**Microbial Matchmakers: Using Machine Learning to Identify Plant Transcription Factor Targets Involved in Host Microbiome Assembly**

Report



Schematic drawing of rhizophere

Ids Hartman  
UvAnetID: 15712745

Amsterdam, 27 June 2025

**Microbial Matchmakers: Using Machine Learning to Identify Plant Transcription Factor Targets Involved in Host Microbiome Assembly**

Report

Ids Hartman

UvAnetID: 15712745

Amsterdam, 27 June 2025

University of Amsterdam  
Green Life Sciences

Extern assessor UvA: Anna Heintz

University of Amsterdam

Biosystems Data Analysis  
Plant Hormone Biology

Amsterdam Science Park 904, 1098 XH Amsterdam  
Supervisor: Aalt-Jan van Dijk

Supervisor: Kathrin Wippel

Signatures with date supervisors:

# Abstract

Developing sustainable agricultural practices requires an understanding of how plants regulate the composition of their root microbiomes. Through a combination of promoter motif discovery, motif enrichment analysis, and gene regulatory network (GRN) inference, this study sought to identify transcription factors (TFs) linked in plant responses to microbial communities. Transcriptome data from *Lotus japonicus* (Lj) and *Arabidopsis thaliana* (At) roots inoculated with synthetic microbial communities (SynComs) were used to identify microbiome-responsive gene clusters and analyze the promoter regions upstream of the genes within them. Using randomized promoter controls, motif enrichment analysis revealed three TFs (bHLH78, NAC010, and NAC013) as potential regulators of microbiome-induced gene expression. GRNBoost2, a machine learning-based technique, was used to infer regulatory networks on a general dataset with all conditions (global network) and a subset with synthetic community treatments (SC-specific network). Comparative research revealed that more generalist TFs dominated the global network, whereas context-specific TFs including WRKY26, MYB51, and NAC103 appeared as important regulators in the SC-specific network. The SC-specific network has greater structural robustness and modularity, according to perturbation analysis. The process is broadly applicable to non-model organisms, despite the fact that the Lj study was constrained by a lack of species-specific genetic data. Overall, the results point to a reconfiguration of transcriptional regulation by context-dependent transcription factors (TFs) during microbiome construction, moving from centralized to dispersed control. This framework lays the groundwork for future research into regulatory mechanisms of plant-microbe interactions and opens new avenues for microbiome-informed crop engineering.

# List of abbreviations used:

**AGL** – AGAMOUS-LIKE (a MADS-box transcription factor family)

**At** – *Arabidopsis thaliana*

**bHLH** – Basic Helix-Loop-Helix (transcription factor family)

**bHLH78** – Specific bHLH transcription factor identified in this study

**CAMTA** – Calmodulin-binding Transcription Activator

**Cytoscape** – Software platform for visualizing molecular interaction networks

**DAG1** – DOF AFFECTING GERMINATION 1

**DASK** – Flexible parallel computing library for distributed analytics

**DEG** – Differentially Expressed Gene

**DOF** – DNA-binding with One Finger (transcription factor family)

**ERF** – Ethylene Responsive Factor

**FIMO** – Find Individual Motif Occurrences (motif scanning tool from MEME Suite)

**FRS4** – FAR-RED ELONGATED HYPOCOTYL3-RELATED SEQUENCE 4

**GO** – Gene Ontology

**GRN** – Gene Regulatory Network

**GRNBoost2** – Gene Regulatory Network inference using Gradient Boosted Trees (version 2)

**ID** – Identifier

**Lj** – *Lotus japonicus*

**LSD1** – LESION SIMULATING DISEASE 1

**MAF1** – MADS AFFECTING FLOWERING 1

**MEME** – Multiple EM for Motif Elicitation (suite of motif tools)

**MYB** – Myeloblastosis (transcription factor family)

**NAC** – NAM, ATAF1/2, and CUC2 (transcription factor family)

**NAC010 / NAC013 / NAC103** – Specific NAC transcription factors identified in this study

**PDF2** – PROTODERMAL FACTOR 2

**Python** – Programming language used for scripting analyses

**SC** – Synthetic Community

**STREME** – Sensitive, Thorough, Rapid, Enriched Motif Elicitation (de novo motif discovery tool)

**TF** – Transcription Factor

**TFBS** – Transcription Factor Binding Site

**TSV** – Tab-Separated Values

**WRKY** – Transcription factor family named after the WRKY domain

# Table of contents

[**1. Introduction 1**](#_m88vtufoporl)

[**2. Materials and Methods 3**](#_3s0f1yjfj3b7)

[2.1 Gene Selection 3](#_7pzo0squu5fy)

[2.2 Promoter Sequence Extraction 3](#_fuz6293hvchu)

[2.3 Motif Discovery 3](#_53mb1z57s0ya)

[2.4 Motif Validation and Enrichment Analysis 3](#_676a5rh217yg)

[2.4.1 Gene Ontology (GO) enrichment analysis 3](#_m3fhgujmr6jg)

[2.4.2 Motif Enrichment Analysis 4](#_52nsxtayb2hl)

[2.4.3 Motif Shuffled Control 4](#_uelt5amzcrf1)

[2.5 Gene Annotation 5](#_zetyag54yh8n)

[2.6 Gene Regulatory Network Inference 5](#_ujjk9qcr8k9d)

[2.7 Network Visualization and Microbiome Linkage 5](#_qae5nsmmhjgp)

[2.8 Statistical Validation and Robustness 6](#_5lzj1x1g4v8h)

[**3. Results 7**](#_fjh5xn6bdm0v)

[3.1 Gene Selection Yields Candidate Microbiome-Responsive Transcription Factors 7](#_y2e722sc3oaa)

[3.2 Discovery of Candidate TF Binding in Promoters 8](#_k2xlzcmh9mm8)

[3.2.1 Arabidopsis Cluster 3 FIMO Results 8](#_qboqw3lzd0au)

[3.2.2 Lotus Cluster 6 FIMO Results 9](#_m6d6fjxyejki)

[3.2.3 De Novo Motif Discovery Using STREME 10](#_x97gbn2idnbg)

[3.3 Motif Enrichment Analysis Validates Biological Signal 10](#_9k5nuny48xzv)

[3.4 Shuffled Promoter Control Supports Motif Specificity 11](#_xx7zqrlhj11s)

[3.5 Construction of the Gene Regulatory Network (GRN) 13](#_jrv006ppibrk)

[3.6 Robustness Testing with Perturbation Analysis Validates Network Architecture 18](#_mpgelqz1oecr)

[**4. Discussion 21**](#_5jv2wlvboly)

[4.1 Microbiome-Responsive Regulatory Elements 21](#_8flt6tos1zyl)

[4.2 Transcriptional Rewiring Under SC Conditions 22](#_b785m5z6pdlo)

[4.3 Network Stability Reveals Different Architectures 22](#_l221s9yjmykw)

[4.4 The Role of Motif Enrichment: Useful, but Limited 23](#_lepgy6im9eh3)

[4.5 Arabidopsis vs. Lotus: The Limits of Cross-Species Tools 23](#_d39ljezbdvr4)

[4.6 Answering the Big Question: Who’s in Charge of Microbiome-Driven Expression? 24](#_nmmebbafb5rt)

[4.7 Future Directions 24](#_dtaxa5ff8bs4)

[**5. Conclusion 25**](#_wa1v4cdbuob)

[**7. Bibliography 26**](#_o688xs5pfwik)

# 1. Introduction

The increase in global population has led to an increase in global food needs, which has put pressure on the agricultural industry to produce more food [1,2]. The use of fertilizers, especially those based on nitrogen, is one way to achieve this [2,3]. However, there are financial and environmental consequences to using synthetic fertilizers excessively [1,2]. Researchers are increasingly moving towards the investigation of the use of plant-associated microbiomes that naturally enhance growth, nutrient uptake and stress resilience to move towards more sustainable agriculture [4,5]. The root microbiomes of plants, which are composed of several microbial communities that dynamically assemble in response to host and environmental cues, are among the most complex and intriguing biological systems. Plants attract and form distinct root microbiomes through species-specific mechanisms that are influenced by host genetics and environmental factors. This work combines two opposing model organisms, Lotus japonicus (Lj) and Arabidopsis thaliana (At), to better understand how various plant species control these microbial relationships. The non-legume At can selectively engage its root microbiota to actively modify it, while the legume Lj can also form nitrogen-fixing symbiosis with rhizobia. Plant transcription factors (TFs), which regulate gene expression in response to microbial signals, are known to play a role in the transcriptional regulation of the assembly of these microbiomes [3, 5–8]. Even though conventional methods like ChIP-seq offer concrete proof of TF-DNA binding, they are frequently restricted to particular tissues or experimental settings and could miss interactions that take place in settings linked to the microbiome or under stress. Furthermore, they usually require a significant amount of experimental work and have low throughput [9]. This study proposes an integrative computational framework that uses machine learning to identify important TFs and and the genes they regulate in the context of microbiome assembly, in order to overcome these limitations. Specifically, gene regulatory networks (GRNs) can be reconstructed from expression data by using the tree-based regression algorithm GRNBoost2 [10]. Numerous biological systems, such as microbial communities and general plant transcriptional networks, demonstrate the successful application of GRNBoost2 [11,12]. In the context of plant-microbiome interactions, prior research has shown that some regulators, such PHR1, influence the root microbiota by combining immune responses and nutritional stress in Arabidopsis thaliana experiments employing synthetic communities (SCs) and ChIP-seq [13]. More generally, recent studies emphasize that genes related to stress signaling pathways, uptake of nutrients, and root exudation affect the composition of rhizosphere communities through transcriptional control. Nonetheless, genome-wide regulatory inference techniques are still rare in this regard [14]. This work fills that gap by using GRNBoost2 to analyze regulatory circuits unique to the microbiota in At and Lj. It can be used to reveal complex transcriptional networks that underlie host-microbe specificity, in part because it captures non-linear relationships between TFs and their target genes. This means the method can detect cases where TF effects depend on combinations of signals or context-specific conditions, features that are common in plant responses to microbial communities [12,15]. The computational pipeline combines motif discovery (using FIMO for known TF binding sites from the JASPAR database [16] and STREME for de novo motifs [17]) and machine learning (GRNBoost2 [10]). Although these tools have been extensively employed in both regulatory genomics and plant biology [18,19], their combined use in microbiome-responsive transcriptional regulation in At and Lj is a novel contribution. Additionally, the predictions are strengthened by statistical validation techniques including motif enrichment testing [16,17] and empirical control using randomized sequences [20]. This project aims to investigate which transcription factors are differentially expressed under microbiome-specific conditions, if these TFs are associated with regulatory motifs in the promoters of microbiome-responsive genes, and if is possible to reconstruct biologically meaningful regulatory networks linking TFs to gene expression patterns involved in microbiome assembly. In order to get insight on the transcriptional control of microbial recruitment, this study uses this process to identify putative TFs, describe their binding locations, and reconstruct GRNs from RNA-seq data from *A. thaliana* and L. *japonicus*. This methodology not only advances our knowledge of host genetic regulation in response to microbial root colonization, but also offers a scalable method for finding candidate genes for sustainable agriculture and microbiome engineering.

# 2. Materials and Methods

All necessary information for the project and associated experiments is detailed below. For a flowchart of the full pipeline, see Figure S1 (supplementary data).

### 2.1 Gene Selection

From the RNA-seq dataset released by Wippel et al. (2021) [5], which examined plant responses to host-derived synthetic microbial communities (SC), gene expression data for At and Lj were retrieved. In the original study, co-expression clusters were generated using Weighted Gene Co-expression Network Analysis (WGCNA) based on normalized transcript counts across all treatment conditions [5]. This method groups genes into modules according to their expression similarity, enabling the identification of gene sets that may share regulatory mechanisms or biological functions. Using annotations from PlantTFDB[[1]](#footnote-1) (Version 5.0), TFs within these clusters were focussed on more in later downstream analysis.

### 2.2 Promoter Sequence Extraction

Promoter sequences (1000 base pairs upstream of transcription start sites) for the selected At genes and Athomologs of the Ljgenes were extracted from the TAIR database[[2]](#footnote-2) (last modified 2023) genome annotation file (see Microbial-Matchmakers GitHub page[[3]](#footnote-3), Extract\_upstream\_promoter\_sequences.py). The following motif analyses used these sequences as the regulatory regions.

### 2.3 Motif Discovery

To identify transcription factor binding sites, both known and de novo motif discovery approaches were employed. Using plant motif databases from JASPAR 2024[[4]](#footnote-4), known motifs were searched using FIMO (Find Individual Motif Occurrences [16]), and novel motifs were found using STREME (Sensitive, Thorough, Rapid, Enriched Motif Elucidation) [17], both of which are components of the MEME Suite (Version 5.5.8). The input consisted of promoter sequences from section 2.2, and motifs were sorted based on statistical significance.

