universität wien

MASTERARBEIT | MASTER'S THESIS

Titel | Title

GATA transcription factors function through GATC motifs in transcribed regions of plants

verfasst von | submitted by

Gabriela Hristova B Sc

angestrebter akademischer Grad | in partial fulfilment of the requirements for the degree of

Master of Science (MSc)

Wien | Vienna, 2024

Studienkennzahl lt. Studienblatt | Degree programme code as it appears on the student record sheet:	UA 066 830
Studienrichtung lt. Studienblatt | Degree programme as it appears on the student record sheet:	Masterstudium Molecular Microbiology, Microbial Ecology and Immunobiology
Betreut von | Supervisor:	Prof. Dr. Magnus Nordborg

Table of contents

Acknowledgments

This Master's thesis was a collaborative effort and I am grateful to the many people who contributed to it. I would like to thank Magnus Nordborg for the opportunity to work in his group, both as an intern and later as a Master's student. I am particularly grateful to Yoav Voichek for designing the project and for his invaluable help, advice, and discussions during this thesis. I want to thank everyone who helped me with the experiments - Almudena Molla Morales, Lidia Bobrovnikova, Viktoria Nizhynska, and Joanna Gunis, as well as Tom Ellis who helped me with the statistical analysis.

I am very grateful to the facilities that made this project possible - the BioOptics facility, especially Pawel Pasierbek, Gabriele Bradamante, and Gerald Schmauß, the NGS facility, and the Media Kitchen. I am also grateful to Professor Saskia Hogenhout (John Innes Centre, UK) for generously providing SAP05 lines and constructs, which were crucial for the first part of this thesis.

Last but not least, I want to thank my family, friends, and especially my boyfriend, Valentin Bauer, for supporting me throughout this whole time.

Abbreviations

AB80 - pea chlorophyll a/b binding protein

ANOVA - Analysis of variance

AuxRE - auxin-responsive element

AY-WB - Aster Yellows phytoplasma strain Witches' Broom

BF - bright-field

BME3 - BLUE MICROPYLAR END 3, GATA8

BR - brassinosteroid

BRE - TFIIB-Recognition Element

Cab-1 - wheat chlorophyll a-b binding protein

CaMV 35S, 35S - 35S RNA gene from Cauliflower mosaic virus

CPE - core promoter element

CRE - cis-regulatory element

DAP-Seq - DNA affinity purification sequencing

DBD - DNA-binding domain

FACS - Fluorescence Activated Cell Sorting

GA - gibberellic acid

GFP - green fluorescent protein

IME - intron-mediated enhancement

Inr - Initiator Element

kb - kilobases

MPRA - massively parallel reporter assay

MS - Murashige and Skoog medium

NGS - Next-generation sequencing

OE - overexpression

PBS - phosphate-buffered saline

PCA - principal component analysis

PEG - polyethylene glycol

PIC - pre-initiation complex

PnWB - peanut witches' broom phytoplasma

QC - quiescent center

rbcS-E9 - small subunit of ribulose-1,5-bisphosphate carboxylase

RNA-Seq - RNA sequencing

SAP - secreted AY-WB protein

SPL - SQUAMOSA promoter binding protein-like

TFs - transcription factors

TFBSs - TF binding sites

TRP1 - *TRYPTOPHAN BIOSYNTHESIS 1*

TSS - transcription start site

UBI4, UBI10 - *UBIQUITIN 4, UBIQUITIN 10*

UTR - untranslated region

WBDL - witches' broom disease of lime phytoplasma

YFP - yellow fluorescent protein

Y Patch - pyrimidine patch

ZnF - zinc finger

Abstract

Previous work in my host lab identified an 8-bp GATC-containing motif downstream of the transcription start site (TSS) that strongly enhances transcription. Published literature suggests that the GATA family of transcription factors (TFs) are the most likely upstream regulators of this motif. This Master's thesis aims to investigate whether GATA factors control the transcription of genes through the GATC motif, employing two approaches: first, through the degradation of GATA factors via SAP05, a parasitic effector protein, and second, by overexpression of GATA TFs in Arabidopsis. The results from the first approach supported our hypothesis, albeit with a modest effect. Conversely, the second approach provided direct evidence that GATC-containing genes are targets of GATA TFs from class A. In summary, these findings indicate that GATA TFs regulate gene expression through a GATC motif located downstream of the TSS.

Zusammenfassung

Frühere Arbeiten in meinem Gastlabor identifizierten ein 8-bp GATC-haltiges Motiv downstream der Transkriptionsstartstelle (TSS), das die Transkription stark fördert. Veröffentlichte Literatur legt nahe, dass die GATA-Familie von Transkriptionsfaktoren (TFs) die wahrscheinlichsten upstream-Regulatoren dieses Motivs sind. Diese Masterarbeit zielt darauf ab zu untersuchen, ob GATA-Faktoren die Transkription von Genen über das GATC-Motiv steuern, und zwar durch zwei Ansätze: den Abbau von GATA-Faktoren mittels eines parasitären Effektorproteins SAP05 und die Überexpressionsanalyse spezifischer GATA-Faktoren in Arabidopsis. Die Ergebnisse des ersten Ansatzes stützten unsere Hypothese, wenn auch mit einer bescheidenen Wirkung. Im Gegensatz dazu lieferte der zweite Ansatz starke Beweise dafür, dass GATC-haltige Gene direkte Ziele der Klasse-A-GATA-TFs sind. Zusammenfassend weisen diese Ergebnisse darauf hin, dass GATA-TFs die Genexpression über ein GATC-Motiv regulieren, das downstream der TSS liegt.

1. Introduction

1.1. Plants and animals have distinct transcriptional regulation

Transcriptional regulation is central to biological processes such as development and environmental responses. Plants and animals have diverged independently for over a billion years (Betts *et al.*, 2018), evolving multicellularity independently (Bonner, 1998; Knoll, 2011). Multicellular organisms require increased transcriptional complexity to support the diverse gene expression profiles of various cell types. Therefore, it is unlikely that plants and animals evolved identical solutions to address this challenge. Yet, basic properties of regulatory sequences are commonly extrapolated from animals to plants (Burgess, Xu and Freeling, 2015). Indeed, evidence suggests that transcriptional regulation in plants has unique properties, as will be elaborated in the following paragraphs.

Regulatory sequences control transcription, determining the timing, strength, direction, and position of transcription initiation (Yamamoto *et al.*, 2007). Regulatory sequences can be separated into core promoters and enhancing elements. The core promoter region is the binding site of the RNA polymerase II preinitiation complex (PIC) and controls basal transcription (Nikolov and Burley, 1997; Berk, 1999), whereas enhancing elements are involved in the temporal and spatial regulation of transcription (Fessele *et al.*, 2002).

The TATA box is the first core promoter element (CPE) discovered in eukaryotic genes (Breathnach and Chambon, 1981). It is recognized by the transcription factor TBP (TATA-binding protein), which initiates transcription (Burley, 1996). The TATA box is found in 24% of yeast, 10% of human (Yang *et al.*, 2007), 29% of Arabidopsis (Molina and Grotewold, 2005), and 19% of rice promoters (Civán and Svec, 2009), respectively. It is highly conserved among eukaryotes and is also found in *Archaea* despite some differences in the TBP-TATA box interaction compared to eukaryotes (Blombach and Grohmann, 2017). Subsequently, several other CPEs have been determined - TFIIB-Recognition Element (BRE), Initiator Element (Inr), CpG islands, Y Patch (pyrimidine patch), and Sp1 element (Smale and Kadonaga, 2003; Yamamoto *et al.*, 2007). An analysis of CPEs in mammalian (human and mouse) and plant (Arabidopsis and rice) genomes revealed that while both groups share the TATA box, and Inr, mammalian genomes lack the Y patch (Yamamoto *et al.*, 2007). Mammals have CpG islands, which are distributed similarly to the Y patch elements in plants (Yamamoto *et al.*, 2007). Furthermore, the Sp1 element was not observed in plant promoters, and genes encoding the Sp1 transcription factor, which binds to this element in mammals (Kriwacki *et al.*, 1992), have not been found in the Arabidopsis or rice genomes (Yamamoto *et al.*, 2007). In addition, a novel class of CPE was discovered in plants - TC elements, found at the TATA-box location in non-TA-rich promoters (Bernard, Brunaud and Lecharny, 2010). Taken together, these findings indicate that plants and animals have evolved distinct sets of core promoter elements.

Another difference between regulatory sequences in plants and animals is their length and spatial organization in the genome around the gene they control. While animal regulatory sequences can extend over tens of kilobases (kb) of DNA (Bonifer, 2000), those in plants are shorter, typically under 1 or 2 kb (Riechmann, 2002). Furthermore, transcription initiation at the transcription start site (TSS) and *cis*-regulatory elements (CRE) is unidirectional in plants, unlike in animals, where it occurs in both directions (Hetzel *et al.*, 2016; Silver *et al.*, 2024). This is supported by the observation that transcription-associated histone modifications accumulate primarily in the gene body in Arabidopsis, rice, and soybean (Silver *et al.*, 2024). In contrast, in *Drosophila melanogaster* and *Homo sapiens*, these modifications are found on both sides of the TSS and CREs (Silver *et al.*, 2024).

Transcription factors (TFs) are proteins that control the transcription rate of associated genes by binding to specific DNA sequences in gene regulatory regions, known as TF binding sites (TFBSs) (Latchman, 1993). They regulate both basal and tissue-specific transcription and are classified into families based on their DNA-binding domain (DBD) (Latchman, 1993). The initial analysis of the Arabidopsis genome identified 1,533 TFs, which comprise ~5.9% of all genes (Riechmann *et al.*, 2000). These TFs were classified into 28 families, 45% (16 families) of which are Arabidopsis-specific and have not been observed in animals (*D. melanogaster* and *C. elegans*) or yeast (*S. cerevisiae*) (Riechmann *et al.*, 2000). Advances in structural biology enabled a new classification of plant TFs based on the structure and binding mode of their DNA-binding domains (Blanc-Mathieu *et al.*, 2024). Direct comparison with a similar classification of mammal TFs revealed eight new classes and 37 new families unique to plants (Blanc-Mathieu *et al.*, 2024). These plant-specific TFs might have evolved *de novo* through rapid sequence evolution of existing DBDs or domestication of transposable elements (Yamasaki *et al.*, 2013; de Mendoza and Sebé-Pedrós, 2019). Moreover, TF families in plants exhibit higher expansion rates compared to those in animals (Shiu, Shih and Li, 2005). This expansion is not solely attributed to whole-genome duplication, as TF genes tend to duplicate more than other plant genes (Shiu, Shih and Li, 2005).

In conclusion, these examples underscore the distinct characteristics of plant transcriptional regulation, as reflected by differences in core promoter elements, transcription factor families, and transcription initiation mechanisms.

1.2. Position-dependent enhancers in flowering plants and the GATC motif

Another key difference in transcriptional regulation between plants and animals is the position dependency of enhancers. Unlike animals, where enhancers can function over long distances and in both directions from the genes they regulate (Zabidi and Stark, 2016), recent research has shown that plant enhancers are position-dependent. One example of this is intron-mediated enhancement (IME), where introns located within the first 1 kb downstream of the TSS significantly boost mRNA levels independently of splicing (Rose, 2019). Mascarenhas *et al.*, 1990 demonstrated that inserting introns 2 and 6 from the maize alcohol dehydrogenase-1 gene into a transgene construct led to a 12- and 20-fold increase in

expression, respectively. However, this enhancement was lost when the introns were moved either upstream of the promoter or downstream of the coding sequence. Similarly, the first intron of the *UBI10* gene was found to increase mRNA accumulation by more than 12-fold (Rose *et al.*, 2008), but when this intron was positioned upstream of the 5'-most TSS in a reporter gene construct, it had no effect on expression (Gallegos and Rose, 2017).

Further evidence of the position-dependent activity of plant enhancers comes from the work of (Jores *et al.*, 2020). They found that the CaMV 35S enhancer exhibits its strongest activity when located just upstream of a minimal promoter. Conversely, when this enhancer is positioned within a transcribed region, such as the 3'UTR or an intron, its activity is completely abolished. A similar pattern was observed for three other enhancers from the pea *AB80* and *rbcS-E9* genes and the wheat *Cab-1* gene. These enhancers showed optimal activity only when placed just upstream of a minimal promoter, becoming inactive when located in the 3'UTR.

Previous work from our lab has expanded this research using a massively parallel reporter assay (MPRA) approach (Jores *et al.*, 2020). They generated a library of 12,000 regulatory sequences derived from regions upstream or downstream of the TSS of highly expressed *A. thaliana* genes (Fig. 1A; Voichek *et al.*, 2023). These sequences were then inserted on either side of the TSS in reporter gene constructs containing one of two different promoters: the minimal CaMV 35S promoter or the Arabidopsis *TRP1* promoter. To ensure the findings were broadly applicable across flowering plants, the library was transiently transformed into four species - protoplasts of *Arabidopsis thaliana*, tomato, and maize, and leaf infiltration of *N. benthamiana*. The effect of these sequences on gene expression was quantified by next-generation sequencing (NGS).

Interestingly, the most significant factor determining fragment activity was its position relative to the TSS, irrespective of species, promoter, or transformation technique (Fig. 1B; Voichek *et al.*, 2023). Moreover, the downstream sequences influenced transcription rates rather than mRNA stability, as evidenced by direct measurements of mRNA synthesis (Fig. 1C; Voichek *et al.*, 2023). The origin of the fragments also played a role, with fragments showing higher activity when positioned in their original location relative to the TSS (Fig. 1D; Voichek *et al.*, 2023). These results reinforce the idea that regulatory sequences in plants are highly position-dependent, emphasizing the critical role of sequences located downstream of the TSS.

An analysis of the strongest downstream sequences revealed an 8 bp motif with G-A-T-C at its core (referred to as the GATC motif) that significantly enhances expression in a dose-dependent manner (Fig. 2A; Voichek *et al.*, 2023). In contrast, removing the motif from fragments by deleting, shuffling, or modifying the core GATC to GATA, resulted in reduced expression of these constructs (Fig. 2B; Voichek *et al.*, 2023). Furthermore, the addition of new GATC motifs to the fragments led to an incremental increase in expression, proving the

sufficiency of the motif to enhance expression (Fig. 2C; Voichek *et al.*, 2023). However, these results were initially observed in a synthetic system. To validate the impact of the GATC motif in vivo, transcriptomic data from Arabidopsis seedlings was analyzed, comparing the expression of endogenous genes containing one or more GATC motifs to those lacking the motif (Fig. 2D; Voichek *et al.*, 2023). The analysis showed a strong correlation between gene expression levels and the number of GATC motifs located within 500 bp downstream of the TSS, thereby supporting the motif's effect in a native context. In addition, the effect of the GATC motif on gene expression occurs through changes in transcription rate, not mRNA stability, as demonstrated by mRNA synthesis rate measurements in the MPRA setup (Fig. 2E; Voichek *et al.*, 2023).

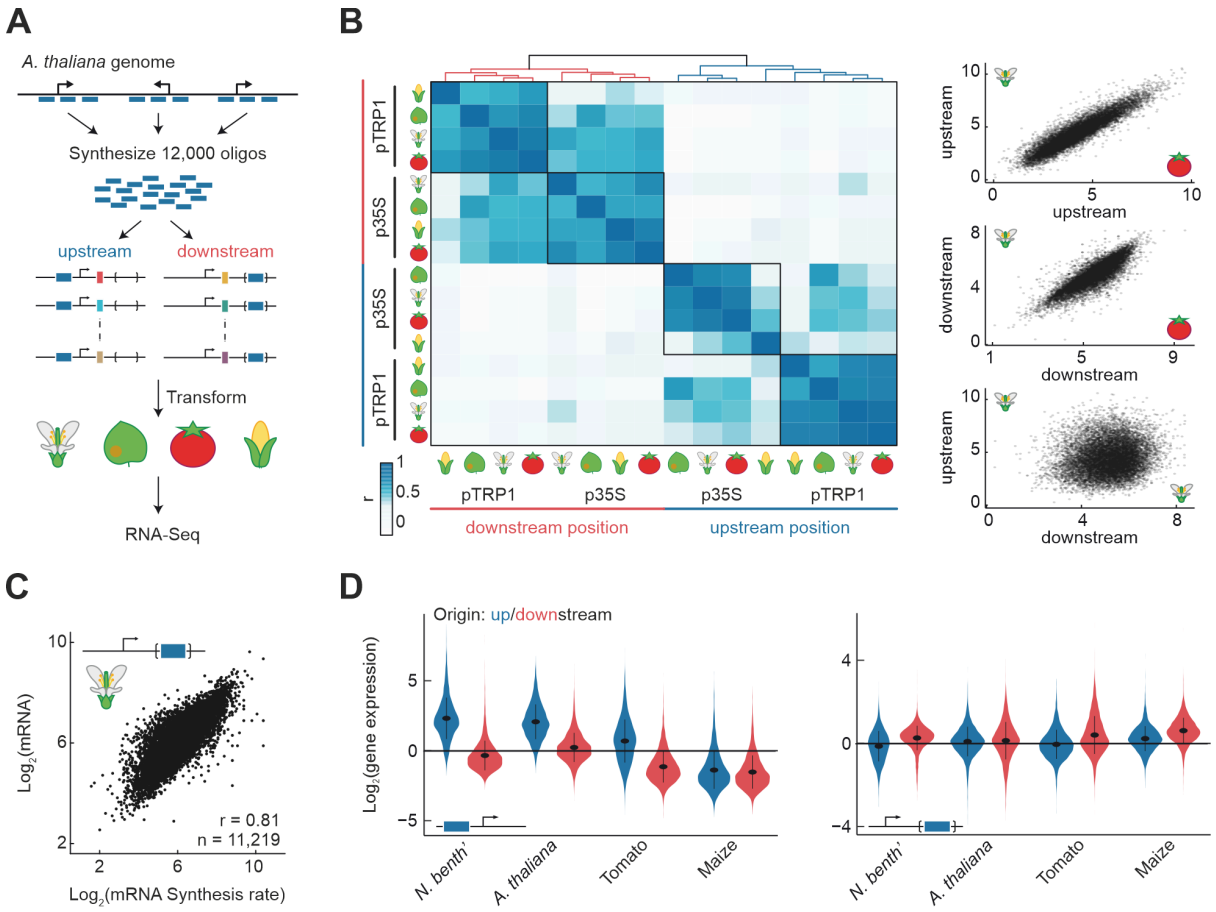


Figure 1: Enhancers in plants are position-dependent

All subfigures are reproduced from Voichek *et al.*, 2023**. (A)** Diagram of the MPRA approach: 12,000 fragments of 160 bp length derived from upstream or downstream of the TSS of Arabidopsis genes were synthesized, pooled, and inserted upstream of the TSS or within the intron of a reporter gene, and each construct was tagged by a unique barcode. The library was transiently transformed into four species, the barcoded RNA was sequenced and gene expression was quantified. **(B)** Left, Pearson's correlation coefficients across all libraries, hierarchically clustered. Insertion position, promoter, and species are indicated. Right, the activity of pTRP1-based constructs, calculated as $\log_2$(gene expression), compared between upstream (top, r=0.92) and downstream (middle, r=0.85)

libraries of Arabidopsis and tomato, and between upstream and downstream libraries of Arabidopsis (bottom, r=0.13). Correlations include 11,736±241 fragments. **(C)** Comparison of mRNA steady-state levels and synthesis rates in pTRP1 constructs with downstream positioned fragments. Pearson's correlation coefficients *r* and number of fragments *n* are indicated. **(D)** Comparison of the activity of upstream- (blue, 3,979±5 fragments) and downstream-derived (red, 7,941±3 fragments) fragments relative to no-insertion constructs, when fragments are positioned upstream (left panel) or downstream (right panel) of the TSS. Constructs are p35S-based. Error bars depict the mean ±1 standard deviation.

