JUSTUS-LIEBIG-
UNIVERSITÄT
GIESSEN

Illuminating the Regulatory Dynamics of Plant Growth: A Study on Cell Cycle Genes and Motif Enrichment in *Arabidopsis thaliana*

Department of Phytopathology
at
Interdisciplinary Research Center for Biosystems,
Land Use and Nutrition (iFZ)
Justus-Liebig-Universität Gießen

Supervisor: Prof. Dr. Patrick Schäfer

Prof. Dr. Sascha Ott

INAUGURAL-DISSERTATION

zur Erlangung des Doktorgrades (Dr. rer. nat.)

vorgelegt von

Wang, Xuesong (王雪松)
aus Jiangsu China

Gießen, 2024

Erklärung gemäß der Promotionsordnung des Fachbereichs 09 vom 07. Juli 2004 § 17 (2)

„Ich erkläre: Ich habe die vorgelegte Dissertation selbständig und ohne unerlaubte fremde Hilfe und nur mit den Hilfen angefertigt, die ich in der Dissertation angegeben habe.

Alle Textstellen, die wörtlich oder sinngemäß aus veröffentlichten Schriften entnommen sind, und alle Angaben, die auf mündlichen Auskünften beruhen, sind als solche kenntlich gemacht.

Bei den von mir durchgeführten und in der Dissertation erwähnten Untersuchungen habe ich die Grundsätze guter wissenschaftlicher Praxis, wie sie in der „Satzung der Justus-Liebig-Universität Gießen zur Sicherung guter wissenschaftlicher Praxis“ niedergelegt sind, eingehalten.“

Mit Genehmigung des Fachbereichs Agrarwissenschaften,

Ökotrophologie und Umweltmanagement der

Justus-Liebig-Universität Gießen

Prüfungskommission:

1. Gutachter(in):	Prof. Dr. Patrick Schäfer
2. Gutachter(in):	Professor Sascha Ott
Prüfer(in):	Prof. Dr. Agnieszka A. Golicz
Prüfer(in):	Prof. Dr. Stefan Janssen
Vorsitzende(r):	Prof. Dr. Joachim Aurbacher

Tag der Disputation: 29.11.2024

Table of Contents

Acknowledgements

Firstly, I would like to express my deep gratitude to my first supervisor, Prof. Dr. agr. Patrick Schäfer. Four years ago, his trust and affirmation set me on the journey of pursuing a PhD, and his guidance helped me navigate through the uncertainties of life. I am immensely grateful for the academic guidance provided by both Patrick and my second supervisor, Prof. Dr. Sascha Ott, during my doctoral studies. Their wisdom and patience have been invaluable assets in my academic growth.

I would also like to extend special thanks to Dr. Ruth Schäfer and other colleagues who have assisted me in experimental research. It is their support that has transformed me from a 'dry' guy with no knowledge of experiments into a 'semi-wet' experimenter.

Most importantly, I want to thank my family for their unwavering support over these four years. Without their encouragement and financial assistance, I would not have achieved what I have today.

Abstract

Plant growth and development rely on the proliferation and expansion of root tip cells, which involves a network of genes participating in the regulation of the cell cycle, namely the mitotic cell cycle and endocycle. Since plants have their sessile nature, they cannot actively avoid adverse factors imposed by the environment, compelling them to develop a suite of adaptive mechanisms to enhance their adaptability and ensure survival. Some of these adaptive mechanisms influence the regulation of the cell cycle, although the underlying mechanistic connections remain elusive. Including the role of transcription factors in regulating the expression of cell cycle genes and their influencing factors, as well as the regulation of the transition between the mitotic cycle and endocycle, remain current research challenges.

Research on plant cell cycle genes reached its peak in the early 21st century when combining cell cycle synchronization in cell cultures with DNA microarray and other technologies. At that time cycle-related genes have been assigned to different cell cycle phases. However, since then, the cell cycle gene networks have hardly been studied or annotated in more detail. Moreover, the ramifications of cell cycle synchronization on the cell cycle dynamics are not fully characterized and quantified. To this end, this work aimed to map and analyze cell cycle gene networks, leveraging more advanced technologies and methodologies, such as single-cell sequencing, to undertake a comprehensive exploration of the cell cycle networks. As for transcriptional regulation mechanisms of cell cycle genes, the focus has primarily been on the issue of transcription factor binding under single motif conditions. This work will focus on the binding characteristics of motif pairs and their role in transcriptional regulation.

This work integrated root single-cell and bulk sequencing data of *Arabidopsis thaliana* to calculate correlations, delineating developmental zones at the single-cell level. Cycling cells in these zones were identified using known cell cycle genes and clustering methods. Differential expression analyses on cells in different cycle phases expanded the cell cycle-related gene set.

The Paired Motif Enrichment Tool (PMET) identifies promoter motif pairs. Building on the existing PMET foundation, this work extends its application to uncover genetic mechanisms regulating the mitotic cycle and endocycle. Using PMET, diverse promoter motif pairs linked to cell cycle-related and stress-induced genes were identified, elucidating gene regulatory patterns involving synergistic transcription factors. Many motif pairs are specific to genes in certain cycle phases and stress conditions, offering new insights into the regulatory mechanisms of cell cycle and stress-induced gene networks.

Abbreviations

A Adenine

A. thaliana Arabidopsis thaliana

ABA Abscisic acid

ACT Arabidopsis co-expression tool

AGI Arabidopsis genome initiative

AAML1 Acyl-CoA metabolism-associated-like 1

ALLR Average log-likelihood ratio

APC/C Anaphase-promoting complex/cyclosom

AUC Area under curve

B1H Bacterial one hybrid selections

BHAT Bhattacharyya coefficient

BioID Proximity-dependent biotinylation

bp Base pair

C Cytosine

C2H2 Cys2his2-type zinc-finger

CCA Canonical correlation analysis

C/EBP CCAAT/Enhancer binding protein

CCAT Combinatorial code analysis tool

CDF Cumulative distribution function

CDK Cyclin-dependent kinase

CDS Coding sequence

CEL-Seq Single-cell expression by linear amplification sequencin

ChIP-seq Chromatin immunoprecipitation coupled with next-generation sequencing

CKI CDK inhibitor

CKL CDK-like kinase

CLI Command line interface

Cister Cis-element cluster finder

CRM Cis-regulatory module

CellMixS Cell mixing score

CYC Cyclin

CytoTRACE Cellular (cyto) trajectory reconstruction analysis

DamID-seq DNA adenine methyltransferase identification sequencing

DNase-seq DNase I hypersensitive sites sequencing

DP Dimerization partner

DZ Dvelopmental/differential zone

DragoNNs Deep regulatory genomic neural networks

EP Enhancer-promoter

ETS Erythroblast transformation specific

EUCL Euclidean distance

EdU 5-ethynyl-2'-deoxyuridine

FACS Fluorescence-activated cell sorting

FDR False discovery rate

G Guanine

GO Gene ontology

GSEA Gene set enrichment analysis

G1 Gap 1

G2 Gap 2

HB Homeobox

HELL Hellinger distance

Hi-C High-throughput chromosome conformation capture

HMM Hidden markov model

HOCOMOCO Homo sapiens comprehensive model collection

HPC High-performance computing

HSF Heat shock factor

HT High-throughput

HVG Highly variable gene

IC Information content

ICK Interactor/inhibitor of cyclin-dependent kinase

iPSC Induced pluripotent stem cell

KL Kullback-leibler divergence

KNN K-nearest neighbors

KRP Kip-related protein

MAN Manhattan distance

M Mitosis

MatInspector Matrix inspector

MATQ-seq Multiple Annealing and Tailing-based Quantitative sequencing

MARS-seq Massively Parallel RNA Single-cell sequencing

MEME Multiple expectation maximizations for motif elicitation

MITOMI Mechanically induced trapping of molecular interactions

MPC Model predictive control

MTC Multiple testing correction

NGS Next-generation sequencing

PAMP Pathogen-associated molecular pattern

PBM Protein binding microarray

PCA Principal component analysis

PCC Pearson correlation

PCD Programmed cell death

PCR Polymerase chain reaction

PC Principal component

PDC Probability density curve

Pfam Protein families

PMET Paired motif enrichment tool

PR Pathogenesis related

PRR Pattern-recognition receptor

PSSM Position specific scoring matrix

PTI Pattern-triggered immunity

PWM Position weight matrix

PW Powdery mildew

Pti4 Pto interacting protein 4

RADAR-seq DNA Repair And Damage-sequencing

RAM Root apical meristem

RB Retinoblastoma

RCJ Root cap junction

Rb Retinoblastoma

RUNx1 Runt-related TF 1

scATAC-seq Single-cell assay for transposase-accessible chromatin sequencing

scDam&T-seq Single-cell DNA adenine methyltransferase identification sequencing and Transcriptome sequencing

scM&T-seq single-cell methylome and transcriptome sequencing

scRNA-seq Single-cell RNA sequencing

SEBF Silencer element binding factor

scM&T-seq Single-cell methylome and transcriptome sequencing

SELEX Systematic evolution of ligands through exponential enrichment

SEUCL Squared euclidean distance

SIM Siamese

SMART-seq Switching Mechanism at 5' end of RNA Template sequencing

SMR Siamese-related protein

SNN Shared nearest neighbor

SNP Single-nucleotide polymorphism

SP3 Specificity protein 3

SUMO Small ubiquitin-related modifier

SW Sandelin-wasserman similarity

T Thymine

TAIR The arabidopsis information resource

TF Transcription factor

TFBS Transcription factor binding site

tSNE T-distributed stochastic neighbor embeddin

TSS Transcription start site

TZ Transition zone

U Uracil

UMAP Uniform manifold approximation and projection

UMI Unique molecular identifier

UTR Untranslated region

WEUCL Weighted euclidean distance

WPCC Weighted pearson correlation

iPSC Induced pluripotent stem cell

mdm4 Mdmx

ssGSEA Single-sample gsea

tSNE T-distributed stochastic neighbor embedding

1 Introduction

1.1 Cell Cycle Dynamics and Plant Stress

1.1.1 Mitotic Cell Cycle and Endocycle

The growth and development of plants are orchestrated by the proliferation and differentiation of newly formed cells, resulting from processes that include cell expansion, division, and differentiation. These processes can be conceptualized within the framework of two main cell cycle models: the mitotic cell cycle and endoreduplication (also referred to as the endocycle). In the model plant *A. thaliana*, like in other plants, the mitotic cell cycle exhibits a characteristic sequence of phases that precede cell division, including the gap 1 (G1), synthesis (S), and gap 2 (G2) phases, followed by mitotic phase (M). For the endocycle, the M phase is absent [1,2,3,4].