### 2.4 Motif Validation and Enrichment Analysis

Several methods were used for statistical validation and enrichment analysis in order to assess the biological importance of the discovered transcription factor binding motifs.

#### 2.4.1 Gene Ontology (GO) enrichment analysis

To evaluate the functional roles of the genes with enriched motifs, gene ontology (GO) enrichment analysis was performed. A statistical overrepresentation test was done using the GO biological process category after genes from the microbiome-responsive clusters (cluster 3 in Atand cluster 6 in Lj) were uploaded to the PANTHER classification system[[5]](#footnote-5) (version 14.0).

#### 2.4.2 Motif Enrichment Analysis

To test for motif overrepresentation, motif frequencies in the foreground gene set (genes from the microbiome-responsive cluster 3) were compared against a custom expression-matched background (Microbial-Matchmakers GitHub, Background\_Arabidopsis\_vs\_Lotus.py) [21]. The background was composed of At genes that were not differentially expressed between At-SC and Lj-SC treatments (log₂FC < 0.5 and adjusted *p* > 0.1), ensuring that they were not responsive to microbiome variation.

To ensure a fair comparison, average expression values under At-SC conditions were calculated for each gene by averaging replicate expression measurements. For each gene in the microbiome-responsive (foreground) set, a gene from the background pool with a similar expression level was selected using the NearestNeighbors algorithm from the scikit-learn Python library[[6]](#footnote-6). This method finds the closest match based on Euclidean distance, which in this case measured the difference in average expression values between genes. By matching in this way, both foreground and background sets had comparable expression under At-SC conditions, allowing motif enrichment to be tested without confounding effects from expression level differences [21].

After background gene selection, promoter sequences (1 kb upstream of the transcription start site) were extracted for both sets. Motif occurrences were identified using FIMO [16], based on known plant TF motifs from the JASPAR database. Enrichment was tested by comparing motif counts in the foreground and background sets using Fisher’s exact test [22] (Microbial-Matchmakers GitHub, Motif\_distribution\_visualization.py) The Benjamini–Hochberg false discovery rate (FDR) method was used to adjust the obtained p-values for multiple testing [23]. The modified values that resulted are shown as q-values. Across all evaluated motifs, these q-values show the statistical significance of motif enrichment after adjusting for false positives.

For clarity, three types of statistical values are reported in this study: **(1)** FIMO match p-values, the raw probability of a sequence matching a given motif, calculated by FIMO based on the motif’s position weight matrix (PWM) [16]. **(2)** Shuffled control empirical p-values, the probability of observing a given motif score in real sequences compared to scores from 100 randomly shuffled versions of the same promoters [17, 24]. **(3)** Enrichment q-values, adjusted *p*-values from Fisher’s exact test comparing motif counts between foreground and expression-matched background sequences [22].

#### 2.4.3 Motif Shuffled Control

A shuffled control study was carried out to see if the observed motif enrichment was caused by significant biological characteristics as opposed to random nucleotide composition. In order to maintain nucleotide composition while upsetting physiologically significant sequence trends, promoter sequences from the foreground gene set were randomly shuffled across 100 repeats [24]. Motif detection was re-run on each shuffled replicate using FIMO (Microbial-Matchmakers GitHub, Shuffled\_control\_At\_100\_times.py).

In FIMO, each motif occurrence is assigned a log-likelihood ratio score, which quantifies how closely a given sequence segment matches the motif’s position weight matrix (PWM) [16, 24]. Higher scores indicate a better match to the expected motif pattern. The scores from real promoter sequences were compared to the distribution of scores from the shuffled replicates to calculate empirical *p*-values, providing additional confidence that the enriched motifs reflect true biological signal rather than random sequence features [24].

### 2.5 Gene Annotation

Using genomic annotation data (in GFF3 format from TAIR), motif occurrences were linked to the genes that corresponded to them (Microbial-Matchmakers GitHub, Gene\_annotation.py). This made it possible to integrate chromosomal location and orientation with promoter motif data, supporting functional interpretation in downstream network analysis (Microbial-Matchmakers GitHub, Gene\_annotation.py).

### 2.6 Gene Regulatory Network Inference

To predict gene regulatory interactions among TFs and their presumed target genes, the GRNBoost2 algorithm was applied from the Arboreto package [10], a tree-based regression method suitable for high-dimensional transcriptomic datasets. Wippel et al.’s (2021) bulk RNA-seq experiment’s adjusted TPM values were used to produce expression data for At and Lj [5]. Genes that had non-zero variation (some degree of variation) in every sample were the only ones kept. GRNs were built in two sets: **(1)** A global GRN that captures larger transcriptional regulatory patterns across species and cell types using the entire expression matrix containing all 16 samples (8 from Lj and 8 from At) (Microbial-Matchmakers GitHub, GRNBoost2\_global\_GRN\_At.py). Due to its massive number of nodes and edges, the resulting network was filtered to keep high-confidence edges (importance score > 2.0). **(2)** To enable a targeted investigation of species-specific regulatory networks, a GRN was constructed using only the 8 root transcriptome samples taken from plants inoculated with their respective SCs: four from At treated with the *Arabidopsis thaliana* synthetic community (At-SC) and four from Lj treated with the *Lotus japonicus* synthetic community (Lj-SC) (Microbial-Matchmakers GitHub, GRNBoost\_AtSC.py). This Synthetic Community-specific GRN (SC-specific GRN) focuses on transcriptional responses that are specific to microbiome colonization in each host species. The PlantTFDB and JASPAR motif databases were used to construct TF identifiers, which were then cross-checked for presence and variance against the expression matrix. The number of distinct target genes found for each GRN was used to establish the top-ranking transcriptional regulators. The data from the motif enrichment study (section 2.4) was cross-referenced in the GRN output to support GRN prediction. This integrative strategy was aimed at identifying candidate master regulators characterized by both high motif enrichment and regulatory influence.

### 2.7 Network Visualization and Microbiome Linkage

TFs with enriched binding motifs and differential expression in response to At-SC and Lj-SC treatments were highlighted in the visualization of the generated GRN using Cytoscape[[7]](#footnote-7) (version 3.10.3). To find possible regulatory hubs, network features including degree and betweenness centrality were calculated using Cytoscape’s NetworkAnalyzer tool, which provides standard topological metrics for directed and undirected networks.

### 2.8 Statistical Validation and Robustness

To evaluate the structural robustness and modular organization of the inferred SC-specific GRN, and global GRN, a targeted perturbation analysis was done (Microbial-Matchmakers GitHub, perturbation\_analysis\_part\_1.py, perturbation\_part\_2\_visualization\_plot.py). This involved simulating the removal of individual transcription factors (TFs) and quantifying the effect on global network connectivity [25]. Using the global GRN and the SC-specific GRN inferred by GRNBoost2, all TFs were ranked by their node degree (the number of outgoing regulatory edges). This analysis was performed on a filtered version of the network that retained only high-confidence regulatory edges (importance score > 2.0). The top 5–20 most connected TFs were further examined after 100 TFs with the highest node degree were chosen as a subset. For every TF, a unique replica of the network was constructed, eliminating all of its outgoing edges. The number of weakly connected components and the size of the largest connected subgraph were calculated using NetworkX (version 3.5). With these metrics it was possible to assess how the network fragments upon removal of individual regulators [25]. A sharp increase in the number of components was interpreted as evidence of the TF's role as a structural connector between modules. This approach tested the network’s sensitivity to disruptions and helped identify TFs that serve as essential hubs for maintaining regulatory cohesion. This strengthens confidence in the inferred TF–target relationships as meaningful contributors to microbiome-associated transcriptional regulation.

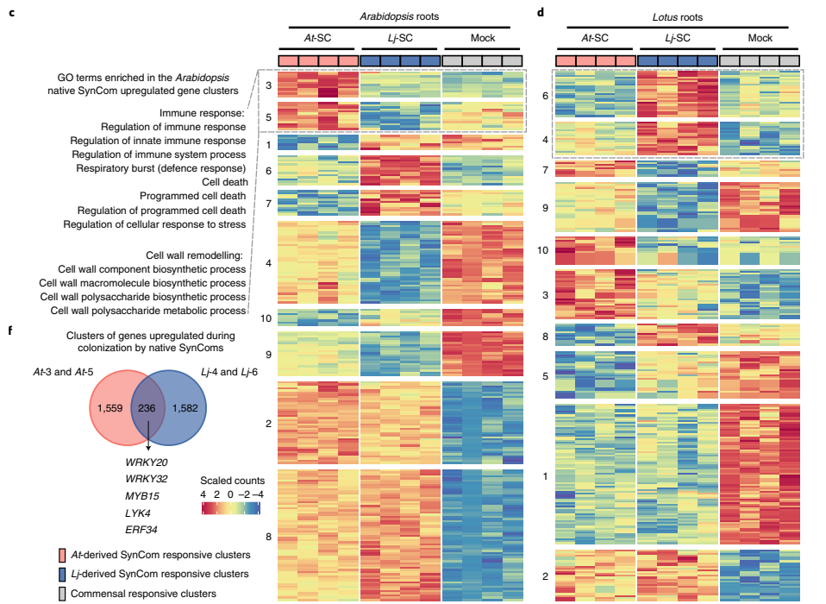
# 3. Results

The aim of this project was to address the gaps in knowledge on host genetic regulation in microbiome assembly, involving the development of a computational pipeline to identify TFs and their regulatory targets involved in plant microbiome recruitment, focusing on the model species Atand Lj. The objective was to integrate motif discovery, GO enrichment analysis and gene regulatory network construction to identify central regulators and evaluate their biological relevance through statistical controls and robustness testing.

Due to limited species-specific genomic resources and the defined timeframe of the project, the analysis of Lj relied on mapping to At homologs and promoter sequences. As a result, findings for Lj should be interpreted with caution, and primarily serve as a comparative reference rather than a complete species-specific analysis.

### 3.1 Gene Selection Yields Candidate Microbiome-Responsive Transcription Factors

To find TFs that respond to stimuli derived from the microbiome and may be involved in host-specific control of root gene expression in Lj and At, gene expression clusters, genes with similar expression patterns for At and Lj were examined using RNA-seq data from Wippel et al. (2021) (Figure 1). These clusters were generated based on co-expression across different microbial treatments, allowing identification of genes that are transcriptionally coordinated in response to synthetic microbial communities [5]. Cluster 3 in Atdemonstrated clear transcriptional activation under the At-SC and suppression under the Lj-SC. Cluster 6 in Ljshowed a mirrored pattern, being suppressed under At-SC and activated under its native Lj-SC. For further analysis, every gene from cluster 3 (At) and cluster 6 (Lj) was kept. These expression patterns indicate a host-specific transcriptional response to microbiome composition [5]. Genes, including TFs, that regulate host-specific microbial interactions are probably present in the clusters. Downstream promoter motif analysis and regulatory network construction were based on the genes in these clusters. They are strong candidates for identifying the transcriptional processes underlying microbiome assembly because of their microbiome-responsive expression patterns.