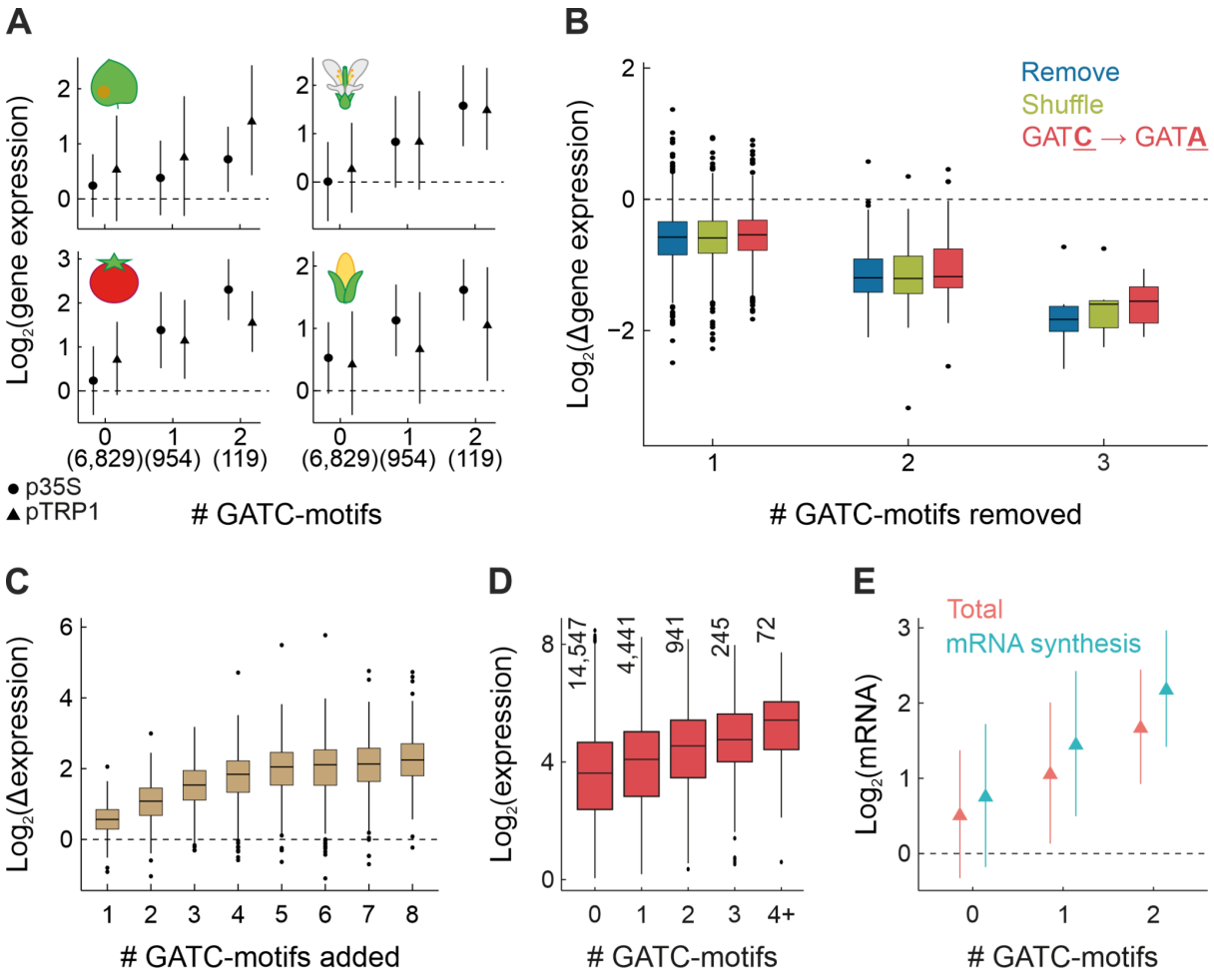


Figure 2: GATC motif downstream of the TSS enhances transcription in a dose-dependent manner
All subfigures are reproduced from Voichek *et al.*, 2023**. (A)** Relative activity of downstream-derived fragments when inserted downstream, as a function of the number of GATC motifs in the tested fragments. Group sizes: 6,855 (no motif), 956 (1 motif), and 119 (2 motifs). Promoter and species are indicated. **(B-C)** Differences in activity of the p35S-downstream library between original and mutated fragments. **(B)** Effect of removing GATC motifs on the activity of 823 fragments. Motifs were removed by deletion, 8 bp shuffle, or GATC-to-GATA mutation. **(C)** Effect of adding GATC motifs for 221 fragments with incremental motif additions. **(D)** Gene expression is positively correlated with the number of GATC motifs within 500 bp downstream of the TSS in the aerial parts of Arabidopsis seedlings (Schmid *et al.*, 2005). Expression levels are shown for different motif counts, with '4+' representing 4 to 9 motifs. A linear fit indicates an effect size of 0.4 per GATC motif (p-value:

$5*10^{-128}$), reflecting the average increase in expression for each additional motif. **(E)** Comparison of mRNA steady-state levels ("Total") and synthesis rates ("mRNA synthesis") in pTRP1 constructs with downstream positioned fragments based on the number of GATC motifs. Group sizes: 6,855 (no motif), 956 (1 motif), 119 (2 motifs). Data represent the average signal from two biological replicates of each library. Error bars represent the mean ± 1 standard deviation.

Taken together, these findings highlight the position-dependency of plant enhancers and uncover a new class of regulatory sequences located downstream of the transcription start site. Among these sequences, the most impactful motif is an 8 bp GATC-containing sequence, which is sufficient to enhance transcription in a dose-dependent manner.

1.3. The GATA family of transcription factors

The effect of the GATC motif on transcription rates suggests that it likely operates through transcription factors, with the GATA family of TFs being the most probable regulators. GATA TFs are found in animals, plants, and fungi. Although GATA factors were initially found to bind the consensus sequence W-G-A-T-A-R (W = A/T; R = A/G), subsequent studies revealed that they can also bind a motif with the core sequence G-A-T-C in other species (Lowry and Atchley, 2000). For example, multiple high-throughput studies have demonstrated that Arabidopsis GATA TFs preferentially bind to the GATC sequence rather than GATA, among other sequences (Franco-Zorrilla *et al.*, 2014; Weirauch *et al.*, 2014; O'Malley *et al.*, 2016). This preference has also been confirmed by targeted studies on specific GATA factors in Arabidopsis (Xu *et al.*, 2017), maize (Tu *et al.*, 2020), tobacco (Sugimoto, Takeda and Hirochika, 2003), and *Catharanthus roseus* (Liu *et al.*, 2019). Collectively, these findings suggest that plant GATA TFs predominantly bind to GATC motifs.

Further supporting these results, the DAP-Seq study by (O'Malley *et al.*, 2016) determined that many GATA factors preferentially bind downstream of the TSS, where the GATC motif is located (Fig. 3; Voichek *et al.*, 2023). This binding pattern aligns with the strong correlation observed between GATA factor gene expression and the genome-wide transcriptional impact of the GATC motif (Fig. 4; Voichek *et al.*, 2023). This correlation is most pronounced for members of the class A GATA TFs (Fig. 4), which also exhibit the highest affinity for the GATC motif (Fig. 3). These observations support the hypothesis that GATA TFs, particularly those from class A, play a role in driving the effects of the GATC motif.

Animal and yeast GATA families are relatively small, comprising 6 GATA factors in humans (Lowry and Atchley, 2000) and 10 GATA factors in *S. cerevisiae (Park et al., 2005)*. In contrast, *A. thaliana* contains 30 GATA TFs, which are categorized into 4 classes - A (14 factors), B (11 factors), C (3 factors), and D (2 factors) - based on the structure of their DNA-binding domain (Fig. 5; Schwechheimer, Schröder and Blaby-Haas, 2022). The GATA DNA-binding domain consists of a type IV zinc finger and a basic region and is the only conserved domain in this family (Lowry and Atchley, 2000). In *Arabidopsis*, GATA transcription factors regulate a wide range of biological processes, including light and hormonal signaling, photomorphogenesis,

development of roots, leaves, and flowers, dormancy control, chlorophyll biosynthesis, chloroplast development, and stomata formation (reviewed in Schwechheimer, Schröder and Blaby-Haas, 2022).

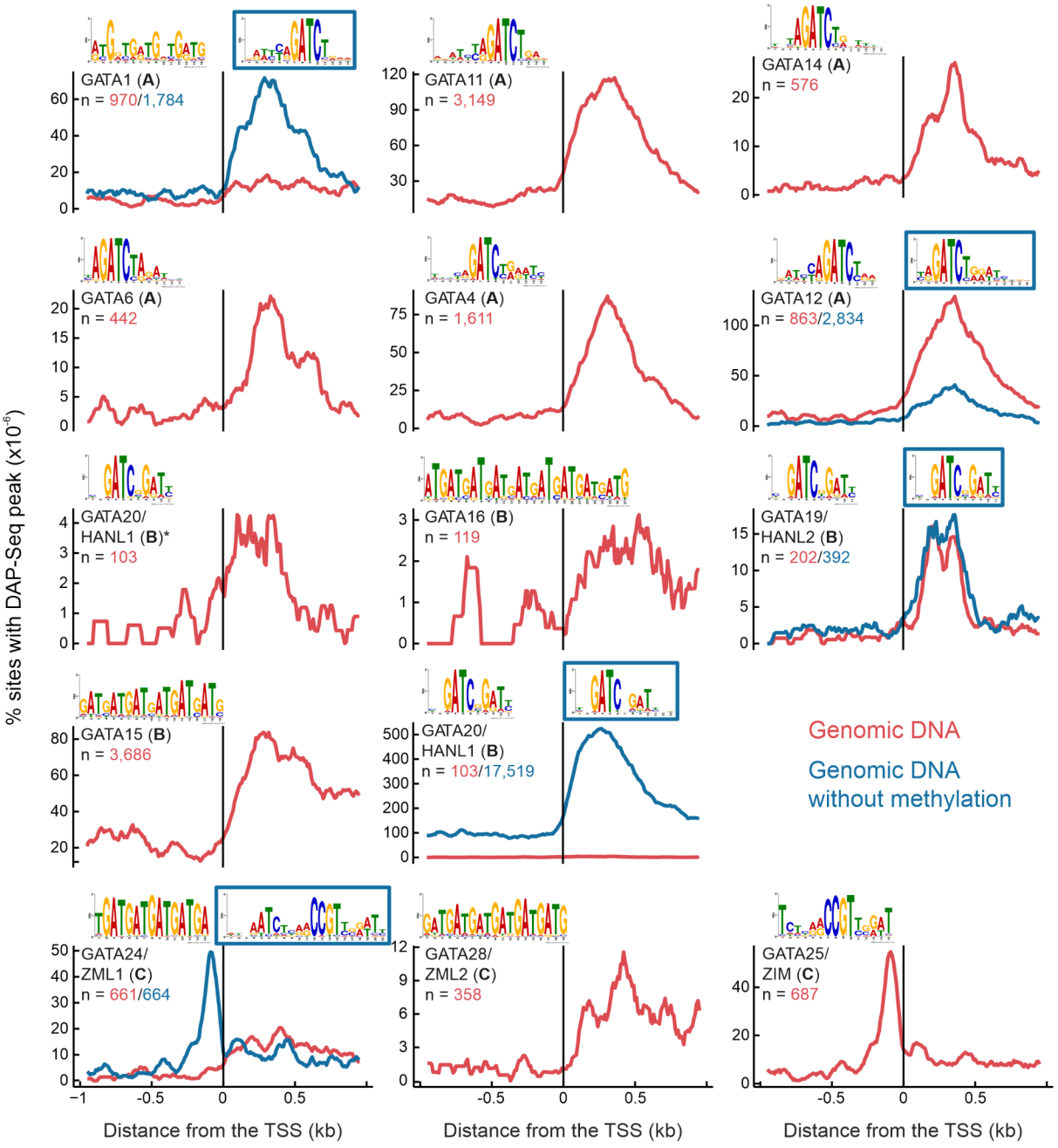


Figure 3: GATA TFs bind to GATC motifs downstream of the TSS

This figure is reproduced from Voichek *et al.*, 2023**.** The enrichment of binding peaks of the 13 GATA TFs analyzed in the DAP-Seq study (O'Malley *et al.*, 2016) is plotted as a function of the distance from the TSS for methylated DNA (red) and non-methylated DNA (blue). The class of the TF is displayed in brackets (Schwechheimer, Schröder and Blaby-Haas, 2022) and the number of binding peaks is shown below each TF's name, color-coded to correspond to the methylated or non-methylated DNA. The binding motif determined for each TF is displayed above its respective plot, with motifs derived

from non-methylated DNA highlighted with a blue frame. GATA20 is represented in two separate plots to account for the significant difference in the dynamic range observed between the methylated and non-methylated DNA.

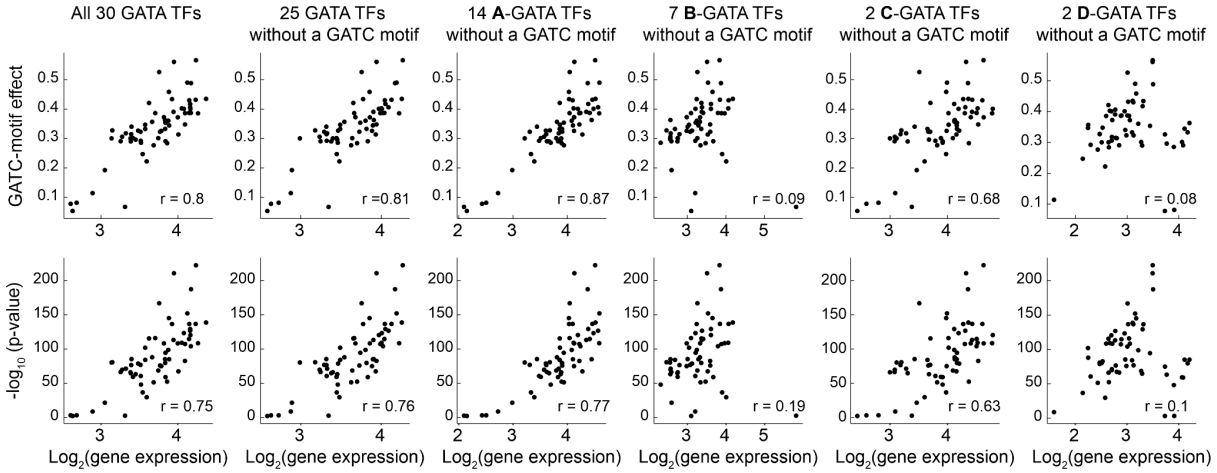


Figure 4: Expression of GATA TFs is correlated with the effect of the GATC motif
This figure is reproduced from Voichek *et al.*, 2023. The correlation between the mean expression levels of GATA TFs and the effect of the GATC motif (upper row), along with the associated p-values (lower row), calculated using transcriptomic data from various tissues (Klepikova *et al.*, 2016). Each column represents a different set of GATA TFs: all 30 GATA factors, 25 GATA factors without a GATC motif, and GATA factors from each class that lack the motif. Excluding GATA TFs containing the GATC motif in their genomic sequence ensures that the motif does not impact the expression of the GATA TFs themselves. The number of TFs included in each analysis is indicated in the column titles, and Pearson's correlation coefficients (r) are displayed on each plot.

Unlike the well-characterized class B GATA TFs, class A GATA TFs remain less thoroughly studied (Schwechheimer, Schröder and Blaby-Haas, 2022). Among them, *AtGATA2* is the most extensively researched and is known as a key transcriptional regulator of photomorphogenesis, where it integrates light and brassinosteroid signaling pathways (Luo *et al.*, 2010). *AtGATA4*, the closest homolog of *AtGATA2* within class A, plays a similar albeit less significant role in photomorphogenesis (Luo *et al.*, 2010). Additionally, *AtGATA2* is expressed in the root, particularly the transition zone between the meristematic and elongation zone (Jiang *et al.*, 2018; refer to Fig. 9A for a diagram of the root). Overexpression of *AtGATA2* severely disrupts root meristem development, while double knockdown mutants *gata2 gata12* display significantly shorter roots with reduced meristematic and transition zones (Jiang *et al.*, 2018). Furthermore, *AtGATA5* and *AtGATA12* are expressed in the vascular cylinder in the root tip, where they control the expression of *VASCULAR-RELATED NAC-DOMAIN7* (VND7), a key player in xylem vessel element differentiation (Endo *et al.*, 2015). In addition to its function in the root, *AtGATA12* also regulates seed dormancy via gibberellic acid (GA) signaling in freshly harvested seeds (Ravindran *et al.*, 2017). Its expression in dry seeds is promoted by RGL2, a repressor of GA signaling, and is suppressed

by high GA levels during imbibition and cold stratification (Ravindran *et al.*, 2017). Conversely, *BLUE MICROPYLAR END 3* (*BME3*), or *AtGATA8*, promotes seed germination (Liu *et al.*, 2005). *BME3* expression increases during cold stratification, whereas *bme3* mutants show reduced germination rates, along with diminished expression of GA biosynthesis genes (Liu *et al.*, 2005). *BME3* may also mediate the cross-talk between auxin, brassinosteroids (BR), and blue light, as these factors influence its expression (Choi *et al.*, 2014). Mutant *bme3* plants exhibit altered levels of auxin-, BR- and light-responsive genes and shorter hypocotyls under dark conditions (Choi *et al.*, 2014).

