

FACULTY OF SCIENCES

Studying the Regulatory Landscape of Flowering Plants

Jan Van de Velde

Promoter: Prof. Dr. Klaas Vandepoele
Co-Promoter: Prof. Dr. Jan Fostier

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: 2016-2017

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. Jan Fostier (co-promoter)

Faculty of Engineering and Architecture, Department of Information Technology (INTEC), Ghent University - iMinds

Prof. Dr. Kerstin Kaufmann

Institute for Biochemistry and Biology, Potsdam University

Prof. Dr. Pieter de Bleser

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

Dr. Vanessa Vermeirssen

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

Dr. Stefanie De Bodt

Crop Science Division, Bayer CropScience SA-NV, Functional Biology

Dr. Inge De Clercq

Department of Animal, Plant and Soil Science, ARC Centre of Excellence in Plant Energy Biology, La Trobe University and Faculty of Sciences, Department of Plant Biotechnology and Bioinformatics, Ghent University

Thank You!

Throughout this PhD I have received a lot of support, therefore there are a number of people I would like to thank.

First of all, I would like to thank Klaas Vandepoele, for his support and guidance. Being a great motivator and a knowledgeable scientist, none of this thesis would have been possible without his help.

I would also like to thank everyone in the CIG and BEG group for four fun years, especially Ken and Bram for always being available for help, tips and jokes.

I would like to thank friends and family for all of the aid, and especially my girlfriend Justyna for being the most supportive person I could imagine.

Everybody knows you can't achieve anything in life if not given the opportunity, so I would like to thank my parents for supporting me in all endeavours undertaken and choices I have made both professionally and personally.

Samenvatting

Transcriptionele regulatie is een dynamisch proces dat een belangrijke rol speelt bij het genereren van genexpressieprofielen tijdens de ontwikkeling van een plant of als reactie op (a)biotische stimuli. De doelstellingen van dit project bestonden uit twee delen: de eerste omvat de studie van transcriptionele regulatie en de manier waarop genexpressie wordt georganiseerd in het genoom. De tweede bestaat uit het toepassen van de verkregen datasets teneinde een functie toe te wijzen aan Arabidopsis transcriptiefactoren en hun doelwit genen die voorheen een onbekende functie hadden.

Het onderzoek beschreven in dit proefschrift begint met de ontwikkeling van een phylogenetic footprinting aanpak voor de identificatie van geconserveerde niet-coderende sequenties (CNSen) in *Arabidopsis thaliana*, die gebruik maakt van de genoominformatie van 12 tweezaadlobbige planten. In deze benadering werden zowel alignerings als niet alignerings gebaseerde technieken toegepast om functionele motieven te identificeren in een set van meerdere organismen. De werkwijze houdt rekening met onvolledige motief conservatie en een hoge sequentie divergentie tussen verwante soorten. In totaal hebben we 69,361 footprints geïdentificeerd gelinkt aan 17,895 genen. Een gen regulatorisch netwerk werd samengesteld door de integratie van gekende transcriptiefactor bindingsplaatsen, verkregen uit de literatuur en experimentele studies. Dit netwerk bestond uit 40,758 interacties, waarvan twee derde in DNase I hypersensitieve plaatsen. Dit netwerk is sterk verrijkt naar in-vivo doelwit genen van gekende regulerend transcriptiefactoren en de algemene kwaliteit ervan werd bevestigd met behulp van vijf verschillende biologische validatie metrices. Tenslotte werd een proof of concept experiment uitgevoerd met gedetailleerde expressie en functie-informatie om aan te tonen hoe statische CNSen kunnen worden omgezet in toestandsafhankelijke gen regulerende netwerken. Dit biedt nieuwe mogelijkheden voor regulerende gen annotatie.

In een daaropvolgende analyse pasten we de bovengenoemde phylogenetic footprinting aanpak toe voor de identificatie van CNSen in tien tweezaadlobbige planten. Dit leverde 1,032,291 CNSen geassocieerd met 243,187 genen op. Om deze CNSen te annoteren met transcriptie factor bindingsplaatsen hebben we gebruik gemaakt van de bindingsplaats informatie van 642 TF's die afkomstig zijn uit 35 TF families in Arabidopsis. Validatie van de verkregen CNSen werd uitgevoerd met TF chromatine immunoprecipitatie gevolgd door sequenering (ChIP-Seq) in drie organismen, dit resulteerde in een significante overlap van de meeste datasets. Ook ultra-geconserveerde CNSen werden geïdentificeerd door het insluiten van genomen van aanvullende plantenfamilies. In totaal werden er 715 bindingsplaatsen voor 501 genen en geïdentificeerd die in tweezaadlobbigen, eenzaadlobbigen, mossen en groene algen geconserveerd waren. Door toepassing van de verkregen CNSen vonden wij dat genen die deel uitmaken van een geconserveerd mini-regulon een grotere samenhang van hun expressieprofiel vertonen dan andere gen paren die dit niet vertonen.

Vervolgens werd een nieuw algoritme ontwikkeld dat zowel alignerings als niet alignerings gebaseerde technieken voor het aflijnen van geconserveerde motieven in de promotersequenties van nauw verwante soorten ondersteunt. Kandidaat motieven zijn exhaustief genumereerd als woorden in het IUPAC alfabet en gescreend voor conservatie door het gebruik van de branch length score. Vanwege het exhaustieve karakter van het algoritme en de grote noodzaak van computationele middelen werd het MapReduce

programmeermodel aangenomen om gebruik te kunnen maken van een cloud computing-infrastructuur. De methode werd toegepast op vier eenzaadlobbige plantensoorten en we waren in staat om aan te tonen dat de hoge scorende motieven aanzienlijk verrijken voor de open chromatine regio's in *Oryza sativa* en transcriptiefactor bindingsplaatsen afgeleid door middel van protein binding microarrays in *Oryza sativa* en *Zea mays*. Verder werd aangetoond dat de werkwijze experimenteel (ChIP-Seq) bepaalde ga2ox1-achtige KN1 bindingsplaatsen kan identificeren in *Zea mays*.

Tenslotte, werd een analyse, die voor 12 NAM-ATAF1/2-CUC2 (NAC) transcriptiefactoren doelwit genen identificeert uitgevoerd. NAC transcriptiefactoren behoren tot de grootste transcriptiefactor families in planten, er is echter beperkte data beschikbaar die het DNA-bindingsdomein individuele leden beschrijven. We gebruikten een transcriptiefactor doelwit gen identificatie workflow gebaseerd op de integratie van nieuwe protein binding microarray data met genexpressie en geconserveerde promoter sequenties om de DNA-bindende voorkeuren te identificeren en de onderliggende gen regulerende netwerken te onthullen. De data biedt hoge resolutie vingerafdrukken voor de meeste bestudeerde transcriptiefactoren en geeft aan dat de NAC DNA bindende voorkeuren zouden voorspeld kunnen worden uit hun DNA bindend sequentie domein. De ontwikkelde methodologie, met de toepassing van complementaire functionele genomische filters, maakt het mogelijk om voor elke transcriptiefactor protein binding microarray data om te zetten in een reeks van doelwit genen met hoge kwaliteit. De NAC doelwit genen gedetecteerd door deze benadering konden bevestigd worden door onafhankelijke in vivo analyses.

Summary

Transcriptional regulation is a dynamic process that plays an important role in establishing gene expression profiles during development or in response to (a)biotic stimuli. The aims of this project consisted of two parts: one is the study of how transcriptional regulation and gene expression is organized across the genome. The second consists of applying the obtained datasets to assign function to TFs and their target genes with previously unknown function.

The research presented in this thesis starts with the development of a phylogenetic footprinting approach for the identification of conserved non-coding sequences (CNSs) in *Arabidopsis thaliana* using genomic information of 12 dicot plants. In this approach both alignment and non-alignment-based techniques were applied to identify functional motifs in a multi-species context. The method accounts for incomplete motif conservation as well as high sequence divergence between related species. In total, we identified 69,361 footprints associated with 17,895 genes. A gene regulatory network was compiled, through the integration of known TFBS obtained from literature and experimental studies, 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, a proof of concept experiment using detailed expression and function information was performed to demonstrate how static CNSs can be converted into condition-dependent regulatory networks, offering new opportunities for regulatory gene annotation.

In a subsequent analysis, we applied the aforementioned phylogenetic footprinting framework to ten dicot plants for the identification of CNSs. This yielded 1,032,291 CNSs associated with 243,187 genes. To annotate these CNSs with TFBSs, we made use of binding site information of 642 TFs originating from 35 TF families in Arabidopsis. Validation of the obtained CNSs was performed using TF chromatin immunoprecipitation sequencing (ChIP-Seq) data from three species, resulting in significant overlap for the majority of datasets. We also identified ultra-conserved CNSs by including genomes of additional plant families and identified 715 binding sites for 501 genes conserved in dicots, monocots, mosses and green algae. Through application of the obtained CNSs we found that genes part of conserved mini-regulons have a higher coherence in their expression profile than other divergent gene pairs.

Next, a novel algorithm was developed that supports both alignment-free and alignment-based conserved motif discovery in the promoter sequences of closely related species. Putative motifs are exhaustively enumerated as words over the IUPAC alphabet and screened for conservation using the branch length score. Because of the exhaustive character of the algorithm and great resource needs, the MapReduce programming model was adopted to take advantage of a cloud computing infrastructure and handle these requirements efficiently. The method was applied to four monocotyledon plant species and we were able to show that high-scoring motifs are significantly enriched for open chromatin regions in *Oryza sativa* and for transcription factor binding sites inferred through protein-binding microarrays in *Oryza sativa* and *Zea mays*. Furthermore, the method was shown to recover ga2ox1-like KN1 binding sites in *Zea mays* experimentally profiled through ChIP-Seq.

Finally, a target gene identification analysis for 12 NAM-ATAF1/2-CUC2 (NAC) transcription factors

was performed. NAC transcription factors are among the largest transcription factor families in plants, yet limited data exists from unbiased approaches to resolve the DNA-binding preferences of individual members. We used a TF-target gene identification workflow based on the integration of novel protein binding microarray data with gene expression and multi-species promoter sequence conservation to identify the DNA-binding specificities and the underlying gene regulatory networks. The data offers specific single base resolution fingerprints for most TFs studied and indicates that NAC DNA binding specificities might be predicted from their DNA binding domain's sequence. The developed methodology, including the application of complementary functional genomics filters, makes it possible to translate, for each TF, protein binding microarray data into a set of high-quality target genes. NAC target genes reported from independent in vivo analyses were confirmed to be detected by this approach.

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 four research chapters, and a general conclusion of the results together with my perspectives on the future of comparative and 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

CDS	Coding sequence
CMM	Conserved motif mapping
CRE	*Cis*-regulatory element
DBD	DNA binding domain
DE	Differential expression
EMSA	Electrophoretic Mobility Shift Assay
ENCODE	ENCyclopedia Of DNA Elements
eQTL	Expression quantitative trait locus
FDR	False discovery rate
GO	Gene Ontology
GRN	Gene regulatory network
miRNA	miRNA
mRNA	Messenger RNA
NAC	NAM/ATAF/CUC
PBM	Protein Binding Microarray
PCC	Pearson correlation coefficient
PMRD	Plant miRNA database
Pol I, II and III	DNA polymerase I, II and III
RNA	Ribonucleic acid
rRNA	Ribosomal RNA
Scmm	Comparative motif mapping score
SMSP	Multi species phylogeny footprinting score
snRNA	Small nuclear RNA
TF	Transcription factor
TFBS	Transcription factor binding site
TPR	True positive rate
tRNA	Transfer RNA
TSS	Transcription start site
UTR	Untranslated region
WGD	Whole genome duplication

CHAPTER 1

Introduction

Development and response mechanisms in eukaryotic organisms are flexible processes that allow them to succeed when exposed to variable environmental stimuli. The phenotype of an organism is the result of a complex combination of different levels of regulation that can be split up in two categories, that of gene expression (transcription and translation) and protein function (post-translational modifications, protein-protein interactions, cofactors etc.). There are a large number of intermediate steps between activation of transcription and the functional protein (Fig. 1.1). Therefore a change in gene expression level does not necessarily indicate a change in protein level/activity and the other way around. Given the scope of the thesis, the following sections will focus on gene expression and the regulation of transcription in particular.

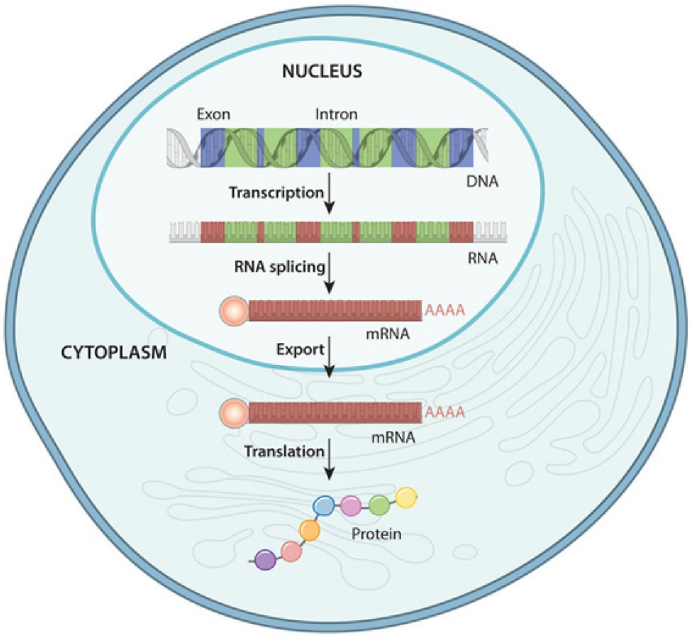

Figure 1.1: **From gene to protein: levels of regulation.** *Eukaryotic gene expression can be regulated at multiple levels. Genomic regulation: chromatin decondensation or condensation, or DNA methylation; transcriptional regulation; RNA processing, RNA turnover in the nucleus and translocation out of the nucleus; translational control; post-translational control (including mRNA turnover in the cytosol, and the folding, assembly, modification, and import of proteins into organelles).* Source: Essentials of Cell Biology

1.1 Gene expression

Gene expression consists out of transcription, ribonucleic acid (RNA) processing and possible translation. Transcription is the biological process in which a complementary RNA copy of a deoxyribonucleic acid (DNA) sequence is synthesized. RNA processing is a maturation step where modifications such as

polyadenylation and 5' capping occur. In translation this mature messenger ribonucleic acid (mRNA) is translated into a protein. All of these steps are very tightly regulated.

The binding of activating transcription factors (TFs) to cis-regulatory elements (CREs) also called target transcription factor binding sites (TFBSs) leads to the recruitment of co-activators and these steps result in the recruitment of a DNA-dependent RNA polymerase through which the DNA is transcribed to RNA (Fig. 1.2). 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). (Sugiura, 1997) (Lee et al., 2004). 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 most studied targets of RNA polymerase III (Pol III) are the different tRNAs. The next step is the assembly of the basal transcription apparatus on the promoter. The promoter of a gene is the sequence located upstream of the transcription start site (TSS). Transcription initiation ends with the incorporation of the first few nucleotides of the mRNA.

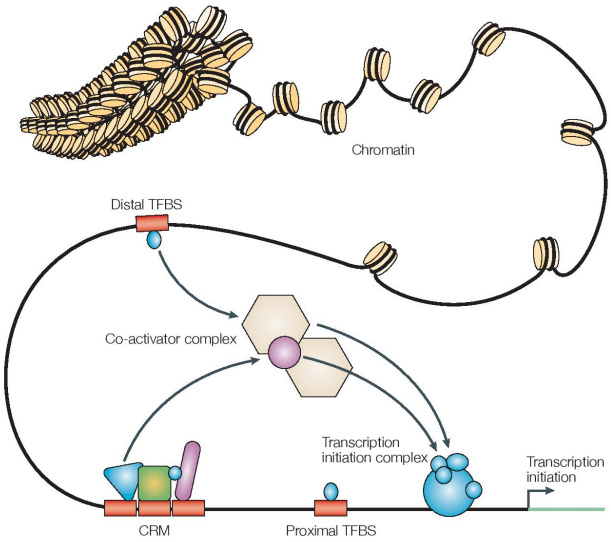

Figure 1.2: **Elements of transcriptional regulation.***Important components of transcription: nucleosome remodelling, TF binding to a TFBS, the recruitment of co-activators and the organisation of regulatory elements in a cis-regulatory module (CRM).* Source: (Wasserman and Sandelin, 2004)

In the next step the initiation complex is activated through phosphorylation and the first 11-15 bases of DNA at the TSS are unwound to introduce a single stranded DNA strand into the active site of Pol II in order to produce a transcribed string. This step is followed by dissociation of Pol II from some of the general TFs also called promoter escape. After promoter escape, Pol II halts at the promoter's proximal pause site, which is located 20-40 nucleotides into the gene. After this pause the last step in transcription starts, which is called promoter elongation in which the transcript is elongated to its full length, this occurs subsequently to a second activation through phosphorylation (Weake and Workman, 2010). The (near complete) assembly of the initiation complex has also been observed as part of a strategy to poise the promoter (Para et al., 2014). This type of preemptive organization can produce a fast response to possible stimuli that require immediate activation of transcription. When the full-length mRNA is formed transcription is terminated by polyadenylation, the 5' end is capped and the polymerase complex and all its co-factors are disassembled. Transcription results in the formation of a pre-mRNA. After this phase the pre-mRNA is spliced into a mature mRNA by removal of the introns. The mature mRNA is read out by the ribosomes to assemble the correct series of amino acids from which peptides and proteins are formed. This process is called translation.

1.2 Transcriptional regulation in plants

The advance in DNA sequencing technologies has provided researchers with complete genome sequences of a rising number of species. These genomic sequences can be used for annotating different gene types. At the level of transcriptional regulation these sequences can be utilized for the prediction of regulatory elements such as TFBSs. Next to complete genome sequences a wealth of sequencing data has become available measuring the expression levels of all genes through RNA-seq. Other methods investigate TF binding to their cognate bindings sites or the genome wide state of the chromatin through a number of techniques like ChIP-seq, Dnase1-seq or ATAC-seq. All of which contribute to the understanding of transcription.

Transcriptional regulation in plants, like in other eukaryotic organisms, is controlled by the interplay of TFs, epigenetic mechanisms and posttranscriptional processes. The complete set of interactions between TFs and their corresponding target genes, given all surrounding regulating mechanisms is called the gene regulatory network. In an attempt to simplify this complex interplay of regulatory levels, studying transcriptional regulation is often simplified to studying the binding of TFs to the cis-regulatory elements of their target genes. The complete set of TFBSs is also referred to as the cistrome encoded in the primary DNA sequence (O'Malley et al., 2016). In terms of TFs it has been suggested that there might be more than 2,300 genes encoding for TFs in *Arabidopsis thaliana*, a flowering plant model organism (Jin et al., 2014). This represents more than 5% of the estimated 30,000 genes in this plant species. All TFs are part of a specific TF-family, these families are based on the presence of a specific DNA binding domain in the coding region of the transcription factor gene. In Arabidopsis there exist 58 TF-families (Jin et al., 2014).

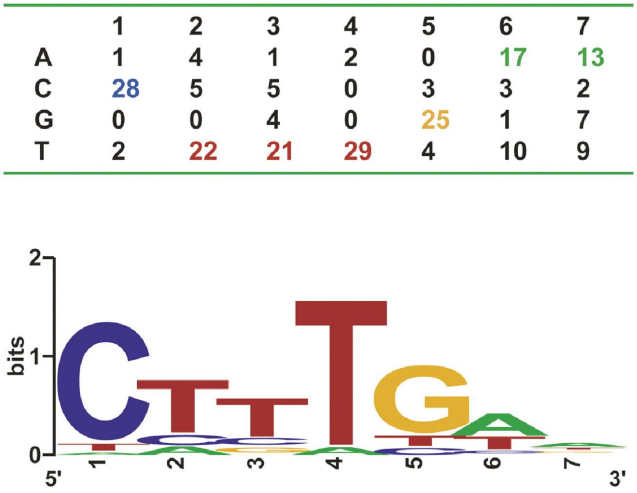

Figure 1.3: **PWM.***Position Weight Matrix (PWM), with the frequency of each nucleotide indicated for each position in the sequence and also visualized as a sequence logo.*

The coding sequence of TFs typically is made up of different domains: a DNA binding domain and a transcription activating domain or a transcription repressing domain. The DNA binding domain defines the genomic sequence to which TFs bind in the promoter of genes. These regions called TFBSs are often represented as a Position Weight Matrix (PWM), with the frequency of each nucleotide indicated for each position in the sequence (Fig. 1.3). There are various algorithms to scan for hits of PWMs in DNA sequences using the nucleotide frequencies independently. Recent research has also pointed out that for some TFs evidence of dependencies between nucleotides could be established whereas for others

this was not possible, leading to an improved detection of binding sites (Tomovic and Oakeley, 2007). Next to nucleotide dependencies within the PWM also DNA shape features (helix twist, minor groove width, propeller twist, and roll) augment the predictive power obtained by DNA sequence-based models (Mathelier et al., 2016). The promoter is typically split in a core promoter and a distal promoter (Molina and Grotewold, 2005). The core promoter is located circa 70 base pairs (bp) upstream from the TSS this is where recruitment of Pol II and assembly of the transcription pre-initiation complex happens. These core promoters often contain conserved CREs recognized by general TFs (Weake and Workman, 2010). The distal promoter contains conserved CREs that are responsible for binding transcription factors (TFs) that govern gene specific transcription regulation (Molina and Grotewold, 2005). The recruited protein complexes bind directly or indirectly to the same genomic segment containing discrete arrangements of TFBSs, also referred to as cis-regulatory modules (CRM). There are different classes of TFs, the first major class is formed by activators and repressors (Weake and Workman, 2010). Binding of these proteins can lead to either gene activation or repression. A specific class of proteins called cofactors, consists of co-activators and co-repressors. These proteins mediate the transcriptional effects of specific activators and repressors. This can possibly happen through chromatin remodelling (Weake and Workman, 2010). This group of TFs is typically not able to bind to DNA on their own, but specificity for a promoter is maintained through protein-protein interactions (PPIs) with specific activators and repressors. Other classes comprise general transcription factors that are important in the formation of the Pol II transcription-initiation complex and transcription factors that are involved in remodeling DNA by inducing bends that facilitate the binding of other proteins to the promoter.

The interaction between a TF and its target genes is not a solely defined by the presence of its corresponding binding site in the promoter of the target gene, it is also determined by how the chromatin is organized. Chromatin is the combination of DNA and all the associated proteins. The fundamental units of chromatin are nucleosomes, which consist out of DNA, wrapped around eight histone core proteins. The most important function of chromatin is compaction of the DNA, allowing the DNA to fit in the nucleus. Compaction of DNA forms heterochromatin and leads to general suppression of gene activity. A relaxed state of the chromatin is also possible and is called euchromatin. This allows TFs to bind and generally promotes gene expression (Roudier et al., 2009). Therefore nucleosome remodeling from heterochromatin to euchromatin is necessary allowing the DNA to relax and the TF to bind (Kaufmann et al., 2010a).

The state of the chromatin is dependent on the different modifications that are present. Chromatin can be modified in different ways (Fig. 1.4). The DNA sequence can be directly modified through the methylation of cytosines, histone tails can be covalently modified and histone variants can be incorporated in the nucleosomes (Kouzarides, 2007; Roudier et al., 2009). Finally, the structure chromatin modifications can also be adjusted through interactions with chromatin remodeling proteins (Kouzarides, 2007). Often recurring combinations of chromatin modifications are called chromatin signatures (Ernst and Kellis, 2010). These chromatin signatures describe specific locations in the genome or functional elements (e.g. the TSS of a gene). As mentioned above certain general TFs also influence the state of the chromatin through nucleosome remodeling. At this point there is an interaction between the epigenetic level and the transcription factors. The state of the chromatin and its effect on promoting or repressing transcriptional activation is dynamic and changes during development and in response to exogenous stimuli: e.g. stress, pathogen attack, temperature and light (Pfluger and Wagner, 2007).

After transcription and splicing, the mRNA is ready to be translated. At this point post-transcriptional regulation can interfere. Post-transcriptional regulation is achieved through the binding of miRNAs or short interfering RNAs (siRNAs) to the mature mRNA. The RNA molecule binds transcripts based on sequence similarity, which allows it to target entire gene families at once.

1.3 Principles of comparative genomics

Comparative genomics is a field in biology in which the genomic sequences of different species are compared. The fundamental assertion in comparative genomics is that all species have evolved from a common ancestor (Lyons and Freeling, 2008). Comparative genomics has recently known a serious boost

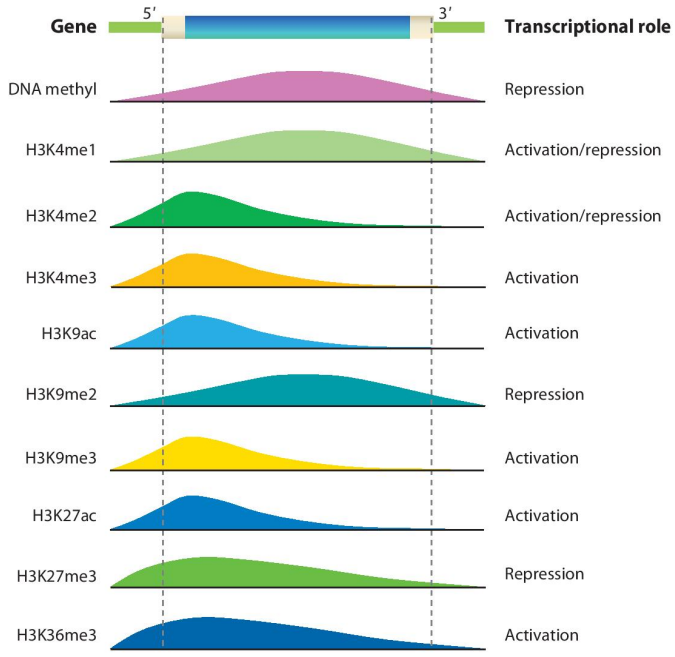

Figure 1.4: **Distribution of chromatin modifications over genes and their relationship with expression.***Chromatin marks analyzed using genome-scale approaches are represented along a schematic Arabidopsis gene with 5' and 3' flanking regions and the coding region indicated.* Source: (Roudier et al., 2009)

due to the fact that the genomic sequences of an increasing number of species are available. Genomes are compared in terms of the size and structure of a genome, the presence or absence of certain functionality, and in terms of how the structure of genomes and function of genes and genomic regions evolve (Hardison, 2003). As genomes evolve different types of rearrangements occur both on a large -and small-scale. Large-scale genomic rearrangements can vary from duplications and inversions to insertions and deletions. These processes alter the structural context of the genome but can also alter the functions of genes by changing the regulatory context. Besides these large-scale processes there are also rearrangements that occur on a smaller scale: nucleotide substitutions, deletions and duplications.

Comparative genomics can be applied on whole genomes to compare large structural properties such as genome size and the number of coding genes. However more in-depth analyses such as comparisons between chromosomes, parts of chromosomes and genes form the majority of what is done in comparative genomics today. When comparing genomic regions or even whole chromosomes two terms are frequently used: synteny and collinearity. Colinearity implies that genes on a chromosome or segment are in the same order and orientation as they were in an ancestor (Lyons and Freeling, 2008). Synteny is the property of being on the same chromosome or being present in the same genomic region through ancestry, but conservation of gene order and orientation is not necessary for a region to be called syntenous.

Homologous genes originate from a common ancestor (Koonin, 2005). A further distinction can be made for homologous genes regarding their origin, namely into orthologous and paralogous genes. Genes originating from a speciation event in the last common ancestor of the compared genomes are called orthologs (Fig. 1.5). An important feature of orthologs regarding the detection of cis-regulatory elements is that they are expected to have equivalent functions. Common ancestry combined with equivalent function is indicative that these genes are also regulated in the same way. If the genes are related due to a duplication event, these genes or regions are called paralogous. Paralogous genes are often subdivided into subgroups called inparalogs and outparalogs. This subdivision can only be made relative to a speciation event. Under these conditions inparalogs are paralogous genes resulting from a duplication after a given speciation event and outparalogs are paralogous genes resulting from a duplication before

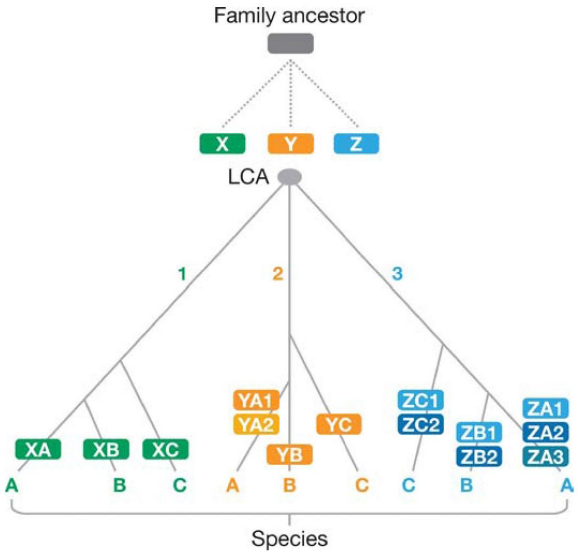

Figure 1.5: **An overview of orthology relationships.** *A hypothetical phylogenetic tree illustrating orthologous and paralogous relationships between three ancestral genes and their descendants in three species. The genes XA, XB and XC are one-to-one orthologs to each other. Each gene in branch 1 is a paralog of each gene in branches 2 and 3. The genes YA1 and YA2 are inparalogs relative to the radiation of species A and B. Branch 3 displays complex many-to-many orthology as frequently observed in plant species. LCA (last common ancestor)* Source: (Koonin, 2005)

a given speciation event. There lies great importance in this distinction between orthologs and paralogs, as it is crucial in the successful functional annotation of genomes and for the reconstruction of genome evolution (Koonin, 2005). The distinction between orthologs and paralogs is also very important in the detection of cis-regulatory elements. This is because of subfunctionalization, which is the selectively neutral tendency of a duplicated gene to lose dispensable cis-regulatory elements on one of both duplicates. Subfunctionalization spreads the functions of an ancestral gene over the duplicated descendant genes. Different approaches exist for identifying orthologous and paralogous genes (tree reconciliation and clustering based approaches). Detecting orthologs using a phylogenetic tree is called tree reconciliation and has the highest confidence to identify speciation events. Reciprocal best-blast-hit is another method that can be used for the detection of orthologs. This method is suited for the detection of orthologs in closely related species but it cannot deal with the complex one-to-many and many-to-many orthologous relationships in distantly related species. Other approaches such as OrthoMCL are based on sequence similarity-based clustering (Li et al., 2003). Van Bel et al have shown that integrating and combining the predictions of these tools leads to a more accurate orthology identification (Van Bel et al., 2012).

An important tool in comparative genomic analyses is sequence alignment. A sequence alignment is best described as a data matrix where each row is a sequence and each column is a position in the sequence. A position in an alignment can represent three things: a match, a mismatch or a gap. The result of a sequence alignment shows the similarity between the sequences that were aligned. The goal of a sequence alignment should be however to detect homology and not similarity (Fig. 1.6).

The neutral theory of molecular evolution states that nucleotide substitutions inherently take place during DNA replication. In the absence of selective constraints, the substitution rate reaches the maximum value set by the mutation rate or the neutral substitution rate. Rates slower than the neutral substitution rate indicate the presence of constraints imposed by negative selection, which rejects and discards deleterious mutations (Jukes and Kimura, 1984). In distantly related species, it is very difficult to align non-coding sequences due to the neutral substitution rate since this leads to saturated substitution. This means that in absence of constraint enough substitutions have occurred so that any position in two aligned sequences has been changed. The neutral substitution rate is often estimated using ancestral repeats or

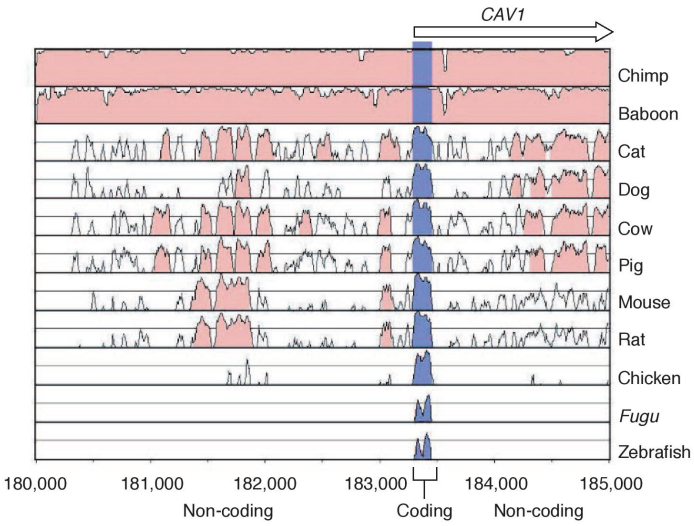

Figure 1.6: **The difference between similarity and constraint.** *Sequence similarity is shown in pink. The high percentage of sequence similarity between Chimp and Baboon is a nice example of sequence conservation due to neutral carry over because of the short divergence time between these two species. Comparing chimp to some other species reveals stretches of conserved non-coding sequence, illustrating constraint. This figure is also illustrative for the fact that when the phylogenetic distance becomes too big, it becomes difficult to pick up the phylogenetic signal, e.g. Chicken.*

Ks. Ks is defined as the number of synonymous substitutions per synonymous site. A synonymous site refers to a position in a codon where redundancy of the genetic code exists so that different nucleotides in this position will result in the same amino acid. An important fact to be noted is that ancestral repeats are non-coding sequences that are assumed to not be under constraint. As Ks is based on coding sequences the assumption of absence of constraint is probably not completely true.

1.4 Identification of transcription factor binding sites

Characterization of TFBS in vivo

Chromatin Immunoprecipitation (ChIP)

ChIP is a broadly adapted technique to identify TF target genes in a direct manner. ChIP consists out of cross-linking the TF to the DNA followed by sonication of the DNA into small pieces. This step is followed by extraction of the TF and the DNA that precipitates through antibodies that directly target the TF (Thibaud-Nissen et al., 2006). In the next step, the cross-linking is reversed and the bound DNA sequences are determined by hybridization on a tiling array or sequencing analysis (ChIP-chip and ChIP-seq, respectively) (Fig. 1.7). The output of a ChIP experiment is further processed computationally and is essentially aimed at mapping the bound DNA sequences back to their respective genomic regions and identifying local enrichments of signal in the sample of interest compared to the control sample. These enriched regions are also referred to as peaks. A common final step in a ChIP analysis is linking the enriched regions to closest gene and as such linking TF bound regions to their respective target genes (Farnham, 2009).

The TF bound regions or peaks generated in a ChIP experiment provide an excellent resource for the identification of the sequence specific binding motif responsible for interaction with the TF, since these peaks should contain the sequence motifs recognized with highest affinity. There are two complementary strategies that can be used for obtaining these binding motifs. The first one is based on the de novo discovery of motifs. This method identifies overrepresented words of different (6-12bp) lengths in the bound sequences. The second one scans the TF bound sequences against databases of known motifs and determines enriched motifs (Higo et al., 1999; Sandelin et al., 2004; Steffens et al., 2004; Yilmaz et al., 2011). A large-scale analysis in *Arabidopsis thaliana* has reported a fairly good agreement between TF-

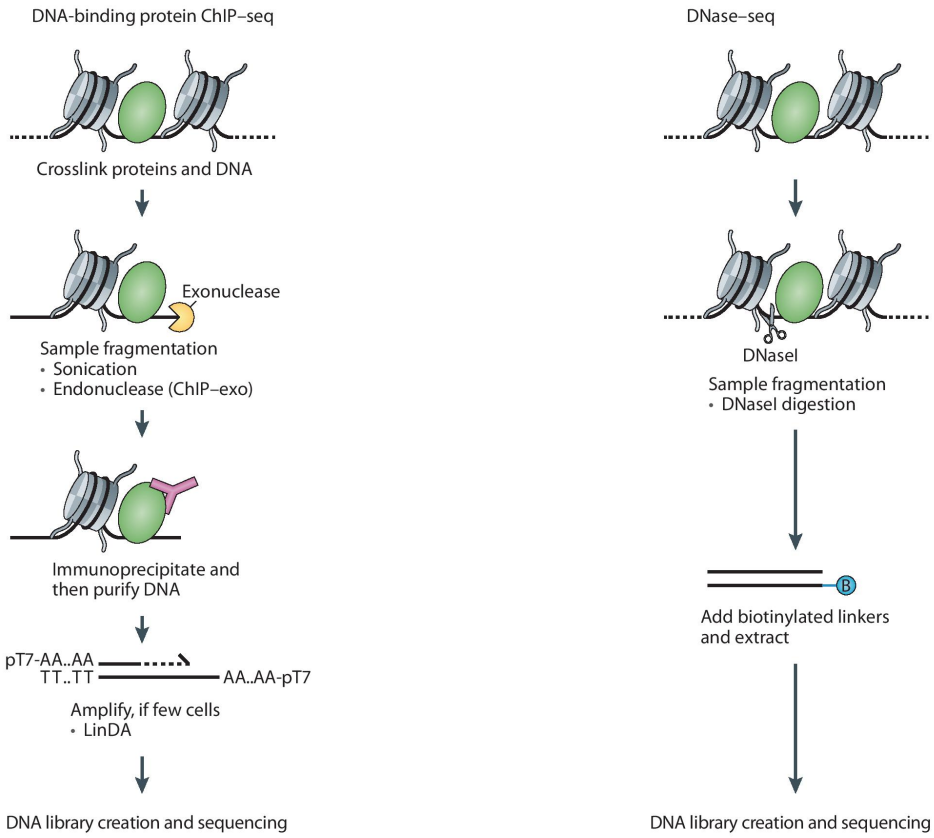

Figure 1.7: **Simplified schematics of the main steps in ChIP-seq and DNase-seq.** *Chromatin immunoprecipitation followed by sequencing (ChIP-seq) for DNA-binding proteins such as transcription factors. DNase-seq relies on digestion by the DNaseI nuclease to identify regions of nucleosome-depleted open chromatin where there are binding sites for all types of factors, but it cannot identify what specific factors are bound.* Source: (Farnham, 2009)

bound DNA motifs in vivo and the consensus DNA sequences derived from in vitro analysis (Heyndrickx et al., 2014). Similar results were obtained for ChIP-based data from the ENCODE Project, in which 55% of TF-occupied DNA segments were enriched in the anticipated DNA-binding motif.

DNase I Hypersensitivity

A downside to ChIP based assays is the necessity of specific, high-grade antibodies that recognize the TFs of interest. A valuable alternative is the identification of nucleosome-free, open chromatin regions that correspond mostly to gene regulatory regions, and more specifically to TFBS. One of these approaches, DNase-seq is a technique for measuring regions of DNase I hypersensitivity genome-wide. This technique involves digestion of chromatin within intact nuclei with DNase I. Cleavage products are selected on their size and sequenced, such that the 5'-most bp of a sequenced read indicates the site at which cleavage by DNase I occurred (Fig. 1.7). Various peak-calling algorithms are used to identify DNase I hypersensitive sites. DNase-seq has been applied successfully in Arabidopsis and rice. These experiments recovered approximately 40,000 and 100,000 DHS per tissue sample, covering 12 and 7% of the Arabidopsis and rice genomes, respectively (Zhang et al., 2012b,a) (Fig. 1.8).

In order to obtain TFBS from these DHS a technique called genomic footprinting is used. DNase I treatment generates a footprint in TF-protected DNA motifs. The algorithms to perform genomic footprinting vary, but all rely on the detection of a cleavage pattern consistent with protein occupancy. This pattern is classically defined as a paucity of cleavage because protein-bound DNA is typically protected from DNase I cleavage.

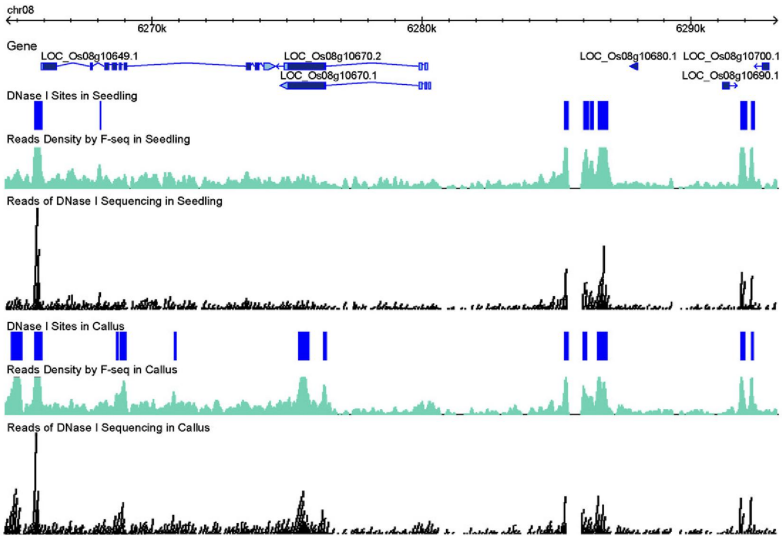

Figure 1.8: **Example of DNase hypersensitive sites in rice.** *This figure displays the transformation of raw sequencing reads obtained by DNase-seq into DNase hypersensitive sites through read density estimation by the F-seq tool.* Source: (Zhang et al., 2012a)

Definition of TF DNA-binding specificities in vitro

Yeast-one-Hybrid (Y1H)

Y1H is not strictly an in vitro system and has been very useful for identifying cis-regulatory sequences. Unlike ChIP based assays, this system is a TF-centered or bottom-up strategy that pursues isolation of TFs that recognize a known prey. 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 allows screening of binding events based on the expression of the reporter gene. Improvements made to the original technique for detection of protein-DNA interactions (Li and Herskowitz, 1993) facilitated the high-throughput and unbiased identification of protein-DNA interactions (Deplancke et al., 2004; Vermeirssen et al., 2007). Hundreds of studies have demonstrated the efficacy of Y1H assays for discovering TF binding with both small (e.g., TFBS), and large DNA fragments (e.g., gene promoters). Y1H was used to construct a gene regulatory network in root (Brady et al., 2011; Gaudinier et al., 2011).

Systematic Evolution of Ligands by Exponential enrichment (SELEX)

SELEX was widely used to determine the DNA binding specificities of TFs (Ellington and Szostak, 1990). The technique is based on the iterative selection and amplification of the DNA sequence with the highest affinity for a given TF. The classical SELEX has two major drawbacks. The first one is that the isolation of TF-DNA complexes from polyacrylamide gels after each round of selection is very laborious. The second reason is that the number of oligonucleotide sequences that could be detected in the final step is limited. Modifying SELEX into a high throughput (HT) analysis has considerably improved the applicability of SELEX-based methods. HT-SELEX utilizes affinity-based steps for isolation of TF-DNA complexes and includes an initial pool of oligonucleotides that is suitable for readout using deep sequencing (Zhao et al., 2009).

Electrophoretic Mobility Shift Assay (EMSA)

EMSA is a gel-based method to separate protein-bound DNA molecules from unbound DNA molecules through the larger weight of the bound complex 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 (Elnitski et al., 2006).

Protein binding microarray (PBM)

PBM is a high-throughput methodology for determining the sequence specificity of a TF. This method utilizes a double-stranded DNA (dsDNA) microarray that is incubated with the TF of interest, and the hybridization signal of fluorescent protein-DNA complexes are visualized in conventional microarray scanners (Berger et al., 2006).

PBMs have a number of advantages that make it preferable over other strategies for identification of DNA motifs in vitro. The optimization of dsDNA microarray designs has created a high throughput technique in a single experiment. The most used microarray contains all possible 10mers compacted in ~44,000 oligonucleotide probes with a length of 35 bp. Consider both strands, this means that every 8mer is represented in ~32 probes (Berger and Bulyk, 2009). PBMs have a universal design containing synthetic DNA probes and as such can be used for any TF independent of selected species(Fig. 1.9). Unlike HT-SELEX, which preferentially recovers high-affinity bound sequences, PBM analysis ranks binding to all possible sequences (generally 8mers) from highest to lowest, allowing identification of low-affinity bound sequences. Over a decade of research on PBMs has lead to a well-established technique with a simplified workflow and advanced algorithms for data analysis (Berger and Bulyk, 2009).

PBMs were applied to determine binding site specificities of MYC2 and ERF1 two TFs in Arabidopsis. this analysis rendered a G-box and the GCC-box as their highest-affinity binding sites respectively. Further analysis of transcriptome datasets revealed that high- and medium-affinity binding sites have biological significance, probably representing relevant cis-acting elements in vivo. Comparison of promoter sequences with putative orthologs from closely related species demonstrated a high degree of conservation of all the identified TFBSs (Godoy et al., 2011).

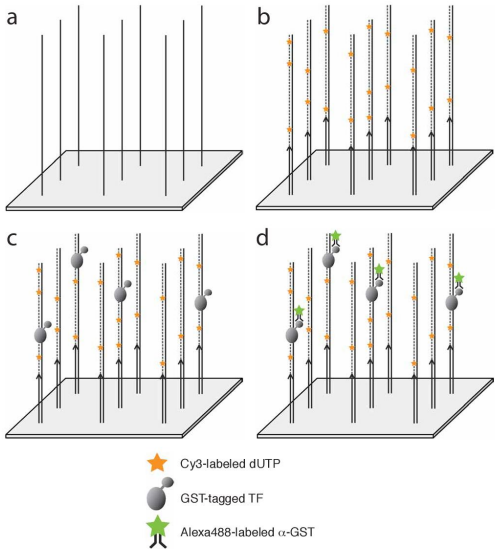

Figure 1.9: **Schematic of universal PBM experiments.** *A commercially synthesized single-stranded DNA microarray is double-stranded by (b) solid-phase primer extension using a small amount of spiked-in fluorescently labeled dUTP. (c) An epitope-tagged TF is bound directly to the DNA on the microarray, and the (d) protein-bound array is labeled with a fluorophore-conjugated antibody.* Source: (Berger and Bulyk, 2009)

1.5 Inferring functional target genes from TFBS data

Identifying groups of genes that collaborate in the development of an organism or in the response to specific stimuli or stress is essential to improve our understanding of these processes. Delineating sets of TF target genes through identification of TFBS in promoter sequences is a powerful strategy for defining specific sets of genes. As described above TF ChIP-seq is an effective technique for obtaining target genes. Despite the advantages of ChIP-based assays, they are nevertheless a snapshot of the genome-wide TF occupancy at a certain point in time, and usually ignore the plasticity in binding events of the TFs that control gene expression. A second downside is that ChIP based assays largely ignore the interactions that exist with other TFs. These downsides are illustrated by the poor correlation between

TF-bound genes and genes transcriptionally regulated by TFs being studied. TFs studied in *Arabidopsis thaliana* report numbers of TF-responding genes between 5%-30% of the total TF-bound genes (Zheng et al., 2009; Winter et al., 2011). These findings illustrate the difficulties in identifying functional TFBS from genome-wide mapping studies, and with them the actual target genes of TFs. Functional TFBSs are defined on the basis of the paramount role in controlling patterned gene expression. Precise definition of TF target genes will require ChIP data combined with comprehensive approaches that monitor gene expression over time, in specific tissues, at different developmental stages, or in changing environmental contexts. An example in Arabidopsis of such an integrative approach was presented with the elucidation of the gene regulatory network that governs the response to ethylene, which depends on the EIN3 TFs. The response follows four temporal waves, each containing subsets of the EIN3-bound targets. This study proposed several layers of transcriptional control, in which a negative feedback loop (which depends on EIN3 binding to targets) is generated to maintain ethylene homeostasis (Chang et al., 2013).

Performing a ChIP assay is an experimental labor-intensive technique and it is at this point not feasible to perform this on a very large scale. Therefore computational methods were developed for defining sets of target genes based on DNA motifs in promoter sequences. There are two main schools in this line of research, one type makes use of evolutionary conservation of TFBS, the other uses transcriptional co-regulation to create gene regulatory networks.

A promising technique for the identification and/or validation of target genes through TFBS is phylogenetic footprinting. Phylogenetic footprints are islands of highly conserved regions embedded within a background of neutrally evolving sequences (Tagle et al., 1988). Regions of non-coding DNA in the genome that are conserved across related species are likely to be under purifying selection because of the fact that functional sequences tend to evolve at a slower rate than nonfunctional sequences (Frazer et al., 2004). The conservation of functional elements is not sufficiently better than that of non-functional elements when the species are too closely related. This is due to the fact that patches of conserved noncoding sequence will happen naturally by neutral carry over from the ancestor (Freeling and Subramaniam, 2009). Oppositely, if two species are only very distantly related most of the evolutionary signal will be lost and will be very difficult to pick up using phylogenetic footprinting. Therefore adequate phylogenetic divergence of species is an important parameter in phylogenetic footprinting.

Phylogenetic footprinting in plants is more complex than in vertebrates due to three reasons. Firstly conserved non-coding sequences are shorter in plants than in vertebrates (Kaplinsky et al., 2002; Guo and Moose, 2003). A second reason is that plants have a higher neutral nucleotide substitution rate over evolutionary time. They mutate faster than vertebrates. This results in a saturated substitution rate for sequences that are not under constraint and also induces more degeneracy in motifs (Freeling and Subramaniam, 2009). The last reason why phylogenetic footprinting is more difficult in plants has to do with the evolution of plant genomes. Plant genomes have been subjected to many genomic rearrangements as well as whole genome duplications. This has resulted in complex ancestry of genes, with many one-to-many and many-to-many orthologous relationships, making the distinction between orthologs and paralogs very important (Proost et al., 2011).

CNS research in the past has been focused on Arabidopsis and grasses (Fig. 1.10) (Kaplinsky et al., 2002; Vandepoele et al., 2006; Freeling et al., 2007; Vandepoele et al., 2009; Baxter et al., 2012; Guo and Moose, 2003; Inada et al., 2003). Exploring CNSs for TFBS identified in vitro or from in vivo assays will help to define target genes of TFs and gene regulatory networks in plants. An attempt to create a conserved TFBS gene regulatory network by looking for TFBSs in conserved genomic non-coding regions has to this point only been performed in Drosophila (Kheradpour et al., 2007).

Another approach used to define target genes is based on the hypothesis of strong co-regulation between a TF of interest and its targets (MacIsaac and Fraenkel, 2006). This assumption is extended to the shared presence of a common TFBS amongst these sets of genes. The presence of TFBS in co-expression clusters can be evaluated through de novo motif finding or bindings site enrichment analysis with a set of known TFBS. Despite the fact that genes with a shared TFBS show on average higher co-expression it was shown that strongly co-expressing do not always share a TFBS (Marco et al., 2009). Microarray or RNA-seq expression datasets often deal with whole organism gene expression levels at different develop-

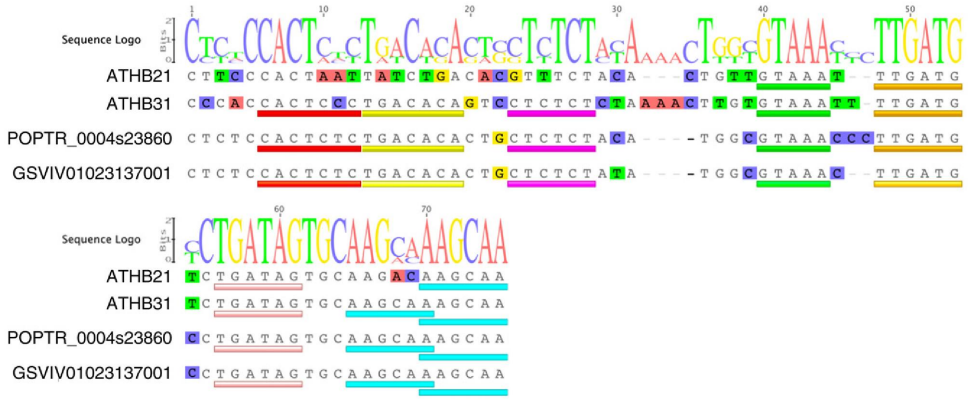

Figure 1.10: **Example of CNSs in Arabidopsis.** *Alignment of CNS depicted in (B). Size of letters in the sequence logo indicates conservation of individual nucleotides. Colored bars indicate positions of potential binding sites based on alignment conservation (yellow, purple, green, and orange bars) and matches with known motifs (P300 in red, GATA in pink, and CBNAC in turquoise.* Source: (Baxter et al., 2012)

mental moments. Given that genes can be transcribed in different tissues simultaneously in multicellular organisms, this aspect is potentially confounding and may be the reason why there is no statistical enrichment of TFBSs in even highly co-expressed genes. The presence of indirect regulatory interactions might also influence the association between co-expression and TFBSs. The differential affinity to binding sites or regulation by miRNA could also explain the partial uncoupling between expression patterns and regulation by common transcription factors. Nevertheless, this method is very popular in regulatory analysis. Predominant TF activator or repressor activities can also be inferred from analysis of co-regulated genes. DNA motifs identified in vitro for transcriptional activators are enriched in the promoters of positively co-regulated genes; conversely, motifs corresponding to repressor TFs are more abundant in promoters of negatively co-regulated genes (Kagale et al., 2010). This approach has also been applied in Arabidopsis to perform motif enrichment in co-expressed gene sets and obtain functionally coherent groups of target genes. (Vandepoele et al., 2009).

CHAPTER 2

Research Aims and Scope

The adaptation of organisms to changes in the surrounding conditions and the environment very often co-occurs with adjustments in the expression pattern of specific genes. These mechanisms are especially important for plants because they are sessile organisms. Because plants cannot simply change their location they need to be ensured that they are able to cope with different conditions each meteorological season by adequate gene expression modifications. An important goal for understanding these biological processes is getting a better insight in the regulation of gene expression. This requires a profound comprehension of how transcription is regulated by TFs on the molecular level. This PhD thesis will focus on transcriptional regulation in plants and will exhibit the application of different approaches to enhance what is currently known on this subject.

The main goal of this PhD is the development of novel methods for the detection of functional and regulatory elements in plant genomes and converting this knowledge into gene regulatory networks and derived applications. The research will be subdivided in two levels: The first level is the detection of binding sites and inference of gene regulatory networks. As many genes in plants lack functional information to this date, the second is the application of the resulting networks to study biological processes and as such provide reliable functional annotation for as many genes as possible.

Initially, a tool to perform phylogenetic footprinting between distantly related genomes will be developed. This tool will be used to delineate conserved non-coding sequences (CNS) through comparative sequence analysis. The genomic properties of the obtained CNSs will be studied in order to obtain novel insights in their role in transcriptional regulation. Through the integration of known transcription factor binding sites obtained from literature and experimental studies, these CNSs will be used to compile gene regulatory networks for a large number of TFs. This network will be validated using known targets and applied as a means to study the transcriptional organization in plants.

The biotechnology industry will become an important factor in both the food and energy industry. Its focus will lie on commercial crops such as the different *Brassicaceae* (dicots), and cereals (monocots). Nevertheless, a lot of fundamental research is being performed on the model organisms *Arabidopsis thaliana* and *Oryza sativa* (rice) for the dicots and monocots respectively. As a consequence, it is essential to know to what degree components and interactions of transcriptional regulation are shared between different species. With respect to these outstanding demands, comparative sequence analysis in non-model species will be used to delineate conserved non-coding sequences in these organisms and study the evolution of TF binding sites. A metric controlling for evolutionary distance will be implemented for analyses in closely related organisms. The results of these analyses will examine the feasibility of transferring knowledge derived from model species to more commercially interesting crop species but the results will also contribute to the limited knowledge of transcription regulation in non-model organisms.

Subsequently, we will study the organization of gene regulation starting from a small set of TFs for which experimentally derived binding specificities are generated. This experimental data will be processed into predicted transcription factor binding sites and integrated with complementary data types to obtain high quality binding sites. These binding sites will be experimentally validated. The resulting network will be evaluated for known targets of the TFs under investigation and through functional en-

richment of the predicted target genes, the biological role of these TFs will be studied using the guilt-by-association principle. As a whole this research chapters serves as an application of previously developed methodologies.

CHAPTER 3

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. (2014). 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. A complete reference can be found in Chapter 8.

3.1 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.(Riechmann and Ratcliffe, 2000) 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 (Riechmann et al., 2000; Jin et al., 2014), knowledge about the genes controlled by different TFs is still very limited. AtRegNet, which is a part of the AGRIS database (Yilmaz et al., 2011), 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 (Garner and Revzin, 1981), systematic evolution of ligands by exponential enrichment (Roulet et al., 2002) and Yeast-one-hybrid (Meng et al., 2005) are labour-intensive and only yield a small number of interactions.(Mejia-Guerra et al., 2012) 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(Ferrier et al., 2011).

Dozens of software tools have been developed to delineate regulatory regions based on experimental features, such as co-regulation, or using advanced computational methods(MacIsaac and Fraenkel, 2006). Although the naive 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 (Tompa et al., 2005). 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 (Zhang et al., 2012b), 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 (Tagle et al., 1988; Handstad et al., 2011). 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(Blanchette and Tompa, 2002; Kaplinsky et al., 2002; Guo and Moose, 2003; Inada et al., 2003; Vandepoele et al., 2006; Thomas et al., 2007; Vandepoele et al., 2009; Baxter et al., 2012). 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 (Siepel et al., 2005), 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.(Siepel et al., 2005) 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.(Hupalo and Kern, 2013) 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.(Reineke et al., 2011) 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.(Van Bel et al., 2012) 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.(De Bodt et al., 2006; Vandepoele et al., 2006)

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 (Picot et al., 2010), to search for CNSs in the 2kb upstream of the transcription start site in Arabidopsis.(Baxter et al., 2012) 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. (2013) generated a whole genome alignment approach using a combination of the LASTZ (Harris, 2007) and MULTIZ (Blanchette et al., 2004) 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 (Pollard et al., 2010) 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 (Siepel et al., 2005) 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.

3.2 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 A.1). The selected comparator dicot species used in this study are reported in Figure A.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 $\geq$ 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 (Subramanian et al., 2008), Sigma (Siddharthan, 2006), ACANA (Huang et al., 2006) and the Seaweeds alignment plot tool.(Picot et al., 2010) 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) (Zhang et al., 2010), 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).(Chaboute et al., 2002; Liu et al., 2010; Wang and Xu, 2010) 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 (Van Bel et al., 2012), 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 A.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 3.1A). All of the significant CNSs were conserved in at least two comparator species while the median number of supporting species was six (Figure 3.1B). 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 3.1C). 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.

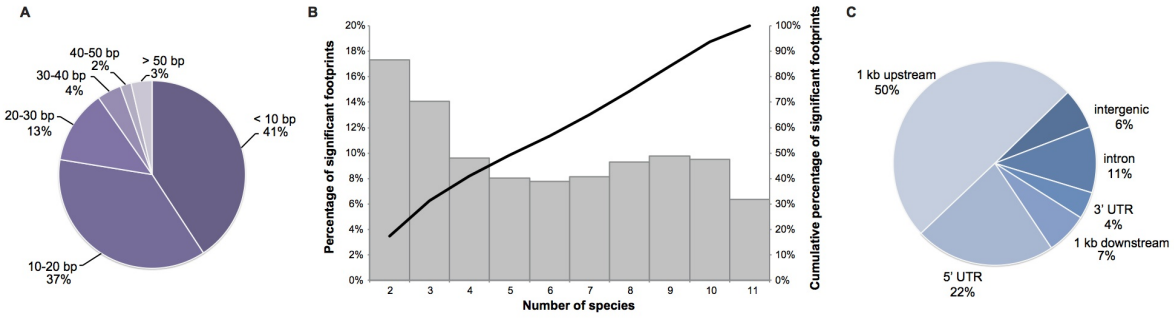

Figure 3.1: **Overview of CNS properties.** *A) Length distribution of significantly conserved footprints. All footprints are grouped in bins of size 10bp. B) Overview of significantly conserved footprints in relation to the number of species in which the footprint was conserved. For all conservation 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[a], 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 A.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 3.2).(Baxter et al., 2012; Haudry et al., 2013; Hupalo and Kern, 2013) 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

[b] http://www.plantcell.org/content/suppl/2014/06/16/tpc.114.127001.DC1/tpc127001_Supplemental_Datasets.xls
[a] https://cb.utdallas.edu/atprobe/index.html

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.