**Figure 1: Transcriptional responses specific to SynComs in roots of *Arabidopsis thaliana and Lotus japonicus*, from Wippel et al (2021) with permission.**

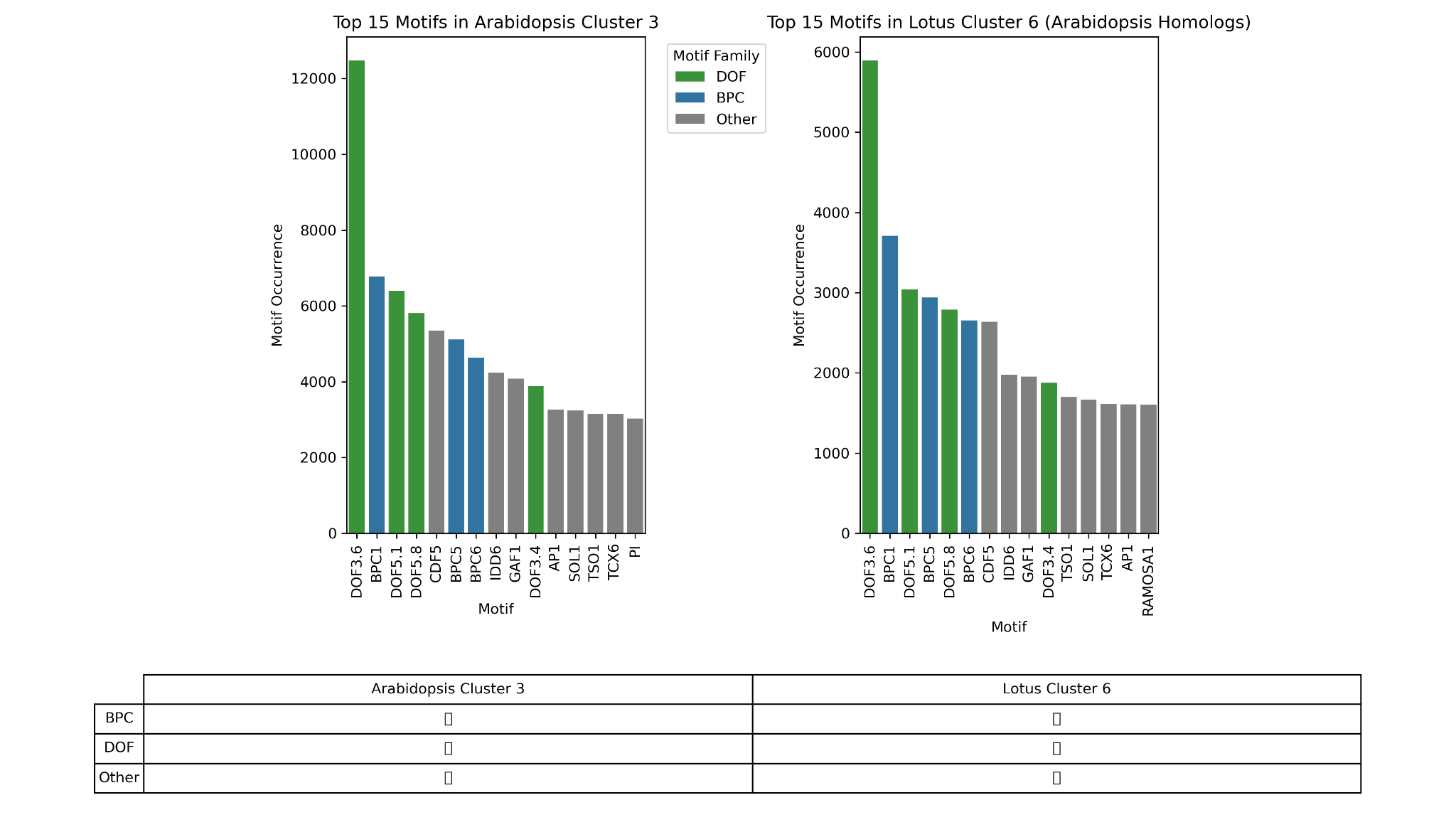
c, d. Heatmaps displaying normalized expression levels of differentially expressed genes grouped by k-means clustering for At (c) and Lj (d), each based on 12 biologically independent samples, following inoculation with host-specific synthetic microbial communities (SynComs). For Lj, a nitrogen-fixing strain from the Lj-SPHERE collection was included in all treatments to minimize the influence of nitrogen starvation or symbiosis on gene expression. f. Shared homologous genes found in corresponding clusters between the two plant species, along with key transcription factors potentially involved in regulating the distinct transcriptional responses.

### 3.2 Discovery of Candidate TF Binding in Promoters

To uncover candidate regulators of microbiome-related gene expression, potential TF binding motifs were identified in the promoter regions of genes responsive to synthetic microbial communities. Promoter sequences (1000 bp upstream) from TAIR were obtained for all genes in At cluster 3 and for the At homologs of Lj cluster 6 genes. These were used as input for motif discovery via two tools from the MEME suite: FIMO, for known motif scanning [16] using the JASPAR 2024 plant database, and STREME, for de novo motif identification [17].

#### 3.2.1 *Arabidopsis* Cluster 3 FIMO Results

In At, FIMO analysis revealed strong enrichment of motifs associated with the DOF (DNA-binding with One Finger) and BPC (BASIC PENTACYSTEINE) transcription factor families, particularly DOF3.6 and BPC1/5/6. These motifs occurred with high frequency in cluster 3 promoters and were supported by highly significant q-values (Benjamini–Hochberg adjusted p-values from Fisher’s exact test; q < 1e−13), based on comparison with an expression-matched background set. Figure 2 summarizes the 15 most frequent motifs detected in cluster 3 promoter regions, highlighting the dominance of DOF and BPC family motifs. Additional support for these motifs was provided by shuffled sequence controls, which showed that real motif scores significantly exceeded those expected from random promoter compositions (empirical p < 0.01 for top motifs).



**Figure 2: Comparative analysis of enriched TF binding motifs in *Arabidopsis* Cluster 3 and *Lotus* Cluster 6 (via *Arabidopsis* homologs).**

Barplots show the top 15 most frequent transcription factor (TF) motifs identified by FIMO in 1 kb upstream promoter regions. The left panel shows At cluster 3, and the right panel shows the corresponding At homologs of Lj cluster 6 genes. Bars are color-coded by TF family (e.g., BPC, DOF, Other) to highlight shared regulatory signatures.

#### 3.2.2 *Lotus* Cluster 6 FIMO Results

FIMO was performed on cluster 6 At homolog promoter sequences from Lj. It's interesting to note that a similar motif pattern surfaced, with DOF and BPC family patterns once more taking center stage (Figure 2). It is important to remember that these findings are based on promoter sequences from At rather than actual Lj genomic areas. These motifs might not represent regulatory activity in the native Lj environment due to functional differences between homologs.

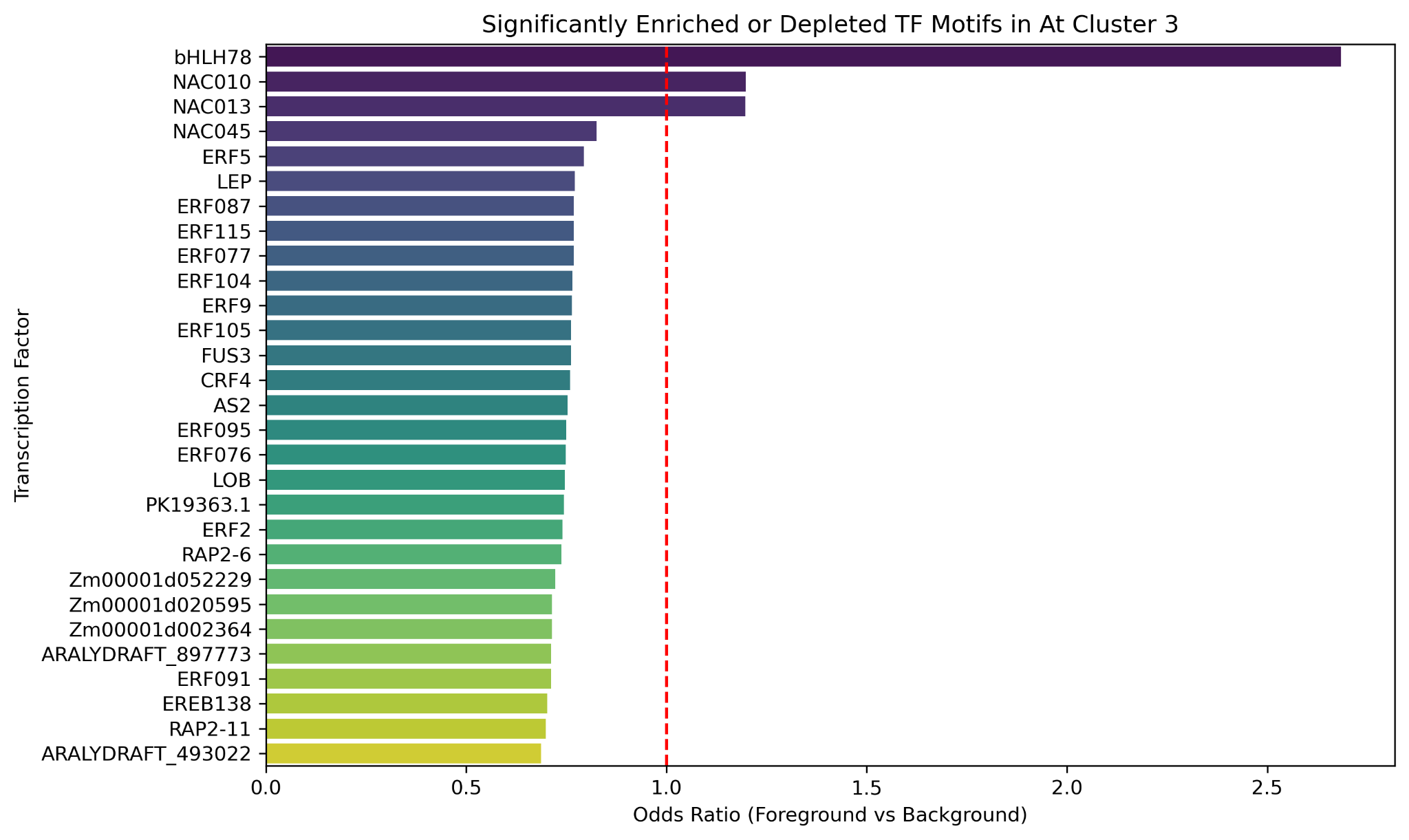
These TFs may control genes involved in transcriptional programs particular to the microbiome, as suggested by the significant occurrence of DOF and BPC motifs in At. Given the possibility of functional divergence of homologous genes across species [26], these findings should be interpreted with caution, even though the recurrence of comparable motifs in At homologs of Lj genes suggests conserved regulatory logic. These motif profiles serve as a reference for the motif enrichment research and regulatory development. The data in At offers candidates for TFs associated with the microbiome. The findings in the Lj example are preliminary and point to directions for further research, ideally utilizing native promoter sequences and motif databases unique to Lj.

#### 3.2.3 De Novo Motif Discovery Using STREME

To find possibly novel regulatory elements in the promoters of microbiome-responsive genes, de novo motif discovery was carried out using STREME in addition to the known motif scanning carried out with FIMO [16,17]. At cluster 3 promoters and the At homolog promoters of Lj cluster 6 genes both had several statistically overrepresented sequence motifs, according to STREME (Table I, in supplementary data). These motifs were found without using existing motif databases and may represent unknown transcription factor binding sites relevant to microbiome-driven transcriptional regulation [17]. Although potential motifs with significant sequence enrichment were found in the STREME output, they were not able to be mapped yet to specific transcription factors or known functional components. This restricts their ability to be incorporated into the more comprehensive regulatory framework developed in this study. Furthermore, the analysis of Lj was conducted using homologous promoter sequences from At, which might not be a true representation of the native regulatory landscape in the Lj genome (Table II, in supplementary data). The motifs identified by STREME were not included in downstream enrichment and network analyses, as they could not yet be linked to known transcription factors or functional annotations.

### 3.3 Motif Enrichment Analysis Validates Biological Signal

To distinguish biologically meaningful regulatory motifs from patterns that commonly occur in promoter sequences, a statistical motif‑enrichment analysis was performed [22]. The goal was to determine whether specific TF binding motifs are over‑represented in genes that respond to microbiome treatment, namely those in At cluster 3. Promoter regions of cluster 3 genes were scanned with FIMO to locate known TF motifs. A suitable background set was required for comparison [22]. Rather than using random sequences or genome‑wide promoters (which could bias the analysis by introducing differences in baseline expression), an expression‑matched background was compiled: genes exhibiting similar baseline expression under At‑SC treatment but showing no significant differential expression between At‑SC and Lj‑SC conditions [5,22]. This background controls for general promoter structure and expression level, thereby isolating motifs specifically associated with microbiome‑responsive regulation. Motif frequencies in the foreground (cluster 3) and background sets were compared using Fisher’s exact test (with FDR correction) (Figure 3). Any motif identified as enriched through this approach is therefore more likely to represent a genuine regulatory signal linked to microbiome response rather than a feature of highly expressed genes in general [22]. The results showed that only three motifs (bHLH78, NAC010, and NAC013) were significantly enriched in the foreground (adjusted p < 0.05). Interestingly, none of these three motifs appeared in the raw FIMO output's top 15 most frequently found motifs (Figure 2), indicating that enrichment was not predicted just by frequency. This finding emphasizes how crucial it is to assess motif relevance using expression-matched background sets. Although they may be found often across promoters, very frequent motifs (such as those from the BPC and DOF families) are not always overrepresented in genes that respond to the microbiome. Despite being less common overall, the enhanced NAC and bHLH motifs seem to be specifically linked to the set of foreground genes and could be more functionally specific regulators. Priority was given to these enriched motifs for the following validation steps, such as integration into empirical testing using shuffled promoters and gene regulatory network analysis. They are a more targeted category of potential regulators that are probably directly involved in transcriptional reactions to the makeup of the microbiome.

