Zhang, J. S. Reece-Hoyes, M. Taylor-Teeples, L. Pu, Z. Liu, G. Breton, J. L. Pruneda-Paz, D. Kim, S. A. Kay, A. J. Walhout, D. Ware, and S. M. Brady. Enhanced y1h assays for arabidopsis. *Nat Methods*, 8(12):1053–5, 2011. ISSN 1548-7105 (Electronic) 1548-7091 (Linking). doi: 10.1038/nmeth.1750. URL `http://www.ncbi.nlm.nih.gov/pubmed/22037706`. Gaudinier, Allison Zhang, Lifang Reece-Hoyes, John S Taylor-Teeples, Mallorie Pu, Li Liu, Zhijie Breton, Ghislain Pruneda-Paz, Jose L Kim, Dahae Kay, Steve A Walhout, Albertha J M Ware, Doreen Brady, Siobhan M eng GM056006/GM/NIGMS NIH HHS/ GM082971/GM/NIGMS NIH HHS/ GM092412/GM/NIGMS

NIH HHS/ R01 GM082971/GM/NIGMS NIH HHS/ Research Support, N.I.H., Extramural Research Support, Non-U.S. Gov't Research Support, U.S. Gov't, Non-P.H.S. 2011/11/01 06:00 Nat Methods. 2011 Oct 30;8(12):1053-5. doi: 10.1038/nmeth.1750.

[66] S. Lindemose, M. K. Jensen, J. V. de Velde, C. O'Shea, K. S. Heyndrickx, C. T. Workman, K. Vandepoele, K. Skriver, and F. D. Masi. A dna-binding-site landscape and regulatory network analysis for nac transcription factors in arabidopsis thaliana. *Nucleic Acids Res*, 2014. ISSN 1362-4962 (Electronic) 0305-1048 (Linking). doi: 10.1093/nar/gku502.

[67] L. Elnitski, V. X. Jin, P. J. Farnham, and S. J. Jones. Locating mammalian transcription factor binding sites: a survey of computational and experimental techniques. *Genome Res*, 16(12): 1455–64, 2006. ISSN 1088-9051 (Print) 1088-9051 (Linking). doi: gr.4140006[pii]10.1101/gr.4140006.

[68] C Tuerk and L Gold. Systematic evolution of ligands by exponential enrichment: RNA ligands to bacteriophage T4 DNA polymerase. *Science*, 249(4968):505–10, 1990.

[69] T. Shimada, N. Fujita, M. Maeda, and A. Ishihama. Systematic search for the cra-binding promoters using genomic selex system. *Genes Cells*, 10(9):907–18, 2005. ISSN 1356-9597 (Print) 1356-9597 (Linking). doi: GTC888[pii]10.1111/j.1365-2443.2005.00888.x.

[70] R. Stoltenburg, C. Reinemann, and B. Strehlitz. Selex–a (r)evolutionary method to generate high-affinity nucleic acid ligands. *Biomol Eng*, 24(4):381–403, 2007. ISSN 1389-0344 (Print) 1389-0344 (Linking). doi: S1389-0344(07)00066-4[pii]10.1016/j.bioeng.2007.06.001.

[71] M. L. Bulyk, E. Gentalen, D. J. Lockhart, and G. M. Church. Quantifying dna-protein interactions by double-stranded dna arrays. *Nat Biotechnol*, 17(6):573–7, 1999. ISSN 1087-0156 (Print) 1087-0156 (Linking). doi: 10.1038/9878. URL `http://www.ncbi.nlm.nih.gov/pubmed/10385322`. Bulyk, M L Gentalen, E Lockhart, D J Church, G M eng Research Support, Non-U.S. Gov't Research Support, U.S. Gov't, Non-P.H.S. 1999/06/29 10:00 Nat Biotechnol. 1999 Jun;17(6):573-7.

[72] A. Marco, C. Konikoff, T. L. Karr, and S. Kumar. Relationship between gene co-expression and sharing of transcription factor binding sites in drosophila melanogaster. *Bioinformatics*, 25 (19):2473–7, 2009. ISSN 1367-4811 (Electronic) 1367-4803 (Linking). doi: btp462[pii]10.1093/bioinformatics/btp462.

[73] K. D. MacIsaac and E. Fraenkel. Practical strategies for discovering regulatory dna sequence motifs. *PLoS Comput Biol*, 2(4):e36, 2006. ISSN 1553-7358 (Electronic). doi: 10.1371/journal.pcbi.0020036.

[74] S. Hannenhalli. Eukaryotic transcription factor binding sites–modeling and integrative search methods. *Bioinformatics*, 24 (11):1325–31, 2008.

[75] D A Tagle, B F Koop, M Goodman, J L Slightom, D L Hess, and R T Jones. Embryonic epsilon and gamma globin genes of a prosimian primate (Galago crassicaudatus). Nucleotide and amino acid sequences, developmental regulation and phylogenetic footprints. *J Mol Biol*, 203(2):439–55, 1988.

[76] Z. Zhang and M. Gerstein. Of mice and men: phylogenetic footprinting aids the discovery of regulatory elements. *J Biol*, 2(2):11, 2003. ISSN 1475-4924 (Electronic) 1475-4924 (Linking). doi: 10.1186/1475-4924-2-11.

[77] X. Xie, J. Lu, E. J. Kulbokas, T. R. Golub, V. Mootha, K. Lindblad-Toh, E. S. Lander, and M. Kellis. Systematic discovery of regulatory motifs in human promoters and 3' utrs by comparison of several mammals. *Nature*, 434(7031):338–45, 2005. ISSN 1476-4687 (Electronic) 0028-0836 (Linking). doi: nature03441[pii]10.1038/nature03441.

[78] Y. Liu, X. S. Liu, L. Wei, R. B. Altman, and S. Batzoglou. Eukaryotic regulatory element conservation analysis and identification using comparative genomics. *Genome Res*, 14(3):451–8, 2004. ISSN 1088-9051 (Print) 1088-9051 (Linking). doi: 10.1101/gr.132760414/3/451[pii].

[79] D. Boffelli, J. McAuliffe, D. Ovcharenko, K. D. Lewis, I. Ovcharenko, L. Pachter, and E. M. Rubin. Phylogenetic shadowing of primate sequences to find functional regions of the human genome. *Science*, 299(5611):1391–4, 2003. ISSN 1095-9203 (Electronic) 0036-8075 (Linking). doi: 10.1126/science.1081331299/5611/1391[pii].

[80] L. E. Chai, S. K. Loh, S. T. Low, M. S. Mohamad, S. Deris, and Z. Zakaria. A review on the computational approaches for gene regulatory network construction. *Comput Biol Med*, 48: 55–65, 2014. ISSN 1879-0534 (Electronic) 0010-4825 (Linking). doi: 10.1016/j.compbiomed.2014.02.011. URL `http://www.ncbi.nlm.nih.gov/pubmed/24637147`. Chai, Lian En Loh, Swee Kuan Low, Swee Thing Mohamad, Mohd Saberi Deris, Safaai Zakaria, Zalmiyah eng Research Support, Non-U.S. Gov't Review 2014/03/19 06:00 Comput Biol Med. 2014 May;48:55-65. doi: 10.1016/j.compbiomed.2014.02.011. Epub 2014 Feb 24.

[81] A. Miura, S. Yonebayashi, K. Watanabe, T. Toyama, H. Shimada, and T. Kakutani. Mobilization of transposons by a mutation abolishing full dna methylation in arabidopsis. *Nature*, 411(6834):212–4, 2001. ISSN 0028-0836 (Print) 0028-0836 (Linking). doi: 10.1038/35075612. URL `http://www.ncbi.nlm.nih.gov/pubmed/11346800`. Miura, A Yonebayashi, S Watanabe, K Toyama, T Shimada, H Kakutani, T eng England 2001/05/11 10:00 Nature. 2001 May 10;411(6834):212-4.

[82] D. Zilberman, M. Gehring, R. K. Tran, T. Ballinger, and S. Henikoff. Genome-wide analysis of arabidopsis thaliana dna methylation uncovers an interdependence between methylation and transcription. *Nat Genet*, 39(1):61–9, 2007. ISSN 1061-4036 (Print) 1061-4036 (Linking). doi: ng1929[pii]10.1038/ng1929.

[83] S. W. Chan, I. R. Henderson, and S. E. Jacobsen. Gardening the genome: Dna methylation in arabidopsis thaliana. *Nat Rev Genet*, 6(5):351–60, 2005. ISSN 1471-0056 (Print) 1471-0056 (Linking). doi: 10.1038/nrg1601. URL `http://www.ncbi.nlm.nih.gov/pubmed/15861207`. Chan, Simon W-L Henderson, Ian R Jacobsen, Steven E eng Research Support, Non-U.S. Gov't Research Support, U.S. Gov't, P.H.S. Review England 2005/04/30 09:00 Nat Rev Genet. 2005 May;6(5):351-60.

[84] C. Becker, J. Hagmann, J. Muller, D. Koenig, O. Stegle, K. Borgwardt, and D. Weigel. Spontaneous epigenetic variation in the arabidopsis thaliana methylome. *Nature*, 480 (7376):245–9, 2011. ISSN 1476-4687 (Electronic) 0028-0836 (Linking). doi: 10.1038/nature10555. URL `http://www.ncbi.nlm.nih.gov/pubmed/22057020http://www.nature.com/nature/journal/v480/n7376/pdf/nature10555.pdf`. Becker, Claude Hagmann, Jorg Muller, Jonas Koenig, Daniel Stegle, Oliver Borgwardt, Karsten Weigel, Detlef eng Research Support, Non-U.S. Gov't England 2011/11/08 06:00 Nature. 2011 Sep 20;480(7376):245-9. doi: 10.1038/nature10555.

[85] R. J. Schmitz, M. D. Schultz, M. G. Lewsey, R. C. O'Malley, M. A. Urich, O. Libiger, N. J. Schork, and J. R. Ecker. Transgenerational epigenetic instability is a source of novel methylation variants. *Science*, 334(6054): 369–73, 2011. ISSN 1095-9203 (Electronic) 0036-8075 (Linking). doi: 10.1126/science.1212959. URL `http://www.ncbi.nlm.nih.gov/pubmed/21921155http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3210014/pdf/nihms325988.pdf`. Schmitz, Robert J Schultz, Matthew D Lewsey, Mathew G O'Malley, Ronan C Urich, Mark A Libiger, Ondrej Schork, Nicholas J Ecker, Joseph R eng F32 HG004830/HG/NHGRI NIH HHS/ F32 HG004830-01/HG/NHGRI NIH HHS/ F32 HG004830-02/HG/NHGRI NIH HHS/ F32 HG004830-03/HG/NHGRI

NIH HHS/ F32-HG004830/HG/NHGRI NIH HHS/ R01 HG003523/HG/NHGRI NIH HHS/ R01 HG003523-01/HG/NHGRI NIH HHS/ R01 HG003523-02/HG/NHGRI NIH HHS/ R01 HG003523-03/HG/NHGRI NIH HHS/ UL1 RR025774/RR/NCRR NIH HHS/ Howard Hughes Medical Institute/ Research Support, N.I.H., Extramural Research Support, Non-U.S. Gov't Research Support, U.S. Gov't, Non-P.H.S. New York, N.Y. 2011/09/17 06:00 Science. 2011 Oct 21;334(6054):369-73. doi: 10.1126/science.1212959. Epub 2011 Sep 15.

[86] Joerg Hagmann, Claude Becker, Jonas Müller, Oliver Stegle, Rhonda C Meyer, Korbinian Schneeberger, Joffrey Fitz, Thomas Altmann, Joy Bergelson, Karsten Borgwardt, and Detlef Weigel. Century-scale methylome stability in a recently diverged arabidopsis thaliana lineage. *bioRxiv*, 2014. doi: 10.1101/009225.

[87] T. Kouzarides. Chromatin modifications and their function. *Cell*, 128(4):693–705, 2007. ISSN 0092-8674 (Print) 0092-8674 (Linking). doi: 10.1016/j.cell.2007.02.005. URL `http://www.ncbi.nlm.nih.gov/pubmed/17320507http://www.cell.com/cell/abstract/S0092-8674(07)00184-5http://ac.els-cdn.com/S0092867407001845/1-s2.0-S0092867407001845-main.pdf?_tid=57c1b8fe-8073-11e4-b128-00000aab0f27&acdnat=1418219541_dfd24b3190395d4715d573f04977bb3a`. Kouzarides, Tony eng Review 2007/02/27 09:00 Cell. 2007 Feb 23;128(4):693-705.

[88] Z. Wang, H. Cao, F. Chen, and Y. Liu. The roles of histone acetylation in seed performance and plant development. *Plant Physiol Biochem*, 84C:125–133, 2014. ISSN 1873-2690 (Electronic) 0981-9428 (Linking). doi: 10.1016/j.plaphy.2014.09.010. URL `http://www.ncbi.nlm.nih.gov/pubmed/25270163`. Wang, Zhi Cao, Hong Chen, Fengying Liu, Yongxiu ENG REVIEW 2014/10/02 06:00 Plant Physiol Biochem. 2014 Nov;84C:125-133. doi: 10.1016/j.plaphy.2014.09.010. Epub 2014 Sep 24.

[89] M. Gentry and L. Hennig. Remodelling chromatin to shape development of plants. *Exp Cell Res*, 321(1):40–6, 2014. ISSN 1090-2422 (Electronic) 0014-4827 (Linking). doi: 10.1016/j.yexcr.2013.11.010. URL `http://www.ncbi.nlm.nih.gov/pubmed/24270012`. Gentry, Matthew Hennig, Lars eng Research Support, Non-U.S. Gov't Review 2013/11/26 06:00 Exp Cell Res. 2014 Feb 1;321(1):40-6. doi: 10.1016/j.yexcr.2013.11.010. Epub 2013 Nov 20.

[90] T. K. To and J. M. Kim. Epigenetic regulation of gene responsiveness in arabidopsis. *Front Plant Sci*, 4:548, 2014. ISSN 1664-462X (Electronic) 1664-462X (Linking). doi: 10.3389/fpls.2013.00548. URL `http://www.ncbi.nlm.nih.gov/pubmed/24432027`. To, Taiko K Kim, Jong Myong eng Review Switzerland 2014/01/17 06:00 Front Plant Sci. 2014 Jan 7;4:548. doi: 10.3389/fpls.2013.00548. eCollection 2014 Jan 7.

[91] G. Rios, C. Leida, A. Conejero, and M. L. Badenes. Epigenetic regulation of bud dormancy events in perennial plants. *Front Plant Sci*, 5:247, 2014. ISSN 1664-462X (Electronic) 1664-462X (Linking). doi: 10.3389/fpls.2014.00247. URL `http://www.ncbi.nlm.nih.gov/pubmed/24917873`. Rios, Gabino Leida, Carmen Conejero, Ana Badenes, Maria Luisa eng Review Switzerland 2014/06/12 06:00 Front Plant Sci. 2014 Jun 3;5:247. doi: 10.3389/fpls.2014.00247. eCollection 2014.

[92] E. S. Gan, J. Huang, and T. Ito. Functional roles of histone modification, chromatin remodeling and micrornas in arabidopsis flower development. *Int Rev Cell Mol Biol*, 305: 115–61, 2013. ISSN 1937-6448 (Print). doi: 10.1016/B978-0-12-407695-2.00003-2. URL `http://www.ncbi.nlm.nih.gov/pubmed/23890381`. Gan, Eng-Seng Huang, Jiangbo Ito, Toshiro eng Review Netherlands 2013/07/31 06:00 Int Rev Cell Mol Biol. 2013;305:115-61. doi: 10.1016/B978-0-12-407695-2.00003-2.

[93] J. Sequeira-Mendes, I. Araguez, R. Peiro, R. Mendez-Giraldez, X. Zhang, S. E. Jacobsen, U. Bastolla, and C. Gutierrez. The functional topography of the arabidopsis genome is organized in a reduced number of linear motifs of chromatin states. *Plant Cell*, 2014. ISSN 1532-298X (Electronic) 1040-4651 (Linking). doi: 10.1105/tpc.114.124578.

[94] G. C. Yuan, Y. J. Liu, M. F. Dion, M. D. Slack, L. F. Wu, S. J. Altschuler, and O. J. Rando. Genome-scale identification of nucleosome positions in s. cerevisiae. *Science*, 309(5734): 626–30, 2005. ISSN 1095-9203 (Electronic) 0036-8075 (Linking). doi: 10.1126/science.1112178. URL `http://www.ncbi.nlm.nih.gov/pubmed/15961632http://www.sciencemag.org/content/309/5734/626.full.pdf`. Yuan, Guo-Cheng Liu, Yuen-Jong Dion, Michael F Slack, Michael D Wu, Lani F Altschuler, Steven J Rando, Oliver J eng Research Support, N.I.H., Extramural Research Support, Non-U.S. Gov't Research Support, U.S. Gov't, P.H.S. New York, N.Y. 2005/06/18 09:00 Science. 2005 Jul 22;309(5734):626-30. Epub 2005 Jun 16.

[95] D. L. Vera, T. F. Madzima, J. D. Labonne, M. P. Alam, G. G. Hoffman, S. B. Girimurugan, J. Zhang, K. M. McGinnis, J. H. Dennis, and H. W. Bass. Differential nuclease sensitivity profiling of chromatin reveals biochemical footprints coupled to gene expression and functional dna elements in maize. *Plant Cell*, 26(10): 3883–93, 2014. ISSN 1532-298X (Electronic) 1040-4651 (Linking). doi: 10.1105/tpc.114.130609. URL `http://www.ncbi.nlm.nih.gov/pubmed/25361955http://www.plantcell.org/content/26/10/3883.full.pdf`. Vera, Daniel L Madzima, Thelma F Labonne, Jonathan D Alam, Mohammad P Hoffman, Gregg G Girimurugan, S B Zhang, Jinfeng McGinnis, Karen M Dennis, Jonathan H Bass, Hank W eng 2014/11/02 06:00 Plant Cell. 2014 Oct;26(10):3883-93. doi: 10.1105/tpc.114.130609. Epub 2014 Oct 31.