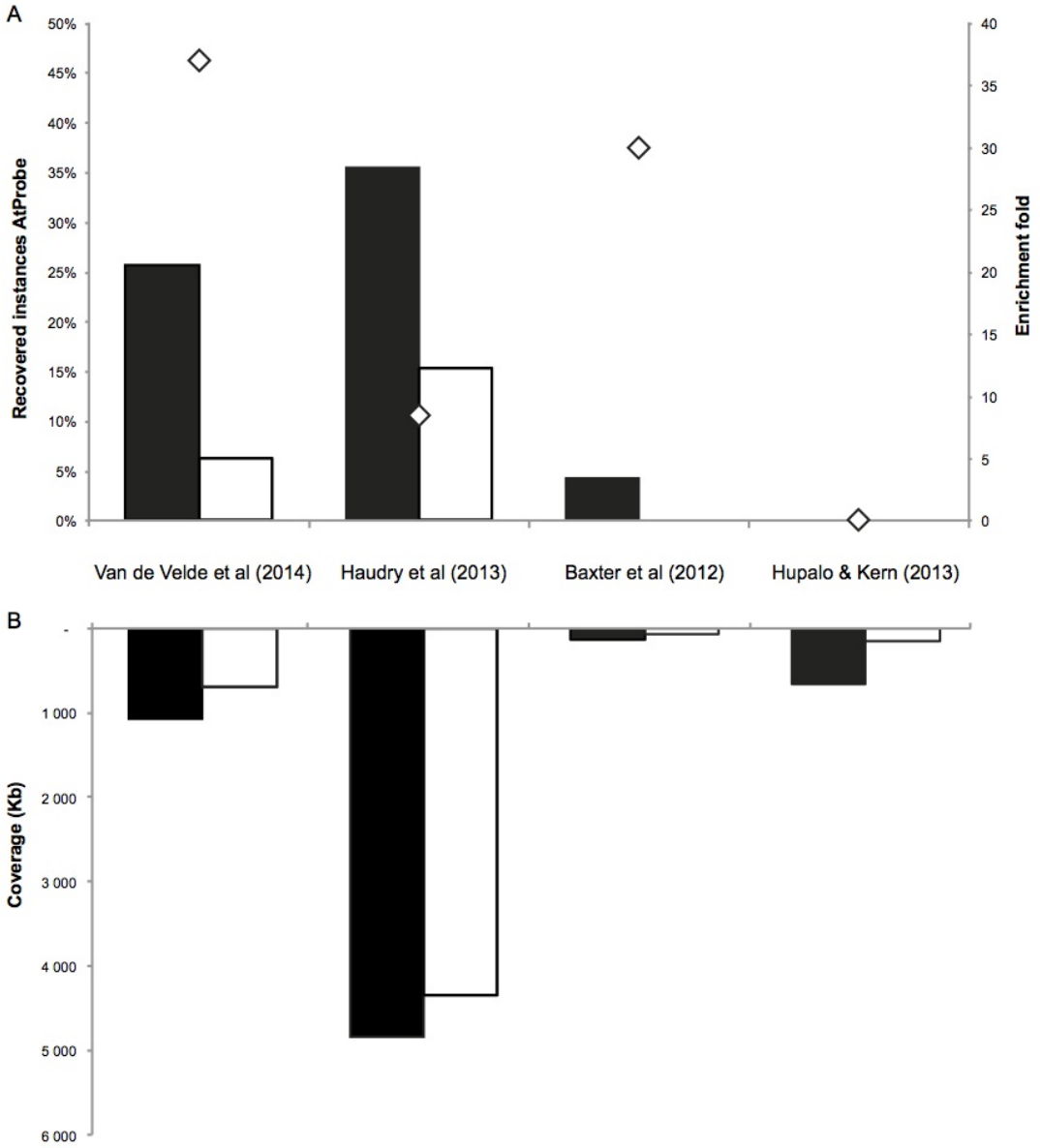

Figure 3.2: **Recovery of AtProbe elements and comparison of CNSs from different phylogenetic footprinting studies.** *(A) Overview of the recovery of experimental AtProbe elements in four different CNS studies. Black boxes show the percentage of recovered elements and white boxes shows the percentage 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.*

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. (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 A.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 3.2). 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 3.2). 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.(Zhang et al., 2012b) This overlap is significant (p-value ≤ 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 A.5). These histone modifications are indicative of active promoters and enhancer elements.(Roudier et al., 2009; He et al., 2011) Interestingly, our regions showed an even higher enrichment for regions where DH sites, H3K4me3, H3K9ac and H3K4me2 coincide (6.3 fold enriched, p-value ≤ 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 3.3).

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 com-

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

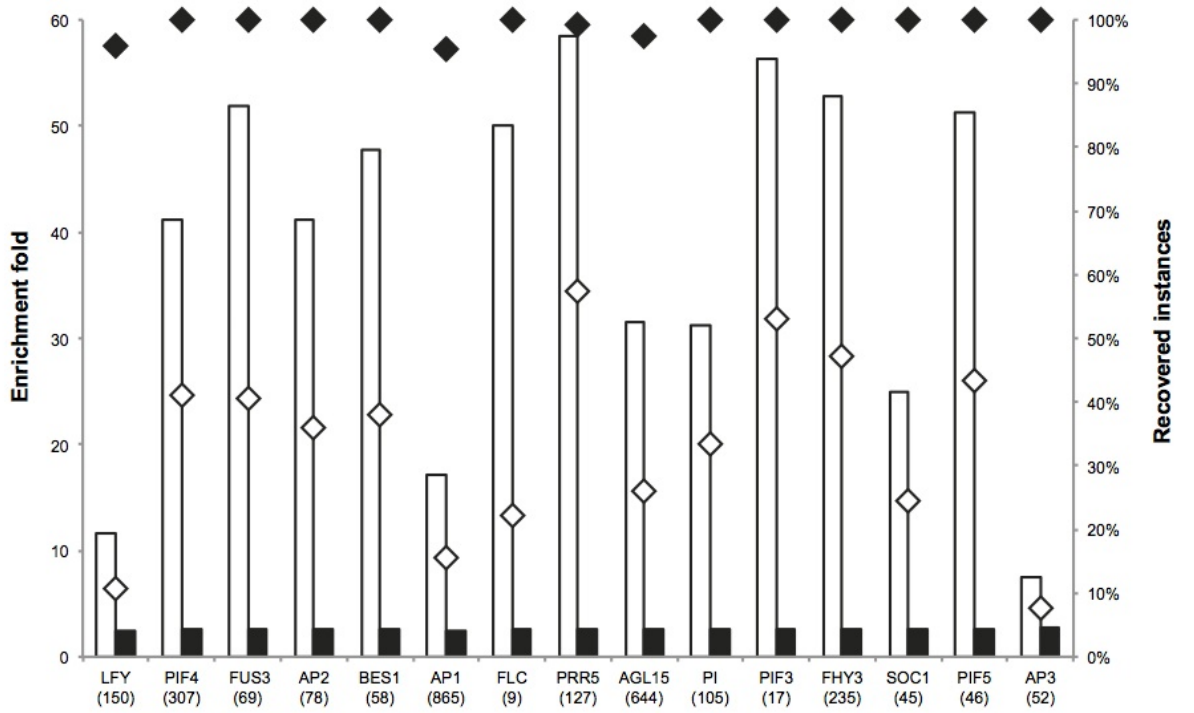

Figure 3.3: **Recovery of *in vivo* functional targets using CNS information.** *White and black boxes show fold enrichments for CNSs and naive motif mapping, respectively. White and black diamonds show the fraction of recovered elements for CNSs and a simple motif mapping approach, respectively.*

prises motif mapping in open non-coding chromatin regions that were identified through DH sites.(Zhang et al., 2012b) 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 3.3, Supplemental Online Data set 3[b] and Figure A.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 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 A.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 (Davuluri et al., 2003) and a collection of regulatory interactions obtained from small-scale studies concerning secondary cell wall metabolism.(Hussey et al., 2013) 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 ≤ 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 ≤ 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).(Hong et al., 2003) 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.(Li et al., 2012b) Another interesting, however unconfirmed, interaction is that between AP2 and mir167a, the latter which is known to play a role in flowering maturation.(Rubio-Somoza and Weigel, 2013)

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) (Ashburner et al., 2000), Mapman (Thimm et al., 2004) and functional gene modules (Heyndrickx and Vandepoele, 2012) 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.(Marbach et al., 2012b) 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 (De Bodt et al., 2010), 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 ≤ 0.05, Figure 3.4). 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 3.4). 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 A.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 (De Bodt et al., 2010), 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) (see Methods). Co-expression be-

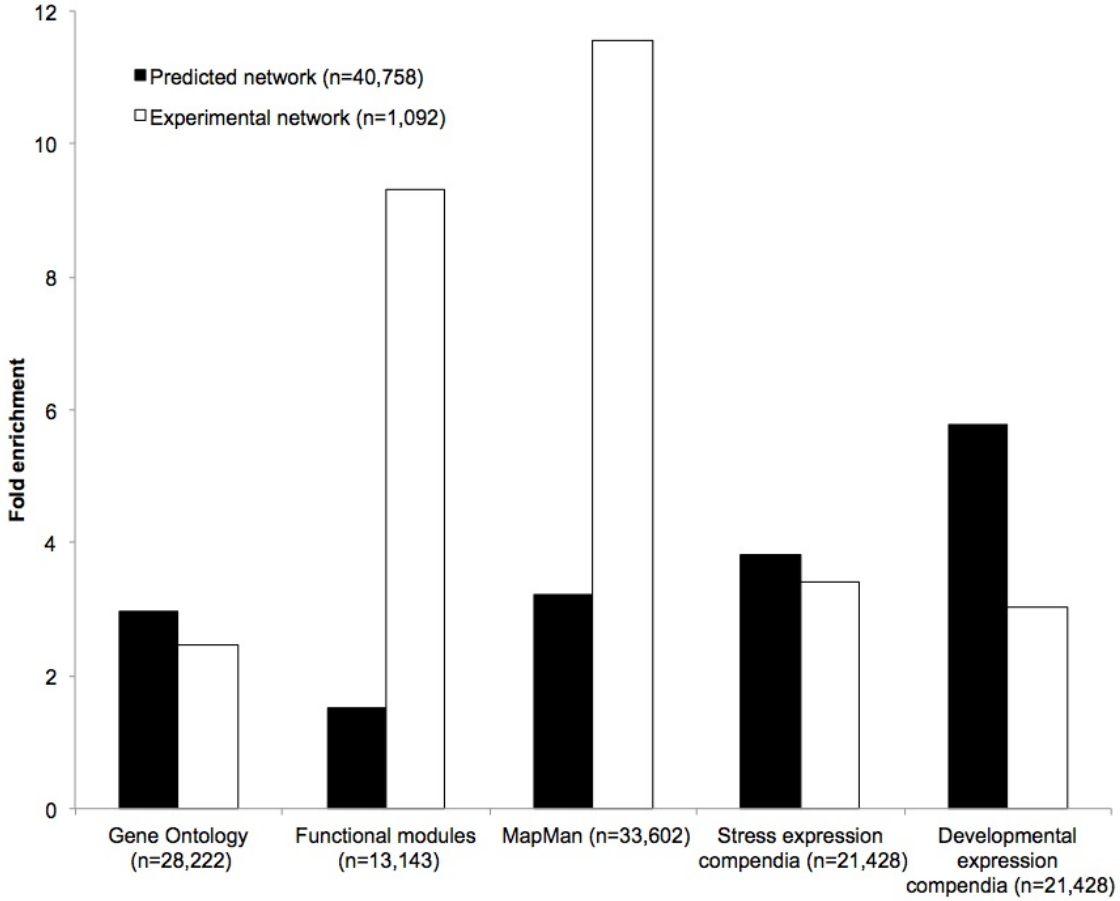

Figure 3.4: **Evaluation of the biological relevance of the predicted network using different biological metrics assessing functional and expression coherence.***Gene Ontology annotations, Mapman annotations and functional modules together with a stress and developmental expression compendium were used 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.*

tween a TF and a predicted target gene can act as a proxy for regulation as both are frequently expressed in the same conditions.(Ma and Wang, 2012) 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 A.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 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.(Zhong et al., 2008) 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;.(Zhong et al., 2008) 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.(Zhong et al., 2010) For VND6 an additional binding site has been described (CTTNAAAGCNA).(Ohashi-Ito et al., 2010) 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; (Zhong et al., 2008) (Figure 3.5). 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.(Kubo et al., 2005) Both TFs reside in the same functional module, which is annotated with the GO term 'response to brassinosteroid stimulus'.(Heyndrickx and Vandepoele, 2012) 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 3.5). 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.(Zhong et al., 2008) 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.(Zhong et al., 2008)

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.(Wuest et al., 2012) AP3 and PI are necessary for the proper development of the petals and stamens.(Jack et al., 1992; Goto and Meyerowitz, 1994) Plant hormones such as jasmonic acid have been shown to play a role in both stamen and petal development.(Brioudes et al., 2009; Song et al., 2013) 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 A.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)(Wuest et al., 2012). 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.(Ito et al., 2007)

3.3 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

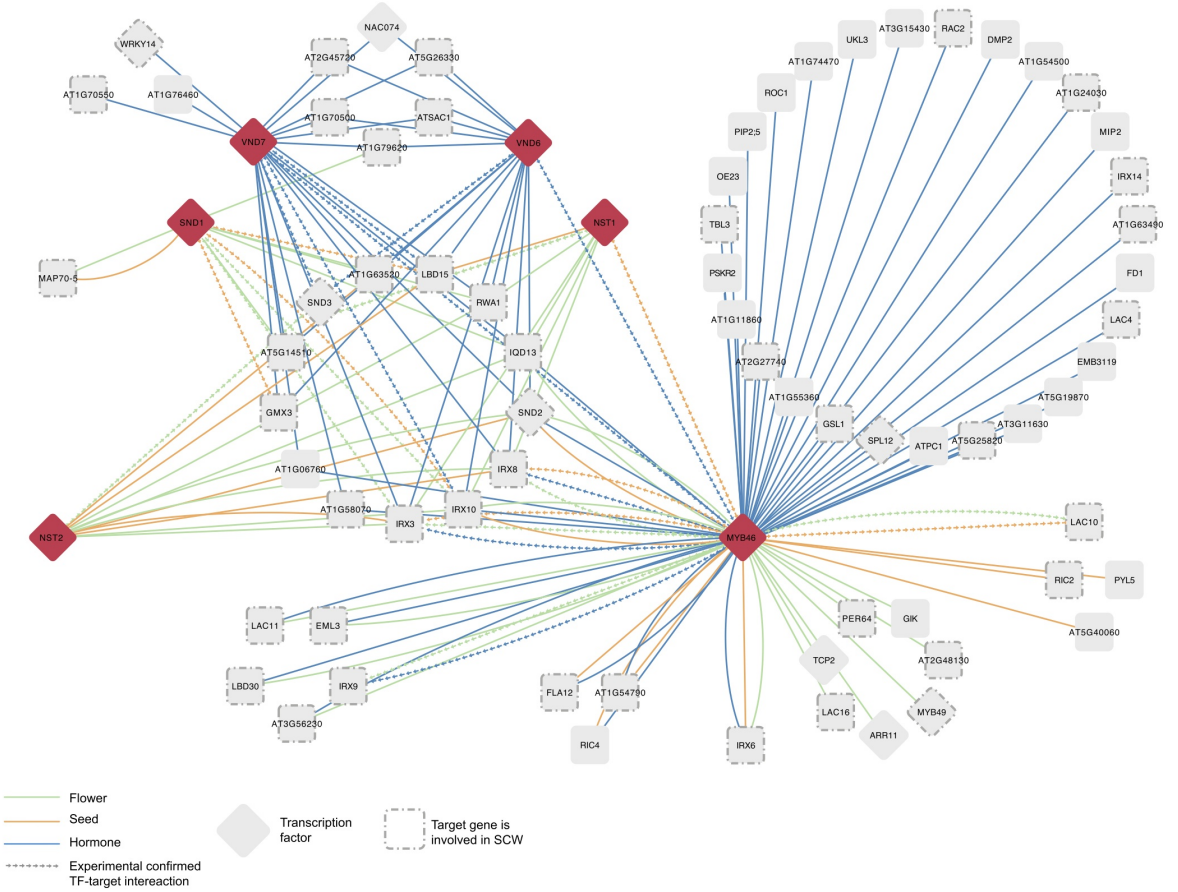

Figure 3.5: **A condition-specific secondary cell wall gene regulatory network.***Nodes and edges depict genes and regulatory interactions, while condition-specific seed, flower and hormone co-expression edges are shown using orange, green and blue lines, respectively. Experimentally confirmed interactions are shown using an arrow line. Red diamonds are the source TFs, grey diamonds are target genes that are TFs and rounded rectangles are other target genes. Target genes with a grey border are known to be involved in secondary cell wall biosynthesis based on Gene Ontology.*

of base pairs upstream of a gene.(Reineke et al., 2011; Korkuc et al., 2014)

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.(Reineke et al., 2011) Phylogenetic footprinting methods that use genome synteny inferred through genome alignments as primary source of orthology information indeed have difficulties integrating distantly related genomes.(Hupalo and Kern, 2013) This is due to the frequent nature of polyploidy and genome rearrangements in dicot plants (Figure A.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 3.1B). Furthermore, approaches based exclusively on pairwise alignments lack the power to detect a large set of our CNSs over a similar evolutionary distance.(Reineke et al., 2011; Baxter et al., 2012)

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 (Baxter et al., 2012; Haudry et al., 2013; Hupalo and Kern, 2013), 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).(Riechmann et al., 1996) 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 (Kaplinsky et al., 2002; Guo and Moose, 2003; Inada et al., 2003; Baxter et al., 2012; Haudry et al., 2013; Hupalo and Kern, 2013), 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 (Franco-Zorrilla et al., 2014; Lindemose et al., 2014) 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.

3.4 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.(Van Bel et al., 2012)The structural annotation of the genomes in PLAZA 2.5 was updated by adding all known miRNAs obtained from the plant microRNA database.(Zhang et al., 2010) miRNA

sequences were downloaded from PMRD and mapped to the genomes using BLASTN (Altschul et al., 1990) and GenomeThreader (-mincoverage 0.89 -minalignmentscore 0.95) (Gremme et al., 2005) 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.(Van Bel et al., 2012) The included orthology detection methods are OrthoMCL (Li et al., 2003), 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.(Van Bel et al., 2012)

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.(Thomas-Chollier et al., 2011) 692 cis-regulatory elements were obtained from AGRIS (Davuluri et al., 2003), PLACE (Higo et al., 1999) and Athamap.(Steffens et al., 2004) 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.(Thomas-Chollier et al., 2011)

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 (Altschul et al., 1990) 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.(Baxter et al., 2012) 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.(Quinlan and Hall, 2010) 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.(Baxter et al., 2012) 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.(Zhang et al., 2012b) The histone modification datasets (H3K4me3, H3K4me2, H3K9ac) were downloaded from the SRA database, GEO accession number GSE28398.(Luo et al., 2013) 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.(Kumar et al., 2013)

ChIP-Seq *in vivo* targets

For the ChIP-Seq datasets (PHYTOCHROME INTERACTING FACTOR 4 [PIF4](Oh et al., 2012), PHYTOCHROME INTERACTING FACTOR 5 [PIF5](Hornitschek et al., 2012), APETALA1 [AP1](Kaufmann et al., 2010b), APETALA2 [AP2](Yant et al., 2010), FLOWERING LOCUS C [FLC](Deng et al., 2011), FAR-RED ELONGATED HYPOCOTYLS 3 [FHY3](Ouyang et al., 2011), PSEUDO RESPONSE REGULATOR 5 [PRR5](Nakamichi et al., 2012), APETALA3 [AP3](Wuest et al., 2012), PISTILLATA [PI](Wuest et al., 2012) and PHYTOCHROME INTERACTING FACTOR 3 [PIF3](Zhang et al., 2013)), 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(Li and Durbin, 2009)). Reads that could not be assigned to a unique position in the genome were removed using samtools (v0.1.18(Li et al., 2009)) 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; (Zhang et al., 2008)). 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)(Yu et al., 2011), SUPPRESSOR OF OVEREXPRESSION OF CO 1 (SOC1)(Tao et al., 2012), AGAMOUS-LIKE 15 (AGL15)(Zheng et al., 2009), LEAFY (LFY)(Winter et al., 2011) and FUSCA 3 (FUS3)(Wang and Perry, 2013), 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. Peak Calling was performed with rMAT.(Droit et al., 2010) 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.(Van Bel et al., 2012) 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.(Thomas-Chollier et al., 2011) The p-value for motif enrichment in the peak set compared with the genomic background was calculated by mapping the motifs using matrix-scan (Turatsinze et al., 2008) (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 ≤ 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 (Seo et al., 2009)).

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.(Marbach et al., 2012b) 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 (Ashburner et al., 2000), Mapman (Thimm et al., 2004), functional modules (Heyndrickx and Vandepoele, 2012) and two expression datasets, a stress expression compendia (336 microarray experiments) and a developmental expression compendia (135 microarray experiments).(De Bodt et al., 2010)

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.(Marbach et al., 2012b) 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.(De Bodt et al., 2010) 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 4

A Collection of Conserved Non-Coding Sequences to Study Gene Regulation in Flowering Plants[a]

Abstract

Transcription factors (TFs) regulate gene expression by binding cis-regulatory elements, of which the identification remains an on-going challenge owing to the prevalence of large numbers of non-functional TF binding sites. Powerful comparative genomics methods, such as phylogenetic footprinting, can be used for the detection of conserved non-coding sequences (CNSs), which are functionally constrained and can greatly help in reducing the number of false-positive elements. In this study, we applied a phylogenetic footprinting approach for the identification of CNSs in ten dicot plants, yielding 1,032,291 CNSs associated with 243,187 genes. To annotate CNSs with TFBSs, we made use of binding site information of 642 TFs originating from 35 TF families in Arabidopsis. In three species, the identified CNSs were evaluated using TF chromatin immunoprecipitation sequencing (ChIP-Seq) data resulting in significant overlap for the majority of datasets. To identify ultra-conserved CNSs, we included genomes of additional plant families and identified 715 binding sites for 501 genes conserved in dicots, monocots, mosses and green algae. Additionally, we found that genes part of conserved mini-regulons have a higher coherence in their expression profile than other divergent gene pairs. All identified CNSs were integrated in the PLAZA 3.0 Dicots comparative genomics platform (`http://bioinformatics.psb.ugent.be/plaza/versions/plaza_v3_dicots/`) together with new functionalities facilitating the exploration of conserved cis-regulatory elements and their associated genes. The availability of this dataset in a user-friendly platform enables the exploration of functional non-coding DNA to study gene regulation in a variety of plant species, including crops.

[a]This chapter is based on Van de Velde et al. (2016) J.V.d.V. and K.V. designed the research methodology; J.V.d.V. and M.V.B. performed data cleaning and analysis; M.V.B. designed the Web site; D.V. created the RNA-Seq expression compendium; J.V.d.V., M.V.B., and K.V. wrote the article. A complete reference can be found in Chapter 8.

4.1 Introduction

DNA sequences that are not actively transcribed and that are conserved across a large number of related species are called conserved non-coding sequences (CNSs). These regions are assumed to have biological relevance because non-functional sequences change at a higher rate during evolution compared to functional sequences (Tagle et al., 1988). The detection of CNSs in plants remains an on-going challenge, because established methods applied in animals or fungi are not always compatible with the properties of plant genomes. The large phylogenetic distance between the currently sequenced dicot plant species hampers the use of lift overs, in which detected transcription factor binding sites (TFBSs) are transferred from one species to another through whole-genome alignments. The potential for this transfer is further decreased by the frequent occurrence of whole-genome duplications and genomic rearrangements in the genomes of flowering plants. Despite these challenges, within the Brassicaceae clade a set of CNSs was successfully identified between closely related species (Haudry et al., 2013). Avoiding the step of whole-genome alignments and replacing it with a multiple pairwise alignment approach has proven to be a useful alternative method to detect CNSs in distantly related plants (Van de Velde et al., 2014). Various software tools have also been developed to identify regulatory regions without using sequence alignments but based on experimental features, such as co-regulation, or using advanced computational methods (MacIsaac and Fraenkel, 2006). Although the naïve mapping of known or de novo found binding sites to promoter regions is frequently used to explore cis-regulatory elements, this approach yields many false positives because TFBSs are often short and typically contain some level of degeneracy in the binding motif (Tompa et al., 2005). The combination of alignment-free binding site detection combined with phylogenetic conservation of these regions has shown great promise, because the application of these methods show significant overlap with experimental TFBSs (Van de Velde et al., 2014).

In Brassicaceae, CNSs have been shown to be under a selective pressure that is comparable to that of protein-coding sequences (Haudry et al., 2013). CNSs are also enriched for regions of open chromatin in Arabidopsis and provide as such a global perspective on possible protein binding to the genome (Van de Velde et al., 2014). In both of the above-mentioned studies, it has also been shown that CNSs greatly overlap with TF ChIP-Seq bound regions. This is an important observation, because TFs play an important role in translating the genotype of plants into their respective phenotype by controlling the spatio-temporal expression of target genes through (combinatorial) binding on TFBSs. A direct application of this feature is the mapping of gene regulatory networks (GRNs) starting from CNSs (Kheradpour et al., 2007; Van de Velde et al., 2014). A GRN is a set of interactions between a TF and target genes and gives a global overview of how transcriptional control is steered in the cell through the activity of TFs.

In the past, much of the CNS research has been focused on Arabidopsis (Kaplinsky et al., 2002; Vandepoele et al., 2006; Thomas et al., 2007; Vandepoele et al., 2009; Baxter et al., 2012; Haudry et al., 2013; Van de Velde et al., 2014) and grasses (Guo and Moose, 2003; Inada et al., 2003; Turco et al., 2013; De Witte et al., 2015) with the exception of the analysis of Baxter et al. (2012), where also footprints were obtained for grape and poplar. Given the limited and biased set of species with available CNSs, there is a great need for CNS detection in other plant species, because these CNSs offer a practical means to enhance the construction of GRNs in crops starting from well-studied model species. An exponent of these CNSs is called ultra-conserved sequences, which are typically long stretches of sequences that are conserved across very large phylogenetic distances. In vertebrates, they are defined as regions that are at least 100 bp long and share 100% sequence identity (Stephen et al., 2008). A pioneering study in plant CNS research suggests that CNSs in grasses (plants) are smaller and far less frequent than those identified in mammalian genes (Kaplinsky et al., 2002). A recent attempt at identifying very deeply conserved CNSs reported that sequences conserved throughout the Eudicot clade of flowering plants could be detected (Burgess and Freeling, 2014). The authors discovered that, based on ten species, a subset of 37 CNSs could be found in all flowering plants. The detected CNSs were functionally similar to vertebrate CNSs, being highly associated with TF-encoding and developmental genes and also enriched in TFBSs (Burgess and Freeling, 2014).

We recently developed a phylogenetic footprinting approach to identify CNSs in Arabidopsis through the comparison with multiple dicot genomes. Comparator species were selected based on the presence

of saturated substitution patterns, which means that non-coding regions that are not under functional constraint will have undergone, on average, one or more mutations. A combination of an alignment-based and a non-alignment-based approach was used to delineate CNSs. The alignment-based approach is best summarized as a multiple local alignment strategy, because local pairwise alignments are first identified and subsequently aggregated on the Arabidopsis reference genome in order to obtain multi-species footprints. The non-alignment-based approach, called Comparative Motif Mapping (CMM), requires a candidate motif (e.g. a TFBS represented as a consensus sequence or position count matrix) as input and assesses the motif conservation in the 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 other species, allowing for incomplete motif conservation (Van de Velde et al., 2014). Here, we applied this methodology to ten dicot genomes and validated the functional importance of these regions by comparing them with experimentally determined TFBSs. We also show that a subset of these CNSs is very deeply conserved in the green plant lineage and can be applied to gain information about the function of TFs through functional enrichment of their predicted target genes.

4.2 Results

Identification of CNSs in Ten Dicot Plant Genomes

A phylogenetic footprinting method that uses an alignment- and non-alignment-based approach was used to detect CNSs in ten dicot species representative for eight plant families (Table 4.1). For each query species, a set of comparator species was selected based on saturated substitution patterns in orthologous gene pairs (Supplemental Table S1[a]).As the PLAZA 3.0 platform was released in the time between the research analyses performed in Chapter 3 and Chapter 4. We wanted to make use of the new and in some cases more qualitative genome assemblies present in PLAZA 3.0 compared to the ones present in PLAZA 2.5. Therefore two metrics were devised to compare the different genomes in the aspects most relevant to the research question. The first metric was assembly completeness assessed by the percentage of truncated genes in the assembly of a genome. The second metric was the completeness of the coding gene space assessed through the presence of a representative gene in each PLAZA gene family. Based on these two metrics the most high quality set of comparator species, that also satisfied the most important requirement of saturated substitution rates for non-coding DNA regions, was selected. (Materials and Methods). Each query species was compared with a set of 13 comparator species including *Arabidopsis thaliana* (Arabidopsis), *Carica papaya* (papaya), *Theobroma cacao* (cocoa tree), *Eucalyptus grandis* (rose gum), *Prunus persica* (peach), *Cucumis melo* (melon), *Glycine max* (soybean), *Populus thrichocarpa* (poplar), *Vitis vinifera* (grape), *Solanum lycopersicum* (tomato), *Beta vulgaris* (beet), *Oryza sativa* (rice) and *Amborella trichopoda* (Amborella). Three different genomic sequence types were defined to identify CNSs (2 kb upstream, 1 kb downstream, and intron). In this analysis, upstream and downstream are used relative to the translation start site and translation stop site, respectively. This is done because it has previously been shown that regulatory elements can be found in the 5' and 3' untranslated region (UTR) (Chaboute et al., 2002; Liu et al., 2010; Wang and Xu, 2010). The second reason to include UTRs is that not all genes have information about their UTR available. Gene orthology information was retrieved with the PLAZA 3.0 integrative orthology method (Van Bel et al., 2012; Proost et al., 2015), 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-many gene relationships (Materials and Methods).

For the detection of CNSs, a multi-species alignment-based approach was applied using the Sigma aligner (Siddharthan, 2006). The CMM approach was used with an enlarged set of 1,211 input sequence motifs and positional weight matrices for 35 TF families (Supplemental Table S2[a] and Materials and Methods). The results of all footprinting analyses are reported in Table I. In total 1,032,291 CNSs were detected for 243,187 genes (Supplemental Dataset S1[b]). To determine whether any of the identified CNSs represent unannotated coding features, we performed a sequence similarity search of all CNSs against a

[a] http://www.plantphysiol.org/content/suppl/2016/06/03/pp.16.00821.DC1/PP2016-00821D_Supplemental_Material.xlsx
[b] http://www.plantphysiol.org/content/suppl/2016/06/03/pp.16.00821.DC1/PP2016-00821D_Supplemental_Dataset_S1.gz

Table 4.1: **Overview of general statistics for the detected CNSs in all investigated plants.**

Species	Plant Family	No. of Genes	Genome Size	No. of CNSs	Coverage	No. of Genes with CNSs	No. of CNSs per Gene	Percentage of Coding CNS	Percentage of CNSs within 500 bp	Median Length of CNS
			Mb		Mb					bp
Arabidopsis thaliana (Arabidopsis)	Brassicaceae	33,602	120	74,381	1.0	19,474	3.82	0.09	62.00	11
Brassica rapa (field mustard)	Brassicaceae	40,998	284	92,578	1.3	29,277	3.16	0.34	60.17	11
Eucalyptus grandis (rose gum)	Myrtaceae	36,493	691	86,434	1.6	23,350	3.70	0.62	50.58	13
Prunus persica (peach)	Rosaceae	27,864	227	109,381	2.3	21,020	5.20	0.66	55.80	15
Cucumis melo (melon)	Cucurbitaceae	28,812	375	63,803	1.2	16,144	3.95	0.45	55.90	14
Glycine max (soybean)	Fabaceae	54,302	974	213,799	3.8	43,198	4.95	0.40	53.25	12
Populus trichocarpa (poplar)	Salicaceae	41,479	417	157,567	3.5	30,662	5.14	0.53	56.68	15
Vitis vinifera (grape)	Vitaceae	26,644	486	105,137	2.1	18,916	5.56	0.56	55.86	13
Solanum lycopersicum (tomato)	Solanaceae	34,859	824	63,428	1.2	19,721	3.22	0.88	53.57	13
Solanum tuberosum (potato)	Solanaceae	35,130	706	65,783	1.2	21,425	3.07	0.71	53.41	13

large set of known plant proteins (Materials and Methods). Across all species only 5,223 CNSs, which corresponds to less than 1% of the total discovered set of CNSs, showed a significant BLASTX hit. These false-positive CNSs were discarded for downstream analysis. The largest number of CNSs was found in soybean (213,799), which has also the largest number of genes with a CNS. The smallest numbers of CNSs were found for tomato (63,428), potato (65,783) and melon (63,803). Both soybean and poplar cover over 3 Mb in CNSs, which is 3.5 to 4 times as much as Arabidopsis, that has the smallest CNS sequence space (1 Mb). The mean number of CNSs per gene varies between 3.07 for potato and 5.56 for grape. The number of CNSs shows a strong correlation with the number of genes in the genome (r2 = 0.74), which is higher than the correlation with the genome size (r2 = 0.33). Because CNSs are detected per gene, this correlation is to be expected. The median length of CNSs per species varies between 11 and 15 bp (Table I). Whereas Arabidopsis and beet have the smallest median CNS length, peach and poplar have the largest. The median number of conserved orthologous species for each CNS is found between four and five comparator species (Fig. 4.1A), which shows that many CNSs are conserved in more than one comparator species and illustrates the multi-species nature of this approach. An evaluation of the location of the CNS relative to the query gene was also performed, revealing that the majority of CNSs are found on the 5' side of the gene (Fig. 4.1B). There are species such as grape, poplar and peach, which have a high fraction of CNSs that are found on the 3' side compared to the other species analyzed. For poplar, this finding is supported by recently performed ChIP-Seq analysis for four TFs, in which three TFs were found to have 19 to 25% of the binding events occurring downstream of a gene (Liu et al., 2015a). In order to further investigate the positional differences of CNSs between species, a density distribution was made for all CNSs up to 2000 bp upstream of a gene, showing three groups of CNS densities in the first 500 bp upstream (Fig. 4.1C). Arabidopsis and field mustard show a high fraction (>= 60%), rose gum shows only 50%, and all other species are found to have a fraction of CNSs in the first 500 bp between 50% and 60%. There is a strong negative correlation (-0.8) between genome size and the percentage of CNSs found in the first 500 bp, which suggests that promoters of species with larger genomes tend to be more stretched out. This result is in agreement with a comparative analysis performed in grasses, where intergenic region expansions from the small Oropetium thomaeum to the

larger Sorghum bicolor were observed (VanBuren et al., 2015).

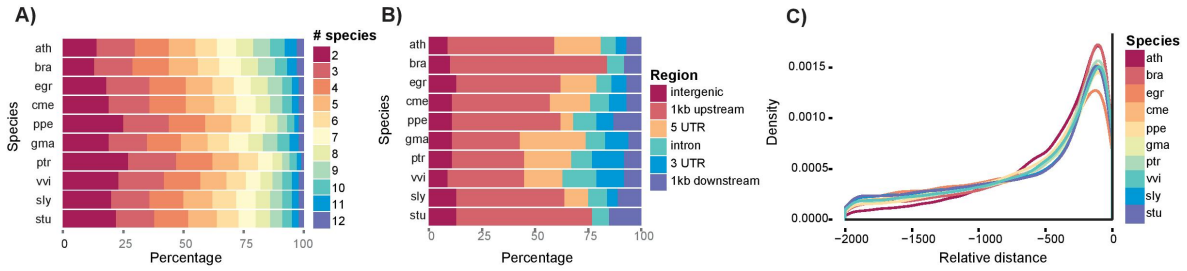

Figure 4.1: **Overview of conserved non-coding sequence properties for all query species.** *A, Overview of significantly conserved footprints in relation to the number of comparator species in which the footprint was conserved. B, Breakdown of CNSs over different structurally annotated genomic regions. C, Density of CNSs across the first 2kb upstream of the translation start site.*

Overlap with TF ChIP-Seq Data

To evaluate the functionality of the identified CNSs and to verify whether these conserved footprints can provide a template to computationally map TF target gene interactions, detailed comparisons of the CNSs were made with publicly available TF ChIP-Seq experiments from tomato, poplar and soybean. The ASR1 TF ChIP-Seq dataset from Ricardi et al. (2014) was used for tomato, a TF ChIP-Seq dataset comprising 2 TFs (NAC and YABBY) was used from Shamimuzzaman and Vodkin (2013) for soybean and two datasets from poplar were used, one containing the ARK1 TF (Liu et al., 2015b) and one containing four TFs (ARK2, PRE, PCN and BLR) (Liu et al., 2015a). The number of overlapping TF ChIP-Seq peaks for each set of CNSs of the corresponding species was determined with the requirement that a CNS had to completely overlap with a TF ChIP-Seq bound region. The overlap of CNSs with the respective TF ChIP-Seq bound regions is shown in Figure 4.2. In poplar, both ARK datasets show a high recovery (62%-64%) of ChIP peaks, opposed to the recovery of PRE, which is rather low (11%). The recovery of the ASR1 dataset is also very low (4%). Certain datasets have a very low (ASR1 and PRE) or high (ARK1 and YABBY) number of bound regions, compared with results from a recent overview study of TF ChIP-Seq analyses in Arabidopsis (Heyndrickx et al., 2014), which might have an influence on the results of the overlap analysis. Additionally, instead of determining the overlapping true positive instances, we also estimated false positives by reshuffling the TF ChIP-Seq genomic locations 1000 times across the genome and determining the overlap with CNSs detected for each species. The estimated number of false positives was used to determine the enrichment for known TF ChIP-Seq bound regions (observed number of elements over expected number of elements, Materials and 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 bound regions, contains only true negatives. However, the shuffled dataset can be used as a proxy to estimate the specificity. Although the recovery rate for individual TFs varied greatly, the enrichment analysis showed that for six out of eight TFs, the number of overlapping peaks was significantly higher compared with those expected by chance ($p<0.001$, Table 4.2).

Quantifying the Evolutionary Conservation of TF Target Gene Interactions

In order to obtain an overview of the evolutionary conservation of TF target gene interactions in the green plant lineage (Viridiplantae), the deep conservation of TFBSs was evaluated. Therefore, the CMM approach was repeated for Arabidopsis, but with a larger number of comparator species (*Physcomitrella patens*, *Chlamydomonas reinhardtii* and *Ostreococcus lucimarinus*) sampling a larger number of plant families (Funariaceae, Chlamydomonadaceae, Bathycoccaceae; Supplemental Dataset S2[a]). This allowed the predicted target genes for each TFBS to be stratified into five phylogenetic clades. The analysis was performed using dicot species as a reference, so the first level of conservation was only within the dicots. If a TFBS was also conserved in Oryza sativa, it was labeled angiosperms, indicating that

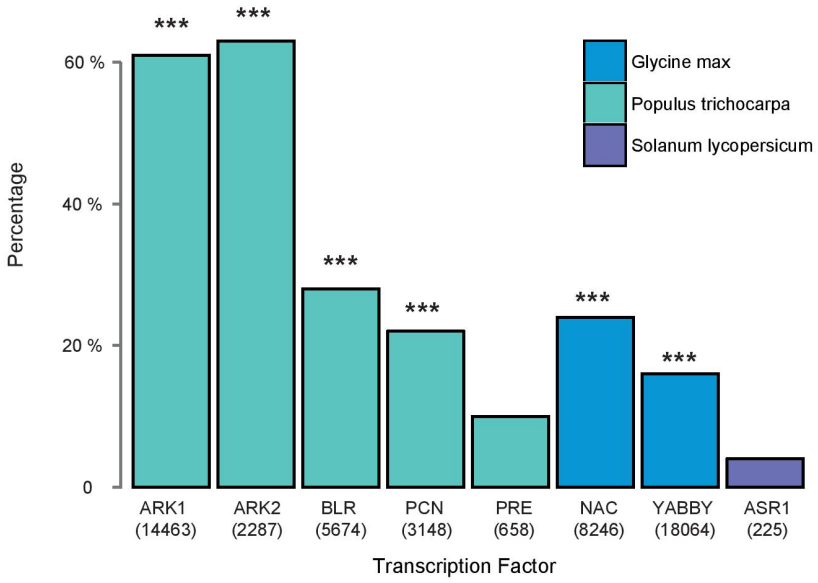

Figure 4.2: **Recovery of TF ChIP-Seq bound regions using CNSs.** *Percentage of TF ChIP-Seq bound regions overlapping with a CNS for each individual TF. *** indicates that the p-value of the enrichment was smaller than 0.001.*

Table 4.2: **Overview of the TF ChIP-Seq overlap analysis including the fold enrichment and p-value.**

Species	ChIP-Seq Data	No. of Peaks	Observed Overlap	Expected Overlap	Enrichment Fold	P
Poplar	ARK1	14,463	8,833	2,286	3.742	0.001
Poplar	ARK2	2,287	1,448	364	3.864	0.001
Poplar	BLR	5,674	1,564	593	2.564	0.001
Poplar	PCN	3,148	705	290	2.362	0.001
Poplar	PRE	658	67	61	1.063	0.243
Soybean	NAC	8,246	1,970	950	2.012	0.001
Soybean	YABBY	18,064	2,913	1,607	1.752	0.001
Tomato	ARS1	225	8	5	1.600	0.114

the binding site was conserved in dicots and monocots. The label Magnoliophyta was given to interactions that were conserved in the flowering plants comprising dicot, monocot and Amborella. The last two clades were Embryophyta and Viridiplantae, if the interactions were conserved in Physcomitrella patens and Chlamydomonas reinhardtii or Ostreococcus lucimarinus, respectively. There are 10,976 genes with at least one conserved element in dicots (44% of genes with conserved orthologs in dicots), 5,788 genes for the angiosperms clade (26%), 2,917 for the Magnoliophyta clade (13%), 1,568 genes for the Embryophyta clade (8%) and 501 genes for the Viridiplantae clade (4%). As expected, these 501 genes cover Gene Ontology (GO) terms related to basal functions such as transport, carbohydrate metabolism and cell cycle. For all above counts, full conservation in all clades was required, but not in all species of that clade. The median number of species in which a binding site was conserved ranged from 87% for Viridiplantae to 40% for dicots (Supplemental Table S3[a]). These numbers illustrate that the early diverging clades have a higher level of species conservation across all species than the younger clades. This could indicate that these evolutionary deeply conserved binding sites are playing a regulatory role in essential biological processes whereas less deeply conserved binding sites are more involved

in clade-specific developmental or responsive processes. In order to further study the evolutionary conservation of regulatory interactions, the evolutionary depth for which an ortholog of the TF that is linked to the conserved binding site could reliably be detected was taken into account. Adding this additional criterion greatly reduced the number of genes for which a conserved interaction could be detected in the distant clades (Supplemental Table S3[a]). The Embryophyta clade contains 334 genes after filtering (365 interactions) and in the Viridiplantae clade only 8 genes remain (10 interactions). These results illustrate that reliably detecting orthologs over very large evolutionary distances is inherently difficult. An example of a regulatory interaction that is conserved in Viridiplantae is the interaction between the E2Fa TF and POL2A, a DNA polymerase epsilon catalytic subunit. This interaction is shown in Figure 4.3 together with other regulatory interactions for E2F TFs that are conserved in angiosperms, Embryophyta or Viridiplantae. To illustrate the validity of these predictions we compared the conserved TF target genes to tandem chromatin affinity purification (TchAP) bound target genes of E2Fa (Verkest et al., 2014) and differentially expressed genes upon overexpression of E2Fa (Naouar et al., 2009). This comparison revealed that in total 108 out of 119 predicted target genes for E2Fa are supported by experimental evidence (82 are bound and regulated, 100 are bound and 90 are regulated). We also integrated the predicted target genes with a set of genes that were deemed to be involved in cell cycle because they display peak expression during specific stages of the cell cycle (Menges et al., 2003). Six of the predicted target genes display this cell cycle dependent expression pattern. Although the majority of these predicted deeply conserved target genes are known to be involved in cell cycle related processes, several genes lack detailed functional annotation. AT4G33870, AT4G23860, AT1G77620 AT3G48540, AT1G61000 and AT3G27640 are all predicted deeply conserved target genes that are also supported by experimental evidence. These genes however lack information about the specific biological processes they are involved, except for AT4G23860 and AT3G27640, which have been assigned to a functional module involved in DNA-dependent DNA replication (Heyndrickx and Vandepoele, 2012). Both the conserved E2F binding sites and the integrated experimental datasets strongly suggest that these genes play an important role in cell cycle related processes.

Obtaining functional annotation through GO enrichment of conserved target genes.

Apart from focusing on deeply conserved CNSs, the large number of binding sites conserved in Dicots can also be used to functionally characterize individual TFs, through GO enrichment of the associated conserved target genes (Materials and Methods). Known functions from literature were used to evaluate if the enriched GO terms were correct. MYB58 and MYB63 activate lignin biosynthesis in fibers and vessels (Zhou et al., 2009) and for both genes, we found the GO term lignin biosynthetic process to be enriched in target gene sets conserved in the dicot clade. Also, both MYB46 and MYB83 are known to be involved in secondary cell wall processes (Zhong and Ye, 2012; Kim et al., 2013). Many enriched GO terms of the target genes of these TFs were related to the regulation of lignin biosynthesis and to cellulose and xylan biosynthetic processes (Supplemental Table S4[a]). MYB84 is part of the set of three regulators of axillary meristem (RAX) genes that are partially redundant regulators of axillary meristem formation (Muller et al., 2006). In the set of target genes, we observed the axillary shoot meristem initiation and meristem maintenance GO terms, confirming this function (Supplemental Table S4[a]). The TF MYB3 represses phenylpropanoid biosynthetic gene expression (Dubos et al., 2008) and we indeed recovered the enriched GO term regulation of phenylpropanoid metabolic process for this gene set. A direct predicted target of MYB3 is MYB4, which together with MYB32 can influence pollen development by changing the flux along the phenylpropanoid biosynthetic pathways, affecting the composition of the pollen wall (Preston et al., 2004). Both TFs showed enrichment toward the GO term regulation of phenylpropanoid metabolic process. MYB4 has also been shown to be involved in the production of UV-protecting sunscreens in Arabidopsis in response to light stress (Jin et al., 2000). The GO term anthocyanin accumulation in tissues in response to UV light was representative for this proposed function. (Supplemental Table S4[a]). The enriched GO terms positive regulation of flavonoid biosynthetic process and flavonol biosynthetic process for MYB111 were a validation of its role in the biosynthesis of flavonol (Stracke et al., 2007). Prevalent throughout the whole GO enrichment table for these MYB TFs were GO terms related to flavonoid biosynthetic processes or related to precursors of

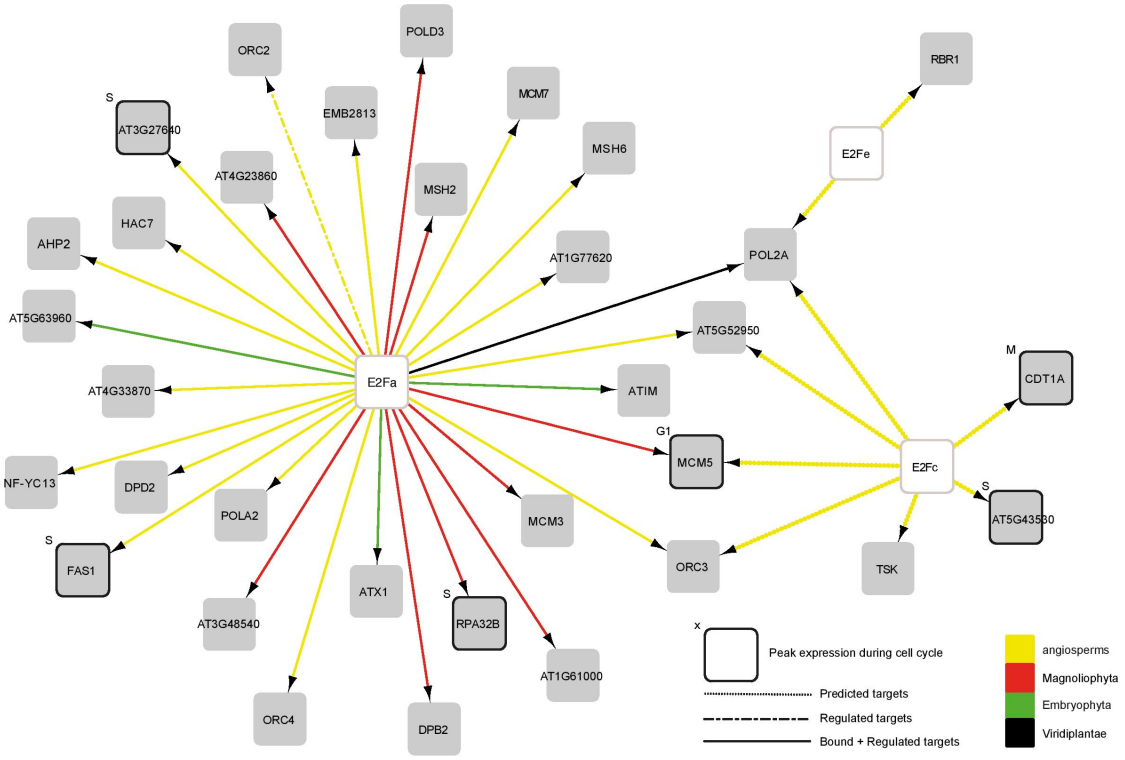

Figure 4.3: **A gene regulatory network of predicted conserved target genes for E2Fa, E2Fc and E2Fe.** *All interactions that are conserved in angiosperms, Embryophyta or Viridiplantae are shown. Interactions conserved up to angiosperms are shown in yellow, up to Magnoliophyta in red, up to Embryophyta in green line and conserved up to Viridiplantae in black. Experimental evidence is indicated by the edge type: a solid line indicates that an interaction is supported by both TchAP-seq binding and differential expression upon TF perturbation, a striped and dotted line indicates that the target gene was only differentially expressed and a dotted line indicate that a prediction was not supported by experimental evidence or that no experimental evidence was available.*

flavonoids. This finding suggests a link between MYB TFs and their role in stress response, which is supported by previous research that has shown that flavonoid biosynthesis is up-regulated in response to a wide range of abiotic stresses, such as cold, salinity and drought (Supplemental Table S4[a]) (Ma et al., 2014).

Discovery and exploration of conserved mini-regulons

Gene rearrangements occur frequently during the evolution of prokaryotic and eukaryotic genomes. The number of rearrangements appears to be a function of the phylogenetic distance between the organisms being studied. Little collinearity in gene order has been observed between Arabidopsis and other dicot genomes due to the large evolutionary distance that separates them (Supplemental Fig. A.1). All previous analyses have focused on linking a conserved binding site of a TF to a target gene. In this paragraph we explore whether we can detect more complex transcriptional units, focusing on divergent gene pairs. Divergent gene pairs are sets of genes that have their TSS directed towards each other and as such have parts of their upstream region shared. These pairs of genes have shown to have a higher correlation in expression than random gene pairs (Krom and Ramakrishna, 2008). First, 6501 divergent gene pairs were identified in the genome of Arabidopsis thaliana. Out of this total set of divergent gene pairs 576 also had a shared conserved cis-regulatory element that was identified for each gene independently. We also checked whether the divergent orientation of these gene pairs was conserved in orthologous gene pairs across other genomes. There were 2238 gene pairs that had their orientation conserved in orthologous gene pairs in one or more other genomes, and 174/2238 gene pairs also had a shared conserved binding site conserved across these orthologous gene pairs. An example of a deep conserved gene pair with a

shared conserved binding site is TOM5 (AT5G08040) and DUF1118 (AT5G08050) that have a conserved PIF1 binding site and conserved orientation in orthologous gene pairs of six other genomes, including rice. In a next step the correlation in expression profile of gene pairs part of these different categories of divergent gene pairs was evaluated using Pearson correlation coefficients (PCCs) based on a RNA-seq expression compendium (Supplemental Table S5[a] and Materials and Methods). A comparison of absolute PCCs for each of the four categories is shown in Figure 4.4A and shows an increase in correlation between gene pairs when a conserved binding site is present. The difference is maximal when both binding site and divergent orientation are conserved across multiple genomes. A significant difference was observed between gene pairs with a conserved binding site and conserved orientation compared to basic divergent gene pairs and divergent gene pairs with conserved orientation. This finding hints towards the existence of conserved mini-regulons where the presence of a conserved regulatory element results in increased co-expression of flanking genes, suggesting tight co-regulation. In order to analyze these 174 mini-regulons in more detail PCCs were also calculated for the TF that is linked to the conserved binding site (Supplemental Table S6[a]). In 12 out of these 174 cases there were also striking similarities in the gene expression profiles of the TF and both gene pairs (Materials and Methods). One example is shown for a conserved BES1 binding site between YLMG2 (AT5G21920) and PAA2 (AT5G21930) conserved in three genomes (cocoa tree, rose gum and grape) and with strong positive PCCs between the genes and TF itself (PCC> 0,70, Fig. 4.4B). A second example is shown for a conserved PIF5 binding site between ANS (AT4G22880) and PGR5-LIKE A (AT4G22890) conserved in four other genomes (papaya, eucalyptus, poplar, and grape) also with a strong positive PCC between the flanking genes and the TF.

Exploration and Visualization of Plant CNSs through the PLAZA 3.0 Dicots Platform

The CNSs detected for all ten dicot query species were uploaded to the PLAZA 3.0 Dicots database and a number of new features were added to facilitate their exploration. On each gene page, a link was added to the toolbox to explore the conserved binding sites (CNSs overlaid with all TFBSs used in this study) for that gene. On this page, a complete overview per investigated region, upstream, downstream or intron is given for all retrieved binding sites per gene. Complementary, conserved binding sites are also visualized using the GenomeView genome browser for all ten species (Abeel et al., 2012). On all Arabidopsis TF-encoding gene pages that have TFBS information, a tab was added containing the associated binding sites for that TF. Besides additions to the gene pages, a binding site page was also created for all motifs and position weight matrices used in this study. On these pages, a common name, description and sequence logo are provided for each binding site, together with the total number of genes associated with this binding (Supplemental Fig. B.1A). Breakdowns of the number of target genes per species and per investigated region are depicted as pie charts (Supplemental Fig. B.1B). Different functionalities are provided in the toolbox section on the binding site page: there is the possibility to explore the associated gene families, as well as GO, MapMan and InterPro functional annotations, based on the conserved target genes. The toolbox also contains the opportunity to look for binding sites with a similar binding profile (Supplemental Fig. B.1C).

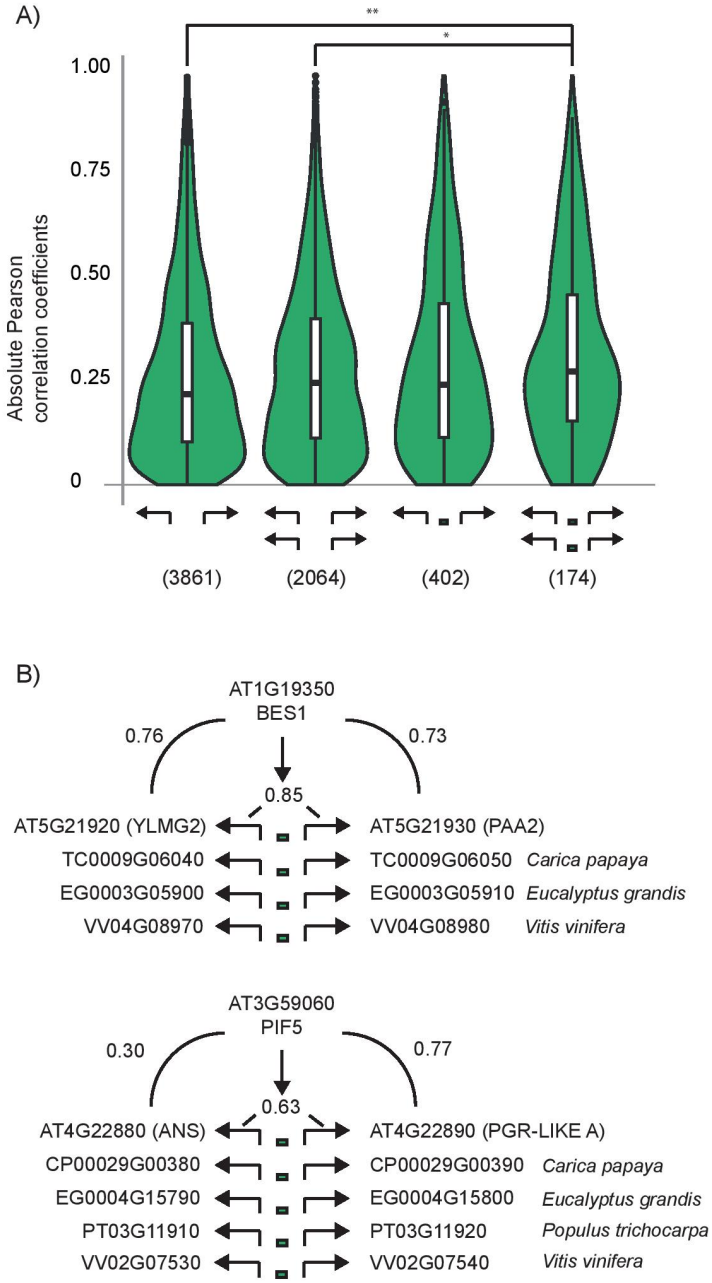

Figure 4.4: **Transcriptional coherence of divergent gene pairs with and without conserved binding sites.** *A, The distribution of absolute Pearson correlation coefficients for all divergent gene pairs is split into four categories. These categories are divergent gene pair, divergent gene pair with orientation conservation, divergent gene pair with conserved binding site and divergent gene pair with conserved binding site and orientation conservation. * and ** indicate that the p-value of the Wilcoxon rank sum test was smaller than 0.05 and 0.01, respectively. B, Two examples of conserved mini-regulons. The conserved binding site is indicated in the center together with the Pearson correlation coefficients between the TF and the divergent genes. Below the divergent genes, the orthologous genes with conserved divergent configurations and the presence of a conserved binding site in other species are shown.*

4.3 Discussion

In this study, we applied a phylogenetic footprinting approach to identify CNSs in ten dicot species. This approach uses both alignment-based and alignment-free techniques and combines different gene orthology prediction methods that do not rely on synteny information. In this manner, it circumvents the step of whole-genome alignment, which is difficult owing to the frequent nature of polyploidy and genome rearrangements in plant genomes. As such, our approach is well suited to incorporate more distantly related species including many-to-many gene orthology relationships. A set of high-quality comparator species was selected for each query species, ensuring that a saturated substitution rate in the absence of selection was present. Across all experiments, 1,032,291 CNSs were detected for 243,187 genes. A strong correlation was detected between the number of CNSs and the total number of genes present in a genome, whereas no strong correlation could be detected between the total number of CNSs and the genome size. There is however another manner in which genome size could be correlated with promoter architecture. The fraction of CNSs in the first 500 bp upstream of the translation start site compared with the total number of CNSs was inversely correlated with genome size. This indicates that smaller genomes have their functional elements packed more closely toward the translation start site of the gene compared with larger genomes and as such have smaller promoters. The detected CNSs were compared with TF ChIP-Seq data from poplar, soybean and tomato. We found that CNSs were enriched for TF bound regions compared with randomly selected regions for six out of eight datasets, illustrating the functional regulatory character of these sequences. Furthermore, using a more extended phylogenetic sampling than used in previous research (Burgess and Freeling, 2014), we were able to discover 715 TFBSs for 501 genes that were conserved from dicots to Viridiplantae. Functions associated with this gene set comprise basal biological processes such as transport, carbohydrate metabolism and cell cycle. The fact that these functions are not highly specialized for flowering plants is in concordance with the predicted age of these interactions. When the presence of the orthologous TF in the comparator species was also taken into account we were able to discover deeply conserved interactions that showed strong experimental support for the E2FA TF. Through GO enrichment of the target genes of TFs, we were able to predict putative functional annotations and confirm known functions for different sets of TFs. This process of assigning functions to the predicted target genes proves useful for genes for which little functional information is available. Assessing the functional coherence of target genes is an alternative manner to validate regulatory interactions and is based on the idea that genes that are part of the same biological pathway are regulated by similar sets of TFs (Marbach et al., 2012b; Lindemose et al., 2014).

The idea that conserved binding sites exert a regulatory role on a bigger scale than only on the closest gene is largely unexplored in plants. To obtain possible mechanistic insights from the presence of conserved binding sites on gene regulation, we investigated the effect on co-expression of a conserved binding site located between divergent gene pairs. Previously Krom and Ramakrishna (2008) reported that specific regulatory elements were overrepresented in divergent or convergent gene pairs with a strong correlation in gene expression. In this analysis we were able to show that the presence of a conserved binding site leads to a significant increase in transcriptional coherence compared to divergent gene pairs that did not share a conserved binding site. This effect became stronger when the divergent gene pair was also conserved in the corresponding genomes where the binding site was conserved. This co-occurrence of binding site conservation and divergent orientation conservation was called a mini-regulon. Finally, through a background model of randomly generated mini-regulons, several cases were discovered where the gene expression profile of the TF was strongly correlated with the divergent gene pair linked to the conserved binding site. These conserved mini-regulons represent examples of spatially-conserved transcriptional units encompassing multiple target genes conserved in multiple plant genomes.

In order to get a better understanding of the organization as well as the function of TFs, it is crucial to study GRNs. CNSs have shown to be important stepping stones for generating functionally relevant GRNs based on TFBSs (Kheradpour et al., 2007; Van de Velde et al., 2014). In the past, much of CNS research has focused on Arabidopsis and grasses. With the availability of CNSs for an increasing number of dicot species, it now becomes possible to leverage existing regulatory annotation approaches in non-model species. A widely used approach to elucidate the function of a TF is to perturb the given

TF and compare expression profiles of the wild-type and perturbed state, leading to lists of differentially expressed genes on which de novo motif finding is often performed to obtain new insights on the regulation of these genes. The combination of these motifs with conservation analysis is a powerful approach to identify genome-wide bona fide target genes with these motifs and can help to unravel the underlying regulatory cascade, as was recently shown for leaf development in maize (Yu et al., 2015). Another approach in which CNSs can play a key role is the translation of existing knowledge of GRNs in model species into economically more interesting species, which is however not trivial owing to the occurrence of evolutionary changes. On the gene level, duplication and loss events play an important role. On the binding site level, the movement, as well as the gain and loss of TFBSs can occur. Both of these types of events can lead to the disappearance or the creation of regulatory interactions (Dermitzakis and Clark, 2002). Given these obstacles, CNSs can provide a useful tool for guiding the delineation of GRNs.