During the G1 phase of the cell cycle, cellular growth occurs as the cell increases in size, produces organelles and other cellular contents. Concurrently, there is an increased demand for the synthesis of essential carbohydrates, proteins, and lipids, which in turn accelerates cellular metabolism [5,6]. The S phase is primarily characterized by DNA replication and is highly conserved among eukaryotes [7,8]. During this phase, DNA undergoes replication, making it particularly susceptible to instability and errors due to the unwinding of the double helix, which leads to the formation of unstable single strands that are prone to mutation [9].

Like the G1 phase, the G2 phase is characterized by further cell growth and the synthesis of RNA and proteins. Additionally, preparations for the M phase include the transformation of chromatin into condensed chromosomes and the disappearance of the nuclear membrane. The key distinction between the G1 and G2 phases lies in the doubling of genetic material in the latter.

The mitotic cycle, highly conserved across animals and plants [4,10,11], is characterized by comparable regulatory mechanisms that ensure the production of two daughter cells through a series of coordinated events governed by cell cycle genes. Through the preceding three phases of the cell cycle, the cellular morphological structure, size, DNA content, and chromatin state reach a state of readiness for division. During the mitotic phase of the M phase, the cell's nucleus divides, producing two identical sets of chromosomes. Subsequently, in the cytokinetic phase, the cytoplasm divides, resulting in the formation of two daughter cells [3,12,13,14].

There are two types of cell division: proliferative division and asymmetric cell division. Proliferative division results in the production of two daughter cells used for growth, development, and tissue repair. In *A. thaliana*, proliferative division mainly occurs in meristematic tissues such as root tips and shoot tips, where cells continuously divide to promote plant growth. Asymmetric cell division generates two different cells, with one retaining stem cell characteristics while the other proliferates and then differentiates to form specific functions such as stomatal cells, root apical meristem (RAM), lateral root formation, and stem cell division in *A. thaliana* [15,16,17,4].

Endocycle refers to a cell cycle characterized by DNA duplication and a subsequent increase in cell size due to the absence of the cell division step following the initial three phases of the mitotic cycle. Endocycle occurs in all eukaryotic cells, including plants and animals [18,19], but it is

particularly common in the developmental processes of plants[20,21,1,22]. While its precise function remains elusive, researchers have proposed several hypotheses[2].

One proposed function is that the endocycle serves as an adaptive mechanism that may offer protection against certain environmental stresses, including those that could potentially lead to DNA damage. De Veylder et al. (2011) reported that the endocycle may assist in the elongation of the hypocotyl in shade to search for light[23], but its specific physiological role in regulating cell size is unclear[24]. Additionally, under conditions of water scarcity and UV-B stress, the endocycle helps plants maintain growth[23]. This is thought to be facilitated by the redundancy of gene copies in polyploid cells[18,25,26,27].

There are also studies showing that the endocycle can enhance tolerance to abiotic stress, such as under salt stress, as reported for *Sorghum bicolor*[28]. However, Francis et al. (2007) expressed skepticism about the idea that the endocycle helps cells resist adverse conditions, questioning the assumption that cells in an endocycle can revert to the mitotic cycle. This notion currently lacks empirical support[29].

Numerous studies have demonstrated the functional significance of the endocycle in cell development as one of the prerequisites for *A. thaliana* hypocotyl and trichome cells[30,31,21,32], in cell expansion[33,34], and in the regulation of metabolic processes, as polyploidy in maize endosperm can enhance metabolic capacities[18].

Several studies have attempted to establish a strong correlation between polyploidy and cell size, such as comparing the sizes of various types of *A. thaliana* cells with different ploidy levels[27]. However, Beemster et al. (2005) argue that the regulatory role of the endocycle in cell size is not universal[35], and there are also reports that question the role of the endocycle in cell size[23,36,20]. Harashima et al. (2010) propose that the endocycle may enhance cellular growth potential, which could be indicative of a cell's capacity for expansion rather than its current size[37].

1.1.2 Cell Cycle Regulation

Cyclin-dependent kinases (CDKs) and their regulatory cyclin subunits (cyclins) are pivotal proteins in the regulation of the cell cycle. Their interactions are crucial for the phosphorylation of key target proteins, enabling the progression of the cell cycle and ensuring the successful proliferation and division of cells[38,39,40,28,2].

CKDs are a type of protein kinase characterized by the requirement for a cyclin to provide the necessary domain to activate their activity. Building on the foundational work of Joubès et al. (2000), who identified and classified 46 putative CDKs from 23 plant species into five types (CDKA to CDKE)[41], Inzé (2007) significantly expanded the catalog to encompass 152 CDKs across 41 plant species. Inzé's research also introduced three novel CDK categories, as proposed by Menges et al. (2005): CDKF, CDKG, and CKLs[39,40]. This expansion reflects a deeper understanding of the diversity and complexity of CDKs in plants. In the *A. thaliana* model organism, this update resulted in a total of 29 CDKs distributed among eight categories, including the 12 core CDKs (CDKA to CDKE) and 17 newly discovered CDKs. As of 2021, the CDK profile in *A. thaliana*

has remained consistent with these findings[28].

CDKBs are the second largest class of plant CDKs and are specific to plants[42]. Apart from groups C and E, other CDKs are involved in cell cycle regulation in various ways[2].

- **CDKA**: *CDKA1;1*
- **CDKB**: *CDKB1;1*, *CDKB1;2*, *CDKB2;1*, *CDKB2;2*
- **CDKC**: *CDKC;1*, *CDKC;2*
- **CDKD**: *CDKD1;1*, *CDKD1;2*, *CDKD1;3*
- **CDKE**: *CDKE;1*
- **CDKF**: *CDKF;1*
- **CDKG**: *CDKG;1*, *CDKG;2*
- **CDKL**: *CDKL;1* to *CDKL;15*

CDK inhibitors (CKIs) can directly influence CDK activity[1,43], thereby indirectly affecting the cell cycle. In plants, there are two types of CKIs: the interactor/inhibitor of cyclin-dependent kinase/Kip-related protein (ICK/KRP) and the plant-specific Siamese (SIM) and Siamese-related (SMR) proteins[44]. *A. thaliana* has seven KRP genes (*KRP1-KRP7*) and 17 SIM/SMR genes (*SIM* and *SMR1-SMR16*).

Cyclins are characterized by their periodic expression throughout different phases of the cell cycle[45], with significant homology observed between plant and animal cyclins. Plant cyclins share homology with mammalian cyclins of types A, B, C, H, and L, although the subgroups within these categories are not conserved between plants and animals[46,47,48,49]. In animals, there are at least 13 cyclin types (A to L and T), whereas over 100 cyclins have been identified in plants. Whole-genome analysis of *A. thaliana* revealed 50 potential cyclins, which can be classified into nine groups (plus two independent subgroups)[47,39]. Five of these groups have not been found in animals, and only cyclins A, B, D, and H are clearly associated with the cell cycle[1,15].

- **Type A**: *CYCA1;1*, *CYCA1;2*, *CYCA2;1*, *CYCA2;2*, *CYCA2;3*, *CYCA3;2*, *CYCA3;3*, *CYCA3;4*, *CYCA4;2*, *CYCA4;3*
- **Type B**: *CYCB1;1*, *CYCB1;2*, *CYCB1;4*, *CYCB1;5*, *CYCB2;1*, *CYCB2;2*, *CYCB2;3*, *CYCB2;4*, *CYCB2;5*, *CYCB3;1*
- **Type C**: *CYCC1;1*, *CYCC1;2*
- **Type D**: *CYCD1;1*, *CYCD2;1*, *CYCD3;1*, *CYCD3;2*, *CYCD3;3*, *CYCD4;1*, *CYCD4;2*, *CYCD5;1*, *CYCD6;1*, *CYCD7;1*
- **Type H**: *CYCH1;1*
- **Type L**: *CYCL1;1*

- **Type P**: *CYCP1;1*, *CYCP2;1*, *CYCP3;1*, *CYCP3;2 CYCP4;1*, *CYCP4;2*, *CYCP4;3*
- **Type Q**: *CYCQ1;1*
- **Type T**: *CYCT1;1*, *CYCT1;2*, *CYCT1;3*, *CYCT1;4*, *CYCT1;5*
- **Others**: *CYL;1*[50], *SDS*[39]

The progression of the G1 phase and the transition from G1 to S phase depend on the interactions between the highly conserved RETINOBLASTOMA (Rb)-E2F regulatory pathway in plants and the CYCD-CDKA complexes (Figure 1). In the G1 phase, in addition to E2F and RBR proteins, CDKs, and cyclins, dimerization partner (DP) proteins are also involved. E2F and DP form heterodimers in a specific pattern: E2Fa/b pairs with DPa, while E2Fc pairs with DPb[51,52,53,54]. The former combination acts as a cell cycle activator, while the latter functions as a repressor[55,1,15].

RBR controls the cell cycle by binding to the E2F-DP heterodimer, masking its transcriptional activation domain and inhibiting E2F-DP-dependent gene expression. Upon interaction with CYCD-CDKA complexes, phosphorylated RBR is released from the RBR-E2F-DP complex[56,57,56]. The freed E2F-DP heterodimer can then promote the transition from G1 to S phase. Unlike E2Fa and E2Fb, E2Fc primarily acts as an inhibitor during the cell cycle, especially during cell arrest[52,58]. CYCD-CDKA can phosphorylate the E2Fc-DPb heterodimer, lifting the repression on S phase genes and promoting the G1/S transition. Del Pozo et al. (2002) reported CYCD-CDKA phosphorylation of E2Fc, while Shimotohno et al. (2021) noted that CYCD-CDKA phosphorylates both E2Fc and DPb[59,28].

The progression of the G2 phase, M phase, and the transition from G2 to M phase in plants depend on the activities of the CYCA/B/D-CDKA/B complexes (Figure 1). In the G2 phase, activator MYB (Act-MYB) TFs, such as MYB3R4, are activated by phosphorylation through CYCA/B/D-CDKA/B, promoting the expression of G2 and G2/M genes. Interestingly, Act-MYB can also enhance the activity of some mitotic cyclins, creating a positive feedback loop between Act-MYB and CYC-CDK, thereby advancing the cell cycle[60]. Repressor MYB (Rep-MYB) functions to inhibit the cell cycle; however, phosphorylation by CYCA/B/D-CDKA/B destabilizes Rep-MYB, facilitating the G2/M transition[28].