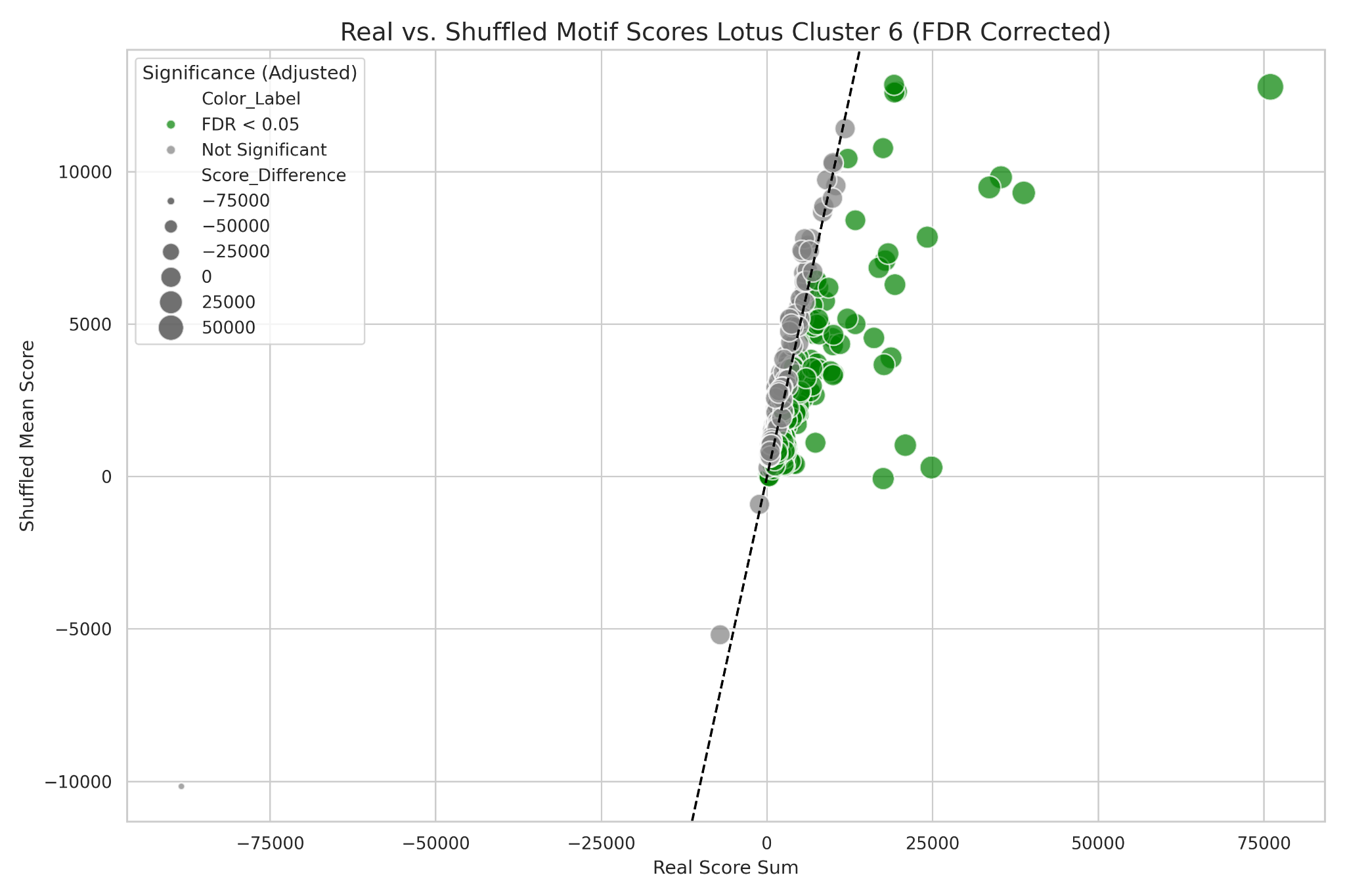


**Figure 3: Significantly enriched or depleted transcription factor (TF) motifs in *Arabidopsis* Cluster 3 promoters.**

Priority was given to these enriched motifs for subsequent validation procedures, such as integration into gene regulatory network analysis and empirical testing using shuffled promoters. They are a more targeted category of potential regulators that are probably directly involved in transcriptional reactions to the makeup of the microbiome. The foreground set (cluster 3 promoters) is compared to an expression-matched background in a bar plot that displays the odds ratios from Fisher's exact test for the presence of specific TF motifs (identified by JASPAR motif IDs). No enrichment is shown by the red dashed line at odds ratio = 1.0. The only motifs displayed are those that have a statistically significant difference (FDR-adjusted p < 0.05). In cluster 3, motifs with odds ratios higher than 1.0 are enriched, whilst those with ratios lower than 1.0 are depleted.

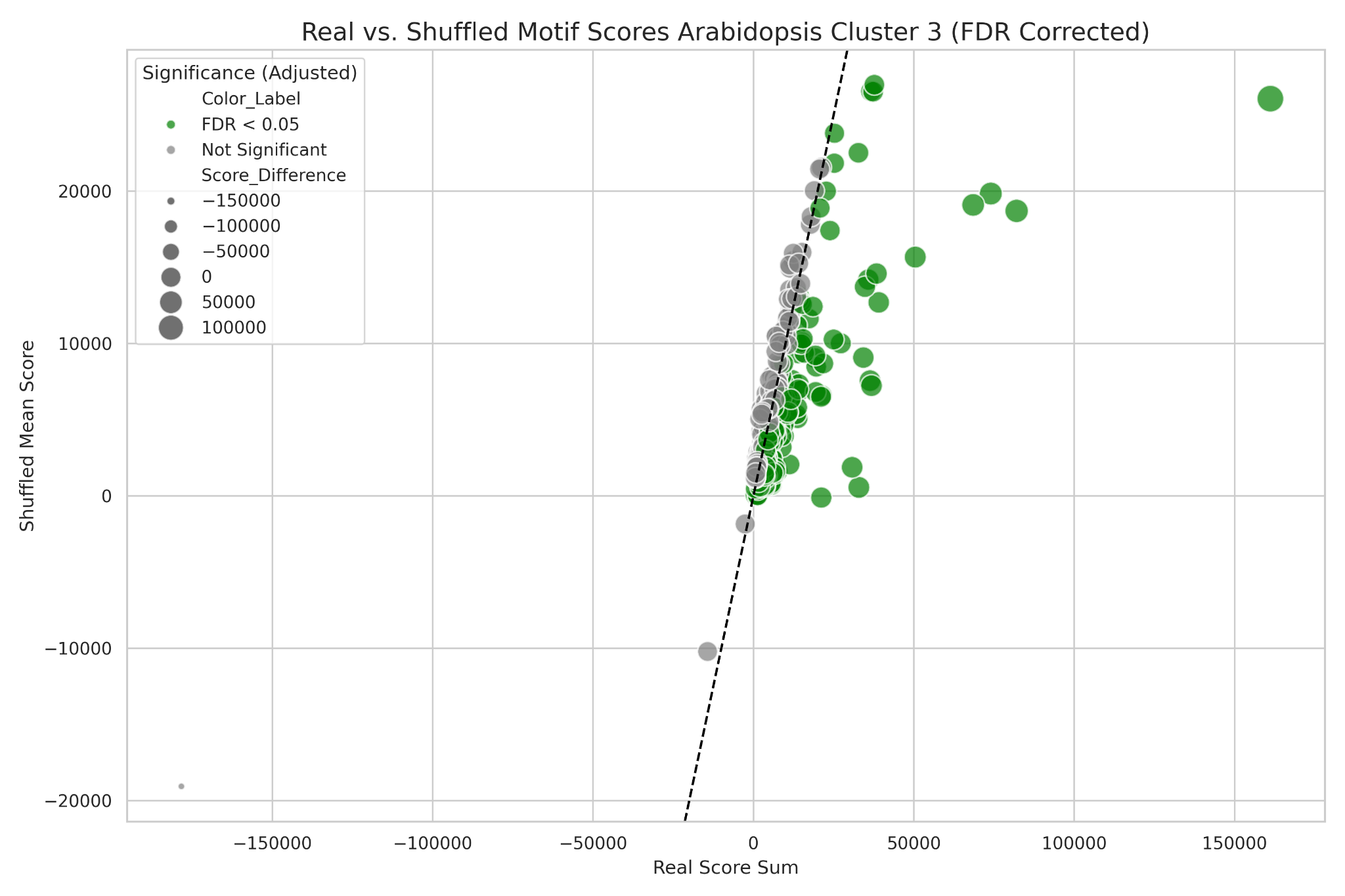
### 3.4 Shuffled Promoter Control Supports Motif Specificity

An extra validation step was carried out to assess the robustness of the results obtained from the enrichment analysis detailed in Section 3.3, which employed expression-matched background promoters to identify motifs notably overrepresented in microbiome-responsive genes. To evaluate whether the identified motif enrichments were driven by true biological sequence features rather than by general nucleotide composition or promoter architecture [24]. To do this, FIMO was used to compare motif scores from the actual cluster 3 promoters to those from 100 shuffled versions of the same sequences. The original nucleotide content was preserved in these shuffled sequences, but biologically significant sequence patterns were broken [24]. By contrasting the distribution of scores derived from the shuffled datasets with the total motif score in genuine promoters, an empirical p-value was determined for each motif. The Benjamini–Hochberg method was applied to control for multiple testing across all tested motifs [23]. Motifs with an FDR-adjusted p-value < 0.05 and a positive score difference (real > shuffled) were considered significantly enriched. These results provide an orthogonal confirmation of the enrichment analysis presented in Figure 3, by showing that the same motifs are unlikely to appear at similar levels by chance alone. This contrast is shown for the homolog promoters of At cluster 3 and the Lj cluster 6 in figures 4 and 5, respectively. The functional importance of important motifs like bHLH78, NAC010, and NAC013 was further supported by the fact that they consistently displayed higher scores in genuine promoters in both figures. Although figures 4 and 5 appear visually similar, this is expected due to the use of At-derived sequences in both cases and the overall similarity of promoter architecture [26]. Nevertheless, the biological context of the two sets is distinct, and the statistical enrichment was evaluated independently for each. The concept that some motifs are conserved regulators of microbiome-responsive transcription across species is supported by the similarity in the results [26]. Overall, the shuffled promoter control reduces the number of potential regulatory elements by removing motifs that might seem important based only on base composition. The high-confidence motifs that passed both the expression-matched enrichment test and the shuffled promoter validation form a robust set of candidates for downstream analysis in the gene regulatory networks.



**Figure 4: Real vs. Shuffled Motif Scores with FDR Correction (*Lotus* Cluster 6)**

Each point represents a motif. The x-axis shows the total FIMO score for the motif in real At homolog (Lj cluster 6) promoters, while the y-axis shows the average FIMO score across 100 shuffled versions of the same sequences. Motifs with FDR-adjusted p-values below 0.05 (green) are significantly enriched in real data. The dashed line represents the null expectation (real = shuffled). Only motifs with greater scores in real data than in shuffled controls are considered enriched.