The integration of this large dataset in the PLAZA 3.0 Dicots platform opens up opportunities for plant scientists to quickly gain information about putative regulators of a gene of interest. It also allows for downstream analysis, such as functional enrichment of target genes of a TF or the investigation of the associated gene families. The presentation of this CNS dataset in an easy accessible form offers advantages for non-computational scientists to access this data and generate new regulatory hypotheses in a diverse set of plant species.

4.4 Material and Methods

Sequence and Orthology Information

The 18 species used in this study were *Arabidopsis thaliana* (TAIR10) (consoritium, 2000), *Brassica rapa* (FPsc v1.3, DOE-JGI) (Wang et al., 2011), *Carica papaya* (Hawaii Agriculture Research Center) (Ming et al., 2008), *Glycine max* (JGI 1.0) (Schmutz et al., 2010), *Populus trichocarpa* (JGI 2.0) (Tuskan et al., 2006), *Theobroma cacao* (CocoaGen v1.0) (Argout et al., 2011), *Vitis vinifera* (Genoscope v1) (Jaillon et al., 2007), *Eucalyptus grandis* (JGI 1.1) (Myburg et al., 2014), *Cucumis melo* (Melonomics v3.5) (Garcia-Mas et al., 2012), *Prunus persica* (JGI 1.0) (The International Peach Genome Initiative et al., 2013), *Solanum lycopersicum* (ITAG 2.3) (Consortium, 2012), *Solanum tuberosum* (ITAG 001) (Consortium, 2011), *Beta vulgaris* (RefBeet 1.1) (Dohm et al., 2014), *Oryza sativa* (MSU RGAP 7) (Project, 2005), *Amborella trichopoda* (Amborella V1.0) (Project, 2013), *Physcomitrella patens* (JGI 1.6) (Rensing et al., 2008), *Ostreococcus lucimarinus* (JGI 2.0) (Palenik et al., 2007) and *Chlamydomonas reinhardtii* (JGI 5.5) (Merchant et al., 2007), and sequences were obtained from the PLAZA 3.0 database (Proost et al., 2015). Three sequence types, i.e. upstream, downstream, and intronic, were used to identify CNSs. Upstream sequences were restricted to the first 1000 or 2000 bp upstream of the translation start site or to a shorter region if the adjacent upstream gene is located within a distance smaller than 1000 or 2000 bp. The 1000- and 2000 bp upstream sequences were processed as two independent runs. Downstream sequences were restricted to the first 1000 bp downstream of the stop codon or to a shorter region if the adjacent downstream gene was within 1000 bp. The intronic sequence type is defined as the complete gene locus starting from the translation start site with exons masked. Orthologs for each gene were determined in 17 species using the PLAZA 3.0 Integrative Orthology method (Proost et al., 2015). The included orthology detection methods are OrthoMCL (Li et al., 2003), phylogenetic tree-based orthologs, and BHIF (Van Bel et al., 2012; Proost et al., 2015). Two orthology definitions were used. The first definition uses a simple best BLAST hit-derived method that includes in-paralogs, called best-hit and in-paralogous families (BHIF), whereas the second definition, called Integrative Orthology, requires that at least two PLAZA detection methods confirm an orthologous gene relationship.

Species Selection

An average pairwise Ks matrix was created with the PLAZA 3.0 platform, by calculating the Ks between all one-to-one collinear homologs of each species combination. Ks is defined as the number of synonymous substitutions per synonymous site. This was done to confirm that all included species have saturated substitution patterns (mean Ks > 1) when comparing orthologous gene pairs with one another (Proost et al., 2015). Saturated substitution patterns indicate that in the absence of selection, the average

position in a DNA sequence stretch has undergone at least one substitution. To detect CNSs in potato, tomato was removed as a comparator species and Arabidopsis was removed when Brassica rapa (field mustard) was analyzed. This was done because substitution rates are not saturated between the genomes of these two combinations of species. To make a more informed decision of which comparator species to include, two other metrics were calculated via the PLAZA 3.0 platform. The first metric was the percentage of protein-coding genes that were not complete (truncated) in the genome assembly. This percentage was assessed by counting for all gene families which genes were removed from the multiple sequence alignment used to generate the phylogenetic tree for each gene family (Proost et al., 2009). The second metric was the percentage of gene families for which a given species did not have a representative gene.

Detection of CNSs using Comparative Motif Mapping and Alignment-Based Phylogenetic Footprinting

The comparative motif mapping algorithm was performed as described in Van de Velde et al. (2014). Known binding sites were mapped on the regions covered by the three sequence types for all included species using DNA-pattern allowing no mismatches (Thomas-Chollier et al., 2011). A total of 690 cis-regulatory elements were obtained from AGRIS (Yilmaz et al., 2011), PLACE (Higo et al., 1999), and Athamap (Steffens et al., 2004). In addition, 44 positional count matrices were obtained from Athamap, and for 15 TFs, positional count matrices were obtained from ChIP-Seq data (Heyndrickx et al., 2014). Finally, 108 and 623 positional weight matrices were obtained from protein-binding microarray studies performed by (Franco-Zorrilla et al., 2014) and (Weirauch et al., 2014), respectively. Positional count matrices were mapped genome-wide using MatrixScan using a p-value cutoff <1e-05 (Thomas-Chollier et al., 2011). The alignment based approach was performed as described in Van de Velde et al. (2014), except that only the Sigma alignment tool (Siddharthan, 2006) was run, with the -x parameter set to 0.5. Pairwise alignments were generated between all query genes and their orthologous genes for all three sequence types. All performed experiments were filtered to retain only regions with a p-value that corresponds to a FDR $\leq 10\%$.

Overlap of CNSs with Benchmarks

TF ChIP-Seq binding location datasets were obtained from the supplementary tables of the respective papers (Shamimuzzaman and Vodkin, 2013; Ricardi et al., 2014; Liu et al., 2015b,a) for all TFs. The benchmark dataset was formatted as a BED file and the overlap was determined using the BEDTools function intersectBed with the -u parameter and the -f parameter set to 1 (Quinlan and Hall, 2010). This means that a TF bound region was considered correctly identified if a CNS was completely overlapping with it. False positives were determined by shuffling the TF bound dataset 1000 times using shuffleBed. The overlap with CNSs was determined for each shuffled file and the median number of recovered elements over 1000 shuffled files was used as a measure for the expected number of overlapping regions. This estimation was used to calculate the fold enrichment, defined as the ratio between observed overlap and expected overlap by chance. RepeatMasker (Smit et al., 2013) was run with default parameters on all three genomes for which TF ChIP-seq data was available and all identified repeat regions were excluded from the sequence space to shuffle the TF bound regions.

Deep Conservation and GO Enrichment

All TFs were categorized according to the TF families described in PlantTFDB 3.0 (Jin et al., 2014). The phylogenetic quantification of TF target genes in their respective TF families was performed based on these TF family annotations. GO annotations for Arabidopsis were obtained from the PLAZA 3.0 database (Proost et al., 2015). Per TF and per phylogenetic group, the enrichment of conserved target genes toward GO annotations (hypergeometric distribution + Bonferroni correction) was determined. The enriched GO terms were made non-redundant by removing enriched parental GO terms, considering the structure of the GO graph. For the gene-GO network, enriched GO terms needed to be supported by at least five target genes. Network visualizations were generated using Cytoscape 3 (Shannon et al., 2003).

Protein-Coding Potential of CNSs

The coding potential of a CNS was determined using BLASTX (Altschul et al., 1990) against the PLAZA 3.0 protein database 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 (Baxter et al., 2012). We then performed the same BLASTX search on the real sequences, using the minimum e-value from the random set (e-value < 0.001) as the cutoff for a significant hit.

RNA-seq compendium

The RNA-Seq expression compendium was built with public datasets from NCBI's Sequence Read Archive (SRA) (Kodama et al., 2012). The compendium contains gene-level expression values for 40 manually selected samples (Supplemental Table S5[a]) of different treatment and tissue combinations. SRA files for each sequencing run were downloaded from the SRA and converted to the FASTQ format using fastq-dump (v2.4.4) from the SRA toolkit. FASTQ files from runs of the same sample were concatenated. Paired-end reads were unpaired by randomly selecting either the forward or reverse read and processing it as single-end. FastQC (v0.9.1) was used to detect overrepresented adapter sequences, which were subsequently clipped with fastx_clipper from the FASTX toolkit (v0.0.13). Nucleotides with Phred quality scores lower than 20 were trimmed with fastq_quality_trimmer from the FASTX toolkit. Reads shorter than 20 nucleotides after quality trimming were discarded. To obtain raw read counts for each transcript in the TAIR10 annotation (Lamesch et al., 2012), Sailfish (v0.6.3) (Patro et al., 2014) was run with a k-mer length of 20. For genes with multiple transcripts, the raw read counts of its transcripts were summed to get a gene-level read count. Counts were then normalized for the entire compendium with the Variance Stabilizing Transformation (VST) from the DESeq R package (v1.14.0) (Anders and Huber, 2010). VST was chosen since it results in correlation coefficients between genes that are most comparable to those obtained with microarray data (Giorgi et al., 2013).

Detecting transcriptionally coherent mini-regulons

A background model of random mini-regulons was created by first sampling a divergent gene pair from the set of 6183 divergent gene pairs for which gene expression data was available in our RNA-seq compendium and randomly assigning a TF to this gene pair (this procedure was repeated 10,000 times). PCCs were determined using the RNA-seq expression compendium between the divergent gene pair and between the assigned TF for each of the random mini-regulon. The harmonic mean was calculated for the three PCCs of each randomly generated mini-regulon. The top five percent highest scores from the resulting distribution were used as cutoff value (0.47) to identify mini-regulons showing strong TF co-expression.

CHAPTER 5

BLSSpeller: exhaustive comparative discovery of conserved cis-regulatory elements[a]

Abstract

The accurate discovery and annotation of regulatory elements remains a challenging problem. The growing number of sequenced genomes creates new opportunities for comparative approaches to motif discovery. Putative binding sites are then considered to be functional if they are conserved in orthologous promoter sequences of multiple related species. Existing methods for comparative motif discovery usually rely on pregenerated multiple sequence alignments, which are difficult to obtain for more diverged species such as plants. As a consequence, misaligned regulatory elements often remain undetected. We present a novel algorithm that supports both alignment-free and alignment-based motif discovery in the promoter sequences of related species. Putative motifs are exhaustively enumerated as words over the IUPAC alphabet and screened for conservation using the branch length score. Additionally, a confidence score is established in a genome-wide fashion. In order to take advantage of a cloud computing infrastructure, the MapReduce programming model is adopted. The method is applied to four monocotyledon plant species and it is shown that high-scoring motifs are significantly enriched for open chromatin regions in *Oryza sativa* and for transcription factor binding sites inferred through protein-binding microarrays in *Oryza sativa* and *Zea mays*. Furthermore, the method is shown to recover experimentally profiled ga2ox1-like KN1 binding sites in *Zea mays*. BLSSpeller was written in Java. Source code and manual are available at urlhttp://bioinformatics.intec.ugent.be/blsspeller

[a]This chapter is based on De Witte et al. (2015). J.F., K.V., P.A., P.D. and B.D. designed the research methodology, D.D.W., D.D., P.A. and J.F implemented the algorithms and generated BLSspeller output datasets, J.V.D.V and M.V.B generated additional datasets for input (M.V.B.) and evaluation (J.V.D.V), D.D.W, J.V.D.V, and J.F performed functional evaluation of the algorithm, D.D.W., J.V.D.V., K.V. and J.F. wrote the manuscript. A complete reference can be found in Chapter 8.

5.1 Introduction

One of the major challenges in systems biology is gaining a full understanding of gene transcriptional regulation. Transcription factors, for which the binding sites are usually hidden in the promoter sequence of the gene, are in this respect of particular importance. Computational approaches for *de novo* motif discovery can be classified in (a) methods to identify binding sites in promoter sequences of co-regulated or co-bound (from e.g. ChIP assays) genes within a single genome and (b) comparative approaches using homologous sequences from multiple related species (Das and Dai, 2007).

The first category uses clusters of co-expressed genes, which are assumed to be regulated by the same set of transcription factors. A drawback of these methods is that the relationship between co-expression and co-regulation relies on complex regulatory mechanisms, making it difficult to assemble reliable datasets since co-expression does not necessarily imply that there is a common binding site involved. Two different algorithmic approaches coexist: the statistical (Bailey et al., 2009; Hughes et al., 2000; Thijs et al., 2002; Liu et al., 2001; Wei and Yu, 2007) and the exhaustive, word-based algorithms. The latter contain graph-based approaches (Eskin and Pevzner, 2002; VijayaSatya and Mukheqee, 2004; Liang et al., 2004) and methods based on index structures (Marsan and Sagot, 2000; Pavesi et al., 2001; Marschall and Rahmann, 2009). Due to the growing availability of genome sequences, a second category of algorithms based on *phylogenetic footprinting* emerged (Blanchette and Tompa, 2002): orthologous regulatory regions from multiple species are compared with the underlying assumption that functional elements evolve at a much slower pace, compared to the non-functional part of the genome, due to selective pressure (Berezikov et al., 2004). Most comparative motif discovery approaches rely in some way on multiple sequence alignments, in which regulatory signals are expected to be well-aligned. Pioneering algorithms in this category are Conreal (Berezikov et al., 2004), Phylonet (Wang and Stormo, 2005) and Phyloscan (Carmack et al., 2007). More recent algorithms relying on alignments are used to study mosquitoes (Sieglaff et al., 2009), Fusarium (Kumar et al., 2010), vertebrates (Ettwiller et al., 2005) and mammals (Xie et al., 2005). It has, however, been shown that known regulatory elements are not always correctly aligned (Siggia, 2005), an issue that is further complicated by the different alignments produced by various alignment programs (Pollard et al., 2004). Transcription factor (TF) binding sites are short, flexible against certain mutations and even mobile which explains why they are sometimes misaligned. Mechanisms have been observed that allow the modification of regulatory sequences without altering their function: *divergence driving words* and *binding site turnover*. Regulatory sequences can diverge freely if the divergence driving words, which are specific short words in the non-coding DNA, are not altered (Bradley et al., 2010). Since a TF can often bind to multiple similar sites, mutations turning one site into another should not affect regulation. Binding site turnover, on the other hand, is the mechanism where the gain of a redundant binding site allows the loss of a previously functional site (Venkataram and Fay, 2010). The corresponding TF can then bind to the new site, maintaining the regulatory interaction. This allows binding sites to relocate within the regulatory sequence, making it difficult for alignment algorithms to correctly align them.

Binding site discovery, especially in plants, has to deal with large divergence times and complex diversification mechanisms such as genome duplications. This makes approaches based on whole genome alignments, often used in *de novo* algorithms, impractical. Some of these problems have been addressed in earlier studies. Stark et al. (2007) used a mixed approach in a study with 12 Drosophila species, starting from whole genome alignments but allowing for limited motif movement within an alignment. Elemento and Tavazoie (2005) designed an alignment-free algorithm to discover overrepresented k-mers over the exact ACGT alphabet in pairs of related genomes. Finally, MDOS (Wu et al., 2008) is a new version of this algorithm with improved statistics.

In this paper, four monocotyledonous plant species are studied using a phylogenetic footprinting approach: *Oryza sativa ssp. indica* (osa), *Brachypodium distachyon* (bdi), *Sorghum bicolor* (sbi) and *Zea mays* (zma). We adopt a gene-centric approach, where the promoter sequences of orthologous genes are grouped into *gene families*. A word-based discovery algorithm was designed to exhaustively report all *genome-wide conserved* motifs. The term *conserved* relates to the occurrence of the motif in multiple promoter sequences of a particular gene family. *Genome-wide conservation* relates to the fact that this

conservation occurs in more gene families than what is expected by chance. Motifs are modeled as words (k-mers) over an alphabet that contains the 4 bases (ACGT) and (optionally) additional degenerate characters from the IUPAC alphabet (Cornish-Bowden, 1985). This degeneracy allows a motif to model a collection of binding sites. The algorithm can be run in both *alignment-free* or *alignment-based* mode. In case of alignment-free discovery, the conservation of a motif is scored irrespective of its orientation or position within a promoter sequence. This relaxed definition of conservation was previously used by Gordan et al. (2010) and is especially relevant when studying more diverged species for which accurate multiple sequence alignments are difficult to generate. Alignment-based discovery adds the constraint that motifs must be aligned, i.e., occur at the same position in the multiple sequence alignment.

Robust algorithms for comparative genomics are expected to gain in power when more related species are added. Most studies so far only consider motifs that are conserved within *all* organisms. The branch length score (BLS) was developed to quantify motif conservation in a biologically meaningful manner and ranges from 0% (not conserved) to 100% (conserved in all sequences). The BLS takes the phylogenetic relationships between the species into account by representing a relative evolutionary distance over which a candidate binding site is conserved within a gene family. The BLS was first used in a comparative study with 12 Drosophila genomes (Stark et al., 2007) and allows studying motifs only conserved in subsets of the organisms. Whereas most current algorithms avoid exploring the full motif space by using greedy algorithms, our method is unique in the sense that it is exhaustive. MDOS (Wu et al., 2008) only processes promising k-mers and gradually adds degeneracy if this improves the conservation score. Kellis et al. (2003) and Stark et al. (2007) use the mini-motifs approach (Helden et al., 2000) only processing promising trinucleotide duos before adding degeneracy. Here, every word that occurs in one of the input sequences, including their degenerate variants, is considered as a candidate motif. The only imposed restrictions are a prespecified minimum and maximum length and a maximum number of degenerate IUPAC characters. The advantage of such exhaustive approach is that the method yields globally optimal results. In order to strongly reduce the runtime and avoid excessive memory requirements, the MapReduce programming model (Dean and Ghemawat, 2008) was adopted as a means to take advantage of a parallel, distributed-memory cloud computing environment. By enabling disk I/O to store intermediate results, the current MapReduce implementation overcomes the memory bottleneck in a prototype implementation of this software that relied on the Message Passing Interface (MPI) for parallelization (De Witte et al., 2014).

5.2 Results and Discussion

BLSSpeller algorithm

The workflow of BLSSpeller is illustrated in Fig. 5.1. The input consists of gene families containing homologous promoter sequences from related species. The algorithm consists of an *intrafamily* and an *interfamily* step with a sorting step in between.

Intrafamily step:

In the intrafamily step, for each gene family individually, all words with a length between $k_{\min} = 6$ and $k_{\max} = 12$ characters that occur in the promoter sequences of that gene family are exhaustively enumerated. Words are spelled in the restricted IUPAC alphabet that consists of 11 characters: 4 base pairs (ACGT), 6 twofold-degenerate characters (RYSWKM) and the 'any' character (N). A maximum of $e_{\max} = 3$ degenerate characters are allowed per candidate motif. The degree of conservation of a word within the gene family is scored using the branch length score (BLS). The intrafamily step can operate in either alignment-free (AF) or alignment-based (AB) mode. In case of AF discovery, the BLS of a word is scored irrespective of its orientation or relative position within the promoter sequences. AB discovery adds the constraint that the words must be aligned in the multiple sequence alignment of the promoter sequences. Words for which the BLS exceeds threshold T are considered to be conserved within the gene family and retained for further processing. Six BLS thresholds T_i (i.e., 15%, 50%, 60%, 70%, 90% and 95%) were used in this study. At the end of this phase an exhaustive list of conserved words has been generated for each gene family individually.

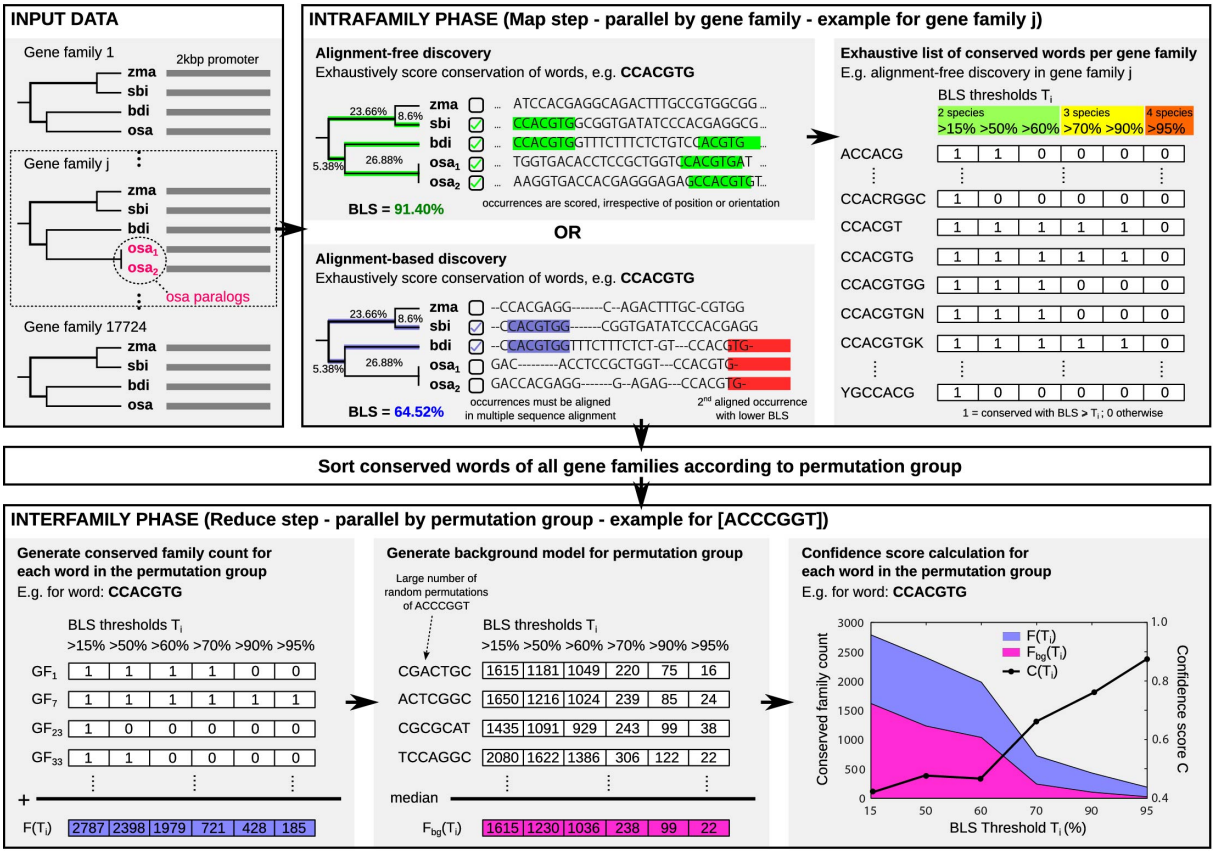

Figure 5.1: Overview of BLSSpeller. The input consists of homologous promoter sequences grouped into gene families. During the **intrafamily phase**, conserved words are exhaustively enumerated for each gene family individually. A word is considered to be conserved in a gene family if its branch length score (BLS) exceeds threshold T. Multiple BLS thresholds T_i can be used in a single run. In the *alignment-free* mode, the BLS of a word is computed irrespective of its orientation or relative position within the promoter sequences. Alternatively, in the *alignment-based* mode, words must appear aligned in the multiple sequence alignment. During the **sorting phase**, conserved words of all gene families are sorted according to permutation group, i.e., words with the same length and base content are grouped together. In the **interfamily phase**, permutation groups are handled individually. First, for each word, the conserved family count $F(T_i)$, i.e., the number of gene families in which the word is conserved with BLS $\geq T_i$, is established for all BLS thresholds T_i. Next, a background model $F_{\text{bg}}(T_i)$ is created by selecting the median value of the conserved family count of a large number of randomly generated instances of the permutation group, again for each threshold T_i. Finally, a confidence score $C(T_i)$ is computed for each T_i. Words for which $F(T_i) \geq F_{\text{thres}}$ and $C(T_i) \geq C_{\text{thres}}$ for any threshold T_i are considered to be genome-wide conserved motifs and are retained.

Interfamily step:

Using the data from the intrafamily step, for each word, the *conserved family count* $F(T_i)$, i.e., the number of gene families in which the word is conserved with a BLS $\geq T_i$, is counted for each BLS threshold T_i. Next, a confidence score $C(T_i)$, adopted from Stark et al. (2007), is established for each candidate motif (see Material and Methods). Two thresholds apply: motifs are only retained when $F(T_i) \geq F_{\text{thres}}$ and $C(T_i) \geq C_{\text{thres}}$ for any of the BLS thresholds T_i. Here, F_{thres} represents a threshold on the conserved family count and is used to eliminate words that are conserved in only few gene families and hence typically do not correspond to TF binding sites. Additionally, C_{thres} ensures that the candidate motif is conserved in a much higher number of gene families than what is expected for such a word (i.e., a word with the same length, base composition and degeneracy) and can hence be considered a potentially functional element. Motifs that satisfy both thresholds are considered to be *genome-wide conserved motifs*.

Note that the branch length score thresholds T_i on the one hand and conserved family count threshold F_{thres} and confidence score threshold C_{thres} on the other hand are independent. The former provides information about the degree of conservation within a single gene family whereas the latter are indicative of the degree of genome-wide conservation. Certain motifs only show up as being genome-wide conserved for high BLS thresholds. This is typically the case for short and/or highly degenerate motifs, where

also permutations of that motif are conserved with a moderate BLS in a rather large number of families, resulting in a low confidence score C. Conversely, a lower BLS threshold allows for the detection of longer motifs with genome-wide conservation in only a subset of the species. Using only a single BLS threshold would therefore limit the sensitivity of the method.

Exhaustive motif discovery in four monocot species

BLSSpeller was applied to four monocot species: *Oryza sativa ssp. indica* (osa), *Brachypodium distachyon* (bdi), *Sorghum bicolor* (sbi) and *Zea mays* (zma). Based on conserved gene content and genome organization, these grass species are considered to be a single genetic system (Bennetzin and Freeling, 1993), making a comparative motif discovery approach feasible. The dataset consists of 17 724 gene families each containing four orthologous genes (one from each organism). Additionally, 10 636 paralogs are taken into account. Hence, a total of 163 064 regulatory sequences (forward and reverse strands) with a length of 2 kbp each, were analyzed. BLSSpeller was run on this dataset using both the alignment-free (AF) and the alignment-based (AB) discovery mode on the Amazon Web Services (Elastic MapReduce) cloud infrastructure using 20 nodes of the type m1.xlarge. On every node, 7 map tasks and 2 reduce tasks were run in parallel. The computational requirements are listed in Suppl. Results[b]. Based on the Amazon pricing of 2014, the financial cost for performing these simulations amounted to 1080$ and 278$ for the AF and AB cases, respectively.

After the intrafamily step and using the AF discovery mode, an aggregated number of 537 billion words were found with a BLS $\geq 15\%$ (i.e., conservation in at least two species) over all 17 724 gene families. Note that these words are not necessarily unique as the same word can be conserved in multiple gene families. Using the AB discovery mode, only 82 billion words were found with a BLS $\geq 15\%$. This is because the AB discovery mode imposes the additional constraint that words should appear aligned in the multiple sequence alignment. After the interfamily step and using $F_{\text{thres}} = 1$ and $C_{\text{thres}} = 0.5$, the number of *genome-wide conserved* motifs amounted to 6.62 and 6.26 billion unique motifs, for the AF and AB discovery mode respectively.

The reason why the number of motifs is high is twofold. First, very relaxed thresholds F_{thres} and C_{thres} were used. It is computationally cheap to further filter this list using more stringent (and biologically meaningful) thresholds (see below). A second reason is the exhaustive, word-based nature of BLSSpeller. If a word is found to be genome-wide conserved, a large number of redundant, highly similar (e.g. slightly more degenerate) variants of that word may also appear in the final output of the method.

Estimation of the False Discovery Rate (FDR)

The output of BLSSpeller consists of a list of motifs, along with the conserved family count $F(T_i)$ and conservation score $C(T_i)$ for the six different BLS thresholds T_i. This list was filtered using more stringent thresholds for F_{thres} (i.e., 1, 10 and 20) and C_{thres} (i.e., 0.5, 0.7 and 0.9). Additionally, the list can be filtered by considering only a (stricter) subset of the BLS thresholds T_i (i.e., all six thresholds $T_1, \ldots, T_6$, three thresholds $T_4, \ldots, T_6$ corresponding to conservation in at least three species, a single threshold T_6 corresponding to conservation in all four species). The number of genome-wide conserved motifs for all 27 parameter combinations is shown in Fig. 5.2 for both AF and AB discovery. Clearly, each of the parameters has a strong influence on the final number of motifs in both the AF and AB discovery.

In order to assess the specificity of the method for the different parameter combinations, we estimate the false discovery rate (FDR) in an empirical fashion by running BLSSpeller on a random dataset generated using a zeroth-order Markov model (preservation of mononucleotide frequencies) as provided by RSAT (Thomas-Chollier et al., 2011). A more detailed version of Fig. 5.2 is available as Fig. C.1. Additional discussion of the limitations of the FDR analysis, higher-order Markov models and FDR analysis as a function of motif length and degeneracy is provided in Suppl. Results[b].

[b] http://bioinformatics.oxfordjournals.org/content/suppl/2015/08/07/btv466.DC1/supplementary_publication.pdf

[b] http://bioinformatics.oxfordjournals.org/content/suppl/2015/08/07/btv466.DC1/supplementary_publication.pdf

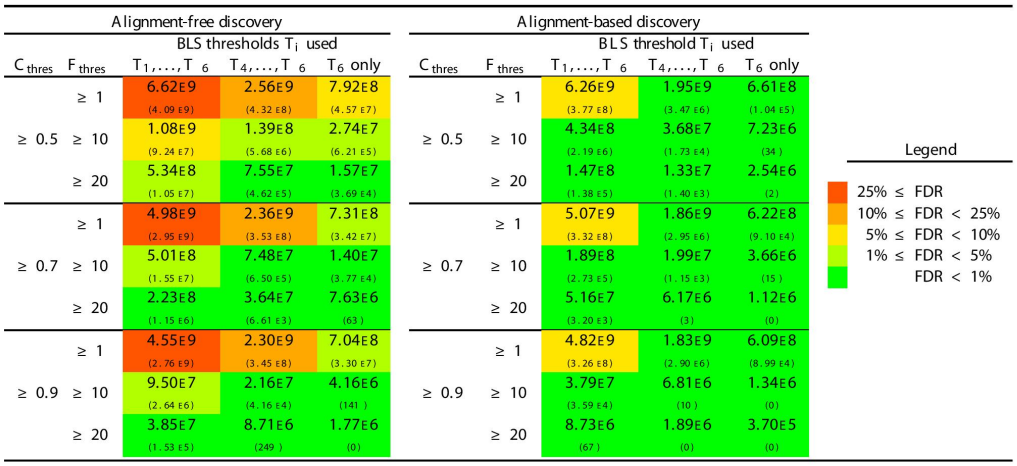

Figure 5.2: Number of genome-wide conserved motifs for both alignment-based and alignment-free discovery for different values of C_{thres} and F_{thres} and different subsets of the six BLS thresholds T_i ($T_1 = 15\%, T_2 = 50\%, T_3 = 60\%, T_4 = 70\%, T_5 = 90\%$ and $T_6 = 95\%$). Top number: real Monocot dataset; bottom number between brackets: random dataset (zeroth-order Markov model). The colors represent the false discovery rate (see legend).

A number of observations can be made. First, for comparable parameter settings, AB discovery has a lower FDR compared to AF discovery. The multiple sequence alignment method increases the specificity for AB discovery as relatively few words will be aligned in random data purely by chance. Second, low values of F_{thres} result in a poor FDR. The reason for this is that in such case, the output consists of a large number of words that are conserved in only a single gene family. If these words are long and/or have low degeneracy, most random permutations of that word will not be conserved in any gene family, resulting in a confidence score $C(T_i) = 1$. We therefore recommend to impose a certain threshold F_{thres} on the conserved family count. As functional transcription factors typically target multiple genes, this appears to be a biologically reasonable approach. Third, a reasonable threshold on the confidence score should be applied. Applying this threshold filters words for which their random permutations are conserved in a comparable number of gene families. This comprises low-complexity motifs and/or highly degenerate motifs. Finally, a more stringent definition of conservation results in an improved FDR. This can be obtained by imposing higher BLS thresholds T_i.

Even though there is a clear correlation between each of the parameters and the FDR, the exact FDR is hard to predict up front and likely also depends on the dataset that is used. We therefore recommend to run BLSSpeller with relaxed parameter settings on both real and random data, and to filter this output using more stringent parameters until a reasonable FDR is obtained.

For reasonably stringent parameter settings where the FDR $< 1\%$, the AF discovery mode reports 3.1 to 6.8 times more motifs compared to the AB discovery. At first glance, this may seem to be a trivial consequence of the relaxed definition of *conservation* in the AF methodology. Indeed, a word that is found to be conserved in a gene family with BLS $\geq T$ using the AB discovery will also be conserved in the AF method. Therefore, $F^{\text{AF}}(T) \geq F^{\text{AB}}(T)$ for each word. However, in order to establish the confidence score $C(T)$, the conserved family count $F(T)$ is compared to the corresponding median value $F_{\text{bg}}(T)$ of the background distribution (see Material and Methods). As $F^{\text{AF}}_{\text{bg}}(T)$ is also computed using the relaxed, alignment-free definition of conservation, it holds that $F^{\text{AF}}_{\text{bg}}(T) \geq F^{\text{AB}}_{\text{bg}}(T)$. Therefore, there is no reason to assume a priori that the AF mode will pick up more motifs than its AB counterpart, as can indeed be observed in Fig. 5.2 for a few parameter combinations, e.g., $F_{\text{thres}} = 1$, $C_{\text{thres}} = 0.7$ and BLS thresholds $T_1 \ldots T_6$. The reason that we do find more genome-wide conserved motifs for most parameter combinations (including those with good FDR) is because we found a significant number of known motif instances to be misaligned in this relatively highly diverged Monocot dataset. This is exemplified

in section 5.2.

Motif instance predictions correlate with experimental cis-regulatory datasets

The genome-wide conserved motifs discovered by BLSSpeller are highly redundant. High-scoring, motifs (AF discovery; BLS $\geq$ 15%, $C \geq 0.9$, $F \geq 20$; 38 462 976 motifs in total) were mapped back to the promoter sequences and were found to cluster around specific genomic regions (see Fig. C.2 and Fig. C.3). Certain loci are covered by thousands of highly similar motif variants. Nevertheless, the high-scoring motifs delineate distinct conserved genomic intervals on the promoter sequences. For these conserved regions, we investigated the accessibility for transcription factor binding in the promoter sequences of rice genes. DNase I hypersensitive 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. Such regions were recently characterized by Zhang et al. (2012a). We performed overlap analysis between conserved genomic regions (as determined by BLSSpeller) and open chromatin regions (see Suppl. Methods[b]). We found a significant enrichment (3.005 fold) of conserved regions for open chromatin regions (p-value < 0.001) (see Table 1). For a stricter subset of motifs (AF discovery; BLS $\geq$ 95%, $C \geq 0.9$, $F \geq 20$; 1 769 963 motifs in total), the fold enrichment increased to 3.796.

Table 5.1: Overlap between conserved genomic regions as identified by BLSSpeller and experimentally profiled open chromatin regions in rice and transcription factor binding sites inferred through protein-binding microarrays in rice and maize. Regions are required to fully overlap in order to be scored.

Overlap with experimentally profiled open chromatin regions (OCR) in *Oryza sativa*					
BLSSpeller thresholds	# conserved regions	# OCR regions	# conserved regions within OCR regions	# rand. conserved regions within OCR regions	enrichment fold
BLS $\geq$ 15%, $C \geq 0.9$, $F \geq$20	754 205	77 247	121 026	40 277	3.005
BLS $\geq$ 95%, $C \geq 0.9$, $F \geq$20	464 229	77 247	98 681	25 996	3.796
Overlap with experimentally profiled TF binding sites (TBS) in *Oryza sativa*					
BLSSpeller thresholds	# conserved regions	# TBS regions	# TBS regions within conserved regions	# TBS regions within rand. conserved regions	enrichment fold
BLS $\geq$ 15%, $C \geq 0.9$, $F \geq$20	754 205	442 506	159 542	42 522	3.752
BLS $\geq$ 95%, $C \geq 0.9$, $F \geq$20	464 229	442 506	37 093	5 689	6.520
Overlap with experimentally profiled TF binding sites (TBS) in *Zea mays*					
BLSSpeller thresholds	# conserved regions	# TBS regions	# TBS regions within conserved regions	# TBS regions within rand. conserved regions	enrichment fold
BLS $\geq$ 15%, $C \geq 0.9$, $F \geq$20	828 400	482 317	156 929	66 564	2.358
BLS $\geq$ 95%, $C \geq 0.9$, $F \geq$20	454 221	482 317	35 710	10 755	3.320

Additionally, we investigated the enrichment of TF binding sites determined in vitro (Weirauch et al., 2014) towards conserved genomic regions in rice and maize. Transcription factor DNA binding specificities are the primary mechanism by which transcription factors recognize genomic features and regulate genes. Recently, a dataset containing a large number of these binding specificities was generated using protein-binding microarrays (PBM) (Weirauch et al., 2014). From this database, PWMs were downloaded for 481 TFs in rice and for 615 TFs in maize. These were mapped onto the respective rice and maize promoters and overlap analysis was performed (see Suppl. Methods[b]). In rice, of the 754 205 constrained genomic regions (BLS $\geq$ 15%), 159 542 contain a PBM-based TF binding site, leading to 3.752 fold enrichment (p-value < 0.001). Again, for the stricter subset of conserved motifs (BLS $\geq$ 95%), fold enrichment increased to 6.520. Maize showed a fold enrichment of 2.358 and 3.320 (p-value < 0.001) respectively. Overall, these analyses revealed that a large part of the conserved non-coding sequences can be accessed by DNA binding proteins and as such can act as functional transcription factor binding sites, and that these conserved non-coding sequences show enrichment for the binding sites of a large number of TFs inferred using PBMs.

Conservation of the ga2ox1-like KN1 binding site

KNOTTED1 (KN1) transcription factors are involved in the establishment and maintenance of plant meristems and are thought to be conserved among the family of grasses (Bolduc and Hake, 2009). Bolduc et al. (2012) profiled KN1 binding sites in *Zea mays* using ChIP-seq experiments. The overlapping loci

[b] `http://bioinformatics.oxfordjournals.org/content/suppl/2015/08/07/btv466.DC1/supplementary_publication.pdf`

in two samples of immature ears were retained and assigned to the nearest gene within a range of 10 kbp. The ChIP-Seq peaks were found to be mainly situated in the 5' en 3' regions extending from the gene but also occur in introns and exons. Thus, a set of 5 118 candidate KN1-regulated maize genes were identified. For approximately 7% of these genes, a binding site reminiscent of the intronic KN1 binding site in ga2ox1, was identified. For these so-called ga2ox1-like KN1 binding sites, a Position Weight Matrix (PWM) was derived by Bolduc et al. (2012). Translated to the IUPAC alphabet, this PWM corresponds to TGAYNGAYDGAY.

We investigate whether BLSSpeller is able to discover the ga2ox1-like KN1 motifs and binding sites through a comparative study of the four monocot species. From the BLSSpeller output, all genome-wide conserved motifs of length 12 that match the ga2ox1-like KN1 PWM identified by Bolduc et al. (2012) were retained. Using alignment-free discovery, and using $F_{\text{thres}} = 20$ and $C_{\text{thres}} = 0.7$ (FDR $\leq 1\%$, see Fig. 5.2), 51 genome-wide conserved motif variants are identified. In total, these motifs are conserved in 165 gene families with a BLS $\geq 15\%$ (i.e., conservation in at least two species). From the 51 identified motif variants, only 19 are required to explain the conservation in all 165 gene families. These essential motifs are listed in Table 5.2 along with their respective metrics. In turn, these gene families contain 213 maize genes in total, 51 of which were also identified in Bolduc et al. (2012). These results were compared to those obtained by Fastcompare (Elemento and Tavazoie, 2005) (see Suppl. Results[b]), a method that also performs motif discovery in an alignment-free and exhaustive manner. However, Fastcompare is limited to the exact ACGT alphabet and pairwise species comparisons. Because of these limitations, Fastcompare could identify only 36 maize gene targets, 10 of which were also identified by Bolduc et al. (2012).

Table 5.2: List of genome-wide conserved ga2ox1-like KN1 motif variants identified by BLSSpeller using both AF and AB discovery. $F(15\%)$ denotes the number of gene families in which the motif is conserved with BLS $\geq 15\%$ while $C(15\%)$ denotes the corresponding confidence score. $\mathscr{M}_{\text{BLS}}$ denotes the number of maize genes contained in the gene families while $\mathscr{M}_{\text{inters}}$ denotes the intersection $\mathscr{M}_{\text{BLS}} \cap \mathscr{M}_{\text{ChIP}}$ with experimentally profiled maize genes.

Alignment-free discovery					Alignment-based discovery				
KN1 motif variant	$F(15\%)$	$C(15\%)$	$\mathscr{M}_{\text{BLS}}$	$\mathscr{M}_{\text{inters}}$	KN1 motif variant	$F(15\%)$	$C(15\%)$	$\mathscr{M}_{\text{BLS}}$	$\mathscr{M}_{\text{inters}}$
TGATNGATKGAY	59	0.93	75	24	TGATNGAYGGAY	11	0.91	10	3
TGATNGAYKGAT	59	0.93	74	20	TGATNGATKGAY	11	0.82	11	3
TGAYNGATKGAT	54	0.93	68	21	TGAYNGACKGAC	10	0.90	11	3
TGATNGAYWGAT	40	0.88	50	11	TGAYGGAYGGAY	9	1.00	9	3
TGAYNGAYTGAT	36	0.89	48	11	TGATNGAYRGAT	9	0.89	10	3
TGAYTGAYTGAY	33	0.97	42	9	TGAYNGAYTGAC	8	0.88	9	2
TGATNGAYTGAY	32	0.88	40	7	TGACNGAYTGAY	8	0.88	10	3
TGAYNGATWGAT	31	0.84	42	12	TGACNGACWGAY	7	0.86	7	2
TGATNGATWGAY	30	0.83	36	9	TGACAGAYRGAY	3	1.00	4	0
TGATNGATRGAY	29	0.86	39	9					
TGAYNGATRGAT	27	0.85	37	9					
TGATNGAYRGAT	26	0.85	35	8					
TGAYNGATTGAY	25	0.84	34	7					
TGAYNGATGGAY	24	0.88	35	9					
TGATNGAYGGAY	24	0.88	31	8					
TGAYTGAYWGAT	22	0.91	27	6					
TGAYNGACTGAY	22	0.91	28	9					
TGAYNGAYTGAC	21	0.90	27	8					
TGAYNGACKGAC	20	0.90	25	10					
Union (all variants)	165	–	213	51	Union (all variants)	37	–	41	10

Similarly, using BLSSpeller's alignment-based discovery mode, conservation with a BLS $\geq 15\%$ is observed in only 37 gene families, even with very relaxed thresholds ($F_{\text{thres}} = 1$ and $C_{\text{thres}} = 0.7$) (FDR $\leq 10\%$). The 9 essential motif variants required to explain this conservation are listed in Table 5.2. The 37 gene families contain 41 maize genes, 10 of which are also reported in Bolduc et al. (2012). Inspection of the promoter sequence alignments of the gene families reveals that the ga2ox1-like KN1 variants are often not aligned, either because the motif instances in the different species are located at entirely different positions in the promoter sequences or because they appear on different strands (see Suppl. Results[b]). Therefore, alignment-based motif discovery approaches such as BLSSpeller in AB mode or the 'mini motifs' approach as used by Stark et al. (2007) suffer from reduced sensitivity on diverged datasets.

5.3 Conclusion

A novel phylogenetic footprinting approach was developed for the sensitive discovery of conserved cis-regulatory elements even in diverged sequences. Using IUPAC strings as motif model and using the MapReduce programming model to enable distributed computing, it was shown that it is feasible to compute all genome-wide conserved words in a large dataset in an exhaustive manner. For a given false discovery rate, it was demonstrated that an alignment-free approach detects more conserved words than an alignment-based approach. Even though millions of genome-wide conserved motifs were identified by our method, mapping of these motifs to the promoter sequences results in constrained conserved genomic regions. It was shown that these conserved regions were significantly enriched for experimentally profiled open chromatin regions in rice and for TF binding sites inferred through protein-binding microarrays in rice and maize. Finally, it was shown that the alignment-free approach shows an improved recovery of the ga2ox1-like KN1 binding site, compared to the alignment-based approach or competing methods.

5.4 Methods

Generation of gene families

The orthology relationships between the genes of the four different monocot plant species were inferred using the 'integrative orthology viewer' in the PLAZA 2.5 platform (Proost et al., 2009; Van Bel et al., 2012). Homologous (i.e. orthologous and paralogous) genes were grouped in gene families and their promoter sequences 2 kbp upstream from the translation start site were extracted. In its most simple form, a family consists of four orthologous genes: one from each organism. In that case, the phylogenetic tree by Reineke et al. (2011) is used. For gene families that comprise one or more paralogs, gene family-specific phylogenetic trees can be constructed that take into account the specific order in which the duplications and speciation events occurred. For simplicity, we assume that all paralogous gene duplications occurred recently. This is modeled by adding a bifurcation with a branch length of zero to the phylogenetic tree which means that only conservation between different species contributes to the branch length score. Note that besides promoter regions, additional homologous sequences of interest (e.g., intronic regions) could be added to the input dataset.

Intrafamily step: conservation within a gene family

For all gene families individually, all words with a length between $k_{\min}$ and $k_{\max}$ characters that occur in any of the sequences are exhaustively enumerated and their degree of conservation within that family is quantified. Words are spelled in the IUPAC alphabet or a subset thereof. Up to $e_{\max}$ degenerate (i.e., non-ACGT) characters are allowed per word. The intrafamily phase can operate in alignment-free or alignment-based mode.

In the alignment-free approach, a generalized suffix tree (GST) is constructed (Giegerich et al., 1999) from the promoter sequences and their reverse complements in the gene family. Using Sagot's Speller algorithm (Marsan and Sagot, 2000), the GST is used to efficiently and exhaustively report all words in the IUPAC alphabet along with the sequences in which they occur. Additional algorithmic details and runtime information are described in Suppl. Methods[b].

The alignment-based mode requires a pregenerated multiple sequence alignment (MSA) of the orthologous promoters in a gene family. Dialign-TX (Subramanian et al., 2008) was chosen to create these MSAs in view of good results on a non-coding alignment benchmark (Pollard et al., 2004). For every position in the alignment, a small GST is generated containing only the suffixes of the sequences that start at that position. The same Speller algorithm is run to report all words and the sequences in which they occur at aligned positions, again using the IUPAC alphabet.

For every word, the degree of conservation in each gene family is quantified using the branch length score (BLS). Given the sequences in which the word occurs, the BLS can be calculated by finding the

[b] http://bioinformatics.oxfordjournals.org/content/suppl/2015/08/07/btv466.DC1/supplementary_publication.pdf

minimum spanning tree that connects these sequences in the phylogenetic tree. The sum of the weights of the horizontal branches in the minimum spanning tree then represents the BLS (Stark et al., 2007). In alignment-based mode, the same motif can occur at multiple aligned positions within a single family; in that case only the highest BLS value is used. Only words for which the BLS exceeds a prespecified threshold T are retained. Such words are said to be *conserved* within the gene family.

Interfamily step: genome-wide conservation

The conserved words of all gene families are sorted according to base content and partitioned into *permutation groups* whose elements are permutations of each other. All words in a permutation group hence have the same length, base content and degeneracy. For example, the words AWTC, WTAC and CAWT belong to the same permutation group.

The number of occurrences for each distinct word within a permutation group is counted. This number corresponds to the number of gene families in which that word is conserved with a BLS $\geq T$ and is referred to as the *conserved family count* $F(T)$. Genome-wide conserved motifs are selected based on the fact that they have a conserved family count $F(T)$ that is (much) higher than the median conserved family count of the member instances of their permutation group. This median value, denoted as $F_{\text{bg}}(T)$ (bg = background) represents the *expected* conserved family count for a word in that permutation group. $F_{\text{bg}}(T)$ is approximated by randomly generating a large number (default=1000) of instances of the permutation group, i.e., random words with the same length and base content and computing the median value for the conserved family count. Note that some of those random instances can have a conserved family count equal to zero.

A confidence score C, adopted from (Stark et al., 2007), is obtained for each word in the permutation group by comparing $F(T)$ and $F_{\text{bg}}(T)$ as follows:

$$C(T) = 1 - \frac{F_{\text{bg}}(T)}{F(T)}$$

Words for which $F(T) \geq F_{\text{thres}}$ and $C(T) \geq C_{\text{thres}}$ are considered *genome-wide conserved* motifs and are retained by the method where F_{thres} and C_{thres} denote user-defined thresholds. The output of the method consists of an exhaustive list of motifs which satisfy these thresholds, along with the $F(T)$ and $C(T)$ metrics. Similar to Stark et al. (2007), rather than using a single threshold T, multiple BLS thresholds T_i can be used in a single run. The confidence score $C(T_i)$ is then computed for all thresholds T_i individually, i.e., $C(T_i) = 1 - \frac{F_{\text{bg}}(T_i)}{F(T_i)}$. Here, $F(T_i)$ denotes the number of families in which the motif is conserved with a BLS higher than the threshold T_i. Similarly, $F_{\text{bg}}(T_i)$ is the corresponding value for the background model. Words for which $F(T_i) \geq F_{\text{thres}}$ and $C(T_i) \geq C_{\text{thres}}$ for *any* of the BLS thresholds T_i are retained.

MapReduce implementation

The method was implemented using the MapReduce (Dean and Ghemawat, 2008) programming model. The map phase corresponds to the intrafamily phase in which the gene families are processed in parallel by the different *mappers*. The reduce phase corresponds to the interfamily phase in which the permutations groups are processed in parallel by the different *reducers*. In between the map and reduce step, the candidate motifs are sorted according to length and base content in order to create the permutation groups.

Acknowledgement

We acknowledge the support of Ghent University (Multidisciplinary Research Partnership 'Bioinformatics: From Nucleotides to Networks') and Dries Vaneechoutte, Kenneth Hoste, Ewan Higgs and Stijn De Weirdt for technical assistance. Part of the computational resources (Stevin Supercomputer Infrastructure) and services used in this work were provided by the VSC (Flemish Supercomputer Center), funded by Ghent University, the Hercules Foundation and the Flemish Government department EWI.

CHAPTER 6

A DNA binding-site landscape and regulatory network analysis for NAC transcription factors in *Arabidopsis thaliana*[a]

6.1 Abstract

Target gene identification for transcription factors is a prerequisite for the systems wide understanding of organismal behavior. NAM-ATAF1/2-CUC2 (NAC) transcription factors are among the largest transcription factor families in plants, yet limited data exists from unbiased approaches to resolve the DNA-binding preferences of individual members. Here, we present a TF-target gene identification workflow based on the integration of novel protein binding microarray data with gene expression and multi-species promoter sequence conservation to identify the DNA-binding specificities and the gene regulatory networks of 12 NAC transcription factors. Our data offer specific single base resolution fingerprints for most TFs studied and indicates that NAC DNA binding specificities might be predicted from their DNA binding domain's sequence. The developed methodology, including the application of complementary functional genomics filters, makes it possible to translate, for each TF, protein binding microarray data into a set of high-quality target genes. With this approach, we confirm NAC target genes reported from independent in vivo analyses. We emphasize that candidate target gene sets together with the workflow associated with functional modules offer a strong resource to unravel the regulatory potential of NAC genes and that this workflow could be used to study other families of transcription factors.

[a]This chapter is based on Lindemose et al. (2014). F.D.M., K.S., K.V., M.K.J. designed the research methodology. S.L., C.S., C.T.W. and F.D.M (PBM experiments) performed experiments. M.K.J. (synthetic promoter & co-expression analyses), J.V.d.V., K.V (target gene detection, data integration and functional module enrichment) and K.S.H (provided functional modules dataset) performed data analyses. F.D.M., K.S., K.V., M.K.J., S.L., J.V.d.V wrote the manuscript. A complete reference can be found in Chapter 8.

6.2 Introduction

Plants use cellular strategies to survive exposure to biotic and abiotic stress. Drought, salt, high temperature and microbial infections are among the most frequent abiotic and biotic stresses encountered by plants (Keurentjes et al., 2011; Lindemose et al., 2013; Matsui et al., 2008; Zheng et al., 2012). Expression of genes that function in stress sensing and tolerance are regulated upon stress exposure by specific TFs (Keurentjes et al., 2011; Lindemose et al., 2013). The NAC (NAM/ATAF/CUC) family of proteins is a major group of plant-specific TFs involved in plant development, senescence, secondary cell wall formation and stress responses (Nakashima et al., 2012; Puranik et al., 2012; Olsen et al., 2005). The well studied model plant *Arabidopsis thaliana*, and economically important crops such as *Nicotiana tabacum*, *Hordeum vulgare* and *Oryza sativa*, each hold the potential to express more than 100 different NAC proteins (Lindemose et al., 2013; Nakashima et al., 2012; Puranik et al., 2012). When genes encoding NAC TFs are over-expressed in plants robust phenotypes including salt and drought tolerance have been observed (Lindemose et al., 2013; Hu et al., 2006; Jeong et al., 2010). Likewise, *nac mutant* plants have been shown to display loss of secondary wall thickening, perturbed resistance towards microbial attack as well as delayed senescence (Lindemose et al., 2013; Nakashima et al., 2012; Puranik et al., 2012; Mitsuda et al., 2005), though functional redundancy often has hampered characterization of individual NAC members. NAC proteins consist of a conserved N-terminal DNA binding domain (DBD), known as the NAC domain, which is also responsible for the oligomerization into dimeric proteins (Olsen et al., 2005; Ernst et al., 2004). The C-terminal region of NAC members is more diverse, intrinsically disordered, and functions as a transcription regulatory domain (Kjaersgaard et al., 2011; Jensen et al., 2010a). Determination of the X-ray structure of the NAC domain from *Arabidopsis thaliana* ANAC019 revealed a novel dimeric DBD predominantly composed of beta-sheets with no well-characterized DNA binding motifs (Ernst et al., 2004). Characterization of the dimerization surface demonstrated that ANAC019 is only able to bind DNA as homo- and hetero-dimers. In addition, the consensus DNA binding sequences of two distantly related NAC TFs, ANAC019 and ANAC092, were identified by in vitro selection (SELEX) and appeared to have minor differences in their DNA-binding specificities (Olsen et al., 2005). For both proteins, the identified core consensus DNA binding sequence was TTNCGT[G/A]. Interestingly, in a recent study it was found that 9 distantly related NAC TFs were able to bind this sequence, though with different affinities (Jensen et al., 2010a). In line with these results, it has been shown that several other NAC TFs bind the core CGT[G/A], but with considerable sequence differences in the flanking bases of the binding site (Xu et al., 2013). Thus, the flanking bases next to the core CGT[G/A] of NAC binding sites in promoters may determine the binding specificities and fine-tune affinity for different NAC TFs in vivo. This effect was recently demonstrated to be highly relevant in the family of basic Helix-Loop-Helix (bHLH) transcription factors (Grove et al., 2009; Masi et al., 2011; Gordan et al., 2013).

Apart from focused dimerization and DNA-binding studies on NAC TFs, global mapping of gene regulatory networks (GRNs) can be facilitated by high-throughput approaches that allow for the discovery and high resolution characterization of genome-wide DNA binding specificities of DNA binding proteins. Protein Binding Microarrays have been widely used as an unbiased and condition independent method for the identification of high resolution DNA specificities for a larger number of TFs from several organisms (Grove et al., 2009; Berger et al., 2008; Badis et al., 2009; Newburger and Bulyk, 2009). PBMs can uncover binding specificities of TFs at the k-mer level, with single-base resolution. Also, PBM data have been shown to strongly correlate with surface plasmon resonance studies of TF-DNA interactions (Berger et al., 2006; Siggers et al., 2011), thus allowing the use of PBM data to analyse biologically relevant data. Further integration of such data with genome annotations, gene expression data and functional modules (Heyndrickx and Vandepoele, 2012), will result in the functional characterization of the mapped observed TF-DNA interactions and possibly the unraveling of TF and condition specific GRNs (Kjaersgaard et al., 2011; Jensen et al., 2010a; Grove et al., 2009; Giorgetti et al., 2010; Wong et al., 2011). In this study, we report the integration of PBM results with co-expression data and functional module enrichment to outline the regulatory network for 12 NAC proteins. Furthermore, we show that this integrative strategy, applicable to any TF target gene analysis, allows for the refinement and increase in significance of TF target genes. We also use our PBM data to motivate mutations in an

element identical to a region of a selected target gene promoter and propose that a simple 2-nucleotide substitution may be exploited to control binding of native TFs to novel promoter elements. Finally, co-expression analysis is used to validate the regulatory potential predicted from our unbiased PBM analysis. This study is the first systems-wide analysis of the NAC family of transcription factors resulting in a global map of the NAC DNA-binding specificities in A. thaliana and we envision the data to be useful for future engineering of improved stress responses in plants.

6.3 Results

DNA binding specificity analysis of individual NAC TFs

Systematic analysis of NAC DNA-binding specificity by Protein Binding Microarrays (Grove et al., 2009; Berger et al., 2008; Berger and Bulyk, 2009) was performed on 12 NAC TFs representing functionally important clades and spanning the phylogenetic diversity of the NAC family (Supp. Figure D.1)(Jensen et al., 2010a). ANAC019 was selected because its NAC domain structure is known (Ernst et al., 2004; Welner et al., 2012), and because it is implicated in networks of stress responses and senescence (Jensen et al., 2010a; Hickman et al., 2013). ANAC055 and ATAF1 are closely related to ANAC019 (Jensen et al., 2010a; Bu et al., 2008), and ATAF1 is a control for the PBM experiments (Jensen et al., 2013). They all cluster together with senescence-associated NAP (Guo and Gan, 2006) based on hormone-dependent gene regulation (Jensen et al., 2010b). Therefore, analysis of these NAC TFs could reveal simple relationships between amino acid sequence and DNA-binding specificity. ANAC092/ORE1 represents a functionally important NAC sub-group (Jensen et al., 2010a; Balazadeh et al., 2010). VND3, VND7, NST2 and SND1 represents a NAC sub-group that is central to secondary cell wall formation (Zhong et al., 2007, 2010; Yamaguchi et al., 2011). NTL8 and NTL6 are transmembrane NAC TFs (Seo and Park, 2010), and NTL6 acts through known binding sites in Pathogenesis-Related (PR) genes (Seo et al., 2010) allowing comparison of PBM and in vivo promoter binding data. The distant NAC members, SOG1 (Yoshiyama et al., 2009), ANAC003 (Jensen et al., 2010a) and VOZ2, were also included. VOZ2 has a zinc finger region N-terminally of the NAC domain (Mitsuda et al., 2004). In the other NAC TFs, the N-terminal NAC domain is followed by various intrinsically disordered transcriptional regulatory domains (TRDs) (Jensen et al., 2010a) (Supp. Figure D.2). Since only the NAC domain is used in this study and since remote disordered regions may fine tune both specificity and affinity of DNA-binding (Fuxreiter et al., 2011) full-length ANAC092 was also used for the PBM experiments. Finally, the WRKY domain of the WRKY1 TF was included due to its well-defined DNA-binding specificity (Turck et al., 2004). We generated a list of 4821 gapped and ungapped 8-mers (Materials and Methods and Supp. Table S2[a]) that showed an Enrichment Score (ES) equal to or greater than 0.40 for at least one tested protein. Clustering of these k-mers revealed that NAC transcription factors can be separated into 3 distinct groups characterized by their DNA specificities (Figure 6.1A). Interestingly, these groups largely match the 3 main branches in the phylogenetic tree shown in Supp. Figure D.1A. Cluster 1, which comprises ANAC019, ANAC055, ANAC092, ATAF1, NAP, NST2, SND1, VND3 and VND7, shows a clear binding preference for the accepted NAC-BS model, T[G/A]CGT (Figure 6.1B) (Olsen et al., 2004). This cluster can be further separated into clusters 1a and 1b. Cluster 1a contains ANAC092, SND1 and NST2 which show a distinctive specificity for TTGCGT. Cluster 1b contains ANAC019, ANAC055, VND7, ATAF1, NAP and VND3, which show a main specificity for the TACGT core motif (Figure 6.1B). This agrees with our earlier results on ATAF1 using a different set of deBruijn sequences and array design (Jensen et al., 2013). Interestingly, VND3 and VND7 are closer, in their sequences, to proteins in Cluster 1a (Supp. Figure D.1A) yet their DNA specificity model groups these TFs with Cluster 1b hinting at minor, yet critical, residue differences that would be able to dictate DNA binding properties of the TF. Reassuringly, Cluster 1a also contained both forms of ANAC092. This observation, together with the logos for both proteins in Figure 6.1B show that the full-length version of ANAC092 binds with higher affinity to an expanded range of k-mers compared to the NAC DBD-only version. Importantly, the DNA-binding specificity was not significantly changed by the disordered C-terminus of ANAC092 (Figure 6.1B, Supp. Figure D.2). This suggests that the intrinsically disordered region of ANAC092 assists the DNA binding giving an overall better binding/higher affinity, possibly

through modulation of conformation, flexibility or spacing within the DNA-protein complex (Fuxreiter et al., 2011). Cluster 2 only contains VOZ2, which distinct preference has a very strong resemblance to a zinc finger motif CCCGCC as shown by, for example, Klf7 (Badis et al., 2009) or Sp1 (Kadonaga et al., 1986). It has been shown that VOZ2's zinc finger is required for DNA binding and this specificity could confirm this requirement (Mitsuda et al., 2004). SOG-1 and ANAC003 failed to generate binding data. Cluster 3 containing NTL6 and NTL8, shows a surprising specificity for k-mers containing TT(A/C/G)CTT (Figure 1B) and, additionally, NTL6 and NTL8 specific k-mers do not appear to show any overlap (Figure 6.1A) with Clusters 1a, 1b or 2. Finally, our PBM data confirms the specificity of WRKY1 for the W box consensus motif TTGACC/T (Figure 6.1A), as previously reported from in vivo ChIP studies (Turck et al., 2004).