As presented in Figure 1, there are other substances in the cells that indirectly regulate the cell cycle by interacting with CYC and/or CDK. WEE1-like protein kinase (WEE1) inhibits CDK activity by phosphorylating it. This phosphorylation suppresses CDK activation, thereby maintaining the cell in the G2 phase of arrest until the cell is ready to enter mitosis (M phase)[61,62,63]. The anaphase-promoting complex/cyclosome (APC/C), a E3 ubiquitin ligase complex, regulates the cell cyle through degradation of important cel cycle regulators[64,28], such as mitotic cyclins leading the exit of mitosis[65,28]. KRP acts as an inhibitor of CDKs, possibly preventing RBR phosphorylation by inactivating the complex containing *CYCD3*[66], leading to the inhibition of G1/S cycle transition. SIM protein contains sequences similar to motifs from KRP , which interact with

D-type cyclins (*CYCD2;1*, *CYCD3;2*he *CYCD4;1*) and *CDKA;1*, which leads to the inhibition of G1/S transition. The inhibition of *CDKA;1* will also promote endocycle[67,68].

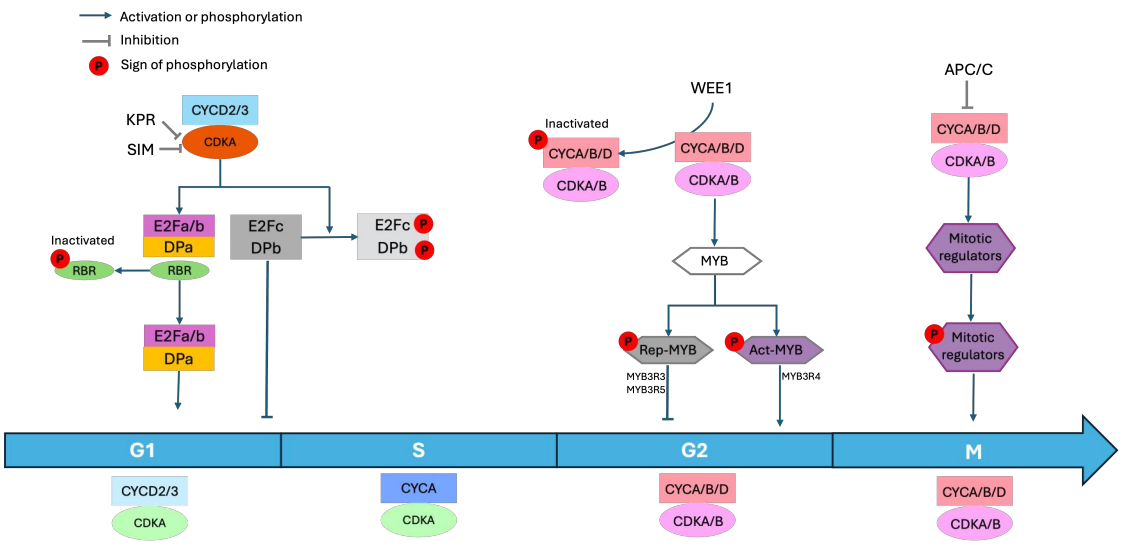


Figure 1: Schematic representation of the mitotic cell cycle in plants involves several key processes. In late G1 phase, the CYCD-CDKA complex phosphorylates the RBR protein, activating the cell cycle-promoting function of the E2Fa/b-DPa heterodimer, thereby advancing the transition from G1 to S phase. The E2Fc-DPb heterodimer, which inhibits S phase genes, is inactivated by phosphorylation through the CYCD-CDKA complex. CKIs such as KRP and SIM also participate in regulating the G1/S transition. In the G2 and M phases, the CYCA/B/D-CDKA/B complexes phosphorylate mitotic regulators like *MYB3R3* and *MYB3R4*, ensuring proper cell cycle progression and successful mitosis. WEE1 and the APC/C inhibit the CYC-CDK complexes, leading the inhibition of cell cyeles.
APC/C, anaphase-promoting complex/cyclosome; CDK, cyclin-dependent kinase; CYC, cyclin; KRP, Kip-related protein; MYB, myeloblastosis; SIM, Siamese

1.1.3 Plant Stress

Plants, as sessile organisms, cannot move to more favorable environments or evade stresses imposed by environmental or anthropogenic factors. These stresses profoundly influence the growth and development of plants. Terrestrial plants face a spectrum of stresses broadly classified into two major categories: biotic and abiotic stresses. Biotic stresses include a range of adversities caused by viruses, microorganisms (such as bacteria, fungi), and macroorganisms (including insects and animals)[69,70,71,72,73]. Abiotic stresses include well-known stressors such as drought[74,75], cold stress[76], heat stress[77,78], salinity[79], soil acidity (or water acidity)[80], ultraviolet radiation intensity[81], and heavy metal accumulation[82]. It is noteworthy that while certain factors may instigate metabolic shifts or alterations in growth patterns in plants, they may not inherently qualify as ”stressors”. For instance, fluctuations in photon flux density or minor variations in temperature or air humidity are common and expected features of the natural environment that plants are adapted to[83].

In response to the stresses, plants have evolved a repertoire of mechanisms to adapt to or cope with environmental pressures. By modulating the expression of cell cycle genes, plants can engage in self-repair and counteract the negative impacts of stress, thereby enhancing their adaptabil-

ity[84,85]. This proactive regulation of cell cycle genes often requires the trade-off of suppressing cell growth and development[84,85]. Under high-temperature conditions, plants may exhibit excessive expression of *ANAC044* and *ANAC085*, thereby inducing cell cycle arrest before division occurs[85]. Additionally, in the presence of stress, abscisic acid (ABA) triggers the expression of the cell cycle inhibitor *KRP*, leading to inhibition of the G1/S transition in rice[86]. Moreover, in maize and *A. thaliana*, various stresses including high temperature, low temperature, and salt stress induces *SIM/SMR*, thereby inhibiting *CDK* activity and consequently suppressing cell division while promoting endocycle[67,87,88,89].

Environmental stresses not only affect the cell cycle but also elicit responses from cycle genes to mitigate such effects. For instance, pathogens can influence cell cycle dynamics, often intertwined with programmed cell death (PCD), a process considered to actively limit pathogen proliferation[90] and regulated by plant homologs of *RBR* and *E2F*[91,92]. Besides, following infection of *A. thaliana* with Cabbage leaf curl virus, *CYCD3;1* and *E2FB* are induced, both serving as activators of the mitotic cycle and endocycle. The resulting polyploidy induced by the endocycle inhibits the virus[93]. Infection of *A. thaliana* with powdery mildew (PW) increases the endocycle and enhances calcium signaling. *MYB3R4*, a mitotic regulator and essential TF for the endocycle, can be targeted for reduced activity and expression as a method to inhibit PW[94].

1.2 Single-Cell Sequencing

1.2.1 What is Single-Cell Sequencing?

Single-cell sequencing is a sophisticated technique that enables the analysis of genomic, transcriptomic, or epigenomic profiles at the resolution of individual cells. However, this basic definition fails to capture the nuanced capabilities and applications of this powerful method. For a better understanding of single-cell sequencing, consider comparing it to bulk sequencing. Bulk sequencing examines cell populations by assessing their combined genomic, transcriptomic, chromatin structure (ATAC-seq[95], Hi-C[96] DNase-seq[97], and flow cytometry[98]), epigenomic profiles[99], and other applications such as DNA damage (RADAR-seq[100]) and TF-TF interactions (ChIP-Seq[101], DNA adenine methyltransferase identification sequencing (DamID-seq)[102]. Single-cell sequencing places particular emphasis on most of these aforementioned topics at the level of individual cells, allowing for detailed insights into cellular heterogeneity and molecular dynamics[103,104,105,106].

In the realm of multicellular organisms, cellular heterogeneity is an undeniable reality[107,108], stemming from a myriad of factors including subtle genomic variations, transcriptional regulations, epigenetic modifications, intercellular interactions, and stochastic influences[109,108]. This heterogeneity manifests in the development of diverse cell types, each exhibiting distinct characteristics and functionalities[110]. Conventional bulk sequencing methodologies simplify complex problems, such as identifying intricate metabolic processes, the transmission of biological signals, and the composition and interaction of transcription networks[108]. Given the ubiquitous nature of cellular heterogeneity, conventional methodologies mask intricacies by averaging genetic infor-

mation such as transcriptional regulations across cell populations when at a fiiner resolution of inspection. While this averaged approach suffices for studying differential gene expression across various tissues or under different conditions, it fails to discern the nuances of gene expression at the level of individual cells or discrete cell subpopulations. Consequently, rare cell populations, such as stem cells residing in the root tip of *A. thaliana*, malignant tumour cells within a tumour nodule[111] or hyper-responsive immune cells within an ostensibly uniform cell population, may be obscured, or altogether overlooked[112,113]. Thus, a more nuanced and comprehensive approach, such as single-cell sequencing, is essential to unveil the real complexity and heterogeneity inherent within biological systems.

Single-cell sequencing is primarily focused on single-cell RNA sequencing (scRNA-seq), complemented by other methodologies tailored to single-cell analyses. These include single-cell methylome and transcriptome sequencing (scM&T-seq)[114], which facilitates the examination of DNA methylation patterns at the single-cell level. Additionally, techniques such as single-cell DamID and Transcriptome sequencing (scDam&T-seq)[115,116], which combines single-cell DamID and cell expression by linear amplification sequencing (CEL-Seq)[117], offer insights into the transcriptional state and protein-DNA interactions within individual cells. Furthermore, single-cell assay for transposase-accessible Chromatin sequencing (scATAC-seq)[118] enables the identification of open chromatin regions, providing valuable information on chromatin accessibility at the single-cell level. These methodologies collectivelyenable studies to dissect the intricacies of cellular heterogeneity and functional diversity with unprecedented resolution.

1.2.2 How to Conduct Single-Cell Sequencing?

The first step of single-cell sequencing is to separate cells from the tissue. Initially, a manual method using a pipette was used to extract a small number of cells[119,120]. However, various technological advancements have since been developed for single-cell extraction, including fluorescence-activated cell sorting (FACS)[121], laser microdissection[122], microfluidics[123,124], and micromanipulation[125,126]. These methods are gradually being phased out due to cumbersome calculations, tedious operations, and low throughput. They were commonly used in the early stages of single-cell sequencing on platforms such as SMART-seq[127], SMART-seq2[128], MATQ-seq[129], CEL-seq[117], and MARS seq[130]. These methods involve physically isolating individual cells, constructing sequencing libraries for each, and sequencing them independently. This ensures high accuracy and detail for each cell but can be labor-intensive and costly.