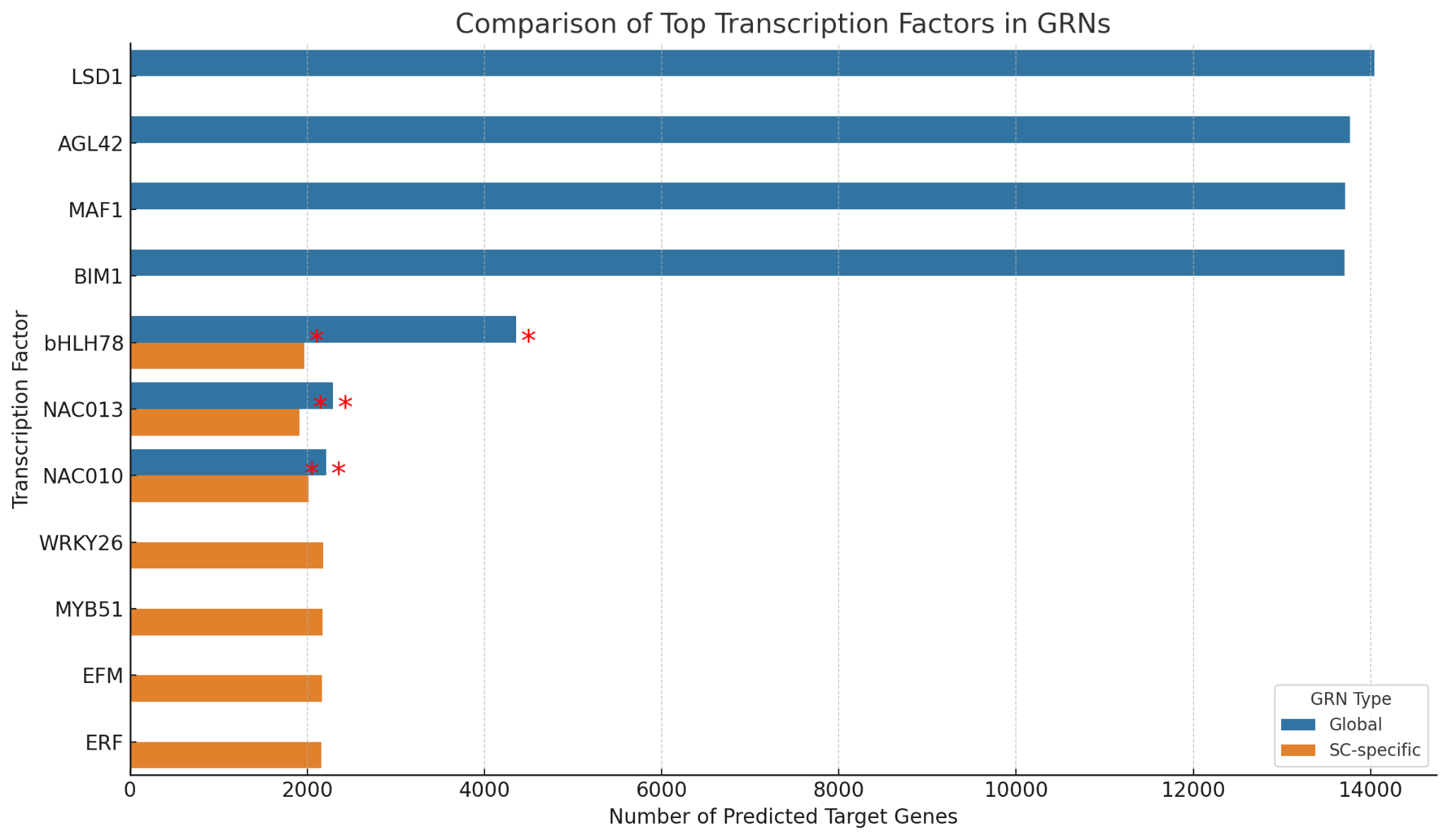


**Figure 5: Real vs. Shuffled Motif Scores with FDR Correction (*Arabidopsis* Cluster 3).**

Each point represents a motif. The x-axis shows the total FIMO score for the motif in real At cluster 3 promoters, while the y-axis shows the average FIMO score across 100 shuffled versions of the same sequences. Motifs with FDR-adjusted p-values below 0.05 (green) are significantly enriched in real data. The dashed line represents the null expectation (real = shuffled). Only motifs with greater scores in real data than in shuffled controls are considered enriched.

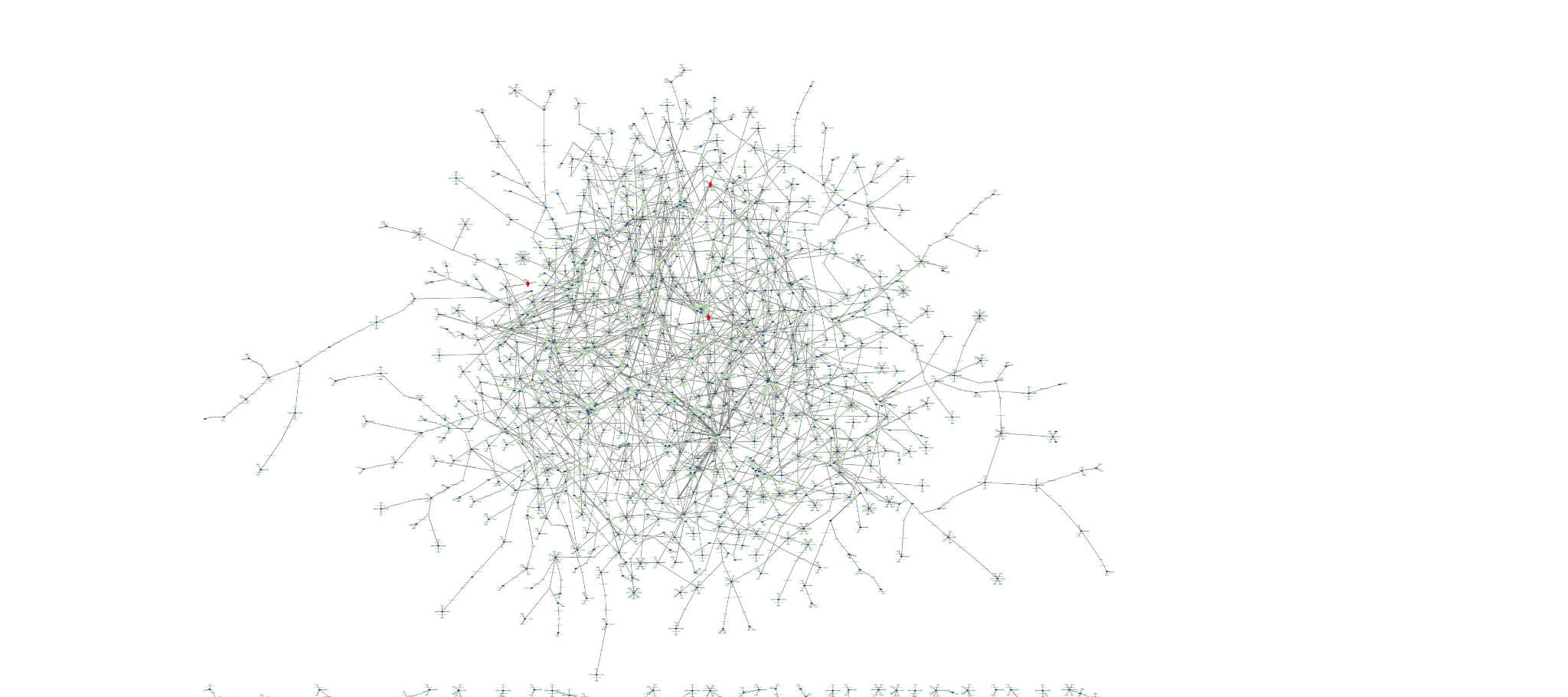
### 3.5 Construction of the Gene Regulatory Network (GRN)

GRNBoost2 was used to predict GRNs from transcriptome data in order to gain a better understanding of how plants regulate gene expression in response to microbiome colonization [10]. Using all 12 samples from all conditions from Wippel et al. (2021), a GRN was first built in order to identify globally significant TFs. In parallel, regulators were identified that may be unique to microbiota-driven expression programs by building a focused GRN solely from data (8 samples) in which the roots of At plants were exposed to their SCs (At and Lj). TFs with a wide-ranging impact were found using the global GRN. The most highly connected node was zinc finger protein LSD1 (AT4G20380), linked to developmental transitions and chromatin remodeling [27], with over 14,000 estimated target genes. Next were broad-acting TFs such as AGL42, MAF1, and BIM1 [28–30]. In many plant tissues, these regulators most likely control general transcriptional maintenance and cellular homeostasis [27–30]. However, a unique regulatory architecture was revealed by the SC-specific GRN. With 2,182 inferred targets, WRKY26 (AT5G07100) was the most highly ranked TF in this case, indicating a crucial function in coordinating responses, particularly during microbial colonization [33]. Other transcription factors that were previously linked to root immunity, secondary metabolism, and environmental response pathways included MYB51, EFM (a member of the MYB family), NAC103, and bHLH115 [34,35]. Thus, a shift in regulatory priorities toward microbial sensing and adaptability is reflected in this SC-specific network. Interestingly, a number of TFs (bHLH78, NAC010, and NAC013) that were found in the motif enrichment analysis (chapter 3.3) were also found in the SC-specific GRN. Although they might not be global hubs, these were ranked 652, 352, and 997, respectively, which are higher ranks than what they were ranked in the global GRN, indicating that they gain importance when considering microbial exposure. The fact that they appear simultaneously in motif enrichment and SC-GRN investigations provides combined evidence for their functional relevance in regulating cluster 3 genes, which are highly enriched in SC-specific responses. A visual illustration of this difference between the global and SC-specific GRNs can be found in Figure 6, which compares the number of estimated target genes for top-ranking transcription factors in both networks. The comparison emphasizes how general TFs are replaced by those with defense, signaling, and environmental adaptation specializations during SC formation, reprogramming transcriptional control. In order to provide more context for these findings, Cytoscape was used to show both the SC-specific GRN (Figure 7 and 8) and the top three TFs of the global GRN (Figure 9). The resulting network features a well-organized architecture with a dense regulatory core composed primarily of high-ranking transcription factors, like WRKY26 and MYB51, which operate as significant hubs with a large number of outgoing edges. This architecture emphasizes the crucial function that these TFs might have in coordinating gene expression responses during SC colonization. The essential role that these TFs may play in coordinating gene expression responses during SC colonization is highlighted by this architecture. The uniqueness of this network to microbiome-induced transcriptional programs is further supported by the fact that outlying nodes represent downstream targets, many of which are specifically regulated in the SC state. Interestingly, Figure 8 shows that the enriched motifs from section 3.3 are part of the downstream interactors. These results collectively imply that plants use a specific group of TFs to modify their transcriptional networks in response to the presence of microbiomes [31]. The findings of high-ranking SC-specific regulators such as WRKY26 and NAC103 provides a path for functional research focused at determining the transcriptional regulation of plant-microbe interactions.

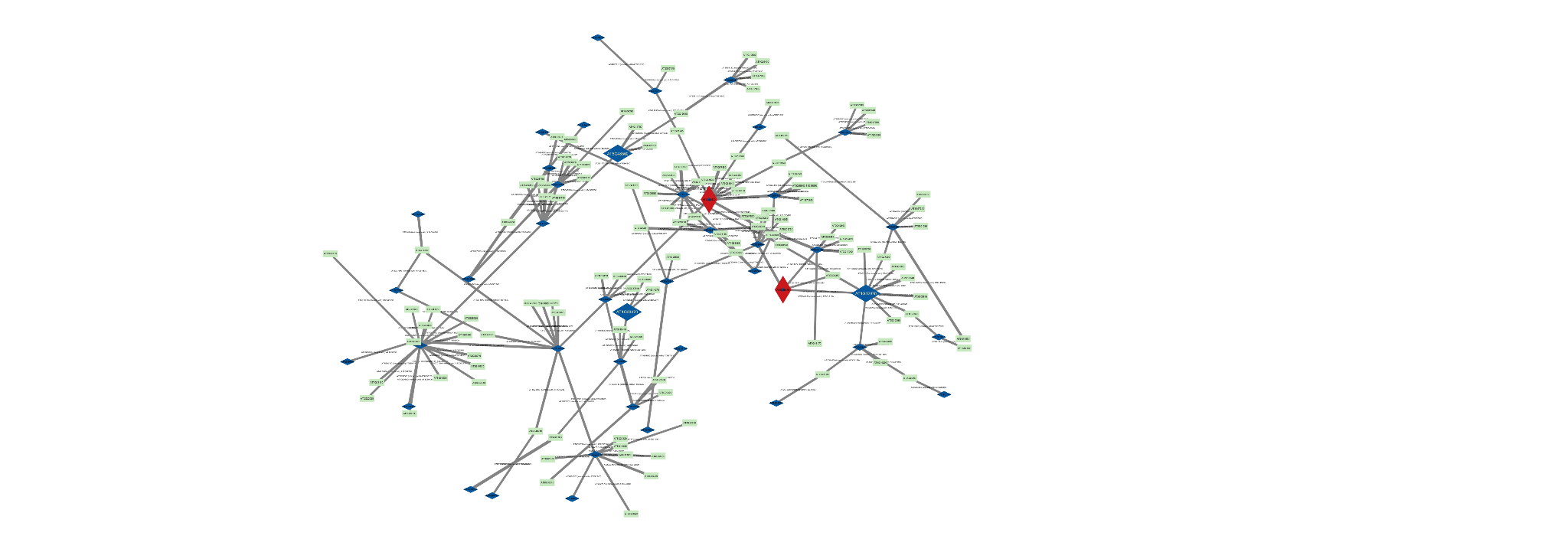


**Figure 6: Visual comparison of TFs between the global GRN and the SC-specific GRN for *Arabidopsis*.**

The orange bars represent TFs prominent in the SC-specific GRN. Blue bars represent TFs from the general, full-expression GRN. The global GRN was constructed using all transcriptomic samples, while the SC-specific GRN was inferred only from samples exposed to synthetic microbial communities. Notably, LSD1, AGL42, and MAF1 dominate the global network, whereas WRKY26, MYB51, and EFM emerge as highly connected TFs in the SC-specific network. Red asterisks indicate transcription factors that match enriched promoter motifs in cluster 3 genes (bHLH78, NAC010, NAC013), suggesting functional involvement in SC-induced gene expression programs.