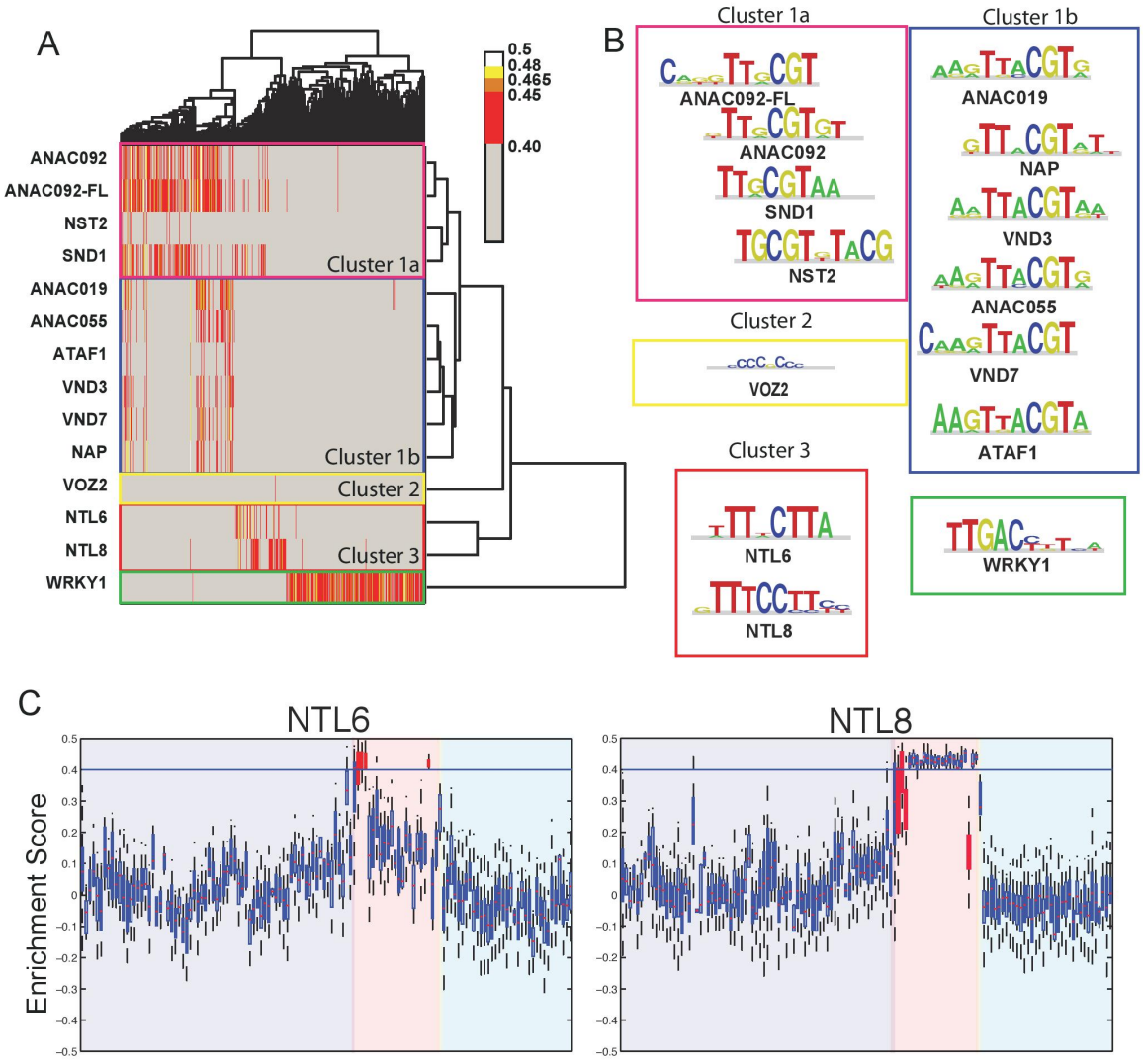

Figure 6.1: **DNA binding profiles of NAC TFs can be separated into 5 specificity clusters.** *A) Bi-dimensional clustergram of the identified 4821 significant k-mers (X axis) vs studied TFs (Y axis). Colored boxes indicate clusters of TFs showing similar DNA specificity profiles at the k-mer level. B) DNA specificities for each TF, grouped in clusters as in Figure 6.1A. C) Enrichment Score distributions for NTL6 and NTL8 shown as boxplots. Red boxes show NTL6 specific k-mer groups. The identity of each k-mer is available in Supp. Fig. D.1B. For each box, the central mark represents the medial value for the distribution, the box edges represent the 25th and 75th percentiles and the whiskers extend to the last non-outlier data point, as described in Matlab's "boxplot" help documentation (`http://www.mathworks.se/help/stats/boxplot.html`).* .

From our PBM analysis, we conclude that NAC proteins show specificities for at least 3 different consensus models, and that the differences in DNA binding specificities largely match the 3 main branches

[a] `http://nar.oxfordjournals.org/content/suppl/2014/06/09/gku502.DC1/nar-00465-z-2014-File008.xlsx`

in the phylogenetic tree shown (Supp. Figure D.1A) This indicates that NAC DNA binding specificities may be estimated from their DNA binding domain's sequence. In order to uncover hidden specificities present in each TF's data, we analysed the available PBM data using shorter word sequences that can represent the full extent of the data in a simple manner. Using a combination of manual and statistical analyses (Jiang et al., 2013) we identified 130 6-mers (ie: ungapped 6-mers and gapped 7-mers) that are able to describe, with high precision, the variation in specificities for each TF at a single base resolution. Additionally, these k-mers allow for the direct comparison of the differences in relative affinity of each protein for each k-mer. Analyses of these comparisons (Figure 6.1C and Supp. Figure D.3) result in the identification of TF-specific k-mers and to have a high resolution fingerprint of the relative affinities of each protein against each key k-mer. For example, NTL6 and NTL8 show similar overall specificity models (Figure 6.1B) yet it is evident that their binding preferences, when looking at shorter k-mers, are dramatically different (Figure 6.1C) and there is no overlap between high-ES k-mers for NTL6 and NTL8, even though their overall specificity models are very similar (Figure 6.1B). Finally, we can rank the individual TFs by overall DNA binding specificity. By simple observation of the boxplots in Supp. Figure S3, we can conclude that ANAC019, ANAC055, ANAC092, SND1 and NTL8 show broad and high specificities, within their subclass (or Cluster) compared to the other NACs.

Our results show that though some NAC TFs share specificities, evident differences among top-ranking k-mers are observed in their binding site preferences. Thus from this detailed analysis we can generate precise specificity models, or fingerprints, for each TF which will uniquely define the spectrum of DNA sequences recognized by each NAC protein.

Identification of direct NAC target genes from DNA-binding data and microarray analysis

Using our PBM results, we next aimed at determining target genes involved in NAC-specific signaling in Arabidopsis. Raw PBM target genes were predicted by initially determining, for each TF, a set of high scoring seed 8-mers and mapping these to the 1kb promoters of all Arabidopsis genes. This resulted in a large number of predicted target genes (P) for the different TF (Supp. Table S3[a] and Supp. Figure D.4). For 3 TFs (ANAC019, ANAC055 and ANAC092), transcriptional profiling of mutant lines resulted in a set of differentially expressed (DE) genes (Hickman et al., 2013; Balazadeh et al., 2010), which were used to evaluate our data processing methodology and to define additional criteria to delineate functional target genes. Although DE genes contain directly as well as indirectly regulated genes, they offer a valuable source of information to assess whether TF binding inferred through PBMs corresponds with TF regulation. As the sets of P target genes showed only moderate enrichment for DE genes in the mutant lines (1.09-1.21 fold enrichment)(Supp. Figure D.5), co-expression and motif conservation information were combined with the PBM data to identify more biologically relevant target genes. Integration of expression data, through enrichment analysis of gene-centric co-expression clusters for P target genes (see Materials and Methods), resulted in a reduced set of predicted + co-expressed PBM target genes (P+COE) (Figure 6.2). For all 3 PBM experiments these candidate target gene sets showed significant overlap with the DE genes yielding higher enrichments (1.68-2.97 fold enrichment) compared with the full set of predicted target genes defined without co-expression information (Supp. Figure D.5). Conservation of PBM motif instances was determined using a multi-species alignment-based phylogenetic footprinting approach with 11 related dicotyledonous species (see Materials and Methods). The inclusion of motif conservation returns a set of target genes conserved within dicot plants (conserved P+COE), for ANAC055 these conserved targets showed an increased enrichment for DE genes (4.78 fold enrichment, see Figure 6.2) compared to only using co-expression as a filter. A similar increase in specificity for functional Gene Ontology (GO) enrichments was observed when comparing the DE gene sets with subsequent filtering of the P target genes using co-expression and motif conservation (data not shown). These results demonstrate that the developed methodology combined with the application of complementary functional genomics filters makes it possible to translate, for each TF, the high-scoring k-mers into a set of high-quality predicted genes, which provide the basis to study different biological processes controlled by several NAC genes. All further analyses are performed using the P+COE target genes because this set has the best balance between sensitivity and specificity. The NAC P+COE target genes were used to generate a gene regulatory network (GRN) comprising 22,489 interactions for 12 TFs

and 9706 P+COE target genes (Supp. Table S3[a] and S4[a]). A set of known TF-target gene interactions curated from literature (Hussey et al., 2013) was used to evaluate the GRN. Experimentally determined target genes were present for 3 TFs (SND1, VND7 and NST2) in our study. Overall, 32% (31/98) of the interactions compiled from different small-scale experiments were recovered by our GRN, indicating that apart from generating many novel interactions, also multiple known interactions were successfully recovered using our approach. Condition and tissue dependent regulation, lack of co-factor data, as well as chromatin state/accessibility information are factors that can interfere with the accurate detection of functional target genes and can cause the mis-identification of a limited set of known regulated genes.

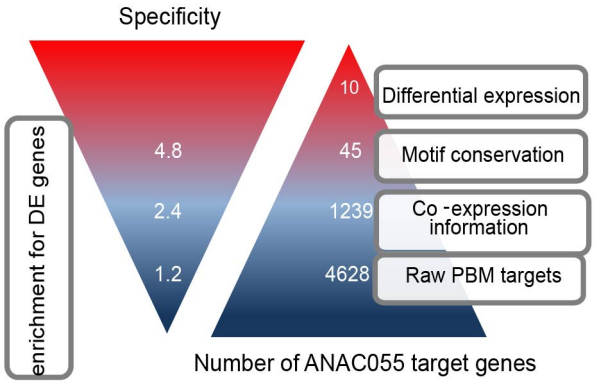

Figure 6.2: **Overview of additional genomic filters leading to TF target genes with increased biological relevance.** *Starting from the predicted PBM ANAC055 target genes, the inclusion of co-expression information and motif conservation leads to a reduced set of target genes (right triangle) showing increased enrichment for DE genes obtained from a ANAC055 perturbation transcript profiling experiment (left triangle). Specificity refers to the enrichment fold for DE genes in the different target gene sets. Whereas motif conservation results in an increased specificity for DE genes compared to predicted PBM targets for ANAC019, ANAC055 and ANAC092, combining co-expression information with motif conservation leads to an additional gain in enrichment for ANAC055.*

To study the overlap of the P+COE target genes, the sets of target genes for the different TFs were compared (see Supp. Figure D.6 and Supp. Table S5[a]). Clustering of the TFs based on the shared target genes revealed two clusters, one containing ANAC092, NST2, ANAC019, ANAC055, NAP, ATAF1, VND3 and VND7, and one containing, SND1, NTL6 and NTL8. Due to the low number of candidate target genes, VOZ2 shows very low overlap scores with the other TFs (Supp. Figure D.6). The high overlap scores between ANAC092, ANAC055 and ANAC019 (> 5-fold enrichment, hypergeometric p-value < 0.01) are in agreement with the significant overlaps between the DE genes obtained from transcript profiling on the corresponding mutants (3-6 fold enrichment, p-value < 0.01; see Supp. Figure D.7), suggesting substantial functional redundancy between those TFs. Functional redundancy between ANAC019 and ANAC055 was previously described in literature (Zheng et al., 2012; Bu et al., 2008; Tran et al., 2004), although some diversity is seen for their senescence associated regulons (Hickman et al., 2013). Furthermore our results can confirm the presence of binding sites for ANAC055 and ANAC019 in the promoter of BSMT1, a salicylic acid methyltransferase, and the highest target gene overlap (84%) was found between ANAC055 and ANAC019. The functional redundancy of P+COE targets was also evaluated through overlap analysis of enriched functional modules. These functional modules comprise a set 13,142 genes (1,562 modules) annotated with specific functional descriptions based on experimental GO information, protein-protein interaction data, protein-DNA interactions or AraNet gene function predictions (Heyndrickx and Vandepoele, 2012). As ANAC019 and ANAC055 also show a significant overlap (80%) of functional modules (p-value < 0.01), these results corroborate the functional redundancy between these two NAC TFs. Other NAC TFs also showed a large overlap in enriched functional modules (Figure 6.3A and Supp. Figure D.8). Comparing the expression profiles of the different TFs during transcript profiling in different stress conditions (Supp. Figure D.9) further supports the functional overlap between ANAC019, ANAC055, ANAC092, ATAF1 and NAP.

To validate the co-binding of different NAC TFs in close proximity through a palindromic binding site, we systematically screened the promoters of ANAC019, ANAC055 and ANAC092 DE genes for palindromic NAC binding sites using the motif CGTN7-8ACG (CGT spacer 7 or 8 nucleotides followed by ACG) (Jensen et al., 2010b; Tran et al., 2004). Only 9%, 15% and 12% respectively of the DE genes contained a palindromic NAC binding site, and for ANAC019 and ANAC092 this overlap was not significant. Based on the PBM binding data, only 2.2-2.8% of the ANAC019/ANAC055/ANAC092 P+COE target genes are bound by two adjacent NAC binding sites (spacer of 7 or 8 nucleotides). Considering all NAC TFs, only 3.7% of the P+COE target genes showed this co-binding pattern, corroborating that in most cases NAC binding and regulation is mediated through an individual binding site.

Overview of functional modules regulated by the different NAC TFs

Apart from comparing the overlap between P+COE genes and DE genes, we also studied the functional landscape of the different TFs using GO and functional modules. Enrichment analysis of P+COE target genes allowed to detect, per TF, the set of modules and associated functions showing significant overlap. The integration of this type of functional datasets can be used to transform the classical GRN into a TF-functional module network from which the diverse functionalities of TFs can be delineated (Supp. Table S6[a] and Figure 6.3). A first set of enriched modules is targeted by multiple TFs (5 or more) and is associated with different stress-related functional descriptions as well as signal transduction, transport and secondary metabolism (Figure 6.3A). The cooperative binding of the genes in these modules mainly comprises known stress-related factors including ANAC019, ANA055, ANAC092, and NAP. The observed association of ATAF1 with growth and development modules is also evident from the vegetative growth phenotypes of plants with perturbed ATAF1 levels (Jensen et al., 2013). A second set of modules is only targeted by a limited number of TFs and the genes in these modules cover a wider variety of biological processes and molecular functions (Figure 6.3B). Examples include previously described functions of SND1 and VND7 in cell wall biosynthesis and a role for NTL8 in embryo development (Zhong et al., 2007, 2010). Furthermore, we found that ANAC092 is linked with multiple transport and signal transduction related modules, which include known DE genes such as RNS1, ILL6 and MAPKKK19 (Balazadeh et al., 2010). Of novel relevance to the secondary cell wall-thickening regulator NST2, we highlight genes responding to nutrient starvation and water deficiency (module 10, Figure 6.3A, Supp. Table S6[a]), whereas novel target genes of VND7 include genes related to defense and programmed cell death (ie. MYB TFs), as well those earlier identified genes related to cell wall biogenesis (Zhong et al., 2010). Likewise, a large part of the verified target genes of secondary cell wall regulator SND1 include genes involved in cell wall biogenesis (ie. SND2 and SND3) and xylem development (ie. IRX genes). Furthermore, we highlight the overrepresentation of functional modules related to transport and senescence to include novel SND1 target genes (Supp. Table S6[a]). Finally, we observed a striking difference in the presence of genes with conserved motifs between the modules that are targeted by a big number of TFs (>5) and the modules that are targeted by a smaller number of TFs (arrow-head lines in Figure 6.3A vs Figure 6.3B), suggesting that the complexly regulated stress modules represent highly conserved regulatory interactions within plants. Obviously, the candidate target gene sets together with the associated functional modules offer a promising resource to unravel the functions of the different NAC genes in more detail.

Using native and synthetic promoter elements to validate PBM results

Binding of TFs to promoter elements is necessary to establish and maintain changes in gene expression levels of target genes (Xiao et al., 2013), and changing the TF-DNA affinity could dramatically affect the regulatory potential of the TF (Chavalit et al., 2013). Acknowledging this, we asked whether it would be possible to turn an element present in a target gene promoter identified from our studies into a synthetic promoter element that would both abrogate binding preferences of one TF and direct binding of another TF. Among our selected NAC TFs, binding site profiles of ANAC092 are most distantly related (ie. most divergent PWMs) to the NTL TFs (Figure 6.1A and Supp. Figure D.10) allowing us to test our hypothesis using these TFs. Firstly, in order to validate our 10-mer PBM data for ANAC092 and NTL6 using EMSA, we used a 30 bp oligonucleotide identical to the promoter of the ANAC092 target gene MYB90

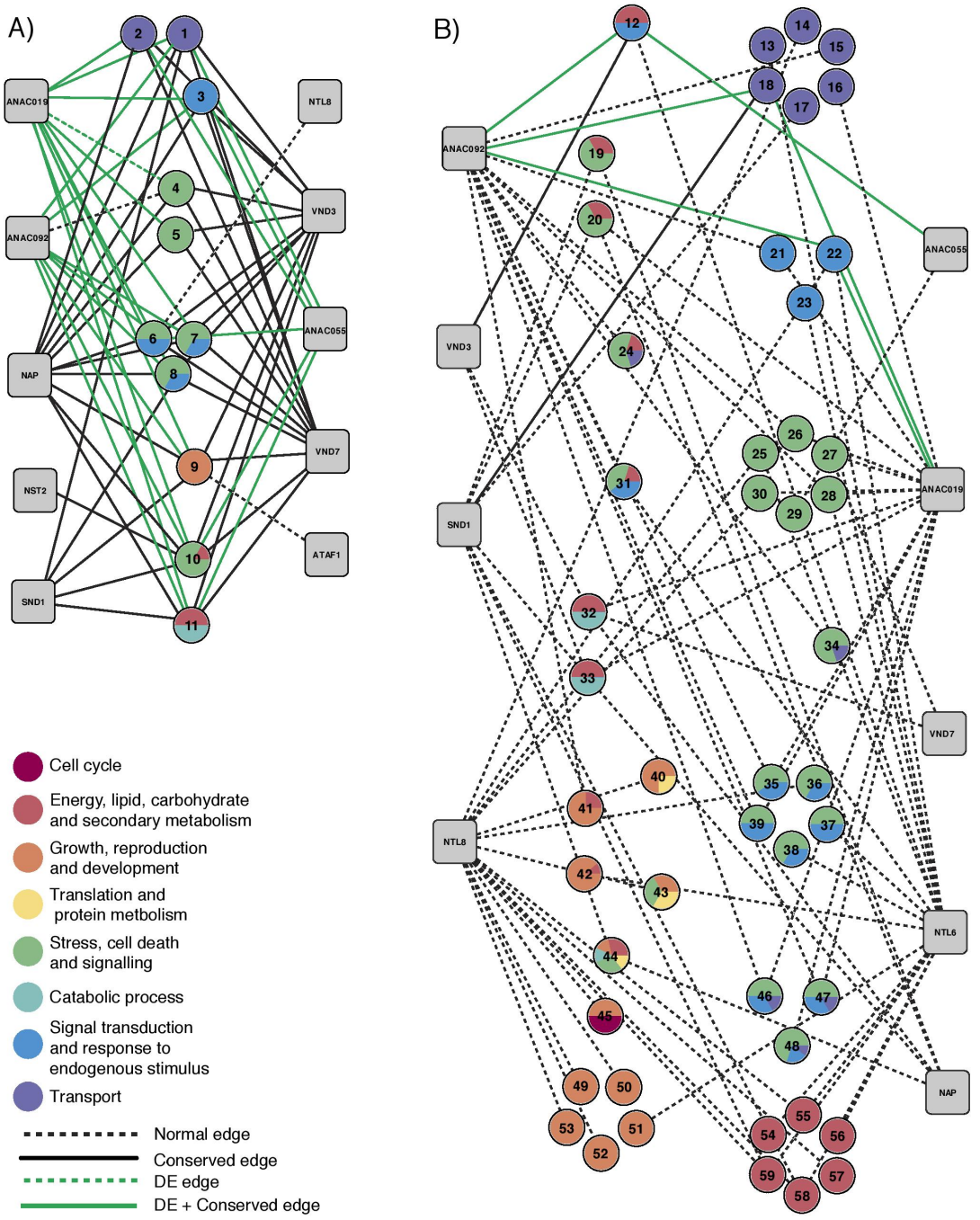

Figure 6.3: **Functional overview of modules enriched for TF-target genes.** *Panel A shows the TF module network for enriched modules that are shared between 5 or more TFs while panel B shows the TF module network for enriched modules that are shared between less than five TFs. Grey boxes represent TFs while colored circles refer to modules attributed to different functional categories. The numbers in the colored circles refer to the functional gene modules described in Supp. Table S6*[a]*. Whereas dotted black edges denote module enrichment for candidate PBM target genes, full black lines denote candidate target genes with a conserved motif and green dotted or full lines indicate that a DE gene for that TF is present in the module.*

involved in activating anthocyanin biosynthesis in response to C and N nutrient status (Gao et al., 2008). MYB90 was chosen as it is one of the 2 genes that passed all filtering tests for ANAC092 (The other one being AT3G02040)(conserved P+COE and DE). The 30 bp oligonucleotide contains a high ES k-mer (TACGTCA.C, 0.46) for ANAC092, yet scores very low for NTL6 (0.02, Supp. Figure D.10). In agreement with our PBM results, our EMSA result shows ANAC092 binding to the 30 bp promoter fragment

spanning the -361 bases upstream of the transcription start site of the MYB90 promoter, whereas no binding was detected using NTL6 (Figure 6.4). Next, using this oligonucleotide we aimed to turn it into a synthetic NTL6-binding promoter element (MYB90Synth) using the smallest Levenshtein distance, representing the minimum number of single-nucleotide changes required to change one sequence into another (Levenshtein, 1966). Using this modified 30 bp oligonucleotide, in which TACGTCA was mutated into a high-ES NTL6 target motif (0.47) TAaGTaA, we observed a lowered affinity of ANAC092 for the MYB90Synth element. This is in accordance with the low PBM derived ES value of ANAC092 for TAAGTAA motifs (0.27, Supp. Figure D.10). Most importantly, NTL6 was observed to bind to the MYB90Synth oligonucleotide with high affinity. As a positive control all proteins were tested for binding to the palindromic NAC-BS consensus (Olsen et al., 2005). Here, ANAC092 showed the strongest affinity. We note that we repeatedly observed two ANAC092-palNACBS and NTL6-MYB90Synth complexes. This could potentially arise from binding of two individual dimers to the DNA fragment, also observed in the co-crystal structure of ANAC019-PalNAC BS (Welner et al., 2012). Taken together, we use a 30 bp oligonucleotide identical to the promoter element of the ANAC092 target gene MYB90 to validate our PBM data for ANAC092. Also, we report a 2-nucleotide substitution of the ANAC092 binding site lowering the affinity of ANAC092 for this synthetic promoter element and turning it into an NTL6-binding element.

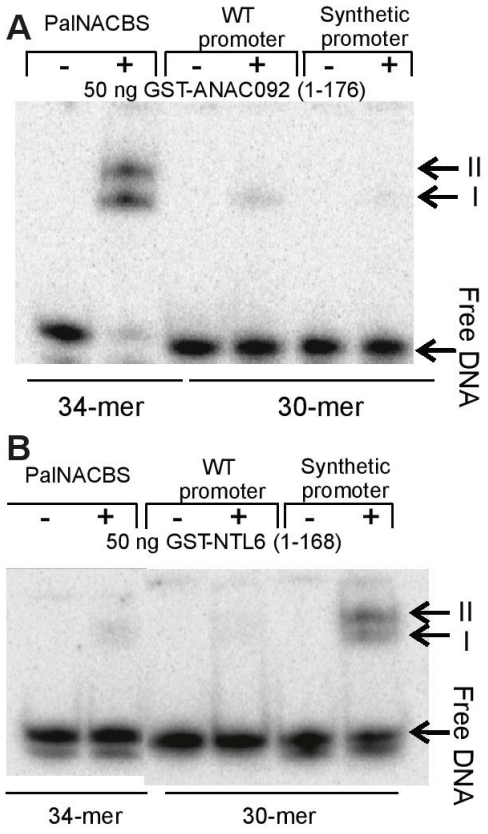

Figure 6.4: **Design of a NTL6 binding site from an ANAC092 promoter.** *ANAC092 (panel A) and NTL6 (panel B) were tested by EMSAs for binding to a known and validated palindromic NAC-BS consensus (palNACBS), a fragment of an identified ANAC092 target promoter (At1g66390; MYB90) (WT promoter) containing the TACGTCA k-mer and a Synthetic promoter where the same k-mer was mutated to TAaGTaA to mimic an NTL6 binding site.*

Using co-expression analysis to uncover the regulatory potential of ANAC092

Co-expression occurs among TFs and target genes (Truman and Glazebrook, 2012). To validate our list of putative target genes for our candidate NAC TFs, we hypothesized that genes controlled by individual

NAC members should be (i) co-expressed during environmental cues known to affect NAC gene expression, and (ii) have one or more NAC consensus binding site(s) in their promoter. For this purpose we performed data-mining on > 3.000 ATH1 microarray samples from wild-type Col-0 plants, deposited at Genevestigator (Hruz et al., 2008) and, using a stringent (> 2-fold regulation, P < 0.05) selection criterion for ANAC092 transcript level perturbations, we found 705 microarray datasets representing 160 perturbations (Figure 6.5). Using these data, we analyzed the co-expression of ANAC092 and the set of 107 putative target genes. From this analysis we identified two major clusters of genes; those with a positive correlation with ANAC092, and those with a negative correlation expression pattern compared to ANAC092. Interestingly, target genes up-regulated in anac092 mutant plants almost perfectly match the genes that are down-regulated when ANAC092 is induced. Vice versa, genes down-regulated in anac092 mutants, show almost perfect co-expression with ANAC092. This indicates that ANAC092 could be both a direct activator and a direct repressor. Moreover, the regulatory potential of ANAC092 is maintained during multiple environmental stresses, and not only during the anac092 vs Col-0 control condition samples reported by Balazadeh and colleagues that we used in this analysis. The strong ANAC092 expression perturbations during environmental stresses observed from our analysis is in agreement with the recent results published by Patil et al. highlighting ANAC092-mediated stress tolerance (Patil et al., 2013).This result suggests ANAC092 as a TF associated with both positive and negative effects on transcription of a large set of stress-related genes.

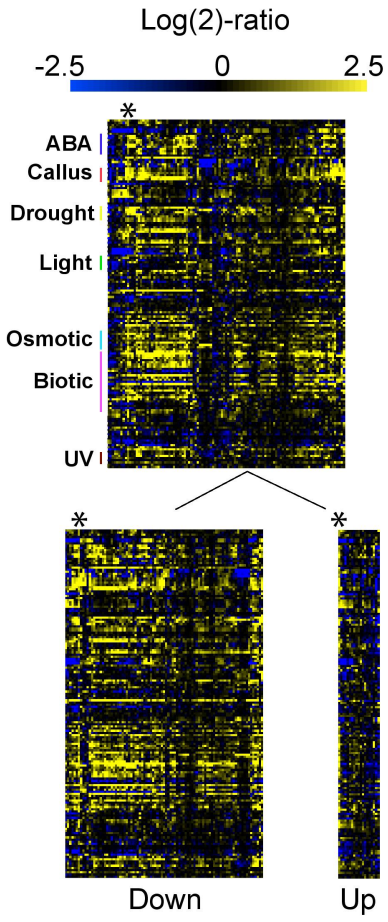

Figure 6.5: **The regulatory potential of ANAC092 is maintained during multiple environmental stresses.** *Top heatmap displays 107 genes differentially regulated in anac092 plants compared to Col-0 wild-type plants, all having ANAC092-BS in their 1 kb promoter. Only conditions affecting ANAC092 expression were included (> 2-fold regulation, P < 0.05, = 160 perturbations, 705 microarrays). Below, 'Down' denotes the 89 genes down-regulated in anac092 mutant plants compared to Col-0 wild-type plants and 'Up' denotes the 18 genes up-regulated in anac092 mutant plants compared to Col-0 wild-type plants. * indicates position of ANAC092. To the left selected conditions perturbing most target genes are high-lighted.*

6.4 Discussion

A major challenge for predicting gene expression is the accurate characterization and design of genetic circuits that regulate single or multiple genes in response to specific environmental, developmental and physiological cues. In the age of synthetic biology, characterization of TF binding preferences and target gene identification offer major advantages towards engineering genetic circuits for optimal fitness in plant responses towards environmental stresses. However, in order to fully understand the regulatory capabilities of any TF, we need to characterize its DNA binding specificities with the highest resolution possible in order to minimize erroneous TF-promoter associations resulting in misleading GRNs. As previously described the CGT[A/G] motif has been identified as the core binding site of stress-inducible NAC TFs (Puranik et al., 2012; Olsen et al., 2005; Tran et al., 2004). However, this motif present in the DNA binding sites of Cluster 1, is also a core binding site for NAC TFs involved in development and secondary wall synthesis (Zhong et al., 2010). The binding sites of Cluster 1 proteins show differences in the flanking regions that mark divergence in the functionality of this cluster's members. These binding differences may be explained by small variations in the DNA-contacting amino acids residues (Supp. Figure D.1B) which, according to the crystallographic model of the ANAC019-DNA complex, are close to the DNA (36). These regions contain both the conserved Arg-88, essential for binding, and the conserved beta strand protruding into the major groove of DNA. ANAC019, ANAC055, NAP and ATAF1, which have similar binding sites, constitute a sub-group based on the sequence regions close to DNA (Supp. Figure D.1B), suggesting that these regions influence DNA-binding specificity. These closely related NAC TFs, however, also show different preferences for A/G of the core binding site which is not easily explained from the sequence alignment. SND1, NST2, VND3 and VND7, involved in secondary wall synthesis (Zhong et al., 2007, 2010) cluster together (Supp. Figure D.1A) (Jensen et al., 2010a) yet the DNA binding specificities of SND1 and NST2 are closer to those of ANAC092 than those of VND3 and VND7. This is unexpected considering that the expected DNA contacting residues for all these TFs are identical. Further analysis will reveal if substitution of single amino acid residue, such as the change of a conserved basic residue to a glutamine (position 127 of VND3), possibly in contact with DNA (Welner et al., 2012), may affect DNA binding specificity. Surprisingly, and in contrast to reports showing that binding of NTL6 to the Pathogenesis-Related (PR) genes depends on the NAC-BS core (Seo et al., 2010) NTL6 and NTL8 do not recognize sequences with the NAC-BS core. We did not observe any overlap between DNA specificities of Cluster 1 and Cluster 3, leading to the hypothesis that these proteins, while members of the same general TF family, are functionally divergent from their paralogues. As seen in bHLH and homeodomain proteins, few amino acids can play a critical role in the definition of DNA specificities for single TFs (Grove et al., 2009; Masi et al., 2011; Berger et al., 2008; Noyes et al., 2008). Indeed, as few as 5 positions show differences between NTL6, NTL8 and the remaining NAC proteins. These are at positions NTL6 74 (Y->F), 102 (R->K), 116 (R->K), 121 (H->Y) and 130 (R->K), with 121 (H->Y) representing the chemically most significant change (Supp. Figure D.1B). While positions 116, 121 and 130 are close to DNA, we can not rule out that positions 74, 102 and additional regions may also influence specificity of these NAC proteins. Although single amino acid residues may dictate DNA binding specificity, conformational changes of for example the DNA-contacting NAC loops (Welner et al., 2012) may also influence DNA-binding specificity (Nakagawa et al., 2013). Clearly, further structural analyses are needed to identify the fine molecular determinants of NAC-DNA binding specificity and affinity even though these presented data can be sufficient to estimate DNA specificities for NAC proteins in terms of Cluster 1, 2 or 3. NAC binding sequences selected in some other studies are palindromic sets of two adjacent sites reflecting that NAC TFs forms and bind DNA as dimers (Olsen et al., 2005; Ernst et al., 2004). However, as seen in this study, single NAC binding sites (NACBSs) can be sufficient for NAC promoter binding. This effect has been also shown to be true from the analysis of ANAC072/019/055 binding to the ERD1 promoter (Tran et al., 2004), ANAC096 binding to the RD29A promoter (Xu et al., 2013) and ATAF1 binding to the 9-cis-epoxycarotenoid dioxygenase (NCED3) promoter (Jensen et al., 2013). In fact, the single ATAF1 binding site identified by PBM analysis was used to identify NCED3 as a direct ATAF1 target gene (Jensen et al., 2013). The fact that a single NAC-BS is sufficient for NAC binding is also supported by in vitro analysis showing that although NAC dimer-

ization is needed for detectable DNA binding, only a single NACBS is needed for binding (Olsen et al., 2005). Furthermore, a recent DNase I footprint of ANAC019 and the palindromic PalNAC BS showed asymmetric protection (i.e. saturation) of the two single binding sites in the palindrome (Welner et al., 2012). Despite this, heterodimerization of NAC TFs (Ernst et al., 2004) may expand the DNA binding specificity spectrum in vitro, as suggested for the bHLH TFs (Grove et al., 2009; Masi et al., 2011). This variability between single or double binding sites can bring yet another level of genetic regulation in NAC dependent stress response in A. thaliana. It is plausible that promoters showing palindromic dimer sites could be differentially regulated by combinations of NAC homo- and hetero-dimers thus expanding on the range of stress signals recognised. To better understand this process a large-scale NAC dimerization screen followed by NAC dimer DNA binding studies would be required. A major challenge for the characterisation of GRNs using high-throughput TF binding data is to properly translate DNA specificities in meaningful lists of potentially regulated genes. Transcription-factor binding affinities determined in vitro have been shown to quantitatively predict the output of complex target promoters (Grove et al., 2009; Rajkumar et al., 2013) yet, the risk of contaminating the target detection analysis with false positives and false negatives is a real threat. By integrating different layers of evidence, such as co-expression information, differential expression in mutant plants, motif conservation and functional gene modules, we were able to obtain meaningful and accurate functional predictions for the studied TFs, including the verification of 31 previously identified NAC TF target genes. This emphasizes the applicability of our workflow using PBM and functional modules to uncover NAC TF target genes. The improved specificity obtained through the integration of complementary functional genomics data sets is in agreement with recent observations from genome-wide chromatin immunoprecipitation experiments, where typically only a minor fraction of bound regions corresponds with bona fide regulated target genes (Ferrier et al., 2011). As a consequence, also for ChIP-chip and ChIP-Seq experiments, detailed motif and expression information are required to define an accurate set of functional in vivo target genes.

Due to the fact that NAC TFs have a large potential in plant engineering and production of more robust economically important crops (Puranik et al., 2012; Jeong et al., 2010; Uauy et al., 2006) detailed knowledge about TF-DNA interfaces and target gene perturbations become crucial knowledge for the exploitation of rationally designed GRNs for improved stress tolerance and other economically important traits. As shown here, the minimal changes in NACBS required to engineer, and potentially redirect, single TF GRNs can hold interesting solutions for future breeding and genome editing projects. For instance, identification of SNPs in TF-BSs of putative orthologous gene promoters related to certain morphological traits, can be harnessed for improving or abrogating TF DNA-binding affinity and thereby transcriptional output. Further away, specific Cas9-based genome editing (Li et al., 2013) could be applied to balance transcriptional output to specific environmental conditions using a one-TF-many-target-genes approach. Using the knowledge and information obtained from this study, we could envision modifying specific NAC binding sites, with great accuracy, to rewire GRNs with the final aim at improving or generate de novo stress responses in A. thaliana and other plants. This novel GRN design could lead to the generation of drought or other climatic-stress resistant crops, that could be designed to contrast desertification and the resulting loss in food production.

6.5 Material and Methods

Sequence analysis of the NAC family

Multiple alignments, phylogenetic tree and the sequence similarity matrix of the DNA binding domains of all proteins were generated using ClustalW (Thompson et al., 1994) and drawn using MatLab (Mathworks, Natick, Ma). BoxShade (`http://www.ch.embnet.org/software/BOX_form.html`) was then used for producing graphical representations of the multiple alignment.

Cloning and recombinant protein production

Oligonucleotides, restriction enzymes and vectors used for cloning of GST-tagged proteins analyzed in this study are listed in Supp. Table S1[a]. Cloning and production of several of the GST-recombinant proteins have already been described (Jensen et al., 2010b). In addition, cDNA clones acquired from

the Arabidopsis Biological Resource Center were amplified by PCR to obtain the region encoding the NAC domain of ANAC055, ANAC072, NAP, and NST2, full-length ANAC092 and the DBD of WRKY1 (Duan et al., 2007). Finally, the NAC domain encoding region of SND1 was synthesized (Eurofins MWG Operon) and used for PCR. The PCR products were inserted into the vectors as shown (Supp. Table S1[a]). For the zinc-finger TFs VOZ2 and WRKY1 50 µM zinc acetate was added to the growth medium. After induction, cells were harvested and sonicated and GST-tagged proteins were purified on glutathione-Sepharose 4B resin (GE Healthcare) as described (Jensen et al., 2010b). Purified recombinant proteins were analysed by SDS-PAGE and absorbance scans. Protein concentrations were estimated from A280 measurements. By using this procedure highly pure GST-tagged recombinant proteins was produced and no further purification was needed. A subset of the NAC proteins described above was also produced by PURExpress In Vitro Protein Synthesis transcription/translation kits (New England Biolabs) according to the manufacturer's instructions. The concentration of purified GST-tagged proteins was quantified by Western blotting using anti-GST antibody (Invitrogen) by comparison to a dilution series of recombinant glutathione-S-transferase (Sigma).

Protein Binding Microarray experiments and data analysis

Oligonucleotide arrays were made double-stranded by primer extension and PBM experiments were performed as described previously using custom 'all 10-mer' array design using the Agilent '4x180K' array format (Agilent Technologies, Inc) (Berger and Bulyk, 2009). All PBM experiments were performed in duplicate at a final protein concentration between 200-500 nM. Microarray scanning, spot quantification, data filtering, normalization and primary analysis were performed as previously described (Grove et al., 2009; Berger and Bulyk, 2009). Significant k-mers were selected by identification of all words showing an Enrichment Score (ES) equal to or greater than 0.4 for at least one studied TF. Contrary to other similar studies, we here retrieved all gapped or un-gapped 8-mers resulting in a final set of 4821 significant k-mers (Supp. Table S2[a]). 'Core words' used for boxplots were identified by a combination of a statistical method ('preferred k-mers') from Jiang et al. (2013) and a visual approach of the previously described matrix. This resulted in the identification of 130 core words of length 6 or 7 that are able to describe the exact specificities of each TF. PWM logos were drawn using the enoLogos engine (Workman et al., 2005). Heatmap figures were made using Matlab (Figure 6.1) and Genesis (Sturn et al., 2002) (Figure D.6 and D.8).

Detection of target genes, integration of co-expression information, gene function enrichment analysis and motif conservation

Target genes were predicted by initially determining for each TF a set of high scoring seed 8-mers (ES > 0.45) and mapping these to the promoters of all Arabidopsis genes (TAIR10). A promoter was defined as the 1000bp upstream of a gene or a shorter region if the adjacent upstream gene is located within a distance smaller than 1kb. To refine the set of PBM predicted (P) target genes, expression data was integrated to define target genes that are also co-expressing with other predicted target genes (P+COE). Based on 14 Affymetrix ATH1 microarray expression compendia delineated by De Bodt and colleagues (De Bodt et al., 2010), we defined for each gene a co-expression cluster by selecting the top-100 co-expressed genes based on Pearson correlation coefficients. A target gene was retained as P+COE target if its co-expression cluster was enriched for target genes of the same TF (hypergeometric distribution, p-value < 0.05). To evaluate the evolutionary conservation of individual k-mer instances, a multi-species phylogenetic footprinting approach was applied. For each Arabidopsis target gene the orthologous genes from 11 other dicot species (Malus domestica, Fragaria vesca, Manihot esculenta, Medicago truncatula, Carica papaya, Glycine max, Lotus japonica, Ricinus communis, Theobroma cacao, Populus trichocarpa and Vitis vinifera; source PLAZA 2.5 (Van Bel et al., 2012)) were retrieved using the PLAZA Integrative Orthology method. First, the 1 kb orthologous promoter sequences were aligned to the query promoter using the Sigma alignment tool (Siddharthan, 2006). Next, all pairwise alignments for each query gene were aggregated on the query sequence generating a multi-species conservation plot that shows for each nucleotide of the investigated region how many species support this nucleotide through pairwise footprints. All footprints for each level of conservation were extracted from the multi-species conservation

plot. Finally, the significance of the observed multi-species footprints, per Arabidopsis target gene, was determined by randomly sampling 1000 non-orthologous gene sets, maintaining the gene and species composition as observed in the real orthologous data set, and scoring in how many random gene sets a footprint with a similar or better multi-species conservation was found. Footprints with a false discovery rate (FDR) < 5% were used to identify conserved PBM motif instances. The significance of the overlap was calculated using the hypergeometric distribution (p-value < 0.05). Fold enrichment was calculated using the formula (k/n)/(K/N) where k is the number of recovered differentially expressed (DE) genes within the predicted target genes, n is the number of predicted target genes, K is the number of DE genes and N is the number of genes in the genome.

Construction and biological evaluation of the NAC gene regulatory network

In order to construct a gene regulatory network all P+COE target genes of all TFs were used. In order to evaluate function of these P+COE target genes we determined, per TF, enriched functional modules for all target genes. The associated GO terms of each enriched functional module were mapped to their parental GO terms, GO slim terms were selected and these GO slim terms were grouped into 10 functional categories. In order to obtain functional categories all GO slim terms were clustered on their enrichment in functional modules and groups of GO slim terms that clustered together were isolated as categories (tropism: tropism; cellular homeostasis: cellular homeostasis; stress cell death and signalling: cell-cell signalling, regulation of gene expression, epigenetic, response to stress, response to biotic stimulus, response to abiotic stimulus, death, cell death, response to external stimulus, cell communication, response to extracellular stimulus; transport: transport; signal transduction and response to endogenous stimulus: signal transduction, response to endogenous stimulus; catabolic process: catabolic process; energy lipid carbohydrate and secondary metabolism : generation of precursor metabolites and energy, photosynthesis, lipid metabolic process, carbohydrate metabolic process, secondary metabolic process; cell cycle: cell cycle; translation and protein metabolism: translation, protein metabolic process; growth reproduction and development: reproduction, multicellular organismal development, anatomical structure morphogenesis, embryo development, post-embryonic development, fruit ripening, abscission, pollination, pollen-pistil interaction, flower development, cellular component organization, cell growth, cell differentiation, growth). The network depicted in Figure 6.3 was constructed using the Node Chart Plugin for Cytoscape 2.8.2 (Smoot et al., 2011). Only modules with enriched GO slim terms are depicted. This plugin allows for a module node to be used as a pie chart and through color-coding for the different functional categories, this allowed visualizing the predicted functional role of each modules associated to each TF.

EMSA (electrophoretic mobility-shift assay)

Purified GST-ANAC092(1-176) and GST-NTL6(1-168) were tested for functionality in EMSAs using a 32P-labelled double stranded oligonucleotide of the palindromic NAC binding site (PalNACBS; Supp. Table S1[a]; (Olsen et al., 2005)), the wild type MYB90 promoter fragment (Supp. Table S1[a]; WT promoter) and the synthetic promoter fragment (Supp. Table S1[a]; Synthetic promoter). EMSAs were performed as described previously (Olsen et al., 2005; Welner et al., 2012). The oligonucleotides used in EMSA were initially pairwise annealed in 100 ţl (20 mM Tris-HCl, pH 8.0, 20 mM MgCl2) by heating the solution to 95řC for 5 min followed by slowly cooling to room temperature, which normally takes hours. Small aliquots were then taken out when needed for labeling, purification and finally EMSA. The DNA concentration in EMSA were kept at 75 pM, which is roughly 1000-fold lower than the estimated Kd for the interaction (Welner et al., 2012).

Co-expression analyses

To investigate if genes differentially expressed in anac092 mutant compared to wild-type Col-0 plants maintain expression perturbations during environmental conditions known to affect ANAC092 levels, we data-mined > 3.000 Col-0 wild-type ATH1 microarray samples from the Genevestigator data repository (Hruz et al., 2008). Using a stringent (> 2-fold regulation, P < 0.05) selection criterion for ANAC092

transcript level perturbations, we found 705 microarray slides from 160 perturbations. This data set was used to perform hierarchal clustering (euclidian distance) of ANAC092 and 107 putative target genes differentially expressed in anac092 mutant compared to wild-type Col-0 plants, all containing ANAC092 BS in their 1 kb promoter.

Accession numbers

Sequence data from this article can be found in TAIR (The Arabidopsis Information Resource) and EMBL (European Molecular Biology Laboratory) data libraries using the nomenclature names, synonyms and accession numbers in Supp. Table S1[a].

6.6 Acknowledgements

We are grateful to Trevor Siggers for his feedback and input on the manuscript.

CHAPTER 7

General Conclusions and Perspectives

7.1 CNS detection and usage for TFBS prediction

Throughout this thesis different approaches were developed for the detection of CNSs. These methods can be split up in alignment based and alignment free. For the CMM approach and alignment based approach developed in Chapter 3 it was shown that both methods have comparable performance and are complementary to each other (Figure A.3). It is however generally considered that alignment free approaches are more sensitive in detecting TFBSs. In Chapter 5 the alignment free approach of BLSspeller was shown to detect more TFBSs associated with TF ChIP-seq peaks than the alignment based approach. This higher sensitivity can be partly explained by the fact that removing the constraint of alignment allows the approach to detect binding site conservation where turnover has occurred. The downside to this flexibility is reflected in the spurious detection of conserved binding sites, because the conservation context is less stringent. Alignment based approaches on the other hand suffer from the fact that many motifs are no longer alignable across very large phylogenetic distances as is the case for a large part of currently sequenced plant genomes. Alignment based approaches are able to pick up longer stretches of conservation. These longer conserved stretches were also shown to contain small non-coding RNAs and as such capture a different spectrum of the conserved non-coding regulatory landscape (Haudry et al., 2013). Another limitation of the alignment free CMM approach is that it relies on known TFBS specificity data as input for the detection of conserved TFBS. This limits the applicability of this approach to TFs for which this information is available. The alignment free approach of the BLSspeller tool elegantly overcomes this limitation by exhaustively enumerating over all possible words in the promoter of the investigated genes. This allows the tool to detect conserved TFBS with high sensitivity and unbiased towards existing data.

The performance of CNS for predicting TFBS was discussed in the different chapters of this thesis. A recurring observation throughout this work is the fact that TFBS predictions based on conservation across larger phylogenetic distances show high specificity but lower recovery. This effect is most visible in (Figure A.6). In this figure a comparison of enrichment towards in vivo functional binding sites (Chapter 3) is made showing that enrichment is higher for conserved binding sites compared to binding sites in open chromatin and binding sites predicted without a functional filter. The recovery of these binding sites shows an opposite trend (data not shown). This leads to the conclusion that conservation is a great functional filter but also has limitations. Rewiring of the gene regulatory network during evolution, caused by small and large scale duplications and the associated sub and neo-functionalization hamper higher recovery rates of TFBSs. Another limitation is the fact that transcriptional regulation of specific biological processes such as adaptive responses to pathogens or environmental changes are simply not conserved.

A similar trend is also visible when comparing TFBS prediction between different methods relying on phylogenetic conservation (Figure 3.2). A larger evolutionary distance considered leads to higher specificity and lower recovery of TFBS. An independent comparison of CNS prediction methods replicated this observation through a comparison of the recovery of and enrichment towards Dnase hypersensitive sites and a manually curated set of CNSs. In this study the CNS detection method presented in Chapter

3 showed a moderate recovery but highest specificity for both the high quality CNS and open chromatin datasets (Burgess, Xu and Freeling, 2015). From this comparison it can be concluded that conservation based methods are subject to a tradeoff between either the specificity provided by large divergence times and the sensitivity provided by the comparison of more closely related species. This poses a challenging problem in plants as the whole genome alignment approaches that are used to create high sensitivity across smaller phylogenetic distances cannot be applied to compare evolutionary more distant genomes. An approach using whole genome alignments across large phylogenetic distances in plants performed very poorly in the same benchmark (Figure 3.2) (Hupalo and Kern, 2013). A potential solution might be found in the use of the BLSspeller tool that can account for both large and small phylogenetic distances.

In line with higher specificity, the yield of detected CNSs is much lower over larger phylogenetic distances. This is evident from comparing the number of CNSs detected by the BLSspeller tool used in monocots (Chapter 5) and the analyses performed in dicots (Chapter 3 & 4) but also in comparison to the work in brassicaceae (Haudry et al., 2013). The conservation that is detected is always with regard to the comparator species used. Different sets of comparator species also capture different spectra of biology, whereas conserved TFBSs detected between closely related species will reflect more lineage or species specific biology, conserved TFBSs across larger distances will represent more core functions. An illustration of this link can be found in Chapter 4, where a small gene regulatory network of conserved E2F interactions, involved in cell cycle regulation, was identified. The most deeply conserved interactions were conserved as far as green algae (Figure 4.3).

7.2 Application of CNSs and Conserved TFBSs

Apart from studying transcriptional regulation through CNSs and determining correct target genes for a TF, investigating the binding pattern of a TF is a powerful method to gain insights in its biology. The function of a TF is largely determined by the processes in which the target genes it controls are involved. The following paragraphs are meant as illustrations and discuss how utilizing different strategies for incorporating CNSs can help to generate leads in unraveling molecular pathways, functionally characterizing TFs and the interpretation of genome-wide association studies (GWAS). 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.

Functional analysis of the Arabidopsis TETRASPANIN gene family (Wang et al., 2015a)

TETRASPANIN (TET) genes encode conserved integral membrane proteins that are known in animals to function in cellular communication during gamete fusion, immunity reaction, and pathogen recognition. In plants, functional information is limited to one of the 17 members of the Arabidopsis TET gene family and to expression data in reproductive stages. Here, the promoter activity of all 17 Arabidopsis TET genes was investigated by reporter lines throughout the life cycle, which predicted functional divergence in the paralogous genes per clade. Mutational analysis showed a role for TET13 in primary root growth and lateral root development and redundant roles for TET5 and TET6 in leaf and root growth through negative regulation of cell proliferation.

In order to further identify environmental and developmental stimuli that might influence TET gene expression cis-regulatory elements in the promoters of TET genes were also explored. To reduce the number of false positives returned by simply mapping known cis-regulatory elements to the 2-kb promoters of the different TET genes, also co-regulatory genes, evolutionary sequence conservation, and information about open chromatin regions were integrated. For each TET gene a co-expression cluster was created using transcriptome samples of tissue and perturbation experiments in which the gene showed a strong response. This was followed by motif enrichment using evolutionary conservation and presence in open chromatin as filters. In total, 71 cis-regulatory elements were identified in the TET promoters and their respective co-regulated genes. Linking TET differential gene expression upon a selection of perturbations with the biological processes in which the known cis-regulatory elements are involved was used to verify the cis-regulatory elements identified in the promoters of the TET genes. The conditions in which the identified cis-regulatory elements were known to play crucial roles had a good

correlation with the conditions in which nine TET genes were differentially expressed. In an attempt to further improve our understanding of the transcriptional regulation of TETs and their positions in molecular pathways, a TF-TET gene regulatory network was built by combining different regulatory data sets (conserved TF-binding sites, ChIP TF-binding and TF expression perturbation information) into a gene regulatory network. Expression measurement data after TF perturbation for nine TFs identified as targeting TET genes confirmed these predictions by significant differential expression of the target genes and identify the respective TET genes as novel components of specific developmental or physiological pathways (Fig. 7.1).

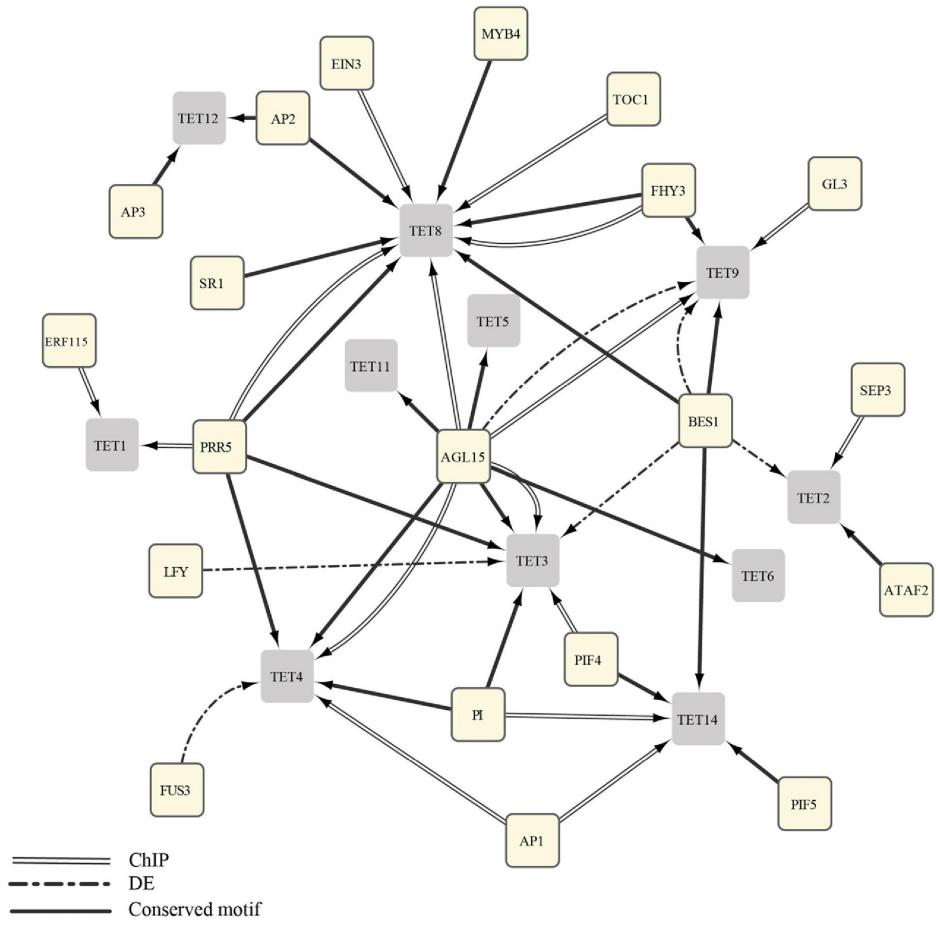

Figure 7.1: **Gene regulatory network for TET genes.** . Source: Wang et al, 2015

The cis-regulatory elements together with transcription factor-binding data provided molecular insight into the sites, conditions, and perturbations that affect TET gene expression and positioned the TET genes in different molecular pathways (Wang et al, 2015a).

Improving energy use efficiency and drought tolerance in Canola (Verkest et al., 2015).

Increasing both the yield potential and stability of crops at the same time is a longstanding challenge in breeding strategies and is mostly approached from a direct genetic point of view, while the utility of epigenetics in complex traits is still unclear. A better understanding of the status of the epigenome and its contribution to agronomic performance would help in developing novel breeding programs and further improve the performance of crops. Starting from isogenic canola (Brassica napus) lines, epilines were generated that display increased energy use efficiency, drought tolerance and nitrogen use efficiency. Transcriptome analysis of the epilines and a line selected for its energy use efficiency solely

revealed common differentially expressed genes related to the onset of stress tolerance-regulating signaling events. Genes related to responses to salt, osmotic, abscisic acid, and drought treatments were specifically differentially expressed in the drought-tolerant epilines. The status of the epigenome, scored as differential histone 3 lysine-4 trimethylation, further supported the phenotype by targeting drought-responsive genes and facilitating the transcription of the differentially expressed (DE) genes.

The enrichment of both kinases and TFs was found in the differentially expressed and differential presence histone 3 lysine-4 trimethylated (DHM) regions hinted at a regulatory role in conferring the improved drought stress tolerance phenotypes. To further investigate this putative regulatory role motif enrichment was performed on the conserved non-coding regions associated to the DE and DHM genes. A significant overrepresentation for W-box, DROUGHT-RESPONSIVE (DRE) elements, and ABSCISIC ACID-RESPONSIVE (ABRE) motifs was observed (Fig. 7.2). Screening of the conserved regions 1,000-bp upstream of all DE and DHM genes with these motifs recovered 23.7% of the DE genes and 18.9% of DE and DHM genes together. This screening also identified putative WRKY, AP2-EREBP DEHYDRATION-RESPONSIVE ELEMENT BINDING (DREB) family, and basic Leu zipper TFs downstream of drought-responsive genes. Finally, DE and DHM genes containing W-box, DRE sites, and ABRE motifs retained TF, kinase, and drought stress-associated GO term enrichments confirming roles for WRKY, AP2-EREBP DREB, and basic Leu zipper TFs as master regulators in the energy use efficiency/PEG selection-improved drought stress tolerance signaling.

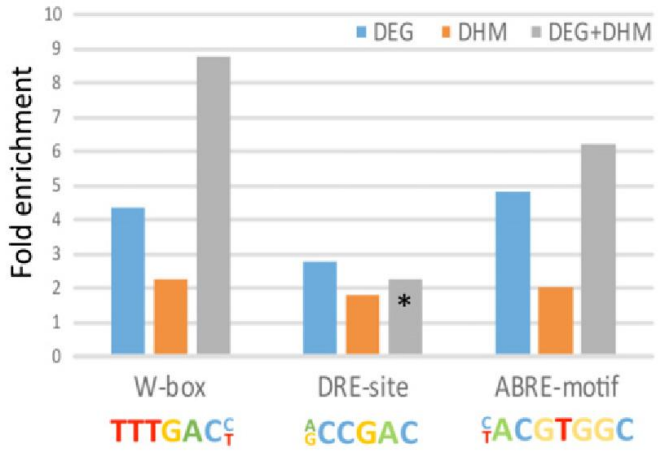

Figure 7.2: **Cis-regulatory element enrichment analysis of DE genes enriched in H3K4me3.** *Enrichment of W-box, DRE site, and ABRE motif occurrence in conserved noncoding sequences upstream of DEGs and/or DHM genes. Sequence logos used in mapping are shown below the graph. *, Not significant ($P > 0.05$).* Source: Verkest et al, 2015

Prioritization of TFs influencing root growth under water deficits in soybean (Lin et al., unpublished)

The reproductive success and survival of plants is determined by their ability to locate and extract water in drying soil. An important adaptation to water-limited conditions shared by several agronomic species, including soybean, is the maintenance of primary root elongation at low water potentials. To reveal possible underlying molecular mechanisms in this process, an integrative methodology was applied. Analysis of soybean root elongation zone spatial and temporal gene expression profiles in response to water deficit conditions, led to the identification of 279 differentially expressed genes. The promoter regions of these genes were screened for over-represented motifs and a total of 89 cis-regulatory elements were discovered. Since motif-finding tools may generate many false positive candidates due to the fact that TF binding sites are often short and certain degeneracy occurs in the binding motif, these cis-regulatory elements were integrated with CNS data for soybean. This integration identified 49 conserved cis-regulatory elements. GO enrichment of the genome-wide conserved target genes of these elements revealed enrichment for abiotic stress response, hormone signaling and development. A number of these

conserved cis-regulatory elements were validated using a dual luciferase assay to confirm the inducibility of selected motifs/promoters by drought-related treatments and were shown to drive gene expression in water limiting conditions.