The second strategy leverages a microfluidic system or small container, often a oil droplet, to encapsulate single cells and form autonomous micro-reactors. Inside each micro-reactor, specific reagents, enzymes, and a small bead are introduced for sequential processes such as cell lysis and subsequent Polymerase Chain Reaction (PCR) amplification. The bead is adorned with numerous specially designed nucleic acid chains, incorporating a primer, a cell barcode, a unique molecular identifier (UMI), and poly(dT). Notably, the cell barcode on each bead is consistent, uniquely identifying each cell, while the UMIs, vary in number, label individual mRNA molecules[131]. When

a cell lyses within the micro-reactor, its mRNA fragments combine with the free nucleic acid chain on the bead, enabling PCR amplification to generate cDNA. The presence of barcode and UMI ensures each amplified cDNA carries a unique identifier, with the barcode indicating the cell source. After removal of the oil, cDNAs from all cells are collected, followed by library construction, sequencing, and the acquisition of the single-cell expression matrix through barcode identification[132]. Platforms with implementations of this strategy include Drop-seq[133] and 10x Genomics[134], which are the current state-of-the-art.

1.2.3 Analysis of Single-Cell RNA Sequencing Data

The first step in the analysis of single-cell sequencing data is the reduction of data dimensionality followed by data clustering. In the process of single-cell analyses, the identification and characterization of cell types, as well as the visualization of cell populations, hinge upon the application of clustering methodologies. The principle behind clustering methodologies underlying mathematical principle involves computing the distances between different cells and grouping them based on these distances. The single-cell expression data can be represented as a two-dimensional table comprising gene expression profiles across individual cells. However, the inherent high dimensionality, often considered as the " curse of dimensionality," and the substantially sparse nature of the expression matrix render the data challenging to interpret and utilize effectively. Moreover, in high-dimensional datasets, the distances between cells tend to be minuscule.

The challenges posed by high-dimensional data necessitates employing mathematical techniques to initially reduce dimensionality. As a result, the number of features is reduced, rendering them independent of one another. Within the realm of single-cell analyses, common methods for dimensionality reduction encompass Principal Component Analysis (PCA), t-distributed Stochastic Neighbor Embedding (tSNE), and Uniform Manifold Approximation and Projection (UMAP). Considering PCA as a quintessential example, the process entails three sequential steps. Initially, highly variable genes (HVGs) are identified as features, aimed at reducing the impact of noise and facilitating the capture of biological variance, thereby enhancing the accuracy and efficiency of subsequent analyses. Then, HVGs are utilized to compute principal components (PCs). These PCs constitute newly established orthogonal variables, representing linear combinations of the original data in directions of maximum variance, with the objective of retaining maximal variance in the dataset. Finally, the appropriate number of PCs is determined, necessitating a balance between preserving biological information (more PCs) and mitigating noise (fewer PCs) as much as possible.

Following dimensionality reduction, the data can undergo distance calculation utilizing various metrics, including Euclidean distance, cosine similarity, Pearson' s correlation, and Spearman' s correlation. These distance metrics are subsequently integrated with clustering techniques such as K-means or hierarchical clustering to partition cells into distinct clusters.

1.2.4 Application of Single-Cell RNA Sequencing

Single-cell technology allows researchers to study gene expression at the individual cell level, providing a higher level of detail. Some common questions addressed whith single-cell studies include:

- **Resolving Cellular Heterogeneity** Critiques of bulk sequencing stem from its tendency to obscure cellular distinctions. Single-cell sequencing, however, offers a solution to this limitation. As demonstrated in cancer research, the heterogeneity among cancer cells can be discerned and quantified by analyzing cellular expression data individually[135,136,137,138]. For example, single-cell sequencing provides novel insights into tumor initiation, intratumoral heterogeneity, cancer metastasis, drug resistance, and the tumor microenvironment[139].

- **Cell Type Identification and Annotation**: Single-cell expression data are utilized in clustering analyses to group cells with similar or identical expression patterns. These groups often reflect specific cell types or biological functions. By comparing the clustering outcomes with known cell types or marker genes, researchers can annotate cell populations. The annotated populations can subsequently undergo further analyses to elucidate their biological functions and states. This includes examining highly expressed genes and conducting differential analysis to identify specific marker genes indicative of particular cell types and their functions[140,141].

- **Inferring Cell Differentiation**: Single-cell trajectory analysis is frequently employed in examining cellular developmental or differentiation pathways. Typical methodologies encompass pseudo-time and RNA velocity analyses[142,143].

1.3 Regulation of Gene Expression at Transcriptional Level

1.3.1 Regulation Mechanisms of Transcription Factors

DNA serves as the repository of genetic information, perpetuated through self-replication, and transcribed into mRNA, which is subsequently translated into proteins that execute various biological functions.

The regulation of mRNA transcription constitutes a central focus of the present analysis. In essence, transcription involves the synthesis of a new single-stranded RNA molecule, which is complementary to the DNA template strand, albeit by the substitution of thymine (T) by uracil (U). Prior to transcription, the DNA double helix near the target gene undergoes unfolding. Subsequently, RNA polymerases bind to the DNA template strand and commence transcription, progressing from the 5' untranslated region (5' UTR) to the 3' untranslated region (3' UTR). The binding region of RNA polymerases, known as the promoter region, dictates the initiation of transcription. Functionally, a promoter is a DNA sequence situated upstream of the transcription start site (TSS) of a gene and may span several thousand base pairs (bp). Unlike bacterial transcription, wherein RNA polymerase directly attaches to the promoter, eukaryotic transcription involves

auxiliary proteins known as TFs that bind to specific sequences within the promoter[144]. These TFs can form stable or transient associations with the promoter to initiate transcription.

This work focuses on the two core functions of TFs.: first, their ability to recognize and bind to short, specific DNA sequences, known as motifs or transcription factor binding sites [TFBSs]) within promoter regions; and second, their capacity to recruit proteins involved in the regulation of transcription. These essential roles position TFs as the key mediators between genetic information encoded in DNA and its functional output. TFs determines the content, timing, and trajectory of the expression of entire gene expression within regulation networks. By orchestrating the activation and suppression of sets of genes, they serve as the initial step in unveiling DNA-encoded information, forming the cornerstone of genetic information transmission within organisms[145]. For example, TFs play critical roles in determining cell fate and controlling developmental patterns[146]. Lee et al. (2013) reported that a small subset of TFs can induce the transformation of various cell types into induced pluripotent stem cells (iPSCs)[147].

The preceding description of TFs may inadvertently convey a bias toward their role as activators. Yet, TFs can also function as transcriptional repressors. For instance, certain TFs exhibit a strong affinity for specific DNA sequences, thereby preventing potential binding by other TFs[148,149]. Consequently, TFs are often categorized as "activators" and "repressors." However, such classifications are not immutable; TFs may be influenced by other regulatory proteins, binding sites, or genes, and can exhibit entirely opposite effects at different loci[150]. Phosphorylation, for example, can convert CCAAT/Enhancer Binding Protein β (*C/EBP*β) from a repressor to an activator[151]. Furthermore, post-translational modifications can also elicit a reversal in TF regulatory directionality. Specifically, the processing of Specificity Protein 3 (SP3) by small ubiquitin-related modifier (SUMO) transforms it into a repressor, while acetylation can promote its role as an activator[152].

In potatoes, the expression direction of the pathogenesis-related-10a (*PR-10a*) gene can also be altered by different TFs. The silencer element binding factor (*SEBF*) and the Pto interacting protein 4 (*Pti4*) act simultaneously on the *PR-10a* promoter through *cis*-regulatory elements, thereby inhibiting its expression. Conversely, *Solanum tuberosum Why1* (*StWhy1*), functioning as a transcriptional activator, can induce the expression of *PR-10a*[150].

1.3.2 Identification of Transcription Factor Binding Site and Motif

When discussing TF and TFBS, the concept of a motif is often mentioned. In bioinformatics, a motif refers to a short sequence pattern or feature that repeatedly occurs in a sequence. In this work, I will use motif to represent TFBS. A commonly used motif model is the probability model. Motifs are represented by "position weight matrices" (PWMs), which indicate the frequency of each nucleotide (A, C, G, and T) at each position within the motif.

In the early 2000s, identifying motifs within sequences was considerably challenging[153] due to short length of sequencing reads and the degeneracy of motifs[154]. The degeneracy means identical or different TFs can bind to motifs with slight differences. Besides, the versatility and flexibility of

a TF determine the diversity of TF bindings, meaning a TF can bind to some atypical motifs. It should be noted that the degeneracy of motifs and diversity of TF binding are two different things. The former emphasizes the variability and substitutability of motifs, while the latter focuses on the characteristics of TFs themselves. Additionally, TFs may not be conserved in different species[154]. The average length of motifs is 9.9 bp for eukaryotes and 15.9 bp for prokaryotes[155]. Sequencing technology generates reads of limited length, which limits the contiguous length of DNA sequence observed. This makes it difficult to determine the precise locations of motifs on sequences, especially in highly repetitive genome regions. Badis et al. (2009) systematically investigated the binding specificities of 104 binding proteins in mouse genome and found that almost all proteins have their unique preferred sequences and about half of them recognized multiple motifs[156]. Similar situations were found in *A. thaliana*[157].

The discovery and identification of binding sites of proteins were typically achieved through methods such as DNase footprinting[158] or electrophoretic mobility shift assays (EMSA)[159,160]. These techniques facilitated the discovery of specific binding proteins using approaches like N-terminal peptide sequencing, or yeast one-hybrid screening[145]. More comprehensive methods for identifying TFBSs were developed, such as DNA affinity purification sequencing (DAP-seq)[161], protein binding microarrays (PBMs)[162], systematic evolution of ligands through exponential enrichment (SELEX) methods[163], high-throughput SELEX (HT-SELEX)[164], HiTS FLIP[165], spec-seq[166], mechanically induced trapping of molecular interactions (MITOMI)[167], bacterial one-hybrid selections (B1H)[168,169], chromatin immunoprecipitation coupled with next-generation sequencing (ChIP-seq)[170,171], and an improved version known as ChIP-exo[172].

Despite this methodological progress, the detection of motifs remains challenging. In addition to experimental approaches, it is also possible to predict motifs on a sequence through bioinformatics. Hashim et al. (2019) reviewed 'motif discovery' tools and benchmarked them according to enumerative, probabilistic, nature inspired, and combinatorial approaches. Overall, enumeration approaches can discover all motifs through the method of exhaustive search, but at the cost of time and complex parameter design. MEME suite is a specific implementation of the probabilistic method and is currently the most common 'motif discovery' tool[173]. Given the constraints of space, it is not feasible to include an exhaustive overview of the numerous tools within this work. For a comprehensive exploration, readers are encouraged to consult the studies conducted by Hashim et al. (2019)[174] and Castellana et al.[175].

Several motif databases have been established through experimental and bioinformatic studies. Relevant databases for plants, particularly *A. thaliana*, include those from Franco-Zorrilla et al. (2014)[157], Jaspar Plants Non-Redundant (2022)[176], Plant Cistrome DB[161], PlantTFDB[177], and CIS BP2[178]. These databases typically store motif information in Position Weight Matrix (PWM) format within ".meme" files.