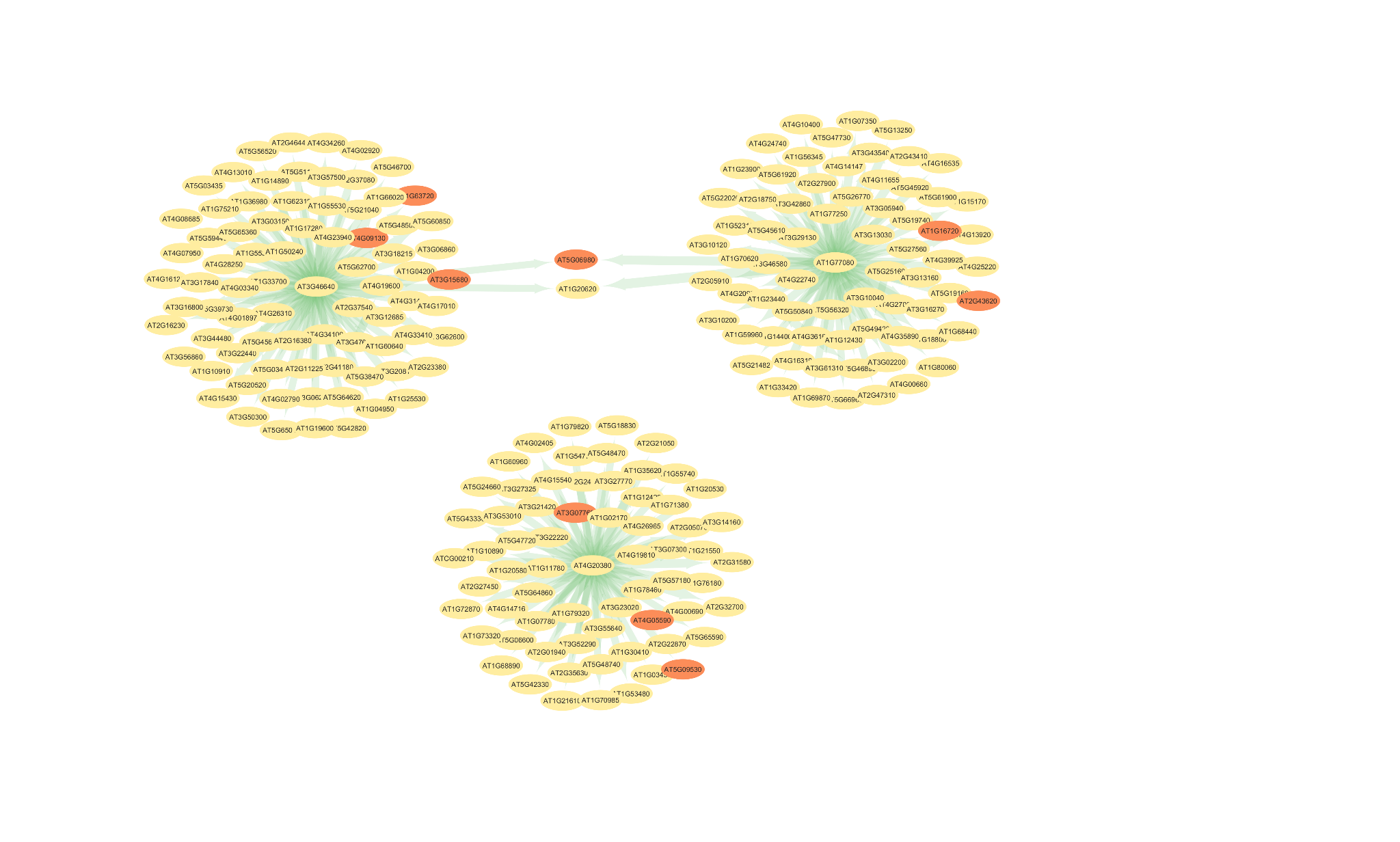


**Figure 7:** **Gene regulatory network inferred from *Arabidopsis* roots inoculated with synthetic communities, visualized in Cytoscape.** Edges represent inferred TF-target relationships from GRNBoost2 with importance scores above 2.0. Node positions are determined by a spring layout algorithm. Dark blue nodes are TFs, while light green nodes are targets of TFs. Inside the figure are three red diamond shaped nodes which represent the three enriched TFs from the enrichment analysis. The dense central cluster reflects a small number of transcription factors, such as WRKY26 and MYB51, acting as network hubs, consistent with their high ranking in Figure 6. This network topology supports a model in which SC-induced transcriptional regulation is coordinated by a defined set of master regulators.



**Figure 8: Focused subnetwork of the Arabidopsis root GRN centered on top two most prominent TFs in the SC-specific GRN visualized in Cytoscape.**

This subnetwork was extracted from the full SC-specific GRN by including the first three levels of downstream connections (incoming and outgoing) from the top two most prominent TFs in the GRN, WKRY26 and MYB51 (shown as red diamond-shaped nodes). Also in the figure are the enriched TFs bHLH78, NAC010, and NAC013 (shown as enlarged blue diamond-shaped nodes). These three TFs were identified as significantly enriched in the TF enrichment analysis and serve as the central nodes in this visualization. Edges represent inferred TF–target relationships predicted by GRNBoost2, filtered for importance scores above 2.0. Node layout was determined using an organic layout algorithm. Dark blue nodes correspond to other TFs, and light green nodes represent target genes. The inclusion of direct targets (1st level), targets of those targets (2nd level), and their downstream interactors (3rd level) provides insight into the broader regulatory influence of these enriched TFs within the SC-induced transcriptional network.



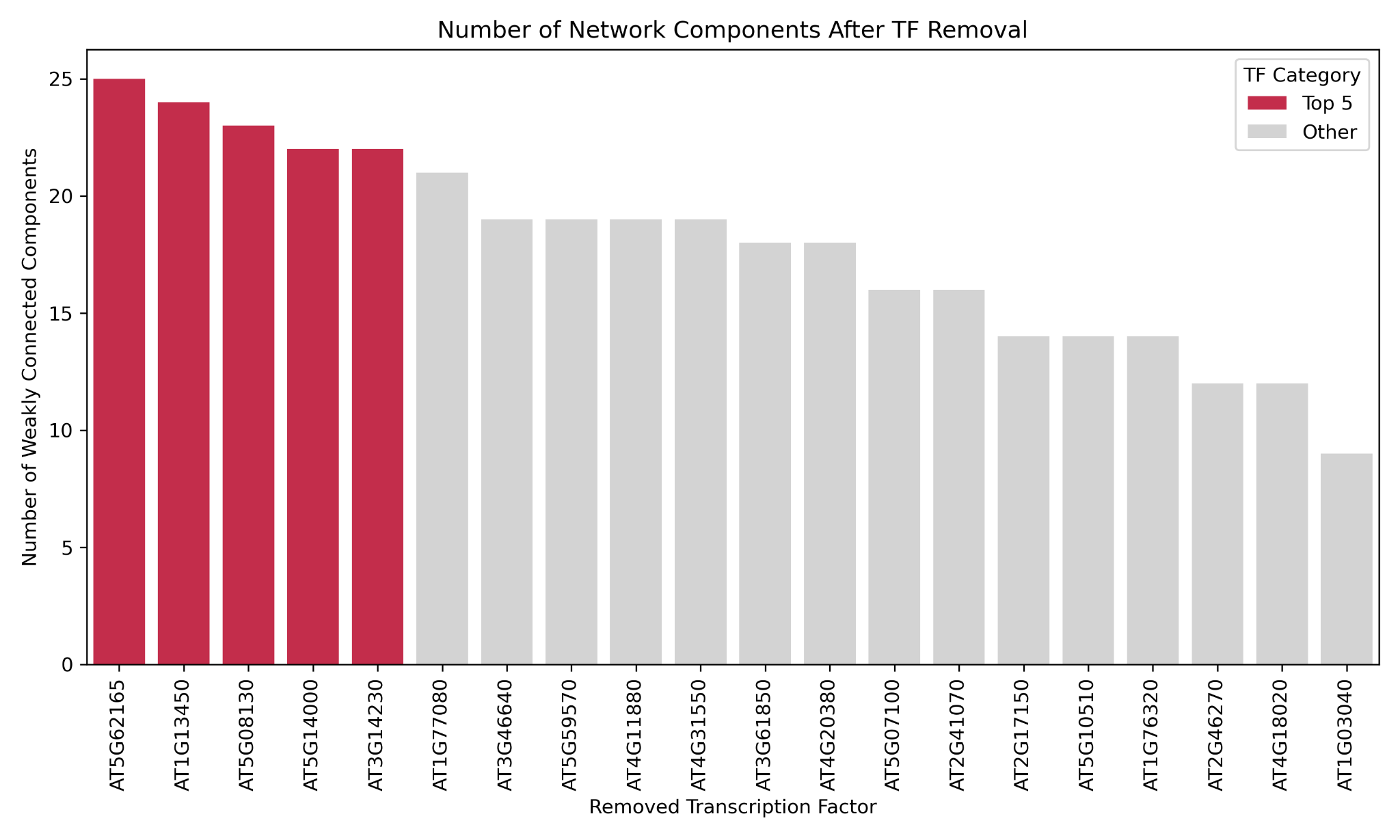
**Figure 9:** **Gene regulatory subnetwork centered on the top three transcription factors (TFs) with the highest number of predicted targets in the *Arabidopsis* global GRN, inferred using GRNBoost2 (importance scores > 2.0).** Each central TF (in the middle) is connected to its direct target genes (light yellow, orange for genes that are included in cluster 3), with edge thickness scaled by interaction importance score. The green edge color and node shading represent the magnitude of network importance. This visualization highlights the breadth of inferred regulatory control for LSD1 (AT4G20380), AGL42 (AT5G62165), and MAF1 (AT1G77080), revealing large and largely non-overlapping regulatory modules.

### 3.6 Robustness Testing with Perturbation Analysis Validates Network Architecture

To evaluate how sensitive the network topology is to perturbation, it was tested how individual TFs contribute to the structural integrity of the global- and SC-specific GRN. This helps identify TFs that function as structurally high connected hubs [25]. This investigation relied on the amount of weakly connected components (WCCs), which are subgraphs where nodes are connected by pathways if edge direction is disregarded. Stated differently, a WCC is a collection of genes and TFs that, even in the absence of regulatory directionality, are nevertheless accessible from one another [25]. Greater network fragmentation is indicated by an increase in WCCs after TF removal, indicating that the lost TF was crucial to preserving overall connectivity. The overall network stability was barely affected when the top 20 TFs (ordered by degree) in the SC-specific GRN were removed. Regardless of which TF was eliminated, the quantity of WCCs constantly remained at ~288 (Figure 10). This occurrence held even for the top five most connected TFs (AT5G62165, AT1G13450, AT5G08130, AT5G14000, and AT3G14230), suggesting a high degree of redundancy in regulatory structure. In contrast, the same perturbation strategy applied to the global GRN revealed a different pattern (Figure 11). Network fragmentation appeared to increase when the top five TFs (LSD1, AGL42, MAF1, BIM1, and bHLH78) were removed; the number of WCCs increased to 20–25, whereas for the majority of other TFs, it was less than 10. This suggests that a small number of TFs are more important to the global GRN's ability to stay connected. These divergent trends indicate that the two networks' organizational structures differ fundamentally. A few highly influential master regulators serve as the main hubs of the global GRN, which seem to be hierarchically organized. The network splits into separate subgraphs when they are removed, indicating that these TFs connect important regulatory modules. On the other hand, the topology of the SC-specific GRN is flatter and more dispersed. No single TF acts as a crucial high connected hub, and even the most connected nodes do not heavily maintain global cohesion. This suggests a more robust and modular architecture in which TFs most likely regulate localized transcriptional programs during the colonization of synthetic communities. Functionally, this implies that while TFs in the SC-GRN may have less important but targeted functions in regulating immunological, metabolic, or developmental responses during microbiome formation, regulators such as LSD1 and AGL42 may be crucial for overall transcriptional homeostasis.



**Figure 10: Comparison of network disruption after individual removal of the top 100 most connected TFs from the SC-specific GRN.** Barplot showing the number of weakly connected components remaining in the GRN after removal of individual TFs. TFs are ranked by their disruptive effect, with the top twenty most fragmenting TFs highlighted in red. All other TFs are shown in grey. Despite TF deletion, the number of weakly connected components remains stable (~288), suggesting general robustness and modularity of the network.



**Figure 11: Impact of transcription factor removal on network connectivity.**

Barplot showing the number of weakly connected components remaining in the global gene regulatory network (GRN) after removal of individual transcription factors (TFs). The top five most fragmenting TFs (AT5G62165, AT1G13450, AT5G08130, AT5G14000, and AT3G14230) are indicated in red. TFs are rated according to their disruptive effect. Every other TF is displayed in grey. When these top TFs were removed, the network broke up into 23–25 fragmented subgraphs, demonstrating their crucial structural function in preserving network cohesiveness. In contrast, removal of less connected TFs resulted in minimal disruption, suggesting functional redundancy or modular confinement of their regulatory influence.