Similarly as the workflow presented in Chapter 6, these three examples illustrate how the usage of conservation as a highly specific filter can help in elucidating molecular pathways and functionally characterizing TFs and their target genes.

Interpretation of noncoding variants associated with traits.

Recently, genotypic data across many accessions as well as the associated phenotypic data (such as trait related and gene expression measurements) have become available and large scale analyses have been applied in order to uncover the genomic loci or SNPs that underlie these different traits (Alonso-Blanco et al., 2016; Gan et al., 2011a; Atwell et al., 2010). Genome-wide association is frequently utilized approach that examines the frequencies of different genotypes with respect to the studied phenotypes and produces lists of significant associations. Most of the genes affected by such a regulatory change have been found to encode transcription factors that regulate the underlying processes. Because of the large and pleiotropic impact of mutations in the DNA binding domain of TFs (trans) a lot of these regulatory changes occur as cis-regulatory mutations. The cis-regulatory mutations usually reside upstream of the targeted gene and often affect spatial and/or temporal expression. The relevance of TFBS in maintaining basal cellular processes is illustrated by the large number of single nucleotide changes in TFBS associated with human diseases and cancer (Maurano et al., 2012). A similar trend is observed in plants where loci with roles in crop domestication involve changes in transcriptional activity. A recent study reported that 67% of genes deemed domestication and/or diversification loci, to be TFs or transcription co-regulators (Meyer and Purugganan, 2013). Loci associated with modern breeding that contribute to crop yield, quality or adaptation to novel environmental conditions also correspond to TFs. 43% of the causative mutations associated with these genes were found to be in cis-regulatory regions (Meyer and Purugganan, 2013). Examples of quantitative trait loci involved in domestication or crop improvement are becoming more prevalent. Kloosterman and co-workers detected a major (QTL) involved in potato tuberization in northern latitudes (Kloosterman et al., 2013). The bitter taste in cucurbits was found to be caused by the Bitter fruit TFs involved in cucurbitacin synthesis (Shang et al., 2014). An example of a cis-regulatory mutation was observed in rice were a causative mutation was pinpointed to the promoter of the GRAIN WIDTH 7 gene which improved both grain appearance and quality without generating a yield penalty (Wang et al., 2015b).

Genome-wide association studies often have a limited ability to pinpoint the causal SNP because of linkage disequilibrium (LD) and offer little insight into the biological mechanisms that mediate the proposed associations. Due to the growing implications that DNA variation within regulatory sequences is causative in the improvement of traits and establishment of new developmental features, these issues could be partially solved through integration with functionally related properties of regulatory sequences. Recent studies have shown that sets of eQTLs are enriched for CNSs, open chromatin regions, TF ChIP-seq bound regions and gene expression promoting histone marks compared to non-regulatory background SNPs (Lappalainen et al., 2013) (Osbourne et al., Unpublished). By making use of this principle Weirauch and colleagues found a causal mechanism for a previously detected eQTL of the AT5G47250 gene (Gan et al., 2011a). The identified SNP leads to a loss of a potential binding site for the VNI2 TF, predicted through PBM TF binding specificity detection and motif screening in promoter of this gene (Weirauch et al., 2014). More advanced approaches have been developed that incorporate functionally relevant annotations and genomic features within the algorithms that attempt to identify the relevant SNPs rather than overlapping the datasets afterwards. All functional annotations are assigned weights that represent the likelihood of having an effect on gene expression and these weights are used to further prioritize SNPs (Gaffney et al., 2012) (Fig. 7.3). The implementation of such a framework in plants, incorporating conserved binding sites as functional annotations among others, could lead to unravelling the causal regulatory mechanisms underlying the variation in phenotype or expression.

Owing to the recent developments in genome editing through technologies such as TRANSCRIP-

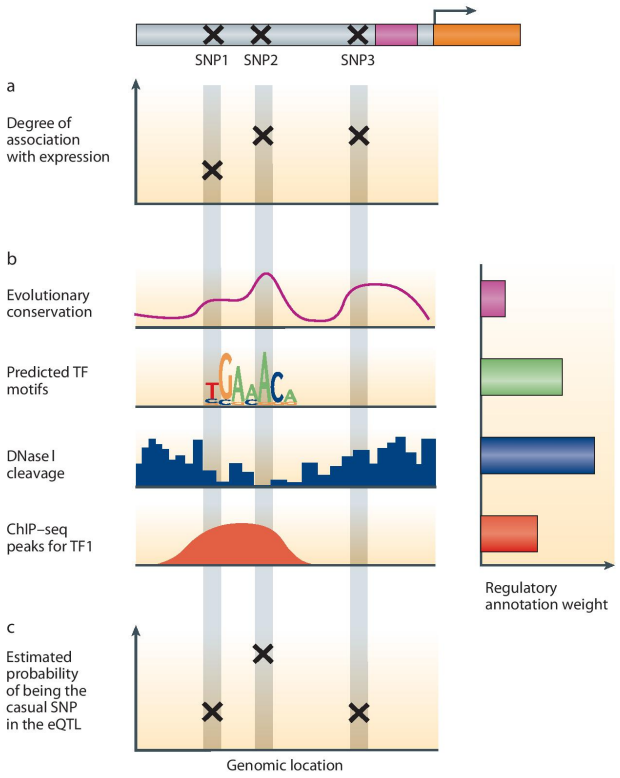

Figure 7.3: **Schematic representation of integration of functional annotations with eQTLs.** *Different functional annotations are assigned weights and are integrated with the predicted eQTLs. This leads to a reprioritization of the eQTLs putatively identifying the causal one.* Source: (Levo and Segal, 2014)

TION ACTIVATOR-LIKE EFFECTOR NUCLEASE (TALEN) and clustered regularly interspaced short palindromic repeat/CRISPR-associated protein 9 (CRISPR/Cas9) this improved understanding of cis-regulatory variation can be applied in breeding strategies for crop improvement. A striking example of such a targeted mutagenesis was shown for the rice gene Os11N3. The pathogen Xanthomonas oryzae uses the TALE effector AvrXa7 binds the effector-binding element in the promoter of the Os11N3 gene in order to activate its expression. Through this mechanism it can abuse the sugar metabolism of the plant for its own benefit. Disruption of this cis-regulatory element through us of TALEN resulted in tolerance to Xanthomonas strains carrying the AvrXa7 effector (Li et al., 2012a).

7.3 Overview and Future Perspectives for Network Inference in Plants

A rise in TFBSs data availability

Throughout this thesis, the number of TFBSs, modeled as PWM or consensus sequence, available for Arabidopsis has greatly expanded due to large-scale efforts to study the DNA binding specificities of TFs. At the time of the research presented in chapter 3 we were able to bring together 436 binding sites for 176 TFs from various sources (Fig. 7.4). This dataset contained mostly consensus sequences but also a few PWMs obtained from the AGRIS (Davuluri et al., 2003), PLACE (Higo et al., 1999) and Athamap (Steffens et al., 2004) databases and a literature study. The emergence of PBMs as a high-throughput method for profiling TF binding specificities substantially enlarged the number of TFBS data and for the research presented in chapter 4 a total of 1,164 binding sites was available for 681 TFs (Weirauch et al., 2014; Franco-Zorrilla et al., 2014; Lindemose et al., 2014). DNA affinity purification sequencing (DAP-seq) is novel high-throughput methodology that was recently published (O'Malley et al., 2016). DAP-seq is a high-throughput TF binding site discovery method that interrogates genomic DNA with in-vitro-expressed TFs. This methodology was performed for 1,812 TFs and retrieved binding site information for 529 TFs. All these datasets together currently result in a total of 1,691 PWMs for 919 TFs. As big

of a leap as this may be, this still does not cover even half of the predicted 2200 TFs in Arabidopsis and ongoing efforts will be required to uncover the complete cistrome(Jin et al., 2014). This increase of TFBS data does however offer opportunities for the inference of gene regulatory networks. The first conserved binding site gene regulatory network presented in Chapter 3 consisted of 40,758 TF-target gene interactions. A recent update of this network was performed integrating all available TFBS data available currently (Fig. 7.4). In this analysis the CMM approach presented in Chapter 4 was applied to this dataset. This resulted in an extended conserved motif network of 340,893 interactions for 874 TFs and as such greatly increasing the size and coverage of the network.

Figure 7.4: **Overview of TFBSs availability throughout this thesis.** *The number of TFBSs and TFs covered by each screen or methodology are shown respectively as blue and green bars. A cumulative line plot is drawn for each illustrating the aggregation of data.*

Evaluating network inference strategies

A central theme across all chapters in this thesis was linking TFs to their correct target genes in order to reveal their biological functions or unraveling the underlying regulatory mechanisms of observed responses to stimuli and phenotypes. Different types of data and methodologies were utilized. In order to obtain an unbiased performance estimate of these approaches, all approaches used in this thesis were compared in a benchmark together with other inference methods. The gold standard to compare all approaches to was constructed from all validated interactions in AtRegNet, a review article on known interactions in secondary cell wall development and large scale data mining effort on 974 peer reviewed articles called Arabidopsis Transcriptional Regulatory Map (ATRM) (Yilmaz et al., 2011; Hussey et al., 2013; Jin et al., 2015). The network inference approaches examined were simple mapping of TFBSs across the promoters of all genes in the *Arabidopsis thaliana* genome and filtering these predicted binding sites with CNS data and open chromatin data. Co-expression approaches measure similarity at the output level of transcriptional regulation. In this analysis two co-expression based approaches were tested, determining the Pearson correlation between a TF and a target gene is very widely used to build co-expression clusters around a TF. As a second approach, GENIE3 a reverse engineering tool for reconstructing regulatory networks based on expression data was used (Huynh-Thu et al., 2010). This tool was shown to be the best performing in a large-scale benchmark (Marbach et al., 2012a). An integration of co-expression clusters with TFBS enrichment was also tested. In this analysis TFBS enrichment was performed on the co-expression cluster around each gene in the genome, all enriched TFs were linked to the genes in the co-expression cluster containing a binding site for this TF. Publicly available data generated by two experimental techniques, Y1H and TF ChIP-seq, was also evaluated next to these com-

putational inference methods(Brady et al., 2011; Li et al., 2014; Tian et al., 2014; Heyndrickx et al., 2014; Taylor-Teeples et al., 2015). All inferred networks were compared to the gold standard in terms of recovery of true edges and the enrichment of a network towards the gold standard (calculated as the ratio between recovery of true edges for the actual network and randomised networks), which functions as a proxy for specificity (Fig. 7.5).

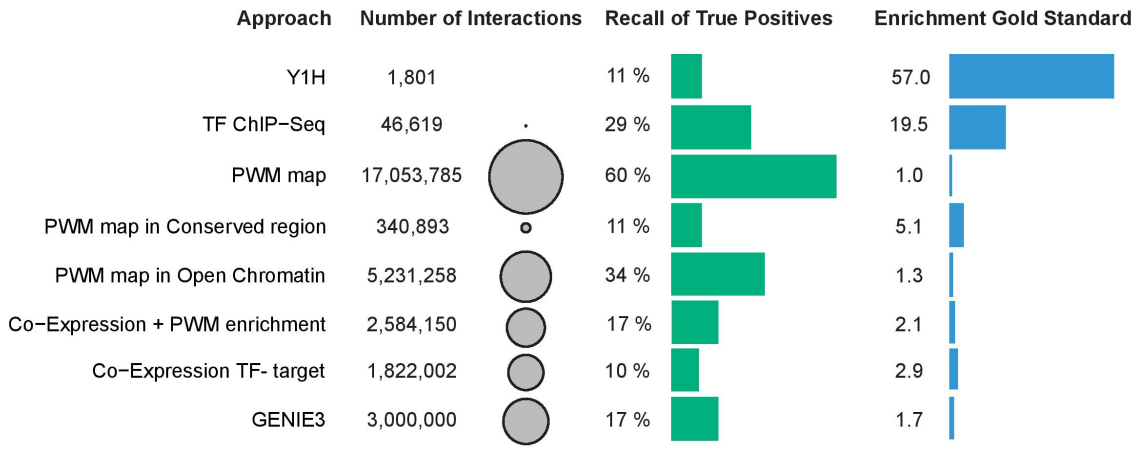

Figure 7.5: **Overview of benchmark of different network inference methods.** *The number of interactions for each approach is displayed both as a number and is also reflected in the size of the circel, the recovery of known interactions is displayed both as a number and as a barplot in green and the enrichment towards known interactions is also displayed as a number and as a bar plot in blue.*

The TF ChIP-seq network seems to perform best with a high recovery and high enrichment of the gold standard, Y1H also performs well with lower recovery but higher specificity. However, both of these experimental techniques have limited coverage. The genome-wide binding profiles of only around 30 TFs have been profiled in Arabidopsis and Y1H is mostly used to screen a set of candidate genes with a library of TFs resulting in a limited set of validated target genes. These limitations are mostly the costs and labor intensiveness of these techniques and are also the main reason why computational predictions are still relevant at this point. The best performing computational method is binding site conservation with a recovery of true interactions at the level of Y1H but strong enrichment towards the gold standard network compared to the other computational methods. Both other TFBS mapping strategies recover larger fractions of true edges but suffer from low specificity, resulting in a low enrichment value. The simple mapping of TFBS is known to result in many false positives because TFBSs are often short and typically contain some level of degeneracy in the binding motif (Tompa et al., 2005). The usage of open chromatin as a filter in this case is an aggregation of data from multiple tissues and conditions and this might, because of the sheer size of genomic regions withheld, introduce false positives in this manner. Open chromatin as guiding data is quite likely best used when the tissue or condition are corresponding to the research question under investigation. The usage of genomic footprints obtained from open chromatin data could offer a solution to this problem but to this date only one such dataset is available for Arabidopsis (Sullivan et al., 2014). An advantage of CNS to filter the TFBSs is that it is unbiased towards specific conditions or tissues. On the other hand, binding site turnover and network rewiring are known limitations that make functional binding sites undetectable through conservation analysis. Co-expression based approaches are also not biased towards condition dependency given the use of sufficiently large and diverse sampling of expression measurements and do not consider TF motif information as such, co-expression based approaches are able to pick up indirect targets regulated by intermediate TFs. Both co-expression based approaches perform relatively well, Pearson correlation co-expression has the second highest enrichment towards the gold standard and GENIE3 displays a good balance between recovery of true edges and enrichment. The integration of TFBS and expression data performs very well displaying a good recovery of true interactions. This approach was frequently used

with success throughout the work presented in this thesis. The enrichment is also higher than other computational methods with the exception of conserved TFBS and Pearson correlation co-expression. In general, methods that were used throughout this thesis have performed well in this benchmark.

Future perspectives for computational network inference in plants

Apart from evaluating the performance of different approaches, the similarity between inference approaches was also analyzed by looking at the enrichment of the delineated networks towards each other (Fig. 7.6). The different approaches show rather low enrichment towards each other with the exception of the two expression based approaches, which is to be expected, and TFBS conservation with TF ChIP-seq. The absence of strong similarities between the different approaches opens possibilities for further integration of these datasets given their individual merits. Previous work has shown that networks derived from different types of data may be affected by different shortcomings and that interactions that were predicted by one approach were not necessarily predictable by the other, further demonstrating the need to integrate complementary data sources (Marbach et al., 2012a). In machine learning integration is often split up in two categories: Supervised learning approaches make efficient use of the available prior knowledge of known interactions to build a model for handling new data, whereas unsupervised learning approaches are unbiased towards prior knowledge. Supervised strategies are generally considered more powerful when a large amount of existing knowledge is available to guide model inference. In organisms with little or no training data unsupervised strategies could prove very useful. Unsupervised integration of predictions at the level of inference methods starting from the same type of input data has shown robust performance on different validation data sets (Marbach et al., 2012b). Similarly predicting regulatory edges by integrating diverse complementary data sets as input features making use of a gold standard of previously validated interactions was shown to be a powerful and robust tool for the inference of transcriptional gene regulatory networks in E. coli and Drosophila melanogaster (Ernst et al., 2008; Marbach et al., 2012b).

As more research in plants is moving towards integrative approaches model-based integration of multiple datasets by is becoming more important (Brady et al., 2011; Taylor-Teeples et al., 2015; Walley et al., 2016). Ideally, emerging research questions in the field of regulatory genomics should be tackled with the generation of complementary datasets, analyzed with integrative methods, actively leveraging the method- and/or data-specific advantages with regard to the gene regulatory networks of interest.

7.4 Network Conservation and Translation Across Species

Although Arabidopsis has been a successful model species, which will continue to lead the forefront in plant research and provide important insights, transferring this knowledge to economically more interesting species will greatly improve the societal value of plant research. The idea that organs with conserved functions across plants (e.g. roots for water and nutrient uptake) should also have conserved functional organization is a paradigm underlying the use of these model organisms. In the case of transcription regulation however, small changes in gene expression control have been shown to be major contributors to phenotypic diversity. Until recently, detection of TFBSs across species was almost exclusively performed using conservation analysis with relative success both in humans, plants and other species (Lindblad-Toh et al., 2011; Haudry et al., 2013; Van de Velde et al., 2014; De Witte et al., 2015; Van de Velde et al., 2016). Nevertheless, conservation analyses lack direct evidence of binding and cannot resolve spatiotemporal patterns of transcription factor binding. Next to these points, these types of analyses also have limited power to detect lineage-specific regulatory changes depending on the set of species included (Alfoldi and Lindblad-Toh, 2013).

Over the last years TFBS conservation has also been assessed experimentally through TF ChIP-seq, tackling the limitations of conservation analyses. The first study to experimentally profile the binding patterns of four orthologous TFs across species revealed that 41 % - 81 % of the individual binding events were species specific (Odom et al., 2007). As more analyses of TF binding conservation accumulate these numbers were confirmed. It seems that greater divergence time leads to lower conservation within groups of species, but that there are differences between groups of species as Drosophila species show

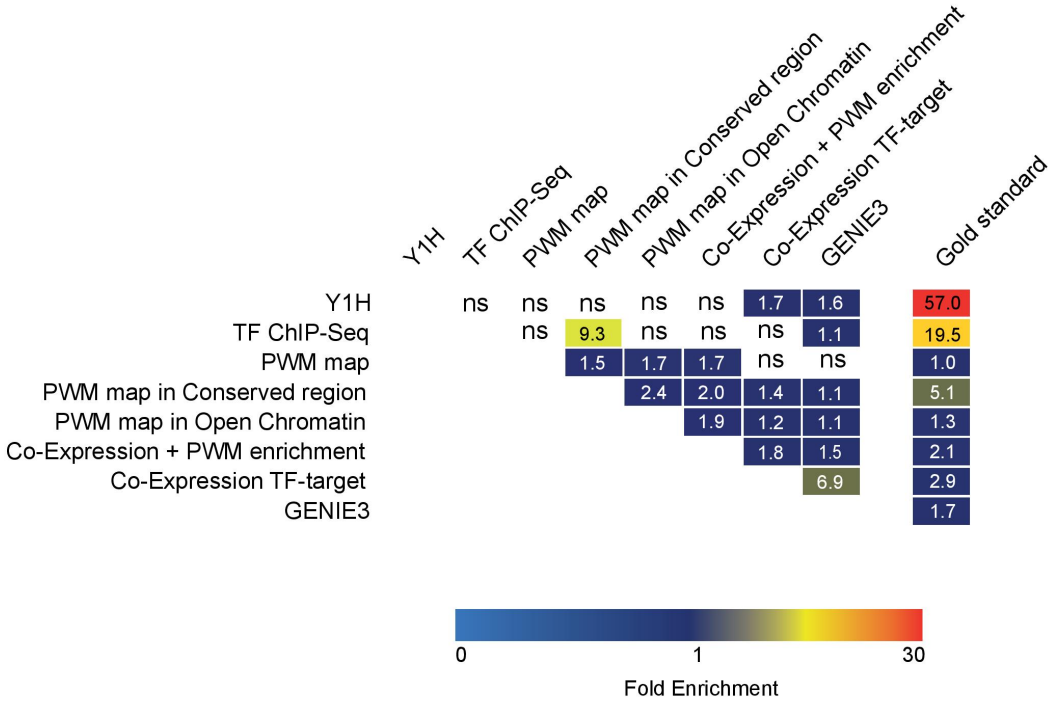

Figure 7.6: **Overview of network enrichment compared to all other networks.** *ns indicates that no significant enrichment could be detected between two networks.*

higher conservation when compared with mammals given a same evolutionary distance (Villar et al., 2014). This same study also concluded that the role of the TF might have an influence on binding site conservation. A comparison between *Arabidopsis thaliana* and *Arabidopsis lyrata* found that around 21 % of binding events were conserved, which is in line with findings from other kingdoms (Muino et al., 2015). To this day it remains challenging to clearly differentiate functional binding from putatively non-functional or background binding (Fisher et al., 2012). As such, the consistent conclusions that many single binding events are not conserved could point towards the non-functionality of certain binding sites. This point is further strengthened by the finding that while many binding events were species-specific, the binding events associated with genes that show a change in expression response upon perturbation of the TF in question were highly conserved (Boj et al., 2009). In a recent publication, 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 indeed highly factor dependent (Cheng et al., 2014).

Next to experimental limitations there are also evolutionary mechanisms that cause changes in TF binding. In plants these evolutionary forces are present both on the binding site level and on the gene level. Mutations have long been thought to be a fundamental cause of binding site turnover, but recently comparative analysis have shown that, at best, a substantial minority of transcription factor binding differences can be attributed to alterations in directly bound genetic sequences. A recent TF ChIP-seq comparative study indicates that sequence changes in the TF binding motif only provide an ex- planation for a minority (12-40 %) of TFBS variation (Villar et al., 2014). Besides the occurrence of mutations in binding sites, another mechanism to create new TFBSs is transposition (Schmidt et al., 2010). In plants, E2F TFBSs might have been amplified by transposon activity in Brassicaceae species (Henaff et al., 2014). Plant genomes have undergone frequent polyploidization and small-scale duplication events, which have resulted in a very complex relationship between orthologous genes (Proost et al., 2009). After duplications, genomic rearrangements and gene loss can occur, but also neo- and subfunctionalization. These changes in gene function can lead to regulatory changes further impeding a straightforward

interpretation of binding site evolution (Airoldi and Davies, 2012; Moghe and Shiu, 2014).

In recent years many computational tools have been developed for reconstructing GRNs and studying network evolution in other model organisms than plants (Thompson et al., 2015). An outstanding challenge in plant regulatory genomics would be to devise an efficient computational framework to study the reorganization of GRNs, as most of the previously developed methods are not adapted to the complex orthology relationships in plants. Such a framework should exploit the wealth of TFBS data that has become available in Arabidopsis to generate species-specific bindings site models through integration with orthology, conservation and dedicated expression data. A large-scale prediction of network rewiring and reconstruction of ancestral states could prove very useful for studying GRNs in crops.

CHAPTER 8

Curriculum Vitae

EXPERIENCE

Data scientist / Bioinformatician Sep 2016 - Present
Biocartis

Predoctoral Fellow / Bioinformatician Dec 2012 - Sep 2016
IWT grant: Regulatory Annotation of Plant Genomes through Functional Chromatin Signatures and Comparative Sequence Analysis.
Department of Plant Systems Biology (VIB) - Ghent University
Promoter: Prof. dr. K. Vandepoele
Co-promoter: Prof. dr. ir. J. Fostier

WORKSHOPS

- Project Management
- English Scientific Writing
- Effective Graphical Displays
- Data Mining and Big Data (Master of Statistical Data Analysis)
- EMBO Practical Course on Genotype to Phenotype Mapping of Complex Traits, European Bioinformatics institute, 2014

EDUCATION

Master in Biochemistry and Biotechnology (Honour) Sep 2010 - Jun 2012
Major Bioinformatics & Systems Biology
Thesis: Conservation Analysis of Experimental Transcription Factor Binding Sites in Plants
Ghent University
Promoter: Prof. dr. K. Vandepoele
Preparation year to Master in Biochemistry and Biotechnology (Honour) Sep 2009 - Jun 2010
Ghent University
Bachelor in Medicine (Satisfaction) Sep 2005 - Jun 2009
Ghent University

SCIENTIFIC PUBLICATIONS

Inference of Transcriptional Networks in Arabidopsis through Conserved Noncoding Sequence Analysis.
Van de Velde J*[a], Heyndrickx KS*, Vandepoele K
Plant Cell 26: 2729-2745, 2014

A Collection of Conserved Noncoding Sequences to Study Gene Regulation in Flowering Plants.
Van de Velde J, Van Bel M, Vaneechoutte D, Vandepoele K
Plant Physiology 171: 2586-2598,2016

BLSSpeller: Exhaustive Comparative Discovery of Conserved Cis-Regulatory Elements.
De Witte D*, Van de Velde J*, Decap D, Van Bel M, Audenaert P, Demeester P, Dhoedt B, Vandepoele K, Fostier J
Bioinformatics 31: 3758-3766, 2015

A DNA-binding-site Landscape and Regulatory Network Analysis for NAC Transcription Factors in Arabidopsis thaliana.
Lindemose S*, Jensen MK*, Van de Velde J*, O'Shea C, Heyndrickx KS, Workman CT, Vandepoele K, Skriver K, De Masi F
Nucleic Acids Res 42: 7681-7693, 2014

A Functional and Evolutionary Perspective on Transcription Factor Binding in Arabidopsis thaliana.
Heyndrickx KS*, Van de Velde J*, Wang C, Weigel D, Vandepoele K
Plant Cell 26: 3894-3910, 2014

Functional Analysis of the Arabidopsis TETRASPANIN Gene Family in Plant Growth and Development.
Wang F, Muto A, Van de Velde J, Neyt P, Himanen K, Vandepoele K, Van Lijsebettens M
Plant Physiology 169: 2200-2214, 2015

Selection for Improved Energy Use Efficiency and Drought Tolerance in Canola Results in Distinct Transcriptome and Epigenome Changes.
Verkest A, Byzova M, Martens C, Willems P, Verwulgen T, Slabbinck B, Rombaut D, Van de Velde J, Vandepoele K, Standaert E, Peeters M, Van Lijsebettens M, Van Breusegem F, De Block M
Plant Physiology 168: 1338-1350, 2015

[a] An asterisk indicates shared author contribution.

Supplemental Data

SUPPLEMENTAL DATA FILE **A**

Detection of Conserved Noncoding Sequences in Arabidopsis

A.1 Supplemental Figures

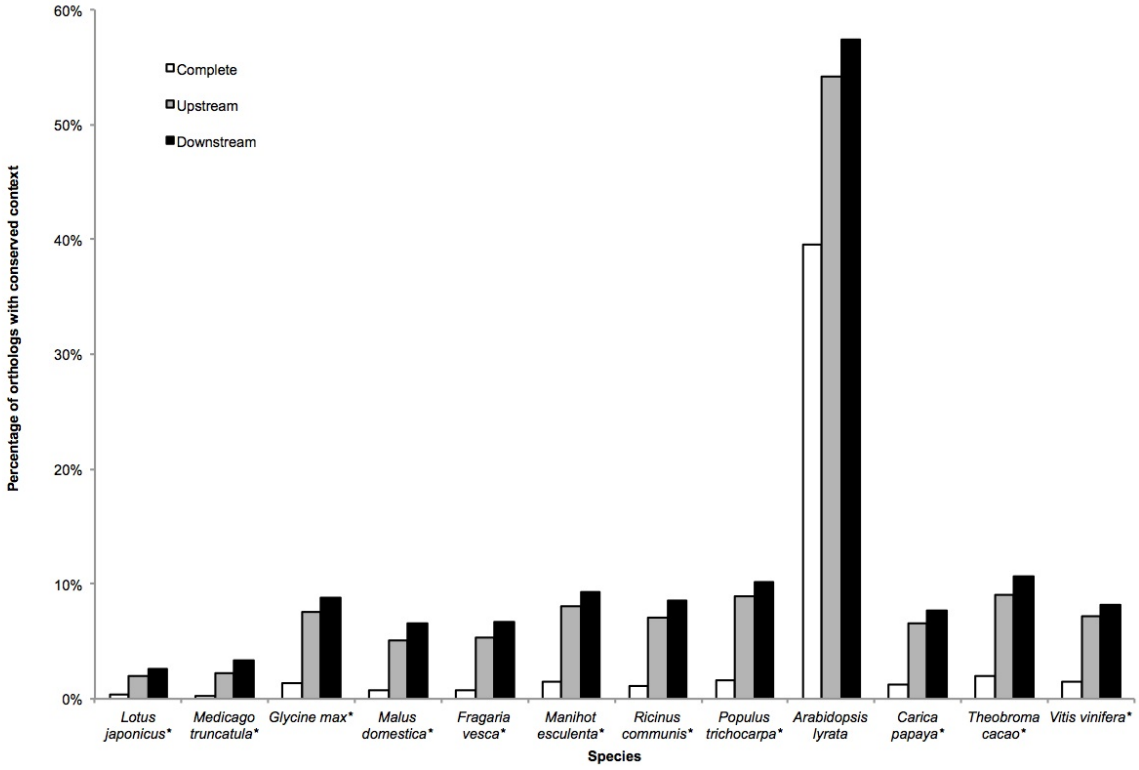

Figure A.1: **Overview of synteny conservation between Arabidopsis and other dicot species.** *This figure shows the percentage of orthologous genes for each Arabidopsis gene for which the flanking genes were conserved by collinearity. Criteria to score collinearity conservation were: 1) whether the 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).*

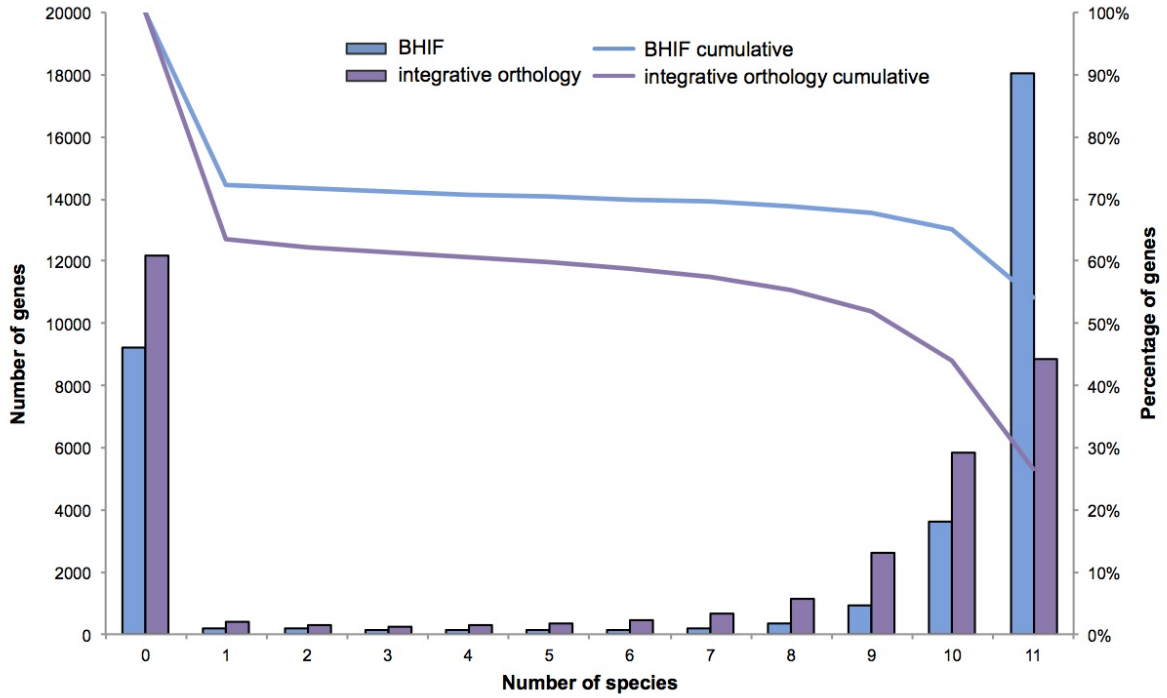

Figure A.2: **Distribution of genes that have orthologs in the dicot comparator species for each orthology detection method.** *The number of Arabidopsis genes with orthologs in different comparator dicot species is depicted for the integrative orthology (purple boxes) and BHIF method (blue boxes), respectively (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).*

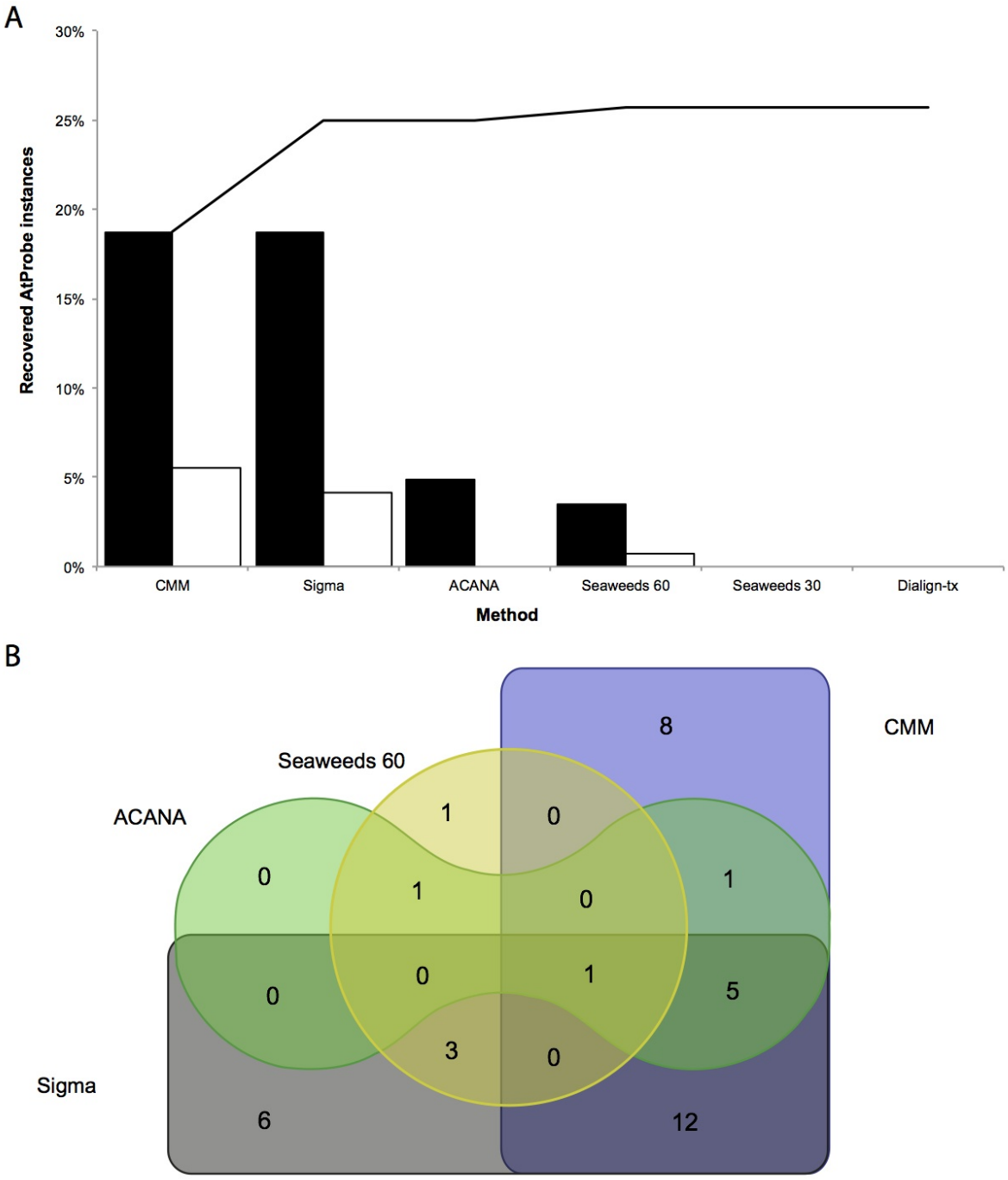

Figure A.3: **Recovery of experimental AtProbe elements using different phylogenetic footprinting approaches.** *A) For the different phylogenetic footprinting approaches developed in this study, the recovery of AtProbe elements was determined. Black boxes show the percentage of recovered elements while 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.*

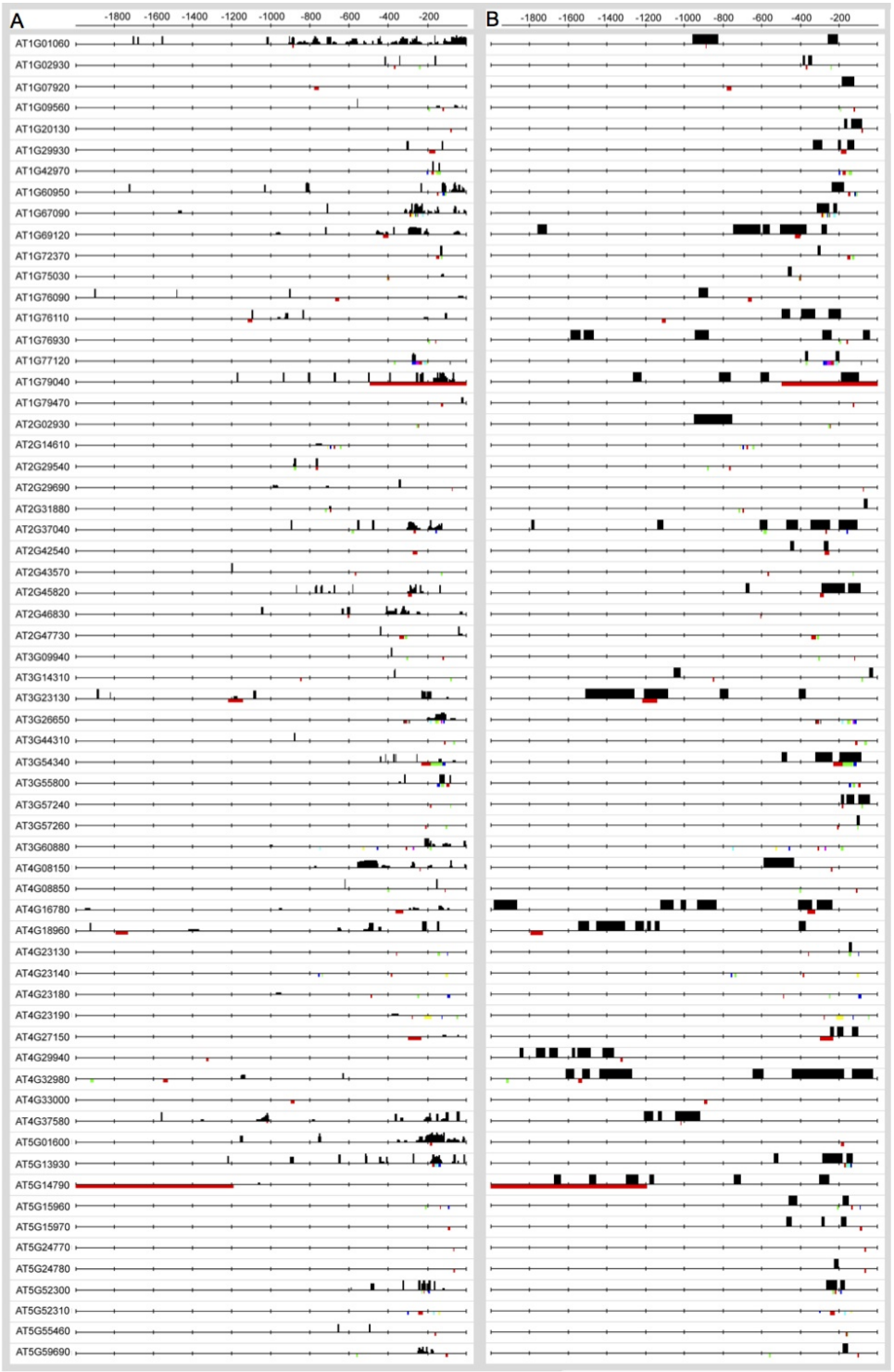

Figure A.4: **Recovery of AtProbe elements for the CNSs described in this paper (A) and by Haudry et al. (2013)(B).** *Black lines denote upstream sequences, colored boxes depict AtProbe elements, and black boxes show significant CNSs.*

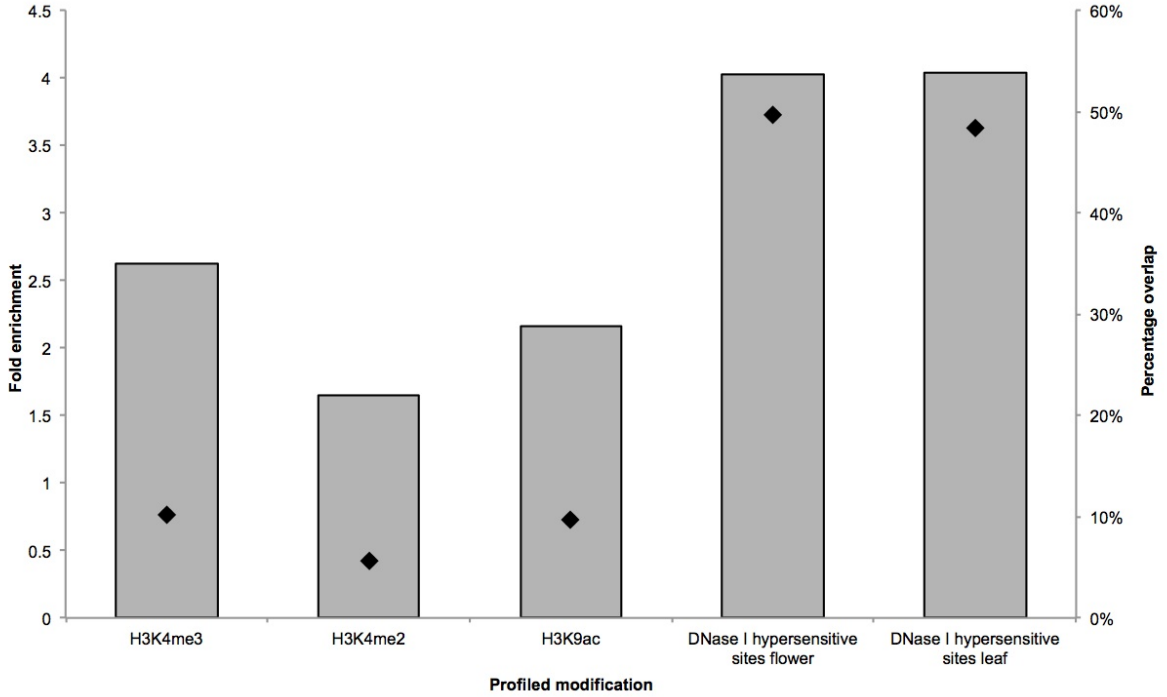

Figure A.5: **Enrichment and overlap of in vivo functional regions with CNSs.** *Grey boxes show the fold enrichment of different histone marks and DH sites. Black diamonds show the percentages of CNSs that overlap with each in vivo functional region dataset.*

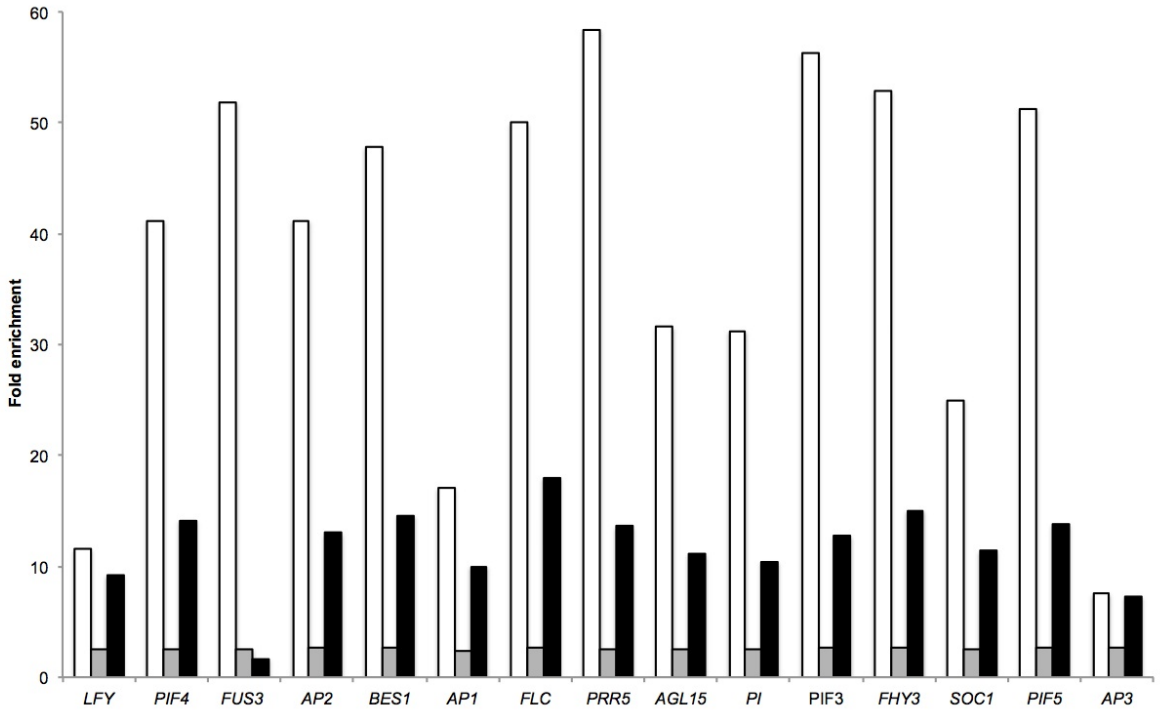

Figure A.6: **Comparison of fold enrichment for in vivo functional binding site regions.** *Fold enrichment for in vivo functional binding sites is shown for our CNSs dataset (white boxes),simple motif mapping (grey boxes) and motif mapping within DH sites (black boxes).*

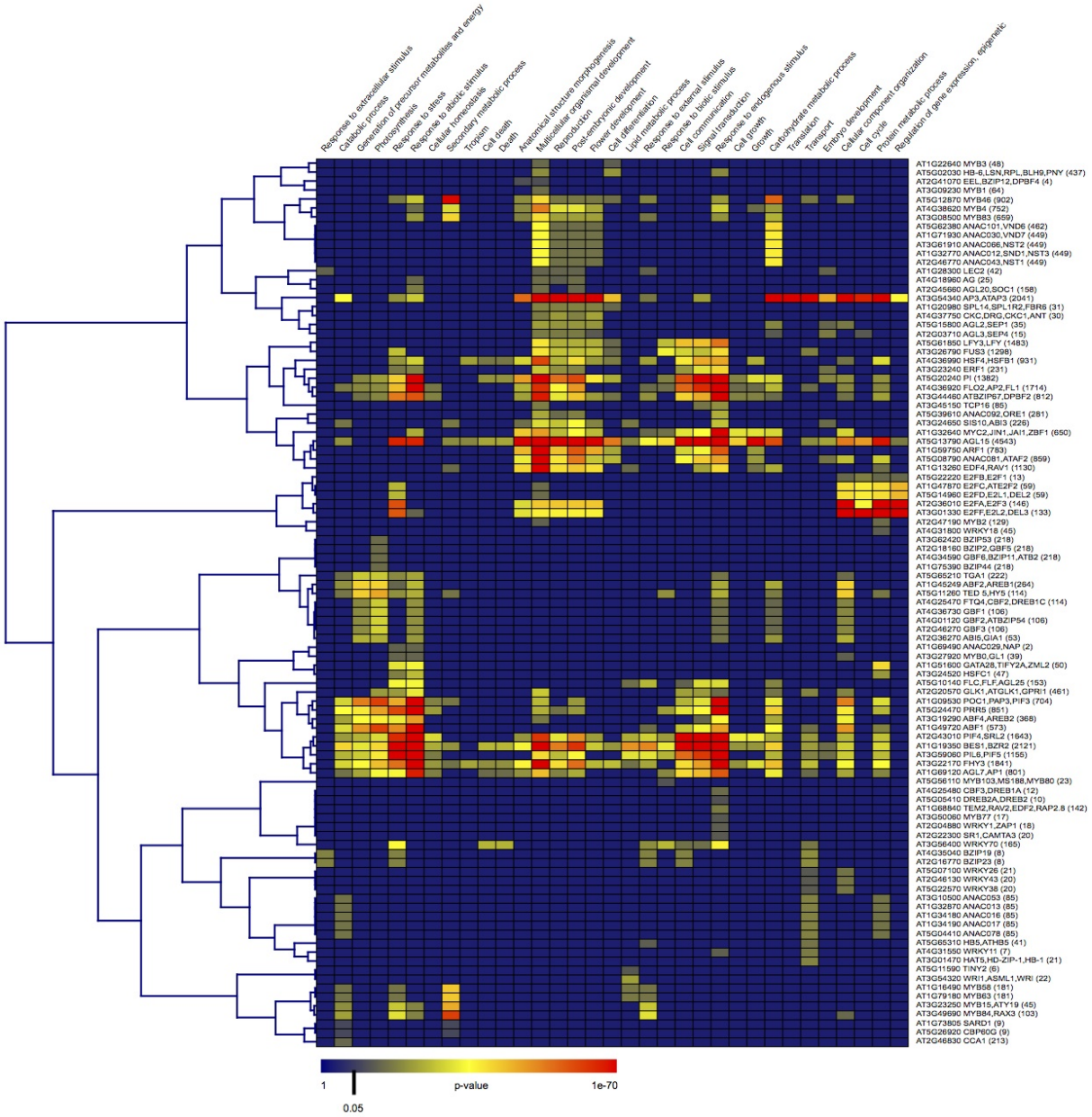

Figure A.7: **GO enrichment for all TF-targets in the predicted GRN.** *A heatmap was generated using Genesis that displays, per TF, the enrichment of target genes towards GO slim annotations (hypergeometric distribution + Bonferroni correction). The number of target genes for each TF is shown in parenthesis.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

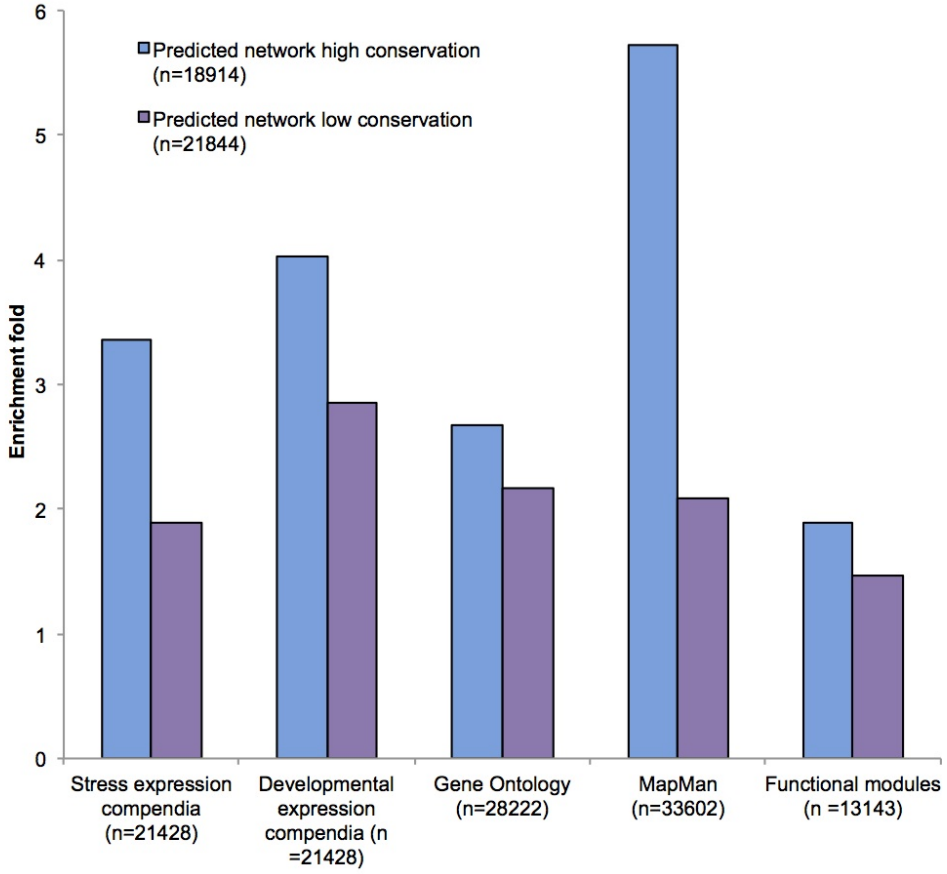

Figure A.8: **Evaluation of the biological relevance of highly and moderately conserved interactions using the biological validation metrics.** *Comparison of the five biological metrics for the predicted sub-networks with highly (blue boxes, >6 species) and moderately (purple boxes, 2-6 species) conserved 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 A.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. Co-expression in this figure was based on a Z-score greater than 2 or smaller than -2, this coincides with the 5% extremest co-expression relationships. Experimental refers to experimentally supported interactions, whereas predicted refers to the interactions that were identified in this chapter. 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.*

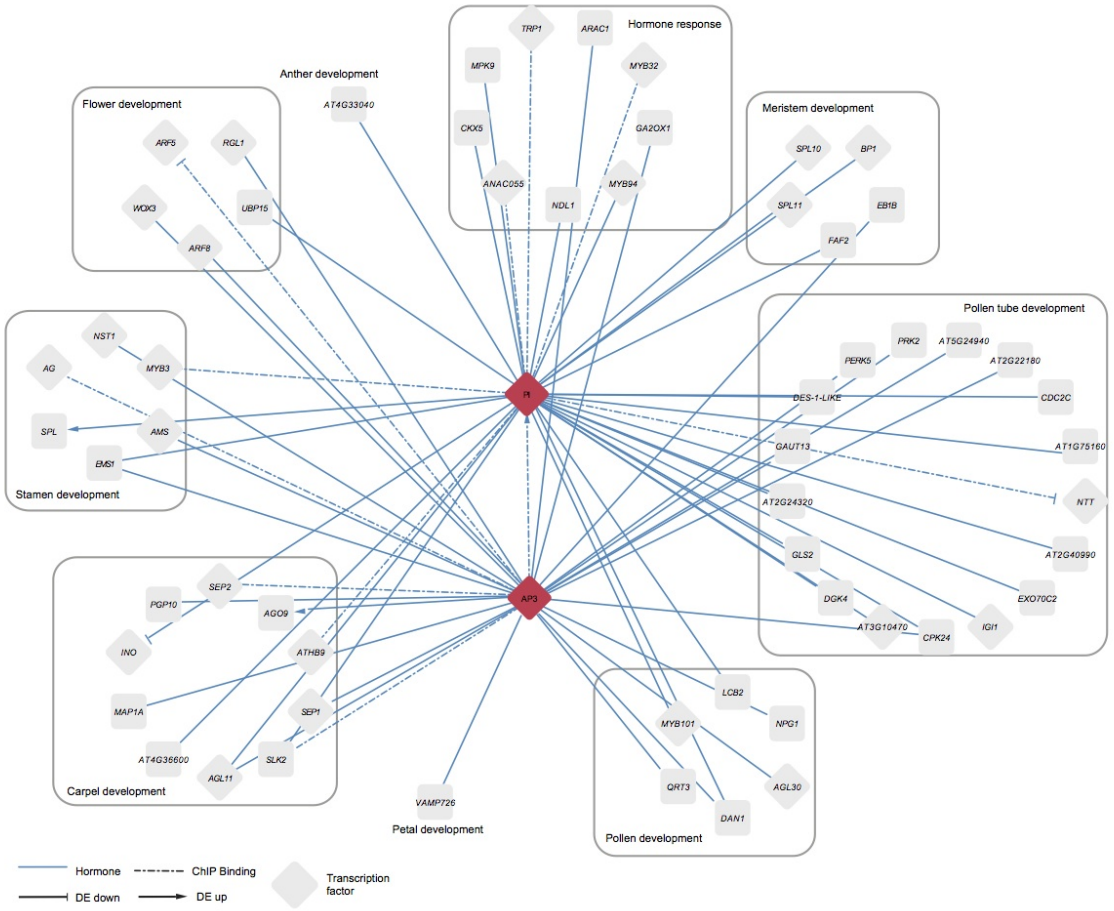

Figure A.10: **A condition-specific GRN for PI and AP3 based on hormone-specific TF-target co-expression edges.** *Genes that have GO annotations related to flower development are displayed. ChIP-bound regions associated with the target gene are shown as dashed lines while differentially expressed genes 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.*

Detection of CNSs in ten dicot species

B.1 Supplemental Figures

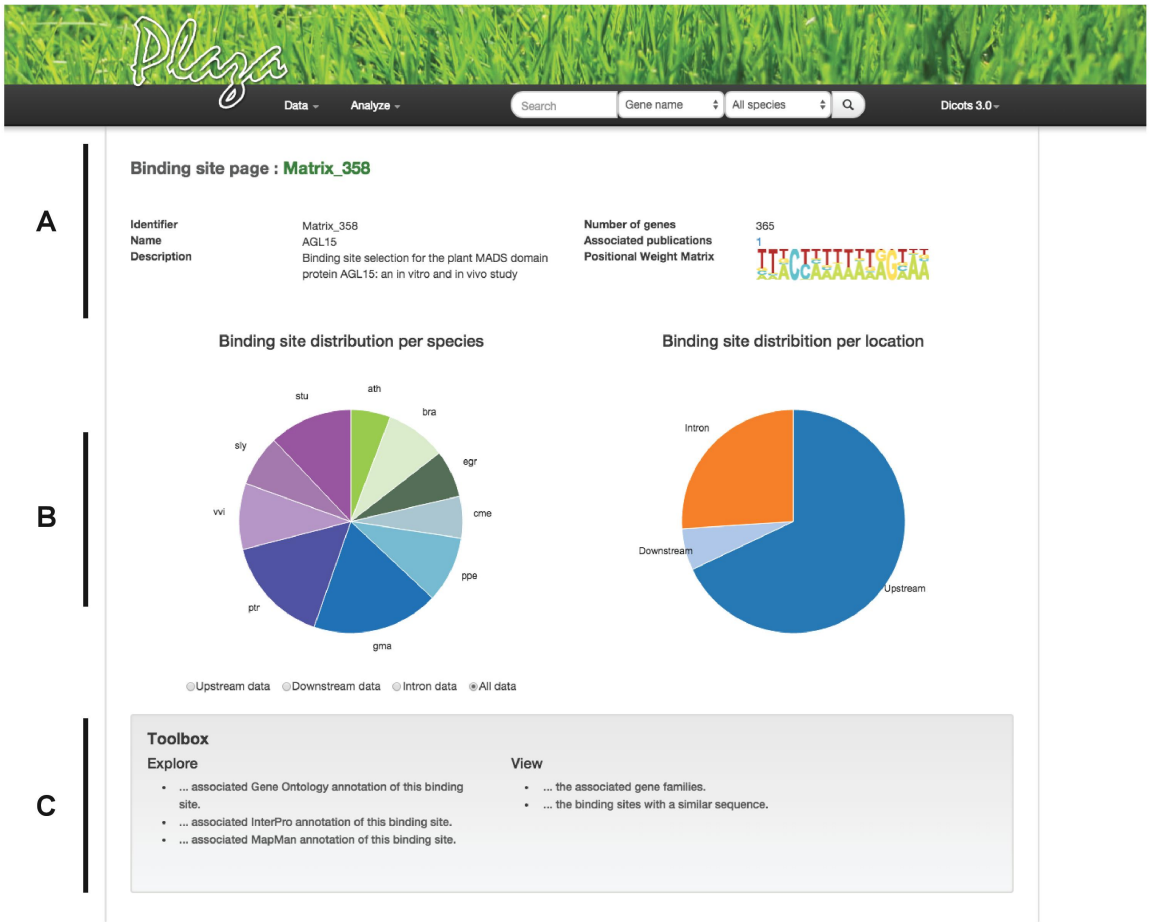

Figure B.1: **Overview of new motif page in PLAZA 3.0 Dicots platform.** *A) Information about the binding site and overview of the total number of target genes. B) Breakdown of the target genes per species and genomic region. C) Toolbox for further downstream analysis.*

Detection of CNSs in Monocot species

C.1 Supplemental Figures

Figure C.1: **Overview of conserved motifs for different thresholds.** *This figure displays the number of genome-wide conserved motifs for both alignment-based and alignment-free discovery for different values of Cthres and Fthres and different subsets of the six BLS thresholds Ti (T1 = 15%, T2 = 50%, T3 = 60%, T4 = 70%, T5 = 90% and T6 = 95%). Top number: real Monocot dataset; bottom number between brackets: random dataset generated using a zeroth-order Markov model (conservation of 1-mer frequencies). The colors represent the false discovery rate.*