Together, these studies demonstrate that GATA factors, specifically those from class A, might operate via the GATC motif, regulating a wide range of physiological and developmental processes in plants.

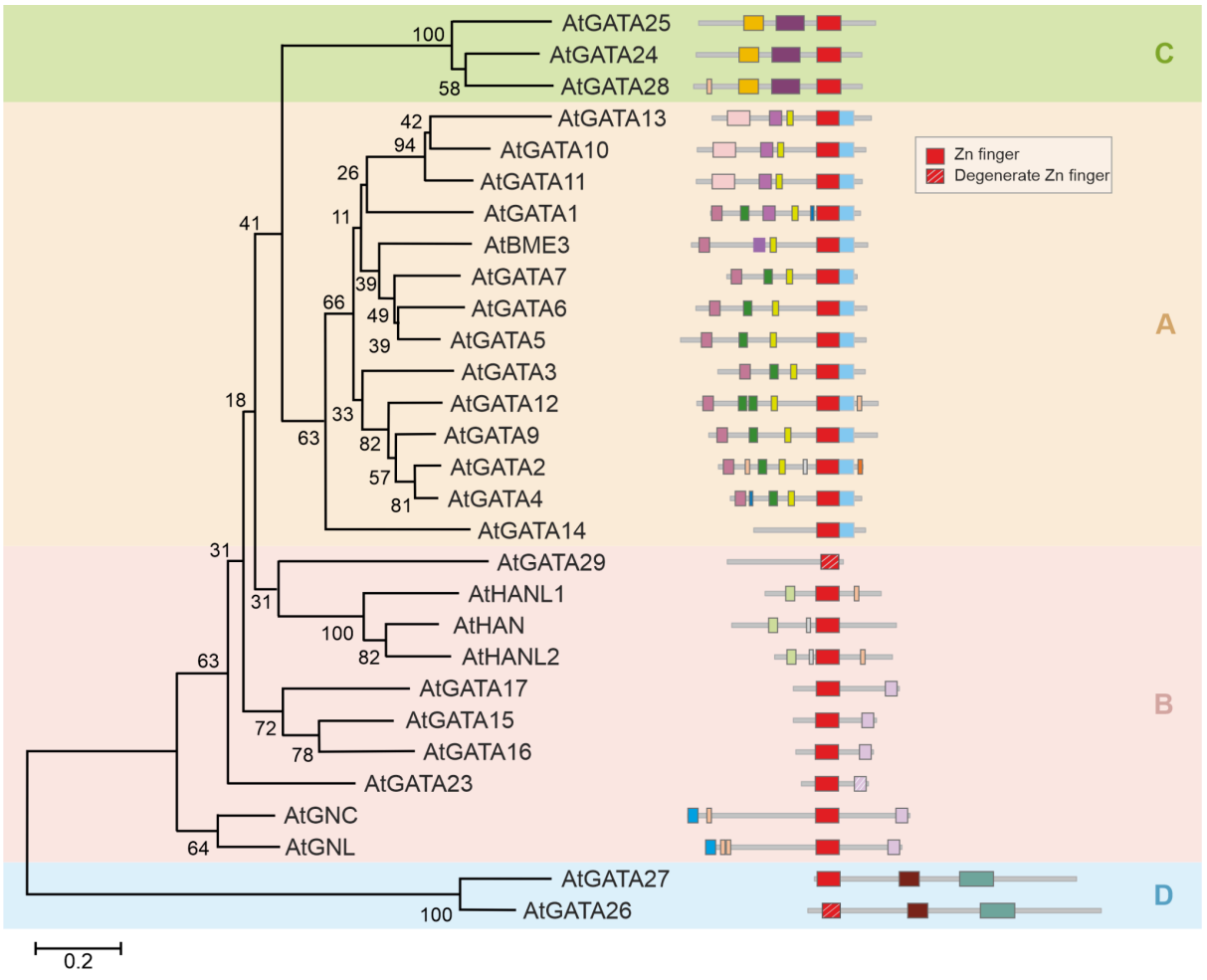


Figure 5: There are 30 GATA TFs in *A. thaliana*

Phylogenetic tree of Arabidopsis GATA factors (excluding GATA17L), adapted from Huang *et al.*, 2021. On the right, the protein structures of each GATA transcription factor are illustrated, with various domains highlighted using color-coded rectangles, adapted from Schwechheimer, Schröder and Blaby-Haas, 2022. The zinc finger domains are marked in red. The GATA TFs are categorized into four

classes: A (orange), B (red), C (green), and D (blue), based on their DNA-binding domain. For detailed descriptions of the other domains, refer to (Schwechheimer, Schröder and Blaby-Haas, 2022).

1.4. Questions and aims of this Master's thesis

The primary objective of this Master's thesis is to investigate whether GATA transcription factors indeed regulate gene expression via the GATC motif located downstream of the TSS and to determine whether this regulatory mechanism applies to all 30 GATA TFs or only a subset. To address this hypothesis, I employed two distinct experimental approaches. First, I utilized an effector protein derived from a phytopathogen, designed to degrade all GATA TFs, to examine the resulting changes in gene expression. In the second approach, I transiently overexpressed three specific GATA TFs in protoplasts to identify their target genes.

2. Results

2.1. Studying the role of GATA TFs in Arabidopsis using an effector protein that degrades all GATA factors

2.1.1. SAP05 is an effector protein that degrades all GATA TFs simultaneously

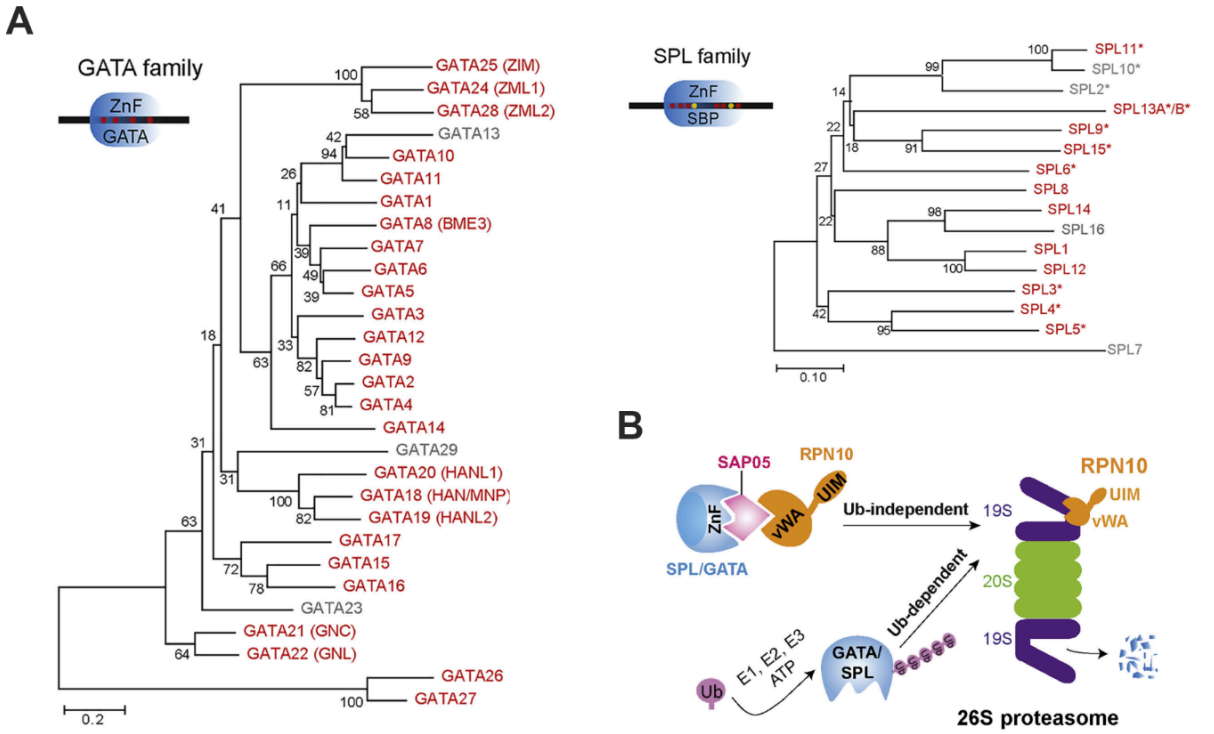
Candidatus Phytoplasma is a genus of pathogenic bacteria that causes insect-vectored diseases in plants (Lee, Davis and Gundersen-Rindal, 2000; The Irpcm Phytoplasma/Spiroplasma Working Team-Phytoplasma Taxonomy Group, 2004), with a lifecycle alternating between plant and insect hosts (Hogenhout *et al.*, 2008). Phytoplasmas infect the phloem sieve cells of the host plant and disperse through the sieve pores (Musetti *et al.*, 2013). *Hemiptera* insects transmit the pathogen by feeding on the phloem of an infected plant (Weintraub and Beanland, 2006). Phytoplasmas induce drastic changes in the plant architecture that include intensive generation of shoots (a phenomenon called "witches' broom") and production of leaf-like flowers (MacLean *et al.*, 2011; Maejima *et al.*, 2014). These morphological changes enable the efficient transmission of the pathogen and inhibit plant reproduction (Du Toit, 2014; MacLean *et al.*, 2014; Rümpler *et al.*, 2015).

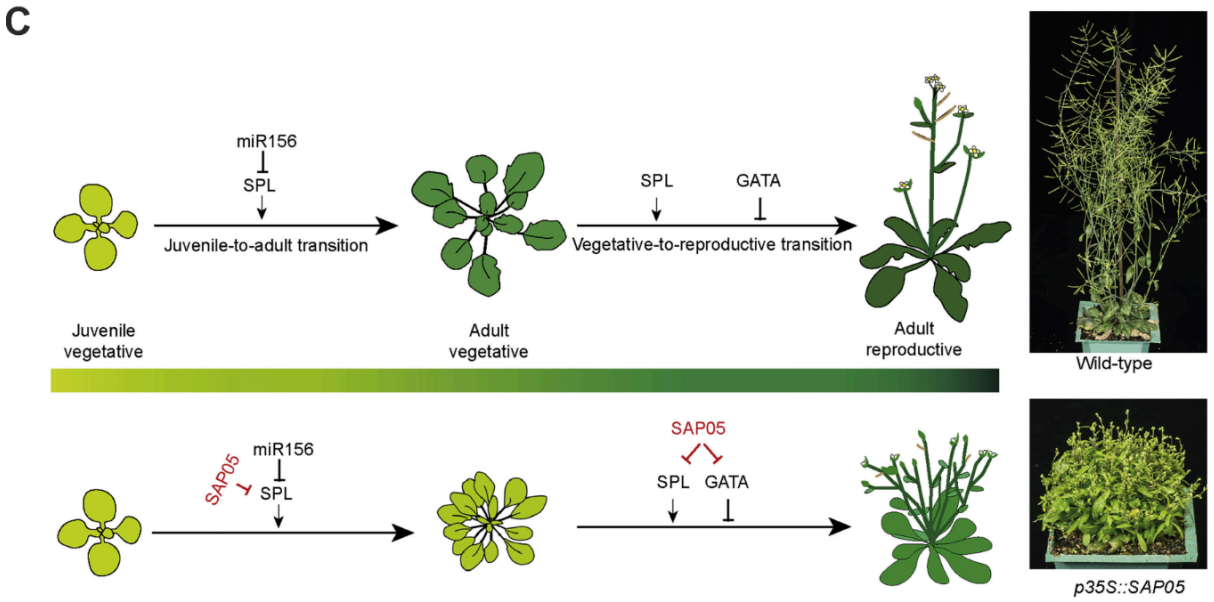
Phytoplasmas promote these effects via the secretion of effector proteins, some of which elicit their effect by degrading transcription factor families in the host. The genome of the well-studied phytoplasma, Aster Yellows Witches' Broom phytoplasma (AY-WB), encodes 56 putative effector genes, called SAP (secreted AY-WB proteins; Bai *et al.*, 2009). Some of them have been characterized: SAP11 targets TCP TFs (Sugio *et al.*, 2011), SAP54 - MADS box TFs (MacLean *et al.*, 2014), and the recently characterized SAP05 degrades both the SPL and the GATA family of TFs (Fig. 6A; Huang *et al.*, 2021). SAP05 facilitates the degradation of SPL and GATA TFs via a ubiquitin-independent pathway that links the TFs to the 26S proteasome component RPN10 (Fig. 6B; Huang *et al.*, 2021). SAP05 binds the GATA TFs through the

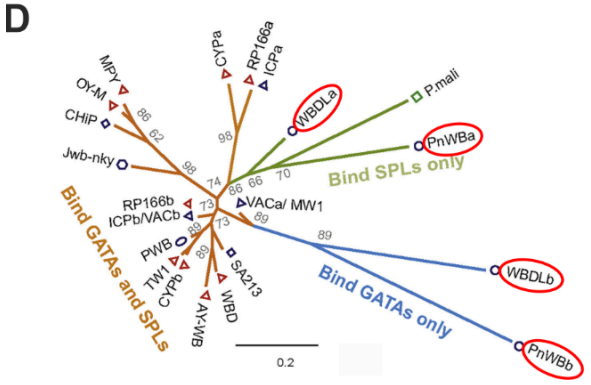
zinc-finger domain, which is the only conserved domain in the GATA family (Lowry and Atchley, 2000; Liu *et al.*, 2023).

SPL TFs regulate the juvenile-to-adult and vegetative-to-reproductive transitions in shoot development (Xu *et al.*, 2016), whereas GATA factors regulate photosynthesis (Bastakis *et al.*, 2018), as well as leaf (Behringer *et al.*, 2014) and flower development (Zhao *et al.*, 2004) (Fig. 6C). By degrading the two families of TFs, SAP05 disrupts the normal developmental transitions, causing plants to maintain their juvenile traits but still bolt to produce sterile flowering shoots. This is manifested by various phenotypic changes in the SAP05-expressing plants compared to wild-type plants: more rosette leaves without serrated edges, more lateral shoots and secondary branches, abnormal flowers with impaired fertility, decreased height, and delayed senescence (Fig. 6C; Huang *et al.*, 2021).

SAP05 homologs from other phytoplasma strains target only SPL TFs (PnWBa, WBDLa) or GATA TFs (PnWBb, WBDLb) (Fig. 6D; Huang *et al.*, 2021). As expected, these homologs also cause weaker phenotypic effects when overexpressed compared to the full witches' broom phenotypes, exhibited by AY-WB SAP05 expressing plants, in which both TFs families are degraded (Fig. 6E; Huang *et al.*, 2021). However, these studies focused solely on above-ground phenotypes and did not include an analysis of gene expression changes.


A
GATA family
ZnF
GATA
GATA25 (ZIM)
GATA24 (ZML1)
GATA28 (ZML2)
GATA13
GATA10
GATA11
GATA1
GATA8 (BME3)
GATA7
GATA6
GATA5
GATA3
GATA12
GATA9
GATA2
GATA4
GATA14
GATA29
GATA20 (HANL1)
GATA18 (HAN/MNP)
GATA19 (HANL2)
GATA17
GATA15
GATA16
GATA23
GATA21 (GNC)
GATA22 (GNL)
GATA26
GATA27
0.2
SPL family
ZnF
SBP
SPL11*
SPL10*
SPL2*
SPL13A*/B*
SPL9*
SPL15*
SPL6*
SPL8
SPL14
SPL16
SPL1
SPL12
SPL3*
SPL4*
SPL5*
SPL7
0.10
B
SAP05
RPN10
ZnF
vWA
UIM
SPL/GATA
Ub-independent
19S
RPN10
UIM
vWA
20S
19S
Ub-dependent
Ub
E1, E2, E3
ATP
GATA/
SPL
26S proteasome


C
miR156
SPL
Juvenile-to-adult transition
SPL
GATA
Vegetative-to-reproductive transition
Juvenile
vegetative
Adult
vegetative
Adult
reproductive
Wild-type
SAP05
miR156
SPL
SAP05
SPL
GATA
p35S::SAP05


D
MPY
OY-M
CHlP
Jwb-nky
RP166b
ICPb/VACb
PWB
TW1
CYPB
AY-WB
WBD
SA213
VACa/ MW1
CYPa
RP166a
ICPa
WBDLa
P.mali
PnWBa
Bind SPLs only
Bind GATAs and SPLs
Bind GATAs only
WBDLb
PnWBb
0.2

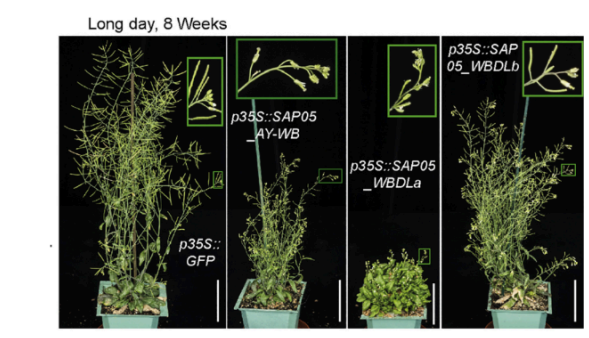

E
Long day, 8 Weeks
p35S::
GFP
p35S::SAP05
_AY-WB
p35S::SAP05
_WBDLa
p35S::SAP
05_WBDLb

Figure 6: SAP05 effector protein degrades GATA and SPL transcription factors

All subfigures are reproduced from Huang *et al.*, 2021. **(A)** Phylogenetic trees of the GATA and SPL families of transcription factors. SAP05 interaction partners in Y2H assays are displayed in red, while untested or autoactivating partners are in gray. The zinc-finger domains of the two families are depicted at the top left corners, with cysteine and histidine residues highlighted in red and yellow. **(B)** Model of the degradation mechanism of SAP05. The effector links the SPL and GATA TFs to RPN10, which induces degradation via the 26S proteasome. The ubiquitin-dependent pathway of degradation is shown for comparison. **(C)** Comparison of the developmental transitions in a wild-type and SAP05-expressing plant. In a wild-type plant, SPL TFs initiate juvenile-to-adult and vegetative-to-reproductive transitions, while GATA TFs suppress bolting. In a SAP05-expressing plant, these transitions are disrupted, leading to the formation of juvenile plants with flowering shoots, which manifests as the witches' broom phenotype. **(D)** Phylogenetic tree of SAP05 homologs from different Phytoplasma species based on alignment of their amino acid sequences. The proteins are grouped by their interaction partners - GATA and SPL TFs, only SPL TFs or only GATA TFs. The two SAP05 homologs of peanut witches' broom (PnWB) and witches' broom disease of lime (WBDL) phytoplasmas are highlighted in red. **(E)** Shoot phenotypes of GFP control line and SAP05 lines grown in long-day conditions. Insets display flowers from each plant. Scale bar, 4 cm.