# 4. Discussion

This study set out to uncover the transcriptional regulatory mechanisms underlying plant microbiome assembly by applying a machine-learning-based GRN inference pipeline to At and Lj transcriptomes exposed to host-derived synthetic microbial communities. Through a combination of motif discovery, enrichment testing, and GRN construction, this work identified both generalist and SC-specific TFs, validated candidate regulators with multiple layers of evidence, and evaluated the robustness of the inferred regulatory architecture.

### 4.1 Microbiome-Responsive Regulatory Elements

This study began by focusing on gene clusters (cluster 3 in At and cluster 6 in Lj) that were clearly upregulated in response to their native synthetic communities (SCs) (see Figure 1). These clusters served as a logical entry point for investigating transcriptional regulation during microbiome colonization.

FIMO was employed to search for known transcription factor binding motifs in promoter regions in order to find potential cis-regulatory elements. To make sure that any motifs enriched in cluster 3 were not only linked to highly expressed genes generally, they were tested for motif enrichment using an expression-matched background (Figure 2), as many motifs are found often throughout the genome [21].

To further validate that our results weren’t due to random chance, a shuffled promoter control was introduced [17]. This test confirmed that the three motifs corresponding to bHLH78, NAC010, and NAC013 were not only enriched compared to background but also not commonly found in randomized sequences, adding confidence to their biological relevance (Figure 5). These three TFs became especially interesting when they also turned up in the SC-specific gene regulatory network (GRN) [32]. Even though they didn’t have the highest number of inferred targets (Figure 6), their combined presence in both motif and network analyses suggests they might play specific roles during microbiome assembly, rather than being general-purpose regulators.

In addition, a STREME analysis was run for *de novo* motif discovery (Table I & II in supplementary data) [17]. While this analysis identified several recurring promoter patterns not captured by the known motif database, most could not be confidently linked to specific transcription factors [17]. Because these novel motifs lack annotation, they were not included in downstream GRN interpretation [32]. Tools such as TOMTOM[[8]](#footnote-8) could be used in future work to compare these *de novo* motifs to known TF binding profiles and assign putative regulators. Still, this exploratory step opens the door for further studies. STREME motifs could be mapped to novel or uncharacterized TFs using experimental techniques such as DAP-seq or TF perturbation studies [19]. In particular, applying this *de novo* discovery pipeline to Lj, where motif databases are less complete, may prove especially valuable for uncovering species-specific regulatory elements.

All things considered, motif enrichment helped to prioritize TFs that could otherwise go unnoticed in network-based analyses alone. Even though not every enriched motif matched a high-degree GRN hub, their overlap with GRN TFs demonstrates the importance of integrating several layers of evidence and validates their regulatory relevance.

### 4.2 Transcriptional Rewiring Under SC Conditions

Interestingly, the "global" GRN, which was constructed using all samples, and the "SC-specific" GRN which only included roots that had received SC treatment, looked very different (Figure 6). TFs like LSD1, AGL42, MAF1, and BIM1, which are probably engaged in basic or housekeeping gene regulation, led the global GRN (Figures 6, 9, 11).

In contrast, in the SC-specific network, TFs like WRKY26, MYB51, NAC103, and bHLH115 took center stage, TFs that are already known to be involved in stress responses and interactions with microbes [33-35]. Visualizing the SC-GRN in Cytoscape helped reinforce this idea (Figure 7). The network showed a dense core of well-connected TFs, many of which are associated with plant immune and metabolic processes. The enriched motif TFs (NAC010, NAC013, and bHLH78) were present in this network too, though not at the very top. This pattern points to a layered regulatory system in which more specialized TFs fine-tune gene expression in specific circumstances, such as microbiome colonization, while more generic TFs coordinate larger responses [36].

### 4.3 Network Stability Reveals Different Architectures

The "robustness" of the GRNs was evaluated by a perturbation study, in which the top TFs were removed to see the evolution of the network structure (Figures 10, 11). Despite removing the top 20 most connected TFs, the SC-GRN's network did not completely collapse. A considerable level of structural redundancy was indicated by the roughly constant number of poorly connected components, which was 288 (Figure 10). This further implies that the SC-GRN is built in a more decentralized manner, with several smaller hubs that can assist one another. Such a setup might be useful for plants, since microbiome conditions can change quickly and unpredictably [37,38].

On the other hand, the global GRN behaved differently. Removing top TFs like LSD1 and AGL42 caused the network to split into many disconnected parts (Figure 11). The number of disconnected components increased to 23–25, highlighting their role as global regulators and highly connected structural hubs [36]. This suggests that the global GRN is more susceptible to disruptions, relying more heavily on a small set of central TFs. It should be mentioned that this conflicting behavior may be caused by variations in network density and size. The structure of the global GRN is more susceptible to node removal, which may be explained by the fact that it contains more nodes and edges than the SC-specific GRN. Even when size is taken into consideration, the SC-specific GRN remains connected even after deleting its most interconnected nodes, suggesting a far more redundant and modular architecture. These findings suggest that plants could improve regulatory resilience in dynamic contexts by switching from a centralized control structure to a more dispersed network to alter transcriptional regulation in microbiome-related scenarios.

### 4.4 The Role of Motif Enrichment: Useful, but Limited

Motif enrichment analysis played a major role in this study by highlighting TFs including bHLH78, NAC010, and NAC013, which were identified as putatively key regulators despite not being among the top hubs in the GRN. These TFs might have gone entirely unnoticed without this step, underscoring the need of combining motif enrichment with additional regulatory inference methods. This is consistent with other studies that combined motif enrichment with co-expression or network data to identify non-obvious candidate regulators [39]. But it was also clear that motif enrichment had its limitations. Significantly enriched motifs in the investigated promoter regions were absent from a large number of the top-ranked TFs in the SC-specific GRN. This result mirrors well-known issues with motif-based prediction, such as the incompleteness of motif databases and the context-dependence of TF binding, which position weight matrices (PWMs) by themselves frequently fail to capture [40]. Although useful, motif enrichment may overlook regulators whose binding is condition-specific or results from cooperative interactions that are not captured by existing techniques.

It's also important to remember that the GRN-based strategy has drawbacks of its own. Direct binding or causality cannot be established by GRN inference from expression data. Instead, it depends on statistical relationships. Furthermore, it might give preference to highly linked TFs based just on their patterns of co-expression rather than their mechanistic centrality.

It is better to think of these two strategies as complementary rather than as superior to one another. GRN inference aids in capturing the larger regulatory architecture and possible hub dynamics, while motif enrichment can identify regulators with sequence-level support, particularly those engaged in localized or condition-specific regulation. Together, they boost candidate TF confidence and aid in ranking them for additional experimental validation. The targeted subnetwork depicted in Figure 8, which focuses on the two most important hub TFs in the SC-specific GRN, WRKY26 and MYB51, further supports this integrative viewpoint. The fact that the TFs (bHLH78, NAC010, and NAC013) identified through motif enrichment appear in this subnetwork as downstream interactors is intriguing. The inclusion of motif-enriched TFs in the regulatory influence of central GRN hubs suggests a layered architecture where they may act as context-dependent or secondary regulators, depending on the situation. This demonstrates how, although it does not establish direct regulatory links, integrating GRN topology with motif-based insights may enhance the identification of physiologically significant transcriptional regulators. Additionally, it provides greater proof of their alleged functional significance.

### 4.5 *Arabidopsis* vs. *Lotus*: The Limits of Cross-Species Tools

A major limitation of this study is the heavy reliance on At resources, particularly for promoter sequences and known transcription factor (TF) motifs. For Lj, no comparable ready-to-use resources were available in a format directly compatible with the computational pipeline developed in this project. While there are a number of publicly available resources for Lj, including the genome located on Phytozome, Lotus Base, and LegumeBase, they have certain drawbacks for motif-level regulation study. For instance, whereas Phytozome offers gene models and genome assemblies, it does not provide pre-extracted promoter sequences or a curated database of TF binding motifs unique to Lotus that may be used with programs like FIMO or GRNBoost2. Similarly, Lotus Base and LegumeBase do not include pre-compiled TF-PWM libraries or standardized, bulk-downloadable upstream regions, but they do provide useful annotations and gene-level information. Therefore, significant pre-processing processes that were outside the purview of this investigation would have been necessary to identify usable promoter sequences and generate motif inputs. Due to the defined time frame, certain steps had to be prioritized, and the Lj component was therefore limited in scope. A more comprehensive, species-specific analysis would require custom data preparation, including extracting upstream promoter sequences (e.g., 1–2 kb) using tools such as pybedtools or gffutils, followed by *de novo* motif discovery with MEME Suite [16,41]. Moreover, either orthology inference using programs like EggNOG-mapper or protein domain-based prediction using InterProScan would be necessary for the annotation of Lj TFs [42, 43].

Even though these procedures add more complexity, they are essential to revealing the regulatory logic unique to Lj, particularly in light of its unique biology, which includes nitrogen-fixing symbiosis and the production of root nodules [19]. Future research that more thoroughly integrates these resources may be able to identify Lj-specific regulatory components that are currently overlooked when using At homologs as stand-ins.

### 4.6 Answering the Big Question: Who’s in Charge of Microbiome-Driven Expression?

This study was motivated by the straightforward question of which transcription factors influence how a plant reacts to its microbiome. The results show that plants activate a different regulatory system when exposed to microbial communities (Figure 6). They enlist a new class of regulators, such as WRKY26 and MYB51, that are more adept at controlling immune responses and microbial signals rather than depending on generalist TFs [33-35]. These TFs coordinate a significant component of the transcriptional response, though others like bHLH78 and NAC010 likely refine it. In addition to affecting which genes are triggered, this regulatory change affects how the plant develops a more resilient and adaptable system to cope with its surroundings [33]. That provides a basic understanding of how plants strike a balance between flexibility and stability.

### 4.7 Future Directions

This study might establish the foundation for determining the transcription factors involved in plant-microbe interactions. These results suggest a scenario in which plants activate a unique set of context-specific regulators to change their transcriptional architecture in response to microbial colonization. Looking towards the future, the involvement of potential TFs such as bHLH78, WRKY26, MYB51, NAC010, NAC013, or NAC103 can be tested using targeted research (e.g., gene knockouts or promoter tests) [44]. Additionally, comparable pipelines can be used across species when more genomic tools for non-model plants, such as Lj, become available. The first step in creating advantageous plant microbiomes for agriculture is ultimately comprehending these transcriptional processes.

# 5. Conclusion

The main goal of the study was to determine which transcription factors manage how a plant responds transcriptionally to its microbiome. Through a combination of robustness testing, network inference, and motif enrichment, this study discovered a distinct regulatory mechanism that activates when plants engage with artificial microbial communities. From generalist TFs like LSD1 and AGL42 in the global network to specialized regulators like WRKY26, MYB51, and NAC103 in the SC-specific context, the results clearly demonstrate a change in regulatory architecture.

Combining motif enrichment and GRN prediction worked very well for finding subtle, potential regulators that could otherwise go unnoticed. NAC010 and bHLH78, for instance, did not rank among the top GRN hubs; yet, their binding motifs passed statistical validation procedures and were considerably enriched in microbiome-responsive promoters. This highlights the value of employing a variety of complementary techniques to identify more focused, condition-specific regulators as well as globally related TFs.

The study's dependence on At-specific resources limited the level of research that Lj could conduct, which was one of its main limitations. Creating unique genomic inputs, such as upstream promoter regions and species-specific motif libraries, will be necessary to completely apply this approach to Lj or other non-model species.

Altogether, the findings support a model in which plants don’t just passively receive microbial signals, but actively rewire their transcriptional programs in response. The SC-specific GRN showed a high level of resistance to disruption in addition to its unique composition, suggesting that plants have evolved a reliable modular regulatory mechanism for managing interactions with the microbiome. This discovery opens the way for further targeted studies and practical uses, such as genetically confirming possible TFs to eventually create microbiome-friendly crops.

# 7. Bibliography

[1] Foley, J. A., Ramankutty, N., Brauman, K. A., Cassidy, E. S., Gerber, J. S., Johnston, M., Mueller, N. D., O’Connell, C., Ray, D. K., West, P. C., Balzer, C., Bennett, E. M., Carpenter, S. R., Hill, J., Monfreda, C., Polasky, S., Rockström, J., Sheehan, J., Siebert, S., . . . Zaks, D. P. M. (2011). Solutions for a cultivated planet. *Nature*, *478*(7369), 337–342.

[2] Zhang, X., Davidson, E. A., Mauzerall, D. L., Searchinger, T. D., Dumas, P., & Shen, Y. (2015). Managing nitrogen for sustainable development. *Nature*, *528*(7580), 51–59.