[96] G. E. Zentner and S. Henikoff. Surveying the epigenomic landscape, one base at a time. *Genome Biol*, 13 (10):250, 2012. ISSN 1465-6914 (Electronic) 1465-6906 (Linking). doi: 10.1186/gb4051. URL `http://www.ncbi.nlm.nih.gov/pubmed/23088423http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3491405/pdf/gb-2012-13-10-250.pdf`. Zentner, Gabriel E Henikoff, Steven eng 5U01 HG004274/HG/NHGRI NIH HHS/ R01 ES020116/ES/NIEHS NIH HHS/ U54CA143862/CA/NCI NIH HHS/ Howard Hughes Medical Institute/ Research Support, N.I.H., Extramural Research Support, Non-U.S. Gov't England 2012/10/24 06:00 Genome Biol. 2012 Oct 22;13(10):250. doi: 10.1186/gb4051.

[97] P. Crevillen, C. Sonmez, Z. Wu, and C. Dean. A gene loop containing the floral repressor flc is disrupted in the early phase of vernalization. *EMBO J*, 32(1):140–8, 2013. ISSN 1460-2075 (Electronic) 0261-4189 (Linking). doi: 10.1038/emboj.2012.324. URL `http://www.ncbi.nlm.nih.gov/pubmed/23222483`. Crevillen, Pedro Sonmez, Cagla Wu, Zhe Dean, Caroline eng BB/G009562/1/Biotechnology and Biological Sciences Research Council/United Kingdom BB/J004588/1/Biotechnology and Biological Sciences Research Council/United Kingdom Research Support, Non-U.S. Gov't England 2012/12/12 06:00 EMBO J. 2013 Jan 9;32(1):140-8. doi: 10.1038/emboj.2012.324. Epub 2012 Dec 7.

[98] J. Dekker, K. Rippe, M. Dekker, and N. Kleckner. Capturing chromosome conformation. *Science*, 295(5558):1306–11, 2002. ISSN 1095-9203 (Electronic) 0036-8075 (Linking). doi: 10.1126/science.1067799. URL `http://www.ncbi.nlm.nih.gov/pubmed/11847345http://www.sciencemag.org/content/295/5558/1306.full.pdf`. Dekker, Job Rippe, Karsten Dekker, Martijn Kleckner, Nancy eng GM25326/GM/NIGMS NIH HHS/ GM44794/GM/NIGMS NIH HHS/ R01 GM025326/GM/NIGMS NIH HHS/ R01

GM044794/GM/NIGMS NIH HHS/ Research Support, Non-U.S. Gov't Research Support, U.S. Gov't, P.H.S. New York, N.Y. 2002/02/16 10:00 Science. 2002 Feb 15;295(5558):1306-11.

[99] M. Simonis, P. Klous, E. Splinter, Y. Moshkin, R. Willemsen, E. de Wit, B. van Steensel, and W. de Laat. Nuclear organization of active and inactive chromatin domains uncovered by chromosome conformation capture-on-chip (4c). *Nat Genet*, 38(11):1348–54, 2006. ISSN 1061-4036 (Print) 1061-4036 (Linking). doi: 10.1038/ng1896. URL `http://www.ncbi.nlm.nih.gov/pubmed/17033623`. Simonis, Marieke Klous, Petra Splinter, Erik Moshkin, Yuri Willemsen, Rob de Wit, Elzo van Steensel, Bas de Laat, Wouter eng Research Support, Non-U.S. Gov't 2006/10/13 09:00 Nat Genet. 2006 Nov;38(11):1348-54. Epub 2006 Oct 8.

[100] J. Dostie, T. A. Richmond, R. A. Arnaout, R. R. Selzer, W. L. Lee, T. A. Honan, E. D. Rubio, A. Krumm, J. Lamb, C. Nusbaum, R. D. Green, and J. Dekker. Chromosome conformation capture carbon copy (5c): a massively parallel solution for mapping interactions between genomic elements. *Genome Res*, 16(10):1299–309, 2006. ISSN 1088-9051 (Print) 1088-9051 (Linking). doi: 10.1101/gr.5571506. URL `http://www.ncbi.nlm.nih.gov/pubmed/16954542http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1581439/pdf/1299.pdf`. Dostie, Josee Richmond, Todd A Arnaout, Ramy A Selzer, Rebecca R Lee, William L Honan, Tracey A Rubio, Eric D Krumm, Anton Lamb, Justin Nusbaum, Chad Green, Roland D Dekker, Job eng CA109597/CA/NCI NIH HHS/ HG003129/HG/NHGRI NIH HHS/ HG003143/HG/NHGRI NIH HHS/ R01 GM078986/GM/NIGMS NIH HHS/ R01 GM078986-01/GM/NIGMS NIH HHS/ R01 HG003143/HG/NHGRI NIH HHS/ R01GM078986/GM/NIGMS NIH HHS/ Research Support, N.I.H., Extramural Research Support, Non-U.S. Gov't Research Support, U.S. Gov't, Non-P.H.S. 2006/09/07 09:00 Genome Res. 2006 Oct;16(10):1299-309. Epub 2006 Sep 5.

[101] E. Lieberman-Aiden, N. L. van Berkum, L. Williams, M. Imakaev, T. Ragoczy, A. Telling, I. Amit, B. R. Lajoie, P. J. Sabo, M. O. Dorschner, R. Sandstrom, B. Bernstein, M. A. Bender, M. Groudine, A. Gnirke, J. Stamatoyannopoulos, L. A. Mirny, E. S. Lander, and J. Dekker. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science*, 326(5950):289–93, 2009. ISSN 1095-9203 (Electronic) 0036-8075 (Linking). doi: 10.1126/science.1181369. URL `http://www.ncbi.nlm.nih.gov/pubmed/19815776`. Lieberman-Aiden, Erez van Berkum, Nynke L Williams, Louise Imakaev, Maxim Ragoczy, Tobias Telling, Agnes Amit, Ido Lajoie, Bryan R Sabo, Peter J Dorschner, Michael O Sandstrom, Richard Bernstein, Bradley Bender, M A Groudine, Mark Gnirke, Andreas Stamatoyannopoulos, John Mirny, Leonid A Lander, Eric S Dekker, Job eng HG003143/HG/NHGRI NIH HHS/ R01 HG003143/HG/NHGRI NIH HHS/ R01 HG003143-06/HG/NHGRI NIH HHS/ R01HL06544/HL/NHLBI NIH HHS/ R37DK44746/DK/NIDDK NIH HHS/ T32 HG002295/HG/NHGRI NIH HHS/ U54HG004592/HG/NHGRI NIH HHS/ Research Support, N.I.H., Extramural Research Support, Non-U.S. Gov't Research Support, U.S. Gov't, Non-P.H.S. New York, N.Y. 2009/10/10 06:00 Science. 2009 Oct 9;326(5950):289-93. doi: 10.1126/science.1181369.

[102] S. Grob, M. W. Schmid, N. W. Luedtke, T. Wicker, and U. Grossniklaus. Characterization of chromosomal architecture in arabidopsis by chromosome conformation capture. *Genome Biol*, 14(11):R129, 2013. ISSN 1465-6914 (Electronic) 1465-6906 (Linking). doi: 10.1186/gb-2013-14-11-r129. URL `http://www.ncbi.nlm.nih.gov/pubmed/24267747http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4053840/pdf/gb-2013-14-11-r129.pdf`. Grob, Stefan Schmid, Marc W Luedtke, Nathan W Wicker, Thomas Grossniklaus, Ueli eng Research Support, Non-U.S. Gov't England 2013/11/26 06:00 Genome Biol. 2013 Nov 24;14(11):R129. doi: 10.1186/gb-2013-14-11-r129.

[103] S. Feng, S. J. Cokus, V. Schubert, J. Zhai, M. Pellegrini, and S. E. Jacobsen. Genome-wide hi-c analyses in wild-type and mutants reveal high-resolution chromatin interactions in arabidopsis. *Mol Cell*, 55(5):694–707, 2014. ISSN 1097-4164 (Electronic) 1097-2765 (Linking). doi: 10.1016/j.molcel.2014.07.008. URL `http://www.ncbi.nlm.nih.gov/pubmed/25132175`. Feng, Suhua Cokus, Shawn J Schubert, Veit Zhai, Jixian Pellegrini, Matteo Jacobsen, Steven E eng GM60398/GM/NIGMS NIH HHS/ Howard Hughes Medical Institute/ Research Support, N.I.H., Extramural Research Support, Non-U.S. Gov't 2014/08/19 06:00 Mol Cell. 2014 Sep 4;55(5):694-707. doi: 10.1016/j.molcel.2014.07.008. Epub 2014 Aug 14.

[104] S. Grob, M. W. Schmid, and U. Grossniklaus. Hi-c analysis in arabidopsis identifies the knot, a structure with similarities to the flamenco locus of drosophila. *Mol Cell*, 55(5):678–93, 2014. ISSN 1097-4164 (Electronic) 1097-2765 (Linking). doi: 10.1016/j.molcel.2014.07.009. URL `http://www.ncbi.nlm.nih.gov/pubmed/25132176http://ac.els-cdn.com/S1097276514006029/1-s2.0-S1097276514006029-main.pdf?_tid=766c0496-8120-11e4-b314-00000aab0f6b&acdnat=1418293895_056a4d2b685aa6e829174a3f11a6710b`. Grob, Stefan Schmid, Marc W Grossniklaus, Ueli eng Research Support, Non-U.S. Gov't 2014/08/19 06:00 Mol Cell. 2014 Sep 4;55(5):678-93. doi: 10.1016/j.molcel.2014.07.009. Epub 2014 Aug 14.

[105] M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin, and G. Sherlock. Gene ontology: tool for the unification of biology. the gene ontology consortium. *Nat Genet*, 25 (1):25–9, 2000. ISSN 1061-4036 (Print) 1061-4036 (Linking). doi: 10.1038/75556.

[106] J. S. Lee, E. Smith, and A. Shilatifard. The language of histone crosstalk. *Cell*, 142(5):682–5, 2010. ISSN 1097-4172 (Electronic) 0092-8674 (Linking). doi: S0092-8674(10)00918-9[pii]10.1016/j.cell.2010.08.011.

[107] S. De Bodt, D. Carvajal, J. Hollunder, J. Van den Cruyce, S. Movahedi, and D. Inze. Cornet: a user-friendly tool for data mining and integration. *Plant physiology*, 152(3):1167–79, 2010. ISSN 1532-2548 (Electronic) 0032-0889 (Linking). doi: pp.109.147215[pii]10.1104/pp.109.147215.

[108] K. Higo, Y. Ugawa, M. Iwamoto, and H. Higo. Place: a database of plant cis-acting regulatory dna elements. *Nucleic Acids Res*, 26(1):358–9, 1998. ISSN 0305-1048 (Print) 0305-1048 (Linking). doi: gkb021[pii].

[109] D. Meinke, R. Muralla, C. Sweeney, and A. Dickerman. Identifying essential genes in arabidopsis thaliana. *Trends Plant Sci*, 13(9):483–91, 2008. ISSN 1360-1385 (Print) 1360-1385 (Linking). doi: S1360-1385(08)00195-7[pii]10.1016/j.tplants.2008.06.003.

[110] M. Zulawski, R. Braginets, and W. X. Schulze. Phosphat goes kinases–searchable protein kinase target information in the plant phosphorylation site database phosphat. *Nucleic acids research*, 41(Database issue):D1176–84, 2013. ISSN 1362-4962 (Electronic) 0305-1048 (Linking). doi: 10.1093/nar/gks1081.

[111] T. Barrett, D. B. Troup, S. E. Wilhite, P. Ledoux, C. Evangelista, I. F. Kim, M. Tomashevsky, K. A. Marshall, K. H. Phillippy, P. M. Sherman, R. N. Muertter, M. Holko, O. Ayanbule, A. Yefanov, and A. Soboleva. Ncbi geo: archive for functional genomics data sets–10 years on. *Nucleic Acids Res*, 39(Database issue):D1005–10, 2011. ISSN 1362-4962 (Electronic) 0305-1048 (Linking). doi: gkq1184[pii]10.1093/nar/gkq1184.

[112] O. Thimm, O. Blasing, Y. Gibon, A. Nagel, S. Meyer, P. Kruger, J. Selbig, L. A. Muller, S. Y. Rhee, and M. Stitt. Mapman: a user-driven tool to display genomics data sets onto diagrams of metabolic pathways and other biological processes. *Plant J*, 37(6):914–39, 2004. ISSN 0960-7412 (Print) 0960-7412 (Linking).

[113] X. Dai and P. X. Zhao. psrnatarget: a plant small rna target analysis server. *Nucleic acids research*, 39(Web Server issue):W155–9, 2011. ISSN 1362-4962 (Electronic) 0305-1048 (Linking). doi: 10.1093/nar/gkr319.

[114] Z. Zhang, J. Yu, D. Li, F. Liu, X. Zhou, T. Wang, Y. Ling, and Z. Su. Pmrd: plant microrna database. *Nucleic acids research*, 38(Database issue):D806–13, 2010. ISSN 1362-4962 (Electronic) 0305-1048 (Linking). doi: 10.1093/nar/gkp818.

[115] A. Ben-Dor, R. Shamir, and Z. Yakhini. Clustering gene expression patterns. *Journal of computational biology : a journal of computational molecular cell biology*, 6(3-4):281–97, 1999. ISSN 1066-5277 (Print) 1066-5277 (Linking). doi: 10.1089/106652799318274.

[116] G. Pavesi, P. Mereghetti, G. Mauri, and G. Pesole. Weeder web: discovery of transcription factor binding sites in a set of sequences from co-regulated genes. *Nucleic Acids Res*, 32(Web Server issue):W199–203, 2004. ISSN 1362-4962 (Electronic) 0305-1048 (Linking). doi: 10.1093/nar/gkh46532/suppl_2/W199[pii].

[117] G. Thijs, M. Lescot, K. Marchal, S. Rombauts, B. De Moor, P. Rouze, and Y. Moreau. A higher-order background model improves the detection of promoter regulatory elements by gibbs sampling. *Bioinformatics*, 17(12):1113–22, 2001. ISSN 1367-4803 (Print) 1367-4803 (Linking).

[118] M. Thomas-Chollier, C. Herrmann, M. Defrance, O. Sand, D. Thieffry, and J. van Helden. Rsat peak-motifs: motif analysis in full-size chip-seq datasets. *Nucleic acids research*, 40(4): e31, 2012. ISSN 1362-4962 (Electronic) 0305-1048 (Linking). doi: 10.1093/nar/gkr1104.

[119] R. Li, C. Yu, Y. Li, T. W. Lam, S. M. Yiu, K. Kristiansen, and J. Wang. Soap2: an improved ultrafast tool for short read alignment. *Bioinformatics*, 25(15):1966–7, 2009. ISSN 1367-4811 (Electronic) 1367-4803 (Linking). doi: btp336[pii]10.1093/bioinformatics/btp336.

[120] S. Mahony and P. V. Benos. Stamp: a web tool for exploring dna-binding motif similarities. *Nucleic acids research*, 35 (Web Server issue):W253–8, 2007. ISSN 1362-4962 (Electronic) 0305-1048 (Linking). doi: 10.1093/nar/gkm272.

[121] A. Droit, C. Cheung, and R. Gottardo. rmat–an r/bioconductor package for analyzing chip-chip experiments. *Bioinformatics*, 26(5):678–9, 2010. ISSN 1367-4811 (Electronic) 1367-4803 (Linking). doi: 10.1093/bioinformatics/btq023.

[122] B. Zacher, P. Torkler, and A. Tresch. Analysis of affymetrix chip-chip data using starr and r/bioconductor. *Cold Spring Harb Protoc*, 2011(5), 2010. ISSN 1559-6095 (Electronic). doi: 2011/5/pdb.top110[pii]10.1101/pdb.top110.

[123] P. Danecek, A. Auton, G. Abecasis, C. A. Albers, E. Banks, M. A. DePristo, R. E. Handsaker, G. Lunter, G. T. Marth, S. T. Sherry, G. McVean, R. Durbin, and Group Genomes Project Analysis. The variant call format and vcftools. *Bioinformatics*, 27(15):2156–8, 2011. ISSN 1367-4811 (Electronic) 1367-4803 (Linking). doi: 10.1093/bioinformatics/btr330.

[124] A. R. Quinlan and I. M. Hall. Bedtools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, 26(6): 841–2, 2010. ISSN 1367-4811 (Electronic) 1367-4803 (Linking). doi: 10.1093/bioinformatics/btq033.

[125] M. Blanchette, W. J. Kent, C. Riemer, L. Elnitski, A. F. Smit, K. M. Roskin, R. Baertsch, K. Rosenbloom, H. Clawson, E. D. Green, D. Haussler, and W. Miller. Aligning multiple genomic sequences with the threaded blockset aligner. *Genome Res*, 14 (4):708–15, 2004. ISSN 1088-9051 (Print) 1088-9051 (Linking). doi: 10.1101/gr.193310414/4/708[pii].

[126] K. S. Pollard, M. J. Hubisz, K. R. Rosenbloom, and A. Siepel. Detection of nonneutral substitution rates on mammalian phylogenies. *Genome Res.*, 20(1):110–121, Jan 2010.