2.1.2. Arabidopsis seedlings expressing the SAP05 effector have deformed roots

In our previous research, the strongest effect of the GATC motif was observed in root tissue (Fig. 7A-B). Therefore, we hypothesized that the degradation of GATA TFs by the SAP05 effector would affect root growth. However, the SAP05 effect on root growth has not been reported previously. To this end, we used two independent homozygous lines for each of the two GATA-specific SAP05 variants (PnWBb and WBDLb) under a CaMV 35S promoter - referred to as PnWBb #1, PnWBb #2, WBDLb #1 and WBDLb #2 in this text. In addition, a line expressing GFP under the CaMV 35S promoter was used as a control. All five lines were grown in parallel and their root length was measured 6 days after germination in three biological replicates (Fig. 8A).

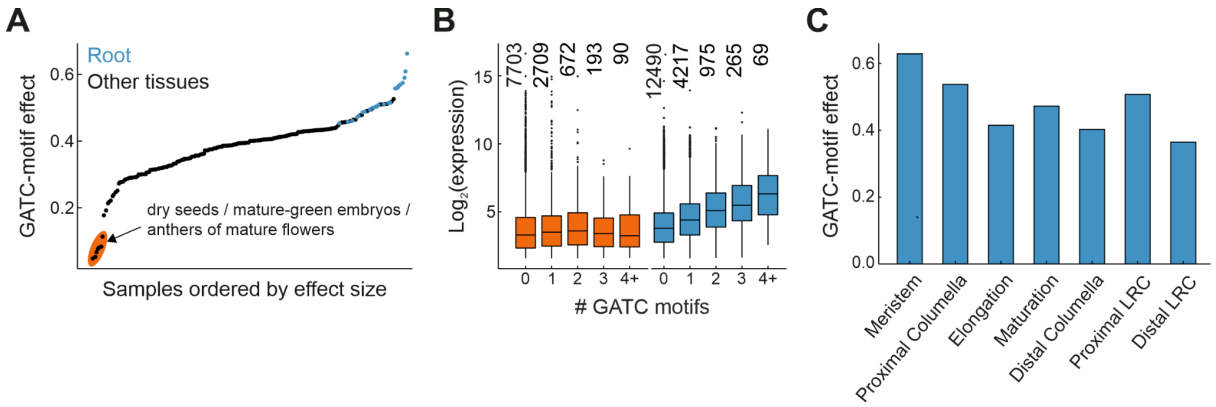


Figure 7: GATC-motif effect is tissue-specific

All subfigures are reproduced from Voichek *et al.*, 2023**. (A)** GATC-motif effect sizes quantified on gene expression from 200 different tissue-specific Arabidopsis samples (Toufighi *et al.*, 2005; Klepikova *et al.*, 2016; Hofmann, Schon and Nodine, 2019). Samples with the lowest and highest

GATC-motif effect sizes are highlighted and annotated. **(B)** Expression values in dry seeds (orange) and seedling roots (blue) (Klepikova *et al.*, 2016), with effect sizes of 0.07 (p-value: $2.6*10^{-3}$) and 0.57 (p-value: $4*10^{-400}$), respectively. The number of genes per category is indicated above each boxplot. **(C)** GATC-effect sizes across different root developmental stages averaged from single-cell expression data (Shahan *et al.*, 2022). LRC - lateral root cap.

As predicted, the lines expressing SAP05 have significantly shorter roots (Fig. 8B). Interestingly, there is variation between independent homozygous lines expressing the same effector, as well as between the two SAP05 variants. This phenotypic difference is consistent across the three replicates and is inherent to each homozygous line. However, there is a larger variation between the two SAP05 variants than between lines expressing the same effector. Therefore, the difference in lines with the same effector might be attributed to variation in the SAP05 gene expression due to different insertion sites. While the differences between SAP05 variants are more likely due to the effect of the protein-coding sequence.

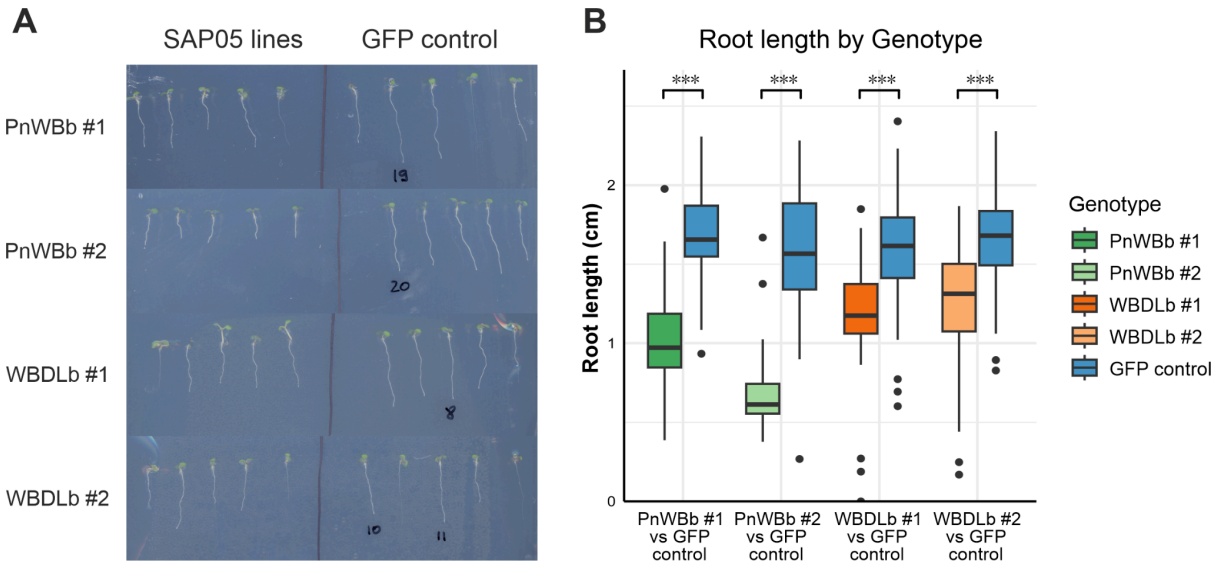


Figure 8: Lines expressing SAP05 effectors have significantly shorter roots than control plants
(A) Representative images of the four SAP05 lines grown on the same plates as the GFP control. The hand-written numbers under some seedlings on the right side mark the seedlings selected for confocal microscopy (shown in Fig. 9). **(B)** Root length quantification of all seedlings from each of the five lines. Each of the SAP05 lines was compared to the GFP control plants grown on the same plates. The boxplots represent combined data from three biological replicates (n = 58-60). p-values: <2*10-16, <2*10-16, 4.319*10-9 and 2.312*10-11, respectively (calculated using Type III ANOVA).

Differences in root length can be due to slower growth in different parts of the root. Root growth primarily happens in the root tip which is composed of a meristematic zone, where cells rapidly divide, an elongation zone where cells increase in size, and a differentiation zone where cells gain their specialized roles, and where root hairs begin to appear (Fig. 9A, Petricka, Winter and Benfey, 2012). In order to understand which zone is affected by the

SAP05 effector, we stained the same seedlings used for root length measurement with a cell wall dye (SCRI Renaissance 2200, Musielak *et al.*, 2016) and imaged the root tip using confocal microscopy.

Root tip images of the SAP05 lines revealed that the tissue organization in the meristematic zone is completely disrupted (Fig. 9B). The columella in the root cap is missing or deformed, as evidenced by the lack of parallel cell files in this region. The quiescent center, from which all other tissues in the root originate (Petricka, Winter and Benfey, 2012), is also missing in the SAP05 seedlings. In addition, the proper cell division in the meristematic zone seems to be disturbed - there are diagonal cell walls and more periclinal divisions, resulting in additional cell files. Lastly, the meristematic zone in the SAP05 lines appears to be shorter with a reduced number of cells compared to the control plants. In support of this observation, we quantified the number of cells in the meristematic zone, which corresponds to the number of cell divisions before cells transition to the elongation zone. All four SAP05 lines demonstrate significantly decreased meristem cell counts relative to the GFP control line (Fig. 9C-D). Therefore, the shorter roots of the SAP05 lines could be attributed to diminished cell division in the meristem and the disrupted tissue organization of the root tip.

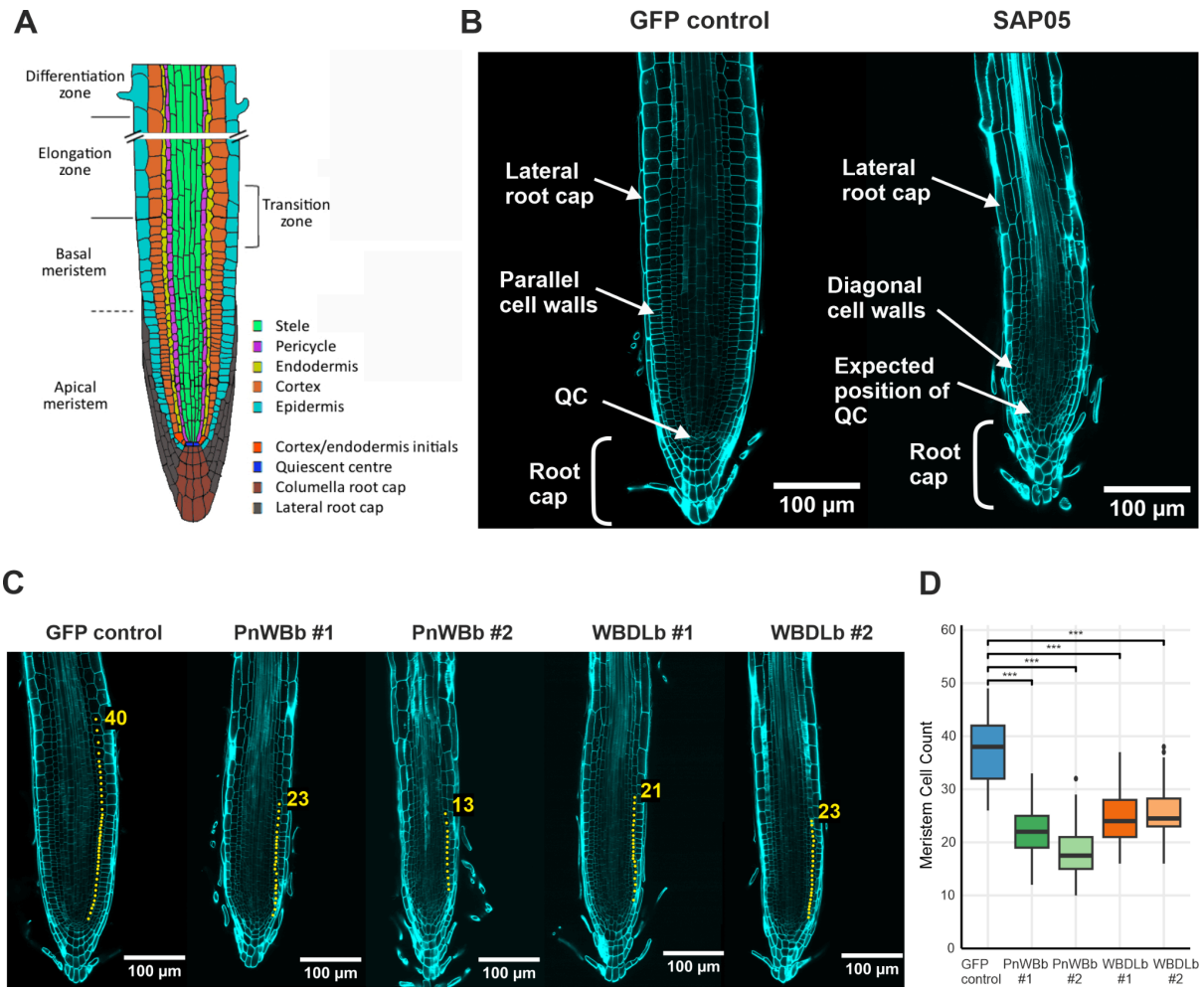


Figure 9: SAP05 lines have disrupted root tip organization and reduced number of root meristem cells

(A) Diagram of the root tip displaying the three zones (meristematic zone, elongation zone, and differentiation zone) and the tissue organization. Figure reproduced from De Smet *et al.*, 2015. **(B)** Representative confocal microscopy images of the root tip of a GFP control seedling and a SAP05 seedling (PnWBb #2). The differences between them are highlighted - in the SAP05 seedling, the root cap columella is missing or deformed, the quiescent center (QC) is missing, there are diagonal cell walls and the lateral root cap is expanded. **(C)** Quantification of the meristem cells of all imaged seedlings from each of the five lines. Each of the SAP05 lines was compared to the GFP control plants. The boxplots represent combined data from three biological replicates (n = 56-59). p-values: $<2*10^{-16}$, $<2*10^{-16}$, $<2*10^{-16}$ and $<2*10^{-16}$, respectively (calculated using Type III ANOVA). **(D)** Representative confocal microscopy images of the root tip of each of the five lines. The quantification of meristem cells is displayed by the yellow dots and the number next to them. The meristem cells were counted only in the meristematic zone of the cortex cell file and primarily on the right side of the root tip. The elongation zone of the cortex file is marked by the first cell with a greater height than width (Perilli and Sabatini, 2010).

Taken together, these findings suggest that SAP05 expression disrupts root development by altering the meristematic zone of the root tip. This aligns with our previous observation that the GATC effect is strongest in the root meristem compared to other tissues in the root (Fig. 7C).

2.1.3. Genome-wide gene expression of SAP05 expressing lines

The severe root phenotypes of the SAP05 expressing lines are consistent with our predicted effect of the GATC motif. However, while the SAP05 effector was shown to degrade GATA TFs previously, the downstream gene expression response to it was not studied. Therefore, to understand the effect of SAP05 on gene expression, we measured the genome-wide gene expression in the SAP05 lines. The four SAP05 lines and the GFP control were grown under the same conditions used for phenotyping, and the roots and shoots were harvested separately for RNA sequencing (RNA-Seq). Ten seedlings from each line were used for each of the three independent biological replicates (Fig. 10A).

We first validated that the SAP05 variants and the GFP are indeed overexpressed in their corresponding line (Fig. 10B-C). Next, we clustered the samples - as expected, roots and shoots clustered together (Fig. 10D), which was also confirmed in the principal component analysis (PCA, Fig. 10E). Surprisingly, gene expression of the SAP05-expressing lines did not cluster separately from the controls (Fig. 10D-E). Moreover, the predicted effect of SAP05 on the expression of genes with a GATC motif was not observed - the expression levels of these genes were comparable to those measured in the GFP control line (Fig. 10F).