1.3.3 Cooperative Transcription Factor Binding

Prokaryotic TFs can recognize and bind to longer motifs[179], and the induction or inhibition of gene expression can be accomplished through one single TF binding. TFs exhibit 1,000-fold greater binding affinities for specific DNA sequences compared to others. Such strong affinity, combined with the enclosure effect (which prevents other TFs from binding), leads to dominant transcriptional regulation by a single TF[145]. However, the situation is more complex in eukaryotes because the binding site specificity is rather loose[180,181], there are more regulatory TFs, and the transcriptional regulation is more sophisticated. The complex regulation network is built through TF binding to multitudinous genes on the genome, while many binding sites are accessible to many TFs simultaneously. The regulation of transcription depends on a process called cooperative TF binding. It includes:

1. The bicorporate TF-TF complex is a precondition for binding to DNA. For example, the affinity of E2F family TFs (*E2F1-6* and *DP1-4*) does not support their direct attachment to DNA, but once *E2F/DP* forms a complex, it exhibits strong DNA affinity[182].

2. DNA-assisted TF-TF interactions require prior involvement of DNA binding because there is no direct integrating affinity between the participating TFs; after TFs bind to DNA, they can interact with each other and then perform their transcriptional functions[183,184,185,186,187,188].

3. DNA-mediated TF-TF complex interactions change the shape or dynamic parameters of DNA to some extent, thereby allowing another TF to physically interact with TFs for cooperative transcriptional regulation. For example, Acyl-CoA metabolism-associated-like 1 (*AAML1*) and Runt-related TF 1 (*RUNx1*) do not directly interact with each other, but only come together through DNA binding[189,190].

4. Indirect cooperativity involves nucleosome intervention, which has effects like direct binding by TFs and can facilitate a very broad range of interactions[191,192,193,194,195,196].

Cooperative binding introduces two important concepts, homotypic and heterotypic motifs. Homotypic motifs refer to the set of motifs that bind with the same TF appearing multiple times on the same sequence, while heterotypic motifs naturally refer to the set of motifs that bind to different TFs within the same sequence. In this work, the focus lies on heterotypic motifs involving a maximum of two distinct TFs to examine the distribution pattern of pairs of different TFs and their motifs within the same sequence.

1.4 Project Aims

In the early years of the 21st century, much effort was devoted into the identification of cell cycle-regulating genes using cell cycle-synchronized cell cultures[197,198,199,200]. While these efforts corroborated each other, they did not lead to significant expansion of the gene list. Notably, few genes related to the G2 phase have been identified, and the catalog of cycle-related genes has seen

little update in the subsequent 20 years. Critically, synchronized cell cultures do not accurately simulate the environment, including intercellular interactions and spatiotemporal effects, which are crucial for root research.

This work aims to identify a comprehensive set of root-specific cell cycle genes in *A. thaliana* and to elucidate their transcriptional regulatory mechanisms through motif pair binding analysis.

1. A new method to discover more cell cycle genes in different developmental zones of *A. thaliana* root tips.
   (a) Cells in single-cell RNA seq data of *A. thaliana* root tips were separated according to their developmental zones.
   (b) Cells within developmental zones were assigned to a cell cycle based on their expression of reference cell cycle genes.
   (c) Differential expression analyses were conducted among cells in the cell cycle to identify new cell cycle genes.
2. Refinement and enhancement of PMET.
   (a) The development of the established PMET was refined to include a user-friendly PMET online version and new features, such as studying the distribution of motif pairs at different promoter positions.
   (b) Gene lists of *A. thaliana* under different stress conditions were exploited to fine-tune the optimal parameters of PMET and study the transcriptional regulatory mechanism of down-regulated genes under heat stress.
   (c) PMET searched for motif pairs on selected novel cell cycle lists to explain the transcriptional regulation of cycle genes through the most enriched TFs.

These endeavors discovered distinct new cell cycle gene lists and cell cycle-specific motif pairs, which contribute to understanding the cell cycle regulatory mechanisms.

2 Paired Motif Enrichment Tool

2.1 Introduction and Motivation

2.1.1 Binding of Transcription Factors

TFs regulate gene expression by binding to cognate DNA motif sequences in *cis*-regulatory modules (CRMs) located upstream of genes[201], which contain multiple types of motif. During the process of transcription process, TFs, as proteins, play a crucial role. Their range of influence can encompass several thousand nucleotides upstream and downstream of a gene. TFs can bind to CRMs, which are functional units within the genome that regulate gene expression. They guide RNA polymerase to the binding sites and initiate transcription. Additionally, TFs can stimulate or suppress gene expression by interacting with enhancer and/or promoter sequences.

The key to a TF's binding to a gene lies in its ability to specifically recognize a variety of short, similar sequences called motifs (8 to 21 nucleotides)[202]. Motifs serve as the footholds for TFs on genes and can be identified *in vivo* or through *in vitro* simulations, often represented as position weight matrices (PWMs), which indicate the probability of the occurrence of any nucleotide at each position in the motif[203].

2.1.2 Cooperative Interaction of Multiple Transcription Factors

Beyond individual TF binding, the cooperative interaction of multiple TFs plays a pivotal role in shaping functional dynamics of genes. Cooperative TFs can form homo- or hetero-dimers, as illustrated in Figure 2. Such a combination of TFs provides an extra layer of specificity and complexity to the gene regulation network[193]. In a homodimer form, two identical TFs combine to create a homotypic complex that recognizes and binds to a homotypic motifs - consecutive or repetitive sequences on DNA with same or similar base patterns. A heterodimer, consisting of two different interacting TFs, recognizes and binds to heterotypic motifs, which are located in adjacent and proximal positions on DNA with different base patterns.

Experimental and bioinformatic studies have consistently demonstrated the tendency of TFBSs to cluster within DNA sequences. Homotypic clusters are localized regions enriched with repetitive binding sites for a single TF in close proximity, while heterotypic clusters, in contrast, involve multiple TFs. CRMs encompass homotypic or heterotypic clusters of motifs that serve as binding sites for regulatory proteins[204]. Homotypic clusters of motifs, characterized by repeated occurrence of motifs, play a significant role in gene regulation. It has been widely observed that a higher frequency of the motif within a homotypic cluster can enhance the binding of TFs to the associated genomic region[205,206,207,208,209]. Heterotypic clusters which incorporate multiple motifs, offer a platform for combinatorial regulation of gene expression, given that distinct TFs may exert cooperative or antagonistic effects on transcription process[204].

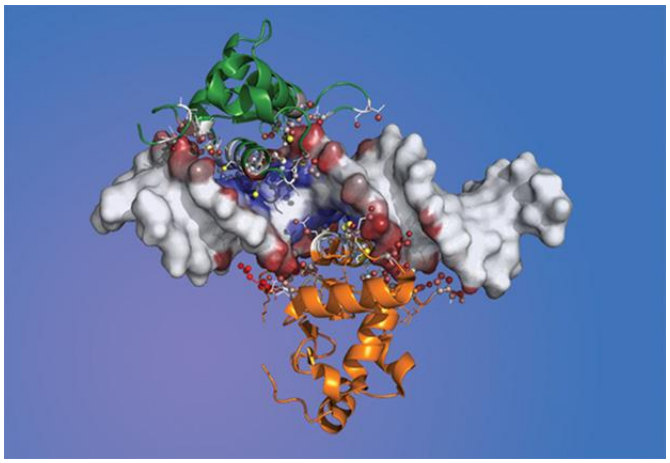


Figure 2: Protein structure of forkhead (green) and erythroblast transformation specific (ETS) (orange) TF binding to DNA (grey) driven by an ETS protein residue (red). IMAGE: Ignacio Ibarra/EMBL.

2.1.3 Experimental Methods for Detecting Transcription Factor Interaction

Current studies of inferring gene regulatory networks mainly rely on gene expression data to map the regulatory relationships between TFs and target genes. Although relying on gene expression data to infer gene regulation has revealed some regulatory mechanisms, this approach is limited and prone to generating false positives due to the lack of consideration of TF interactions[210]. To addres this limitation, researchers have used multiple experiments techniques. Chromatin immunoprecipitation sequencing (ChIP-seq), is an advanced technique for identifying the DNA-binding sites of specific proteins across the entire genome. Essentially, ChIP-seq captures a snapshot of the specific genomic locations where a protein either actively bound or has the potential to bind. While ChIP-seq is primarily used to identify the binding sites of individual TFs, it also holds promise for elucidating TF-TF interactions by comparing binding sites from two separate ChIP assays. If considerable overlap is observed between the binding sites of two TFs, it may suggest a potential cooperative interaction. ChIP-seq can be expanded to study TF-TF interactions through exploiting sequential ChIP (ChIP-re-ChIP). This method involves a two-step immunoprecipitation process, to identify TFs that co-occupy the same genomic region. On the other hand, proximity-dependent biotinylation (BioID) is particularly effective for detecting transient TF-TF interactions, especially when dealing with weak, unstable or transiently interacting proteins.

Although experimental assays offer valuable insights, they are not without limitations. They often involve labor-intensive procedures, rely on costly reagents and equipment; and can be constrained by experimental conditions. Furthermore, detecting certain cooperative interactions between TFs are can be challenging. Some of these interactions might only manifest under physiological or pathological conditions; this makes their detection and replication even harder.

2.1.4 Bioinformatic Methods for Detecting Transcription Factor Interaction

Given these challenges, bioinformatic methods have become indispensable, effectively addressing many of aforementioned limitations. From a biological standpoint, analyzing paired motifs is akin to investigating potential TF-TF interactions, as each TF is associated with specific motifs. From a computational perspective, searching and identifying motifs within sequences is both efficient and straightforward. Most importantly, bioinformatic techniques facilitate a more profound understanding of regulatory contexts. For instance, motif searches can be specifically tailored to the promoter or enhancer regions of genes within a particular cell type, thereby enhancing the understanding the biological significance and implications of TF-TF interactions.

Cis-element Cluster Finder (Cister)[211] utilizes a Hidden Markov Model (HMM) to detect and delineate *cis*-element clusters in DNA sequences. This approach is designed to mitigate the challenge of excessive and presumably false positive predictions often encountered when using position specific scoring matrices (PSSMs), such as MatInspector[212], to search for signature sequence patterns or motifs. However, Cister is not without limitations. It operates on the presumption that all types of *cis*-elements occur with equal frequency, a necessary assumption due to the scarcity of experimental data to reliably estimate their relative frequencies. There is a prior assumption that all types of *cis*-elements have the same frequencies due to the lack of experimental data to estimate reliably the relative frequencies. Furthermore, the application of Cister to full-size genomes can result in compromised precision and practicability[211].