[127] A. Siepel, G. Bejerano, J. S. Pedersen, A. S. Hinrichs, M. Hou, K. Rosenbloom, H. Clawson, J. Spieth, L. W. Hillier, S. Richards, G. M. Weinstock, R. K. Wilson, R. A. Gibbs, W. J. Kent, W. Miller, and D. Haussler. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res.*, 15(8):1034–1050, Aug 2005.

[128] A. R. Subramanian, M. Kaufmann, and B. Morgenstern. Dialign-tx: greedy and progressive approaches for segment-based multiple sequence alignment. *Algorithms for molecular biology : AMB*, 3:6, 2008. ISSN 1748-7188 (Electronic) 1748-7188 (Linking). doi: 10.1186/1748-7188-3-6.

[129] R. Siddharthan. Sigma: multiple alignment of weakly-conserved non-coding dna sequence. *BMC Bioinformatics*, 7: 143, 2006. ISSN 1471-2105 (Electronic) 1471-2105 (Linking). doi: 10.1186/1471-2105-7-143.

[130] W. Huang, D. M. Umbach, and L. Li. Accurate anchoring alignment of divergent sequences. *Bioinformatics*, 22(1):29–34, Jan 2006.

[131] E. Picot, P. Krusche, A. Tiskin, I. Carre, and S. Ott. Evolutionary analysis of regulatory sequences (ears) in plants. *Plant J*, 2010. ISSN 1365-313X (Electronic) 0960-7412 (Linking). doi: 10.1111/j.1365-313X.2010.04314.x.

[132] S. Farrona, F. L. Thorpe, J. Engelhorn, J. Adrian, X. Dong, L. Sarid-Krebs, J. Goodrich, and F. Turck. Tissue-specific expression of flowering locus t in arabidopsis is maintained independently of polycomb group protein repression. *Plant Cell*, 23 (9):3204–14, 2011. ISSN 1532-298X (Electronic) 1040-4651 (Linking). doi: 10.1105/tpc.111.087809.

[133] C. A. Esmon, U. V. Pedmale, and E. Liscum. Plant tropisms: providing the power of movement to a sessile organism. *Int. J. Dev. Biol.*, 49(5-6):665–674, 2005.

[134] K. S. Heyndrickx and K. Vandepoele. Systematic identification of functional plant modules through the integration of complementary data sources. *Plant physiology*, 159(3):884–901, 2012. ISSN 1532-2548 (Electronic) 0032-0889 (Linking). doi: 10.1104/pp.112.196725.

[135] S. M. Brady and N. J. Provart. Web-queryable large-scale data sets for hypothesis generation in plant biology. *Plant Cell*, 21 (4):1034–51, 2009. ISSN 1040-4651 (Print) 1040-4651 (Linking). doi: tpc.109.066050[pii]10.1105/tpc.109.066050.

[136] Y. A. Kourmpetis, A. D. van Dijk, R. C. van Ham, and C. J. ter Braak. Genome-wide computational function prediction of arabidopsis proteins by integration of multiple data sources. *Plant Physiol*, 155(1):271–81, 2010. ISSN 1532-2548 (Electronic) 0032-0889 (Linking). doi: pp.110.162164[pii]10.1104/pp.110.162164.

[137] A. L. Barabasi and Z. N. Oltvai. Network biology: understanding the cell's functional organization. *Nat Rev Genet*, 5(2): 101–13, 2004. ISSN 1471-0056 (Print) 1471-0056 (Linking). doi: 10.1038/nrg1272nrg1272[pii].

[138] J. Lisso, D. Steinhauser, T. Altmann, J. Kopka, and C. Mussig. Identification of brassinosteroid-related genes by means of transcript co-response analyses. *Nucleic Acids Res*, 33 (8):2685–96, 2005. ISSN 1362-4962 (Electronic) 0305-1048 (Linking). doi: 33/8/2685[pii]10.1093/nar/gki566.

[139] K. Horan, C. Jang, J. Bailey-Serres, R. Mittler, C. Shelton, J. F. Harper, J. K. Zhu, J. C. Cushman, M. Gollery, and T. Girke. Annotating genes of known and unknown function by large-scale coexpression analysis. *Plant Physiol*, 147(1): 41–57, 2008. ISSN 0032-0889 (Print) 0032-0889 (Linking). doi: pp.108.117366[pii]10.1104/pp.108.117366.

[140] O. Atias, B. Chor, and D. A. Chamovitz. Large-scale analysis of arabidopsis transcription reveals a basal co-regulation network. *BMC Syst Biol*, 3:86, 2009. ISSN 1752-0509 (Electronic) 1752-0509 (Linking). doi: 1752-0509-3-86[pii]10.1186/1752-0509-3-86.

[141] J. Geisler-Lee, N. O'Toole, R. Ammar, N. J. Provart, A. H. Millar, and M. Geisler. A predicted interactome for arabidopsis. *Plant Physiol*, 145(2):317–29, 2007. ISSN 0032-0889 (Print) 0032-0889 (Linking). doi: pp.107.103465[pii]10.1104/pp.107.103465.

[142] J. Boruc, H. Van den Daele, J. Hollunder, S. Rombauts, E. Mylle, P. Hilson, D. Inze, L. De Veylder, and E. Russinova. Functional modules in the arabidopsis core cell cycle binary protein-protein interaction network. *Plant Cell*, 22(4):1264–80, 2010. ISSN 1532-298X (Electronic) 1040-4651 (Linking). doi: tpc.109.073635[pii]10.1105/tpc.109.073635.

[143] Arabidopsis Interactome Mapping Consortium. Evidence for network evolution in an arabidopsis interactome map. *Science*, 333(6042):601–7, 2011. ISSN 1095-9203 (Electronic) 0036-8075 (Linking). doi: 333/6042/601[pii]10.1126/science.1203877.

[144] T. Ferrier, J. T. Matus, J. Jin, and J. L. Riechmann. Arabidopsis paves the way: genomic and network analyses in crops. *Curr Opin Biotechnol*, 22(2):260–70, 2011. ISSN 1879-0429 (Electronic) 0958-1669 (Linking). doi: S0958-1669(10)00228-4[pii]10.1016/j.copbio.2010.11.010.

[145] S. Persson, H. Wei, J. Milne, G. P. Page, and C. R. Somerville. Identification of genes required for cellulose synthesis by regression analysis of public microarray data sets. *Proc Natl Acad Sci U S A*, 102(24):8633–8, 2005. ISSN 0027-8424 (Print) 0027-8424 (Linking). doi: 0503392102[pii]10.1073/pnas.0503392102.

[146] H. Wei, S. Persson, T. Mehta, V. Srinivasasainagendra, L. Chen, G. P. Page, C. Somerville, and A. Loraine. Transcriptional coordination of the metabolic network in arabidopsis. *Plant Physiol*, 142(2):762–74, 2006. ISSN 0032-0889 (Print) 0032-0889 (Linking). doi: pp.106.080358[pii]10.1104/pp.106.080358.

[147] C. J. Wolfe, I. S. Kohane, and A. J. Butte. Systematic survey reveals general applicability of "guilt-by-association" within gene coexpression networks. *BMC Bioinformatics*, 6:227, 2005. ISSN 1471-2105 (Electronic) 1471-2105 (Linking). doi: 1471-2105-6-227[pii]10.1186/1471-2105-6-227.

[148] K. Vandepoele, T. Casneuf, and Y. Van de Peer. Identification of novel regulatory modules in dicotyledonous plants using expression data and comparative genomics. *Genome Biol*, 7(11): R103, 2006. ISSN 1465-6914 (Electronic) 1465-6906 (Linking). doi: 10.1186/gb-2006-7-11-r103.

[149] S. Ma and H. J. Bohnert. Integration of arabidopsis thaliana stress-related transcript profiles, promoter structures, and cell-specific expression. *Genome Biol*, 8(4):R49, 2007. ISSN 1465-6914 (Electronic) 1465-6906 (Linking). doi: gb-2007-8-4-r49[pii]10.1186/gb-2007-8-4-r49.

[150] K. Vandepoele, M. Quimbaya, T. Casneuf, L. De Veylder, and Y. Van de Peer. Unraveling transcriptional control in arabidopsis using cis-regulatory elements and coexpression networks. *Plant Physiol*, 150(2):535–46, 2009. ISSN 0032-0889 (Print) 0032-0889 (Linking). doi: pp.109.136028[pii]10.1104/pp.109.136028.

[151] T. P. Michael, T. C. Mockler, G. Breton, C. McEntee, A. Byer, J. D. Trout, S. P. Hazen, R. Shen, H. D. Priest, C. M. Sullivan, S. A. Givan, M. Yanovsky, F. Hong, S. A. Kay, and J. Chory. Network discovery pipeline elucidates conserved time-of-day-specific cis-regulatory modules. *PLoS Genet*, 4(2): e14, 2008. ISSN 1553-7404 (Electronic) 1553-7390 (Linking). doi: 07-PLGE-RA-0600[pii]10.1371/journal.pgen.0040014.

[152] M. Tompa, N. Li, T. L. Bailey, G. M. Church, B. De Moor, E. Eskin, A. V. Favorov, M. C. Frith, Y. Fu, W. J. Kent, V. J. Makeev, A. A. Mironov, W. S. Noble, G. Pavesi, G. Pesole, M. Regnier, N. Simonis, S. Sinha, G. Thijs, J. van Helden, M. Vandenbogaert, Z. Weng, C. Workman, C. Ye, and Z. Zhu. Assessing computational tools for the discovery of transcription factor binding sites. *Nat Biotechnol*, 23(1):137–44, 2005. ISSN 1087-0156 (Print) 1087-0156 (Linking). doi: nbt1053[pii]10.1038/nbt1053.

[153] J. M. Stuart, E. Segal, D. Koller, and S. K. Kim. A gene-coexpression network for global discovery of conserved genetic modules. *Science*, 302(5643):249–55, 2003. ISSN 1095-9203 (Electronic) 0036-8075 (Linking). doi: 10.1126/science.10874471087447[pii].

[154] J. Lee, K. He, V. Stolc, H. Lee, P. Figueroa, Y. Gao, W. Tongprasit, H. Zhao, I. Lee, and X. W. Deng. Analysis of transcription factor hy5 genomic binding sites revealed its hierarchical role in light regulation of development. *Plant Cell*, 19(3):731–49, 2007. ISSN 1040-4651 (Print) 1040-4651 (Linking). doi: tpc.106.047688[pii]10.1105/tpc.106.047688.

[155] K. Morohashi and E. Grotewold. A systems approach reveals regulatory circuitry for arabidopsis trichome initiation by the gl3 and gl1 selectors. *PLoS Genet*, 5(2):e1000396, 2009. ISSN 1553-7404 (Electronic) 1553-7390 (Linking). doi: 10.1371/journal.pgen.1000396.

[156] W. Busch, A. Miotk, F. D. Ariel, Z. Zhao, J. Forner, G. Daum, T. Suzaki, C. Schuster, S. J. Schultheiss, A. Leibfried, S. Haubeiss, N. Ha, R. L. Chan, and J. U. Lohmann. Transcriptional control of a plant stem cell niche. *Dev Cell*, 18(5): 849–61, 2010. ISSN 1878-1551 (Electronic) 1534-5807 (Linking). doi: S1534-5807(10)00151-6[pii]10.1016/j.devcel.2010.03.012.

[157] L. Yant, J. Mathieu, T. T. Dinh, F. Ott, C. Lanz, H. Wollmann, X. Chen, and M. Schmid. Orchestration of the floral transition and floral development in arabidopsis by the bifunctional transcription factor apetala2. *Plant Cell*, 22(7):2156–70, 2010. ISSN 1532-298X (Electronic) 1040-4651 (Linking). doi: tpc.110.075606[pii]10.1105/tpc.110.075606.

[158] K. Kaufmann, J. M. Muino, R. Jauregui, C. A. Airoldi, C. Smaczniak, P. Krajewski, and G. C. Angenent. Target genes of the mads transcription factor sepallata3: integration of developmental and hormonal pathways in the arabidopsis flower. *PLoS Biol*, 7(4):e1000090, 2009. ISSN 1545-7885 (Electronic) 1544-9173 (Linking). doi: 08-PLBI-RA-4384[pii]10.1371/journal.pbio.1000090.

[159] J. Mathieu, L. J. Yant, F. Murdter, F. Kuttner, and M. Schmid. Repression of flowering by the mir172 target smz. *PLoS Biol*, 7 (7):e1000148, 2009. ISSN 1545-7885 (Electronic) 1544-9173 (Linking). doi: 10.1371/journal.pbio.1000148.

[160] K. Kaufmann, F. Wellmer, J. M. Muino, T. Ferrier, S. E. Wuest, V. Kumar, A. Serrano-Mislata, F. Madueno, P. Krajewski, E. M. Meyerowitz, G. C. Angenent, and J. L. Riechmann. Orchestration of floral initiation by apetala1. *Science*, 328(5974):85–9, 2010. ISSN 1095-9203 (Electronic) 0036-8075 (Linking). doi: 328/5974/85[pii]10.1126/science.1185244.

[161] Y. Fujikawa and N. Kato. Split luciferase complementation assay to study protein-protein interactions in arabidopsis protoplasts. *Plant J*, 52(1):185–95, 2007. ISSN 0960-7412 (Print) 0960-7412 (Linking). doi: TPJ3214[pii]10.1111/j.1365-313X.2007.03214.x.

[162] J. Van Leene, H. Stals, D. Eeckhout, G. Persiau, E. Van De Slijke, G. Van Isterdael, A. De Clercq, E. Bonnet, K. Laukens, N. Remmerie, K. Henderickx, T. De Vijlder, A. Abdelkrim, A. Pharazyn, H. Van Onckelen, D. Inze, E. Witters, and G. De Jaeger. A tandem affinity purification-based technology platform to study the cell cycle interactome in arabidopsis thaliana. *Mol Cell Proteomics*, 6(7):1226–38, 2007. ISSN 1535-9476 (Print) 1535-9476 (Linking). doi: M700078-MCP200[pii]10.1074/mcp.M700078-MCP200.

[163] J. F. Li, J. Bush, Y. Xiong, L. Li, and M. McCormack. Large-scale protein-protein interaction analysis in arabidopsis mesophyll protoplasts by split firefly luciferase complementation. *PLoS One*, 6(11):e27364, 2011. ISSN 1932-6203 (Electronic) 1932-6203 (Linking). doi: 10.1371/journal.pone.0027364PONE-D-11-13286[pii].

[164] S. De Bodt, S. Proost, K. Vandepoele, P. Rouze, and Y. Van de Peer. Predicting protein-protein interactions in arabidopsis thaliana through integration of orthology, gene ontology and co-expression. *BMC Genomics*, 10:288, 2009. ISSN 1471-2164 (Electronic) 1471-2164 (Linking). doi: 1471-2164-10-288[pii]10.1186/1471-2164-10-288.

[165] D. Warde-Farley, S. L. Donaldson, O. Comes, K. Zuberi, R. Badrawi, P. Chao, M. Franz, C. Grouios, F. Kazi, C. T. Lopes, A. Maitland, S. Mostafavi, J. Montojo, Q. Shao, G. Wright, G. D. Bader, and Q. Morris. The genemania prediction server: biological network integration for gene prioritization and predicting gene function. *Nucleic Acids Research*, 38(Web Server issue):W214–20, 2010. ISSN 1362-4962 (Electronic) 0305-1048 (Linking). doi: 10.1093/nar/gkq537.

[166] J. R. Bradford, C. J. Needham, P. Tedder, M. A. Care, A. J. Bulpitt, and D. R. Westhead. Go-at: in silico prediction of gene function in arabidopsis thaliana by combining heterogeneous data. *Plant J*, 61(4):713–21, 2010. ISSN 1365-313X (Electronic) 0960-7412 (Linking). doi: TPJ4097[pii]10.1111/j.1365-313X.2009.04097.x.

[167] S. P. Ficklin and F. A. Feltus. Gene coexpression network alignment and conservation of gene modules between two grass species: Maize and rice. *Plant Physiol*, 156(3):1244–56, 2011. ISSN 1532-2548 (Electronic) 0032-0889 (Linking). doi: pp.111.173047[pii]10.1104/pp.111.173047.

[168] S. Movahedi, Y. Van de Peer, and K. Vandepoele. Comparative network analysis reveals that tissue specificity and gene function are important factors influencing the mode of expression evolution in arabidopsis and rice. *Plant Physiol*, 156(3):1316–30, 2011. ISSN 1532-2548 (Electronic) 0032-0889 (Linking). doi: pp.111.177865[pii]10.1104/pp.111.177865.

[169] M. Mutwil, S. Klie, T. Tohge, F. M. Giorgi, O. Wilkins, M. M. Campbell, A. R. Fernie, B. Usadel, Z. Nikoloski, and S. Persson. Planet: combined sequence and expression comparisons across plant networks derived from seven species. *Plant Cell*, 23(3):895–910, 2011. ISSN 1532-298X (Electronic) 1040-4651 (Linking). doi: tpc.111.083667[pii]10.1105/tpc.111.083667.

[170] S. Movahedi, M. Van Bel, K. S. Heyndrickx, and K. Vandepoele. Comparative co-expression analysis in plant biology. *Plant, cell & environment*, 35(10):1787–98, 2012. ISSN 1365-3040 (Electronic) 0140-7791 (Linking). doi: 10.1111/j.1365-3040.2012.02517.x.