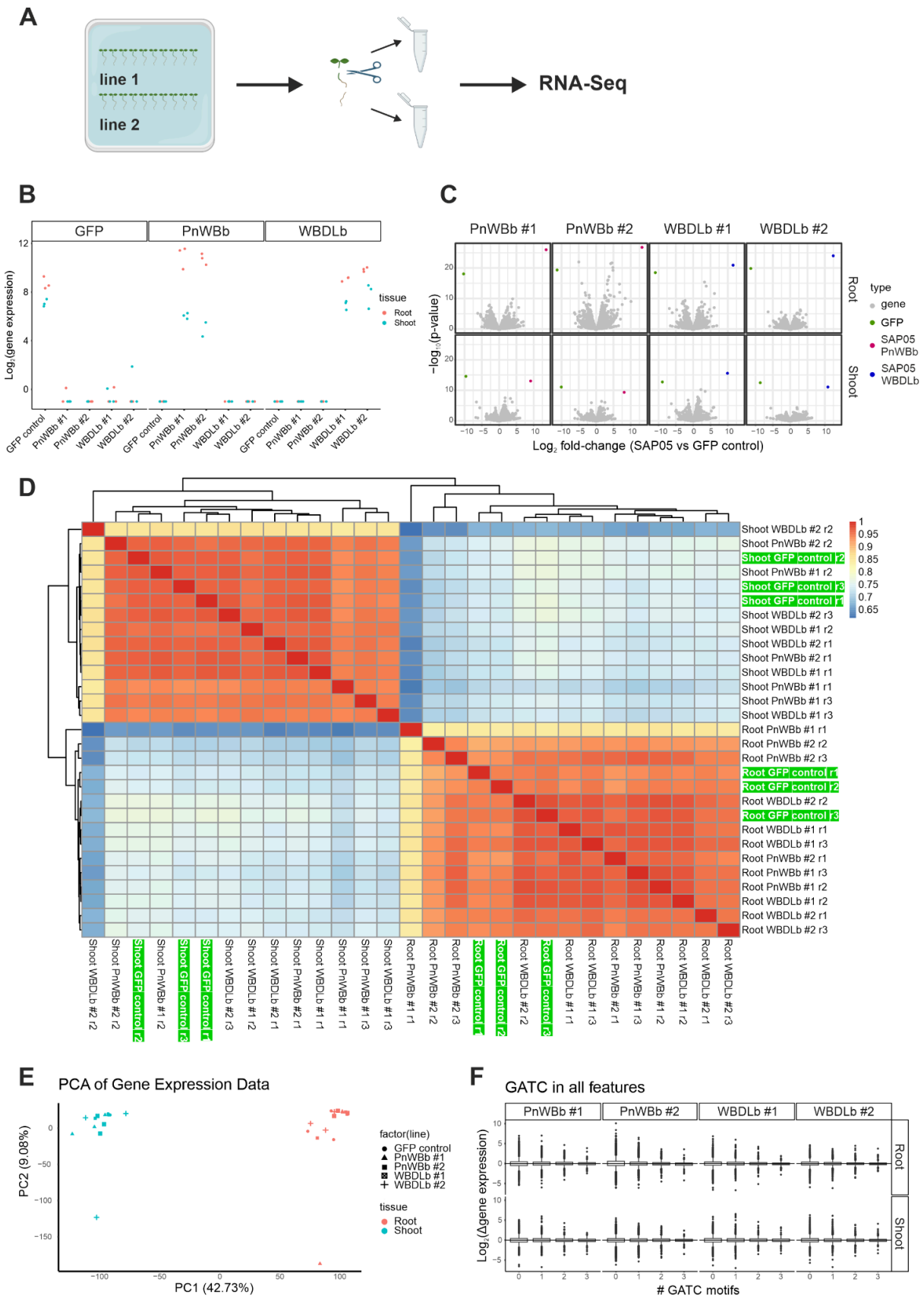


Figure 10: Genome-wide gene expression in SAP05 lines

(A) Experimental setup: 10 seedlings from each of the five lines (PnWBb #1, PnWBb #2, WBDLb #1, WBDLb #2, and GFP control) were grown for 6 days. The shoots and roots of all seedlings from each

line were harvested separately, followed by RNA extraction and RNA-seq. The experiment was done in three biological replicates. **(B)** Verification of our lines: Gene expression levels of GFP, SAP05-PnWBb, and SAP05-WBDLb (as indicated in the panel title) are shown. Expression levels are log2-transformed and plotted for the five lines across root (red) and shoot (blue) tissue, and three replicates. **(C)** Differential expression analysis: A volcano plot shows the -log10 p-values relative to the log2 fold changes in gene expression of the four SAP05 lines compared to the GFP control line. Calculations were performed using DESeq2 (Love, Huber and Anders, 2014) based on three replicates per genotype. The GFP, SAP05-PnWBb, and SAP05-WBDLb genes are highlighted in green, red, and blue respectively. **(D)** Pearson's correlation coefficients of all samples, hierarchically clustered. The GFP control samples are highlighted in green. r1, r2, and r3 - replicate 1, 2 and 3. **(E)** Principal component analysis of all samples. Genotypes are coded by symbol shape, and tissues - by color (red - root, blue - shoot). **(F)** Gene expression of the four SAP05 lines compared to the GFP control line across the root (top panels) and the shoot (bottom panels) tissue, plotted for genes with 0, 1, 2, or ≥3 GATC motifs in the 500 bp downstream of their TSS.

2.1.4. Genome-wide gene expression response to transient SAP05 expression

The lack of significant differences in gene expression between the SAP05 lines and the control was unexpected, given the pronounced phenotypic differences. One possible explanation is that the effect of SAP05 may be specific to the meristematic zone but is diluted by the more abundant cells in the elongation and differentiation zones. Additionally, we did not observe the anticipated decrease in expression of GATC-containing genes. This could be due to SAP05's impact on root development, as the detected gene expression changes might not be a direct response to SAP05 expression. To better understand SAP05's direct effect on gene expression, we decided to measure gene expression upon transient expression of SAP05. This approach will eliminate the issues of signal dilution as well as downstream changes due to developmental effects.

Transient expression of external DNA in Arabidopsis can be achieved using protoplasts. Protoplasts are plant cells without a cell wall, which allows the efficient delivery of external DNA (Fig. 11). Protoplasts have been used for over 60 years to study different aspects of plant biology (Cocking, 1960), leading to numerous fundamental discoveries. For instance, key properties of splicing in plants were identified through transient transformation of *N. plumbaginifolia* protoplasts (Goodall and Filipowicz, 1989). Additionally, the first Auxin-responsive element (AuxRE) was characterized in the auxin-induced gene, *PS-IAA4/5*, using a transient assay in pea epicotyl protoplasts (Ballas, Wong and Theologis, 1993). Moreover, multiple techniques for direct visualization of protein-protein interactions were developed in protoplasts (Subramaniam *et al.*, 2001; Walter *et al.*, 2004), enabling the rapid verification of these interactions in vivo. Therefore, protoplasts prove to be an essential tool for advancing our understanding of plant biology.

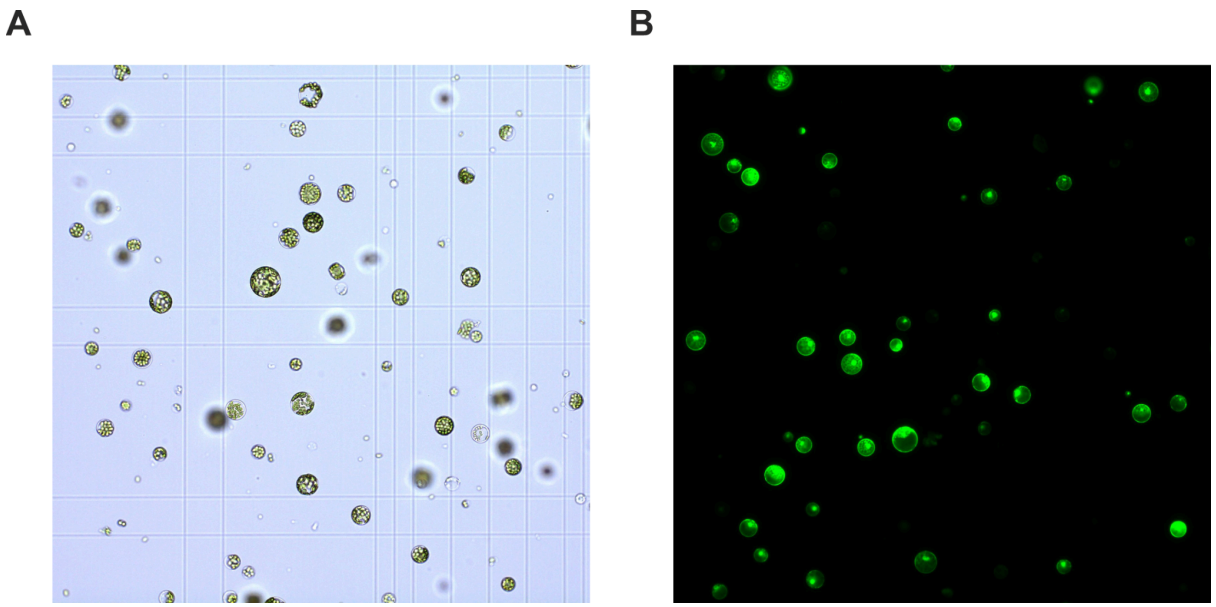


Figure 11: Transient transformation of Arabidopsis protoplasts
(A) Image of protoplasts isolated from leaf mesophyll tissue of *Arabidopsis thaliana* obtained by bright-field microscope under 10x magnification. The protoplasts are counted on a hemacytometer before transformation. **(B)** Successfully transformed protoplasts expressing GFP with a nuclear localization signal driven by *UBI4* promoter after overnight incubation. The image is taken on an epifluorescent microscope under 10x magnification.

To quantify the direct effect of SAP05 on gene expression, we transformed leaf-derived Arabidopsis protoplasts with either of the two SAP05 variants driven by a double 35S promoter, a GFP control plasmid driven by a *UBI4* promoter (positive control), and a negative control with the same volume of elution buffer. As our transformation efficiency was approximately 50%, bulk gene-expression measurement will dilute the SAP05 effect due to the presence of untransformed protoplasts. Therefore, we selected only transformed protoplasts by Fluorescence activated cell sorting (FACS). To this end, we cloned a GFP gene under the control of *UBI4* promoter on the same plasmids as the SAP05 variants. In this way, the transformed protoplasts could be distinguished by their fluorescence and sorted out by FACS. As a control for the effects of FACS on gene expression, the same number of samples was harvested without prior sorting. The experiment was performed in three biological replicates (Fig. 12A).

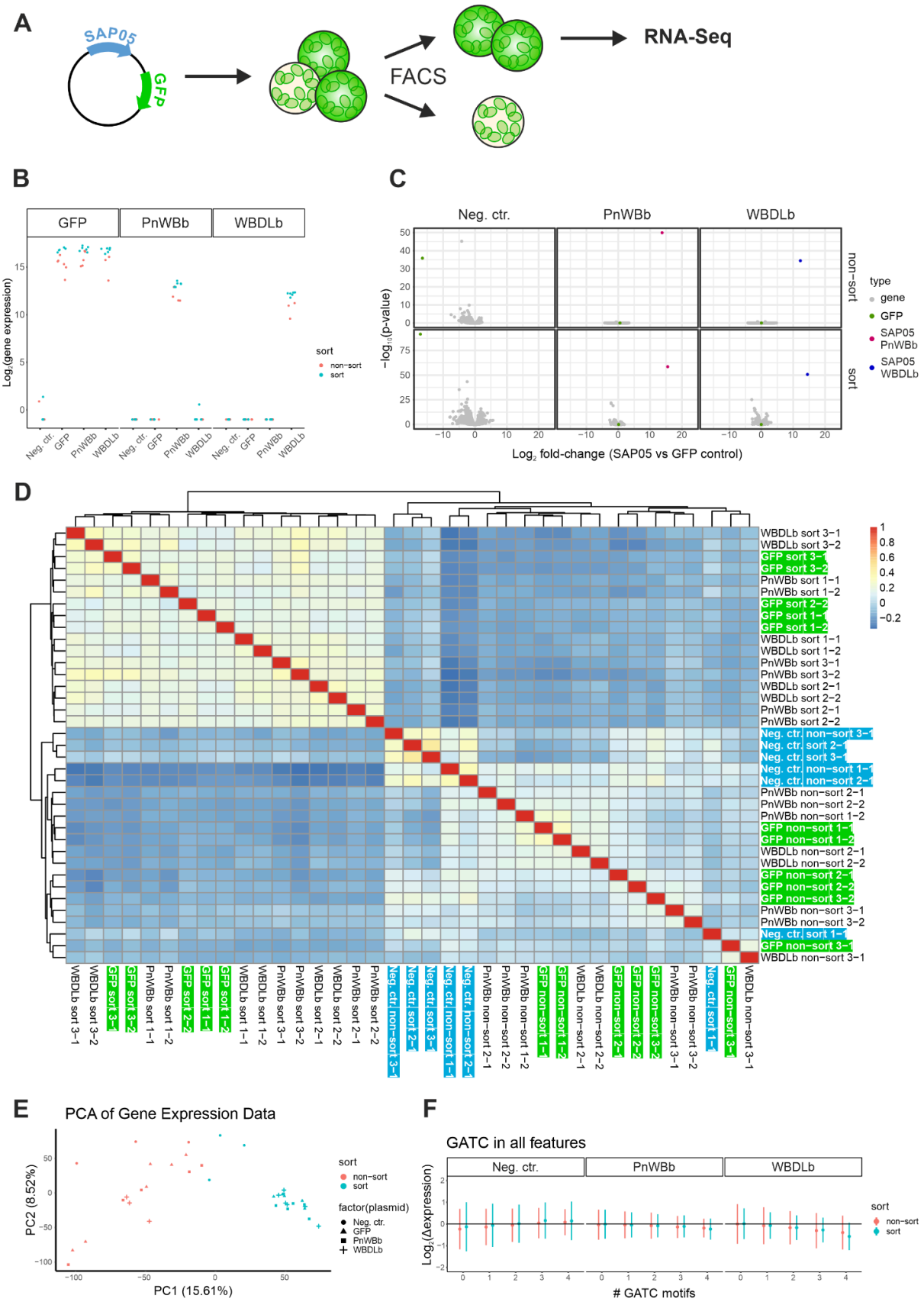


Figure 12: Transient overexpression of SAP05 genes in protoplasts

(A) Experimental setup: Arabidopsis leaf protoplasts were isolated and transformed by PEG-mediated delivery with plasmids carrying SAP05-PnWBb or SAP05-WBDLb under the control of a double 35S

promoter and a GFP reporter gene under the control of *UBI4* promoter, on the same plasmid. A positive control was transformed with a plasmid carrying GFP driven by *UBI4* promoter and a negative control was transformed with the same volume of elution buffer. 21 hours post-transformation, half of the samples ("non-sorted") were harvested for RNA extraction. The other half of the samples were sorted with Fluorescence Activated Cell Sorting (FACS) and GFP-positive cells were harvested for SAP05 or GFP control. Negative control samples also went through the FACS and were harvested as a control. RNA was extracted immediately after harvesting. Three biological repeats of protoplast isolation were done, with each plasmid transformed four times per isolation replicate, resulting in six replicates per plasmid for each condition (sorted/non-sorted samples). In each biological replicate, two negative controls were transformed - one for the sorted and one for the non-sorted samples. **(B)** Verification of overexpression: Gene expression levels of GFP, SAP05-PnWBb, and SAP05-WBDLb (as indicated in the panel title) are shown. Log2-transformed expression levels are plotted for the three plasmids (with six replicates each) and the negative control (with three replicates) across non-sorted (red) and sorted (blue) samples. **(C)** Differential expression analysis: A volcano plot shows the -log10 p-values relative to the log2 fold changes in gene expression of the SAP05 samples and the negative control compared to the positive controls. The top and bottom panels show the non-sorted and the sorted samples, respectively. Calculations were performed using DESeq2 (Love, Huber and Anders, 2014) based on six replicates per plasmid and three replicates of the negative control. The GFP, SAP05-PnWBb, and SAP05-WBDLb genes are highlighted in green, red, and blue respectively. **(D)** Pearson's correlation coefficients of all samples, hierarchically clustered. The positive and negative controls are highlighted in green and blue, respectively. The biological replicate is indicated with the first number after the sample names, while the technical replicate with the second number. **(E)** Principal component analysis of all samples. Plasmids are coded by symbol shape, and conditions - by color (red - non-sorted, blue - sorted). **(F)** Log fold change of gene expression of SAP05 samples or negative controls compared to the positive controls. Data is shown separately for non-sorted (red) and sorted (blue) samples, plotted for genes with 0, 1, 2, 3, or ≥4 GATC motifs downstream of the TSS.

Analysis of the expression levels of GFP and the two SAP05 variants (PnWBb and WBDLb) validated our experimental setup, demonstrating the expression of each plasmid in the correct samples (Fig. 12B-C). The sorting itself had a consistent effect on gene expression, as both correlation analysis and PCA analysis showed that the sorted samples clustered separately from the non-sorted ones (Fig. 12D-E). Interestingly, also in these experiments, the SAP05 samples did not cluster together. However, the SAP05-expressing protoplasts showed a decreased expression of GATC-containing genes compared to the positive control (Fig. 12F). The decrease in expression was proportional to the number of GATC motifs downstream of the gene's TSS, with the most significant effect observed in genes containing four or more motifs. Notably, the two SAP05 variants differ in their impact: WBDLb expression resulted in a stronger downregulation of GATC genes compared to PnWBb. This contrasts with observations in the SAP05 lines, where PnWBb-expressing lines exhibited a stronger root phenotype than WBDLb lines (Fig. 8B, 9C).

In conclusion, the results from the transient expression of SAP05 effectors in protoplasts provide support for our hypothesis that GATA transcription factors regulate the expression of

GATC-containing genes. However, the effect observed is not as strong as expected from the increase in expression we have previously observed in synthetic systems when adding GATC motifs. We can only speculate on the source of these differences: first, it is not clear how much of the GATA TFs are left in the SAP05 expressing protoplasts. Second, the SAP05 works through the degradation of the GATA TFs, and thus the effect we measure on gene expression is due to the degradation of mRNA. This process is slower than changes due to activation of transcription and might be even slower in the protoplast system, where cells are not dividing. Therefore, as a complementary approach, we decided to use an activating system as explained below.

2.2. Overexpression of GATA factors leads to upregulation of genes with a GATC motif

Degrading all GATA TFs with the SAP05 effector indicated that GATAs regulate the expression of genes with a GATC motif. To validate these findings, we aimed to complement the SAP05 experiments by transiently overexpressing GATA TFs in protoplasts and measuring the gene expression response. To determine the optimal incubation period for transformed protoplasts, we conducted a time-course experiment using a GFP control plasmid. We assessed the percentage of GFP-expressing protoplasts at nine time points post-transformation: 40 minutes, 1h 20 minutes, 2h, 3h, 4h, 5h, 6h, 7h, and 8h (Fig. 13, 14A). Fluorescent protoplasts were first detected at 2h post-transformation, with a steady increase in fluorescence over the next 6 hours, reaching 46% at 8h (Fig. 14A). Based on these results, we selected an 8-hour incubation period for the transient overexpression experiments.

We chose three GATA TFs (GATA1, GATA4, and GATA6) based on their high expression levels in leaf tissue and confirmed their binding preference to sequences with GATC motifs in the DAP-Seq study (Fig. 3; O'Malley *et al.*, 2016). Arabidopsis leaf protoplasts were then transformed with plasmids encoding either GATA1, GATA4, GATA6, or GFP, driven by a double 35S promoter. Following an 8-hour incubation, the protoplasts were harvested for RNA-Seq. The experiment was conducted in four replicates per plasmid (Fig. 14B).