Cluster-Buster, an advanced iteration of Cister, is engineered to identify clusters of pre-specified motifs within DNA sequences. It utilizes a probabilistic model to discern regions in the sequence that statistically align with a motif cluster model, distinguishing them from 'background DNA'. Subsequently, forward and Backward algorithms are applied to refine the regions identified. However, Cluster-Buster is far from being perfect. The tool is not applicable to sequences longer than a few kilobases in length and requires the initial hypothesis that a particular set of motifs is clustered[213].

Other computational tools, such as Co-Bind[214], LOGOS[215], ModuleSearcher[216], CisModule[217], and CMA[218], confront challenges akin to the ones previously discussed. Some of these tools require additional parameters, including the expected number of modules/clusters, anticipated motif spacing, and presumed distances between modules. Central to these tools is the presumption that cooperative TF binding sites are situated in close vicinity to one another. However, recent studies have challenged this assumption. Specifically, cooperative TFs have been observed to exhibit two distinct interaction patterns: short-range interactions with one orientation favoring a particular orientation (of TFs); and long-range interactions with alternative orientations[193]. This concept aligns with research on paired motifs that drive enhancer-promoter (EP) communications. While paired motifs might be present in EP pairs, the exact spacing between enhancers and promoters varies. This variability is attributed to the positioning of enhancers, which can be located either upstream or downstream of the target gene, or even nestled within introns[219].

Combinatorial code analysis tool (CCAT) is a tool to predict genome-wide co-binding among

TFs. Unlike traditional methods, which primarily rely on sequence information, CCAT integrates publicly available TF binding specificity data with DNaseI chromatin accessibility data[220]. Similarly, TFcoop predicts cooperative TF binding sites, but it achieves this by aligning the nucleotide content of the sequences with the binding affinities of all identified cooperating TFs[221]. While both of these combination-wise approaches demonstrate efficacy, the accessibility of additional data can impair the applicability and adaptability of the methods.

To Address the complexities and limitations of existing methodologies, Rich et al., a team from Prof. Dr. Sascha Ott's research group, introduced PMET, a novel computational framework designed to elucidate the co-localization patterns of TF binding motifs within specified gene sets[222]. Building upon this foundation, this study aims to enhance and expand capabilites of PMET, applying it to dissect the regulatory dynamics of genes associated with cell cycle and heat stress response.

2.2 PMET Methodology

PMET is a powerful computational tool designed to analyze the paired binding site of TFs or co-occurrence of motifs within gene promoter regions. It integrates two types of motifs: homotypic motifs, which refer to the recurrence of the same TF binding site, and heterotypic motifs, which involve binding sites of different TFs. Such reoccurrence of the motif can enhance the stability or binding affinity of TF binding, thereby positively effecting gene expression. These TFs interact with one another due to functional similarity or physical proximity, to fulfill their biological functions for more complex gene regulation. Leveraging its bespoke PMET indexing algorithm, the tool meticulously scans and indexes promoter sequences for individual motifs, accessing their presence and calculating a binomial p-value to quantify their statistical significance. Subsequently, PMET investigates the enrichment of co-occurrence of motifs in promoters of genes, with a particular focus on clusters of homotypic and heterotypic motifs (Figure 3).

2.2.1 Genome-wide Motif Indexing

1. **Sorting annotation**: Sorting sequence coordinates in .gff3 files with "GFF3sort" facilitates efficient gene sequence retrieval and analysis. such as quickly locating the nearest CDS to the transcription start site (TSS).

2. **Extracting promoter**: Inferring promoter regions starting from TSS in the genomic annotation file and extracting them using 'bedtools flank provides sequences to search for motifs.

3. **Removing overlapping promoter chunks**: Removing overlapping promoter with "bedtools subtract" enhances the clarity and independence of regulatory elements for downstream analysis and experimental accuracy;

4. **Adding 5'UTR**: Locating the start position of the CDS near the TSS, approximating the end of the 5'UTR, and adding it to the promoter as needed helps understand translation regulation.

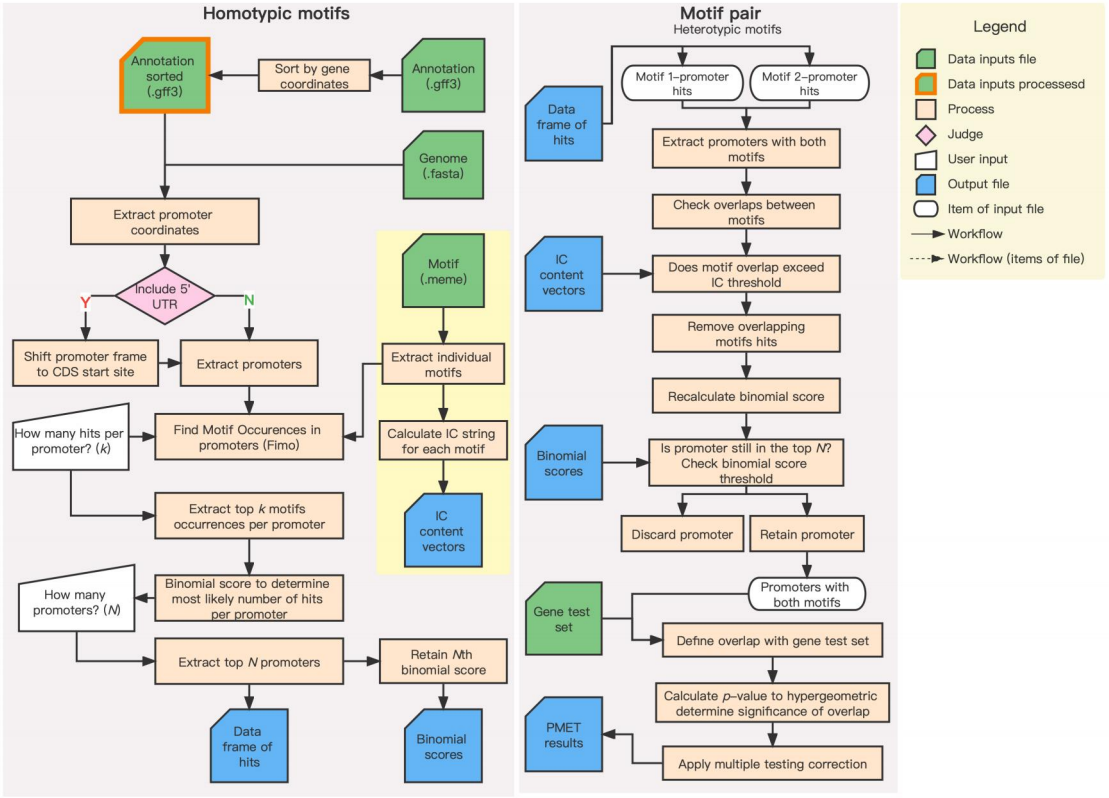


Figure 3: Workflow of PMET procedures to identify homotypic and heterotypic motifs.

5. **Calculating information content of motifs**: Information content measures the conservation of nucleotides in motifs, aiding in PMET pairing;

6. **Finding homotypic motif matches**: FIMO from the MEME Suite identifies homotypic motif clusters by scanning sequences and assessing the significance of occurrences with p-values.

7. **Indexing homotypic motif matches**: Evaluating the statistical significance of observing up to k (user-defined) motif matches within a promoter. This work calculates the complement of the cumulative distribution function (CDF) for the binomial distribution for a series of hit counts using the geometric mean of the observed motif matches' p-values, represented as Formula 1.

$$Score_n = 1 - \text{BinomialCDF}(n; N, gm) = \sum_{i=0}^{n} \binom{N}{i} gm^i (1-gm)^{N-i} \tag{1}$$

where $n \leq k$.

This approach identifies the hit count n, least likely to occur by chance, providing a metric (lowestScore) to assess the affinity of homotypic motifs in promoters. The optimal number of hits per motif is determined using the lowestScore, because motif mathes with p-value smaller than lowestScore will be removed. And the lowestScore will also serve as a measure

of the current motif's affinity to a promoter, for ranking promoters.

N denotes the total number of potential matches for a motif from both directions on a promoter, given by Formula 2, and gm is the geometric mean of the p-values for k motif matches, calculated as Formula 3. The lowestScore is defined as Formula 4.

$$N = 2 \times (\text{promoterLength} - \text{motifLength} + 1) \tag{2}$$

$$gm = \sqrt[k]{p_1 p_2 \cdots p_k} \tag{3}$$

where p_i is the p-value of the ith hit.

$$\text{lowestScore} = \min(Score_n) \tag{4}$$

8. **Selecting promoters with the highest affinity for motif**: LowestScore quantifies and ranks promoter affinity, with an optimal promoter count of 5,000 balancing hit detection and computational efficiency.

2.2.2 Cooperative Motif Pairing

The cooperative Motif Pairing is designed to detect heterotypic motifs that are distinct and co-occur on the same promoter—and to quantify the enrichment of these co-occurrences.

1. **Checking overlaps of Co-Occurrences**: Eliminating unqualified motifs. Partial overlaps of heterotypic motifs are permitted, as supported by studies indicating that TF-TF interactions involve significant binding overlaps[193,223]. However, complete overlaps are permissible (see Figure 4c). When co-occurring motifs exhibit overlaps, as shown in Figure 4b, the information content (IC) of overlap is calculated using the IC values stored during the PMET indexing. This calculated IC for the overlap is then compared against a user-defined threshold (ICthreshold) to determine if the co-occurrence ia retained. This approach ensures that only biologically plausible motif pairs are selected, enhancing the reliability of PMET's insights into TF interactions and regulatory mechanisms.

2. **Hypergeometric Test**: A hypergeometric test is employed to evaluate the enrichment of genes with both motifs within a specific gene cluster, calculating the probability of such co-occurrences as compared to random distribution across the genome.

2.3 PMET: From Implementation to Application

2.3.1 Software Development Pipeline

PMET, initially developed by Rich-Griffin et al. (2020), primarily identifies paired TF motif binding in upstream sequences of TSS[224]. PMET has been further expanded to consider motif pairs

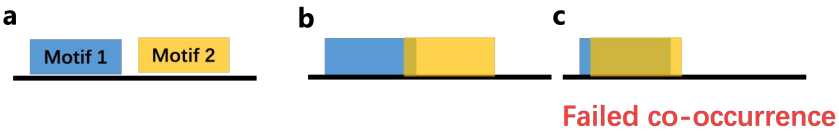


Figure 4: Overview of different degrees of motif co-occurrence and overlap. (a) No overlap between the motif 1 and the motif 2, suggesting a possible co-regulatory interaction between TFs; (b) Edge-to-edge overlap of motifs, which still allows for co-regulatory TF interaction; (c) The motif 1 and the motif 2 significantly or completely overlap, indicating a lack of distinct binding sites for each motif.

in other sequences such as UTR and mRNA, uncovering potentially overlooked regulatory mechanisms. Additionally, the differential enrichment of motif pairs in different portions of regions of upstream sequences of TSS was investigated. PMET retains its command-line interface, providing professional users with flexibility, customizability, scalability, and automation capabilities. Furthermore, a web-based Shiny app is available for interactive data exploration, streamlining PMET utilization and enhancing user engagement through a responsive interface while ensuring robust data handling.