[171] T. Z. Berardini, S. Mundodi, L. Reiser, E. Huala, M. Garcia-Hernandez, P. Zhang, L. A. Mueller, J. Yoon, A. Doyle, G. Lander, N. Moseyko, D. Yoo, I. Xu, B. Zoeckler, M. Montoya, N. Miller, D. Weems, and S. Y. Rhee. Functional annotation of the arabidopsis genome using controlled vocabularies. *Plant Physiol*, 135(2):745–55, 2004. ISSN 0032-0889 (Print) 0032-0889 (Linking). doi: 10.1104/pp.104.040071pp.104.040071[pii].

[172] B. Usadel, T. Obayashi, M. Mutwil, F. M. Giorgi, G. W. Bassel, M. Tanimoto, A. Chow, D. Steinhauser, S. Persson, and N. J. Provart. Co-expression tools for plant biology: opportunities for hypothesis generation and caveats. *Plant Cell Environ*, 32(12):1633–51, 2009. ISSN 1365-3040 (Electronic) 0140-7791 (Linking). doi: PCE2040[pii]10.1111/j.1365-3040.2009.02040.x.

[173] M. Mutwil, B. Usadel, M. Schutte, A. Loraine, O. Ebenhoh, and S. Persson. Assembly of an interactive correlation network for the arabidopsis genome using a novel heuristic clustering algorithm. *Plant Physiol*, 152(1):29–43, 2010. ISSN 1532-2548 (Electronic) 0032-0889 (Linking). doi: pp.109.145318[pii]10.1104/pp.109.145318.

[174] G. Pavesi, G. Mauri, and G. Pesole. An algorithm for finding signals of unknown length in dna sequences. *Bioinformatics*, 17 Suppl 1:S207–14, 2001. ISSN 1367-4803 (Print) 1367-4803 (Linking).

[175] G. Thijs, K. Marchal, M. Lescot, S. Rombauts, B. De Moor, P. Rouze, and Y. Moreau. A gibbs sampling method to detect overrepresented motifs in the upstream regions of coexpressed genes. *J Comput Biol*, 9(2):447–64, 2002. ISSN 1066-5277 (Print) 1066-5277 (Linking). doi: 10.1089/10665270252935566.

[176] K. Higo, Y. Ugawa, M. Iwamoto, and T. Korenaga. Plant cis-acting regulatory dna elements (place) database: 1999. *Nucleic acids research*, 27(1):297–300, 1999. ISSN 0305-1048 (Print) 0305-1048 (Linking). doi: gkc003[pii].

[177] L. Mao, J. L. Van Hemert, S. Dash, and J. A. Dickerson. Arabidopsis gene co-expression network and its functional modules. *BMC Bioinformatics*, 10:346, 2009. ISSN 1471-2105 (Electronic) 1471-2105 (Linking). doi: 1471-2105-10-346[pii]10.1186/1471-2105-10-346.

[178] W. S. Chao, M. E. Foley, M. Dogramaci, J. V. Anderson, and D. P. Horvath. Alternating temperature breaks dormancy in leafy spurge seeds and impacts signaling networks associated with hy5. *Funct Integr Genomics*, 11(4):637–49, 2011. ISSN 1438-7948 (Electronic) 1438-793X (Linking). doi: 10.1007/s10142-011-0253-0.

[179] G. E. Zinman, S. Zhong, and Z. Bar-Joseph. Biological interaction networks are conserved at the module level. *BMC systems biology*, 5:134, 2011. ISSN 1752-0509 (Electronic) 1752-0509 (Linking). doi: 10.1186/1752-0509-5-134.

[180] M. Van Bel, S. Proost, E. Wischnitzki, S. Movahedi, C. Scheerlinck, Y. Van de Peer, and K. Vandepoele. Dissecting plant genomes with the plaza comparative genomics platform. *Plant physiology*, 158(2):590–600, 2012. ISSN 1532-2548 (Electronic) 0032-0889 (Linking). doi: 10.1104/pp.111.189514.

[181] G. W. Bassel, H. Lan, E. Glaab, D. J. Gibbs, T. Gerjets, N. Krasnogor, A. J. Bonner, M. J. Holdsworth, and N. J. Provart. Genome-wide network model capturing seed germination reveals coordinated regulation of plant cellular phase transitions. *Proc Natl Acad Sci U S A*, 108(23):9709–14, 2011. ISSN 1091-6490 (Electronic) 0027-8424 (Linking). doi: 1100958108[pii]10.1073/pnas.1100958108.

[182] J. Zhou, C. Lee, R. Zhong, and Z. H. Ye. Myb58 and myb63 are transcriptional activators of the lignin biosynthetic pathway during secondary cell wall formation in arabidopsis. *Plant Cell*, 21(1):248–66, 2009. ISSN 1040-4651 (Print) 1040-4651 (Linking). doi: tpc.108.063321[pii]10.1105/tpc.108.063321.

[183] D. Brown, R. Wightman, Z. Zhang, L. D. Gomez, I. Atanassov, J. P. Bukowski, T. Tryfona, S. J. McQueen-Mason, P. Dupree, and S. Turner. Arabidopsis genes irregular xylem (irx15) and irx15l encode duf579-containing proteins that are essential for normal xylan deposition in the secondary cell wall. *Plant J*, 66(3):401–13, 2011. ISSN 1365-313X (Electronic) 0960-7412 (Linking). doi: 10.1111/j.1365-313X.2011.04501.x.

[184] R. Zhong, C. Lee, J. Zhou, R. L. McCarthy, and Z. H. Ye. A battery of transcription factors involved in the regulation of secondary cell wall biosynthesis in arabidopsis. *Plant Cell*, 20(10): 2763–82, 2008. ISSN 1040-4651 (Print) 1040-4651 (Linking). doi: tpc.108.061325[pii]10.1105/tpc.108.061325.

[185] C. Lee, Q. Teng, R. Zhong, and Z. H. Ye. The four arabidopsis reduced wall acetylation genes are expressed in secondary wall-containing cells and required for the acetylation of xylan. *Plant Cell Physiol*, 52(8):1289–301, 2011. ISSN 1471-9053 (Electronic) 0032-0781 (Linking). doi: pcr075[pii]10.1093/pcp/pcr075.

[186] C. Yanhui, Y. Xiaoyuan, H. Kun, L. Meihua, L. Jigang, G. Zhaofeng, L. Zhiqiang, Z. Yunfei, W. Xiaoxiao, Q. Xiaoming, S. Yunping, Z. Li, D. Xiaohui, L. Jingchu, D. Xing-Wang, C. Zhangliang, G. Hongya, and Q. Li-Jia. The myb transcription factor superfamily of arabidopsis: expression analysis and phylogenetic comparison with the rice myb family. *Plant Mol Biol*, 60(1):107–24, 2006. ISSN 0167-4412 (Print) 0167-4412 (Linking). doi: 10.1007/s11103-005-2910-y.

[187] Y. Hirakawa, H. Shinohara, Y. Kondo, A. Inoue, I. Nakanomyo, M. Ogawa, S. Sawa, K. Ohashi-Ito, Y. Matsubayashi, and H. Fukuda. Non-cell-autonomous control of vascular stem cell fate by a cle peptide/receptor system. *Proc Natl Acad Sci U S A*, 105(39):15208–13, 2008. ISSN 1091-6490 (Electronic) 0027-8424 (Linking). doi: 0808444105[pii]10.1073/pnas.0808444105.

[188] M. Quimbaya, K. Vandepoele, E. Raspe, M. Matthijs, S. Dhondt, G. T. Beemster, G. Berx, and L. De Veylder. Identification of putative cancer genes through data integration and comparative genomics between plants and humans. *Cell Mol Life Sci*, 2012. ISSN 1420-9071 (Electronic) 1420-682X (Linking). doi: 10.1007/s00018-011-0909-x.

[189] J. D. Leverson, H. K. Huang, S. L. Forsburg, and T. Hunter. The schizosaccharomyces pombe aurora-related kinase ark1 interacts with the inner centromere protein pic1 and mediates chromosome segregation and cytokinesis. *Mol Biol Cell*, 13(4):1132–43, 2002. ISSN 1059-1524 (Print) 1059-1524 (Linking). doi: 10.1091/mbc.01-07-0330.

[190] B. Farinas and P. Mas. Functional implication of the myb transcription factor rve8/lcl5 in the circadian control of histone acetylation. *Plant J*, 66(2):318–29, 2011. ISSN 1365-313X (Electronic) 0960-7412 (Linking). doi: 10.1111/j.1365-313X.2011.04484.x.

[191] M. V. Arana, N. Marin-de la Rosa, J. N. Maloof, M. A. Blazquez, and D. Alabadi. Circadian oscillation of gibberellin signaling in arabidopsis. *Proc Natl Acad Sci U S A*, 108(22):9292–7, 2011. ISSN 1091-6490 (Electronic) 0027-8424 (Linking). doi: 1101050108[pii]10.1073/pnas.1101050108.

[192] A. Lysenko, M. Defoin-Platel, K. Hassani-Pak, J. Taubert, C. Hodgman, C. J. Rawlings, and M. Saqi. Assessing the functional coherence of modules found in multiple-evidence networks from arabidopsis. *BMC Bioinformatics*, 12(1):203, 2011. ISSN 1471-2105 (Electronic) 1471-2105 (Linking). doi: 1471-2105-12-203[pii]10.1186/1471-2105-12-203.

[193] J. H. Cho, K. Wang, and D. J. Galas. An integrative approach to inferring biologically meaningful gene modules. *BMC Syst Biol*, 5(1):117, 2011. ISSN 1752-0509 (Electronic) 1752-0509 (Linking). doi: 1752-0509-5-117[pii]10.1186/1752-0509-5-117.

[194] P. Adler, R. Kolde, M. Kull, A. Tkachenko, H. Peterson, J. Reimand, and J. Vilo. Mining for coexpression across hundreds of datasets using novel rank aggregation and visualization methods. *Genome biology*, 10(12):R139, 2009. ISSN 1465-6914 (Electronic) 1465-6906 (Linking). doi: gb-2009-10-12-r139[pii]10.1186/gb-2009-10-12-r139.

[195] M. Freeling, L. Rapaka, E. Lyons, B. Pedersen, and B. C. Thomas. G-boxes, bigfoot genes, and environmental response: characterization of intragenomic conserved noncoding sequences in arabidopsis. *Plant Cell*, 19(5):1441–57, 2007. ISSN 1040-4651 (Print) 1040-4651 (Linking). doi: tpc.107.050419[pii]10.1105/tpc.107.050419.

[196] R. A. Irizarry, B. Hobbs, F. Collin, Y. D. Beazer-Barclay, K. J. Antonellis, U. Scherf, and T. P. Speed. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics*, 4(2):249–64, 2003. ISSN 1465-4644 (Print) 1465-4644 (Linking). doi: 10.1093/biostatistics/4.2.2494/2/249[pii].

[197] S. Proost, M. Van Bel, L. Sterck, K. Billiau, T. Van Parys, Y. Van de Peer, and K. Vandepoele. Plaza: a comparative genomics resource to study gene and genome evolution in plants. *The Plant cell*, 21(12):3718–31, 2009. ISSN 1532-298X (Electronic) 1040-4651 (Linking). doi: tpc.109.071506[pii]10.1105/tpc.109.071506.

[198] Y. Pilpel, P. Sudarsanam, and G. M. Church. Identifying regulatory networks by combinatorial analysis of promoter elements. *Nature genetics*, 29(2):153–9, 2001. ISSN 1061-4036 (Print) 1061-4036 (Linking). doi: 10.1038/ng724ng724[pii].

[199] J. D. Storey and R. Tibshirani. Statistical significance for genomewide studies. *Proceedings of the National Academy of Sciences of the United States of America*, 100(16):9440–5, 2003. ISSN 0027-8424 (Print) 0027-8424 (Linking). doi: 10.1073/pnas.15305091001530509100[pii].

[200] P. Shannon, A. Markiel, O. Ozier, N. S. Baliga, J. T. Wang, D. Ramage, N. Amin, B. Schwikowski, and T. Ideker. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome research*, 13(11):2498–504, 2003. ISSN 1088-9051 (Print) 1088-9051 (Linking). doi: 10.1101/gr.1239303.

[201] K. S. Heyndrickx, J. Van de Velde, C. Wang, D. Weigel, and K. Vandepoele. A functional and evolutionary perspective on transcription factor binding in arabidopsis thaliana. *Plant Cell*, 26(10):3894–910, 2014. ISSN 1532-298X (Electronic) 1040-4651 (Linking). doi: 10.1105/tpc.114.130591. URL `http://www.ncbi.nlm.nih.gov/pubmed/25361952`. Heyndrickx, Ken S Van de Velde, Jan Wang, Congmao Weigel, Detlef Vandepoele, Klaas eng 2014/11/02 06:00 Plant Cell. 2014 Oct;26(10):3894-910. doi: 10.1105/tpc.114.130591. Epub 2014 Oct 31.

[202] G. A. Wray, M. W. Hahn, E. Abouheif, J. P. Balhoff, M. Pizer, M. V. Rockman, and L. A. Romano. The evolution of transcriptional regulation in eukaryotes. *Mol Biol Evol*, 20(9):1377–419, 2003. ISSN 0737-4038 (Print) 0737-4038 (Linking). doi: 10.1093/molbev/msg140msg140[pii].

[203] S. MacArthur, X. Y. Li, J. Li, J. B. Brown, H. C. Chu, L. Zeng, B. P. Grondona, A. Hechmer, L. Simirenko, S. V. Keranen, D. W. Knowles, M. Stapleton, P. Bickel, M. D. Biggin, and M. B. Eisen. Developmental roles of 21 drosophila transcription factors are determined by quantitative differences in binding to an overlapping set of thousands of genomic regions. *Genome biology*, 10(7):R80, 2009. ISSN 1465-6914 (Electronic) 1465-6906 (Linking). doi: 10.1186/gb-2009-10-7-r80.

[204] L. Teytelman, D. M. Thurtle, J. Rine, and A. van Oudenaarden. Highly expressed loci are vulnerable to misleading chip localization of multiple unrelated proteins. *Proceedings of the National Academy of Sciences of the United States of America*, 2013. ISSN 1091-6490 (Electronic) 0027-8424 (Linking). doi: 10.1073/pnas.1316064110.

[205] E. L. Van Nostrand and S. K. Kim. Integrative analysis of c. elegans modencode chip-seq data sets to infer gene regulatory interactions. *Genome research*, 23(6):941–53, 2013. ISSN 1549-5469 (Electronic) 1088-9051 (Linking). doi: 10.1101/gr.152876.112.

[206] M. B. Gerstein, Z. J. Lu, E. L. Van Nostrand, C. Cheng, B. I. Arshinoff, T. Liu, K. Y. Yip, R. Robilotto, A. Rechtsteiner, K. Ikegami, P. Alves, A. Chateigner, M. Perry, M. Morris, R. K. Auerbach, X. Feng, J. Leng, A. Vielle, W. Niu, K. Rhrissorrakrai, A. Agarwal, R. P. Alexander, G. Barber, C. M. Brdlik, J. Brennan, J. J. Brouillet, A. Carr, M. S. Cheung, H. Clawson, S. Contrino, L. O. Dannenberg, A. F. Dernburg, A. Desai, L. Dick, A. C. Dose, J. Du, T. Egelhofer, S. Ercan, G. Euskirchen, B. Ewing, E. A. Feingold, R. Gassmann, P. J. Good, P. Green, F. Gullier, M. Gutwein, M. S. Guyer, L. Habegger, T. Han, J. G. Henikoff, S. R. Henz, A. Hinrichs, H. Holster, T. Hyman, A. L. Iniguez, J. Janette, M. Jensen, M. Kato, W. J. Kent, E. Kephart, V. Khivansara, E. Khurana, J. K. Kim, P. Kolasinska-Zwierz, E. C. Lai, I. Latorre,

A. Leahey, S. Lewis, P. Lloyd, L. Lochovsky, R. F. Lowdon, Y. Lubling, R. Lyne, M. MacCoss, S. D. Mackowiak, M. Mangone, S. McKay, D. Mecenas, G. Merrihew, 3rd Miller, D. M., A. Muroyama, J. I. Murray, S. L. Ooi, H. Pham, T. Phippen, E. A. Preston, N. Rajewsky, G. Ratsch, H. Rosenbaum, J. Rozowsky, K. Rutherford, P. Ruzanov, M. Sarov, R. Sasidharan, A. Sboner, P. Scheid, E. Segal, H. Shin, C. Shou, F. J. Slack, et al. Integrative analysis of the caenorhabditis elegans genome by the modencode project. *Science*, 330(6012):1775–87, 2010. ISSN 1095-9203 (Electronic) 0036-8075 (Linking). doi: 10.1126/science.1196914.

[207] C. Cheng, K. K. Yan, W. Hwang, J. Qian, N. Bhardwaj, J. Rozowsky, Z. J. Lu, W. Niu, P. Alves, M. Kato, M. Snyder, and M. Gerstein. Construction and analysis of an integrated regulatory network derived from high-throughput sequencing data. *PLoS computational biology*, 7(11):e1002190, 2011. ISSN 1553-7358 (Electronic) 1553-734X (Linking). doi: 10.1371/journal.pcbi.1002190.