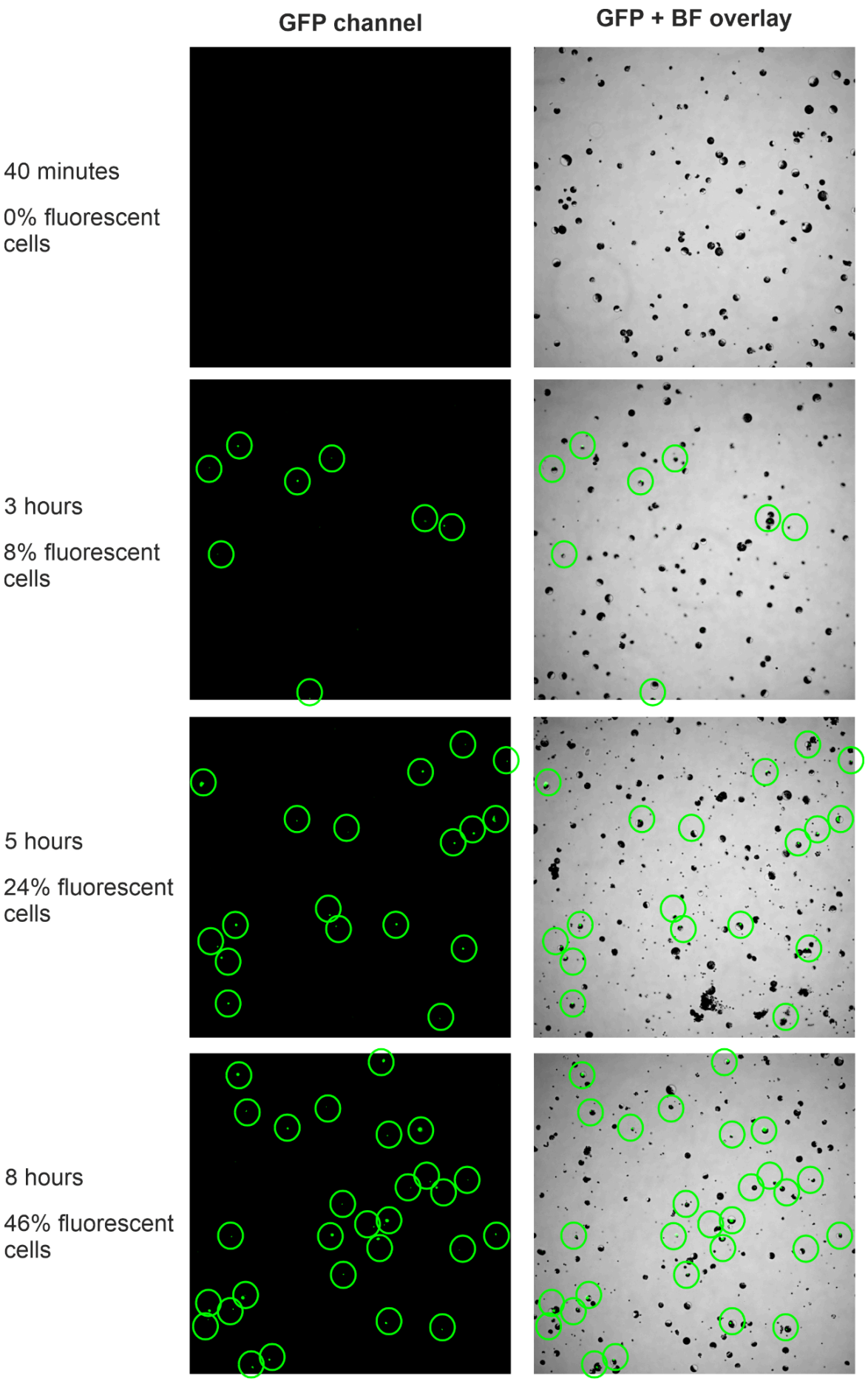

GFP channel
GFP + BF overlay
40 minutes
0% fluorescent cells
3 hours
8% fluorescent cells
5 hours
24% fluorescent cells
8 hours
46% fluorescent cells

Figure 13: Transient overexpression of GFP in protoplasts over 8 hours

1.8 million leaf-derived Arabidopsis protoplasts were transformed with a plasmid encoding GFP driven by a double 35S promoter. The same number of protoplasts were transformed with an equal volume of elution buffer, serving as a negative control. The two samples were each split into 9 aliquots of 0.2 million protoplasts which were incubated for 9 different periods - 40', 1h 20', 2h, 3h, 4h, 5h, 6h, 7h, and 8h. The respective GFP-transformed sample and the negative control were imaged under a fluorescent microscope at each time point. The percentage of fluorescent cells was quantified for each GFP-transformed sample as the average ratio of fluorescent cells to all cells visible in the frame (3-6 frames per time point). Representative images of four time points are depicted - 40', 3h, 5h, and 8h. On the left are images taken in the GFP channel, showing the fluorescent cells, while on the right are overlay images combining the GFP and bright-field (BF) channel, showing all cells. The fluorescent cells are highlighted with green circles and the percentages of fluorescent cells are depicted under each time point.

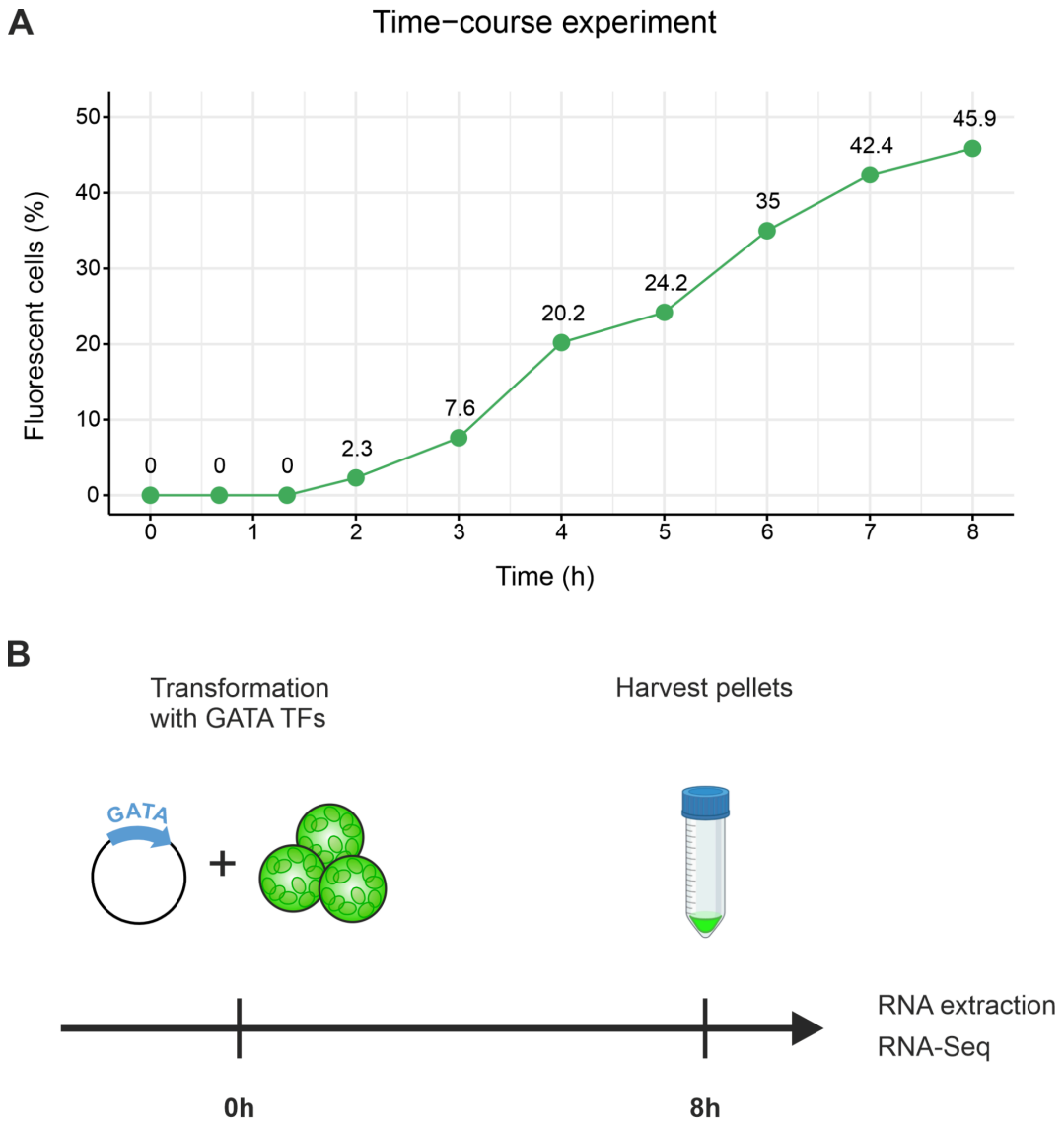


Figure 14: Time-course quantification and experimental setup for GATA overexpression

(A) Percentage of fluorescent cells in the GFP-transformed samples for all time points. **(B)** Experimental setup for GATA overexpression experiment: Arabidopsis leaf protoplasts were isolated and transformed with plasmids encoding GATA1, GATA4, GATA6 TFs, or GFP (positive control) driven by a double 35S promoter. A positive and negative imaging control were transformed with the GFP plasmid and an equal volume of elution buffer, respectively. The imaging controls were imaged with a fluorescent microscope after 7h incubation to determine the transformation efficiency. All other samples were harvested for RNA extraction after 8h incubation. Two biological repeats of protoplast isolation were done, with each plasmid transformed two times per isolation replicate, resulting in four replicates per plasmid.

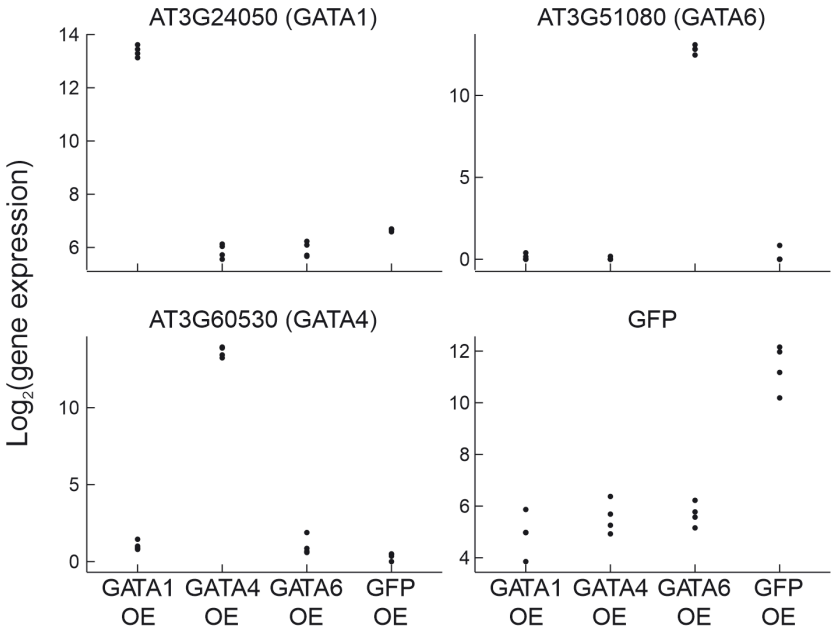


Figure 15: Verification of overexpression (OE) experiments
Gene expression levels of GATA1, GATA4, GATA6, and GFP (as indicated in the panel title) are shown. Expression levels are log2-transformed and plotted for each sample across four replicates.

The three GATA TFs and GFP were overexpressed in the appropriate samples, validating our experimental setup (Fig. 15). The plasmids used in this experiment also contain a selection cassette encoding YFP under the control of a seed coat promoter. This resulted in a low-level "leaky" expression of YFP across all samples, which was detected as GFP expression due to cross-alignment of sequencing reads (Fig. 15, bottom right panel).

Differential expression analysis identified numerous genes that were differentially expressed in response to GATA overexpression (Fig. 16A). A closer look at the upregulated genes revealed significant overlap with GATA TF targets identified in the DAP-Seq study (Fig. 16B). Notably, over 50% of the upregulated genes contained at least one GATC motif, compared to only about 25% of the genes with unchanged expression (Fig. 17A). In contrast, the percentage of downregulated genes containing a GATC motif was significantly lower than in the unchanged genes (Fig. 17A). Furthermore, the degree of upregulation correlated with the number of GATC motifs in the genes, with those containing three or more GATC motifs showing the highest increase in expression (Fig. 17B).

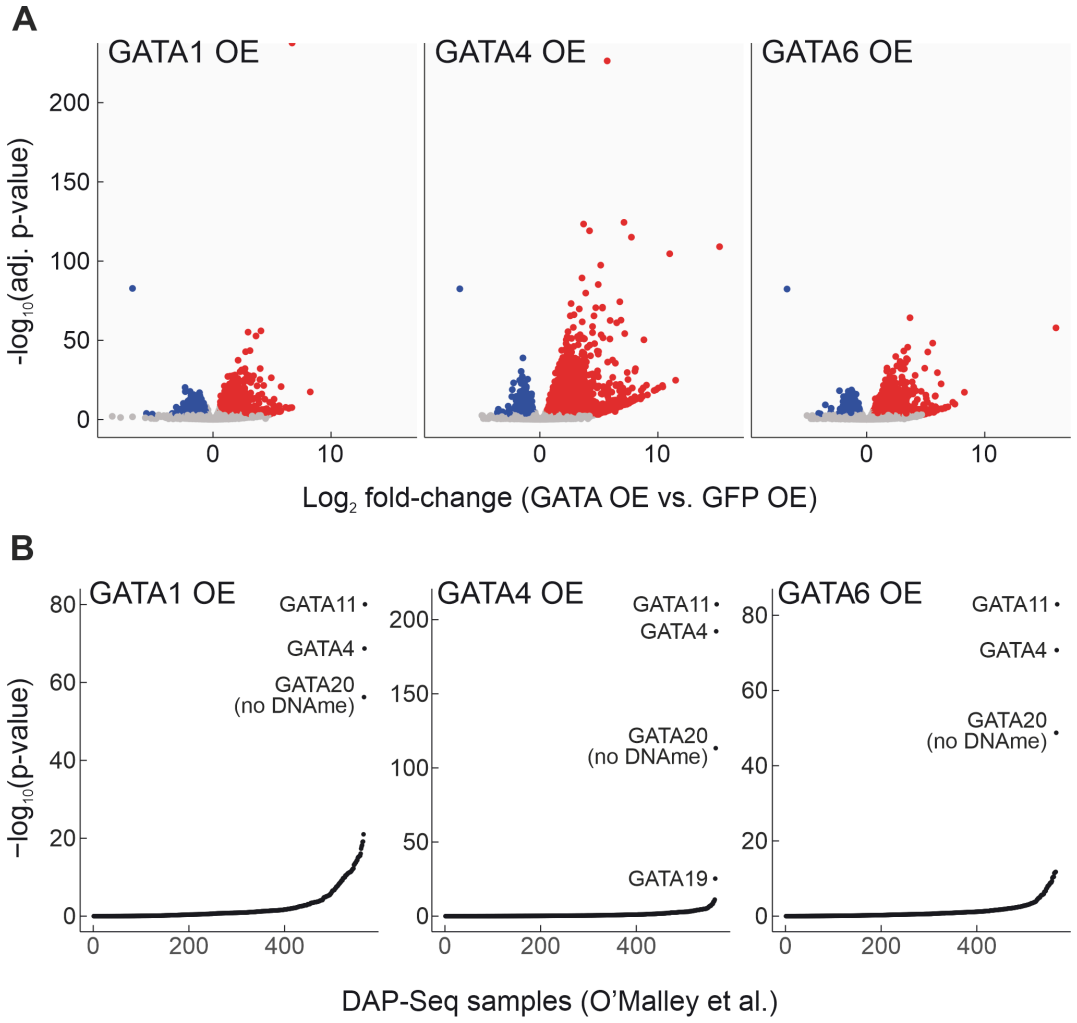


Figure 16: GATA overexpression (OE) causes upregulation of genes identified as GATA targets by DAP-Seq

(A) Differential expression analysis: A volcano plot illustrates the -log10 adjusted p-values against the log2 fold changes in gene expression after OE of each GATA TF compared to GFP OE. The analysis was conducted using DESeq2 (Love, Huber and Anders, 2014), based on four replicates per sample. Genes that exhibited a change in expression of at least 50% with an adjusted p-value of less than 0.001 were considered differentially expressed. These genes are highlighted in blue (downregulated) and red (upregulated), while all other genes are represented in gray. **(B)** Enrichment analysis: Gene targets identified by DAP-Seq for all TFs in the study were compared with genes upregulated (as defined in A) following overexpression of the GATA TFs. Enrichment p-values were calculated using a hypergeometric test and presented for each TF sample previously analyzed by DAP-Seq. The results indicate that in the three GATA overexpression experiments, the upregulated genes are most enriched with the targets of GATA TFs in the DAP-Seq data.