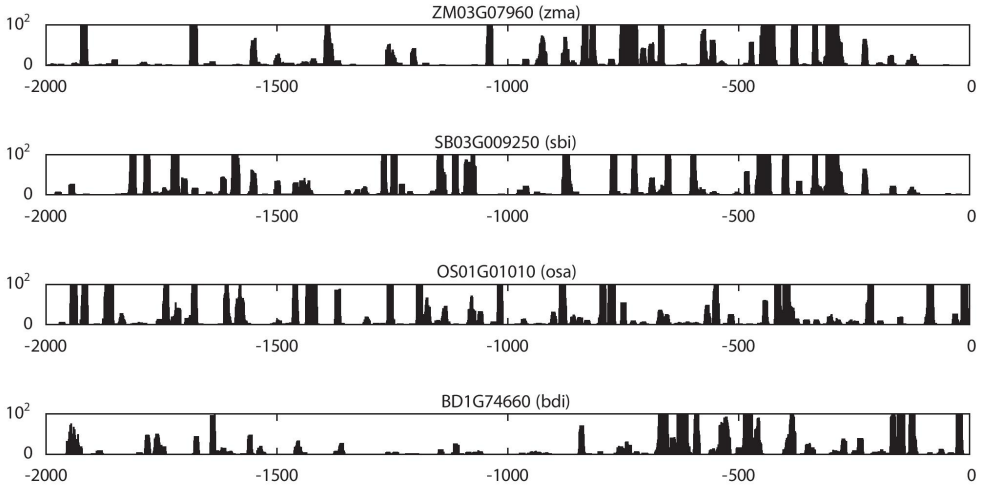

Figure C.2: **Example of CNSs for an orthologous family.** *Conserved regions in the promoters of the genes in gene family iORTHO00001 corresponding to motif instances with BLS $\geq$ 15%, F $\geq$ 20 and C $\geq$ 0.9, i.e., high-scoring motifs that are conserved in at least two species. The height of the bars corresponds to the number of distinct motif variants that map to that location. Note that the y-axis has been truncated at 100: certain loci in this gene family are covered with up to 18 418 distinct motif variants.*

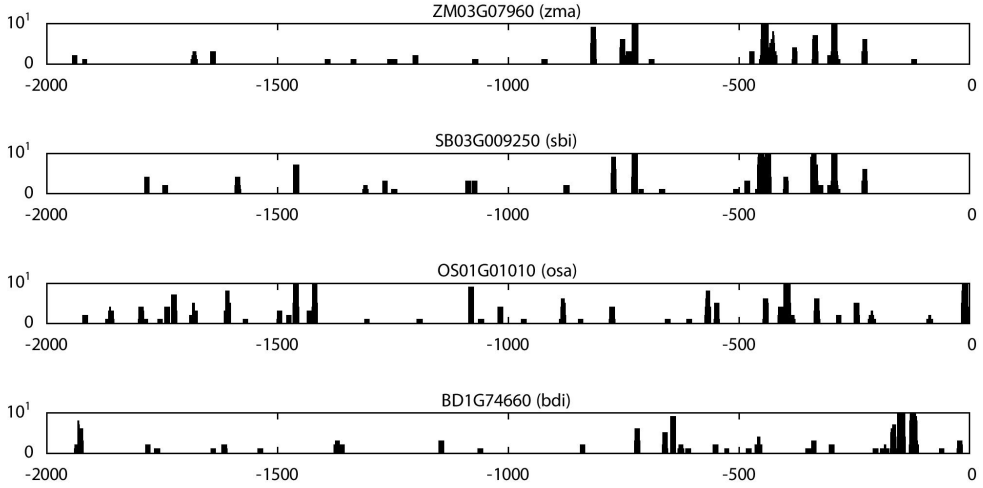

Figure C.3: **Example of CNSs for an orthologous family.** *Conserved regions in the promoters of the genes in gene family iORTHO00001 corresponding to motifs instances with BLS $\geq$ 95%, $F(95\%) \geq 50$ and $C(95\%) \geq 0.9$, i.e., motifs conserved in all four species. The height of the bars corresponds to the number of distinct motif variants that map to that location. Note that the y-axis has been truncated at 10: certain loci in this gene family are covered with up to 568 distinct motif variants.*

A NAC binding site landscape in Arabidopsis

D.1 Supplemental Figures

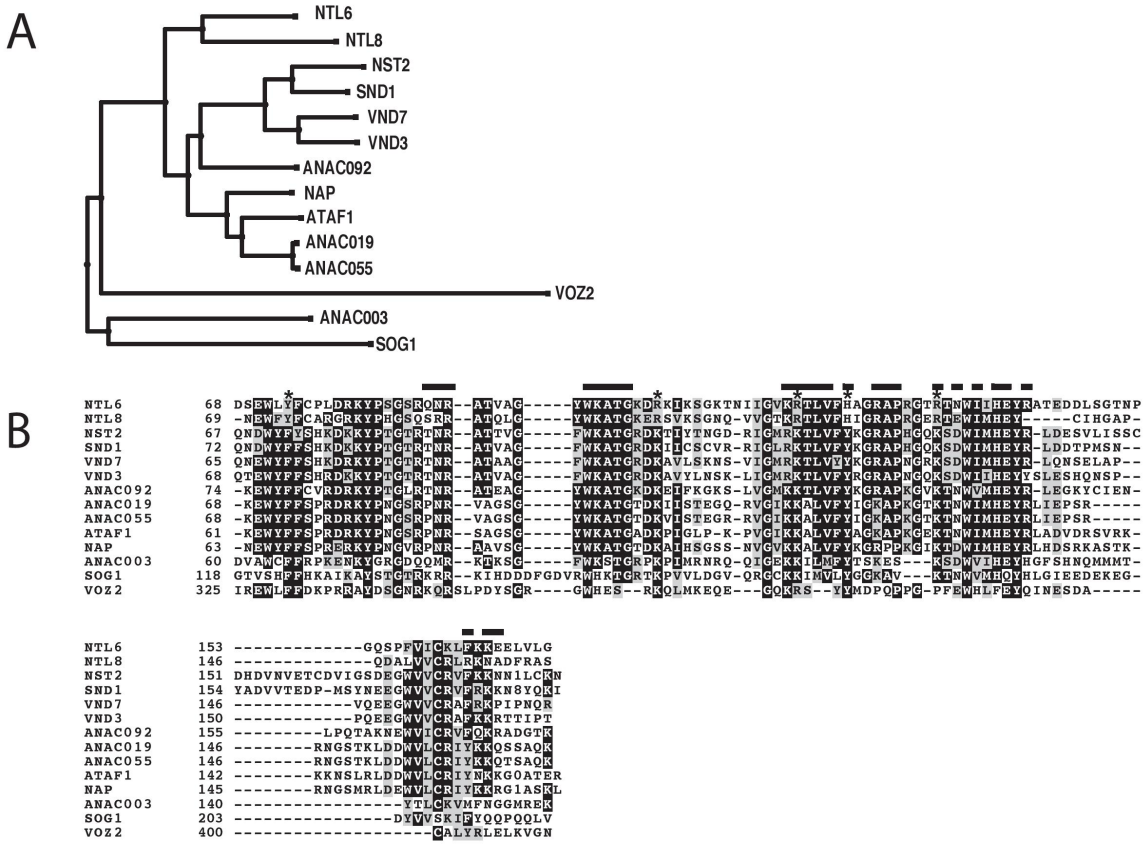

Figure D.1: **Sequence conservation and identity of NAC protein DNA binding domains.** *A) NAC DNA binding domain sequence similarity tree for all studied NAC proteins shows 3 main clusters for our candidate TFs. Cluster I contains ANAC092, NST2, SND1, VND3, VND7, ANAC019, ANAC055, ATAF1 and NAP; Cluster II contains NTL6 and NTL8 and, finally, Cluster III contains VOZ2, ANAC003 and SOG1. B) Multiple sequence alignment of the DNA binding regions of the selected 14 NAC proteins. Residues that based on the x-ray model of the ANAC019-DNA complex are close to DNA are shown by a bar. Residues marked with black boxes are common to at least half of the sequences and residues marked in grey boxes are chemically similar in half of the sequences. Asterisks highlight those residues showing remarkable divergence between NTL6, NTL8 and the remaining NAC proteins.*

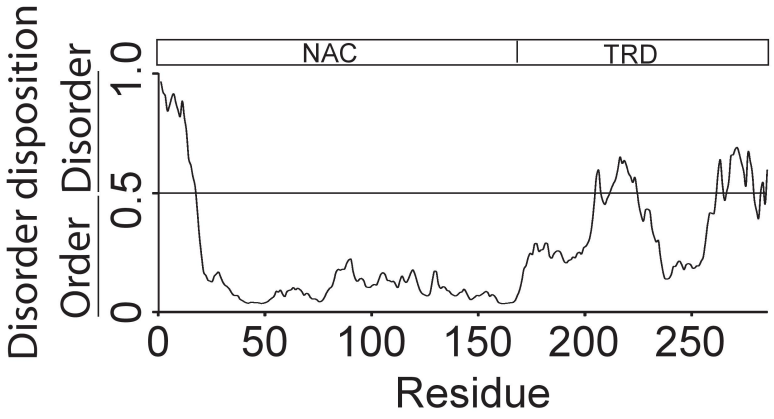

Figure D.2: **Disorder prediction in ANAC092.** *ID prediction of ANAC092 was performed using PONDR-FIT (Xue et al. 2010). A threshold is applied with disorder assigned to values ≥ 0.5.*

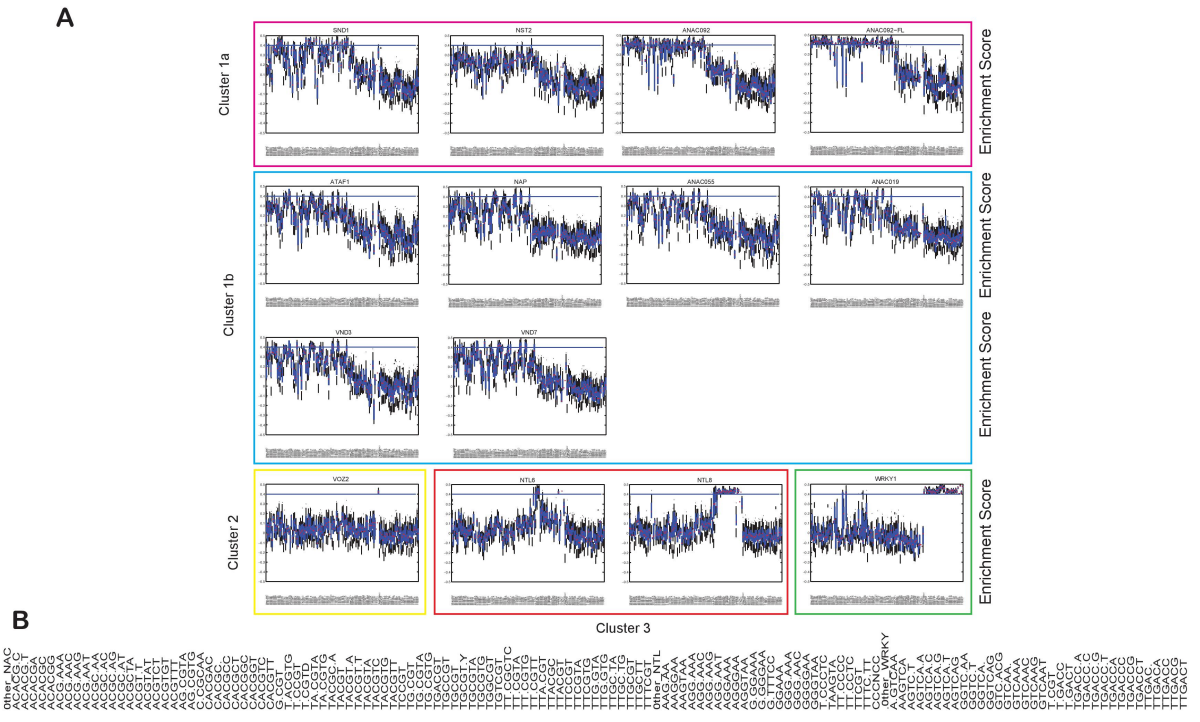

Figure D.3: **Overview of ES distributions.** *A) Boxplots of ES distributions for 130 signature 6-mers for all tested TFs. TF boxplots are grouped according to clusters in Figure 1A. B) List of the 130 key 6-mers describing NAC DNA specificities as presented on the x-axis of all boxplots.*

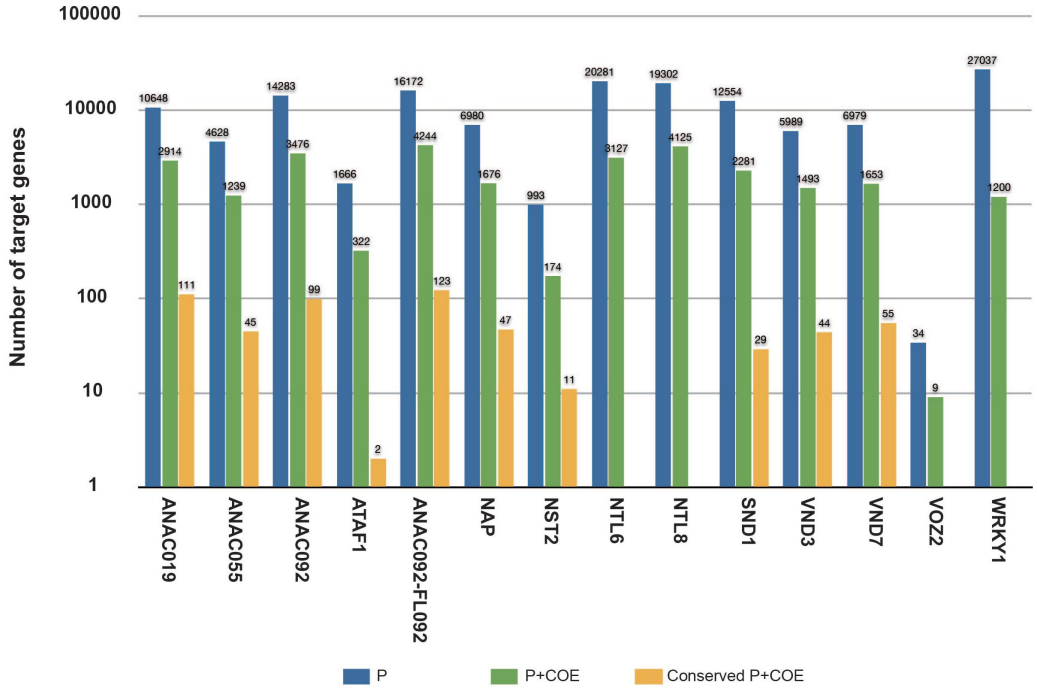

Figure D.4: **Overview of the number of target genes for different NAC TFs and filtering approaches.** *The number of target genes is shown for each TF. Blue bars indicate the number of P target genes through simple screening of promoters with high scoring k-mers. Green and yellow bars show the number of target genes when integrating co-expression information and conserved motif information, respectively.*

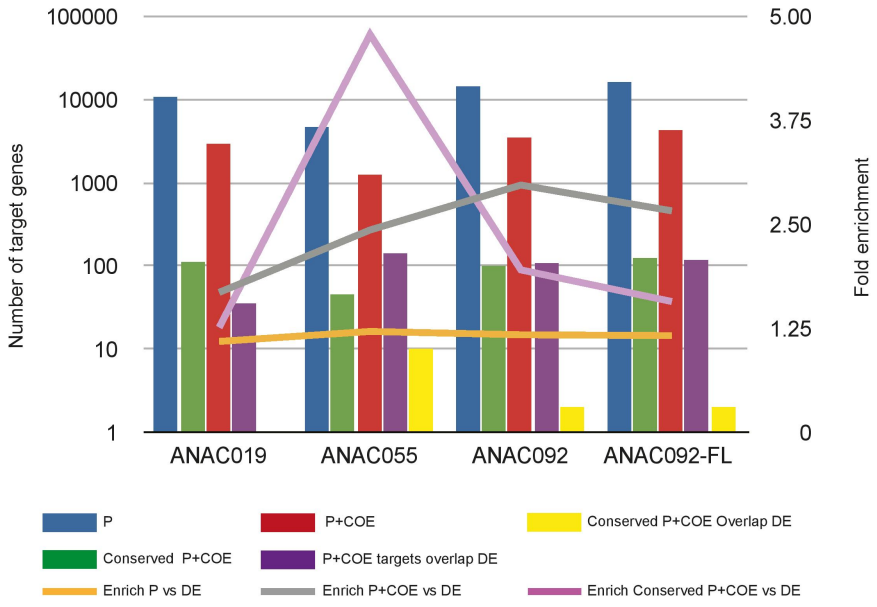

Figure D.5: **Overlap analysis of different sets of PBM target genes and enrichment analysis for differentially expressed genes.** *Bars show the number of target genes for different PBM filtering approaches. COE and Conserved refer to the integration of co-expression and motif conservation information, respectively. DE refers to integration of differentially expressed gene sets. Purple, grey and yellow lines display enrichment for DE genes for P targets, integration with co-expression and integration with conserved motif information, respectively.*

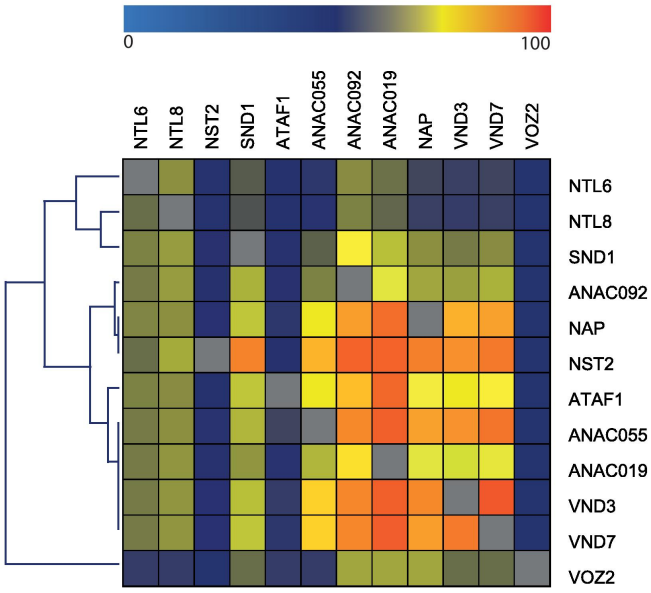

Figure D.6: **Clustering of NAC TFs based on shared target genes.** *The overlap between P+COE target genes for all TFs was measured and clustered using hierarchical clustering with complete linkage (with Pearson correlation as a distance metric). Overlap was defined as the fraction of target genes for TF row shared with TF column and ranges from 0 to 100%.*

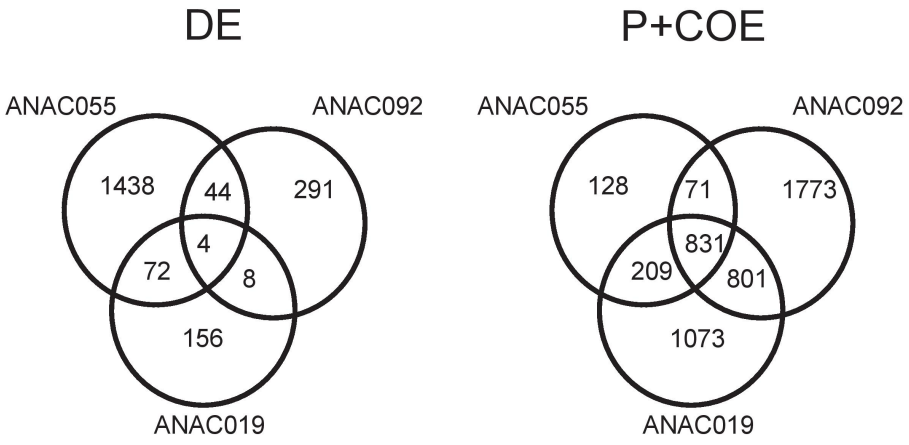

Figure D.7: **Overlap analysis of DE genes and P+COE target genes for ANAC055, ANA092 and ANAC019.** *Overlap between DE genes and P+COE target genes illustrates the redundancy in target genes that exists between these TFs. P+COE refers to predicted target genes combined with co-expression information.*

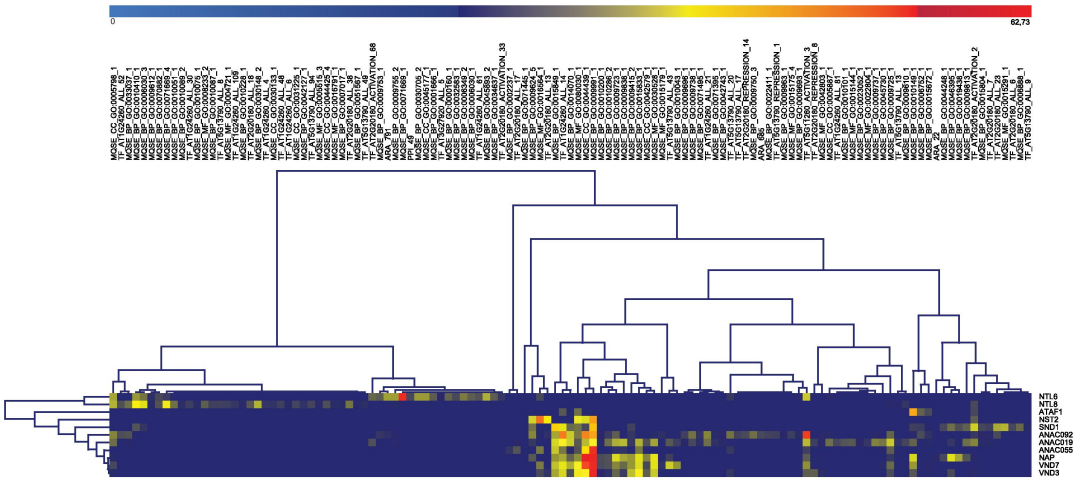

Figure D.8: **Clustering of NAC TFs based on shared target gene modules.** *Based on the enriched modules per TF (P+COE target genes), hierarchical clustering was performed on both TFs and modules. Enrichment folds are displayed as different colors in the heatmap.*

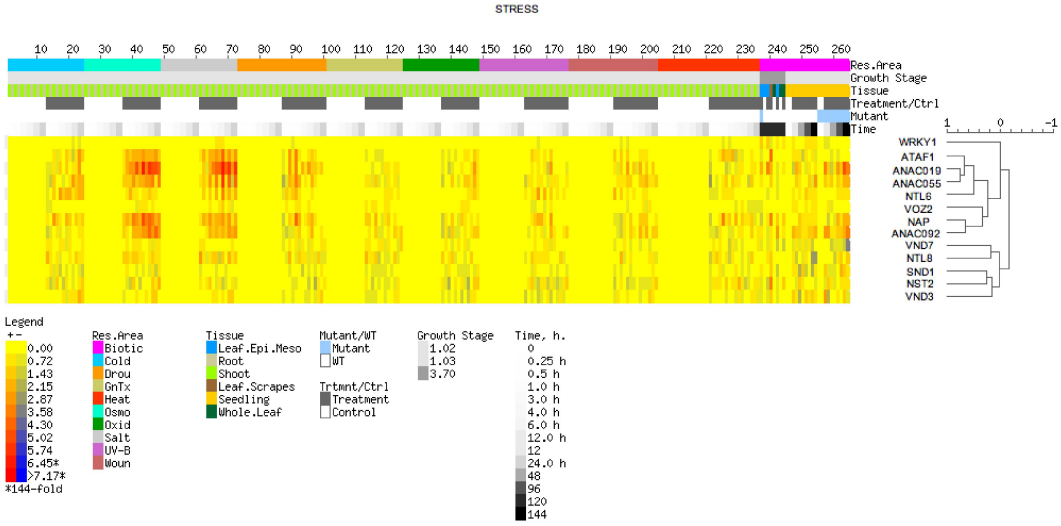

Figure D.9: **Overview of stress-responsive expression profiles for the different NAC TFs.** *Expression levels in different stress conditions were clustered for all TFs using the BAR Toronto Expression Browser (Stress series).*

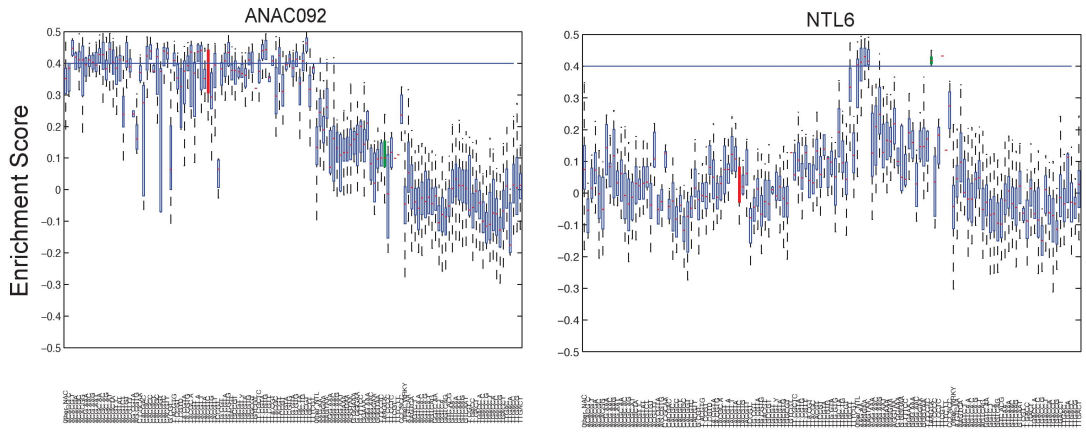

Figure D.10: **ANAC092 and NTL6 show non overlapping k-mer signatures.** *For each protein, the red box shows the ES distribution of k-mers containing the TACGTC key k-mer, which is specific for Cluster 1a and 1b proteins. The green box shows the ES distribution of k-mers containing the TAAGTA key k-mer, specific for Cluster 3 proteins.*

Bibliography

Thomas Abeel, Thomas Van Parys, Yvan Saeys, James Galagan, and Yves Van de Peer. GenomeView: a next-generation genome browser. *Nucleic Acids Research*, 40(2):e12–e12, January 2012. ISSN 0305-1048, 1362-4962. doi: 10.1093/nar/gkr995. URL http://nar.oxfordjournals.org/content/40/2/e12.

Chiara A. Airoldi and Brendan Davies. Gene duplication and the evolution of plant MADS-box transcription factors. *Journal of Genetics and Genomics = Yi Chuan Xue Bao*, 39(4):157–165, April 2012. ISSN 1673-8527. doi: 10.1016/j.jgg.2012.02.008.

Jessica Alfoldi and Kerstin Lindblad-Toh. Comparative genomics as a tool to understand evolution and disease. *Genome Research*, 23(7): 1063–1068, July 2013. ISSN 1088-9051. doi: 10.1101/gr.157503.113. URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3698499/.

Carlos Alonso-Blanco, Jorge Andrade, Claude Becker, Felix Bemm, Joy Bergelson, Karsten M. Borgwardt, Jun Cao, Eunyoung Chae, Todd M. Dezwaan, Wei Ding, Joseph R. Ecker, Moises Exposito-Alonso, Ashley Farlow, Joffrey Fitz, Xiangchao Gan, Dominik G. Grimm, Angela M. Hancock, Stefan R. Henz, Svante Holm, Matthew Horton, Mike Jarsulic, Randall A. Kerstetter, Arthur Korte, Pamela Korte, Christa Lanz, Cheng-Ruei Lee, Dazhe Meng, Todd P. Michael, Richard Mott, Ni Wayan Muliyati, Thomas Nagele, Matthias Nagler, Viktoria Nizhynska, Magnus Nordborg, Polina Yu Novikova, F. Xavier Piceş, Alexander Platzer, Fernando A. Rabanal, Alex Rodriguez, Beth A. Rowan, Patrice A. Salome, Karl J. Schmid, Robert J. Schmitz, umit Seren, Felice Gianluca Sperone, Mitchell Sudkamp, Hannes Svardal, Matt M. Tanzer, Donald Todd, Samuel L. Volchenboum, Congmao Wang, George Wang, Xi Wang, Wolfram Weckwerth, Detlef Weigel, and Xuefeng Zhou. 1,135 Genomes Reveal the Global Pattern of Polymorphism in Arabidopsis thaliana. *Cell*, 166(2):481–491, July 2016. ISSN 0092-8674, 1097-4172. doi: 10.1016/j.cell.2016.05.063. URL http://www.cell.com/cell/abstract/S0092-8674(16)30667-5.

Stephen F. Altschul, Warren Gish, Webb Miller, Eugene W. Myers, and David J. Lipman. Basic local alignment search tool. *Journal of Molecular Biology*, 215(3):403–410, October 1990. ISSN 0022-2836. doi: 10.1016/S0022-2836(05)80360-2. URL http://www.sciencedirect.com/science/article/pii/S0022283605803602.

Simon Anders and Wolfgang Huber. Differential expression analysis for sequence count data. *Genome Biology*, 11:R106, 2010. ISSN 1474-760X. doi: 10.1186/gb-2010-11-10-r106. URL http://dx.doi.org/10.1186/gb-2010-11-10-r106.

Xavier Argout, Jerome Salse, Jean-Marc Aury, Mark J. Guiltinan, Gaetan Droc, Jerome Gouzy, Mathilde Allegre, Cristian Chaparro, Thierry Legavre, Siela N. Maximova, Michael Abrouk, Florent Murat, Olivier Fouet, Julie Poulain, Manuel Ruiz, Yolande Roguet, Maguy Rodier-Goud, Jose Fernandes Barbosa-Neto, Francois Sabot, Dave Kudrna, Jetty Siva S. Ammiraju, Stephan C. Schuster, John E. Carlson, Erika Sallet, Thomas Schiex, Anne Dievart, Melissa Kramer, Laura Gelley, Zi Shi, Aurelie Bocrard, Christopher Viot, Michel Boccara, Ange Marie Risterucci, Valentin Guignon, Xavier Sabau, Michael J. Axtell, Zhaorong Ma, Yufan Zhang, Spencer Brown, Mickael Bourge, Wolfgang Golser, Xiang Song, Didier Clement, Ronan Rivallan, Mathias Tahi, Joseph Moroh Akaza, Bertrand Pitollat, Karina Gramacho, Angelique D'Hont, Dominique Brunel, Diogenes Infante, Ismael Kebe, Pierre Costet, Rod Wing, W. Richard McCombie, Emmanuel Guiderdoni, Francis Quetier, Olivier Panaud, Patrick Wincker, Stephanie Bocs, and Claire Lanaud. The genome of Theobroma cacao. *Nature Genetics*, 43(2):101–108, February 2011. ISSN 1061-4036. doi: 10.1038/ng.736. URL http://www.nature.com/ng/journal/v43/n2/full/ng.736.html.

Michael Ashburner, Catherine A. Ball, Judith A. Blake, David Botstein, Heather Butler, J. Michael Cherry, Allan P. Davis, Kara Dolinski, Selina S. Dwight, Janan T. Eppig, Midori A. Harris, David P. Hill, Laurie Issel-Tarver, Andrew Kasarskis, Suzanna Lewis, John C. Matese, Joel E. Richardson, Martin Ringwald, Gerald M. Rubin, and Gavin Sherlock. Gene Ontology: tool for the unification of biology. *Nature Genetics*, 25(1):25–29, May 2000. ISSN 1061-4036. doi: 10.1038/75556. URL http://www.nature.com/doifinder/10.1038/75556.

Susanna Atwell, Yu S. Huang, Bjarni J. Vilhjalmsson, Glenda Willems, Matthew Horton, Yan Li, Dazhe Meng, Alexander Platt, Aaron M. Tarone, Tina T. Hu, Rong Jiang, N. Wayan Muliyati, Xu Zhang, Muhammad Ali Amer, Ivan Baxter, Benjamin Brachi, Joanne Chory, Caroline Dean, Marilyne Debieu, Juliette de Meaux, Joseph R. Ecker, Nathalie Faure, Joel M. Kniskern, Jonathan D. G. Jones, Todd Michael, Adnane Nemri, Fabrice Roux, David E. Salt, Chunlao Tang, Marco Todesco, M. Brian Traw, Detlef Weigel, Paul Marjoram, Justin O. Borevitz, Joy Bergelson, and Magnus Nordborg. Genome-wide association study of 107 phenotypes in Arabidopsis thaliana inbred lines. *Nature*, 465(7298):627–631, June 2010. ISSN 1476-4687. doi: 10.1038/nature08800.

Gwenael Badis, Michael F. Berger, Anthony A. Philippakis, Shaheynoor Talukder, Andrew R. Gehrke, Savina A. Jaeger, Esther T. Chan, Genita Metzler, Anastasia Vedenko, Xiaoyu Chen, Hanna Kuznetsov, Chi-Fong Wang, David Coburn, Daniel E. Newburger, Quaid Morris, Timothy R. Hughes, and Martha L. Bulyk. Diversity and Complexity in DNA Recognition by Transcription Factors. *Science*, 324(5935):1720–1723, June 2009. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.1162327. URL http://science.sciencemag.org/content/324/5935/1720.

Timothy L. Bailey, Mikael Boden, Fabian A. Buske, Martin Frith, Charles E. Grant, Luca Clementi, Jingyuan Ren, Wilfred W. Li, and William S. Noble. MEME SUITE: tools for motif discovery and searching. *Nucleic Acids Research*, 37(Web Server issue):W202–208, July 2009. ISSN 1362-4962. doi: 10.1093/nar/gkp335.

Salma Balazadeh, Hamad Siddiqui, Annapurna D. Allu, Lilian P. Matallana-Ramirez, Camila Caldana, Mohammad Mehrnia, Maria-Inas Zanor, Barbara Kohler, and Bernd Mueller-Roeber. A gene regulatory network controlled by the NAC transcription factor ANAC092/AtNAC2/ORE1 during salt-promoted senescence. *The Plant Journal*, 62(2):250–264, April 2010. ISSN 1365-313X. doi: 10.1111/j.1365-313X.2010.04151.x. URL http://onlinelibrary.wiley.com/doi/10.1111/j.1365-313X.2010.04151.x/abstract.

Laura Baxter, Aleksey Jironkin, Richard Hickman, Jay Moore, Christopher Barrington, Peter Krusche, Nigel P. Dyer, Vicky Buchanan-Wollaston, Alexander Tiskin, Jim Beynon, Katherine Denby, and Sascha Ott. Conserved Noncoding Sequences Highlight Shared Components of Regulatory Networks in Dicotyledonous Plants. *The Plant Cell*, 24(10):3949–3965, October 2012. ISSN , 1532-298X. doi: 10.1105/tpc.112.103010. URL http://www.plantcell.org/content/24/10/3949.

Jeffrey L. Bennetzin and Michael Freeling. Grasses as a single genetic system: genome composition, collinearity and compatibility. *Trends in Genetics*, 9(8):259–261, August 1993. ISSN 0168-9525. doi: 10.1016/0168-9525(93)90001-X. URL http://www.sciencedirect.com/science/article/pii/016895259390001X.

Eugene Berezikov, Victor Guryev, Ronald H.A. Plasterk, and Edwin Cuppen. CONREAL: Conserved Regulatory Elements Anchored Alignment Algorithm for Identification of Transcription Factor Binding Sites by Phylogenetic Footprinting. *Genome Research*, 14(1):170–178, January 2004. ISSN 1088-9051. doi: 10.1101/gr.1642804. URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC314294/.

Michael F. Berger and Martha L. Bulyk. Universal protein-binding microarrays for the comprehensive characterization of the DNA-binding specificities of transcription factors. *Nature Protocols*, 4 (3):393–411, March 2009. ISSN 1754-2189. doi: 10.1038/nprot.2008.195. URL http://www.nature.com/nprot/journal/v4/n3/full/nprot.2008.195.html.

Michael F. Berger, Anthony A. Philippakis, Aaron M. Qureshi, Fangxue S. He, Preston W. Estep, and Martha L. Bulyk. Compact, universal DNA microarrays to comprehensively determine transcription-factor binding site specificities. *Nature Biotechnology*, 24(11):1429–1435, November 2006. ISSN 1087-0156. doi: 10.1038/nbt1246. URL http://www.nature.com/nbt/journal/v24/n11/full/nbt1246.html.

Michael F. Berger, Gwenael Badis, Andrew R. Gehrke, Shaheynoor Talukder, Anthony A. Philippakis, Lourdes Pena-Castillo, Trevis M. Alleyne, Sanie Mnaimneh, Olga B. Botvinnik, Esther T. Chan, Faiqua Khalid, Wen Zhang, Daniel Newburger, Savina A. Jaeger, Quaid D. Morris, Martha L. Bulyk, and Timothy R. Hughes. Variation in Homeodomain DNA Binding Revealed by High-Resolution Analysis of Sequence Preferences. *Cell*, 133(7):1266–1276, June 2008. ISSN 0092-8674, 1097-4172. doi: 10.1016/j.cell.2008.05.024. URL http://www.cell.com/cell/abstract/S0092-8674(08)00683-1.

Mathieu Blanchette and Martin Tompa. Discovery of Regulatory Elements by a Computational Method for Phylogenetic Footprinting. *Genome Research*, 12(5):739–748, May 2002. ISSN 1088-9051, 1549-5469. doi: 10.1101/gr.6902. URL http://genome.cshlp.org/content/12/5/739.

Mathieu Blanchette, W. James Kent, Cathy Riemer, Laura Elnitski, Arian F. A. Smit, Krishna M. Roskin, Robert Baertsch, Kate Rosenbloom, Hiram Clawson, Eric D. Green, David Haussler, and Webb Miller. Aligning Multiple Genomic Sequences With the Threaded Blockset Aligner. *Genome Research*, 14(4):708–715, April 2004. ISSN 1088-9051, 1549-5469. doi: 10.1101/gr.1933104. URL http://genome.cshlp.org/content/14/4/708.

Sylvia F. Boj, Joan Marc Servitja, David Martin, Martin Rios, Iannis Talianidis, Roderic Guigo, and Jorge Ferrer. Functional Targets of the Monogenic Diabetes Transcription Factors HNF-1Îś and HNF-4Îś Are Highly Conserved Between Mice and Humans. *Diabetes*, 58(5):1245–1253, May 2009. ISSN 0012-1797. doi: 10.2337/db08-0812. URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2671044/.

Nathalie Bolduc and Sarah Hake. The Maize Transcription Factor KNOTTED1 Directly Regulates the Gibberellin Catabolism Gene ga2ox1. *The Plant Cell*, 21(6):1647–1658, June 2009. ISSN , 1532-298X. doi: 10.1105/tpc.109.068221. URL http://www.plantcell.org/content/21/6/1647.

Nathalie Bolduc, Alper Yilmaz, Maria Katherine Mejia-Guerra, Kengo Morohashi, Devin O'Connor, Erich Grotewold, and Sarah Hake. Unraveling the KNOTTED1 regulatory network in maize meristems. *Genes & Development*, 26(15):1685–1690, August 2012. ISSN 1549-5477. doi: 10.1101/gad.193433.112.

Robert K. Bradley, Xiao-Yong Li, Cole Trapnell, Stuart Davidson, Lior Pachter, Hou Cheng Chu, Leath A. Tonkin, Mark D. Biggin, and Michael B. Eisen. Binding Site Turnover Produces Pervasive Quantitative Changes in Transcription Factor Binding between Closely Related Drosophila Species. *PLOS Biol*, 8(3): e1000343, 2010. ISSN 1545-7885. doi: 10.1371/journal.pbio.1000343. URL http://journals.plos.org/plosbiology/article?id=10.1371/journal.pbio.1000343.

Siobhan M. Brady, Lifang Zhang, Molly Megraw, Natalia J. Martinez, Eric Jiang, Charles S. Yi, Weilin Liu, Anna Zeng, Mallorie Taylor-Teeples, Dahae Kim, Sebastian Ahnert, Uwe Ohler, Doreen Ware, Albertha J. M. Walhout, and Philip N. Benfey. A stele-enriched gene regulatory network in the Arabidopsis root. *Molecular Systems Biology*, 7(1):459, January 2011. ISSN 1744-4292, 1744-4292. doi: 10.1038/msb.2010.114. URL http://msb.embopress.org/content/7/1/459.

Florian Brioudes, Caroline Joly, Judit Szacsi, Emilie Varaud, Julie Leroux, Floriant Bellvert, Cedric Bertrand, and Mohammed Bendahmane. Jasmonate controls late development stages of petal growth in Arabidopsis thaliana. *The Plant Journal*, 60(6):1070–1080, December 2009. ISSN 1365-313X. doi: 10.1111/j.1365-313X.2009.04023.x. URL http://onlinelibrary.wiley.com/doi/10.1111/j.1365-313X.2009.04023.x/abstract.

Qingyun Bu, Hongling Jiang, Chang-Bao Li, Qingzhe Zhai, Jie Zhang, Xiaoyan Wu, Jiaqiang Sun, Qi Xie, and Chuanyou Li. Role of the Arabidopsis thaliana NAC transcription factors ANAC019 and ANAC055 in regulating jasmonic acid-signaled defense responses. *Cell Research*, 18(7):756–767, July 2008. ISSN 1001-0602. doi: 10.1038/cr.2008.53. URL http://www.nature.com/cr/journal/v18/n7/full/cr200853a.html.

Diane Burgess and Michael Freeling. The Most Deeply Conserved Noncoding Sequences in Plants Serve Similar Functions to Those in Vertebrates Despite Large Differences in Evolutionary Rates. *The Plant Cell Online*, page tpc.113.121905, March 2014. ISSN , 1532-298X. doi: 10.1105/tpc.113.121905. URL http://www.plantcell.org/content/early/2014/03/27/tpc.113.121905.

C. Steven Carmack, Lee Ann McCue, Lee A. Newberg, and Charles E. Lawrence. PhyloScan: identification of transcription factor binding sites using cross-species evidence. *Algorithms for molecular biology: AMB*, 2:1, 2007. ISSN 1748-7188. doi: 10.1186/1748-7188-2-1.

Marie-Edith Chaboute, Bernadette Clement, and Gabriel Philipps. S Phase and Meristem-specific Expression of the TobaccoRNR1b Gene Is Mediated by an E2f Element Located in the 5 Leader Sequence. *Journal of Biological Chemistry*, 277(20):17845–17851, May 2002. ISSN 0021-9258, 1083-351X. doi: 10.1074/jbc.M200959200. URL http://www.jbc.org/content/277/20/17845.

Katherine Noelani Chang, Shan Zhong, Matthew T. Weirauch, Gary Hon, Mattia Pelizzola, Hai Li, Shao-shan Carol Huang, Robert J. Schmitz, Mark A. Urich, Dwight Kuo, Joseph R. Nery, Hong Qiao, Ally Yang, Abdullah Jamali, Huaming Chen, Trey Ideker, Bing Ren, Ziv Bar-Joseph, Timothy R. Hughes, and Joseph R. Ecker. Temporal transcriptional response to ethylene gas drives growth hormone cross-regulation in Arabidopsis. *eLife*, 2:e00675, June 2013. ISSN 2050-084X. doi: 10.7554/eLife.00675. URL https://elifesciences.org/content/2/e00675v1.

Tanit Chavalit, Pinnara Rojvirat, Sureeporn Muangsawat, and Sarawut Jitrapakdee. Hepatocyte nuclear factor 4 alfa regulates the expression of the murine pyruvate carboxylase gene through the HNF4-specific binding motif in its proximal promoter. *Biochimica et Biophysica Acta (BBA) - Gene Regulatory Mechanisms*, 1829 (10):987–999, October 2013. ISSN 1874-9399. doi: 10.1016/j.bbagrm.2013.05.001. URL http://www.sciencedirect.com/science/article/pii/S1874939913000898.

Yong Cheng, Zhihai Ma, Bong-Hyun Kim, Weisheng Wu, Philip Cayting, Alan P. Boyle, Vasavi Sundaram, Xiaoyun Xing, Nergiz Dogan, Jingjing Li, Ghia Euskirchen, Shin Lin, Yiing Lin, Axel Visel, Trupti Kawli, Xinqiong Yang, Dorrelyn Patacsil, Cheryl A. Keller, Belinda Giardine, Mouse ENCODE Consortium, Anshul Kundaje, Ting Wang, Len A. Pennacchio, Zhiping Weng, Ross C. Hardison, and Michael P. Snyder. Principles of regulatory information conservation between mouse and human. *Nature*, 515 (7527):371–375, November 2014. ISSN 1476-4687. doi: 10.1038/nature13985.

Arabidopsis consoritium. Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. *Nature*, 408(6814): 796–815, December 2000. ISSN 0028-0836. doi: 10.1038/35048692. URL http://www.nature.com/nature/journal/v408/n6814/full/408796a0.html.

The Potato Genome Sequencing Consortium. Genome sequence and analysis of the tuber crop potato. *Nature*, 475 (7355):189–195, July 2011. ISSN 0028-0836. doi: 10.1038/nature10158. URL http://www.nature.com/nature/journal/v475/n7355/full/nature10158.html.

The Tomato Genome Consortium. The tomato genome sequence provides insights into fleshy fruit evolution. *Nature*, 485 (7400):635–641, May 2012. ISSN 0028-0836. doi: 10.1038/nature11119. URL http://www.nature.com/nature/journal/v485/n7400/full/nature11119.html.

A. Cornish-Bowden. Nomenclature for incompletely specified bases in nucleic acid sequences: recommendations 1984. *Nucleic Acids Research*, 13(9):3021–3030, May 1985. ISSN 0305-1048.

Modan K. Das and Ho-Kwok Dai. A survey of DNA motif finding algorithms. *BMC Bioinformatics*, 8(7):S21, 2007. ISSN 1471-2105. doi: 10.1186/1471-2105-8-S7-S21. URL http://dx.doi.org/10.1186/1471-2105-8-S7-S21.

Ramana V. Davuluri, Hao Sun, Saranyan K. Palaniswamy, Nicole Matthews, Carlos Molina, Mike Kurtz, and Erich Grotewold. AGRIS: Arabidopsis Gene Regulatory Information Server, an information resource of Arabidopsis cis-regulatory elements and transcription factors. *BMC Bioinformatics*, 4:25, 2003. ISSN 1471-2105. doi: 10.1186/1471-2105-4-25. URL http://dx.doi.org/10.1186/1471-2105-4-25.

Stefanie De Bodt, Guenter Theissen, and Yves Van de Peer. Promoter Analysis of MADS-Box Genes in Eudicots Through Phylogenetic Footprinting. *Molecular Biology and Evolution*, 23 (6):1293–1303, June 2006. ISSN 0737-4038, 1537-1719. doi: 10.1093/molbev/msk016. URL http://mbe.oxfordjournals.org/content/23/6/1293.

Stefanie De Bodt, Diana Carvajal, Jens Hollunder, Joost Van den Cruyce, Sara Movahedi, and Dirk Inze. CORNET: A User-Friendly Tool for Data Mining and Integration. *Plant Physiology*, 152(3): 1167–1179, March 2010. ISSN , 1532-2548. doi: 10.1104/pp.109.147215. URL http://www.plantphysiol.org/content/152/3/1167.

Dieter De Witte, Michiel Van Bel, Pieter Audenaert, Piet Demeester, Bart Dhoedt, Klaas Vandepoele, and Jan Fostier. A Parallel, Distributed-Memory Framework for Comparative Motif Discovery. In Roman Wyrzykowski, Jack Dongarra, Konrad Karczewski, and Jerzy Wazniewski, editors, *Parallel Processing and Applied Mathematics*, volume 8385, pages 268–277. Springer Berlin Heidelberg, Berlin, Heidelberg, 2014. ISBN 978-3-642-55194-9 978-3-642-55195-6. URL http://link.springer.com/10.1007/978-3-642-55195-6_25.

Dieter De Witte, Jan Van de Velde, Dries Decap, Michiel Van Bel, Pieter Audenaert, Piet Demeester, Bart Dhoedt, Klaas Vandepoele, and Jan Fostier. BLSSpeller: exhaustive comparative discovery of conserved cis-regulatory elements. *Bioinformatics*, 31(23):3758–3766, December 2015. ISSN 1367-4803, 1460-2059. doi: 10.1093/bioinformatics/btv466. URL http://bioinformatics.oxfordjournals.org/content/31/23/3758.

Jeffrey Dean and Sanjay Ghemawat. MapReduce: Simplified Data Processing on Large Clusters. *Commun. ACM*, 51(1):107–113, January 2008. ISSN 0001-0782. doi: 10.1145/1327452.1327492. URL http://doi.acm.org/10.1145/1327452.1327492.

Weiwei Deng, Hua Ying, Chris A. Helliwell, Jennifer M. Taylor, W. James Peacock, and Elizabeth S. Dennis. FLOWERING LOCUS C (FLC) regulates development pathways throughout the life cycle of Arabidopsis. *Proceedings of the National Academy of Sciences*, 108(16):6680–6685, April 2011. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.1103175108. URL http://www.pnas.org/content/108/16/6680.

Bart Deplancke, Denis Dupuy, Marc Vidal, and Albertha J. M. Walhout. A Gateway-Compatible Yeast One-Hybrid System. *Genome Research*, 14(10b):2093–2101, October 2004. ISSN 1088-9051, 1549-5469. doi: 10.1101/gr.2445504. URL http://genome.cshlp.org/content/14/10b/2093.

Emmanouil T. Dermitzakis and Andrew G. Clark. Evolution of Transcription Factor Binding Sites in Mammalian Gene Regulatory Regions: Conservation and Turnover. *Molecular Biology and Evolution*, 19(7):1114–1121, July 2002. ISSN 0737-4038, 1537-1719. URL http://mbe.oxfordjournals.org/content/19/7/1114.

Juliane C. Dohm, AndrÃl' E. Minoche, Daniela HoltgrÃd'we, Salvador Capella-GutiÃl'rrez, Falk Zakrzewski, Hakim Tafer, Oliver Rupp, Thomas Rosleff SÃűrensen, Ralf Stracke, Richard Reinhardt, Alexander Goesmann, Thomas Kraft, Britta Schulz, Peter F. Stadler, Thomas Schmidt, Toni Gabaldon, Hans Lehrach, Bernd Weisshaar, and Heinz Himmelbauer. The genome of the recently domesticated crop plant sugar beet (Beta vulgaris). *Nature*, 505(7484):546–549, January 2014. ISSN 0028-0836. doi: 10.1038/nature12817. URL http://www.nature.com/nature/journal/v505/n7484/full/nature12817.html.

Arnaud Droit, Charles Cheung, and Raphael Gottardo. rMAT - an R/Bioconductor package for analyzing ChIP-chip experiments. *Bioinformatics*, 26(5):678–679, March 2010. ISSN 1367-4803, 1460-2059. doi: 10.1093/bioinformatics/btq023. URL http://bioinformatics.oxfordjournals.org/content/26/5/678.

Ming-Rui Duan, Jie Nan, Yu-He Liang, Peng Mao, Lu Lu, Lanfen Li, Chunhong Wei, Luhua Lai, Yi Li, and Xiao-Dong Su. DNA binding mechanism revealed by high resolution crystal structure of Arabidopsis thaliana WRKY1 protein. *Nucleic Acids Research*, 35(4): 1145–1154, February 2007. ISSN 0305-1048, 1362-4962. doi: 10.1093/nar/gkm001. URL http://nar.oxfordjournals.org/content/35/4/1145.

Christian Dubos, Jose Le Gourrierec, Antoine Baudry, Gunnar Huep, Elodie Lanet, Isabelle Debeaujon, Jean-Marc Routaboul, Alessandro Alboresi, Bernd Weisshaar, and LoÃíc Lepiniec. MYBL2 is a new regulator of flavonoid biosynthesis in Arabidopsis thaliana. *The Plant Journal: For Cell and Molecular Biology*, 55(6): 940–953, September 2008. ISSN 1365-313X. doi: 10.1111/j.1365-313X.2008.03564.x.

Olivier Elemento and Saeed Tavazoie. Fast and systematic genome-wide discovery of conserved regulatory elements using a non-alignment based approach. *Genome Biology*, 6(2):R18, 2005. ISSN 1474-760X. doi: 10.1186/gb-2005-6-2-r18.

Andrew D. Ellington and Jack W. Szostak. In vitro selection of RNA molecules that bind specific ligands. *Nature*, 346(6287): 818–822, August 1990. ISSN 0028-0836. doi: 10.1038/346818a0. URL http://www.nature.com/nature/journal/v346/n6287/abs/346818a0.html.

Laura Elnitski, Victor X. Jin, Peggy J. Farnham, and Steven J. M. Jones. Locating mammalian transcription factor binding sites: A survey of computational and experimental techniques. *Genome Research*, 16(12):1455–1464, December 2006. ISSN 1088-9051, 1549-5469. doi: 10.1101/gr.4140006. URL http://genome.cshlp.org/content/16/12/1455.

Heidi A Ernst, Addie Nina Olsen, Karen Skriver, Sine Larsen, and Leila Lo Leggio. Structure of the conserved domain of ANAC, a member of the NAC family of transcription factors. *EMBO reports*, 5(3):297–303, March 2004. ISSN 1469-3178. doi: 10.1038/sj.embor.7400093. URL http://onlinelibrary.wiley.com/doi/10.1038/sj.embor.7400093/abstract.

Jason Ernst and Manolis Kellis. Discovery and characterization of chromatin states for systematic annotation of the human genome. *Nature Biotechnology*, 28(8):817–825, August 2010. ISSN 1087-0156. doi: 10.1038/nbt.1662. URL http://www.nature.com/nbt/journal/v28/n8/full/nbt.1662.html.

Jason Ernst, Qasim K. Beg, Krin A. Kay, Gabor Balazsi, Zoltan N. Oltvai, and Ziv Bar-Joseph. A semi-supervised method for predicting transcription factor-gene interactions in Escherichia coli. *PLoS computational biology*, 4(3):e1000044, March 2008. ISSN 1553-7358. doi: 10.1371/journal.pcbi.1000044.

Eleazar Eskin and Pavel A. Pevzner. Finding composite regulatory patterns in DNA sequences. *Bioinformatics (Oxford, England)*, 18 Suppl 1:S354–363, 2002. ISSN 1367-4803.

Laurence Ettwiller, Benedict Paten, Marcel Souren, Felix Loosli, Jochen Wittbrodt, and Ewan Birney. The discovery, positioning and verification of a set of transcription-associated motifs in vertebrates. *Genome Biology*, 6:R104, 2005. ISSN 1474-760X. doi: 10.1186/gb-2005-6-12-r104. URL `http://dx.doi.org/10.1186/gb-2005-6-12-r104`.

Peggy J. Farnham. Insights from genomic profiling of transcription factors. *Nature Reviews Genetics*, 10(9):605–616, September 2009. ISSN 1471-0056. doi: 10.1038/nrg2636. URL `http://www.nature.com/nrg/journal/v10/n9/abs/nrg2636.html`.

Thilia Ferrier, Jose Tomas Matus, Jian Jin, and Jose Luis Riechmann. Arabidopsis paves the way: genomic and network analyses in crops. *Current Opinion in Biotechnology*, 22(2):260–270, April 2011. ISSN 0958-1669. doi: 10.1016/j.copbio.2010.11.010. URL `http://www.sciencedirect.com/science/article/pii/S0958166910002284`.

William W. Fisher, Jingyi Jessica Li, Ann S. Hammonds, James B. Brown, Barret D. Pfeiffer, Richard Weiszmann, Stewart MacArthur, Sean Thomas, John A. Stamatoyannopoulos, Michael B. Eisen, Peter J. Bickel, Mark D. Biggin, and Susan E. Celniker. DNA regions bound at low occupancy by transcription factors do not drive patterned reporter gene expression in Drosophila. *Proceedings of the National Academy of Sciences of the United States of America*, 109(52):21330–21335, December 2012. ISSN 1091-6490. doi: 10.1073/pnas.1209589110.

Jose M. Franco-Zorrilla, Irene Lopez-Vidriero, Jose L. Carrasco, Marta Godoy, Pablo Vera, and Roberto Solano. DNA-binding specificities of plant transcription factors and their potential to define target genes. *Proceedings of the National Academy of Sciences*, 111(6):2367–2372, February 2014. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.1316278111. URL `http://www.pnas.org/content/111/6/2367`.

Kelly A. Frazer, Heng Tao, Kazutoyo Osoegawa, Pieter J. de Jong, Xiyin Chen, Mark F. Doherty, and David R. Cox. Noncoding Sequences Conserved in a Limited Number of Mammals in the SIM2 Interval are Frequently Functional. *Genome Research*, 14(3):367–372, March 2004. ISSN 1088-9051. doi: 10.1101/gr.1961204. URL `http://www.ncbi.nlm.nih.gov/pmc/articles/PMC353216/`.

Michael Freeling and Shabarinath Subramaniam. Conserved noncoding sequences (CNSs) in higher plants. *Current Opinion in Plant Biology*, 12(2):126–132, April 2009. ISSN 1369-5266. doi: 10.1016/j.pbi.2009.01.005. URL `http://www.sciencedirect.com/science/article/pii/S1369526609000089`.

Michael Freeling, Lakshmi Rapaka, Eric Lyons, Brent Pedersen, and Brian C. Thomas. G-Boxes, Bigfoot Genes, and Environmental Response: Characterization of Intragenomic Conserved Noncoding Sequences in Arabidopsis. *The Plant Cell*, 19(5):1441–1457, May 2007. ISSN 1040-4651. doi: 10.1105/tpc.107.050419. URL `http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1913728/`.

Monika Fuxreiter, Istvan Simon, and Sarah Bondos. Dynamic protein-DNA recognition: beyond what can be seen. *Trends in Biochemical Sciences*, 36(8):415–423, August 2011. ISSN 0968-0004. doi: 10.1016/j.tibs.2011.04.006. URL `http://www.sciencedirect.com/science/article/pii/S0968000411000594`.

Daniel J. Gaffney, Jean-Baptiste Veyrieras, Jacob F. Degner, Roger Pique-Regi, Athma A. Pai, Gregory E. Crawford, Matthew Stephens, Yoav Gilad, and Jonathan K. Pritchard. Dissecting the regulatory architecture of gene expression QTLs. *Genome Biology*, 13(1):R7, 2012. ISSN 1474-760X. doi: 10.1186/gb-2012-13-1-r7.

Xiangchao Gan, Oliver Stegle, Jonas Behr, Joshua G. Steffen, Philipp Drewe, Katie L. Hildebrand, Rune Lyngsoe, Sebastian J. Schultheiss, Edward J. Osborne, Vipin T. Sreedharan, Andre Kahles, Regina Bohnert, Geraldine Jean, Paul Derwent, Paul Kersey, Eric J. Belfield, Nicholas P. Harberd, Eric Kemen, Christopher Toomajian, Paula X. Kover, Richard M. Clark, Gunnar Ratsch, and Richard Mott. Multiple reference genomes and transcriptomes for Arabidopsis thaliana. *Nature*, 477(7365):419–423, September 2011a. ISSN 0028-0836. doi: 10.1038/nature10414. URL `http://www.nature.com/nature/journal/v477/n7365/full/nature10414.html?cookies=accepted`.

Peng Gao, Zeyu Xin, and Zhi-Liang Zheng. The OSU1/QUA2/TSD2-Encoded Putative Methyltransferase Is a Critical Modulator of Carbon and Nitrogen Nutrient Balance Response in Arabidopsis. *PLoS ONE*, 3(1), January 2008. ISSN 1932-6203. doi: 10.1371/journal.pone.0001387. URL `http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2148111/`.

Jordi Garcia-Mas, Andrej Benjak, Walter Sanseverino, Michael Bourgeois, Gisela Mir, Victor M. Gonzalez, Elizabeth Henaff, Francisco Cimara, Luca Cozzuto, Ernesto Lowy, Tyler Alioto, Salvador Capella-Gutierrez, Jose Blanca, Joaquin Canizares, Pello Ziarsolo, Daniel Gonzalez-Ibeas, Luis Rodriguez-Moreno, Marcus Droege, Lei Du, Miguel Alvarez-Tejado, Belen Lorente-Galdos, Marta Mela, Luming Yang, Yiqun Weng, Arcadi Navarro, Tomas Marques-Bonet, Miguel A. Aranda, Fernando Nuez, Belon Pices, Toni Gabaldon, Guglielmo Roma, Roderic Guigo, Josep M. Casacuberta, Pere Aras, and Pere Puigdomenech. The genome of melon (Cucumis melo L.). *Proceedings of the National Academy of Sciences*, 109(29):11872–11877, July 2012. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.1205415109. URL `http://www.pnas.org/content/109/29/11872`.

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):3047–3060, July 1981. ISSN 0305-1048. URL `http://www.ncbi.nlm.nih.gov/pmc/articles/PMC327330/`.

Allison Gaudinier, Lifang Zhang, John S. Reece-Hoyes, Mallorie Taylor-Teeples, Li Pu, Zhijie Liu, Ghislain Breton, Jose L. Pruneda-Paz, Dahae Kim, Steve A. Kay, Albertha J. M. Walhout, Doreen Ware, and Siobhan M. Brady. Enhanced Y1h assays for Arabidopsis. *Nature Methods*, 8(12):1053–1055, December 2011. ISSN 1548-7091. doi: 10.1038/nmeth.1750. URL `http://www.nature.com/nmeth/journal/v8/n12/full/nmeth.1750.html`.

Robert Giegerich, Stefan Kurtz, and Jens Stoye. Efficient Implementation of Lazy Suffix Trees. In Jeffrey S. Vitter and Christos D. Zaroliagis, editors, *Algorithm Engineering*, number 1668 in Lecture Notes in Computer Science, pages 30–42. Springer Berlin Heidelberg, July 1999. ISBN 978-3-540-66427-7 978-3-540-48318-2. URL `http://link.springer.com/chapter/10.1007/3-540-48318-7_5`. DOI: 10.1007/3-540-48318-7_5.