[3] Mus, F., Crook, M. B., Garcia, K., Costas, A. G., Geddes, B. A., Kouri, E. D., Paramasivan, P., Ryu, M., Oldroyd, G. E. D., Poole, P. S., Udvardi, M. K., Voigt, C. A., Ané, J., & Peters, J. W. (2016). Symbiotic Nitrogen Fixation and the Challenges to Its Extension to Nonlegumes. *Applied And Environmental Microbiology*, *82*(13), 3698–3710.

[4] Fitzpatrick, C. R., Salas-González, I., Conway, J. M., Finkel, O. M., Gilbert, S., Russ, D., Teixeira, P. J. P. L., & Dangl, J. L. (2020). The Plant Microbiome: From Ecology to Reductionism and Beyond. *Annual Review Of Microbiology*, *74*(1), 81–100.

[5] Wippel, K., Tao, K., Niu, Y., Zgadzaj, R., Kiel, N., Guan, R., Dahms, E., Zhang, P., Jensen, D. B., Logemann, E., Radutoiu, S., Schulze-Lefert, P., & Garrido-Oter, R. (2021). Host preference and invasiveness of commensal bacteria in the Lotus and Arabidopsis root microbiota. *Nature Microbiology*, *6*(9), 1150–1162.

[6] Yang, L., Qian, X., Zhao, Z., Wang, Y., Ding, G., & Xing, X. (2024). Mechanisms of rhizosphere plant-microbe interactions: molecular insights into microbial colonization. *Frontiers in Plant Science*, *15*.

[7] Hacquard, S., Spaepen, S., Garrido-Oter, R., & Schulze-Lefert, P. (2017). Interplay Between Innate Immunity and the Plant Microbiota. *Annual Review Of Phytopathology*, *55*(1), 565–589.

[8] Castrillo, G., Teixeira, P. J. P. L., Paredes, S. H., Law, T. F., De Lorenzo, L., Feltcher, M. E., Finkel, O. M., Breakfield, N. W., Mieczkowski, P., Jones, C. D., Paz-Ares, J., & Dangl, J. L. (2017). Root microbiota drive direct integration of phosphate stress and immunity. *Nature*, *543*(7646), 513–518.

[9] Weirauch, M. T., Yang, A., Albu, M., Cote, A. G., Montenegro-Montero, A., Drewe, P., Najafabadi, H. S., Lambert, S. A., Mann, I., Cook, K., Zheng, H., Goity, A., Van Bakel, H., Lozano, J., Galli, M., Lewsey, M. G., Huang, E., Mukherjee, T., Chen, X., . . . Hughes, T. R. (2014). Determination and Inference of Eukaryotic Transcription Factor Sequence Specificity. *Cell*, *158*(6), 1431–1443.

[10] Moerman, T., Santos, S. A., González-Blas, C. B., Simm, J., Moreau, Y., Aerts, J., & Aerts, S. (2018). GRNBoost2 and Arboreto: efficient and scalable inference of gene regulatory networks. *Bioinformatics*, *35*(12), 2159–2161.

[11] Pratapa, A., Jalihal, A. P., Law, J. N., Bharadwaj, A., & Murali, T. M. (2020). Benchmarking algorithms for gene regulatory network inference from single-cell transcriptomic data. *Nature Methods*, *17*(2), 147–154.

[12] Greenfield, A., Hafemeister, C., & Bonneau, R. (2013). Robust data-driven incorporation of prior knowledge into the inference of dynamic regulatory networks. *Bioinformatics*, *29*(8), 1060–1067.

[13] Castrillo, G., Teixeira, P. J. P. L., Paredes, S. H., Law, T. F., De Lorenzo, L., Feltcher, M. E., Finkel, O. M., Breakfield, N. W., Mieczkowski, P., Jones, C. D., Paz-Ares, J., & Dangl, J. L. (2017b). Root microbiota drive direct integration of phosphate stress and immunity. *Nature*, *543*(7646), 513–518.

[14] Liu, Q., Cheng, L., Nian, H., Jin, J., & Lian, T. (2022). Linking plant functional genes to rhizosphere microbes: a review. *Plant Biotechnology Journal*, *21*(5), 902–917.

[15] Huynh-Thu, V. A., Irrthum, A., Wehenkel, L., & Geurts, P. (2010). Inferring Regulatory Networks from Expression Data Using Tree-Based Methods. *PLoS ONE*, *5*(9), e12776.

[16] Grant, C. E., Bailey, T. L., & Noble, W. S. (2011). FIMO: scanning for occurrences of a given motif. *Bioinformatics*, *27*(7), 1017–1018.

[17] Bailey, T. L. (2021). STREME: accurate and versatile sequence motif discovery. *Bioinformatics*, *37*(18), 2834–2840.

[18] O’Malley, R. C., Huang, S. C., Song, L., Lewsey, M. G., Bartlett, A., Nery, J. R., Galli, M., Gallavotti, A., & Ecker, J. R. (2016). Cistrome and Epicistrome Features Shape the Regulatory DNA Landscape. *Cell*, *165*(5), 1280–1292.

[19] Franco-Zorrilla, J. M., López-Vidriero, I., Carrasco, J. L., Godoy, M., Vera, P., & Solano, R. (2014). DNA-binding specificities of plant transcription factors and their potential to define target genes. *Proceedings Of The National Academy Of Sciences*, *111*(6), 2367–2372.

[20] MacIsaac, K. D., & Fraenkel, E. (2006). Practical Strategies for Discovering Regulatory DNA Sequence Motifs. *PLoS Computational Biology*, *2*(4), e36.

[21] Zambelli, F., Pesole, G., & Pavesi, G. (2012). Motif discovery and transcription factor binding sites before and after the next-generation sequencing era. *Briefings in Bioinformatics*, *14*(2), 225–237.

[22] Kim, H. (2017). Statistical notes for clinical researchers: Chi-squared test and Fisher’s exact test. *Restorative Dentistry & Endodontics*, *42*(2), 152.

[23] Glueck, D. H., Mandel, J., Karimpour-Fard, A., Hunter, L., & Muller, K. E. (2008). Exact Calculations of Average Power for the Benjamini-Hochberg Procedure. *The International Journal Of Biostatistics*, *4*(1).

[24] Jiang, P., & Singh, M. (2013). CCAT: Combinatorial Code Analysis Tool for transcriptional regulation. *Nucleic Acids Research*, *42*(5), 2833–2847.

[25] Caldu-Primo, J. L., Alvarez-Buylla, E. R., & Davila-Velderrain, J. (2018). Structural robustness of mammalian transcription factor networks reveals plasticity across development. *Scientific Reports*, *8*(1).

[26] Nadimpalli, S., Persikov, A. V., & Singh, M. (2015). Pervasive Variation of Transcription Factor Orthologs Contributes to Regulatory Network Evolution. *PLoS Genetics*, *11*(3), e1005011.

[27] Dietrich, R. A., Richberg, M. H., Schmidt, R., Dean, C., & Dangl, J. L. (1997). A Novel Zinc Finger Protein Is Encoded by the Arabidopsis LSD1 Gene and Functions as a Negative Regulator of Plant Cell Death. *Cell*, *88*(5), 685–694.

[28] Dorca‐Fornell, C., Gregis, V., Grandi, V., Coupland, G., Colombo, L., & Kater, M. M. (2011). The Arabidopsis SOC1‐like genes AGL42, AGL71 and AGL72 promote flowering in the shoot apical and axillary meristems. *The Plant Journal*, *67*(6), 1006–1017.

[29] Ahn, C. S., Lee, D., & Pai, H. (2018). Characterization of Maf1 in Arabidopsis: function under stress conditions and regulation by the TOR signaling pathway. *Planta*, *249*(2), 527–542.

[30] Liang, T., Mei, S., Shi, C., Yang, Y., Peng, Y., Ma, L., Wang, F., Li, X., Huang, X., Yin, Y., & Liu, H. (2018). UVR8 Interacts with BES1 and BIM1 to Regulate Transcription and Photomorphogenesis in Arabidopsis. *Developmental Cell*, *44*(4), 512-523.e5.

[31] Keppler, A., Roulier, M., Pfeilmeier, S., Petti, G. C., Sintsova, A., Maier, B. A., Bortfeld-Miller, M., Sunagawa, S., Zipfel, C., & Vorholt, J. A. (2024). Plant microbiota feedbacks through dose-responsive expression of general non-self response genes. *Nature Plants*.

[32] Marbach, D., Costello, J. C., Küffner, R., Vega, N. M., Prill, R. J., Camacho, D. M., Allison, K. R., Kellis, M., Collins, J. J., & Stolovitzky, G. (2012). Wisdom of crowds for robust gene network inference. *Nature Methods*, *9*(8), 796–804.

[33] Eulgem, T., & Somssich, I. E. (2007). Networks of WRKY transcription factors in defense signaling. *Current Opinion in Plant Biology*, *10*(4), 366–371.

[34] Frerigmann, H., & Gigolashvili, T. (2014). MYB34, MYB51, and MYB122 Distinctly Regulate Indolic Glucosinolate Biosynthesis in Arabidopsis thaliana. *Molecular Plant*, *7*(5), 814–828.

[35] Liang, G., Zhang, H., Li, X., Ai, Q., & Yu, D. (2017). bHLH transcription factor bHLH115 regulates iron homeostasis in Arabidopsis thaliana. *Journal Of Experimental Botany*, *68*(7), 1743–1755.

[36] Heyndrickx, K. S., Van de Velde, J., Wang, C., Weigel, D., & Vandepoele, K. (2014). A Functional and Evolutionary Perspective on Transcription Factor Binding in Arabidopsis thaliana. *The Plant Cell*, *26*(10), 3894–3910.

[37] Birkenbihl, R. P., Kracher, B., Ross, A., Kramer, K., Finkemeier, I., & Somssich, I. E. (2018). Principles and characteristics of the Arabidopsis WRKY regulatory network during early MAMP‐triggered immunity. *The Plant Journal*, *96*(3), 487–502.

[38] Thiergart, T., Durán, P., Ellis, T., Vannier, N., Garrido-Oter, R., Kemen, E., Roux, F., Alonso-Blanco, C., Ågren, J., Schulze-Lefert, P., & Hacquard, S. (2019). Root microbiota assembly and adaptive differentiation among European Arabidopsis populations. *Nature Ecology & Evolution*, *4*(1), 122–131.

[39] Ma, S., Shah, S., Bohnert, H. J., Snyder, M., & Dinesh-Kumar, S. P. (2013). Incorporating Motif Analysis into Gene Co-expression Networks Reveals Novel Modular Expression Pattern and New Signaling Pathways. *PLoS Genetics*, *9*(10), e1003840.

[40] Inukai, S., Kock, K. H., & Bulyk, M. L. (2017). Transcription factor–DNA binding: beyond binding site motifs. *Current Opinion in Genetics & Development*, *43*, 110–119.

[41] Dale, R. K., Pedersen, B. S., & Quinlan, A. R. (2011). Pybedtools: a flexible Python library for manipulating genomic datasets and annotations. *Bioinformatics*, *27*(24), 3423–3424.

[42] Cantalapiedra, C. P., Hernández-Plaza, A., Letunic, I., Bork, P., & Huerta-Cepas, J. (2021). eggNOG-mapper v2: Functional Annotation, Orthology Assignments, and Domain Prediction at the Metagenomic Scale. *Molecular Biology And Evolution*, *38*(12), 5825–5829.

[43] Mulder, N., & Apweiler, R. (2007). InterPro and InterProScan. *Methods in Molecular Biology*, 59–70.

[44] Zhang, Y., Tessaro, M. J., Lassner, M., & Li, X. (2003). Knockout Analysis of Arabidopsis Transcription Factors TGA2, TGA5, and TGA6 Reveals Their Redundant and Essential Roles in Systemic Acquired Resistance. *The Plant Cell*, *15*(11), 2647–2653.

1. <https://planttfdb.gao-lab.org/> Consulted on 20-06-2025 [↑](#footnote-ref-1)
2. <https://www.arabidopsis.org/> Consulted on 20-06-2025 [↑](#footnote-ref-2)
3. <https://github.com/IdsMe-2/Microbial-Matchmakers.git> Consulted on 26-06-2025 [↑](#footnote-ref-3)
4. <https://jaspar.elixir.no/> Consulted on 20-06-2025 [↑](#footnote-ref-4)
5. <https://www.pantherdb.org/> Consulted on 20-06-2025 [↑](#footnote-ref-5)
6. <https://scikit-learn.org/stable/modules/neighbors.html> Consulted on 20-6-2025 [↑](#footnote-ref-6)
7. <https://cytoscape.org/> Consulted on 20-06-2025 [↑](#footnote-ref-7)
8. <https://meme-suite.org/meme/tools/tomtom> Consulted 20-06-2025 [↑](#footnote-ref-8)