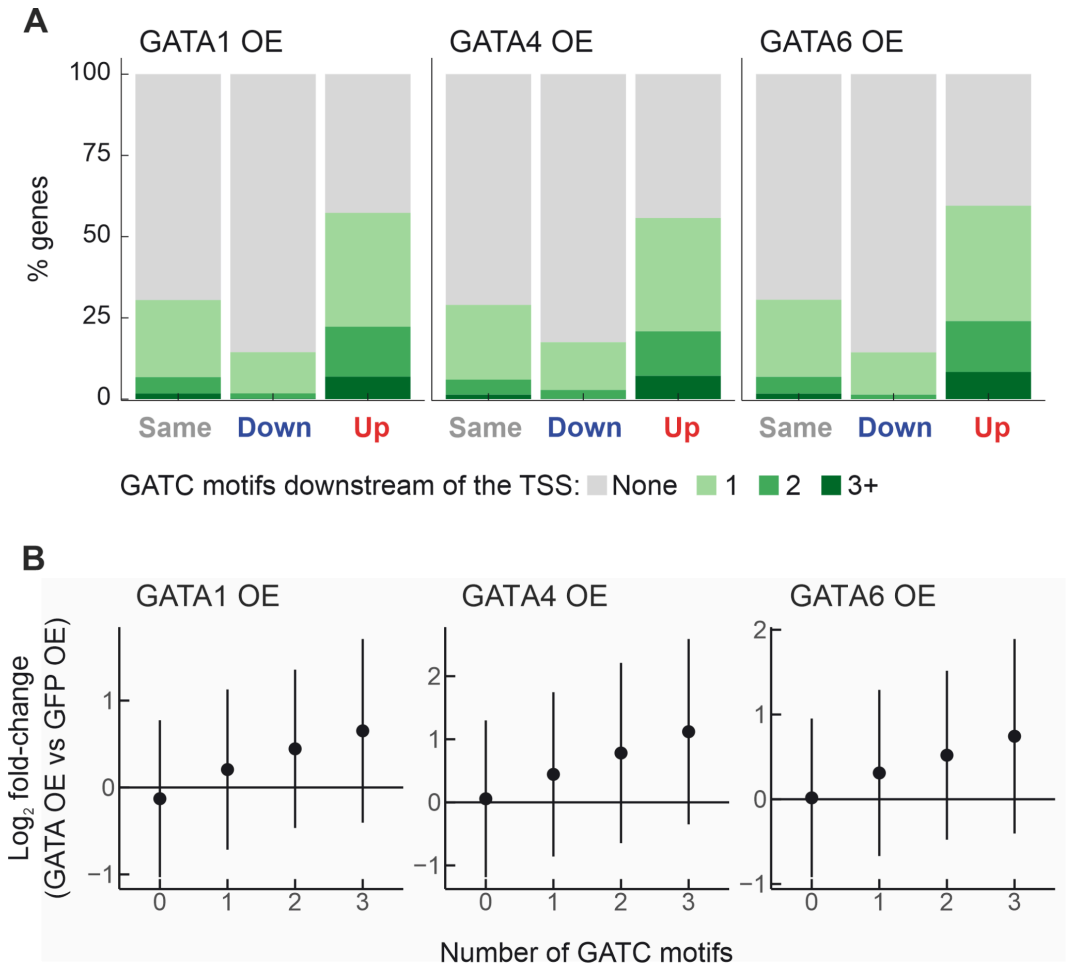


Figure 17: GATA overexpression upregulates genes proportionally to the number of GATC motifs
(A) Proportion of expressed genes in each OE experiment based on the number of GATC motifs within 500 bp downstream of the TSS in them (depicted in gray for no GATC motifs and different green hues for 1, 2, and 3+ motifs). The genes are divided according to their expression relative to the GFP OE samples - upregulated ("Up"), downregulated ("Down"), and unchanged ("Same"), as defined in Fig. 16A. **(B)** Gene expression response to overexpression of GATA TFs compared to GFP, plotted for genes with 0, 1, 2, or ≥3 GATC motifs within 500 bp downstream of the TSS. The groups contain 11,885±420, 4,174±124, 973±27, and 350±7 genes, respectively.

To conclude, these findings suggest a direct role for GATA TFs in regulating gene expression through GATC motifs, with the extent of gene activation strongly linked to the number of these motifs.

3. Discussion

In this thesis, I employed two distinct approaches to investigate the relationship between GATA TFs and the GATC motif. The first approach, involving an effector protein, provided only limited support for this connection. In contrast, the second, more direct approach offered

compelling evidence that GATA TFs regulate the expression of genes through this motif. Below, I will discuss these findings and explore their possible explanations.

The root phenotypes observed in SAP05-expressing lines align with the finding that the GATC motif effect is most pronounced in roots. However, gene expression analysis in these lines does not reveal significant differences from the control line. This discrepancy could be attributed to the potential dilution of gene expression signals from other cell types within the root, especially since the phenotype is confined to the root tip. Furthermore, the SAP05 lines did not show the expected decrease in expression of GATC-containing genes, which could be explained by SAP05's effect on root development, leading to downstream changes in gene expression.

When SAP05 was transiently expressed, the expression of genes with a GATC motif was reduced, supporting our hypothesis. However, the effect was relatively subtle, and the samples were segregated by sorting treatment rather than transformed plasmid. This could be due to the insufficient incubation period; SAP05 requires time to be expressed and degrade the GATA TFs, which will lead to reduced transcription of GATC-containing genes. Since we measure gene expression rather than transcription rates, the effects will only appear after mRNA degradation has occurred. Therefore, a longer incubation period might be necessary for the impact on the expression of GATC genes to become more evident.

Transient overexpression of GATA TFs demonstrated that GATC-containing genes were direct targets of GATA TFs, which is consistent with our hypothesis. However, these results were limited to only GATA TFs from class A. Further investigation of the other three classes is needed to determine how other GATA TFs influence gene expression, which will provide a more comprehensive understanding of their mode of action.

In conclusion, this thesis presents evidence that GATA factors in Arabidopsis regulate gene expression through a GATC motif located within 500 bp downstream of the TSS. This motif is found in 7,397 genes, enriched for the gene ontology terms of the Golgi apparatus, endoplasmic reticulum, endosomes, and vesicle-mediated transport (Voichek *et al.*, 2023). Considering the diverse roles of GATA factors in plants, it is plausible that they perform their functions by regulating fundamental cellular processes such as secretion. The dose-dependent effect of the GATC motif may have evolved as a mechanism to fine-tune the expression of the numerous genes controlled by GATA factors. Moreover, the expansion of GATA factors in plants reflects their crucial roles, providing redundancy for robustness and enabling precise spatial and temporal regulation of gene expression.

4. Methods and materials

4.1. Plant material and growth conditions

Four SAP05-expressing lines (PnWBb-SAP05 #1, PnWBb-SAP05 #2, WBDLb-SAP05 #1, WBDLb-SAP05 #2) and a GFP control line in *A. thaliana* Col-0 background were generously provided by Professor Saskia Hogenhout (John Innes Centre, United Kingdom). Seeds were sterilized with 70% [v/v] ethanol and 6.5% [v/v] bleach and sown on square plates (12x12 cm) containing ½ MS, 1% [w/v] agar, and 12 mg/L BASTA (DL-Phosphinothricin/Glufosinate ammonium). The plants were grown under long-day conditions (21°C, 85 µmol/m2/s) for 6 days, followed by a selection of homozygous lines.

For root phenotyping and gene expression experiments, homozygous seeds were sterilized as described above and sown on square plates (12x12 cm) containing ½ MS, 1% [w/v] sucrose, and 0.8% [w/v] agar. The plants were grown under long-day conditions for 6 days.

For the protoplast experiments, *A. thaliana* Col-0 plants were grown as described in Voichek *et al.*, 2023 ("Plant material and growth conditions").

4.2. Phenotyping of SAP05 lines

4.2.1. Root length assay

After 6 days of growth, plates with SAP05 and control lines were scanned and the root length of each germinated seedling was measured using the segmented line tool in Fiji software (Schindelin *et al.*, 2012). The root length values of each SAP05 line and the control were compared using Type III ANOVA in R ('Companion to Applied Regression [R package car version 3.1-2]', 2023).

4.2.2. Fixation and confocal imaging of the root tip

For each biological replicate, 20 seedlings from each line were transferred to 6-well plates (10 seedlings per well) and fixed with 3 mL/well 4% [w/v] paraformaldehyde in phosphate-buffered saline (PBS) solution for 1h in a vacuum chamber. The fixed seedlings were then washed twice with 3 mL/well PBS for 1 min and 3 mL staining solution was added to each well. The plates were wrapped in aluminum foil and incubated at 4°C for 3-4 days until imaging. The staining solution contains 0.1% [v/v] SCRI Renaissance 2200 (Musielak *et al.*, 2016) in ClearSee solution (Kurihara *et al.*, 2015).

Before imaging, the staining solution was removed from each well and the seedlings were washed with 3 mL/well ClearSee solution. Strips of transparent plastic tape were added to the long edges of each microscope slide, so the coverslip was slightly raised. 90 µL ClearSee solution was spread on the microscope slide and 5 seedlings were mounted on each slide. The cotyledons were removed with a scalpel blade before putting the coverslip and sealing it with clear nail polish. The microscope slides were imaged on LSM780 Axio Imager with

Plan-Apochromat 40x/1.3 Oil DIC M27 objective using 0.03% Laser Diode 405 25mW. For each seedling, the middle plane of the root (the widest part) from the root tip to the beginning of the elongation zone was imaged.

4.2.3. Quantification of root meristem cells

The meristem cells in the cortex cell file of each confocal root tip image were counted using the multi-point tool in Fiji software (Schindelin *et al.*, 2012), according to a previously published protocol (Perilli and Sabatini, 2010). The meristem cell counts of each SAP05 line and the control were compared using Type III ANOVA in R ('Companion to Applied Regression [R package car version 3.1-2]', 2023).

4.3. RNA extraction from shoots and roots

For each biological replicate, the shoots (cotyledons) and roots (~1 cm from the root tip) of 10 seedlings from each line were harvested separately in 1.5 mL Eppendorf tubes with 6-7 medium-sized glass beads, followed by immediate snap-freezing in liquid nitrogen. The tubes were stored at -70°C until RNA extraction.

Total RNA from shoots and roots was extracted using the Monarch® Total RNA Miniprep Kit (New England Biolabs, T2010S). The frozen tissues were disrupted mechanically using a bead mill and then the protocol for "Tough-to-Lyse Samples" in part 1 of the kit manual was followed by resuspending each sample with 800 µL 1X DNA/RNA Protection Reagent and proceeding to part 2.

4.4. Cloning plasmids encoding SAP05 variants and GATA TFs

For the transient transformation of SAP05, three plasmids were constructed using the GreenGate system (Lampropoulos *et al.*, 2013): (1) a plasmid encoding *35S-35S*::*SAP05-PnWBb* and P_{UBI4}::*GFP* reporter gene ("*PnWBb* plasmid"), (2) a plasmid encoding *35S-35S*::*SAP05-WBDLb* and P_{UBI4}::*GFP* reporter gene ("*WBDLb* plasmid"), and (3) a plasmid encoding only P_{UBI4}::*GFP* ("positive control"). The original plasmids encoding *SAP05-PnWBb* and *SAP05-WBDLb* were generously provided by Professor Saskia Hogenhout (John Innes Centre, United Kingdom).

In the GATA overexpression experiments, four plasmids were constructed using the GreenGate system (Lampropoulos et al., 2013). These plasmids encoded either GATA1, GATA4, GATA6 or a GFP reporter gene, all driven by a double 35S promoter. The plasmids containing the coding sequences for the three GATA transcription factors were obtained from Twist Bioscience.

4.5. Protoplast production and transformation

The protocol was adapted from Voichek *et al.*, 2023 ("MPRA assay in tomato and Arabidopsis protoplasts"). After counting, the protoplasts were resuspended in MMG

solution to a concentration of 2 million cells/mL. For each biological replicate, 500,000 protoplasts (in 250 µL) were individually transformed with 25 µg of either the *PnWBb* plasmid, *WBDLb* plasmid, or the control plasmid (defined above), each at a concentration of 1 µg/µL. Four independent samples were transformed with each plasmid. Two negative samples were each transformed with 25 µL elution buffer (QIAGEN Plasmid Plus Midi Kit, 12943). In the final step, each sample was resuspended in 250 µL W1 solution and transferred to a separate well in a 24-well plate. The samples were incubated for 21h at 25°C under constant light (85 µmol/m^2/s) before sorting.

4.6. FACS of transformed protoplasts and RNA extraction

For each biological replicate, half of the samples ("non-sorted") were harvested without prior sorting by centrifugation at 450xg for 3 minutes, followed by removal of the supernatant and snap-freezing in liquid nitrogen. These samples were stored at -70°C until RNA extraction. The other half of the samples ("sorted") were sorted with Fluorescence Activated Cell Sorter (FACS) following the protocol described in (Bargmann and Birnbaum, 2009). Protoplasts were sorted with FACS Aria III (BD Biosciences) using a 100 µm nozzle, PBS as sheath fluid, and 20 psi as sheath pressure. A 488-nm blue laser was used for GFP excitation with emission measured at 530/30 nm, while a 561-nm yellow-green laser was used for chlorophyll autofluorescence with emission measured at 610/20 nm. GFP-positive cells were sorted from *PnWBb*, *WBDLb*, and positive control samples, while GFP-negative cells were sorted from the negative sample. All samples were sorted in 500 µL Lysis buffer (Monarch® Total RNA Miniprep Kit).

4.7. Transient overexpression of GATA factors in protoplasts

The protocol described in Voichek *et al.*, 2023 ("Transient overexpression of GATA TFs in Arabidopsis protoplasts") was followed.

4.8. RNA extraction from protoplasts

RNA was extracted from all protoplast experiments using the Monarch® Total RNA Miniprep Kit. For the SAP05 experiments, RNA extraction was carried out immediately after sorting, for both sorted and non-sorted samples. The frozen pellets from the non-sorted samples were briefly thawed on ice and resuspended in 500 µL of Lysis buffer. In the GATA overexpression experiments, the frozen pellets were similarly thawed on ice and resuspended in 400 µL of Lysis buffer.

4.9. RNA sequencing and analysis of RNA-seq data

The protocols described in Voichek *et al.*, 2023 ("RNA sequencing (RNA-Seq)" and "Processing of RNA-Seq") were followed.

References

Bai, X. *et al.* (2009) 'AY-WB phytoplasma secretes a protein that targets plant cell nuclei', *Molecular plant-microbe interactions: MPMI*, 22(1), pp. 18–30. Available at: https://doi.org/10.1094/MPMI-22-1-0018.

Ballas, N., Wong, L.M. and Theologis, A. (1993) 'Identification of the auxin-responsive element, AuxRE, in the primary indoleacetic acid-inducible gene, PS-IAA4/5, of pea (Pisum sativum)', *Journal of molecular biology*, 233(4), pp. 580–596. Available at: https://doi.org/10.1006/jmbi.1993.1537.

Bargmann, B.O.R. and Birnbaum, K.D. (2009) 'Positive fluorescent selection permits precise, rapid, and in-depth overexpression analysis in plant protoplasts', *Plant physiology*, 149(3), pp. 1231–1239. Available at: https://doi.org/10.1104/pp.108.133975.

Bastakis, E. *et al.* (2018) 'LLM-Domain B-GATA Transcription Factors Play Multifaceted Roles in Controlling Greening in Arabidopsis', *The Plant cell*, 30(3), pp. 582–599. Available at: https://doi.org/10.1105/tpc.17.00947.

Behringer, C. *et al.* (2014) 'Functional diversification within the family of B-GATA transcription factors through the leucine-leucine-methionine domain', *Plant physiology*, 166(1), pp. 293–305. Available at: https://doi.org/10.1104/pp.114.246660.

Berk, A.J. (1999) 'Activation of RNA polymerase II transcription', *Current opinion in cell biology*, 11(3), pp. 330–335. Available at: https://doi.org/10.1016/S0955-0674(99)80045-3.

Bernard, V., Brunaud, V. and Lecharny, A. (2010) 'TC-motifs at the TATA-box expected position in plant genes: a novel class of motifs involved in the transcription regulation', *BMC genomics*, 11(1), p. 166. Available at: https://doi.org/10.1186/1471-2164-11-166.

Betts, H.C. *et al.* (2018) 'Integrated genomic and fossil evidence illuminates life's early evolution and eukaryote origin', *Nature ecology & evolution*, 2(10), pp. 1556–1562. Available at: https://doi.org/10.1038/s41559-018-0644-x.

Blanc-Mathieu, R. *et al.* (2024) 'Plant-TFClass: a structural classification for plant transcription factors', *Trends in plant science*, 29(1), pp. 40–51. Available at: https://doi.org/10.1016/j.tplants.2023.06.023.

Blombach, F. and Grohmann, D. (2017) 'Same same but different: The evolution of TBP in archaea and their eukaryotic offspring', *Transcription*, 8(3), pp. 162–168. Available at: https://doi.org/10.1080/21541264.2017.1289879.

Bonifer, C. (2000) 'Developmental regulation of eukaryotic gene loci: which cis-regulatory information is required?', *Trends in genetics: TIG*, 16(7), pp. 310–315. Available at: https://doi.org/10.1016/s0168-9525(00)02029-1.

Bonner, J.T. (1998) 'The origins of multicellularity', *Integrative Biology*, 1(1), pp. 27–36. Available at: https://doi.org/10.1002/(sici)1520-6602(1998)1:1<27::aid-inbi4>3.3.co;2-y.

Breathnach, R. and Chambon, P. (1981) 'Organization and expression of eucaryotic split genes coding for proteins', *Annual review of biochemistry*, 50, pp. 349–383. Available at: https://doi.org/10.1146/annurev.bi.50.070181.002025.

Burgess, D.G., Xu, J. and Freeling, M. (2015) 'Advances in understanding cis regulation of the plant gene with an emphasis on comparative genomics', *Current opinion in plant biology*, 27, pp. 141–147. Available at: https://doi.org/10.1016/j.pbi.2015.07.001.

Burley, S.K. (1996) 'The TATA box binding protein', *Current opinion in structural biology*, 6(1), pp. 69–75. Available at: https://doi.org/10.1016/s0959-440x(96)80097-2.

Choi, D. *et al.* (2014) 'iNID: an analytical framework for identifying network models for interplays among developmental signaling in Arabidopsis', *Molecular plant*, 7(5), pp. 792–813. Available at: https://doi.org/10.1093/mp/sst173.

Civán, P. and Svec, M. (2009) 'Genome-wide analysis of rice (Oryza sativa L. subsp. japonica) TATA box and Y Patch promoter elements', *Genome / National Research Council Canada = Genome / Conseil national de recherches Canada*, 52(3), pp. 294–297. Available at: https://doi.org/10.1139/G09-001.

Cocking, E.C. (1960) 'A Method for the Isolation of Plant Protoplasts and Vacuoles', *Nature*, 187(4741), pp. 962–963. Available at: https://doi.org/10.1038/187962a0.

'Companion to Applied Regression [R package car version 3.1-2]' (2023). Available at: https://CRAN.R-project.org/package=car (Accessed: 19 August 2024).

De Smet, S. *et al.* (2015) 'Gene Networks Involved in Hormonal Control of Root Development in Arabidopsis thaliana: A Framework for Studying Its Disturbance by Metal Stress', *International journal of molecular sciences*, 16(8), pp. 19195–19224. Available at: https://doi.org/10.3390/ijms160819195.