Luca Giorgetti, Trevor Siggers, Guido Tiana, Greta Caprara, Samuele Notarbartolo, Teresa Corona, Manolis Pasparakis, Paolo Milani, Martha L. Bulyk, and Gioacchino Natoli. Noncooperative Interactions between Transcription Factors and Clustered DNA Binding Sites Enable Graded Transcriptional Responses to Environmental Inputs. *Molecular Cell*, 37(3):418–428, February 2010. ISSN 1097-2765. doi: 10.1016/j.molcel.2010.01.016. URL `http://www.cell.com/molecular-cell/abstract/S1097-2765(10)00045-6`.

Federico M. Giorgi, Cristian Del Fabbro, and Francesco Licausi. Comparative study of RNA-seq- and Microarray-derived coexpression networks in Arabidopsis thaliana. *Bioinformatics*, page btt053, February 2013. ISSN 1367-4803, 1460-2059. doi: 10.1093/bioinformatics/btt053. URL `http://bioinformatics.oxfordjournals.org/content/early/2013/02/01/bioinformatics.btt053`.

Marta Godoy, Jose M. Franco-Zorrilla, Julian Perez-Perez, Juan C. Oliveros, Oscar Lorenzo, and Roberto Solano. Improved protein-binding microarrays for the identification of DNA-binding specificities of transcription factors. *The Plant Journal*, 66(4):700–711, May 2011. ISSN 1365-313X. doi: 10.1111/j.1365-313X.2011.04519.x. URL http://onlinelibrary.wiley.com/doi/10.1111/j.1365-313X.2011.04519.x/abstract.

Raluca Gordan, Leelavati Narlikar, and Alexander J. Hartemink. Finding regulatory DNA motifs using alignment-free evolutionary conservation information. *Nucleic Acids Research*, 38(6):e90–e90, April 2010. ISSN 0305-1048, 1362-4962. doi: 10.1093/nar/gkp1166. URL http://nar.oxfordjournals.org/content/38/6/e90.

Raluca Gordan, Ning Shen, Iris Dror, Tianyin Zhou, John Horton, Remo Rohs, and Martha L. Bulyk. Genomic Regions Flanking E-Box Binding Sites Influence DNA Binding Specificity of bHLH Transcription Factors through DNA Shape. *Cell Reports*, 3(4):1093–1104, April 2013. ISSN 2211-1247. doi: 10.1016/j.celrep.2013.03.014. URL http://www.cell.com/cell-reports/abstract/S2211-1247(13)00121-6.

K. Goto and E. M. Meyerowitz. Function and regulation of the Arabidopsis floral homeotic gene PISTILLATA. *Genes & Development*, 8(13):1548–1560, July 1994. ISSN 0890-9369, 1549-5477. doi: 10.1101/gad.8.13.1548. URL http://genesdev.cshlp.org/content/8/13/1548.

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, December 2005. ISSN 0950-5849. doi: 10.1016/j.infsof.2005.09.005. URL http://www.sciencedirect.com/science/article/pii/S0950584905001345.

Christian A. Grove, Federico De Masi, M. Inmaculada Barrasa, Daniel E. Newburger, Mark J. Alkema, Martha L. Bulyk, and Albertha J. M. Walhout. A Multiparameter Network Reveals Extensive Divergence between C. elegans bHLH Transcription Factors. *Cell*, 138(2):314–327, July 2009. ISSN 0092-8674, 1097-4172. doi: 10.1016/j.cell.2009.04.058. URL http://www.cell.com/cell/abstract/S0092-8674(09)00519-4.

Hena Guo and Stephen P. Moose. Conserved Noncoding Sequences among Cultivated Cereal Genomes Identify Candidate Regulatory Sequence Elements and Patterns of Promoter Evolution. *The Plant Cell*, 15(5):1143–1158, May 2003. ISSN , 1532-298X. doi: 10.1105/tpc.010181. URL http://www.plantcell.org/content/15/5/1143.

Yongfeng Guo and Susheng Gan. AtNAP, a NAC family transcription factor, has an important role in leaf senescence. *The Plant Journal*, 46(4):601–612, May 2006. ISSN 1365-313X. doi: 10.1111/j.1365-313X.2006.02723.x. URL http://onlinelibrary.wiley.com/doi/10.1111/j.1365-313X.2006.02723.x/abstract.

Ross C. Hardison. Comparative Genomics. *PLOS Biol*, 1(2):e58, November 2003. ISSN 1545-7885. doi: 10.1371/journal.pbio.0000058. URL http://journals.plos.org/plosbiology/article?id=10.1371/journal.pbio.0000058.

Annabelle Haudry, Adrian E. Platts, Emilio Vello, Douglas R. Hoen, Mickael Leclercq, Robert J. Williamson, Ewa Forczek, Zoe Joly-Lopez, Joshua G. Steffen, Khaled M. Hazzouri, Ken Dewar, John R. Stinchcombe, Daniel J. Schoen, Xiaowu Wang, Jeremy Schmutz, Christopher D. Town, Patrick P. Edger, J. Chris Pires, Karen S. Schumaker, David E. Jarvis, Terezie Mandakova, Martin A. Lysak, Erik van den Bergh, M. Eric Schranz, Paul M. Harrison, Alan M. Moses, Thomas E. Bureau, Stephen I. Wright, and Mathieu Blanchette. An atlas of over 90,000 conserved noncoding sequences provides insight into crucifer regulatory regions. *Nature Genetics*, 45(8):891–898, August 2013. ISSN 1061-4036. doi: 10.1038/ng.2684. URL http://www.nature.com/ng/journal/v45/n8/full/ng.2684.html.

Guangming He, Axel A. Elling, and Xing Wang Deng. The Epigenome and Plant Development. *Annual Review of Plant Biology*, 62(1):411–435, 2011. doi: 10.1146/annurev-arplant-042110-103806. URL http://dx.doi.org/10.1146/annurev-arplant-042110-103806.

Jacques van Helden, Alma. F. Rios, and Julio Collado-Vides. Discovering regulatory elements in non-coding sequences by analysis of spaced dyads. *Nucleic Acids Research*, 28(8):1808–1818, April 2000. ISSN 0305-1048. URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC102821/.

Ken S. Heyndrickx and Klaas Vandepoele. Systematic Identification of Functional Plant Modules through the Integration of Complementary Data Sources. *Plant Physiology*, 159(3):884–901, July 2012. ISSN , 1532-2548. doi: 10.1104/pp.112.196725. URL http://www.plantphysiol.org/content/159/3/884.

Ken S. Heyndrickx, Jan Van de Velde, Congmao Wang, Detlef Weigel, and Klaas Vandepoele. A Functional and Evolutionary Perspective on Transcription Factor Binding in Arabidopsis thaliana. *The Plant Cell*, 26(10):3894–3910, October 2014. ISSN , 1532-298X. doi: 10.1105/tpc.114.130591. URL http://www.plantcell.org/content/26/10/3894.

Richard Hickman, Claire Hill, Christopher A. Penfold, Emily Breeze, Laura Bowden, Jonathan D. Moore, Peijun Zhang, Alison Jackson, Emma Cooke, Findlay Bewicke-Copley, Andrew Mead, Jim Beynon, David L. Wild, Katherine J. Denby, Sascha Ott, and Vicky Buchanan-Wollaston. A local regulatory network around three NAC transcription factors in stress responses and senescence in Arabidopsis leaves. *The Plant Journal*, 75(1):26–39, July 2013. ISSN 1365-313X. doi: 10.1111/tpj.12194. URL http://onlinelibrary.wiley.com/doi/10.1111/tpj.12194/abstract.

Kenichi Higo, Yoshihiro Ugawa, Masao Iwamoto, and Tomoko Korenaga. Plant cis-acting regulatory DNA elements (PLACE) database: 1999. *Nucleic Acids Research*, 27(1):297–300, January 1999. ISSN 0305-1048, 1362-4962. doi: 10.1093/nar/27.1.297. URL http://nar.oxfordjournals.org/content/27/1/297.

Ray L. Hong, Lynn Hamaguchi, Maximilian A. Busch, and Detlef Weigel. Regulatory Elements of the Floral Homeotic Gene AGAMOUS Identified by Phylogenetic Footprinting and Shadowing. *The Plant Cell*, 15(6):1296–1309, June 2003. ISSN , 1532-298X. doi: 10.1105/tpc.009548. URL http://www.plantcell.org/content/15/6/1296.

Patricia Hornitschek, Markus V. Kohnen, SÃl'verine Lorrain, Jacques Rougemont, Karin Ljung, Irene LÃşpez-Vidriero, JosÃl' M. Franco-Zorrilla, Roberto Solano, Martine Trevisan, Sylvain Pradervand, Ioannis Xenarios, and Christian Fankhauser. Phytochrome interacting factors 4 and 5 control seedling growth in changing light conditions by directly controlling auxin signaling. *The Plant Journal*, 71(5):699–711, September 2012. ISSN 1365-313X. doi: 10.1111/j.1365-313X.2012.05033.x. URL http://onlinelibrary.wiley.com/doi/10.1111/j.1365-313X.2012.05033.x/abstract.

Tomas Hruz, Oliver Laule, Gabor Szabo, Frans Wessendorp, Stefan Bleuler, Lukas Oertle, Peter Widmayer, Wilhelm Gruissem, and Philip Zimmermann. Genevestigator V3: A Reference Expression Database for the Meta-Analysis of Transcriptomes. *Advances in Bioinformatics*, 2008:e420747, July 2008. ISSN 1687-8027. doi: 10.1155/2008/420747. URL http://www.hindawi.com/journals/abi/2008/420747/abs/.

Honghong Hu, Mingqiu Dai, Jialing Yao, Benze Xiao, Xianghua Li, Qifa Zhang, and Lizhong Xiong. Overexpressing a NAM, ATAF, and CUC (NAC) transcription factor enhances drought resistance and salt tolerance in rice. *Proceedings of the National Academy of Sciences*, 103(35):12987–12992, August 2006. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.0604882103. URL http://www.pnas.org/content/103/35/12987.

Weichun Huang, David M. Umbach, and Leping Li. Accurate anchoring alignment of divergent sequences. *Bioinformatics*, 22 (1):29–34, January 2006. ISSN 1367-4803, 1460-2059. doi: 10.1093/bioinformatics/bti772. URL http://bioinformatics.oxfordjournals.org/content/22/1/29.

J. D. Hughes, P. W. Estep, S. Tavazoie, and G. M. Church. Computational identification of cis-regulatory elements associated with groups of functionally related genes in Saccharomyces cerevisiae. *Journal of Molecular Biology*, 296(5):1205–1214, March 2000. ISSN 0022-2836. doi: 10.1006/jmbi.2000.3519.

Daniel Hupalo and Andrew D. Kern. Conservation and functional element discovery in 20 angiosperm plant genomes. *Molecular Biology and Evolution*, 30(7):1729–1744, July 2013. ISSN 1537-1719. doi: 10.1093/molbev/mst082.

Steven Grant Hussey, Eshchar Mizrachi, Nicky M. Creux, and Alexander A. Myburg. Navigating the transcriptional roadmap regulating plant secondary cell wall deposition. *Plant Cell Biology*, 4:325, 2013. doi: 10.3389/fpls.2013.00325. URL http://journal.frontiersin.org/article/10.3389/fpls.2013.00325/full.

Van Anh Huynh-Thu, Alexandre Irrthum, Louis Wehenkel, and Pierre Geurts. Inferring Regulatory Networks from Expression Data Using Tree-Based Methods. *PLOS ONE*, 5(9):e12776, September 2010. ISSN 1932-6203. doi: 10.1371/journal.pone.0012776. URL http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0012776.

Tony Handstad, Morten Beck Rye, Finn Drablos, and PÃěl Saetrom. A ChIP-Seq Benchmark Shows That Sequence Conservation Mainly Improves Detection of Strong Transcription Factor Binding Sites. *PLOS ONE*, 6(4):e18430, April 2011. ISSN 1932-6203. doi: 10.1371/journal.pone.0018430. URL http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0018430.

Elizabeth Henaff, Cristina Vives, Benedicte Desvoyes, Ankita Chaurasia, Jordi Payet, Crisanto Gutierrez, and Josep M. Casacuberta. Extensive amplification of the E2f transcription factor binding sites by transposons during evolution of Brassica species. *The Plant Journal*, 77(6):852–862, March 2014. ISSN 1365-313X. doi: 10.1111/tpj.12434. URL http://onlinelibrary.wiley.com/doi/10.1111/tpj.12434/abstract.

Dan Choffnes Inada, Ali Bashir, Chunghau Lee, Brian C. Thomas, Cynthia Ko, Stephen A. Goff, and Michael Freeling. Conserved Noncoding Sequences in the Grasses4. *Genome Research*, 13 (9):2030–2041, September 2003. ISSN 1088-9051, 1549-5469. doi: 10.1101/gr.1280703. URL http://genome.cshlp.org/content/13/9/2030.

Toshiro Ito, Kian-Hong Ng, Tze-Soo Lim, Hao Yu, and Elliot M. Meyerowitz. The Homeotic Protein AGAMOUS Controls Late Stamen Development by Regulating a Jasmonate Biosynthetic Gene in Arabidopsis. *The Plant Cell*, 19(11):3516–3529, November 2007. ISSN , 1532-298X. doi: 10.1105/tpc.107.055467. URL http://www.plantcell.org/content/19/11/3516.

Thomas Jack, Laura L. Brockman, and Elliot 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, February 1992. ISSN 0092-8674. doi: 10.1016/0092-8674(92)90144-2. URL http://www.sciencedirect.com/science/article/pii/0092867492901442.

Olivier Jaillon, Jean-Marc Aury, Benjamin Noel, Alberto Policriti, Christian Clepet, Alberto Casagrande, Nathalie Choisne, Sebastien Aubourg, Nicola Vitulo, Claire Jubin, Alessandro Vezzi, Fabrice Legeai, Philippe Hugueney, Corinne Dasilva, David Horner, Erica Mica, Delphine Jublot, Julie Poulain, Clemence Bruyere, Alain Billault, Beatrice Segurens, Michel Gouyvenoux, Edgardo Ugarte, Federica Cattonaro, Veronique Anthouard, Virginie Vico, Cristian Del Fabbro, Michael Alaux, Gabriele Di Gaspero, Vincent Dumas, Nicoletta Felice, Sophie Paillard, Irena Juman, Marco Moroldo, Simone Scalabrin, Aurelie Canaguier, Isabelle Le Clainche, Giorgio Malacrida, Eleonore Durand, Graziano Pesole, Valerie Laucou, Philippe Chatelet, Didier Merdinoglu, Massimo Delledonne, Mario Pezzotti, Alain Lecharny, Claude Scarpelli, Francois Artiguenave, M. Enrico Pe, Giorgio Valle, Michele Morgante, Michel Caboche, Anne-Francoise Adam-Blondon, Jean Weissenbach, Francis Quatier, and Patrick Wincker. The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. *Nature*, 449 (7161):463–467, September 2007. ISSN 0028-0836. doi: 10.1038/nature06148. URL http://www.nature.com/nature/journal/v449/n7161/abs/nature06148.html.

Michael K. Jensen, Trine Kjaersgaard, Michael M. Nielsen, Pernille Galberg, Klaus Petersen, Charlotte O'Shea, and Karen Skriver. The Arabidopsis thaliana NAC transcription factor family: structure-function relationships and determinants of ANAC019 stress signalling. *Biochemical Journal*, 426(2):183–196, March 2010a. ISSN 0264-6021, 1470-8728. doi: 10.1042/BJ20091234. URL http://www.biochemj.org/content/426/2/183.

Michael Krogh Jensen, Trine Kjaersgaard, Klaus Petersen, and Karen Skriver. NAC genes. *Plant Signaling & Behavior*, 5(7):907–910, July 2010b. ISSN 1559-2316. URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3014545/.

Michael Krogh Jensen, Soren Lindemose, Federico de Masi, Julia J. Reimer, Michael Nielsen, Venura Perera, Chris T. Workman, Franziska Turck, Murray R. Grant, John Mundy, Morten Petersen, and Karen Skriver. ATAF1 transcription factor directly regulates abscisic acid biosynthetic gene NCED3 in Arabidopsis thaliana. *FEBS Open Bio*, 3:321–327, 2013. ISSN 2211-5463. doi: 10.1016/j.fob.2013.07.006. URL http://www.sciencedirect.com/science/article/pii/S2211546313000399.

Jin Seo Jeong, Youn Shic Kim, Kwang Hun Baek, Harin Jung, Sun-Hwa Ha, Yang Do Choi, Minkyun Kim, Christophe Reuzeau, and Ju-Kon Kim. Root-Specific Expression of OsNAC10 Improves Drought Tolerance and Grain Yield in Rice under Field Drought Conditions. *Plant Physiology*, 153(1):185–197, May 2010. ISSN , 1532-2548. doi: 10.1104/pp.110.154773. URL http://www.plantphysiol.org/content/153/1/185.

Bo Jiang, Jun S. Liu, and Martha L. Bulyk. Bayesian hierarchical model of protein-binding microarray k-mer data reduces noise and identifies transcription factor subclasses and preferred k-mers. *Bioinformatics*, 29(11):1390–1398, June 2013. ISSN 1367-4803, 1460-2059. doi: 10.1093/bioinformatics/btt152. URL http://bioinformatics.oxfordjournals.org/content/29/11/1390.

H. Jin, E. Cominelli, P. Bailey, A. Parr, F. Mehrtens, J. Jones, C. Tonelli, B. Weisshaar, and C. Martin. Transcriptional repression by AtMYB4 controls production of UV-protecting sunscreens in Arabidopsis. *The EMBO journal*, 19(22):6150–6161, November 2000. ISSN 0261-4189. doi: 10.1093/emboj/19.22.6150.

Jinpu Jin, He Zhang, Lei Kong, Ge Gao, and Jingchu Luo. PlantTFDB 3.0: a portal for the functional and evolutionary study of plant transcription factors. *Nucleic Acids Research*, 42(Database issue):D1182–D1187, January 2014. ISSN 0305-1048. doi: 10.1093/nar/gkt1016. URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3965000/.

Jinpu Jin, Kun He, Xing Tang, Zhe Li, Le Lv, Yi Zhao, Jingchu Luo, and Ge Gao. An Arabidopsis Transcriptional Regulatory Map Reveals Distinct Functional and Evolutionary Features of Novel Transcription Factors. *Molecular Biology and Evolution*, page msv058, March 2015. ISSN 0737-4038, 1537-1719. doi: 10.1093/molbev/msv058. URL http://mbe.oxfordjournals.org/content/early/2015/04/01/molbev.msv058.

Thomas H. Jukes and Motoo Kimura. Evolutionary constraints and the neutral theory. *Journal of Molecular Evolution*, 21(1): 90–92, 1984. ISSN 0022-2844, 1432-1432. doi: 10.1007/BF02100633. URL http://link.springer.com/article/10.1007/BF02100633.

James T. Kadonaga, Katherine A. Jones, and Robert Tjian. Promoter-specific activation of RNA polymerase II transcription by Sp1. *Trends in Biochemical Sciences*, 11(1):20–23, January 1986. ISSN 0968-0004. doi: 10.1016/0968-0004(86)90226-4. URL http://www.sciencedirect.com/science/article/pii/0968000486902264.

Sateesh Kagale, Matthew G. Links, and Kevin Rozwadowski. Genome-Wide Analysis of Ethylene-Responsive Element Binding Factor-Associated Amphiphilic Repression Motif-Containing Transcriptional Regulators in Arabidopsis. *Plant Physiology*, 152 (3):1109–1134, March 2010. ISSN , 1532-2548. doi: 10.1104/pp.109.151704. URL http://www.plantphysiol.org/content/152/3/1109.

Nicholas J. Kaplinsky, David M. Braun, Jon Penterman, Stephen A. Goff, and Michael Freeling. Utility and distribution of conserved noncoding sequences in the grasses. *Proceedings of the National Academy of Sciences*, 99(9):6147–6151, April 2002. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.052139599. URL http://www.pnas.org/content/99/9/6147.

Kerstin Kaufmann, Alice Pajoro, and Gerco C. Angenent. Regulation of transcription in plants: mechanisms controlling developmental switches. *Nature Reviews Genetics*, 11(12):830–842, December 2010a. ISSN 1471-0056. doi: 10.1038/nrg2885. URL http://www.nature.com/nrg/journal/v11/n12/full/nrg2885.html.

Kerstin Kaufmann, Frank Wellmer, Jose M. Muino, Thilia Ferrier, Samuel E. Wuest, Vijaya Kumar, Antonio Serrano-Mislata, Francisco Madueno, Pawel Krajewski, Elliot M. Meyerowitz, Gerco C. Angenent, and Jose Luis Riechmann. Orchestration of Floral Initiation by APETALA1. *Science*, 328(5974):85–89, April 2010b. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.1185244. URL http://science.sciencemag.org/content/328/5974/85.

Manolis Kellis, Nick Patterson, Matthew Endrizzi, Bruce Birren, and Eric S. Lander. Sequencing and comparison of yeast species to identify genes and regulatory elements. *Nature*, 423(6937):241–254, May 2003. ISSN 0028-0836. doi: 10.1038/nature01644.

Joost J. B. Keurentjes, Gerco C. Angenent, Marcel Dicke, Vitor A. P. Martins Dos Santos, Jaap Molenaar, Wim H. van der Putten, Peter C. de Ruiter, Paul C. Struik, and Bart P. H. J. Thomma. Redefining plant systems biology: from cell to ecosystem. *Trends in Plant Science*, 16(4):183–190, April 2011. ISSN 1360-1385. doi: 10.1016/j.tplants.2010.12.002. URL http://www.cell.com/trends/plant-science/abstract/S1360-1385(10)00255-4.

Pouya Kheradpour, Alexander Stark, Sushmita Roy, and Manolis Kellis. Reliable prediction of regulator targets using 12 Drosophila genomes. *Genome Research*, 17(12):1919–1931, December 2007. ISSN 1088-9051, 1549-5469. doi: 10.1101/gr.7090407. URL http://genome.cshlp.org/content/17/12/1919.

Won-Chan Kim, Joo-Yeol Kim, Jae-Heung Ko, Jungmook Kim, and Kyung-Hwan Han. Transcription factor MYB46 is an obligate component of the transcriptional regulatory complex for functional expression of secondary wall-associated cellulose synthases in Arabidopsis thaliana. *Journal of Plant Physiology*, 170(15):1374–1378, October 2013. ISSN 1618-1328. doi: 10.1016/j.jplph.2013.04.012.

Trine Kjaersgaard, Michael K. Jensen, Michael W. Christiansen, Per Gregersen, Birthe B. Kragelund, and Karen Skriver. Senescence-associated Barley NAC (NAM, ATAF1,2, CUC) Transcription Factor Interacts with Radical-induced Cell Death 1 through a Disordered Regulatory Domain. *Journal of Biological Chemistry*, 286 (41):35418–35429, October 2011. ISSN 0021-9258, 1083-351X. doi: 10.1074/jbc.M111.247221. URL http://www.jbc.org/content/286/41/35418.

Bjorn Kloosterman, Jose A. Abelenda, Mara del Mar Carretero Gomez, Marian Oortwijn, Jan M. de Boer, Krissana Kowitwanich, Beatrix M. Horvath, Herman J. van Eck, Cezary Smaczniak, Salome Prat, Richard G. F. Visser, and Christian W. B. Bachem. Naturally occurring allele diversity allows potato cultivation in northern latitudes. *Nature*, 495 (7440):246–250, March 2013. ISSN 0028-0836. doi: 10.1038/nature11912. URL http://www.nature.com/nature/journal/v495/n7440/full/nature11912.html.

Yuichi Kodama, Martin Shumway, Rasko Leinonen, and International Nucleotide Sequence Database Collaboration. The Sequence Read Archive: explosive growth of sequencing data. *Nucleic Acids Research*, 40(Database issue):D54–56, January 2012. ISSN 1362-4962. doi: 10.1093/nar/gkr854.

Eugene V. Koonin. Orthologs, Paralogs, and Evolutionary Genomics. *Annual Review of Genetics*, 39(1):309–338, 2005. doi: 10.1146/annurev.genet.39.073003.114725. URL http://dx.doi.org/10.1146/annurev.genet.39.073003.114725.

Paula Korkuc, Jos H. M. Schippers, and Dirk Walther. Characterization and Identification of cis-Regulatory Elements in Arabidopsis Based on Single-Nucleotide Polymorphism Information. *Plant Physiology*, 164(1):181–200, January 2014. ISSN , 1532-2548. doi: 10.1104/pp.113.229716. URL http://www.plantphysiol.org/content/164/1/181.

Tony Kouzarides. Chromatin Modifications and Their Function. *Cell*, 128(4):693–705, February 2007. ISSN 0092-8674, 1097-4172. doi: 10.1016/j.cell.2007.02.005. URL http://www.cell.com/cell/abstract/S0092-8674(07)00184-5.

Nicholas Krom and Wusirika Ramakrishna. Comparative analysis of divergent and convergent gene pairs and their expression patterns in rice, Arabidopsis, and populus. *Plant Physiology*, 147(4):1763–1773, August 2008. ISSN 0032-0889. doi: 10.1104/pp.108.122416.

Minoru Kubo, Makiko Udagawa, Nobuyuki Nishikubo, Gorou Horiguchi, Masatoshi Yamaguchi, Jun Ito, Tetsuro Mimura, Hiroo Fukuda, and Taku Demura. Transcription switches for protoxylem and metaxylem vessel formation. *Genes & Development*, 19(16):1855–1860, August 2005. ISSN 0890-9369. doi: 10.1101/gad.1331305.

Lokesh Kumar, Andrew Breakspear, Corby Kistler, Li-Jun Ma, and Xiaohui Xie. Systematic discovery of regulatory motifs in Fusarium graminearum by comparing four Fusarium genomes. *BMC Genomics*, 11:208, 2010. ISSN 1471-2164. doi: 10.1186/1471-2164-11-208. URL http://dx.doi.org/10.1186/1471-2164-11-208.

Vibhor Kumar, Masafumi Muratani, Nirmala Arul Rayan, Petra Kraus, Thomas Lufkin, Huck Hui Ng, and Shyam Prabhakar. Uniform, optimal signal processing of mapped deep-sequencing data. *Nature Biotechnology*, 31(7):615–622, July 2013. ISSN 1087-0156. doi: 10.1038/nbt.2596. URL http://www.nature.com/nbt/journal/v31/n7/full/nbt.2596.html.

Philippe Lamesch, Tanya Z. Berardini, Donghui Li, David Swarbreck, Christopher Wilks, Rajkumar Sasidharan, Robert Muller, Kate Dreher, Debbie L. Alexander, Margarita Garcia-Hernandez, Athikkattuvalasu S. Karthikeyan, Cynthia H. Lee, William D. Nelson, Larry Ploetz, Shanker Singh, April Wensel, and Eva Huala. The Arabidopsis Information Resource (TAIR): improved gene annotation and new tools. *Nucleic Acids Research*, 40(Database issue):D1202–1210, January 2012. ISSN 1362-4962. doi: 10.1093/nar/gkr1090.

Tuuli Lappalainen, Michael Sammeth, Marc R. Friedlander, Peter A. C. 't Hoen, Jean Monlong, Manuel A. Rivas, Mar Gonzalez-Porta, Natalja Kurbatova, Thasso Griebel, Pedro G. Ferreira, Matthias Barann, Thomas Wieland, Liliana Greger, Maarten van Iterson, Jonas Almlof, Paolo Ribeca, Irina Pulyakhina, Daniela Esser, Thomas Giger, Andrew Tikhonov, Marc Sultan, Gabrielle Bertier, Daniel G. MacArthur, Monkol Lek, Esther Lizano, Henk P. J. Buermans, Ismael Padioleau, Thomas Schwarzmayr, Olof Karlberg, Halit Ongen, Helena Kilpinen, Sergi Beltran, Marta Gut, Katja Kahlem, Vyacheslav Amstislavskiy, Oliver Stegle, Matti Pirinen, Stephen B. Montgomery, Peter Donnelly, Mark I.

McCarthy, Paul Flicek, Tim M. Strom, Geuvadis Consortium, Hans Lehrach, Stefan Schreiber, Ralf Sudbrak, Angel Carracedo, Stylianos E. Antonarakis, Robert Hasler, Ann-Christine Syvonen, Gert-Jan van Ommen, Alvis Brazma, Thomas Meitinger, Philip Rosenstiel, Roderic Guigo, Ivo G. Gut, Xavier Estivill, and Emmanouil T. Dermitzakis. Transcriptome and genome sequencing uncovers functional variation in humans. *Nature*, 501(7468): 506–511, September 2013. ISSN 1476-4687. doi: 10.1038/nature12531.

Yoontae Lee, Minju Kim, Jinju Han, Kyu-Hyun Yeom, Sanghyuk Lee, Sung Hee Baek, and V. Narry Kim. MicroRNA genes are transcribed by RNA polymerase II. *The EMBO journal*, 23(20):4051–4060, October 2004. ISSN 0261-4189. doi: 10.1038/sj.emboj.7600385.

V. I. Levenshtein. Binary codes capable of correcting deletions, insertions and reverseals. January 1966. URL `https://www.scienceopen.com/document?vid=f6e0fb0c-e8b1-4eb8-b01a-c44dddbf7840`.

Michal Levo and Eran Segal. In pursuit of design principles of regulatory sequences. *Nature Reviews Genetics*, 15 (7):453–468, July 2014. ISSN 1471-0056. doi: 10.1038/nrg3684. URL `http://www.nature.com/nrg/journal/v15/n7/abs/nrg3684.html`.

Baohua Li, Allison Gaudinier, Michelle Tang, Mallorie Taylor-Teeples, Ngoc T. Nham, Cyrus Ghaffari, Darik Scott Benson, Margaret Steinmann, Jennifer A. Gray, Siobhan M. Brady, and Daniel J. Kliebenstein. Promoter-Based Integration in Plant Defense Regulation1[W][OPEN]. *Plant Physiology*, 166(4):1803–1820, December 2014. ISSN 0032-0889. doi: 10.1104/pp.114.248716. URL `http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4256871/`.

Heng Li and Richard Durbin. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, 25(14): 1754–1760, July 2009. ISSN 1367-4803, 1460-2059. doi: 10.1093/bioinformatics/btp324. URL `http://bioinformatics.oxfordjournals.org/content/25/14/1754`.

Heng Li, Bob Handsaker, Alec Wysoker, Tim Fennell, Jue Ruan, Nils Homer, Gabor Marth, Goncalo Abecasis, Richard Durbin, and 1000 Genome Project Data Processing Subgroup. The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, 25(16): 2078–2079, August 2009. ISSN 1367-4803, 1460-2059. doi: 10.1093/bioinformatics/btp352. URL `http://bioinformatics.oxfordjournals.org/content/25/16/2078`.

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):1870–1874, December 1993. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.8266075. URL `http://science.sciencemag.org/content/262/5141/1870`.

Jian-Feng Li, Julie E. Norville, John Aach, Matthew McCormack, Dandan Zhang, Jenifer Bush, George M. Church, and Jen Sheen. Multiplex and homologous recombination-mediated genome editing in Arabidopsis and Nicotiana benthamiana using guide RNA and Cas9. *Nature Biotechnology*, 31(8):688–691, August 2013. ISSN 1087-0156. doi: 10.1038/nbt.2654. URL `http://www.nature.com/nbt/journal/v31/n8/full/nbt.2654.html`.

Li Li, Christian J. Stoeckert, and David S. Roos. OrthoMCL: Identification of Ortholog Groups for Eukaryotic Genomes. *Genome Research*, 13(9):2178–2189, September 2003. ISSN 1088-9051, 1549-5469. doi: 10.1101/gr.1224503. URL `http://genome.cshlp.org/content/13/9/2178`.

Ting Li, Bo Liu, Martin H. Spalding, Donald P. Weeks, and Bing Yang. High-efficiency TALEN-based gene editing produces disease-resistant rice. *Nature Biotechnology*, 30(5):390–392, May 2012a. ISSN 1087-0156. doi: 10.1038/nbt.2199. URL `http://www.nature.com/nbt/journal/v30/n5/full/nbt.2199.html`.

Wei Li, Xiao Cui, Zhaolu Meng, Xiahe Huang, Qi Xie, Heng Wu, Hailing Jin, Dabing Zhang, and Wanqi Liang. Transcriptional Regulation of Arabidopsis MIR168a and ARGONAUTE1 Homeostasis in Abscisic Acid and Abiotic Stress Responses. *Plant Physiology*, 158(3):1279–1292, March 2012b. ISSN , 1532-2548. doi: 10.1104/pp.111.188789. URL `http://www.plantphysiol.org/content/158/3/1279`.

S. Liang, M. P. Samanta, and B. A. Biegel. cWINNOWER algorithm for finding fuzzy dna motifs. *Journal of Bioinformatics and Computational Biology*, 2(1):47–60, March 2004. ISSN 0219-7200.

Kerstin Lindblad-Toh, Manuel Garber, Or Zuk, Michael F. Lin, Brian J. Parker, Stefan Washietl, Pouya Kheradpour, Jason Ernst, Gregory Jordan, Evan Mauceli, Lucas D. Ward, Craig B. Lowe, Alisha K. Holloway, Michele Clamp, Sante Gnerre, Jessica Alfoldi, Kathryn Beal, Jean Chang, Hiram Clawson, James Cuff, Federica Di Palma, Stephen Fitzgerald, Paul Flicek, Mitchell Guttman, Melissa J. Hubisz, David B. Jaffe, Irwin Jungreis, W. James Kent, Dennis Kostka, Marcia Lara, Andre L. Martins, Tim Massingham, Ida Moltke, Brian J. Raney, Matthew D. Rasmussen, Jim Robinson, Alexander Stark, Albert J. Vilella, Jiayu Wen, Xiaohui Xie, Michael C. Zody, Broad Institute Sequencing Platform and Whole Genome Assembly Team, Jen Baldwin, Toby Bloom, Chee Whye Chin, Dave Heiman, Robert Nicol, Chad Nusbaum, Sarah Young, Jane Wilkinson, Kim C. Worley, Christie L. Kovar, Donna M. Muzny, Richard A. Gibbs, Baylor College of Medicine Human Genome Sequencing Center Sequencing Team, Andrew Cree, Huyen H. Dihn, Gerald Fowler, Shalili Jhangiani, Vandita Joshi, Sandra Lee, Lora R. Lewis, Lynne V. Nazareth, Geoffrey Okwuonu, Jireh Santibanez, Wesley C. Warren, Elaine R. Mardis, George M. Weinstock, Richard K. Wilson, Genome Institute at Washington University, Kim Delehaunty, David Dooling, Catrina Fronik, Lucinda Fulton, Bob Fulton, Tina Graves, Patrick Minx, Erica Sodergren, Ewan Birney, Elliott H. Margulies, Javier Herrero, Eric D. Green, David Haussler, Adam Siepel, Nick Goldman, Katherine S. Pollard, Jakob S. Pedersen, Eric S. Lander, and Manolis Kellis. A high-resolution map of human evolutionary constraint using 29 mammals. *Nature*, 478(7370):476–482, October 2011. ISSN 1476-4687. doi: 10.1038/nature10530.

Soren Lindemose, Charlotte O'Shea, Michael Krogh Jensen, and Karen Skriver. Structure, Function and Networks of Transcription Factors Involved in Abiotic Stress Responses. *International Journal of Molecular Sciences*, 14(3):5842–5878, March 2013. doi: 10.3390/ijms14035842. URL `http://www.mdpi.com/1422-0067/14/3/5842`.

Soren Lindemose, Michael K. Jensen, Jan Van de Velde, Charlotte O'Shea, Ken S. Heyndrickx, Christopher T. Workman, Klaas Vandepoele, Karen Skriver, and Federico De Masi. A DNA-binding-site landscape and regulatory network analysis for NAC transcription factors in Arabidopsis thaliana. *Nucleic Acids Research*, 42(12):7681–7693, August 2014. ISSN 0305-1048. doi: 10.1093/nar/gku502. URL `http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4081100/`.

Lijun Liu, Trevor Ramsay, Matthew Zinkgraf, David Sundell, Nathaniel Robert Street, Vladimir Filkov, and Andrew Groover. A resource for characterizing genome-wide binding and putative target genes of transcription factors expressed during secondary growth and wood formation in Populus. *The Plant Journal: For Cell and Molecular Biology*, 82(5):887–898, June 2015a. ISSN 1365-313X. doi: 10.1111/tpj.12850.

Lijun Liu, Matthew Zinkgraf, H. Earl Petzold, Eric P. Beers, Vladimir Filkov, and Andrew Groover. The Populus ARBORKNOX1 homeodomain transcription factor regulates woody growth through binding to evolutionarily conserved target genes of diverse function. *The New Phytologist*, 205(2):682–694, January 2015b. ISSN 1469-8137. doi: 10.1111/nph.13151.

Wen Xian Liu, Hua Liang Liu, Zhi Jian Chai, Xiu Ping Xu, Yan Ru Song, and Le Qing Qu. Evaluation of seed storage-protein gene 5Õ untranslated regions in enhancing gene expression in transgenic rice seed. *Theoretical and Applied Genetics*, 121(7):1267–1274, June 2010. ISSN 0040-5752, 1432-2242. doi: 10.

1007/s00122-010-1386-6. URL http://link.springer.com/article/10.1007/s00122-010-1386-6.

X. Liu, D. L. Brutlag, and J. S. Liu. BioProspector: discovering conserved DNA motifs in upstream regulatory regions of co-expressed genes. *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, pages 127–138, 2001. ISSN 2335-6936.

Chongyuan Luo, David J. Sidote, Yi Zhang, Randall A. Kerstetter, Todd P. Michael, and Eric Lam. Integrative analysis of chromatin states in Arabidopsis identified potential regulatory mechanisms for natural antisense transcript production. *The Plant Journal*, 73(1):77–90, January 2013. ISSN 1365-313X. doi: 10.1111/tpj.12017. URL http://onlinelibrary.wiley.com/doi/10.1111/tpj.12017/abstract.

Eric Lyons and Michael Freeling. How to usefully compare homologous plant genes and chromosomes as DNA sequences. *The Plant Journal*, 53(4):661–673, February 2008. ISSN 1365-313X. doi: 10.1111/j.1365-313X.2007.03326.x. URL http://onlinelibrary.wiley.com/doi/10.1111/j.1365-313X.2007.03326.x/abstract.

Chuang Ma and Xiangfeng Wang. Application of the Gini Correlation Coefficient to Infer Regulatory Relationships in Transcriptome Analysis. *Plant Physiology*, 160(1):192–203, September 2012. ISSN , 1532-2548. doi: 10.1104/pp.112.201962. URL http://www.plantphysiol.org/content/160/1/192.

Dongyun Ma, Dexiang Sun, Chenyang Wang, Yaoguang Li, and Tiancai Guo. Expression of flavonoid biosynthesis genes and accumulation of flavonoid in wheat leaves in response to drought stress. *Plant physiology and biochemistry: PPB / SociÃl'tÃl' franÃğaise de physiologie vÃl'gÃl'tale*, 80:60–66, July 2014. ISSN 1873-2690. doi: 10.1016/j.plaphy.2014.03.024.

Kenzie D. MacIsaac and Ernest Fraenkel. Practical Strategies for Discovering Regulatory DNA Sequence Motifs. *PLOS Comput Biol*, 2(4):e36, April 2006. ISSN 1553-7358. doi: 10.1371/journal.pcbi.0020036. URL http://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.0020036.

Daniel Marbach, James C. Costello, Robert Koffner, Nicole M. Vega, Robert J. Prill, Diogo M. Camacho, Kyle R. Allison, DREAM5 Consortium, Manolis Kellis, James J. Collins, and Gustavo Stolovitzky. Wisdom of crowds for robust gene network inference. *Nature Methods*, 9(8):796–804, August 2012a. ISSN 1548-7105. doi: 10.1038/nmeth.2016.

Daniel Marbach, Sushmita Roy, Ferhat Ay, Patrick E. Meyer, Rogerio Candeias, Tamer Kahveci, Christopher A. Bristow, and Manolis Kellis. Predictive regulatory models in Drosophila melanogaster by integrative inference of transcriptional networks. *Genome Research*, 22(7):1334–1349, July 2012b. ISSN 1088-9051, 1549-5469. doi: 10.1101/gr.127191.111. URL http://genome.cshlp.org/content/22/7/1334.

Antonio Marco, Charlotte Konikoff, Timothy L. Karr, and Sudhir Kumar. Relationship between gene co-expression and sharing of transcription factor binding sites in Drosophila melanogaster. *Bioinformatics*, 25(19):2473–2477, October 2009. ISSN 1367-4803, 1460-2059. doi: 10.1093/bioinformatics/btp462. URL http://bioinformatics.oxfordjournals.org/content/25/19/2473.

Laurent Marsan and Marie-France Sagot. Extracting Structured Motifs Using a Suffix Tree-Algorithms and Application to Promoter Consensus Identification. In *Proceedings of the Fourth Annual International Conference on Computational Molecular Biology*, RECOMB '00, pages 210–219, New York, NY, USA, 2000. ACM. ISBN 978-1-58113-186-4. doi: 10.1145/332306.332553. URL http://doi.acm.org/10.1145/332306.332553.

Tobias Marschall and Sven Rahmann. Efficient exact motif discovery. *Bioinformatics*, 25(12):i356–i364, June 2009. ISSN 1367-4803, 1460-2059. doi: 10.1093/bioinformatics/btp188. URL http://bioinformatics.oxfordjournals.org/content/25/12/i356.

Federico De Masi, Christian A. Grove, Anastasia Vedenko, Andreu Alibos, Stephen S. Gisselbrecht, Luis Serrano, Martha L. Bulyk, and Albertha J. M. Walhout. Using a structural and logics systems approach to infer bHLHâĂŞDNA binding specificity determinants. *Nucleic Acids Research*, 39(11):4553–4563, June 2011. ISSN 0305-1048, 1362-4962. doi: 10.1093/nar/gkr070. URL http://nar.oxfordjournals.org/content/39/11/4553.

Akihiro Matsui, Junko Ishida, Taeko Morosawa, Yoshiki Mochizuki, Eli Kaminuma, Takaho A. Endo, Masanori Okamoto, Eiji Nambara, Maiko Nakajima, Makiko Kawashima, Masakazu Satou, Jong-Myong Kim, Norio Kobayashi, Tetsuro Toyoda, Kazuo Shinozaki, and Motoaki Seki. Arabidopsis Transcriptome Analysis under Drought, Cold, High-Salinity and ABA Treatment Conditions using a Tiling Array. *Plant and Cell Physiology*, 49(8):1135–1149, August 2008. ISSN 0032-0781, 1471-9053. doi: 10.1093/pcp/pcn101. URL http://pcp.oxfordjournals.org/content/49/8/1135.

Matthew T. Maurano, Richard Humbert, Eric Rynes, Robert E. Thurman, Eric Haugen, Hao Wang, Alex P. Reynolds, Richard Sandstrom, Hongzhu Qu, Jennifer Brody, Anthony Shafer, Fidencio Neri, Kristen Lee, Tanya Kutyavin, Sandra Stehling-Sun, Audra K. Johnson, Theresa K. Canfield, Erika Giste, Morgan Diegel, Daniel Bates, R. Scott Hansen, Shane Neph, Peter J. Sabo, Shelly Heimfeld, Antony Raubitschek, Steven Ziegler, Chris Cotsapas, Nona Sotoodehnia, Ian Glass, Shamil R. Sunyaev, Rajinder Kaul, and John A. Stamatoyannopoulos. Systematic Localization of Common Disease-Associated Variation in Regulatory DNA. *Science*, 337(6099):1190–1195, September 2012. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.1222794. URL http://science.sciencemag.org/content/337/6099/1190.

Maria Katherine Mejia-Guerra, Marcelo Pomeranz, Kengo Morohashi, and Erich Grotewold. From plant gene regulatory grids to network dynamics. *Biochimica et Biophysica Acta (BBA) - Gene Regulatory Mechanisms*, 1819(5):454–465, May 2012. ISSN 1874-9399. doi: 10.1016/j.bbagrm.2012.02.016. URL http://www.sciencedirect.com/science/article/pii/S1874939912000685.

Xiangdong Meng, Michael H. Brodsky, and Scot A. Wolfe. A bacterial one-hybrid system for determining the DNA-binding specificity of transcription factors. *Nature Biotechnology*, 23(8):988–994, August 2005. ISSN 1087-0156. doi: 10.1038/nbt1120. URL http://www.nature.com/nbt/journal/v23/n8/full/nbt1120.html.

Margit Menges, Lars Hennig, Wilhelm Gruissem, and James A. H. Murray. Genome-wide gene expression in an Arabidopsis cell suspension. *Plant Molecular Biology*, 53(4):423–442, November 2003. ISSN 0167-4412. doi: 10.1023/B:PLAN.0000019059.56489.ca.

Sabeeha S. Merchant, Simon E. Prochnik, Olivier Vallon, Elizabeth H. Harris, Steven J. Karpowicz, George B. Witman, Astrid Terry, Asaf Salamov, Lillian K. Fritz-Laylin, Laurence Marechal-Drouard, Wallace F. Marshall, Liang-Hu Qu, David R. Nelson, Anton A. Sanderfoot, Martin H. Spalding, Vladimir V. Kapitonov, Qinghu Ren, Patrick Ferris, Erika Lindquist, Harris Shapiro, Susan M. Lucas, Jane Grimwood, Jeremy Schmutz, Pierre Cardol, Heriberto Cerutti, Guillaume Chanfreau, Chun-Long Chen, Valerie Cognat, Martin T. Croft, Rachel Dent, Susan Dutcher, Emilio Fernandez, Hideya Fukuzawa, David Gonzalez-Ballester, Diego GonzÃąlez-Halphen, Armin Hallmann, Marc Hanikenne, Michael Hippler, William Inwood, Kamel Jabbari, Ming Kalanon, Richard Kuras, Paul A. Lefebvre, Stephane D. Lemaire, Alexey V. Lobanov, Martin Lohr, Andrea Manuell, Iris Meier, Laurens Mets, Maria Mittag, Telsa Mittelmeier, James V. Moroney, Jeffrey Moseley, Carolyn Napoli, Aurora M. Nedelcu, Krishna Niyogi, Sergey V. Novoselov, Ian T. Paulsen, Greg Pazour, Saul Purton, Jean-Philippe Ral, Diego Mauricio Riaso-Pachan, Wayne Riekhof, Linda Rymarquis, Michael Schroda, David Stern, James Umen, Robert Willows, Nedra Wilson, Sara Lana Zimmer, Jens Allmer, Janneke Balk, Katerina Bisova, Chong-Jian Chen, Marek Elias, Karla Gendler, Charles Hauser, Mary Rose Lamb, Heidi Ledford, Joanne C. Long, Jun Minagawa, M. Dudley Page, Junmin Pan, Wirulda Pootakham, Sanja

Roje, Annkatrin Rose, Eric Stahlberg, Aimee M. Terauchi, Pinfen Yang, Steven Ball, Chris Bowler, Carol L. Dieckmann, Vadim N. Gladyshev, Pamela Green, Richard Jorgensen, Stephen Mayfield, Bernd Mueller-Roeber, Sathish Rajamani, Richard T. Sayre, Peter Brokstein, Inna Dubchak, David Goodstein, Leila Hornick, Y. Wayne Huang, Jinal Jhaveri, Yigong Luo, Diego MartÃ■nez, Wing Chi Abby Ngau, Bobby Otillar, Alexander Poliakov, Aaron Porter, Lukasz Szajkowski, Gregory Werner, Kemin Zhou, Igor V. Grigoriev, Daniel S. Rokhsar, and Arthur R. Grossman. The Chlamydomonas genome reveals the evolution of key animal and plant functions. *Science (New York, N.Y.)*, 318(5848):245–250, October 2007. ISSN 1095-9203. doi: 10.1126/science.1143609.

Rachel S. Meyer and Michael D. Purugganan. Evolution of crop species: genetics of domestication and diversification. *Nature Reviews Genetics*, 14(12):840–852, December 2013. ISSN 1471-0056. doi: 10.1038/nrg3605. URL `http://www.nature.com/nrg/journal/v14/n12/full/nrg3605.html`.

Ray Ming, Shaobin Hou, Yun Feng, Qingyi Yu, Alexandre Dionne-Laporte, Jimmy H. Saw, Pavel Senin, Wei Wang, Benjamin V. Ly, Kanako L. T. Lewis, Steven L. Salzberg, Lu Feng, Meghan R. Jones, Rachel L. Skelton, Jan E. Murray, Cuixia Chen, Wubin Qian, Junguo Shen, Peng Du, Moriah Eustice, Eric Tong, Haibao Tang, Eric Lyons, Robert E. Paull, Todd P. Michael, Kerr Wall, Danny W. Rice, Henrik Albert, Ming-Li Wang, Yun J. Zhu, Michael Schatz, Niranjan Nagarajan, Ricelle A. Acob, Peizhu Guan, Andrea Blas, Ching Man Wai, Christine M. Ackerman, Yan Ren, Chao Liu, Jianmei Wang, Jianping Wang, Jong-Kuk Na, Eugene V. Shakirov, Brian Haas, Jyothi Thimmapuram, David Nelson, Xiyin Wang, John E. Bowers, Andrea R. Gschwend, Arthur L. Delcher, Ratnesh Singh, Jon Y. Suzuki, Savarni Tripathi, Kabi Neupane, Hairong Wei, Beth Irikura, Maya Paidi, Ning Jiang, Wenli Zhang, Gernot Presting, Aaron Windsor, Rafael Navajas-PÃl'rez, Manuel J. Torres, F. Alex Feltus, Brad Porter, Yingjun Li, A. Max Burroughs, Ming-Cheng Luo, Lei Liu, David A. Christopher, Stephen M. Mount, Paul H. Moore, Tak Sugimura, Jiming Jiang, Mary A. Schuler, Vikki Friedman, Thomas Mitchell-Olds, Dorothy E. Shippen, Claude W. dePamphilis, Jeffrey D. Palmer, Michael Freeling, Andrew H. Paterson, Dennis Gonsalves, Lei Wang, and Maqsudul Alam. The draft genome of the transgenic tropical fruit tree papaya (Carica papaya Linnaeus). *Nature*, 452(7190):991–996, April 2008. ISSN 1476-4687. doi: 10.1038/nature06856.

Nobutaka Mitsuda, Toru Hisabori, Kunio Takeyasu, and Masa H. Sato. VOZ; Isolation and Characterization of Novel Vascular Plant Transcription Factors with a One-Zinc Finger from Arabidopsis thaliana. *Plant and Cell Physiology*, 45(7):845–854, July 2004. ISSN 0032-0781, 1471-9053. doi: 10.1093/pcp/pch101. URL `http://pcp.oxfordjournals.org/content/45/7/845`.

Nobutaka Mitsuda, Motoaki Seki, Kazuo Shinozaki, and Masaru Ohme-Takagi. The NAC Transcription Factors NST1 and NST2 of Arabidopsis Regulate Secondary Wall Thickenings and Are Required for Anther Dehiscence. *The Plant Cell*, 17(11):2993–3006, November 2005. ISSN , 1532-298X. doi: 10.1105/tpc.105.036004. URL `http://www.plantcell.org/content/17/11/2993`.

Gaurav D. Moghe and Shin-Han Shiu. The causes and molecular consequences of polyploidy in flowering plants. *Annals of the New York Academy of Sciences*, 1320:16–34, July 2014. ISSN 1749-6632. doi: 10.1111/nyas.12466.

Carlos Molina and Erich Grotewold. Genome wide analysis of Arabidopsis core promoters. *BMC Genomics*, 6:25, 2005. ISSN 1471-2164. doi: 10.1186/1471-2164-6-25. URL `http://dx.doi.org/10.1186/1471-2164-6-25`.

Jose M. Muino, Suzanne de Bruijn, Alice Pajoro, Koen Geuten, Martin Vingron, Gerco C. Angenent, and Kerstin Kaufmann. Evolution of DNA-binding sites of a floral master regulatory transcription factor. *Molecular Biology and Evolution*, page msv210, October 2015. ISSN 0737-4038, 1537-1719. doi: 10.1093/molbev/msv210. URL `http://mbe.oxfordjournals.org/content/early/2015/10/01/molbev.msv210`.

Alexander A. Myburg, Dario Grattapaglia, Gerald A. Tuskan, Uffe Hellsten, Richard D. Hayes, Jane Grimwood, Jerry Jenkins, Erika Lindquist, Hope Tice, Diane Bauer, David M. Goodstein, Inna Dubchak, Alexandre Poliakov, Eshchar Mizrachi, Anand R. K. Kullan, Steven G. Hussey, Desre Pinard, Karen van der Merwe, Pooja Singh, Ida van Jaarsveld, Orzenil B. Silva-Junior, Roberto C. Togawa, Marilia R. Pappas, Danielle A. Faria, Carolina P. Sansaloni, Cesar D. Petroli, Xiaohan Yang, Priya Ranjan, Timothy J. Tschaplinski, Chu-Yu Ye, Ting Li, Lieven Sterck, Kevin Vanneste, Florent Murat, Marcal Soler, Holone San Clemente, Naijib Saidi, Hua Cassan-Wang, Christophe Dunand, Charles A. Hefer, Erich Bornberg-Bauer, Anna R. Kersting, Kelly Vining, Vindhya Amarasinghe, Martin Ranik, Sushma Naithani, Justin Elser, Alexander E. Boyd, Aaron Liston, Joseph W. Spatafora, Palitha Dharmwardhana, Rajani Raja, Christopher Sullivan, Elisson Romanel, Marcio Alves-Ferreira, Carsten Kulheim, William Foley, Victor Carocha, Jorge Paiva, David Kudrna, Sergio H. Brommonschenkel, Giancarlo Pasquali, Margaret Byrne, Philippe Rigault, Josquin Tibbits, Antanas Spokevicius, Rebecca C. Jones, Dorothy A. Steane, Renee. Vaillancourt, Brad M. Potts, Fourie Joubert, Kerrie Barry, Georgios J. Pappas, Steven H. Strauss, Pankaj Jaiswal, Jacqueline Grima-Pettenati, Jerome Salse, Yves Van de Peer, Daniel S. Rokhsar, and Jeremy Schmutz. The genome of Eucalyptus grandis. *Nature*, 510(7505):356–362, June 2014. ISSN 0028-0836. doi: 10.1038/nature13308. URL `http://www.nature.com/nature/journal/v510/n7505/full/nature13308.html`.

Duarte Muller, Gregor Schmitz, and Klaus Theres. Blind Homologous R2r3 Myb Genes Control the Pattern of Lateral Meristem Initiation in Arabidopsis. *The Plant Cell*, 18(3):586–597, March 2006. ISSN 1040-4651. doi: 10.1105/tpc.105.038745. URL `http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1383635/`.

So Nakagawa, Stephen S. Gisselbrecht, Julia M. Rogers, Daniel L. Hartl, and Martha L. Bulyk. DNA-binding specificity changes in the evolution of forkhead transcription factors. *Proceedings of the National Academy of Sciences*, 110(30):12349–12354, July 2013. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.1310430110. URL `http://www.pnas.org/content/110/30/12349`.

Norihito Nakamichi, Takatoshi Kiba, Mari Kamioka, Takamasa Suzuki, Takafumi Yamashino, Tetsuya Higashiyama, Hitoshi Sakakibara, and Takeshi Mizuno. Transcriptional repressor PRR5 directly regulates clock-output pathways. *Proceedings of the National Academy of Sciences*, 109(42):17123–17128, October 2012. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.1205156109. URL `http://www.pnas.org/content/109/42/17123`.

Kazuo Nakashima, Hironori Takasaki, Junya Mizoi, Kazuo Shinozaki, and Kazuko Yamaguchi-Shinozaki. NAC transcription factors in plant abiotic stress responses. *Biochimica et Biophysica Acta (BBA) - Gene Regulatory Mechanisms*, 1819(2):97–103, February 2012. ISSN 1874-9399. doi: 10.1016/j.bbagrm.2011.10.005. URL `http://www.sciencedirect.com/science/article/pii/S1874939911001817`.

Nazra Naouar, Klaas Vandepoele, Tim Lammens, Tineke Casneuf, Georg Zeller, Paul van Hummelen, Detlef Weigel, Gunnar Ratsch, Dirk Inze, Martin Kuiper, Lieven De Veylder, and Marnik Vuylsteke. Quantitative RNA expression analysis with Affymetrix Tiling 1.0r arrays identifies new E2f target genes. *The Plant Journal: For Cell and Molecular Biology*, 57(1):184–194, January 2009. ISSN 1365-313X. doi: 10.1111/j.1365-313X.2008.03662.x.

Daniel E. Newburger and Martha L. Bulyk. UniPROBE: an online database of protein binding microarray data on proteinâĂŞDNA interactions. *Nucleic Acids Research*, 37(suppl 1):D77–D82, January 2009. ISSN 0305-1048, 1362-4962. doi: 10.1093/nar/gkn660. URL `http://nar.oxfordjournals.org/content/37/suppl_1/D77`.

Marcus B. Noyes, Ryan G. Christensen, Atsuya Wakabayashi, Gary D. Stormo, Michael H. Brodsky, and Scot A. Wolfe. Analysis of Homeodomain Specificities Allows the Family-wide Prediction of Preferred Recognition Sites. *Cell*, 133(7):1277–1289, June 2008. ISSN 0092-8674, 1097-4172. doi: 10.1016/j.cell.

2008.05.023. URL http://www.cell.com/cell/abstract/S0092-8674(08)00682-X.

Duncan T. Odom, Robin D. Dowell, Elizabeth S. Jacobsen, William Gordon, Timothy W. Danford, Kenzie D. MacIsaac, P. Alexander Rolfe, Caitlin M. Conboy, David K. Gifford, and Ernest Fraenkel. Tissue-Specific Transcriptional Regulation has Diverged Significantly between Human and Mouse. *Nature genetics*, 39(6):730–732, June 2007. ISSN 1061-4036. doi: 10.1038/ng2047. URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3797512/.

Eunkyoo Oh, Jia-Ying Zhu, and Zhi-Yong Wang. Interaction between BZR1 and PIF4 integrates brassinosteroid and environmental responses. *Nature Cell Biology*, 14(8):802–809, August 2012. ISSN 1465-7392. doi: 10.1038/ncb2545. URL http://www.nature.com/ncb/journal/v14/n8/full/ncb2545.html.

Kyoko Ohashi-Ito, Yoshihisa Oda, and Hiroo Fukuda. Arabidopsis VASCULAR-RELATED NAC-DOMAIN6 Directly Regulates the Genes That Govern Programmed Cell Death and Secondary Wall Formation during Xylem Differentiation. *The Plant Cell*, 22(10):3461–3473, October 2010. ISSN , 1532-298X. doi: 10.1105/tpc.110.075036. URL http://www.plantcell.org/content/22/10/3461.

Addie N. Olsen, Heidi A. Ernst, Leila Lo Leggio, and Karen Skriver. DNA-binding specificity and molecular functions of NAC transcription factors. *Plant Science*, 169(4):785–797, October 2005. ISSN 0168-9452. doi: 10.1016/j.plantsci.2005.05.035. URL http://www.sciencedirect.com/science/article/pii/S0168945205002049.

Addie Nina Olsen, Heidi Asschenfeldt Ernst, Leila Lo Leggio, Eva Johansson, Sine Larsen, and Karen Skriver. Preliminary crystallographic analysis of the NAC domain of ANAC, a member of the plant-specific NAC transcription factor family. *Acta Crystallographica Section D Biological Crystallography*, 60(1):112–115, January 2004. ISSN 0907-4449. doi: 10.1107/S0907444903022029. URL http://scripts.iucr.org/cgi-bin/paper?S0907444903022029.

Ronan C. O'Malley, Shao-shan Carol Huang, Liang Song, Mathew G. Lewsey, Anna Bartlett, Joseph R. Nery, Mary Galli, Andrea Gallavotti, and Joseph R. Ecker. Cistrome and Epicistrome Features Shape the Regulatory DNA Landscape. *Cell*, 165(5):1280–1292, May 2016. ISSN 0092-8674, 1097-4172. doi: 10.1016/j.cell.2016.04.038. URL http://www.cell.com/cell/abstract/S0092-8674(16)30481-0.

Xinhao Ouyang, Jigang Li, Gang Li, Bosheng Li, Beibei Chen, Huaishun Shen, Xi Huang, Xiaorong Mo, Xiangyuan Wan, Rongcheng Lin, Shigui Li, Haiyang Wang, and Xing Wang Deng. Genome-Wide Binding Site Analysis of FAR-RED ELONGATED HYPOCOTYL3 Reveals Its Novel Function in Arabidopsis Development. *The Plant Cell*, 23(7):2514–2535, July 2011. ISSN , 1532-298X. doi: 10.1105/tpc.111.085126. URL http://www.plantcell.org/content/23/7/2514.

Brian Palenik, Jane Grimwood, Andrea Aerts, Pierre Rouze, Asaf Salamov, Nicholas Putnam, Chris Dupont, Richard Jorgensen, Evelyne Derelle, Stephane Rombauts, Kemin Zhou, Robert Otillar, Sabeeha S. Merchant, Sheila Podell, Terry Gaasterland, Carolyn Napoli, Karla Gendler, Andrea Manuell, Vera Tai, Olivier Vallon, Gwenael Piganeau, Severine Jancek, Marc Heijde, Kamel Jabbari, Chris Bowler, Martin Lohr, Steven Robbens, Gregory Werner, Inna Dubchak, Gregory J. Pazour, Qinghu Ren, Ian Paulsen, Chuck Delwiche, Jeremy Schmutz, Daniel Rokhsar, Yves Van de Peer, Herve Moreau, and Igor V. Grigoriev. The tiny eukaryote Ostreococcus provides genomic insights into the paradox of plankton speciation. *Proceedings of the National Academy of Sciences*, 104(18):7705–7710, May 2007. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.0611046104. URL http://www.pnas.org/content/104/18/7705.