Command Line Interface (CLI) For users with basic computational skills, a user-friendly Bash script is provided, complete with an intuitive guide and colorful hits for ease of use (Figure 5).

a

```
└─ scripts/PMETindex_promoters.sh
No arguments supplied
Usage: PMETindexgenome [options] <genome> <gff3> <memefile>


Options:
  -r <PMETindex_path>       Full path of python scripts.                 Required.
  -i <gff3_identifier>      Gene identifier in gff3 file.                Required.
  -o <output_directory>     Output directory for results.                Default: Current Directory
  -n <topn>                 Top n promoter hits per motif.               Default: 5000
  -k <maxk>                 Max motif hits within each promoter.         Default: 5
  -p <promoter_length>      Length of promoter in bp for motif detection. Default: 1000
  -v <include_overlaps>     Handle promoter overlaps with sequences.     Default: AllowOverlap
  -u <include_UTR>          Include 5' UTR sequence?                     Default: No
  -f <fimo_threshold>       Minimum quality for hits by fimo.            Default: 0.05
  -t <threads>              Number of threads.                           Default: 1
  -d <delete>               Delete unnecessary files.                    Default: No


Use this script to create PMET index for Paired Motif Enrichment Test using genome files.
```

b

```
└─ scripts/pmetParallel
          Input parameters
-----------------------------------
Input Directory:                    ./
Gene list file:                     input.txt
IC threshold:                       4
Threads used                        16
Promoter lengths:                   promoter_lengths.txt
Binomial threshold values:          binomial_thresholds.txt
Motif IC values:                    IC.txt
Fimo files:                         ./fimohits/
Output directory:                   ./
-----------------------------------

Reading input files...
Error: Cannot open file ./promoter_lengths.txt
```

Figure 5: Diagram of PMET command-line usage and options. Command-line output highlights various arguments in distinctive colors with indentation for clarity. (a) PMET indexing, which involves searching for homotypic motifs, and (b) PMET pairing, which involves searching for heterotypic motifs, are demonstrated.

An illustrative display of PMET's command-line usage and options. The interface highlights various arguments in distinctive colors for clarity. This organized presentation aids users in efficiently configuring the PMET tool for creating indexes tailored for the Paired Motif Enrichment Test using specified genome files.

1. **Performance and Efficiency**: Generally, CLI offers better performance compared to graphical user interfaces (GUIs) or web-based tools, as it allocates all computational power to the task at hand. In contrast, GUIs or web-based environments consume additional system resources, including CPU power and memory, which is particularly critical for memory-intensive bioinformatics tools with massive data.

2. **Flexibility and Customizability**: GUIs or web-based tools may limit users with their structured menus, potentially causing confusion with multiple levels of options. CLI, on the other hand, presents options in a straightforward manner, allowing for more precise control of PMET's functionalities. PMET's CLI can easily incorporate new functions or parameters, enhancing its adaptability for active research environments. Users can also modify and compile the code to customize PMET, further integrating it into broader bioinformatics workflows.

3. **Automation**: CLI has a significant advantage in terms of automation, facilitating the integration of tools into loops or parallel processes in bioinformatics pipelines. This capability is particularly valuable for benchmarking different tools or parameter sets.

4. **Scalability**: PMET's CLI is scalable, suitable for debugging on personal computers or large-scale analysis on clusters. It can leverage multi-core processing to expedite tasks, and physical memory limitations can be addressed by adding more memory or transitioning to high-performance computing (HPC) environments.

5. **Quick Access to Data**: Biological data, typically large and stored locally, can be directly accessed via CLI without the need for data uploading.

6. **Demonstration with real data**: To improve user comprehension and engagement, a step-by-step demonstration of PMET's command-line interface (CLI) is provided using real data and executable commands. This demonstration is intended to support users who may not initially possess a comprehensive understanding of PMET's functionalities, allowing them to build familiarity progressively. Executing the provided commands yields informative console outputs that guide users through the analysis process, facilitating both learning and interpretation. Ultimately, the demonstration produces biologically meaningful results, reinforcing users' conceptual understanding through PMET practical application.

PMET in a Shiny app In addition to the CLI, an integrated C++ program with an R Shiny app is deployed on a web server. This setup facilitates user interaction with the program via a web browser for an intuitive experience.

1. **Usability and Accessibility**: The PMET Shiny app provides a user-friendly interface, enabling users to interact with the program easily. It allows users to adjust program parameters without the need for complex command-line knowledge. With visual cues and textual instructions, users can efficiently provide accurate input and parameters. This synergy ensures that the high-performance attributes of C++ are maintained while benefiting from Shiny's interactive features.

2. **Advantages of web-based tool**: Using PMET Shiny app eliminates the need for local installation and compilation of dependencies. This is particularly advantageous for users not familiar with bash programming or those without the resources for intensive computational tasks.

3. **Rendering capabilities** The inclusion of a visual heatmap not only condenses the discoveries into an easily digestible format but also enriches the text-based output. It makes finding biological association easier.

In the page of 'Run job', users have the options to select their preferred mode for conducting PMET analysis. In the mode of **Promoters (Pre-computed species)** mode, this work has prepared a comprehensive and pre-processed dataset encompassing promoters extracted from 21 plant species and conducted homotypic motifs searching or PMET indexing within these promoters across five well-established motif databases, using default parameters. For general purposes, The expended the plant list includes *A. thaliana*, *Brachypodium distachyon*, *Brassica napus*, *Glycine max*, *Hordeum vulgare*, *Hordeum vulgare Morex V3*, *Hordeum vulgare goldenpromise*, *Hordeumvulgare R1*, *Hordeum vulgare v082214v1*, *Medicago truncatula*, *Oryza sativa indica 9311*, *Oryza sativa indica IR8*, *Oryza sativa indica MH63*, *Oryza sativa indica ZS97*, *Oryza sativa japonica Ensembl*, *oryza sativa japonica ensembl*, *Oryza sativa japonica Kitaake*, *Oryza sativa japonica Nipponbare*, *Oryza sativa japonica V7.1*, *Solanum lycopersicum*, *Solanum tuberosum*, *Triticum aestivum* and *Zea mays*. The intuitive and interactive selections not only offer species and motif databases of interest but also reduce the running time by 95%. After uploading gene clusters for specific regulatory, PMET pairing is performed to search motif pairs based on results of (Figure 6). The **Promoters** mode fully unleashes the power and versatility of PMET Shiny, thanks to its compatibility to work with the genomes of all species and a variety of motif databases. The parameter panel allows for fine-tuning to achieve optimal settings tailored for different scenarios. In the **Genomic intervals** mode, PMET conducts homotypic indexing directly on user-provided sequences, bypassing the need to extract promoters. This mode is particularly suitable for analyzing Assays for Transposase-Accessible Chromatin using sequencing (ATAC-seq) data.

Upon submitting a task in PMET Shiny, users can enjoy the convenience of the web-based tool, as they simply need to check their email for notifications and download the results directly, bypassing the complexities of local command-line operations.

The **Visualize results** tab in PMET Shiny demonstrates its usability and rendering capabilities. This work has incorporated an adjustable heatmap generator, which summarizes the findings

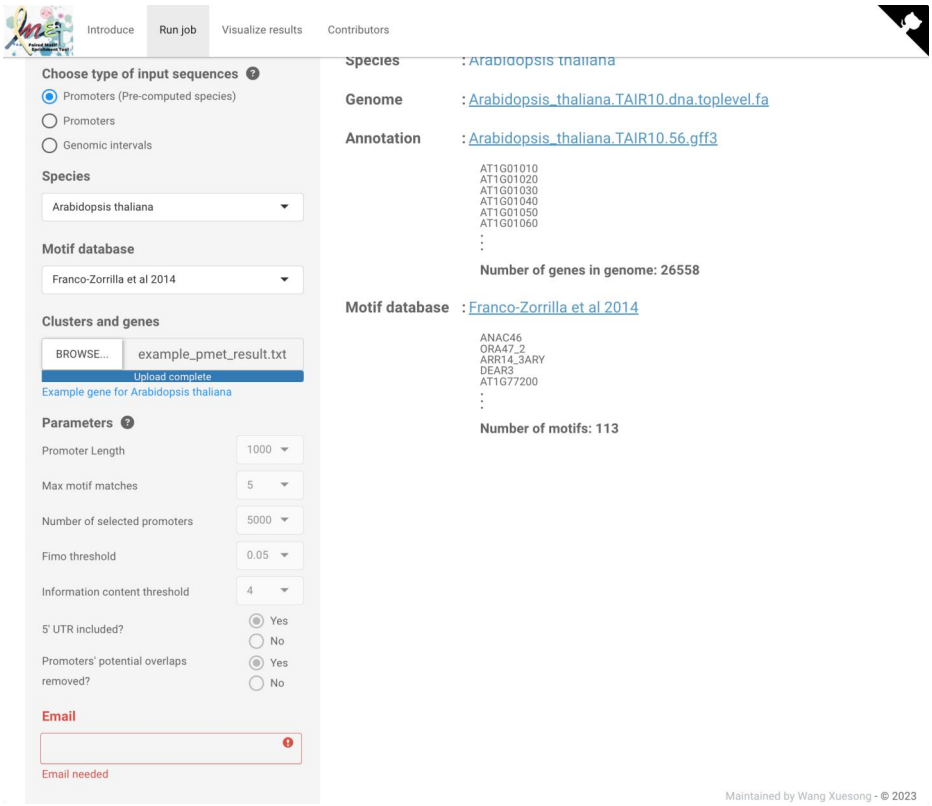


Figure 6: PMET Shiny application interface. Left panel shows settings and parameters adjustable by users for PMET analyses. Right panel presents related information based on user adjustments.

into an easily digestible format. This visual enhancement complements the text-based output by providing a graphical representation that accelerates data interpretation. Researchers can adjust the thresholds on this heatmap, fostering exploration of various motif combinations associated with specific gene clusters. This adaptability allows for a more detailed examination into motifs of the greatest statistical and biological relevance (Figure 7).

After conducting PMET analysis on gene sets specific to certain clusters (i.e., cell types), the enrichment of genes in particular motif pairs was obtained. The process for handling PMET results is as follows:

1. A threshold of 0.01 is applied to filter the Motif Pair Set(MPS) derived from PMET results, and the filtered outcomes are categorized by gene clusters, denoted as $\mathrm{MPS}_{\mathrm{cluster}}$.