[208] S. Roy, J. Ernst, P. V. Kharchenko, P. Kheradpour, N. Negre, M. L. Eaton, J. M. Landolin, C. A. Bristow, L. Ma, M. F. Lin, S. Washietl, B. I. Arshinoff, F. Ay, P. E. Meyer, N. Robine, N. L. Washington, L. Di Stefano, E. Berezikov, C. D. Brown, R. Candeias, J. W. Carlson, A. Carr, I. Jungreis, D. Marbach, R. Sealfon, M. Y. Tolstorukov, S. Will, A. A. Alekseyenko, C. Artieri, B. W. Booth, A. N. Brooks, Q. Dai, C. A. Davis, M. O. Duff, X. Feng, A. A. Gorchakov, T. Gu, J. G. Henikoff, P. Kapranov, R. Li, H. K. MacAlpine, J. Malone, A. Minoda, J. Nordman, K. Okamura, M. Perry, S. K. Powell, N. C. Riddle, A. Sakai, A. Samsonova, J. E. Sandler, Y. B. Schwartz, N. Sher, R. Spokony, D. Sturgill, M. van Baren, K. H. Wan, L. Yang, C. Yu, E. Feingold, P. Good, M. Guyer, R. Lowdon, K. Ahmad, J. Andrews, B. Berger, S. E. Brenner, M. R. Brent, L. Cherbas, S. C. Elgin, T. R. Gingeras, R. Grossman, R. A. Hoskins, T. C. Kaufman, W. Kent, M. I. Kuroda, T. Orr-Weaver, N. Perrimon, V. Pirrotta, J. W. Posakony, B. Ren, S. Russell, P. Cherbas, B. R. Graveley, S. Lewis, G. Micklem, B. Oliver, P. J. Park, S. E. Celniker, S. Henikoff, G. H. Karpen, E. C. Lai, D. M. MacAlpine, L. D. Stein, K. P. White, and M. Kellis. Identification of functional elements and regulatory circuits by drosophila modencode. *Science*, 330(6012):1787–97, 2010. ISSN 1095-9203 (Electronic) 0036-8075 (Linking). doi: 10.1126/science.1198374.

[209] N. Negre, C. D. Brown, L. Ma, C. A. Bristow, S. W. Miller, U. Wagner, P. Kheradpour, M. L. Eaton, P. Loriaux, R. Sealfon, Z. Li, H. Ishii, R. F. Spokony, J. Chen, L. Hwang, C. Cheng, R. P. Auburn, M. B. Davis, M. Domanus, P. K. Shah, C. A. Morrison, J. Zieba, S. Suchy, L. Senderowicz, A. Victorsen, N. A. Bild, A. J. Grundstad, D. Hanley, D. M. MacAlpine, M. Mannervik, K. Venken, H. Bellen, R. White, M. Gerstein, S. Russell, R. L. Grossman, B. Ren, J. W. Posakony, M. Kellis, and K. P. White. A cis-regulatory map of the drosophila genome. *Nature*, 471(7339):527–31, 2011. ISSN 1476-4687 (Electronic) 0028-0836 (Linking). doi: 10.1038/nature09990.

[210] B. E. Bernstein, E. Birney, I. Dunham, E. D. Green, C. Gunter, and M. Snyder. An integrated encyclopedia of dna elements in the human genome. *Nature*, 489(7414):57–74, 2012. ISSN 1476-4687 (Electronic) 0028-0836 (Linking). doi: 10.1038/nature11247.

[211] M. B. Gerstein, A. Kundaje, M. Hariharan, S. G. Landt, K. K. Yan, C. Cheng, X. J. Mu, E. Khurana, J. Rozowsky, R. Alexander, R. Min, P. Alves, A. Abyzov, N. Addleman, N. Bhardwaj, A. P. Boyle, P. Cayting, A. Charos, D. Z. Chen, Y. Cheng, D. Clarke, C. Eastman, G. Euskirchen, S. Frietze, Y. Fu, J. Gertz, F. Grubert, A. Harmanci, P. Jain, M. Kasowski, P. Lacroute, J. Leng, J. Lian, H. Monahan, H. O'Geen, Z. Ouyang, E. C. Partridge, D. Patacsil, F. Pauli, D. Raha, L. Ramirez, T. E. Reddy, B. Reed, M. Shi, T. Slifer, J. Wang, L. Wu, X. Yang, K. Y. Yip, G. Zilberman-Schapira, S. Batzoglou, A. Sidow, P. J. Farnham, R. M. Myers, S. M. Weissman, and M. Snyder. Architecture of the human regulatory network derived from encode data. *Nature*, 489(7414):91–100, 2012. ISSN 1476-4687 (Electronic) 0028-0836 (Linking). doi: 10.1038/nature11245.

[212] X. Wang and L. Ma. Unraveling the circadian clock in arabidopsis. *Plant signaling & behavior*, 8(2), 2012. ISSN 1559-2324 (Electronic) 1559-2316 (Linking).

[213] D. Marbach, S. Roy, F. Ay, P. E. Meyer, R. Candeias, T. Kahveci, C. A. Bristow, and M. Kellis. Predictive regulatory models in drosophila melanogaster by integrative inference of transcriptional networks. *Genome research*, 22(7):1334–49, 2012. ISSN 1549-5469 (Electronic) 1088-9051 (Linking). doi: 10.1101/gr.127191.111.

[214] M. K. Mejia-Guerra, M. Pomeranz, K. Morohashi, and E. Grotewold. From plant gene regulatory grids to network dynamics. *Biochimica et biophysica acta*, 1819(5):454–465, 2012. ISSN 0006-3002 (Print) 0006-3002 (Linking). doi: 10.1016/j.bbagrm.2012.02.016.

[215] G. Li, X. Ruan, R. K. Auerbach, K. S. Sandhu, M. Zheng, P. Wang, H. M. Poh, Y. Goh, J. Lim, J. Zhang, H. S. Sim, S. Q. Peh, F. H. Mulawadi, C. T. Ong, Y. L. Orlov, S. Hong, Z. Zhang, S. Landt, D. Raha, G. Euskirchen, C. L. Wei, W. Ge, H. Wang, C. Davis, K. I. Fisher-Aylor, A. Mortazavi, M. Gerstein, T. Gingeras, B. Wold, Y. Sun, M. J. Fullwood, E. Cheung, E. Liu, W. K. Sung, M. Snyder, and Y. Ruan. Extensive promoter-centered chromatin interactions provide a topological basis for transcription regulation. *Cell*, 148(1-2):84–98, 2012. ISSN 1097-4172 (Electronic) 0092-8674 (Linking). doi: 10.1016/j.cell.2011.12.014. URL `http://www.ncbi.nlm.nih.gov/pubmed/22265404`. Li, Guoliang Ruan, Xiaoan Auerbach, Raymond K Sandhu, Kuljeet Singh Zheng, Meizhen Wang, Ping Poh, Huay Mei Goh, Yufen Lim, Joanne Zhang, Jingyao Sim, Hui Shan Peh, Su Qin Mulawadi, Fabianus Hendriyan Ong, Chin Thing Orlov, Yuriy L Hong, Shuzhen Zhang, Zhizhuo Landt, Steve Raha, Debasish Euskirchen, Ghia Wei, Chia-Lin Ge, Weihong Wang, Huaien Davis, Carrie Fisher-Aylor, Katherine I Mortazavi, Ali Gerstein, Mark Gingeras, Thomas Wold, Barbara Sun, Yi Fullwood, Melissa J Cheung, Edwin Liu, Edison Sung, Wing-Kin Snyder, Michael Ruan, Yijun eng HG004456/HG/NHGRI NIH HHS/ R01 HG004456/HG/NHGRI NIH HHS/ R01 HG004456-03/HG/NHGRI NIH HHS/ Research Support, N.I.H., Extramural Research Support, Non-U.S. Gov't 2012/01/24 06:00 Cell. 2012 Jan 20;148(1-2):84-98. doi: 10.1016/j.cell.2011.12.014.

[216] W. Zhang, T. Zhang, Y. Wu, and J. Jiang. Genome-wide identification of regulatory dna elements and protein-binding footprints using signatures of open chromatin in arabidopsis. *The Plant cell*, 2012. ISSN 1532-298X (Electronic) 1040-4651 (Linking). doi: 10.1105/tpc.112.098061.

[217] J. A. Higgins, P. C. Bailey, and D. A. Laurie. Comparative genomics of flowering time pathways using brachypodium distachyon as a model for the temperate grasses. *PloS one*, 5(4): e10065, 2010. ISSN 1932-6203 (Electronic) 1932-6203 (Linking). doi: 10.1371/journal.pone.0010065.

[218] A. B. Stergachis, E. Haugen, A. Shafer, W. Fu, B. Vernot, A. Reynolds, A. Raubitschek, S. Ziegler, E. M. LeProust, J. M. Akey, and J. A. Stamatoyannopoulos. Exonic transcription factor binding directs codon choice and affects protein evolution. *Science*, 342(6164):1367–72, 2013. ISSN 1095-9203 (Electronic) 0036-8075 (Linking). doi: 10.1126/science.1243490.

[219] L. Bulow, J. C. Bolivar, J. Ruhe, Y. Brill, and R. Hehl. 'microrna targets', a new athamap web-tool for genome-wide identification of mirna targets in arabidopsis thaliana. *BioData mining*, 5(1):7, 2012. ISSN 1756-0381 (Electronic) 1756-0381 (Linking). doi: 10.1186/1756-0381-5-7.

[220] J. Adrian, S. Farrona, J. J. Reimer, M. C. Albani, G. Coupland, and F. Turck. cis-regulatory elements and chromatin state coordinately control temporal and spatial expression of flowering locus t in arabidopsis. *Plant Cell*, 22(5):1425–40, 2010. ISSN 1532-298X (Electronic) 1040-4651 (Linking). doi: 10.1105/tpc.110.074682.

[221] S. Cao, R. W. Kumimoto, N. Gnesutta, A. M. Calogero, R. Mantovani, and 3rd Holt, B. F. A distal ccaat/nuclear factor y complex promotes chromatin looping at the flowering locus t promoter and regulates the timing of flowering in arabidopsis. *Plant Cell*, 26(3):1009–17, 2014. ISSN 1532-298X (Electronic) 1040-4651 (Linking). doi: 10.1105/tpc.113.120352.

[222] D. Weigel and R. Mott. The 1001 genomes project for arabidopsis thaliana. *Genome Biol*, 10(5):107, 2009. ISSN 1465-6914 (Electronic) 1465-6906 (Linking). doi: 10.1186/gb-2009-10-5-107.

[223] M. Nei and W. H. Li. Mathematical model for studying genetic variation in terms of restriction endonucleases. *Proc Natl Acad Sci U S A*, 76(10):5269–73, 1979. ISSN 0027-8424 (Print) 0027-8424 (Linking).

[224] A. Haudry, A. E. Platts, E. Vello, D. R. Hoen, M. Leclercq, R. J. Williamson, E. Forczek, Z. Joly-Lopez, J. G. Steffen, K. M. Hazzouri, K. Dewar, J. R. Stinchcombe, D. J. Schoen, X. Wang, J. Schmutz, C. D. Town, P. P. Edger, J. C. Pires, K. S. Schumaker, D. E. Jarvis, T. Mandakova, M. A. Lysak, E. van den Bergh, M. E. Schranz, P. M. Harrison, A. M. Moses, T. E. Bureau, S. I. Wright, and M. Blanchette. An atlas of over 90,000 conserved noncoding sequences provides insight into crucifer regulatory regions. *Nat Genet*, 45(8):891–8, 2013. ISSN 1546-1718 (Electronic) 1061-4036 (Linking). doi: 10.1038/ng.2684.

[225] J. Van de Velde, K. S. Heyndrickx, and K. Vandepoele. Inference of transcriptional networks in arabidopsis through conserved noncoding sequence analysis. *Plant Cell*, 2014. ISSN 1532-298X (Electronic) 1040-4651 (Linking). doi: 10.1105/tpc.114.127001.

[226] M. E. Williams, R. Foster, and N. H. Chua. Sequences flanking the hexameric g-box core cacgtg affect the specificity of protein binding. *Plant Cell*, 4(4):485–96, 1992. ISSN 1040-4651 (Print) 1040-4651 (Linking). doi: 10.1105/tpc.4.4.485.

[227] K. M. Catron, N. Iler, and C. Abate. Nucleotides flanking a conserved taat core dictate the dna binding specificity of three murine homeodomain proteins. *Molecular and cellular biology*, 13(4):2354–65, 1993. ISSN 0270-7306 (Print) 0270-7306 (Linking).

[228] M. Suzuki, M. G. Ketterling, and D. R. McCarty. Quantitative statistical analysis of cis-regulatory sequences in aba/vp1- and cbf/dreb1-regulated genes of arabidopsis. *Plant physiology*, 139(1):437–47, 2005. ISSN 0032-0889 (Print) 0032-0889 (Linking). doi: 10.1104/pp.104.058412.

[229] S. de Folter, R. G. Immink, M. Kieffer, L. Parenicova, S. R. Henz, D. Weigel, M. Busscher, M. Kooiker, L. Colombo, M. M. Kater, B. Davies, and G. C. Angenent. Comprehensive interaction map of the arabidopsis mads box transcription factors. *Plant Cell*, 17(5):1424–33, 2005. ISSN 1040-4651 (Print) 1040-4651 (Linking). doi: 10.1105/tpc.105.031831.

[230] R. Shalgi, D. Lieber, M. Oren, and Y. Pilpel. Global and local architecture of the mammalian microrna-transcription factor regulatory network. *PLoS Comput Biol*, 3(7):e131, 2007. ISSN 1553-7358 (Electronic). doi: 06-PLCB-RA-0543[pii] 10.1371/journal.pcbi.0030131.

[231] J. W. Foley and A. Sidow. Transcription-factor occupancy at hot regions quantitatively predicts rna polymerase recruitment in five human cell lines. *BMC Genomics*, 14(1):720, 2013. ISSN 1471-2164 (Electronic) 1471-2164 (Linking). doi: 10.1186/1471-2164-14-720.

[232] K. Y. Yip, C. Cheng, N. Bhardwaj, J. B. Brown, J. Leng, A. Kundaje, J. Rozowsky, E. Birney, P. Bickel, M. Snyder, and M. Gerstein. Classification of human genomic regions based on experimentally determined binding sites of more than 100 transcription-related factors. *Genome Biol*, 13(9):R48, 2012. ISSN 1465-6914 (Electronic) 1465-6906 (Linking). doi: 10.1186/gb-2012-13-9-r48.

[233] J. M. Franco-Zorrilla, I. Lopez-Vidriero, J. L. Carrasco, M. Godoy, P. Vera, and R. Solano. Dna-binding specificities of plant transcription factors and their potential to define target genes. *Proc Natl Acad Sci U S A*, 111(6):2367–72, 2014. ISSN 1091-6490 (Electronic) 0027-8424 (Linking). doi: 10.1073/pnas.1316278111.

[234] D. L. Wheeler, T. Barrett, D. A. Benson, S. H. Bryant, K. Canese, V. Chetvernin, D. M. Church, M. Dicuccio, R. Edgar, S. Federhen, M. Feolo, L. Y. Geer, W. Helmberg, Y. Kapustin, O. Khovayko, D. Landsman, D. J. Lipman, T. L. Madden, D. R. Maglott, V. Miller, J. Ostell, K. D. Pruitt, G. D. Schuler, M. Shumway, E. Sequeira, S. T. Sherry, K. Sirotkin, A. Souvorov, G. Starchenko, R. L. Tatusov, T. A. Tatusova, L. Wagner, and E. Yaschenko. Database resources of the national center for biotechnology information. *Nucleic Acids Res*, 36(Database issue):D13–21, 2008. ISSN 1362-4962 (Electronic) 0305-1048 (Linking). doi: 10.1093/nar/gkm1000.

[235] K. N. Chang, S. Zhong, M. T. Weirauch, G. Hon, M. Pelizzola, H. Li, S. S. Huang, R. J. Schmitz, M. A. Urich, D. Kuo, J. R. Nery, H. Qiao, A. Yang, A. Jamali, H. Chen, T. Ideker, B. Ren, Z. Bar-Joseph, T. R. Hughes, and J. R. Ecker. Temporal transcriptional response to ethylene gas drives growth hormone cross-regulation in arabidopsis. *eLife*, 2:e00675, 2013. ISSN 2050-084X (Electronic). doi: 10.7554/eLife.00675.

[236] R. C. Gentleman, V. J. Carey, D. M. Bates, B. Bolstad, M. Dettling, S. Dudoit, B. Ellis, L. Gautier, Y. Ge, J. Gentry, K. Hornik, T. Hothorn, W. Huber, S. Iacus, R. Irizarry, F. Leisch, C. Li, M. Maechler, A. J. Rossini, G. Sawitzki, C. Smith, G. Smyth, L. Tierney, J. Y. Yang, and J. Zhang. Bioconductor: open software development for computational biology and bioinformatics. *Genome biology*, 5(10):R80, 2004. ISSN 1465-6914 (Electronic) 1465-6906 (Linking). doi: 10.1186/gb-2004-5-10-r80.

[237] J. V. Turatsinze, M. Thomas-Chollier, M. Defrance, and J. van Helden. Using rsat to scan genome sequences for transcription factor binding sites and cis-regulatory modules. *Nature protocols*, 3(10):1578–88, 2008. ISSN 1750-2799 (Electronic) 1750-2799 (Linking). doi: 10.1038/nprot.2008.97.

[238] R. K. Dale, B. S. Pedersen, and A. R. Quinlan. Pybedtools: a flexible python library for manipulating genomic datasets and annotations. *Bioinformatics*, 27(24):3423–4, 2011. ISSN 1367-4811 (Electronic) 1367-4803 (Linking). doi: 10.1093/bioinformatics/btr539.

[239] S. De Bodt, J. Hollunder, H. Nelissen, N. Meulemeester, and D. Inze. Cornet 2.0: integrating plant coexpression, protein-protein interactions, regulatory interactions, gene associations and functional annotations. *The New phytologist*, 195(3):707–20, 2012. ISSN 1469-8137 (Electronic) 0028-646X (Linking). doi: 10.1111/j.1469-8137.2012.04184.x.

[240] E. Jones, E. Oliphant, and P. Peterson. Scipy: Open source scientific tools for python. 2001.