Du Toit, A. (2014) 'Bacterial pathogenicity: Phytoplasma converts plants into zombies', *Nature reviews. Microbiology*, 12(6), p. 393. Available at: https://doi.org/10.1038/nrmicro3283.

Endo, H. *et al.* (2015) 'Multiple classes of transcription factors regulate the expression of VASCULAR-RELATED NAC-DOMAIN7, a master switch of xylem vessel differentiation', *Plant & cell physiology*, 56(2), pp. 242–254. Available at: https://doi.org/10.1093/pcp/pcu134.

Fessele, S. *et al.* (2002) 'Regulatory context is a crucial part of gene function', *Trends in genetics: TIG*, 18(2), pp. 60–63. Available at: https://doi.org/10.1016/s0168-9525(02)02591-x.

Franco-Zorrilla, J.M. *et al.* (2014) 'DNA-binding specificities of plant transcription factors and their potential to define target genes', *Proceedings of the National Academy of Sciences of the United States of America*, 111(6), pp. 2367–2372. Available at: https://doi.org/10.1073/pnas.1316278111.

Gallegos, J.E. and Rose, A.B. (2017) 'Intron DNA sequences can be more important than the proximal promoter in determining the site of transcript initiation', *The plant cell*, 29(4), pp. 843–853. Available at: https://doi.org/10.1105/tpc.17.00020.

Goodall, G.J. and Filipowicz, W. (1989) 'The AU-rich sequences present in the introns of plant nuclear pre-mRNAs are required for splicing', *Cell*, 58(3), pp. 473–483. Available at: https://doi.org/10.1016/0092-8674(89)90428-5.

Hetzel, J. *et al.* (2016) 'Nascent RNA sequencing reveals distinct features in plant transcription', *Proceedings of the National Academy of Sciences of the United States of America*, 113(43), pp. 12316–12321. Available at: https://doi.org/10.1073/pnas.1603217113.

Hofmann, F., Schon, M.A. and Nodine, M.D. (2019) 'The embryonic transcriptome of Arabidopsis thaliana', *Plant reproduction*, 32(1), pp. 77–91. Available at: https://doi.org/10.1007/s00497-018-00357-2.

Hogenhout, S.A. *et al.* (2008) 'Phytoplasmas: bacteria that manipulate plants and insects', *Molecular plant pathology*, 9(4), pp. 403–423. Available at: https://doi.org/10.1111/j.1364-3703.2008.00472.x.

Huang, W. *et al.* (2021) 'Parasitic modulation of host development by ubiquitin-independent protein degradation', *Cell*, 184(20), pp. 5201–5214.e12. Available at: https://doi.org/10.1016/j.cell.2021.08.029.

Jiang, K. *et al.* (2018) 'Longitudinal patterning in roots: a GATA2-auxin interaction underlies and maintains the root transition domain', *Planta*, 247(4), pp. 831–843. Available at: https://doi.org/10.1007/s00425-017-2831-4.

Jores, T. *et al.* (2020) 'Identification of plant enhancers and their constituent elements by STARR-seq in tobacco leaves', *The plant cell*, 32(7), pp. 2120–2131. Available at: https://doi.org/10.1105/tpc.20.00155.

Klepikova, A.V. *et al.* (2016) 'A high resolution map of the Arabidopsis thaliana developmental transcriptome based on RNA-seq profiling', *The Plant journal: for cell and molecular biology*, 88(6), pp. 1058–1070. Available at: https://doi.org/10.1111/tpj.13312.

Knoll, A.H. (2011) 'The Multiple Origins of Complex Multicellularity', *Annual review of earth and planetary sciences*, 39(Volume 39, 2011), pp. 217–239. Available at: https://doi.org/10.1146/annurev.earth.031208.100209.

Kriwacki, R.W. *et al.* (1992) 'Sequence-specific recognition of DNA by zinc-finger peptides derived from the transcription factor Sp1', *Proceedings of the National Academy of Sciences of the United States of America*, 89(20), pp. 9759–9763. Available at: https://doi.org/10.1073/pnas.89.20.9759.

Kurihara, D. *et al.* (2015) 'ClearSee: a rapid optical clearing reagent for whole-plant fluorescence imaging', *Development (Cambridge, England)*, 142(23), pp. 4168–4179. Available at: https://doi.org/10.1242/dev.127613.

Lampropoulos, A. *et al.* (2013) 'GreenGate---a novel, versatile, and efficient cloning system for plant transgenesis', *PloS one*, 8(12), p. e83043. Available at: https://doi.org/10.1371/journal.pone.0083043.

Latchman, D.S. (1993) 'Transcription factors: an overview', *International journal of experimental pathology*, 74(5), pp. 417–422. Available at: https://www.ncbi.nlm.nih.gov/pubmed/8217775.

Lee, I.M., Davis, R.E. and Gundersen-Rindal, D.E. (2000) 'Phytoplasma: phytopathogenic mollicutes', *Annual review of microbiology*, 54, pp. 221–255. Available at: https://doi.org/10.1146/annurev.micro.54.1.221.

Liu, P.-P. *et al.* (2005) 'The BME3 (Blue Micropylar End 3) GATA zinc finger transcription factor is a positive regulator of Arabidopsis seed germination: GATA zinc finger in seed germination', *The Plant journal: for cell and molecular biology*, 44(6), pp. 960–971. Available at: https://doi.org/10.1111/j.1365-313X.2005.02588.x.

Liu, Q. *et al.* (2023) 'Bimodular architecture of bacterial effector SAP05 that drives ubiquitin-independent targeted protein degradation', *Proceedings of the National Academy of Sciences of the United States of America*, 120(49), p. e2310664120. Available at: https://doi.org/10.1073/pnas.2310664120.

Liu, Y. *et al.* (2019) 'GATA and Phytochrome Interacting Factor transcription factors regulate light-induced vindoline biosynthesis in Catharanthus roseus', *Plant physiology*, 180(3), pp.

1336–1350. Available at: https://doi.org/10.1104/pp.19.00489.

Love, M.I., Huber, W. and Anders, S. (2014) 'Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2', *Genome biology*, 15(12), p. 550. Available at: https://doi.org/10.1186/s13059-014-0550-8.

Lowry, J.A. and Atchley, W.R. (2000) 'Molecular evolution of the GATA family of transcription factors: conservation within the DNA-binding domain', *Journal of molecular evolution*, 50(2), pp. 103–115. Available at: https://doi.org/10.1007/s002399910012.

Luo, X.-M. *et al.* (2010) 'Integration of light- and brassinosteroid-signaling pathways by a GATA transcription factor in Arabidopsis', *Developmental cell*, 19(6), pp. 872–883. Available at: https://doi.org/10.1016/j.devcel.2010.10.023.

MacLean, A.M. *et al.* (2011) 'Phytoplasma Effector SAP54 Induces Indeterminate Leaf-Like Flower Development in Arabidopsis Plants', *Plant physiology*, 157(2), pp. 831–841. Available at: https://doi.org/10.1104/pp.111.181586.

MacLean, A.M. *et al.* (2014) 'Phytoplasma effector SAP54 hijacks plant reproduction by degrading MADS-box proteins and promotes insect colonization in a RAD23-dependent manner', *PLoS biology*, 12(4), p. e1001835. Available at: https://doi.org/10.1371/journal.pbio.1001835.

Maejima, K. *et al.* (2014) 'Recognition of floral homeotic MADS domain transcription factors by a phytoplasmal effector, phyllogen, induces phyllody', *The Plant journal: for cell and molecular biology*, 78(4), pp. 541–554. Available at: https://doi.org/10.1111/tpj.12495.

Mascarenhas, D. *et al.* (1990) 'Intron-mediated enhancement of heterologous gene expression in maize', *Plant molecular biology*, 15(6), pp. 913–920. Available at: https://doi.org/10.1007/bf00039430.

de Mendoza, A. and Sebé-Pedrós, A. (2019) 'Origin and evolution of eukaryotic transcription factors', *Current opinion in genetics & development*, 58-59, pp. 25–32. Available at: https://doi.org/10.1016/j.gde.2019.07.010.

Molina, C. and Grotewold, E. (2005) 'Genome wide analysis of Arabidopsis core promoters', *BMC genomics*, 6, p. 25. Available at: https://doi.org/10.1186/1471-2164-6-25.

Musetti, R. *et al.* (2013) 'Phytoplasma-triggered Ca(2+) influx is involved in sieve-tube blockage', *Molecular plant-microbe interactions: MPMI*, 26(4), pp. 379–386. Available at: https://doi.org/10.1094/MPMI-08-12-0207-R.

Musielak, T. *et al.* (2016) 'Use of SCRI renaissance 2200 (SR2200) as a versatile dye for imaging of developing embryos, whole ovules, pollen tubes and roots', *Bio-protocol*, 6(18). Available at: https://doi.org/10.21769/bioprotoc.1935.

Nikolov, D.B. and Burley, S.K. (1997) 'RNA polymerase II transcription initiation: a structural view', *Proceedings of the National Academy of Sciences of the United States of America*, 94(1), pp. 15–22. Available at: https://doi.org/10.1073/pnas.94.1.15.

O'Malley, R.C. *et al.* (2016) 'Cistrome and epicistrome features shape the regulatory DNA landscape', *Cell*, 165(5), pp. 1280–1292. Available at: https://doi.org/10.1016/j.cell.2016.04.038.

Park, J. *et al.* (2005) 'A Comparative Genome-Wide Analysis of GATA Transcription Factors in Fungi', *Genomics & informatics*, 4. Available at: https://www.researchgate.net/publication/241511816

(Accessed: 17 August 2024).

Perilli, S. and Sabatini, S. (2010) 'Analysis of root meristem size development', *Methods in molecular biology (Clifton, N.J.)*. Edited by L. Hennig and C. Köhler, 655, pp. 177–187. Available at: https://doi.org/10.1007/978-1-60761-765-5_12.

Petricka, J.J., Winter, C.M. and Benfey, P.N. (2012) 'Control of Arabidopsis root development', *Annual review of plant biology*, 63, pp. 563–590. Available at: https://doi.org/10.1146/annurev-arplant-042811-105501.

Ravindran, P. *et al.* (2017) 'A novel RGL2-DOF6 complex contributes to primary seed dormancy in Arabidopsis thaliana by regulating a GATA transcription factor', *Molecular plant*, 10(10), pp. 1307–1320. Available at: https://doi.org/10.1016/j.molp.2017.09.004.

Riechmann, J.L. *et al.* (2000) 'Arabidopsis transcription factors: genome-wide comparative analysis among eukaryotes', *Science*, 290(5499), pp. 2105–2110. Available at: https://doi.org/10.1126/science.290.5499.2105.

Riechmann, J.L. (2002) 'Transcriptional regulation: a genomic overview', *The Arabidopsis book / American Society of Plant Biologists*, 1, p. e0085. Available at: https://doi.org/10.1199/tab.0085.

Rose, A.B. *et al.* (2008) 'Promoter-proximal introns in Arabidopsis thaliana are enriched in dispersed signals that elevate gene expression', *The plant cell*, 20(3), pp. 543–551. Available at: https://doi.org/10.1105/tpc.107.057190.

Rose, A.B. (2019) 'Introns as gene regulators: A brick on the accelerator', *Frontiers in genetics*, 9, p. 672. Available at: https://doi.org/10.3389/fgene.2018.00672.

Rümpler, F. *et al.* (2015) 'Did Convergent Protein Evolution Enable Phytoplasmas to Generate "Zombie Plants"?', *Trends in plant science*, 20(12), pp. 798–806. Available at: https://doi.org/10.1016/j.tplants.2015.08.004.

Schindelin, J. *et al.* (2012) 'Fiji: an open-source platform for biological-image analysis', *Nature methods*, 9(7), pp. 676–682. Available at: https://doi.org/10.1038/nmeth.2019.

Schmid, M. *et al.* (2005) 'A gene expression map of Arabidopsis thaliana development', *Nature genetics*, 37(5), pp. 501–506. Available at: https://doi.org/10.1038/ng1543.

Schwechheimer, C., Schröder, P.M. and Blaby-Haas, C.E. (2022) 'Plant GATA factors: Their biology, phylogeny, and phylogenomics', *Annual review of plant biology*, 73(1), pp. 123–148. Available at: https://doi.org/10.1146/annurev-arplant-072221-092913.

Shahan, R. *et al.* (2022) 'A single-cell Arabidopsis root atlas reveals developmental trajectories in wild-type and cell identity mutants', *Developmental cell*, 57(4), pp. 543–560.e9. Available at: https://doi.org/10.1016/j.devcel.2022.01.008.

Shiu, S.-H., Shih, M.-C. and Li, W.-H. (2005) 'Transcription factor families have much higher expansion rates in plants than in animals', *Plant physiology*, 139(1), pp. 18–26. Available at: https://doi.org/10.1104/pp.105.065110.

Silver, B.D. *et al.* (2024) 'Differences in transcription initiation directionality underlie distinctions between plants and animals in chromatin modification patterns at genes and cis-regulatory elements', *G3*, 14(3). Available at: https://doi.org/10.1093/g3journal/jkae016.

Smale, S.T. and Kadonaga, J.T. (2003) 'The RNA polymerase II core promoter', *Annual review of biochemistry*, 72, pp. 449–479. Available at: https://doi.org/10.1146/annurev.biochem.72.121801.161520.

Subramaniam, R. *et al.* (2001) 'Direct visualization of protein interactions in plant cells', *Nature biotechnology*, 19(8), pp. 769–772. Available at: https://doi.org/10.1038/90831.

Sugimoto, K., Takeda, S. and Hirochika, H. (2003) 'Transcriptional activation mediated by binding of a plant GATA-type zinc finger protein AGP1 to the AG-motif (AGATCCAA) of the wound-inducible Myb gene NtMyb2', *The Plant journal: for cell and molecular biology*, 36(4), pp. 550–564. Available at: https://doi.org/10.1046/j.1365-313x.2003.01899.x.

Sugio, A. *et al.* (2011) 'Phytoplasma protein effector SAP11 enhances insect vector reproduction by manipulating plant development and defense hormone biosynthesis', *Proceedings of the National Academy of Sciences of the United States of America*, 108(48), pp. E1254–63. Available at: https://doi.org/10.1073/pnas.1105664108.

The Irpcm Phytoplasma/Spiroplasma Working Team-Phytoplasma Taxonomy Group (2004) '"Candidatus Phytoplasma", a taxon for the wall-less, non-helical prokaryotes that colonize plant phloem and insects', *International journal of systematic and evolutionary microbiology*, 54(Pt 4), pp. 1243–1255. Available at: https://doi.org/10.1099/ijs.0.02854-0.

Toufighi, K. *et al.* (2005) 'The Botany Array Resource: e-Northerns, Expression Angling, and promoter analyses', *The Plant journal: for cell and molecular biology*, 43(1), pp. 153–163. Available at: https://doi.org/10.1111/j.1365-313X.2005.02437.x.

Tu, X. *et al.* (2020) 'Reconstructing the maize leaf regulatory network using ChIP-seq data of 104 transcription factors', *Nature communications*, 11(1), p. 5089. Available at: https://doi.org/10.1038/s41467-020-18832-8.

Voichek, Y. *et al.* (2023) 'Widespread position-dependent transcriptional regulatory sequences in plants', *bioRxiv*. Available at: https://doi.org/10.1101/2023.09.15.557872.

Walter, M. *et al.* (2004) 'Visualization of protein interactions in living plant cells using bimolecular fluorescence complementation: Protein interaction in plant cells', *The Plant journal: for cell and molecular biology*, 40(3), pp. 428–438. Available at: https://doi.org/10.1111/j.1365-313X.2004.02219.x.

Weintraub, P.G. and Beanland, L. (2006) 'Insect vectors of phytoplasmas', *Annual review of entomology*, 51, pp. 91–111. Available at: https://doi.org/10.1146/annurev.ento.51.110104.151039.

Weirauch, M.T. *et al.* (2014) 'Determination and inference of eukaryotic transcription factor sequence specificity', *Cell*, 158(6), pp. 1431–1443. Available at: https://doi.org/10.1016/j.cell.2014.08.009.

Xu, M. *et al.* (2016) 'Developmental Functions of miR156-Regulated SQUAMOSA PROMOTER BINDING PROTEIN-LIKE (SPL) Genes in Arabidopsis thaliana', *PLoS genetics*, 12(8), p. e1006263. Available at: https://doi.org/10.1371/journal.pgen.1006263.

Xu, Z. *et al.* (2017) 'Genome-wide binding analysis of AtGNC and AtCGA1 demonstrates their cross-regulation and common and specific functions', *Plant direct*, 1(4), p. e00016. Available at: https://doi.org/10.1002/pld3.16.

Yamamoto, Y.Y. *et al.* (2007) 'Differentiation of core promoter architecture between plants and mammals revealed by LDSS analysis', *Nucleic acids research*, 35(18), pp. 6219–6226. Available at:

https://doi.org/10.1093/nar/gkm685.

Yamasaki, K. *et al.* (2013) 'DNA-binding domains of plant-specific transcription factors: structure, function, and evolution', *Trends in plant science*, 18(5), pp. 267–276. Available at: https://doi.org/10.1016/j.tplants.2012.09.001.

Yang, C. *et al.* (2007) 'Prevalence of the initiator over the TATA box in human and yeast genes and identification of DNA motifs enriched in human TATA-less core promoters', *Gene*, 389(1), pp. 52–65. Available at: https://doi.org/10.1016/j.gene.2006.09.029.

Zabidi, M.A. and Stark, A. (2016) 'Regulatory enhancer-core-promoter communication via transcription factors and cofactors', *Trends in genetics: TIG*, 32(12), pp. 801–814. Available at: https://doi.org/10.1016/j.tig.2016.10.003.

Zhao, Y. *et al.* (2004) 'HANABA TARANU is a GATA transcription factor that regulates shoot apical meristem and flower development in Arabidopsis', *The Plant cell*, 16(10), pp. 2586–2600. Available at: https://doi.org/10.1105/tpc.104.024869.