Alessia Para, Ying Li, Amy Marshall-Colin, Kranthi Varala, Nancy J. Francoeur, Tara M. Moran, Molly B. Edwards, Christopher Hackley, Bastiaan O. R. Bargmann, Kenneth D. Birnbaum, W. Richard McCombie, Gabriel Krouk, and Gloria M. Coruzzi. Hit-and-run transcriptional control by bZIP1 mediates rapid nutrient signaling in Arabidopsis. *Proceedings of the National Academy of Sciences of the United States of America*, 111(28):10371–10376, July 2014. ISSN 0027-8424. doi: 10.1073/pnas.1404657111. URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4104873/.

Mahesh Patil, S. V. Ramu, P. Jathish, Rohini Sreevathsa, P. Chandrashekar Reddy, T. G. Prasad, and M. Udayakumar. Overexpression of AtNAC2 (ANAC092) in groundnut (Arachis hypogaea L.) improves abiotic stress tolerance. *Plant Biotechnology Reports*, 8(2):161–169, November 2013. ISSN 1863-5466, 1863-5474. doi: 10.1007/s11816-013-0305-0. URL http://link.springer.com/article/10.1007/s11816-013-0305-0.

Rob Patro, Stephen M. Mount, and Carl Kingsford. Sailfish enables alignment-free isoform quantification from RNA-seq reads using lightweight algorithms. *Nature Biotechnology*, 32(5):462–464, May 2014. ISSN 1087-0156. doi: 10.1038/nbt.2862. URL http://www.nature.com/nbt/journal/v32/n5/full/nbt.2862.html.

G. Pavesi, G. Mauri, and G. Pesole. An algorithm for finding signals of unknown length in DNA sequences. *Bioinformatics (Oxford, England)*, 17 Suppl 1:S207–214, 2001. ISSN 1367-4803.

Jennifer Pfluger and Doris Wagner. Histone modifications and dynamic regulation of genome accessibility in plants. *Current opinion in plant biology*, 10(6):645–652, December 2007. ISSN 1369-5266. doi: 10.1016/j.pbi.2007.07.013. URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2140274/.

Emma Picot, Peter Krusche, Alexander Tiskin, Isabelle CarrÃl', and Sascha Ott. Evolutionary analysis of regulatory sequences (EARS) in plants. *The Plant Journal*, 64(1):165–176, October 2010. ISSN 1365-313X. doi: 10.1111/j.1365-313X.2010.04314.x. URL http://onlinelibrary.wiley.com/doi/10.1111/j.1365-313X.2010.04314.x/abstract.

Daniel A Pollard, Casey M Bergman, Jens Stoye, Susan E Celniker, and Michael B Eisen. Benchmarking tools for the alignment of functional noncoding DNA. *BMC Bioinformatics*, 5:6, January 2004. ISSN 1471-2105. doi: 10.1186/1471-2105-5-6. URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC344529/.

Katherine S. Pollard, Melissa J. Hubisz, Kate R. Rosenbloom, and Adam Siepel. Detection of nonneutral substitution rates on mammalian phylogenies. *Genome Research*, 20(1):110–121, January 2010. ISSN 1088-9051, 1549-5469. doi: 10.1101/gr.097857.109. URL http://genome.cshlp.org/content/20/1/110.

Jeremy Preston, Janet Wheeler, Joshua Heazlewood, Song Feng Li, and Roger W. Parish. AtMYB32 is required for normal pollen development in Arabidopsis thaliana. *The Plant Journal: For Cell and Molecular Biology*, 40(6):979–995, December 2004. ISSN 0960-7412. doi: 10.1111/j.1365-313X.2004.02280.x.

Amborella Genome Project. The Amborella Genome and the Evolution of Flowering Plants. *Science*, 342(6165):1241089, December 2013. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.1241089. URL http://science.sciencemag.org/content/342/6165/1241089.

International Rice Genome Sequencing Project. The map-based sequence of the rice genome. *Nature*, 436(7052):793–800, August 2005. ISSN 0028-0836. doi: 10.1038/nature03895. URL http://www.nature.com/nature/journal/v436/n7052/abs/nature03895.html.

Sebastian Proost, Michiel Van Bel, Lieven Sterck, Kenny Billiau, Thomas Van Parys, Yves Van de Peer, and Klaas Vandepoele. PLAZA: a comparative genomics resource to study gene and genome evolution in plants. *The Plant Cell*, 21(12):3718–3731, December 2009. ISSN 1532-298X. doi: 10.1105/tpc.109.071506.

Sebastian Proost, Pedro Pattyn, Tom Gerats, and Yves Van de Peer. Journey through the past: 150 million years of plant genome evolution. *The Plant Journal*, 66(1):58–65, April

2011. ISSN 1365-313X. doi: 10.1111/j.1365-313X.2011.04521.x. URL http://onlinelibrary.wiley.com/doi/10.1111/j.1365-313X.2011.04521.x/abstract.

Sebastian Proost, Michiel Van Bel, Dries Vaneechoutte, Yves Van de Peer, Dirk Inze, Bernd Mueller-Roeber, and Klaas Vandepoele. PLAZA 3.0: an access point for plant comparative genomics. *Nucleic Acids Research*, 43(Database issue):D974–981, January 2015. ISSN 1362-4962. doi: 10.1093/nar/gku986.

Swati Puranik, Pranav Pankaj Sahu, Prem S. Srivastava, and Manoj Prasad. NAC proteins: regulation and role in stress tolerance. *Trends in Plant Science*, 17(6):369–381, June 2012. ISSN 1360-1385. doi: 10.1016/j.tplants.2012.02.004. URL http://www.cell.com/trends/plant-science/abstract/S1360-1385(12)00042-8.

Aaron R. Quinlan and Ira M. Hall. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, 26(6):841–842, March 2010. ISSN 1367-4803, 1460-2059. doi: 10.1093/bioinformatics/btq033. URL http://bioinformatics.oxfordjournals.org/content/26/6/841.

Arun S. Rajkumar, Nicolas Denervaud, and Sebastian J. Maerkl. Mapping the fine structure of a eukaryotic promoter input-output function. *Nature Genetics*, 45(10):1207–1215, October 2013. ISSN 1061-4036. doi: 10.1038/ng.2729. URL http://www.nature.com/ng/journal/v45/n10/full/ng.2729.html.

Anna R. Reineke, Erich Bornberg-Bauer, and Jenny Gu. Evolutionary divergence and limits of conserved non-coding sequence detection in plant genomes. *Nucleic Acids Research*, 39(14):6029–6043, August 2011. ISSN 0305-1048, 1362-4962. doi: 10.1093/nar/gkr179. URL http://nar.oxfordjournals.org/content/39/14/6029.

Stefan A. Rensing, Daniel Lang, Andreas D. Zimmer, Astrid Terry, Asaf Salamov, Harris Shapiro, Tomoaki Nishiyama, Pierre-Francois Perroud, Erika A. Lindquist, Yasuko Kamisugi, Takako Tanahashi, Keiko Sakakibara, Tomomichi Fujita, Kazuko Oishi, Tadasu Shin-I, Yoko Kuroki, Atsushi Toyoda, Yutaka Suzuki, Shin-Ichi Hashimoto, Kazuo Yamaguchi, Sumio Sugano, Yuji Kohara, Asao Fujiyama, Aldwin Anterola, Setsuyuki Aoki, Neil Ashton, W. Brad Barbazuk, Elizabeth Barker, Jeffrey L. Bennetzen, Robert Blankenship, Sung Hyun Cho, Susan K. Dutcher, Mark Estelle, Jeffrey A. Fawcett, Heidrun Gundlach, Kousuke Hanada, Alexander Heyl, Karen A. Hicks, Jon Hughes, Martin Lohr, Klaus Mayer, Alexander Melkozernov, Takashi Murata, David R. Nelson, Birgit Pils, Michael Prigge, Bernd Reiss, Tanya Renner, Stephane Rombauts, Paul J. Rushton, Anton Sanderfoot, Gabriele Schween, Shin-Han Shiu, Kurt Stueber, Frederica L. Theodoulou, Hank Tu, Yves Van de Peer, Paul J. Verrier, Elizabeth Waters, Andrew Wood, Lixing Yang, David Cove, Andrew C. Cuming, Mitsuyasu Hasebe, Susan Lucas, Brent D. Mishler, Ralf Reski, Igor V. Grigoriev, Ralph S. Quatrano, and Jeffrey L. Boore. The Physcomitrella genome reveals evolutionary insights into the conquest of land by plants. *Science (New York, N.Y.)*, 319(5859):64–69, January 2008. ISSN 1095-9203. doi: 10.1126/science.1150646.

Martiniano M. Ricardi, Rodrigo M. Gonzalez, Silin Zhong, Pia G. Dominguez, Tomas Duffy, Pablo G. Turjanski, Juan D. Salgado Salter, Karina Alleva, Fernando Carrari, James J. Giovannoni, Jose M. Estevez, and Norberto D. Iusem. Genome-wide data (ChIP-seq) enabled identification of cell wall-related and aquaporin genes as targets of tomato ASR1, a drought stress-responsive transcription factor. *BMC Plant Biology*, 14:29, 2014. ISSN 1471-2229. doi: 10.1186/1471-2229-14-29. URL http://dx.doi.org/10.1186/1471-2229-14-29.

J L Riechmann, M Wang, and E M Meyerowitz. DNA-binding properties of Arabidopsis MADS domain homeotic proteins APETALA1, APETALA3, PISTILLATA and AGAMOUS. *Nucleic Acids Research*, 24(16):3134–3141, August 1996. ISSN 0305-1048. URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC146081/.

J. L. Riechmann, J. Heard, G. Martin, L. Reuber, C.-Z. Jiang, J. Keddie, L. Adam, O. Pineda, O. J. Ratcliffe, R. R. Samaha, R. Creelman, M. Pilgrim, P. Broun, J. Z. Zhang, D. Ghandehari, B. K. Sherman, and G.-L. Yu. Arabidopsis Transcription Factors: Genome-Wide Comparative Analysis Among Eukaryotes. *Science*, 290 (5499):2105–2110, December 2000. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.290.5499.2105. URL http://science.sciencemag.org/content/290/5499/2105.

Jose Luis Riechmann and Oliver J Ratcliffe. A genomic perspective on plant transcription factors. *Current Opinion in Plant Biology*, 3(5):423–434, October 2000. ISSN 1369-5266. doi: 10.1016/S1369-5266(00)00107-2. URL http://www.sciencedirect.com/science/article/pii/S1369526600001072.

Francois Roudier, Felipe Karam Teixeira, and Vincent Colot. Chromatin indexing in Arabidopsis: an epigenomic tale of tails and more. *Trends in Genetics*, 25(11):511–517, November 2009. ISSN 0168-9525. doi: 10.1016/j.tig.2009.09.013. URL http://www.cell.com/trends/genetics/abstract/S0168-9525(09)00186-3.

Emmanuelle Roulet, Stephane Busso, Anamaria A. Camargo, Andrew J. G. Simpson, Nicolas Mermod, and Philipp Bucher. High-throughput SELEX-SAGE method for quantitative modeling of transcription-factor binding sites. *Nature Biotechnology*, 20(8):831–835, August 2002. ISSN 1087-0156. doi: 10.1038/nbt718. URL http://www.nature.com/nbt/journal/v20/n8/full/nbt718.html.

Ignacio Rubio-Somoza and Detlef Weigel. Coordination of Flower Maturation by a Regulatory Circuit of Three MicroRNAs. *PLOS Genet*, 9(3):e1003374, 2013. ISSN 1553-7404. doi: 10.1371/journal.pgen.1003374. URL http://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.1003374.

Albin Sandelin, Wynand Alkema, Par Engstrom, Wyeth W. Wasserman, and Boris Lenhard. JASPAR: an open-access database for eukaryotic transcription factor binding profiles. *Nucleic Acids Research*, 32(suppl 1):D91–D94, January 2004. ISSN 0305-1048, 1362-4962. doi: 10.1093/nar/gkh012. URL http://nar.oxfordjournals.org/content/32/suppl_1/D91.

Dominic Schmidt, Michael D. Wilson, Benoit Ballester, Petra C. Schwalie, Gordon D. Brown, Aileen Marshall, Claudia Kutter, Stephen Watt, Celia P. Martinez-Jimenez, Sarah Mackay, Iannis Talianidis, Paul Flicek, and Duncan T. Odom. Five-vertebrate ChIP-seq reveals the evolutionary dynamics of transcription factor binding. *Science (New York, N.Y.)*, 328(5981):1036–1040, May 2010. ISSN 1095-9203. doi: 10.1126/science.1186176.

Jeremy Schmutz, Steven B. Cannon, Jessica Schlueter, Jianxin Ma, Therese Mitros, William Nelson, David L. Hyten, Qijian Song, Jay J. Thelen, Jianlin Cheng, Dong Xu, Uffe Hellsten, Gregory D. May, Yeisoo Yu, Tetsuya Sakurai, Taishi Umezawa, Madan K. Bhattacharyya, Devinder Sandhu, Babu Valliyodan, Erika Lindquist, Myron Peto, David Grant, Shengqiang Shu, David Goodstein, Kerrie Barry, Montona Futrell-Griggs, Brian Abernathy, Jianchang Du, Zhixi Tian, Liucun Zhu, Navdeep Gill, Trupti Joshi, Marc Libault, Anand Sethuraman, Xue-Cheng Zhang, Kazuo Shinozaki, Henry T. Nguyen, Rod A. Wing, Perry Cregan, James Specht, Jane Grimwood, Dan Rokhsar, Gary Stacey, Randy C. Shoemaker, and Scott A. Jackson. Genome sequence of the palaeopolyploid soybean. *Nature*, 463 (7278):178–183, January 2010. ISSN 0028-0836. doi: 10.1038/nature08670. URL http://www.nature.com/nature/journal/v463/n7278/full/nature08670.html.

Eunjoo Seo, Horim Lee, Jin Jeon, Hanna Park, Jungmook Kim, Yoo-Sun Noh, and Ilha Lee. Crosstalk between Cold Response and Flowering in Arabidopsis Is Mediated through the Flowering-Time Gene SOC1 and Its Upstream Negative Regulator FLC. *The Plant Cell*, 21(10):3185–3197, October 2009. ISSN , 1532-298X. doi: 10.1105/tpc.108.063883. URL http://www.plantcell.org/content/21/10/3185.

Pil Joon Seo and Chung-Mo Park. A membrane-bound NAC transcription factor as an integrator of biotic and abiotic stress signals. *Plant Signaling & Behavior*, 5(5):481–483, May 2010. ISSN null. doi: 10.4161/psb.11083. URL http://dx.doi.org/10.4161/psb.11083.

Pil Joon Seo, Mi Jung Kim, Ju-Young Park, Sun-Young Kim, Jin Jeon, Yong-Hwan Lee, Jungmook Kim, and Chung-Mo Park. Cold activation of a plasma membrane-tethered NAC transcription factor induces a pathogen resistance response in Arabidopsis. *The Plant Journal*, 61(4):661–671, February 2010. ISSN 1365-313X. doi: 10.1111/j.1365-313X.2009.04091.x. URL http://onlinelibrary.wiley.com/doi/10.1111/j.1365-313X.2009.04091.x/abstract.

Md Shamimuzzaman and Lila Vodkin. Genome-wide identification of binding sites for NAC and YABBY transcription factors and co-regulated genes during soybean seedling development by ChIP-Seq and RNA-Seq. *BMC Genomics*, 14:477, 2013. ISSN 1471-2164. doi: 10.1186/1471-2164-14-477. URL http://dx.doi.org/10.1186/1471-2164-14-477.

Yi Shang, Yongshuo Ma, Yuan Zhou, Huimin Zhang, Lixin Duan, Huiming Chen, Jianguo Zeng, Qian Zhou, Shenhao Wang, Wenjia Gu, Min Liu, Jinwei Ren, Xingfang Gu, Shengping Zhang, Ye Wang, Ken Yasukawa, Harro J. Bouwmeester, Xiaoquan Qi, Zhonghua Zhang, William J. Lucas, and Sanwen Huang. Plant science. Biosynthesis, regulation, and domestication of bitterness in cucumber. *Science (New York, N.Y.)*, 346(6213):1084–1088, November 2014. ISSN 1095-9203. doi: 10.1126/science.1259215.

Paul Shannon, Andrew Markiel, Owen Ozier, Nitin S. Baliga, Jonathan T. Wang, Daniel Ramage, Nada Amin, Benno Schwikowski, and Trey Ideker. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Research*, 13(11):2498–2504, November 2003. ISSN 1088-9051. doi: 10.1101/gr.1239303.

Rahul Siddharthan. Sigma: multiple alignment of weakly-conserved non-coding DNA sequence. *BMC Bioinformatics*, 7:143, 2006. ISSN 1471-2105. doi: 10.1186/1471-2105-7-143. URL http://dx.doi.org/10.1186/1471-2105-7-143.

Douglas H. Sieglaff, W. Augustine Dunn, Xiaohui S. Xie, Karyn Megy, Osvaldo Marinotti, and Anthony A. James. Comparative genomics allows the discovery of cis-regulatory elements in mosquitoes. *Proceedings of the National Academy of Sciences of the United States of America*, 106(9):3053–3058, March 2009. ISSN 0027-8424. doi: 10.1073/pnas.0813264106. URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2640218/.

Adam Siepel, Gill Bejerano, Jakob S. Pedersen, Angie S. Hinrichs, Minmei Hou, Kate Rosenbloom, Hiram Clawson, John Spieth, LaDeana W. Hillier, Stephen Richards, George M. Weinstock, Richard K. Wilson, Richard A. Gibbs, W. James Kent, Webb Miller, and David Haussler. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Research*, 15(8):1034–1050, August 2005. ISSN 1088-9051, 1549-5469. doi: 10.1101/gr.3715005. URL http://genome.cshlp.org/content/15/8/1034.

Trevor Siggers, Michael H. Duyzend, Jessica Reddy, Sidra Khan, and Martha L. Bulyk. Non-DNA-binding cofactors enhance DNA-binding specificity of a transcriptional regulatory complex. *Molecular Systems Biology*, 7(1):555, January 2011. ISSN 1744-4292, 1744-4292. doi: 10.1038/msb.2011.89. URL http://msb.embopress.org/content/7/1/555.

Eric D. Siggia. Computational methods for transcriptional regulation. *Current Opinion in Genetics & Development*, 15(2):214–221, April 2005. ISSN 0959-437X. doi: 10.1016/j.gde.2005.02.004.

Michael E. Smoot, Keiichiro Ono, Johannes Ruscheinski, Peng-Liang Wang, and Trey Ideker. Cytoscape 2.8: new features for data integration and network visualization. *Bioinformatics*, 27(3):431–432, February 2011. ISSN 1367-4803, 1460-2059. doi: 10.1093/bioinformatics/btq675. URL http://bioinformatics.oxfordjournals.org/content/27/3/431.

Susheng Song, Tiancong Qi, Huang Huang, and Daoxin Xie. Regulation of Stamen Development by Coordinated Actions of Jasmonate, Auxin, and Gibberellin in Arabidopsis. *Molecular Plant*, 6(4):1065–1073, July 2013. ISSN 1674-2052. doi: 10.1093/mp/sst054. URL http://www.sciencedirect.com/science/article/pii/S1674205214609023.

Alexander Stark, Michael F. Lin, Pouya Kheradpour, Jakob S. Pedersen, Leopold Parts, Joseph W. Carlson, Madeline A. Crosby, Matthew D. Rasmussen, Sushmita Roy, Ameya N. Deoras, J. Graham Ruby, Julius Brennecke, Harvard FlyBase curators, Berkeley Drosophila Genome Project, Emily Hodges, Angie S. Hinrichs, Anat Caspi, Benedict Paten, Seung-Won Park, Mira V. Han, Morgan L. Maeder, Benjamin J. Polansky, Bryanne E. Robson, Stein Aerts, Jacques van Helden, Bassem Hassan, Donald G. Gilbert, Deborah A. Eastman, Michael Rice, Michael Weir, Matthew W. Hahn, Yongkyu Park, Colin N. Dewey, Lior Pachter, W. James Kent, David Haussler, Eric C. Lai, David P. Bartel, Gregory J. Hannon, Thomas C. Kaufman, Michael B. Eisen, Andrew G. Clark, Douglas Smith, Susan E. Celniker, William M. Gelbart, and Manolis Kellis. Discovery of functional elements in 12 Drosophila genomes using evolutionary signatures. *Nature*, 450(7167):219–232, November 2007. ISSN 1476-4687. doi: 10.1038/nature06340.

Nils Ole Steffens, Claudia Galuschka, Martin Schindler, Lorenz Bulow, and Reinhard Hehl. AthaMap: an online resource for in silico transcription factor binding sites in the Arabidopsis thaliana genome. *Nucleic Acids Research*, 32(suppl 1):D368–D372, January 2004. ISSN 0305-1048, 1362-4962. doi: 10.1093/nar/gkh017. URL http://nar.oxfordjournals.org/content/32/suppl_1/D368.

Stuart Stephen, Michael Pheasant, Igor V. Makunin, and John S. Mattick. Large-scale appearance of ultraconserved elements in tetrapod genomes and slowdown of the molecular clock. *Molecular Biology and Evolution*, 25(2):402–408, February 2008. ISSN 1537-1719. doi: 10.1093/molbev/msm268.

Ralf Stracke, Hirofumi Ishihara, Gunnar Huep, Aiko Barsch, Frank Mehrtens, Karsten Niehaus, and Bernd Weisshaar. Differential regulation of closely related R2r3-MYB transcription factors controls flavonol accumulation in different parts of the Arabidopsis thaliana seedling. *The Plant Journal: For Cell and Molecular Biology*, 50(4):660–677, May 2007. ISSN 0960-7412. doi: 10.1111/j.1365-313X.2007.03078.x.

Alexander Sturn, John Quackenbush, and Zlatko Trajanoski. Genesis: cluster analysis of microarray data. *Bioinformatics*, 18(1):207–208, January 2002. ISSN 1367-4803, 1460-2059. doi: 10.1093/bioinformatics/18.1.207. URL http://bioinformatics.oxfordjournals.org/content/18/1/207.

Amarendran R Subramanian, Michael Kaufmann, and Burkhard Morgenstern. DIALIGN-TX: greedy and progressive approaches for segment-based multiple sequence alignment. *Algorithms for Molecular Biology : AMB*, 3:6, May 2008. ISSN 1748-7188. doi: 10.1186/1748-7188-3-6. URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2430965/.

Masahiro Sugiura. Plant in Vitro Transcription Systems. *Annual Review of Plant Physiology and Plant Molecular Biology*, 48(1):383–398, 1997. doi: 10.1146/annurev.arplant.48.1.383. URL http://dx.doi.org/10.1146/annurev.arplant.48.1.383.

Alessandra M. Sullivan, Andrej A. Arsovski, Janne Lempe, Kerry L. Bubb, Matthew T. Weirauch, Peter J. Sabo, Richard Sandstrom, Robert E. Thurman, Shane Neph, Alex P. Reynolds, Andrew B. Stergachis, Benjamin Vernot, Audra K. Johnson, Eric Haugen, Shawn T. Sullivan, Agnieszka Thompson, Fidencio V. Neri III, Molly Weaver, Morgan Diegel, Sanie Mnaimneh, Ally Yang, Timothy R. Hughes, Jennifer L. Nemhauser, Christine Queitsch, and John A. Stamatoyannopoulos. Mapping and Dynamics of Regulatory DNA and Transcription Factor Networks in A. thaliana. *Cell Reports*, 8(6):2015–2030, September 2014. ISSN 2211-1247. doi: 10.1016/j.celrep.2014.08.019. URL http://www.sciencedirect.com/science/article/pii/S2211124714006822.

Danilo A. Tagle, Ben F. Koop, Morris Goodman, Jerry L. Slightom, David L. Hess, and Richard T. Jones. Embryonic alfa and beta globin genes of a prosimian primate (Galago crassicaudatus). *Journal of Molecular Biology*, 203(2):439–455, September 1988. ISSN 0022-2836. doi: 10.1016/0022-2836(88)90011-3. URL http://www.sciencedirect.com/science/article/pii/0022283688900113.

Zhen Tao, Lisha Shen, Chang Liu, Lu Liu, Yuanyuan Yan, and Hao Yu. Genome-wide identification of SOC1 and SVP targets during the floral transition in Arabidopsis. *The Plant Journal*, 70(4):549–561, May 2012. ISSN 1365-313X. doi: 10.1111/j.1365-313X.2012.04919.x. URL http://onlinelibrary.wiley.com/doi/10.1111/j.1365-313X.2012.04919.x/abstract.

M. Taylor-Teeples, L. Lin, M. de Lucas, G. Turco, T. W. Toal, A. Gaudinier, N. F. Young, G. M. Trabucco, M. T. Veling, R. Lamothe, P. P. Handakumbura, G. Xiong, C. Wang, J. Corwin, A. Tsoukalas, L. Zhang, D. Ware, M. Pauly, D. J. Kliebenstein, K. Dehesh, I. Tagkopoulos, G. Breton, J. L. Pruneda-Paz, S. E. Ahnert, S. A. Kay, S. P. Hazen, and S. M. Brady. An Arabidopsis gene regulatory network for secondary cell wall synthesis. *Nature*, 517(7536):571–575, January 2015. ISSN 0028-0836. doi: 10.1038/nature14099. URL http://www.nature.com/nature/journal/v517/n7536/abs/nature14099.html.

The International Peach Genome Initiative, Ignazio Verde, Albert G. Abbott, Simone Scalabrin, Sook Jung, Shengqiang Shu, Fabio Marroni, Tatyana Zhebentyayeva, Maria Teresa Dettori, Jane Grimwood, Federica Cattonaro, Andrea Zuccolo, Laura Rossini, Jerry Jenkins, Elisa Vendramin, Lee A. Meisel, Veronique Decroocq, Bryon Sosinski, Simon Prochnik, Therese Mitros, Alberto Policriti, Guido Cipriani, Luca Dondini, Stephen Ficklin, David M. Goodstein, Pengfei Xuan, Cristian Del Fabbro, Valeria Aramini, Dario Copetti, Susana Gonzalez, David S. Horner, Rachele Falchi, Susan Lucas, Erica Mica, Jonathan Maldonado, Barbara Lazzari, Douglas Bielenberg, Raul Pirona, Mara Miculan, Abdelali Barakat, Raffaele Testolin, Alessandra Stella, Stefano Tartarini, Pietro Tonutti, Pere Aris, Ariel Orellana, Christina Wells, Dorrie Main, Giannina Vizzotto, Herman Silva, Francesco Salamini, Jeremy Schmutz, Michele Morgante, and Daniel S. Rokhsar. The high-quality draft genome of peach (Prunus persica) identifies unique patterns of genetic diversity, domestication and genome evolution. *Nature Genetics*, 45(5):487–494, May 2013. ISSN 1061-4036. doi: 10.1038/ng.2586. URL http://www.nature.com/ng/journal/v45/n5/full/ng.2586.html.

Francoise Thibaud-Nissen, Hank Wu, Todd Richmond, Julia C. Redman, Christopher Johnson, Roland Green, Jonathan Arias, and Christopher 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. *The Plant Journal*, 47(1):152–162, July 2006. ISSN 1365-313X. doi: 10.1111/j.1365-313X.2006.02770.x. URL http://onlinelibrary.wiley.com/doi/10.1111/j.1365-313X.2006.02770.x/abstract.

Gert Thijs, Kathleen Marchal, Magali Lescot, Stephane Rombauts, Bart De Moor, Pierre RouzÃl', and Yves Moreau. A Gibbs sampling method to detect overrepresented motifs in the upstream regions of coexpressed genes. *Journal of Computational Biology: A Journal of Computational Molecular Cell Biology*, 9(2):447–464, 2002. ISSN 1066-5277. doi: 10.1089/10665270252935566.

Oliver Thimm, Oliver Blasing, Yves Gibon, Axel Nagel, Svenja Meyer, Peter Kruger, Joachim Selbig, Lukas A. Muller, Seung Y. Rhee, and Mark Stitt. mapman: a user-driven tool to display genomics data sets onto diagrams of metabolic pathways and other biological processes. *The Plant Journal*, 37(6):914–939, March 2004. ISSN 1365-313X. doi: 10.1111/j.1365-313X.2004.02016.x. URL http://onlinelibrary.wiley.com/doi/10.1111/j.1365-313X.2004.02016.x/abstract.

Brian C. Thomas, Lakshmi Rapaka, Eric Lyons, Brent Pedersen, and Michael Freeling. Arabidopsis intragenomic conserved noncoding sequence. *Proceedings of the National Academy of Sciences*, 104(9):3348–3353, February 2007. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.0611574104. URL http://www.pnas.org/content/104/9/3348.

Morgane Thomas-Chollier, Carl Herrmann, Matthieu Defrance, Olivier Sand, Denis Thieffry, and Jacques van Helden. RSAT peak-motifs: motif analysis in full-size ChIP-seq datasets. *Nucleic Acids Research*, page gkr1104, December 2011. ISSN 0305-1048, 1362-4962. doi: 10.1093/nar/gkr1104. URL http://nar.oxfordjournals.org/content/early/2011/12/08/nar.gkr1104.

Dawn Thompson, Aviv Regev, and Sushmita Roy. Comparative analysis of gene regulatory networks: from network reconstruction to evolution. *Annual Review of Cell and Developmental Biology*, 31:399–428, 2015. ISSN 1530-8995. doi: 10.1146/annurev-cellbio-100913-012908.

Julie D. Thompson, Desmond G. Higgins, and Toby J. Gibson. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Research*, 22(22):4673–4680, November 1994. ISSN 0305-1048, 1362-4962. doi: 10.1093/nar/22.22.4673. URL http://nar.oxfordjournals.org/content/22/22/4673.

Caihuan Tian, Xiaoni Zhang, Jun He, Haopeng Yu, Ying Wang, Bihai Shi, Yingying Han, Guoxun Wang, Xiaoming Feng, Cui Zhang, Jin Wang, Jiyan Qi, Rong Yu, and Yuling Jiao. An organ boundary-enriched gene regulatory network uncovers regulatory hierarchies underlying axillary meristem initiation. *Molecular Systems Biology*, 10(10):1–2, October 2014. ISSN 1744-4292. doi: 10.15252/msb.20145470. URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4299377/.

Martin Tompa, Nan Li, Timothy L. Bailey, George M. Church, Bart De Moor, Eleazar Eskin, Alexander V. Favorov, Martin C. Frith, Yutao Fu, W. James Kent, Vsevolod J. Makeev, Andrei A. Mironov, William Stafford Noble, Giulio Pavesi, Graziano Pesole, Mireille RÃl'gnier, Nicolas Simonis, Saurabh Sinha, Gert Thijs, Jacques van Helden, Mathias Vandenbogaert, Zhiping Weng, Christopher Workman, Chun Ye, and Zhou Zhu. Assessing computational tools for the discovery of transcription factor binding sites. *Nature Biotechnology*, 23(1):137–144, January 2005. ISSN 1087-0156. doi: 10.1038/nbt1053. URL http://www.nature.com/nbt/journal/v23/n1/full/nbt1053.html.

Lam-Son Phan Tran, Kazuo Nakashima, Yoh Sakuma, Sean D. Simpson, Yasunari Fujita, Kyonoshin Maruyama, Miki Fujita, Motoaki Seki, Kazuo Shinozaki, and Kazuko Yamaguchi-Shinozaki. Isolation and Functional Analysis of Arabidopsis Stress-Inducible NAC Transcription Factors That Bind to a Drought-Responsive cis-Element in the early responsive to dehydration stress 1 Promoter. *The Plant Cell*, 16(9):2481–2498, September 2004. ISSN , 1532-298X. doi: 10.1105/tpc.104.022699. URL http://www.plantcell.org/content/16/9/2481.

William Truman and Jane Glazebrook. Co-expression analysis identifies putative targets for CBP60g and SARD1 regulation. *BMC Plant Biology*, 12:216, 2012. ISSN 1471-2229. doi: 10.1186/1471-2229-12-216. URL http://dx.doi.org/10.1186/1471-2229-12-216.

Jean-Valery Turatsinze, Morgane Thomas-Chollier, Matthieu Defrance, and Jacques van Helden. Using RSAT to scan genome sequences for transcription factor binding sites and cis-regulatory modules. *Nature Protocols*, 3(10):1578–1588, September 2008. ISSN 1754-2189. doi: 10.1038/nprot.2008.97. URL http://www.nature.com/nprot/journal/v3/n10/full/nprot.2008.97.html.

Franziska Turck, Aifen Zhou, and Imre E. Somssich. Stimulus-Dependent, Promoter-Specific Binding of Transcription Factor WRKY1 to Its Native Promoter and the Defense-Related Gene PcPR1-1 in Parsley. *The Plant Cell*, 16(10):2573–2585, October 2004. ISSN , 1532-298X. doi: 10.1105/tpc.104.024810. URL http://www.plantcell.org/content/16/10/2573.

Gina Turco, James C. Schnable, Brent Pedersen, and Michael Freeling. Automated conserved non-coding sequence (CNS) discovery reveals differences in gene content and promoter evolution among grasses. *Frontiers in Plant Science*, 4:170, 2013. ISSN 1664-462X. doi: 10.3389/fpls.2013.00170.

G. A. Tuskan, S. Difazio, S. Jansson, J. Bohlmann, I. Grigoriev, U. Hellsten, N. Putnam, S. Ralph, S. Rombauts, A. Salamov, J. Schein, L. Sterck, A. Aerts, R. R. Bhalerao, R. P. Bhalerao, D. Blaudez, W. Boerjan, A. Brun, A. Brunner, V. Busov, M. Campbell, J. Carlson, M. Chalot, J. Chapman, G.-L. Chen, D. Cooper, P. M. Coutinho, J. Couturier, S. Covert, Q. Cronk, R. Cunningham, J. Davis, S. Degroeve, A. Dujardin, C. Depamphilis, J. Detter, B. Dirks, I. Dubchak, S. Duplessis, J. Ehlting, B. Ellis, K. Gendler, D. Goodstein, M. Gribskov, J. Grimwood, A. Groover, L. Gunter, B. Hamberger, B. Heinze, Y. Helariutta, B. Henrissat, D. Holligan, R. Holt, W. Huang, N. Islam-Faridi, S. Jones, M. Jones-Rhoades, R. Jorgensen, C. Joshi, J. Kangasjarvi, J. Karlsson, C. Kelleher, R. Kirkpatrick, M. Kirst, A. Kohler, U. Kalluri, F. Larimer, J. Leebens-Mack, J.-C. Leplu, P. Locascio, Y. Lou, S. Lucas, F. Martin, B. Montanini, C. Napoli, D. R. Nelson, C. Nelson, K. Nieminen, O. Nilsson, V. Pereda, G. Peter, R. Philippe, G. Pilate, A. Poliakov, J. Razumovskaya, P. Richardson, C. Rinaldi, K. Ritland, P. Rouze, D. Ryaboy, J. Schmutz, J. Schrader, B. Segerman, H. Shin, A. Siddiqui, F. Sterky, A. Terry, C.-J. Tsai, E. Uberbacher, P. Unneberg, J. Vahala, K. Wall, S. Wessler, G. Yang, T. Yin, C. Douglas, M. Marra, G. Sandberg, Y. Van de Peer, and D. Rokhsar. The genome of black cottonwood, Populus trichocarpa (Torr. & Gray). *Science (New York, N.Y.)*, 313(5793):1596–1604, September 2006. ISSN 1095-9203. doi: 10.1126/science.1128691.

Cristobal Uauy, Assaf Distelfeld, Tzion Fahima, Ann Blechl, and Jorge Dubcovsky. A NAC Gene Regulating Senescence Improves Grain Protein, Zinc, and Iron Content in Wheat. *Science*, 314 (5803):1298–1301, November 2006. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.1133649. URL `http://science.sciencemag.org/content/314/5803/1298`.

Michiel Van Bel, Sebastian Proost, Elisabeth Wischnitzki, Sara Movahedi, Christopher Scheerlinck, Yves Van de Peer, and Klaas Vandepoele. Dissecting Plant Genomes with the PLAZA Comparative Genomics Platform. *Plant Physiology*, 158(2):590–600, February 2012. ISSN , 1532-2548. doi: 10.1104/pp.111.189514. URL `http://www.plantphysiol.org/content/158/2/590`.

Jan Van de Velde, Ken S. Heyndrickx, and Klaas Vandepoele. Inference of Transcriptional Networks in Arabidopsis through Conserved Noncoding Sequence Analysis. *The Plant Cell Online*, page tpc.114.127001, July 2014. ISSN , 1532-298X. doi: 10.1105/tpc.114.127001. URL `http://www.plantcell.org/content/early/2014/07/02/tpc.114.127001`.

Jan Van de Velde, Michiel Van Bel, Dries Vaneechoutte, and Klaas Vandepoele. A Collection of Conserved Noncoding Sequences to Study Gene Regulation in Flowering Plants. *Plant Physiology*, 171(4):2586–2598, August 2016. ISSN , 1532-2548. doi: 10.1104/pp.16.00821. URL `http://www.plantphysiol.org/content/171/4/2586`.

Robert VanBuren, Doug Bryant, Patrick P. Edger, Haibao Tang, Diane Burgess, Dinakar Challabathula, Kristi Spittle, Richard Hall, Jenny Gu, Eric Lyons, Michael Freeling, Dorothea Bartels, Boudewijn Ten Hallers, Alex Hastie, Todd P. Michael, and Todd C. Mockler. Single-molecule sequencing of the desiccation-tolerant grass Oropetium thomaeum. *Nature*, 527 (7579):508–511, November 2015. ISSN 0028-0836. doi: 10.1038/nature15714. URL `http://www.nature.com/nature/journal/v527/n7579/full/nature15714.html`.

Klaas Vandepoele, Tineke Casneuf, and Yves Van de Peer. Identification of novel regulatory modules in dicotyledonous plants using expression data and comparative genomics. *Genome Biology*, 7: R103, 2006. ISSN 1474-760X. doi: 10.1186/gb-2006-7-11-r103. URL `http://dx.doi.org/10.1186/gb-2006-7-11-r103`.

Klaas Vandepoele, Mauricio Quimbaya, Tine Casneuf, Lieven De Veylder, and Yves Van de Peer. Unraveling Transcriptional Control in Arabidopsis Using cis-Regulatory Elements and Coexpression Networks. *Plant Physiology*, 150(2):535–546, June 2009. ISSN , 1532-2548. doi: 10.1104/pp.109.136028. URL `http://www.plantphysiol.org/content/150/2/535`.

Sandeep Venkataram and Justin C. Fay. Is transcription factor binding site turnover a sufficient explanation for cis-regulatory sequence divergence? *Genome Biology and Evolution*, 2:851–858, 2010. ISSN 1759-6653. doi: 10.1093/gbe/evq066.

Aurine Verkest, Thomas Abeel, Ken S. Heyndrickx, Jelle Van Leene, Christa Lanz, Eveline Van De Slijke, Nancy De Winne, Dominique Eeckhout, Geert Persiau, Frank Van Breusegem, Dirk Inze, Klaas Vandepoele, and Geert De Jaeger. A generic tool for transcription factor target gene discovery in Arabidopsis cell suspension cultures based on tandem chromatin affinity purification. *Plant Physiology*, 164(3):1122–1133, March 2014. ISSN 1532-2548. doi: 10.1104/pp.113.229617.

Aurine Verkest, Marina Byzova, Cindy Martens, Patrick Willems, Tom Verwulgen, Bram Slabbinck, Debbie Rombaut, Jan Van de Velde, Klaas Vandepoele, Evi Standaert, Marrit Peeters, Mieke Van Lijsebettens, Frank Van Breusegem, and Marc De Block. Selection for Improved Energy Use Efficiency and Drought Tolerance in Canola Results in Distinct Transcriptome and Epigenome Changes1[OPEN]. *Plant Physiology*, 168(4): 1338–1350, August 2015. ISSN 0032-0889. doi: 10.1104/pp.15.00155. URL `http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4528734/`.

Vanessa Vermeirssen, Bart Deplancke, M. Inmaculada Barrasa, John S. Reece-Hoyes, H. Efsun Arda, Christian A. Grove, Natalia J. Martinez, Reynaldo Sequerra, Lynn Doucette-Stamm, Michael R. Brent, and Albertha J. M. Walhout. Matrix and Steiner-triple-system smart pooling assays for high-performance transcription regulatory network mapping. *Nature Methods*, 4 (8):659–664, August 2007. ISSN 1548-7091. doi: 10.1038/nmeth1063. URL `http://www.nature.com/nmeth/journal/v4/n8/full/nmeth1063.html`.

R. VijayaSatya and A. Mukheqee. PRUNER: algorithms for finding monad patterns in DNA sequences. In *2004 IEEE Computational Systems Bioinformatics Conference, 2004. CSB 2004. Proceedings*, pages 662–665, August 2004. doi: 10.1109/CSB.2004.1332537.

Diego Villar, Paul Flicek, and Duncan T. Odom. Evolution of transcription factor binding in metazoans - mechanisms and functional implications. *Nature Reviews. Genetics*, 15(4):221–233, April 2014. ISSN 1471-0064. doi: 10.1038/nrg3481.

Justin W. Walley, Ryan C. Sartor, Zhouxin Shen, Robert J. Schmitz, Kevin J. Wu, Mark A. Urich, Joseph R. Nery, Laurie G. Smith, James C. Schnable, Joseph R. Ecker, and Steven P. Briggs. Integration of omic networks in a developmental atlas of maize. *Science (New York, N.Y.)*, 353(6301):814–818, August 2016. ISSN 1095-9203. doi: 10.1126/science.aag1125.

Chun-Tao Wang and Yi-Nong Xu. The 5Õ untranslated region of the FAD3 mRNA is required for its translational enhancement at low temperature in Arabidopsis roots. *Plant Science*, 179(3): 234–240, September 2010. ISSN 0168-9452. doi: 10.1016/j.plantsci.2010.05.008. URL `http://www.sciencedirect.com/science/article/pii/S0168945210001470`.

Fangfang Wang and Sharyn E. Perry. Identification of Direct Targets of FUSCA3, a Key Regulator of Arabidopsis Seed Development. *Plant Physiology*, 161(3):1251–1264, March 2013. ISSN , 1532-2548. doi: 10.1104/pp.112.212282. URL `http://www.plantphysiol.org/content/161/3/1251`.

Feng Wang, Antonella Muto, Jan Van de Velde, Pia Neyt, Kristiina Himanen, Klaas Vandepoele, and Mieke Van Lijsebettens. Functional Analysis of the Arabidopsis TETRASPANIN Gene Family in Plant Growth and Development1[OPEN]. *Plant Physiology*, 169(3):2200–2214, November 2015a. ISSN 0032-0889. doi: 10.1104/pp.15.01310. URL `http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4634101/`.

Shaokui Wang, Shan Li, Qian Liu, Kun Wu, Jianqing Zhang, Shuansuo Wang, Yi Wang, Xiangbin Chen, Yi Zhang, Caixia Gao, Feng Wang, Haixiang Huang, and Xiangdong Fu. The OsSPL16-GW7 regulatory module determines grain shape and simultaneously improves rice yield and grain quality. *Nature Genetics*, 47 (8):949–954, August 2015b. ISSN 1061-4036. doi: 10.1038/ng.3352. URL http://www.nature.com/ng/journal/v47/n8/full/ng.3352.html.

Ting Wang and Gary D. Stormo. Identifying the conserved network of cis-regulatory sites of a eukaryotic genome. *Proceedings of the National Academy of Sciences of the United States of America*, 102(48):17400–17405, November 2005. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.0505147102. URL http://www.pnas.org/content/102/48/17400.

Xiaowu Wang, Hanzhong Wang, Jun Wang, Rifei Sun, Jian Wu, Shengyi Liu, Yinqi Bai, Jeong-Hwan Mun, Ian Bancroft, Feng Cheng, Sanwen Huang, Xixiang Li, Wei Hua, Junyi Wang, Xiyin Wang, Michael Freeling, J. Chris Pires, Andrew H. Paterson, Boulos Chalhoub, Bo Wang, Alice Hayward, Andrew G. Sharpe, Beom-Seok Park, Bernd Weisshaar, Binghang Liu, Bo Li, Bo Liu, Chaobo Tong, Chi Song, Christopher Duran, Chunfang Peng, Chunyu Geng, Chushin Koh, Chuyu Lin, David Edwards, Desheng Mu, Di Shen, Eleni Soumpourou, Fei Li, Fiona Fraser, Gavin Conant, Gilles Lassalle, Graham J. King, Guusje Bonnema, Haibao Tang, Haiping Wang, Harry Belcram, Heling Zhou, Hideki Hirakawa, Hiroshi Abe, Hui Guo, Hui Wang, Huizhe Jin, Isobel A. P. Parkin, Jacqueline Batley, Jeong-Sun Kim, Jeremy Just, Jianwen Li, Jiaohui Xu, Jie Deng, Jin A. Kim, Jingping Li, Jingyin Yu, Jinling Meng, Jinpeng Wang, Jiumeng Min, Julie Poulain, Jun Wang, Katsunori Hatakeyama, Kui Wu, Li Wang, Lu Fang, Martin Trick, Matthew G. Links, Meixia Zhao, Mina Jin, Nirala Ramchiary, Nizar Drou, Paul J. Berkman, Qingle Cai, Quanfei Huang, Ruiqiang Li, Satoshi Tabata, Shifeng Cheng, Shu Zhang, Shujiang Zhang, Shunmou Huang, Shusei Sato, Silong Sun, Soo-Jin Kwon, Su-Ryun Choi, Tae-Ho Lee, Wei Fan, Xiang Zhao, Xu Tan, Xun Xu, Yan Wang, Yang Qiu, Ye Yin, Yingrui Li, Yongchen Du, Yongcui Liao, Yongpyo Lim, Yoshihiro Narusaka, Yupeng Wang, Zhenyi Wang, Zhenyu Li, Zhiwen Wang, Zhiyong Xiong, Zhonghua Zhang, and Brassica rapa Genome Sequencing Project Consortium. The genome of the mesopolyploid crop species Brassica rapa. *Nature Genetics*, 43(10):1035–1039, October 2011. ISSN 1546-1718. doi: 10.1038/ng.919.

Wyeth W. Wasserman and Albin Sandelin. Applied bioinformatics for the identification of regulatory elements. *Nature Reviews Genetics*, 5(4):276–287, April 2004. ISSN 1471-0056. doi: 10.1038/nrg1315. URL http://www.nature.com/nrg/journal/v5/n4/full/nrg1315.html.

Vikki M. Weake and Jerry L. Workman. Inducible gene expression: diverse regulatory mechanisms. *Nature Reviews Genetics*, 11(6): 426–437, June 2010. ISSN 1471-0056. doi: 10.1038/nrg2781. URL http://www.nature.com/nrg/journal/v11/n6/full/nrg2781.html.

Wei Wei and Xiao-Dan Yu. Comparative analysis of regulatory motif discovery tools for transcription factor binding sites. *Genomics, Proteomics & Bioinformatics*, 5(2):131–142, May 2007. ISSN 1672-0229. doi: 10.1016/S1672-0229(07)60023-0.

Matthew T. Weirauch, Ally Yang, Mihai Albu, Atina G. Cote, Alejandro Montenegro-Montero, Philipp Drewe, Hamed S. Najafabadi, Samuel A. Lambert, Ishminder Mann, Kate Cook, Hong Zheng, Alejandra Goity, Harm van Bakel, Jean-Claude Lozano, Mary Galli, Mathew G. Lewsey, Eryong Huang, Tuhin Mukherjee, Xiaoting Chen, John S. Reece-Hoyes, Sridhar Govindarajan, Gad Shaulsky, Albertha J. M. Walhout, Francois-Yves Bouget, Gunnar Ratsch, Luis F. Larrondo, Joseph R. Ecker, and Timothy R. Hughes. Determination and inference of eukaryotic transcription factor sequence specificity. *Cell*, 158(6):1431–1443, September 2014. ISSN 1097-4172. doi: 10.1016/j.cell.2014.08.009.

Ditte H. Welner, Soren Lindemose, J. Gunter Grossmann, Niels Erik Myllegaard, Addie N. Olsen, Charlotte Helgstrand, Karen Skriver, and Leila Lo Leggio. DNA binding by the plant-specific NAC transcription factors in crystal and solution: a firm link to WRKY and GCM transcription factors. *Biochemical Journal*, 444(3):395–404, June 2012. ISSN 0264-6021, 1470-8728. doi: 10.1042/BJ20111742. URL http://www.biochemj.org/content/444/3/395.

Cara Winter, Ryan Austin, Servane Blanvillain-Baufume, Maxwell A. Reback, Marie Monniaux, Miin-Feng Wu, Yi Sang, Ayako Yamaguchi, Nobutoshi Yamaguchi, Jane E. Parker, Francois Parcy, Shane T. Jensen, Hongzhe Li, and Doris Wagner. LEAFY Target Genes Reveal Floral Regulatory Logic, cis Motifs, and a Link to Biotic Stimulus Response. *Developmental Cell*, 20(4):430–443, April 2011. ISSN 1534-5807. doi: 10.1016/j.devcel.2011.03.019. URL http://www.sciencedirect.com/science/article/pii/S1534580711001250.

Daniel Wong, Ana Teixeira, Spyros Oikonomopoulos, Peter Humburg, Imtiaz Nisar Lone, David Saliba, Trevor Siggers, Martha Bulyk, Dimitar Angelov, Stefan Dimitrov, Irina A. Udalova, and Jiannis Ragoussis. Extensive characterization of NF-ÎžB binding uncovers non-canonical motifs and advances the interpretation of genetic functional traits. *Genome Biology*, 12:R70, 2011. ISSN 1474-760X. doi: 10.1186/gb-2011-12-7-r70. URL http://dx.doi.org/10.1186/gb-2011-12-7-r70.

Christopher T. Workman, Yutong Yin, David L. Corcoran, Trey Ideker, Gary D. Stormo, and Panayiotis V. Benos. enoLOGOS: a versatile web tool for energy normalized sequence logos. *Nucleic Acids Research*, 33(suppl 2):W389–W392, July 2005. ISSN 0305-1048, 1362-4962. doi: 10.1093/nar/gki439. URL http://nar.oxfordjournals.org/content/33/suppl_2/W389.

Jie Wu, Douglas H. Sieglaff, Joshua Gervin, and Xiaohui S. Xie. Discovering regulatory motifs in the Plasmodium genome using comparative genomics. *Bioinformatics (Oxford, England)*, 24(17): 1843–1849, September 2008. ISSN 1367-4811. doi: 10.1093/bioinformatics/btn348.

Samuel E. Wuest, Diarmuid S. O Maoileidigh, Liina Rae, Kamila Kwasniewska, Andrea Raganelli, Katarzyna Hanczaryk, Amanda J. Lohan, Brendan Loftus, Emmanuelle Graciet, and Frank Wellmer. Molecular basis for the specification of floral organs by APETALA3 and PISTILLATA. *Proceedings of the National Academy of Sciences*, 109(33):13452–13457, August 2012. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.1207075109. URL http://www.pnas.org/content/109/33/13452.

Jun Xiao, Yongchun Zhou, Hao Lai, Shi Lei, Lisa H. Chi, and Xianwei Mo. Transcription Factor NF-Y Is a Functional Regulator of the Transcription of Core Clock Gene Bmal1. *Journal of Biological Chemistry*, 288(44):31930–31936, November 2013. ISSN 0021-9258, 1083-351X. doi: 10.1074/jbc.M113.507038. URL http://www.jbc.org/content/288/44/31930.

Xiaohui Xie, Jun Lu, E. J. Kulbokas, Todd R. Golub, Vamsi Mootha, Kerstin Lindblad-Toh, Eric S. Lander, and Manolis Kellis. Systematic discovery of regulatory motifs in human promoters and 3' UTRs by comparison of several mammals. *Nature*, 434(7031):338–345, March 2005. ISSN 1476-4687. doi: 10.1038/nature03441.

Zheng-Yi Xu, Soo Youn Kim, Do Young Hyeon, Dae Heon Kim, Ting Dong, Youngmin Park, Jing Bo Jin, Se-Hwan Joo, Seong-Ki Kim, Jong Chan Hong, Daehee Hwang, and Inhwan Hwang. The Arabidopsis NAC Transcription Factor ANAC096 Cooperates with bZIP-Type Transcription Factors in Dehydration and Osmotic Stress Responses. *The Plant Cell Online*, page tpc.113.119099, November 2013. ISSN , 1532-298X. doi: 10.1105/tpc.113.119099. URL http://www.plantcell.org/content/early/2013/11/27/tpc.113.119099.

Masatoshi Yamaguchi, Nobutaka Mitsuda, Misato Ohtani, Masaru Ohme-Takagi, Ko Kato, and Taku Demura. VASCULAR-RELATED NAC-DOMAIN 7 directly regulates the expression of a broad range of genes for xylem vessel formation. *The Plant Journal*, 66(4):579–590, May 2011. ISSN 1365-313X. doi: 10.1111/j.1365-313X.2011.04514.x. URL http://onlinelibrary.wiley.com/doi/10.1111/j.1365-313X.2011.04514.x/abstract.

Levi Yant, Johannes Mathieu, Thanh Theresa Dinh, Felix Ott, Christa Lanz, Heike Wollmann, Xuemei Chen, and Markus Schmid. Orchestration of the Floral Transition and Floral Development in Arabidopsis by the Bifunctional Transcription Factor APETALA2. *The Plant Cell*, 22(7):2156–2170, July 2010. ISSN , 1532-298X. doi: 10.1105/tpc.110.075606. URL http://www.plantcell.org/content/22/7/2156.

Alper Yilmaz, Maria Katherine Mejia-Guerra, Kyle Kurz, Xiaoyu Liang, Lonnie Welch, and Erich Grotewold. AGRIS: the Arabidopsis Gene Regulatory Information Server, an update. *Nucleic Acids Research*, 39(suppl 1):D1118–D1122, January 2011. ISSN 0305-1048, 1362-4962. doi: 10.1093/nar/gkq1120. URL http://nar.oxfordjournals.org/content/39/suppl_1/D1118.

Kaoru Yoshiyama, Phillip A. Conklin, Neil D. Huefner, and Anne B. Britt. Suppressor of gamma response 1 (SOG1) encodes a putative transcription factor governing multiple responses to DNA damage. *Proceedings of the National Academy of Sciences*, 106 (31):12843–12848, August 2009. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.0810304106. URL http://www.pnas.org/content/106/31/12843.

Chun-Ping Yu, Sean Chun-Chang Chen, Yao-Ming Chang, Wen-Yu Liu, Hsin-Hung Lin, Jinn-Jy Lin, Hsiang June Chen, Yu-Ju Lu, Yi-Hsuan Wu, Mei-Yeh Jade Lu, Chen-Hua Lu, Arthur Chun-Chieh Shih, Maurice Sun-Ben Ku, Shin-Han Shiu, Shu-Hsing Wu, and Wen-Hsiung Li. Transcriptome dynamics of developing maize leaves and genomewide prediction of cis elements and their cognate transcription factors. *Proceedings of the National Academy of Sciences of the United States of America*, 112(19):E2477–2486, May 2015. ISSN 1091-6490. doi: 10.1073/pnas.1500605112.

Xiaofei Yu, Lei Li, Jaroslaw Zola, Maneesha Aluru, Huaxun Ye, Andrew Foudree, Hongqing Guo, Sarah Anderson, Srinivas Aluru, Peng Liu, Steve Rodermel, and Yanhai Yin. A brassinosteroid transcriptional network revealed by genome-wide identification of BESI target genes in Arabidopsis thaliana. *The Plant Journal*, 65(4):634–646, February 2011. ISSN 1365-313X. doi: 10.1111/j.1365-313X.2010.04449.x. URL http://onlinelibrary.wiley.com/doi/10.1111/j.1365-313X.2010.04449.x/abstract.

Wenli Zhang, Yufeng Wu, James C. Schnable, Zixian Zeng, Michael Freeling, Gregory E. Crawford, and Jiming Jiang. High-resolution mapping of open chromatin in the rice genome. *Genome Research*, 22(1):151–162, January 2012a. ISSN 1088-9051, 1549-5469. doi: 10.1101/gr.131342.111. URL http://genome.cshlp.org/content/22/1/151.

Wenli Zhang, Tao Zhang, Yufeng Wu, and Jiming Jiang. Genome-wide identification of regulatory DNA elements and protein-binding footprints using signatures of open chromatin in Arabidopsis. *The Plant Cell*, 24(7):2719–2731, July 2012b. ISSN 1532-298X. doi: 10.1105/tpc.112.098061.

Yong Zhang, Tao Liu, Clifford A. Meyer, Jerome Eeckhoute, David S. Johnson, Bradley E. Bernstein, Chad Nusbaum, Richard M. Myers, Myles Brown, Wei Li, and X. Shirley Liu. Model-based Analysis of ChIP-Seq (MACS). *Genome Biology*, 9:R137, 2008. ISSN 1474-760X. doi: 10.1186/gb-2008-9-9-r137. URL http://dx.doi.org/10.1186/gb-2008-9-9-r137.

Yu Zhang, Oleg Mayba, Anne Pfeiffer, Hui Shi, James M. Tepperman, Terence P. Speed, and Peter 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 Genet*, 9(1):e1003244, January 2013. ISSN 1553-7404. doi: 10.1371/journal.pgen.1003244. URL http://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.1003244.

Zhenhai Zhang, Jingyin Yu, Daofeng Li, Zuyong Zhang, Fengxia Liu, Xin Zhou, Tao Wang, Yi Ling, and Zhen Su. PMRD: plant microRNA database. *Nucleic Acids Research*, 38(suppl 1):D806–D813, January 2010. ISSN 0305-1048, 1362-4962. doi: 10.1093/nar/gkp818. URL http://nar.oxfordjournals.org/content/38/suppl_1/D806.

Yue Zhao, David Granas, and Gary D. Stormo. Inferring Binding Energies from Selected Binding Sites. *PLOS Comput Biol*, 5(12):e1000590, December 2009. ISSN 1553-7358. doi: 10.1371/journal.pcbi.1000590. URL http://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1000590.

Xiao-yu Zheng, Natalie Weaver Spivey, Weiqing Zeng, Po-Pu Liu, Zheng Qing Fu, Daniel F. Klessig, Sheng Yang He, and Xinnian Dong. Coronatine Promotes Pseudomonas syringae Virulence in Plants by Activating a Signaling Cascade that Inhibits Salicylic Acid Accumulation. *Cell Host & Microbe*, 11(6):587–596, June 2012. ISSN 1931-3128. doi: 10.1016/j.chom.2012.04.014. URL http://www.cell.com/cell-host-microbe/abstract/S1931-3128(12)00160-6.

Yumei Zheng, Na Ren, Huai Wang, Arnold J. Stromberg, and Sharyn E. Perry. Global Identification of Targets of the Arabidopsis MADS Domain Protein AGAMOUS-Like15. *The Plant Cell*, 21(9):2563–2577, September 2009. ISSN , 1532-298X. doi: 10.1105/tpc.109.068890. URL http://www.plantcell.org/content/21/9/2563.

Ruiqin Zhong and Zheng-Hua Ye. MYB46 and MYB83 bind to the SMRE sites and directly activate a suite of transcription factors and secondary wall biosynthetic genes. *Plant & Cell Physiology*, 53(2):368–380, February 2012. ISSN 1471-9053. doi: 10.1093/pcp/pcr185.

Ruiqin Zhong, Elizabeth A. Richardson, and Zheng-Hua Ye. Two NAC domain transcription factors, SND1 and NST1, function redundantly in regulation of secondary wall synthesis in fibers of Arabidopsis. *Planta*, 225(6):1603–1611, March 2007. ISSN 0032-0935, 1432-2048. doi: 10.1007/s00425-007-0498-y. URL http://link.springer.com/article/10.1007/s00425-007-0498-y.

Ruiqin Zhong, Chanhui Lee, Jianli Zhou, Ryan L. McCarthy, and Zheng-Hua Ye. A Battery of Transcription Factors Involved in the Regulation of Secondary Cell Wall Biosynthesis in Arabidopsis. *The Plant Cell*, 20(10):2763–2782, October 2008. ISSN , 1532-298X. doi: 10.1105/tpc.108.061325. URL http://www.plantcell.org/content/20/10/2763.

Ruiqin Zhong, Chanhui Lee, and Zheng-Hua Ye. Global Analysis of Direct Targets of Secondary Wall NAC Master Switches in Arabidopsis. *Molecular Plant*, 3(6):1087–1103, November 2010. ISSN 1674-2052. doi: 10.1093/mp/ssq062. URL http://www.cell.com/molecular-plant/abstract/S1674-2052(14)60556-6.

Jianli Zhou, Chanhui Lee, Ruiqin Zhong, and Zheng-Hua Ye. MYB58 and MYB63 are transcriptional activators of the lignin biosynthetic pathway during secondary cell wall formation in Arabidopsis. *The Plant Cell*, 21(1):248–266, January 2009. ISSN 1040-4651. doi: 10.1105/tpc.108.063321.