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

[242] M. M. Garner and A. Revzin. A gel electrophoresis method for quantifying the binding of proteins to specific dna regions: application to components of the escherichia coli lactose operon regulatory system. *Nucleic acids research*, 9(13):3047–60, 1981. ISSN 0305-1048 (Print) 0305-1048 (Linking).

[243] E. Roulet, S. Busso, A. A. Camargo, A. J. Simpson, N. Mermod, and P. Bucher. High-throughput SELEX SAGE method for quantitative modeling of transcription-factor binding sites. *Nat. Biotechnol.*, 20(8):831–835, Aug 2002.

[244] X. Meng, M. H. Brodsky, and S. A. Wolfe. A bacterial one-hybrid system for determining the DNA-binding specificity of transcription factors. *Nat. Biotechnol.*, 23(8):988–994, Aug 2005.

[245] T. Ferrier, J. T. Matus, J. Jin, and J. L. Riechmann. Arabidopsis paves the way: genomic and network analyses in crops. *Curr Opin Biotechnol*, 22(2):260–70, 2011. ISSN 1879-0429 (Electronic) 0958-1669 (Linking). doi: S0958-1669(10) 00228-4[pii]10.1016/j.copbio.2010.11.010.

[246] T. Handstad, M. B. Rye, F. Drablos, and P. Saetrom. A chip-seq benchmark shows that sequence conservation mainly improves detection of strong transcription factor binding sites. *PLoS One*, 6(4):e18430, 2011. ISSN 1932-6203 (Electronic) 1932-6203 (Linking). doi: 10.1371/journal.pone.0018430.

[247] N. J. Kaplinsky, D. M. Braun, J. Penterman, S. A. Goff, and M. Freeling. Utility and distribution of conserved noncoding sequences in the grasses. *Proceedings of the National Academy of Sciences of the United States of America*, 99(9):6147–51, 2002. ISSN 0027-8424 (Print) 0027-8424 (Linking). doi: 10.1073/pnas.052139599.

[248] H. Guo and S. P. Moose. Conserved noncoding sequences among cultivated cereal genomes identify candidate regulatory sequence elements and patterns of promoter evolution. *Plant Cell*, 15(5):1143–58, 2003. ISSN 1040-4651 (Print) 1040-4651 (Linking).

[249] D. C. Inada, A. Bashir, C. Lee, B. C. Thomas, C. Ko, S. A. Goff, and M. Freeling. Conserved noncoding sequences in the grasses. *Genome research*, 13(9):2030–41, 2003. ISSN 1088-9051 (Print) 1088-9051 (Linking). doi: 10.1101/gr.1280703.

[250] B. C. Thomas, L. Rapaka, E. Lyons, B. Pedersen, and M. Freeling. Arabidopsis intragenomic conserved noncoding sequence. *Proc. Natl. Acad. Sci. U.S.A.*, 104(9):3348–3353, Feb 2007.

[251] L. Baxter, A. Jironkin, R. Hickman, J. Moore, C. Barrington, P. Krusche, N. P. Dyer, V. Buchanan-Wollaston, A. Tiskin, J. Beynon, K. Denby, and S. Ott. Conserved noncoding sequences highlight shared components of regulatory networks in dicotyledonous plants. *The Plant cell*, 2012. ISSN 1532-298X (Electronic) 1040-4651 (Linking). doi: 10.1105/tpc.112.103010.

[252] D. Hupalo and A. D. Kern. Conservation and functional element discovery in 20 angiosperm plant genomes. *Mol. Biol. Evol.*, 30(7):1729–1744, Jul 2013.

[253] A. R. Reineke, E. Bornberg-Bauer, and J. Gu. Evolutionary divergence and limits of conserved non-coding sequence detection in plant genomes. *Nucleic Acids Res.*, 39(14):6029–6043, Aug 2011.

[254] S. De Bodt, G. Theissen, and Y. Van de Peer. Promoter analysis of MADS-box genes in eudicots through phylogenetic footprinting. *Mol. Biol. Evol.*, 23(6):1293–1303, Jun 2006.

[255] M. E. Chaboute, B. Clement, and G. Philipps. S phase and meristem-specific expression of the tobacco RNR1b gene is mediated by an E2F element located in the 5' leader sequence. *J. Biol. Chem.*, 277(20):17845–17851, May 2002.

[256] W. X. Liu, H. L. Liu, Z. J. Chai, X. P. Xu, Y. R. Song, and l. e. Q. Qu. Evaluation of seed storage-protein gene 5' untranslated regions in enhancing gene expression in transgenic rice seed. *Theor. Appl. Genet.*, 121(7):1267–1274, Nov 2010.

[257] J. Wang, M. Lu, C. Qiu, and Q. Cui. Transmir: a transcription factor-microrna regulation database. *Nucleic acids research*, 38(Database issue):D119–22, 2010. ISSN 1362-4962 (Electronic) 0305-1048 (Linking). doi: 10.1093/nar/gkp803.

[258] Q. He, A. F. Bardet, B. Patton, J. Purvis, J. Johnston, A. Paulson, M. Gogol, A. Stark, and J. Zeitlinger. High conservation of transcription factor binding and evidence for combinatorial regulation across six drosophila species. *Nat Genet*, 43(5):414–20, 2011. ISSN 1546-1718 (Electronic) 1061-4036 (Linking). doi: ng.808[pii]10.1038/ng.808.

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

[260] S. G. Hussey, E. Mizrachi, N. M. Creux, and A. A. Myburg. Navigating the transcriptional roadmap regulating plant secondary cell wall deposition. *Front Plant Sci*, 4:325, 2013.

[261] R. L. Hong, L. Hamaguchi, M. A. Busch, and D. Weigel. Regulatory elements of the floral homeotic gene AGAMOUS identified by phylogenetic footprinting and shadowing. *Plant Cell*, 15(6):1296–1309, Jun 2003.

[262] I. Rubio-Somoza and D. Weigel. Coordination of flower maturation by a regulatory circuit of three microRNAs. *PLoS Genet.*, 9(3):e1003374, Mar 2013.

[263] X. Ma and L. Gao. Biological network analysis: insights into structure and functions. *Briefings in functional genomics*, 11 (6):434–42, 2012. ISSN 2041-2657 (Electronic) 2041-2649 (Linking). doi: 10.1093/bfgp/els045.

[264] R. Zhong, C. Lee, and Z. H. Ye. Global analysis of direct targets of secondary wall NAC master switches in Arabidopsis. *Mol Plant*, 3(6):1087–1103, Nov 2010.

[265] K. Ohashi-Ito, Y. Oda, and H. Fukuda. Arabidopsis VASCULAR-RELATED NAC-DOMAIN6 directly regulates the genes that govern programmed cell death and secondary wall formation during xylem differentiation. *Plant Cell*, 22 (10):3461–3473, Oct 2010.

[266] M. Kubo, M. Udagawa, N. Nishikubo, G. Horiguchi, M. Yamaguchi, J. Ito, T. Mimura, H. Fukuda, and T. Demura. Transcription switches for protoxylem and metaxylem vessel formation. *Genes Dev.*, 19(16):1855–1860, Aug 2005.

[267] S. E. Wuest, D. S. O'Maoileidigh, L. Rae, K. Kwasniewska, A. Raganelli, K. Hanczaryk, A. J. Lohan, B. Loftus, E. Graciet, and F. Wellmer. Molecular basis for the specification of floral organs by apetala3 and pistillata. *Proceedings of the National Academy of Sciences of the United States of America*, 109(33):13452–7, 2012. ISSN 1091-6490 (Electronic) 0027-8424 (Linking). doi: 10.1073/pnas.1207075109.

[268] T. Jack, L. L. Brockman, and E. M. Meyerowitz. The homeotic gene APETALA3 of Arabidopsis thaliana encodes a MADS box and is expressed in petals and stamens. *Cell*, 68(4):683–697, Feb 1992.

[269] K. Goto and E. M. Meyerowitz. Function and regulation of the Arabidopsis floral homeotic gene PISTILLATA. *Genes Dev.*, 8 (13):1548–1560, Jul 1994.

[270] F. Brioudes, C. Joly, J. Szecsi, E. Varaud, J. Leroux, F. Bellvert, C. Bertrand, and M. Bendahmane. Jasmonate controls late development stages of petal growth in Arabidopsis thaliana. *Plant J.*, 60(6):1070–1080, Dec 2009.

[271] S. Song, T. Qi, H. Huang, and D. Xie. Regulation of stamen development by coordinated actions of jasmonate, auxin, and gibberellin in Arabidopsis. *Mol Plant*, 6(4):1065–1073, Jul 2013.

[272] T. Ito, K. H. Ng, T. S. Lim, H. Yu, and E. M. Meyerowitz. The homeotic protein AGAMOUS controls late stamen development by regulating a jasmonate biosynthetic gene in Arabidopsis. *Plant Cell*, 19(11):3516–3529, Nov 2007.

[273] P. Korkuc, J. H. Schippers, and D. Walther. Characterization and identification of cis-regulatory elements in Arabidopsis based on single-nucleotide polymorphism information. *Plant Physiol.*, 164(1):181–200, Jan 2014.

[274] J. L. Riechmann, B. A. Krizek, and E. M. Meyerowitz. Dimerization specificity of arabidopsis mads domain homeotic proteins apetala1, apetala3, pistillata, and agamous. *Proc Natl Acad Sci U S A*, 93(10):4793–8, 1996. ISSN 0027-8424 (Print) 0027-8424 (Linking).

[275] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman. Basic local alignment search tool. *J Mol Biol*, 215 (3):403–10, 1990. ISSN 0022-2836 (Print) 0022-2836 (Linking). doi: 10.1016/S0022-2836(05)80360-2S0022-2836(05) 80360-2[pii].

[276] Gordon Gremme, Volker Brendel, Michael E. Sparks, and Stefan Kurtz. Engineering a software tool for gene structure prediction in higher organisms. *Information and Software Technology*, 47(15):965–978, dec 2005. doi: 10.1016/j.infsof.2005.09.005. URL `http://dx.doi.org/10.1016/j.infsof.2005.09.005`.

[277] L. Li, Jr. Stoeckert, C. J., and D. S. Roos. Orthomcl: identification of ortholog groups for eukaryotic genomes. *Genome Res*, 13(9):2178–89, 2003. ISSN 1088-9051 (Print) 1088-9051 (Linking). doi: 10.1101/gr.122450313/9/2178[pii].

[278] M. Thomas-Chollier, O. Sand, J. V. Turatsinze, R. Janky, M. Defrance, E. Vervisch, S. Brohee, and J. van Helden. RSAT: regulatory sequence analysis tools. *Nucleic Acids Res.*, 36(Web Server issue):W119–127, Jul 2008.

[279] N. O. Steffens, C. Galuschka, M. Schindler, L. Bulow, and R. Hehl. Athamap: an online resource for in silico transcription factor binding sites in the arabidopsis thaliana genome. *Nucleic Acids Res*, 32:D368–D372, 2004.

[280] C. Luo, D. J. Sidote, Y. Zhang, R. A. Kerstetter, T. P. Michael, and E. Lam. Integrative analysis of chromatin states in arabidopsis identified potential regulatory mechanisms for natural antisense transcript production. *The Plant journal : for cell and molecular biology*, 2012. ISSN 1365-313X (Electronic) 0960-7412 (Linking). doi: 10.1111/tpj.12017.

[281] V. Kumar, M. Muratani, N. A. Rayan, P. Kraus, T. Lufkin, H. H. Ng, and S. Prabhakar. Uniform, optimal signal processing of mapped deep-sequencing data. *Nat. Biotechnol.*, 31(7):615–622, Jul 2013.

[282] E. Oh, J. Y. Zhu, and Z. Y. Wang. Interaction between bzr1 and pif4 integrates brassinosteroid and environmental responses. *Nature cell biology*, 14(8):802–9, 2012. ISSN 1476-4679 (Electronic) 1465-7392 (Linking). doi: 10.1038/ncb2545.

[283] P. Hornitschek, M. V. Kohnen, S. Lorrain, J. Rougemont, K. Ljung, I. Lopez-Vidriero, J. M. Franco-Zorrilla, R. Solano, M. Trevisan, S. Pradervand, I. Xenarios, and C. Fankhauser. Phytochrome interacting factors 4 and 5 control seedling growth in changing light conditions by directly controlling auxin signaling. *The Plant journal : for cell and molecular biology*, 2012. ISSN 1365-313X (Electronic) 0960-7412 (Linking). doi: 10.1111/j.1365-313X.2012.05033.x.

[284] W. Deng, H. Ying, C. A. Helliwell, J. M. Taylor, W. J. Peacock, and E. S. Dennis. Flowering locus c (flc) regulates development pathways throughout the life cycle of arabidopsis. *Proceedings of the National Academy of Sciences of the United States of America*, 108(16):6680–5, 2011. ISSN 1091-6490 (Electronic) 0027-8424 (Linking). doi: 10.1073/pnas.1103175108.

[285] X. Ouyang, J. Li, G. Li, B. Li, B. Chen, H. Shen, X. Huang, X. Mo, X. Wan, R. Lin, S. Li, H. Wang, and X. W. Deng. Genome-wide binding site analysis of far-red elongated hypocotyl3 reveals its novel function in arabidopsis development. *The Plant cell*, 23(7):2514–35, 2011. ISSN 1532-298X (Electronic) 1040-4651 (Linking). doi: 10.1105/tpc.111.085126.

[286] N. Nakamichi, T. Kiba, M. Kamioka, T. Suzuki, T. Yamashino, T. Higashiyama, H. Sakakibara, and T. Mizuno. Transcriptional repressor prr5 directly regulates clock-output pathways. *Proceedings of the National Academy of Sciences of the United States of America*, 2012. ISSN 1091-6490 (Electronic) 0027-8424 (Linking). doi: 10.1073/pnas.1205156109.

[287] Y. Zhang, O. Mayba, A. Pfeiffer, H. Shi, J. M. Tepperman, T. P. Speed, and P. H. Quail. A quartet of pif bhlh factors provides a transcriptionally centered signaling hub that regulates seedling morphogenesis through differential expression-patterning of shared target genes in arabidopsis. *PLoS genetics*, 9(1):e1003244, 2013. ISSN 1553-7404 (Electronic) 1553-7390 (Linking). doi: 10.1371/journal.pgen.1003244.

[288] X. Yu, L. Li, J. Zola, M. Aluru, H. Ye, A. Foudree, H. Guo, S. Anderson, S. Aluru, P. Liu, S. Rodermel, and Y. Yin. A brassinosteroid transcriptional network revealed by genome-wide identification of besi target genes in arabidopsis thaliana. *Plant J*, 65(4):634–46, 2011. ISSN 1365-313X (Electronic) 0960-7412 (Linking). doi: 10.1111/j.1365-313X.2010.04449.x.

[289] Z. Tao, L. Shen, C. Liu, L. Liu, Y. Yan, and H. Yu. Genome-wide identification of soc1 and svp targets during the floral transition in arabidopsis. *The Plant journal : for cell and molecular biology*, 70(4):549–61, 2012. ISSN 1365-313X (Electronic) 0960-7412 (Linking). doi: 10.1111/j.1365-313X.2012.04919.x.

[290] Y. Zheng, N. Ren, H. Wang, A. J. Stromberg, and S. E. Perry. Global identification of targets of the arabidopsis mads domain protein agamous-like15. *Plant Cell*, 21(9):2563–77, 2009. ISSN 1532-298X (Electronic) 1040-4651 (Linking). doi: tpc.109.068890[pii]10.1105/tpc.109.068890.

[291] C. M. Winter, R. S. Austin, S. Blanvillain-Baufume, M. A. Reback, M. Monniaux, M. F. Wu, Y. Sang, A. Yamaguchi, N. Yamaguchi, J. E. Parker, F. Parcy, S. T. Jensen, H. Li, and D. Wagner. Leafy target genes reveal floral regulatory logic, cis motifs, and a link to biotic stimulus response. *Developmental cell*, 20(4):430–43, 2011. ISSN 1878-1551 (Electronic) 1534-5807 (Linking). doi: 10.1016/j.devcel.2011.03.019.

[292] X. Wang, Y. Bian, K. Cheng, L. F. Gu, M. Ye, H. Zou, S. S. Sun, and J. X. He. A large-scale protein phosphorylation analysis reveals novel phosphorylation motifs and phosphoregulatory networks in arabidopsis. *Journal of proteomics*, 78:486–98, 2013. ISSN 1876-7737 (Electronic). doi: 10.1016/j.jprot.2012.10.018.

[293] E. Seo, H. Lee, J. Jeon, H. Park, J. Kim, Y. S. Noh, and I. Lee. Crosstalk between cold response and flowering in arabidopsis is mediated through the flowering-time gene soc1 and its upstream negative regulator flc. *Plant Cell*, 21(10):3185–97, 2009. ISSN 1532-298X (Electronic) 1040-4651 (Linking). doi: 10.1105/tpc.108.063883.

[294] Y. Huang, C. Y. Li, Y. Qi, S. Park, and S. I. Gibson. Sis8, a putative mitogen-activated protein kinase kinase kinase, regulates sugar-resistant seedling development in arabidopsis. *Plant J*, 77(4):577–88, 2014. ISSN 1365-313X (Electronic) 0960-7412 (Linking). doi: 10.1111/tpj.12404. URL `http://www.ncbi.nlm.nih.gov/pubmed/24320620`. Huang, Yadong Li, Chun Yao Qi, Yiping Park, Sungjin Gibson, Susan I eng Research Support, Non-U.S. Gov't England 2013/12/11 06:00 Plant J. 2014 Feb;77(4):577-88. doi: 10.1111/tpj.12404. Epub 2014 Jan 23.

[295] T. J. Heisel, C. Y. Li, K. M. Grey, and S. I. Gibson. Mutations in histone acetyltransferase1 affect sugar response and gene expression in arabidopsis. *Front Plant Sci*, 4:245, 2013. ISSN 1664-462X (Electronic) 1664-462X (Linking). doi: 10.3389/fpls.2013.00245. URL `http://www.ncbi.nlm.nih.gov/pubmed/23882272`. Heisel, Timothy J Li, Chun Yao Grey, Katia M Gibson, Susan I eng Switzerland 2013/07/25 06:00 Front Plant Sci. 2013 Jul 17;4:245. doi: 10.3389/fpls.2013.00245. eCollection 2013.

[296] M. Vanstraelen, M. Baloban, O. Da Ines, A. Cultrone, T. Lammens, V. Boudolf, S. C. Brown, L. De Veylder, P. Mergaert, and E. Kondorosi. Apc/c-ccs52a complexes control meristem maintenance in the arabidopsis root. *Proc Natl Acad Sci U S A*, 106(28):11806–11, 2009. ISSN 1091-6490 (Electronic) 0027-8424 (Linking). doi: 10.1073/pnas.0901193106. URL `http:`

//www.ncbi.nlm.nih.gov/pubmed/19553203. Vanstraelen, Marleen Baloban, Mikhail Da Ines, Olivier Cultrone, Antonietta Lammens, Tim Boudolf, Veronique Brown, Spencer C De Veylder, Lieven Mergaert, Peter Kondorosi, Eva eng Research Support, Non-U.S. Gov't 2009/06/26 09:00 Proc Natl Acad Sci U S A. 2009 Jul 14;106(28):11806-11. doi: 10.1073/pnas.0901193106. Epub 2009 Jun 24.

[297] L. Vercruyssen, A. Verkest, N. Gonzalez, K. S. Heyndrickx, D. Eeckhout, S. K. Han, T. Jegu, R. Archacki, J. Van Leene, M. Andriankaja, S. De Bodt, T. Abeel, F. Coppens, S. Dhondt, L. De Milde, M. Vermeersch, K. Maleux, K. Gevaert, A. Jerzmanowski, M. Benhamed, D. Wagner, K. Vandepoele, G. De Jaeger, and D. Inze. Angustifolia3 binds to swi/snf chromatin remodeling complexes to regulate transcription during arabidopsis leaf development. *Plant Cell*, 2014. ISSN 1532-298X (Electronic) 1040-4651 (Linking). doi: 10.1105/tpc.113.115907.

[298] P. M. Donnelly, D. Bonetta, H. Tsukaya, R. E. Dengler, and N. G. Dengler. Cell cycling and cell enlargement in developing leaves of arabidopsis. *Dev Biol*, 215 (2):407–19, 1999. ISSN 0012-1606 (Print) 0012-1606 (Linking). doi: 10.1006/dbio.1999.9443. URL http://www.ncbi.nlm.nih.gov/pubmed/10545247http://ac.els-cdn.com/S0012160699994435/1-s2.0-S0012160699994435-main.pdf?_tid=fb19b4e6-8827-11e4-af34-00000aacb35d&acdnat=1419066782_9f55e4e90a6fabda19414ff1a48d5fe8. Donnelly, P M Bonetta, D Tsukaya, H Dengler, R E Dengler, N G eng 1999/11/05 Dev Biol. 1999 Nov 15;215(2):407-19.

[299] M. Andriankaja, S. Dhondt, S. De Bodt, H. Vanhaeren, F. Coppens, L. De Milde, P. Muhlenbock, A. Skirycz, N. Gonzalez, G. T. Beemster, and D. Inze. Exit from proliferation during leaf development in arabidopsis thaliana: a not-so-gradual process. *Dev Cell*, 22(1):64–78, 2012. ISSN 1878-1551 (Electronic) 1534-5807 (Linking). doi: 10.1016/j.devcel.2011.11.011. URL http://www.ncbi.nlm.nih.gov/pubmed/22227310. Andriankaja, Megan Dhondt, Stijn De Bodt, Stefanie Vanhaeren, Hannes Coppens, Frederik De Milde, Liesbeth Muhlenbock, Per Skirycz, Aleksandra Gonzalez, Nathalie Beemster, Gerrit T S Inze, Dirk eng Research Support, Non-U.S. Gov't 2012/01/10 06:00 Dev Cell. 2012 Jan 17;22(1):64-78. doi: 10.1016/j.devcel.2011.11.011. Epub 2012 Jan 5.

[300] N. Gonzalez, L. Pauwels, A. Baekelandt, L. Cuellar Perez De Milde, A. A. Nagels Durand, R. De Clercq, E. Van De Slijke, R. Vanden Bossche, J. Van Leene, K.S. Heyndrickx, D. Eeckhout, K. Gevaert, K. Vandepoele, G. De Jaeger, A. Goossens, and D. Inze. A repressor protein complex regulating leaf growth in the dicot arabidopsis. *Unpublished*.

[301] D. C. Bergmann and F. D. Sack. Stomatal development. *Annu Rev Plant Biol*, 58:163–81, 2007. ISSN 1543-5008 (Print) 1543-5008 (Linking). doi: 10.1146/annurev.arplant.58.032806.104023. URL http://www.ncbi.nlm.nih.gov/pubmed/17201685http://www.annualreviews.org/doi/abs/10.1146/annurev.arplant.58.032806.104023. Bergmann, Dominique C Sack, Fred D eng Research Support, U.S. Gov't, Non-P.H.S. Review 2007/01/05 09:00 Annu Rev Plant Biol. 2007;58:163-81.

[302] L. J. Pillitteri and K. U. Torii. Mechanisms of stomatal development. *Annu Rev Plant Biol*, 63:591–614, 2012. ISSN 1545-2123 (Electronic) 1543-5008 (Linking). doi: 10.1146/annurev-arplant-042811-105451. URL http://www.ncbi.nlm.nih.gov/pubmed/22404473. Pillitteri, Lynn Jo Torii, Keiko U eng Research Support, Non-U.S. Gov't Research Support, U.S. Gov't, Non-P.H.S. Review 2012/03/13 06:00 Annu Rev Plant Biol. 2012;63:591-614. doi: 10.1146/annurev-arplant-042811-105451. Epub 2012 Jan 30.

[303] D. W. White. Peapod regulates lamina size and curvature in arabidopsis. *Proc Natl Acad Sci U S A*, 103(35):13238–43, 2006. ISSN 0027-8424 (Print) 0027-8424 (Linking). doi: 10.1073/pnas.0604349103. URL http://www.ncbi.nlm.nih.gov/pubmed/16916932. White, Derek W R eng Research Support, Non-U.S. Gov't 2006/08/19 09:00 Proc Natl Acad Sci U S A. 2006 Aug 29;103(35):13238-43. Epub 2006 Aug 17.

[304] W. Dewitte, S. Scofield, A. A. Alcasabas, S. C. Maughan, M. Menges, N. Braun, C. Collins, J. Nieuwland, E. Prinsen, V. Sundaresan, and J. A. Murray. Arabidopsis cycd3 d-type cyclins link cell proliferation and endocycles and are rate-limiting for cytokinin responses. *Proc Natl Acad Sci U S A*, 104(36): 14537–42, 2007. ISSN 0027-8424 (Print) 0027-8424 (Linking). doi: 10.1073/pnas.0704166104. URL http://www.ncbi.nlm.nih.gov/pubmed/17726100. Dewitte, Walter Scofield, Simon Alcasabas, Annette A Maughan, Spencer C Menges, Margit Braun, Nils Collins, Carl Nieuwland, Jeroen Prinsen, Els Sundaresan, Venkatesan Murray, James A H eng Research Support, Non-U.S. Gov't 2007/08/30 09:00 Proc Natl Acad Sci U S A. 2007 Sep 4;104(36):14537-42. Epub 2007 Aug 28.

[305] X. Zhou, K. F. Benson, H. R. Ashar, and K. Chada. Mutation responsible for the mouse pygmy phenotype in the developmentally regulated factor hmgi-c. *Nature*, 376(6543): 771–4, 1995. ISSN 0028-0836 (Print) 0028-0836 (Linking). doi: 10.1038/376771a0. URL http://www.ncbi.nlm.nih.gov/pubmed/7651535. Zhou, X Benson, K F Ashar, H R Chada, K eng Research Support, U.S. Gov't, P.H.S. ENGLAND 1995/08/31 Nature. 1995 Aug 31;376(6543):771-4.

[306] A. Ritter, P. Fernandez-Calvo, K.S. Heyndrickx, A. Nagels Durand, D. Gasperini, R. Vanden Bossche, R. De Clercq, D. Eeckhout, E.E. Farmer, K. Gevaert, G. De Jaeger, K. Vandepoele, L. Pauwels, and A. Goossens. Frs7 and frs12 are transcriptional repressor proteins modulating red light signaling in arabidopsis. *Unpublished*.

[307] Todd P Michael, Ghislain Breton, Samuel P Hazen, Henry Priest, Todd C Mockler, Steve A Kay, and Joanne Chory. A morning-specific phytohormone gene expression program underlying rhythmic plant growth. *PLoS Biol*, 6(9):e225, 2008.

[308] Todd P Michael, Todd C Mockler, Ghislain Breton, Connor McEntee, Amanda Byer, Jonathan D Trout, Samuel P Hazen, Rongkun Shen, Henry D Priest, Christopher M Sullivan, Scott A Givan, Marcelo Yanovsky, Fangxin Hong, Steve A Kay, and Joanne Chory. Network discovery pipeline elucidates conserved time-of-day-specific cis-regulatory modules. *PLoS Genet*, 4(2):e14, 2008.

[309] Mark Spensley, Jae-Yean Kim, Emma Picot, John Reid, Sascha Ott, Chris Helliwell, and Isabelle A Carre. Evolutionarily conserved regulatory motifs in the promoter of the Arabidopsis clock gene LATE ELONGATED HYPOCOTYL. *Plant Cell*, 21(9):2606–23, 2009.

[310] E. de Wit and W. de Laat. A decade of 3c technologies: insights into nuclear organization. *Genes Dev*, 26(1):11–24, 2012. ISSN 1549-5477 (Electronic) 0890-9369 (Linking). doi: 10.1101/gad.179804.111. URL http://www.ncbi.nlm.nih.gov/pubmed/22215806http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3258961/pdf/11.pdf. de Wit, Elzo de Laat, Wouter eng Research Support, Non-U.S. Gov't Review 2012/01/05 06:00 Genes Dev. 2012 Jan 1;26(1):11-24. doi: 10.1101/gad.179804.111.

[311] M. J. Fullwood, M. H. Liu, Y. F. Pan, J. Liu, H. Xu, Y. B. Mohamed, Y. L. Orlov, S. Velkov, A. Ho, P. H. Mei, E. G. Chew, P. Y. Huang, W. J. Welboren, Y. Han, H. S. Ooi, P. N. Ariyaratne, V. B. Vega, Y. Luo, P. Y. Tan, P. Y. Choy, K. D. Wansa, B. Zhao, K. S. Lim, S. C. Leow, J. S. Yow, R. Joseph, H. Li, K. V. Desai, J. S. Thomsen, Y. K. Lee, R. K. Karuturi, T. Herve, G. Bourque, H. G. Stunnenberg, X. Ruan, V. Cacheux-Rataboul, W. K. Sung, E. T. Liu, C. L. Wei, E. Cheung, and Y. Ruan. An oestrogen-receptor-alpha-bound human chromatin interactome. *Nature*, 462(7269):58–64, 2009. ISSN 1476-4687 (Electronic) 0028-0836 (Linking). doi:

10.1038/nature08497. URL http://www.ncbi.nlm.nih.gov/pubmed/19890323http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2774924/pdf/nihms-145406.pdf. Fullwood, Melissa J Liu, Mei Hui Pan, You Fu Liu, Jun Xu, Han Mohamed, Yusoff Bin Orlov, Yuriy L Velkov, Stoyan Ho, Andrea Mei, Poh Huay Chew, Elaine G Y Huang, Phillips Yao Hui Welboren, Willem-Jan Han, Yuyuan Ooi, Hong Sain Ariyaratne, Pramila N Vega, Vinsensius B Luo, Yanquan Tan, Peck Yean Choy, Pei Ye Wansa, K D Senali Abayratna Zhao, Bing Lim, Kar Sian Leow, Shi Chi Yow, Jit Sin Joseph, Roy Li, Haixia Desai, Kartiki V Thomsen, Jane S Lee, Yew Kok Karuturi, R Krishna Murthy Herve, Thoreau Bourque, Guillaume Stunnenberg, Hendrik G Ruan, Xiaoan Cacheux-Rataboul, Valere Sung, Wing-Kin Liu, Edison T Wei, Chia-Lin Cheung, Edwin Ruan, Yijun eng 1U54HG004557-01/HG/NHGRI NIH HHS/ R01 HG004456/HG/NHGRI NIH HHS/ R01 HG004456-01/HG/NHGRI NIH HHS/ R01 HG004456-02/HG/NHGRI NIH HHS/ R01 HG004456-03/HG/NHGRI NIH HHS/ R01HG003521-01/HG/NHGRI NIH HHS/ R01HG004456-01/HG/NHGRI NIH HHS/ U54 HG004557/HG/NHGRI NIH HHS/ U54 HG004557-01/HG/NHGRI NIH HHS/ U54 HG004557-02/HG/NHGRI NIH HHS/ U54 HG004557-03/HG/NHGRI NIH HHS/ U54 HG004557-04/HG/NHGRI NIH HHS/ Research Support, N.I.H., Extramural Research Support, Non-U.S. Gov't England 2009/11/06 06:00 Nature. 2009 Nov 5;462(7269):58-64. doi: 10.1038/nature08497.

[312] N. Heidari, D. H. Phanstiel, C. He, F. Grubert, F. Jahanbani, M. Kasowski, M. Q. Zhang, and M. P. Snyder. Genome-wide map of regulatory interactions in the human genome. *Genome Res*, 24(12):1905–17, 2014. ISSN 1549-5469 (Electronic) 1088-9051 (Linking). doi: 10.1101/gr.176586.114. URL http://www.ncbi.nlm.nih.gov/pubmed/25228660http://genome.cshlp.org/content/24/12/1905.full.pdf. Heidari, Nastaran Phanstiel, Douglas H He, Chao Grubert, Fabian Jahanbani, Fereshteh Kasowski, Maya Zhang, Michael Q Snyder, Michael P eng 2014/09/18 06:00 Genome Res. 2014 Dec;24(12):1905-17. doi: 10.1101/gr.176586.114. Epub 2014 Sep 16.

[313] D. Chen, L. Y. Fu, Z. Zhang, G. Li, H. Zhang, L. Jiang, A. P. Harrison, H. P. Shanahan, C. Klukas, H. Y. Zhang, Y. Ruan, L. L. Chen, and M. Chen. Dissecting the chromatin interactome of microrna genes. *Nucleic Acids Res*, 42(5):3028–43, 2014. ISSN 1362-4962 (Electronic) 0305-1048 (Linking). doi: 10.1093/nar/gkt1294. URL http://www.ncbi.nlm.nih.gov/pubmed/24357409. Chen, Dijun Fu, Liang-Yu Zhang, Zhao Li, Guoliang Zhang, Hang Jiang, Li Harrison, Andrew P Shanahan, Hugh P Klukas, Christian Zhang, Hong-Yu Ruan, Yijun Chen, Ling-Ling Chen, Ming eng Research Support, Non-U.S. Gov't England 2013/12/21 06:00 Nucleic Acids Res. 2014 Mar;42(5):3028-43. doi: 10.1093/nar/gkt1294. Epub 2013 Dec 18.

[314] V. Schubert, R. Rudnik, and I. Schubert. Chromatin associations in arabidopsis interphase nuclei. *Front Genet*, 5:389, 2014. ISSN 1664-8021 (Electronic) 1664-8021 (Linking). doi: 10.3389/fgene.2014.00389. URL http://www.ncbi.nlm.nih.gov/pubmed/25431580http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4230181/pdf/fgene-05-00389.pdf. Schubert, Veit Rudnik, Radoslaw Schubert, Ingo eng Switzerland 2014/11/29 06:00 Front Genet. 2014 Nov 13;5:389. doi: 10.3389/fgene.2014.00389. eCollection 2014.

[315] X. Zhang, J. Xia, Y. E. Lii, B. E. Barrera-Figueroa, X. Zhou, S. Gao, L. Lu, D. Niu, Z. Chen, C. Leung, T. Wong, H. Zhang, J. Guo, Y. Li, R. Liu, W. Liang, J. K. Zhu, W. Zhang, and H. Jin. Genome-wide analysis of plant nat-sirnas reveals insights into their distribution, biogenesis and function. *Genome Biol*, 13(3):R20, 2012. ISSN 1465-6914 (Electronic) 1465-6906 (Linking). doi: 10.1186/gb-2012-13-3-r20. URL http://www.ncbi.nlm.nih.gov/pubmed/22439910. Zhang, Xiaoming Xia, Jing Lii, Yifan E Barrera-Figueroa, Blanca E Zhou, Xuefeng Gao, Shang Lu, Lu Niu, Dongdong Chen, Zheng Leung, Christy Wong, Timothy Zhang, Huiming Guo, Jianhua Li, Yi Liu, Renyi Liang, Wanqi Zhu, Jian-Kang Zhang, Weixiong Jin, Hailing eng R01 GM070795/GM/NIGMS NIH HHS/ R01 GM093008/GM/NIGMS NIH HHS/ R01GM059138/GM/NIGMS NIH HHS/ R01GM070795/GM/NIGMS NIH HHS/ R01GM086412/GM/NIGMS NIH HHS/ R01GM093008/GM/NIGMS NIH HHS/ RC1AR058681/AR/NIAMS NIH HHS/ Research Support, N.I.H., Extramural Research Support, U.S. Gov't, Non-P.H.S. England 2012/03/24 06:00 Genome Biol. 2012;13(3):R20. doi: 10.1186/gb-2012-13-3-r20.

[316] A. T. Wierzbicki. The role of long non-coding rna in transcriptional gene silencing. *Curr Opin Plant Biol*, 15(5):517–22, 2012. ISSN 1879-0356 (Electronic) 1369-5266 (Linking). doi: 10.1016/j.pbi.2012.08.008. URL http://www.ncbi.nlm.nih.gov/pubmed/22960034. Wierzbicki, Andrzej T eng Research Support, U.S. Gov't, Non-P.H.S. Review England 2012/09/11 06:00 Curr Opin Plant Biol. 2012 Nov;15(5):517-22. doi: 10.1016/j.pbi.2012.08.008. Epub 2012 Sep 6.

[317] J. Liu, C. Jung, J. Xu, H. Wang, S. Deng, L. Bernad, C. Arenas-Huertero, and N. H. Chua. Genome-wide analysis uncovers regulation of long intergenic noncoding rnas in arabidopsis. *The Plant cell*, 2012. ISSN 1532-298X (Electronic) 1040-4651 (Linking). doi: 10.1105/tpc.112.102855. URL http://www.ncbi.nlm.nih.gov/pubmed/23136377http://www.plantcell.org/content/early/2012/11/05/tpc.112.102855.full.pdf. Liu, Jun Jung, Choonkyun Xu, Jun Wang, Huan Deng, Shulin Bernad, Lucia Arenas-Huertero, Catalina Chua, Nam-Hai Plant Cell. 2012 Nov 6.

[318] J. Jin, J. Liu, H. Wang, L. Wong, and N. H. Chua. Plncdb: plant long non-coding rna database. *Bioinformatics*, 29(8):1068–71, 2013. ISSN 1367-4811 (Electronic) 1367-4803 (Linking). doi: 10.1093/bioinformatics/btt107. URL http://www.ncbi.nlm.nih.gov/pubmed/23476021. Jin, Jingjing Liu, Jun Wang, Huan Wong, Limsoon Chua, Nam-Hai eng Research Support, Non-U.S. Gov't England Oxford, England 2013/03/12 06:00 Bioinformatics. 2013 Apr 15;29(8):1068-71. doi: 10.1093/bioinformatics/btt107. Epub 2013 Mar 7.

[319] C. P. Ponting, P. L. Oliver, and W. Reik. Evolution and functions of long noncoding rnas. *Cell*, 136(4):629–41, 2009. ISSN 1097-4172 (Electronic) 0092-8674 (Linking). doi: 10.1016/j.cell.2009.02.006. URL http://www.ncbi.nlm.nih.gov/pubmed/19239885.

[320] S. Swiezewski, F. Liu, A. Magusin, and C. Dean. Cold-induced silencing by long antisense transcripts of an arabidopsis polycomb target. *Nature*, 462(7274):799–802, 2009. ISSN 1476-4687 (Electronic) 0028-0836 (Linking). doi: 10.1038/nature08618. URL http://www.ncbi.nlm.nih.gov/pubmed/20010688. Swiezewski, Szymon Liu, Fuquan Magusin, Andreas Dean, Caroline eng BB/D010799/1/Biotechnology and Biological Sciences Research Council/United Kingdom Research Support, Non-U.S. Gov't England 2009/12/17 06:00 Nature. 2009 Dec 10;462(7274):799-802. doi: 10.1038/nature08618.

[321] J. B. Heo and S. Sung. Vernalization-mediated epigenetic silencing by a long intronic noncoding rna. *Science*, 331(6013):76–9, 2011. ISSN 1095-9203 (Electronic) 0036-8075 (Linking). doi: 10.1126/science.1197349. URL http://www.ncbi.nlm.nih.gov/pubmed/21127216http://www.sciencemag.org/content/331/6013/76.full.pdf. Heo, Jae Bok Sung, Sibum eng Research Support, Non-U.S. Gov't Research Support, U.S. Gov't, Non-P.H.S. New York, N.Y. 2010/12/04 06:00 Science. 2011 Jan 7;331(6013):76-9. doi: 10.1126/science.1197349. Epub 2010 Dec 2.

[322] X. Ma, C. Shao, Y. Jin, H. Wang, and Y. Meng. Long noncoding rnas: a novel endogenous source for the generation of dicer-like 1-dependent small rnas in arabidopsis thaliana. *RNA Biol*, 11(4):373–90, 2014. ISSN 1555-8584 (Electronic) 1547-6286 (Linking). doi: 10.4161/rna.28725. URL http://www.ncbi.nlm.nih.gov/pubmed/24717238. Ma, Xiaoxia Shao, Chaogang Jin, Yongfeng Wang, Huizhong Meng, Yijun eng Research Support, Non-U.S. Gov't 2014/04/11 06:00 RNA Biol. 2014 Apr;11(4):373-90. doi: 10.4161/rna.28725. Epub 2014 Apr 4.

[323] H. J. Wu, Z. M. Wang, M. Wang, and X. J. Wang. Widespread long noncoding rnas as endogenous target mimics for micrornas in plants. *Plant Physiol*, 161(4):1875–84, 2013. ISSN 1532-2548 (Electronic) 0032-0889 (Linking). doi: 10.1104/pp.113.215962. URL http://www.ncbi.nlm.nih.gov/pubmed/23429259. Wu, Hua-Jun Wang, Zhi-Min Wang, Meng Wang, Xiu-Jie eng Research Support, Non-U.S. Gov't 2013/02/23 06:00 Plant Physiol. 2013 Apr;161(4):1875-84. doi: 10.1104/pp.113.215962. Epub 2013 Feb 21.

[324] F. Bardou, F. Ariel, C. G. Simpson, N. Romero-Barrios, P. Laporte, S. Balzergue, J. W. Brown, and M. Crespi. Long noncoding rna modulates alternative splicing regulators in arabidopsis. *Dev Cell*, 30(2):166–76, 2014. ISSN 1878-1551 (Electronic) 1534-5807 (Linking). doi: 10.1016/j.devcel.2014.06.017. URL http://www.ncbi.nlm.nih.gov/pubmed/25073154http://ac.els-cdn.com/S1534580714004067/1-s2.0-S1534580714004067-main.pdf?_tid=3773d37e-933e-11e4-9998-00000aacb361&acdnat=1420285795_8437db1ca1dff6e7fad402f1fe303678. Bardou, Florian Ariel, Federico Simpson, Craig G Romero-Barrios, Natali Laporte, Philippe Balzergue, Sandrine Brown, John W S Crespi, Martin eng BB/G024979/1/Biotechnology and Biological Sciences Research Council/United Kingdom Biotechnology and Biological Sciences Research Council/United Kingdom Research Support, Non-U.S. Gov't 2014/07/30 06:00 Dev Cell. 2014 Jul 28;30(2):166-76. doi: 10.1016/j.devcel.2014.06.017.

[325] U. Johanson, J. West, C. Lister, S. Michaels, R. Amasino, and C. Dean. Molecular analysis of frigida, a major determinant of natural variation in arabidopsis flowering time. *Science*, 290(5490):344–7, 2000. ISSN 0036-8075 (Print) 0036-8075 (Linking). URL http://www.ncbi.nlm.nih.gov/pubmed/11030654. Johanson, U West, J Lister, C Michaels, S Amasino, R Dean, C eng Research Support, Non-U.S. Gov't New York, N.Y. 2000/10/13 11:00 Science. 2000 Oct 13;290(5490):344-7.

[326] P. Li, D. Filiault, M. S. Box, E. Kerdaffrec, C. van Oosterhout, A. M. Wilczek, J. Schmitt, M. McMullan, J. Bergelson, M. Nordborg, and C. Dean. Multiple flc haplotypes defined by independent cis-regulatory variation underpin life history diversity in arabidopsis thaliana. *Genes Dev*, 28(15): 1635–40, 2014. ISSN 1549-5477 (Electronic) 0890-9369 (Linking). doi: 10.1101/gad.245993.114. URL http://www.ncbi.nlm.nih.gov/pubmed/25035417. Li, Peijin Filiault, Daniele Box, Mathew S Kerdaffrec, Envel van Oosterhout, Cock Wilczek, Amity M Schmitt, Johanna McMullan, Mark Bergelson, Joy Nordborg, Magnus Dean, Caroline eng BB/I007857/1/Biotechnology and Biological Sciences Research Council/United Kingdom BB/J004588/1/Biotechnology and Biological Sciences Research Council/United Kingdom Research Support, Non-U.S. Gov't Research Support, U.S. Gov't, Non-P.H.S. 2014/07/19 06:00 Genes Dev. 2014 Aug 1;28(15):1635-40. doi: 10.1101/gad.245993.114. Epub 2014 Jul 17.

[327] U. Rosas, Y. Mei, Q. Xie, J. A. Banta, R. W. Zhou, G. Seufferheld, S. Gerard, L. Chou, N. Bhambhra, J. D. Parks, J. M. Flowers, C. R. McClung, Y. Hanzawa, and M. D. Purugganan. Variation in arabidopsis flowering time associated with cis-regulatory variation in constans. *Nat Commun*, 5:3651, 2014. ISSN 2041-1723 (Electronic) 2041-1723 (Linking). doi: 10.1038/ncomms4651. URL http://www.ncbi.nlm.nih.gov/pubmed/24736505. Rosas, Ulises Mei, Yu Xie, Qiguang Banta, Joshua A Zhou, Royce W Seufferheld, Gabriela Gerard, Silvia Chou, Lucy Bhambhra, Naeha Parks, Jennifer Deane Flowers, Jonathan M McClung, C Robertson Hanzawa, Yoshie Purugganan, Michael D eng Research Support, Non-U.S. Gov't Research Support, U.S. Gov't, Non-P.H.S. England 2014/04/17 06:00 Nat Commun. 2014 Apr 16;5:3651. doi: 10.1038/ncomms4651.

[328] L. Suter, M. Ruegg, N. Zemp, L. Hennig, and A. Widmer. Gene regulatory variation mediates flowering responses to vernalization along an altitudinal gradient in arabidopsis. *Plant Physiol*, 166(4):1928–42, 2014. ISSN 1532-2548 (Electronic) 0032-0889 (Linking). doi: 10.1104/pp.114.247346. URL http://www.ncbi.nlm.nih.gov/pubmed/25339407. Suter, Leonie Ruegg, Marlene Zemp, Niklaus Hennig, Lars Widmer, Alex eng 2014/10/24 06:00 Plant Physiol. 2014 Dec;166(4):1928-42. doi: 10.1104/pp.114.247346. Epub 2014 Oct 22.

[329] 000 rice genomes project The 3. The 3,000 rice genomes project. *Gigascience*, 3:7, 2014. ISSN 2047-217X (Electronic) 2047-217X (Linking). doi: 10.1186/2047-217X-3-7. URL http://www.ncbi.nlm.nih.gov/pubmed/24872877. eng England 2014/05/30 06:00 Gigascience. 2014 May 28;3:7. doi: 10.1186/2047-217X-3-7. eCollection 2014.

[330] D. T. Odom, R. D. Dowell, E. S. Jacobsen, W. Gordon, T. W. Danford, K. D. MacIsaac, P. A. Rolfe, C. M. Conboy, D. K. Gifford, and E. Fraenkel. Tissue-specific transcriptional regulation has diverged significantly between human and mouse. *Nat Genet*, 39(6):730–2, 2007. ISSN 1061-4036 (Print) 1061-4036 (Linking). doi: 10.1038/ng2047. URL http://www.ncbi.nlm.nih.gov/pubmed/17529977. Odom, Duncan T Dowell, Robin D Jacobsen, Elizabeth S Gordon, William Danford, Timothy W MacIsaac, Kenzie D Rolfe, P Alexander Conboy, Caitlin M Gifford, David K Fraenkel, Ernest eng 15603/Cancer Research UK/United Kingdom A15603/Cancer Research UK/United Kingdom DK076284/DK/NIDDK NIH HHS/ DK68655/DK/NIDDK NIH HHS/ DK70813/DK/NIDDK NIH HHS/ DK92310/DK/NIDDK NIH HHS/ Comparative Study Research Support, N.I.H., Extramural Research Support, Non-U.S. Gov't 2007/05/29 09:00 Nat Genet. 2007 Jun;39(6):730-2. Epub 2007 May 21.

[331] S. F. Boj, J. M. Servitja, D. Martin, M. Rios, I. Talianidis, R. Guigo, and J. Ferrer. Functional targets of the monogenic diabetes transcription factors hnf-1alpha and hnf-4alpha are highly conserved between mice and humans. *Diabetes*, 58 (5):1245–53, 2009. ISSN 1939-327X (Electronic) 0012-1797 (Linking). doi: 10.2337/db08-0812. URL http://www.ncbi.nlm.nih.gov/pubmed/19188435http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2671044/pdf/zdb1245.pdf. Boj, Sylvia F Servitja, Joan Marc Martin, David Rios, Martin Talianidis, Iannis Guigo, Roderic Ferrer, Jorge eng Research Support, Non-U.S. Gov't 2009/02/04 09:00 Diabetes. 2009 May;58(5):1245-53. doi: 10.2337/db08-0812. Epub 2009 Feb 2.

[332] Y. Cheng, Z. Ma, B. H. Kim, W. Wu, P. Cayting, A. P. Boyle, V. Sundaram, X. Xing, N. Dogan, J. Li, G. Euskirchen, S. Lin, Y. Lin, A. Visel, T. Kawli, X. Yang, D. Patacsil, C. A. Keller, B. Giardine, Encode Consortium Mouse, A. Kundaje, T. Wang, L. A. Pennacchio, Z. Weng, R. C. Hardison, and M. P. Snyder. Principles of regulatory information conservation between mouse and human. *Nature*, 515(7527):371–5, 2014. ISSN 1476-4687 (Electronic) 0028-0836 (Linking). doi: 10.1038/nature13985. URL http://www.ncbi.nlm.nih.gov/pubmed/25409826. Cheng, Yong Ma, Zhihai Kim, Bong-Hyun Wu, Weisheng Cayting, Philip Boyle, Alan P Sundaram, Vasavi Xing, Xiaoyun Dogan, Nergiz Li, Jingjing Euskirchen, Ghia Lin, Shin Lin, Yiing Visel, Axel Kawli, Trupti Yang, Xinqiong Patacsil, Dorrelyn Keller, Cheryl A Giardine, Belinda Kundaje, Anshul Wang, Ting Pennacchio, Len A Weng, Zhiping Hardison, Ross C Snyder, Michael P eng 1U54HG00699/HG/NHGRI

NIH HHS/ 3RC2HG005602/HG/NHGRI NIH HHS/ 5U54HG006996/HG/NHGRI NIH HHS/ R01 ES024992/ES/NIEHS NIH HHS/ R01 HG007175/HG/NHGRI NIH HHS/ R01 HG007354/HG/NHGRI NIH HHS/ R01DK065806/DK/NIDDK NIH HHS/ R01HG003988/HG/NHGRI NIH HHS/ RC2HG005573/HG/NHGRI NIH HHS/ U54HG006997/HG/NHGRI NIH HHS/ Research Support, American Recovery and Reinvestment Act Research Support, N.I.H., Extramural England 2014/11/21 06:00 Nature. 2014 Nov 20;515(7527):371-5. doi: 10.1038/nature13985.

[333] S. Islam, U. Kjallquist, A. Moliner, P. Zajac, J. B. Fan, P. Lonnerberg, and S. Linnarsson. Highly multiplexed and strand-specific single-cell rna 5' end sequencing. *Nat Protoc*, 7(5): 813–28, 2012. ISSN 1750-2799 (Electronic) 1750-2799 (Linking). doi: 10.1038/nprot.2012.022. URL http://www.ncbi.nlm.nih.gov/pubmed/22481528. Islam, Saiful Kjallquist, Una Moliner, Annalena Zajac, Pawel Fan, Jian-Bing Lonnerberg, Peter Linnarsson, Sten eng 261063/European Research Council/International Research Support, Non-U.S. Gov't England 2012/04/07 06:00 Nat Protoc. 2012 Apr 5;7(5):813-28. doi: 10.1038/nprot.2012.022.

[334] A. K. Lane, C. E. Niederhuth, L. Ji, and R. J. Schmitz. pencode: A plant encyclopedia of dna elements. *Annu Rev Genet*, 48:49–70, 2014. ISSN 1545-2948 (Electronic) 0066-4197 (Linking). doi: 10.1146/annurev-genet-120213-092443. URL http://www.ncbi.nlm.nih.gov/pubmed/25149370http://www.annualreviews.org/doi/abs/10.1146/annurev-genet-120213-092443http://www.annualreviews.org/doi/pdf/10.1146/annurev-genet-120213-092443. Lane, Amanda K Niederhuth, Chad E Ji, Lexiang Schmitz, Robert J eng 2014/08/26 06:00 Annu Rev Genet. 2014 Nov 23;48:49-70. doi: 10.1146/annurev-genet-120213-092443. Epub 2014 Aug 15.

[335] T. T. Dinh, T. Girke, X. Liu, L. Yant, M. Schmid, and X. Chen. The floral homeotic protein apetala2 recognizes and acts through an at-rich sequence element. *Development*, 139 (11):1978–86, 2012. ISSN 1477-9129 (Electronic) 0950-1991 (Linking). doi: 10.1242/dev.077073.