2. Duplicates of motif pairs across two or more gene clusters are removed to ensure their uniqueness within each gene cluster. For example, motif pairs exclusive to the cortex-specific gene set are denoted as $\mathrm{MPS}^{\mathrm{unique}}_{\mathrm{cortex}}$.

3. Motif pairs within each gene cluster are sorted by their adjusted p-value (p.adj), and the top n most significant motif pairs are selected as the most significant. The most significant set for the cortex is denoted as $\mathrm{MPS}^{\mathrm{unique\ n}}_{\mathrm{cortex}}$.

4. Subsequently, motifs ($\mathrm{ML}_{\mathrm{cluster}}$) are extracted from the set of n unique motif pairs ($\mathrm{MPS}^{\mathrm{unique\ n}}_{\mathrm{cluster}}$).

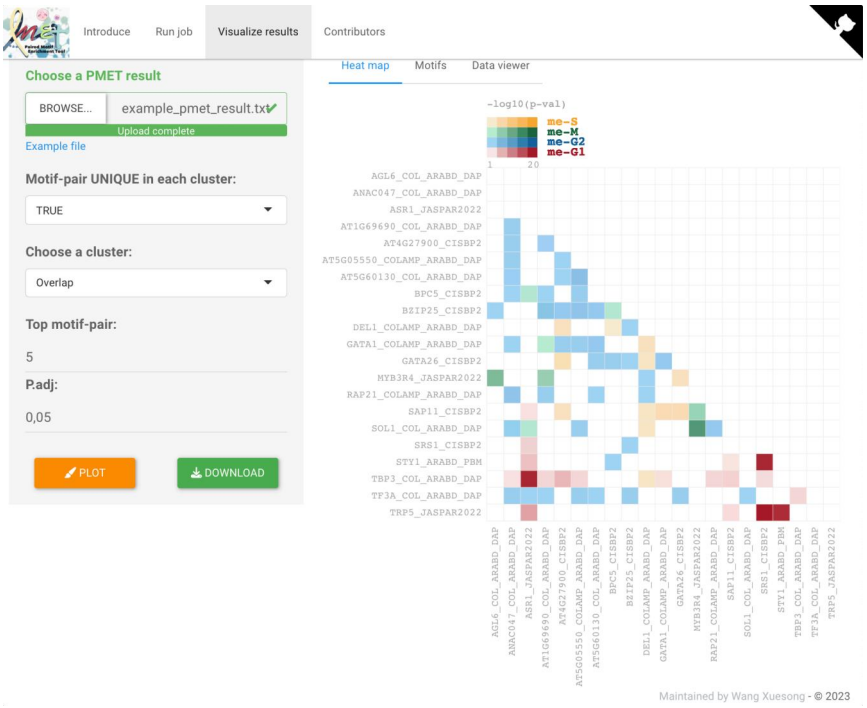


Figure 7: PMET Shiny visualization interface. Left panel displays user inputs and parameters for the visualization. Right panel shows the heatmap of uploaded results.

5. motifs ($\mathrm{ML}_{\mathrm{cluster}}$) of different clusters are combined to form a final motif list (ML).

6. Motif pairs are constructed using motifs from the motif list (ML) and identified within their exclusive sets ($\mathrm{MPS}^{\mathrm{unique}}_{\mathrm{cluster}}$).

7. Using the motif list (ML) as both the horizontal and vertical axes coordinates, a heatmap is plotted to visualize the data of the identified motif pairs. Different colors in the heatmap represent the clusters of genes, with colors shades indicating varying levels of motif pair enrichment.

This process provides a clear methodology for filtering and identifying significantly enriched in specific clusters, which aids in gaining a deeper understanding of the regulatory networks of different gene clusters. This approach also allows for a focus on regulatory elements of particular importance to each gene cluster.

2.3.2 Performance Evaluation

The analysis delivers results in two formats: a comprehensive text file detailing numerical data and a visual heatmap offering a clear, graphical representation of motif interrelationships. Within the text file, PMET analysis results are presented in a tab-delimited layout including columns such as cluster identifiers, motif pairs specific to each sequence cluster, associated P-values, and the sequence sets housing this pairs. This textual format establishes a solid foundation for subsequent investigations, enabling researchers to delve into the data and discern the biological implications associated with individual clusters, motif synergies, and their respective sequence sets.

The inclusion of a visual heatmap not only condenses the discoveries into an easily digestible format but also enriches the text-based output. This graphical representation accelerates data interpretation, enabling researchers to flexibly adjustment the thresholds. This adaptability facilitates the exploration of varied motif pairs associated with specific gene clusters, homing in on the motifs with the greatest statistical and biological relevance.

2.4 Processing and Analysis of Motifs

In the pursuit of understanding gene regulation within plants such as *A. thaliana*, PMET has been meticulously crafted to discern the presence and enrichment of TF motif pairs within specific gene clusters. To facilitate this analysis, PMET leverages a curated selection of motif databases that include a breadth of experimental evidence. These databases, derived from high-throughput techniques like SELEX-seq, protein binding microarray (PBM), ChIP-seq and DNA affinity purification and sequencing (DAP-seq), provide a robust foundation for analyzing TF binding across various biological contexts–ranging from distinct cell types to varied environmental conditions, such as stress or treatments. With the support of these *in vivo* resources, the *in silico* PMET methods can proivde s a detailed and integrated picture of the regulatory networks.

2.4.1 Compiling Motif Databases

This work aggregated motifs from multiple established databases, creating a dataset rich in comprehensiveness but also in redundancy. This collection includes 3,043 motifs from the following sources: Franco-Zorrilla et al. (2014)[157], Jaspar Plants Non-Redundant (2022)[176], Plant Cistrome DB[161], PlantTFDB[177], and CIS-BP2[178], offering a broad spectrum of motif recognition patterns (Table 1). The UpSet plot depicted in Figure 8 illustrates the intersections of motif alternative names across these databases, revealing shared identities that reflect biological similarities or functions. These alternative names often are not unique and correspond to broader categories or specific TFs. For instance, a motif identified as "MA0001.1" with AGL3 as its alternative name, which is linked to *AGL3* TF under *MADS* box factor class and *MIKC* family. To provide a clear biological context and ensure the uniqueness of each entry, the alternative names for the 3,043 motifs were used as identifiers, supplemented with the corresponding database name. In instances of duplication, a unique index was appended to distinguish each motif.

Table 1: Summary of motif databases used for PMET analyses

Motif Database	Types of Motif	Number of Motifs
Franco-Zorrilla et al., 2014	Experimentally Defined (SELEX-seq and PBM)	113
Jaspar Plants Non-Redundant 2022	Manually Curated Motifs for Plants Based on Published Collections and de Novo Generated	656
Plant Cistrome DB	Experimentally Defined (DAP-seq)	872
PlantTFDB	Manually Curated Motifs of Published Collections (ampDAP, ChIP-chip/seq, DAP, PBM and SELEX)	619
CIS-BP2	Experimentally Defined (PBM) and Published Collections	783

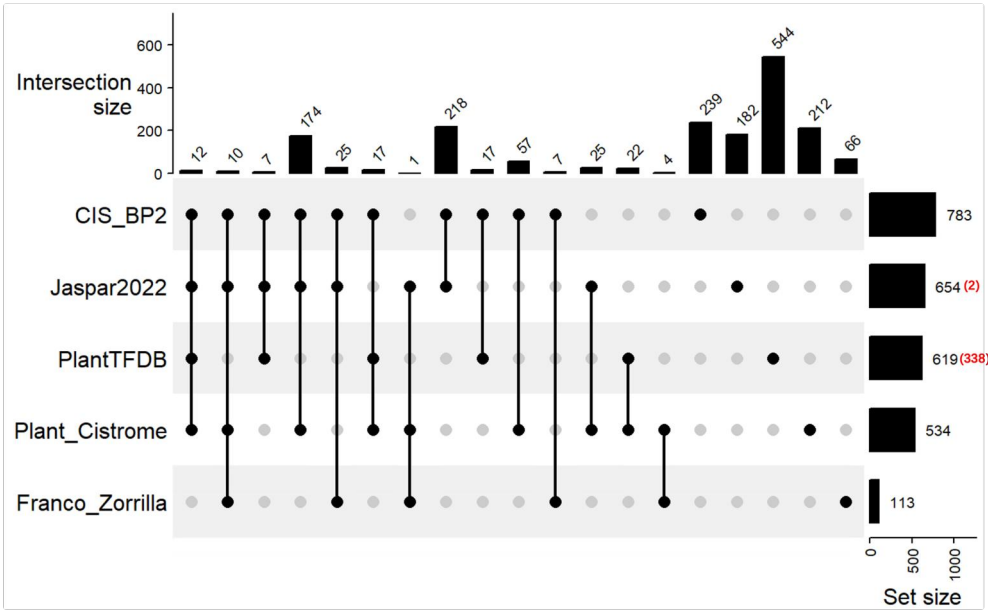


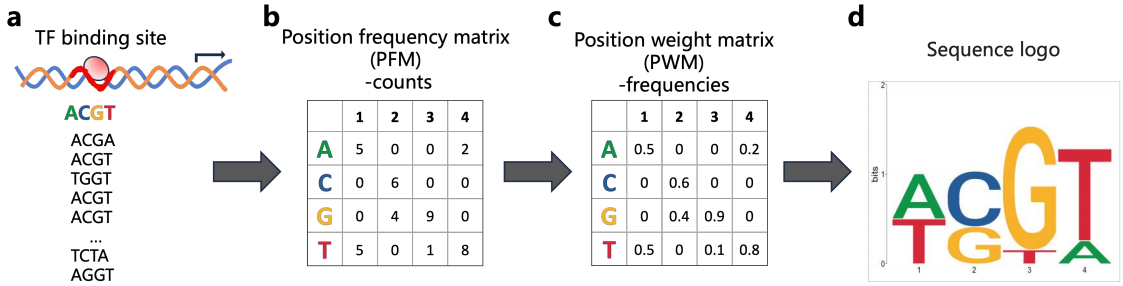
Figure 8: Intersections of motifs shared between five different motif databases. Each row represents a database, and bars with linked dots indicate unique combinations of databases sharing motifs. Intersection sizes are represented by the height of bars along the top. For instance, the first row indicates 783 motifs in CIS-BP2, with dots representing potential overlaps with other databases. The first column, marked by a line with 4 dots, denotes that 12 motifs with the same names are shared across four databases, including CIS-BP2, Jaspar Plants Non-Redundant 2022, PlantTFDB, and Plant Cistrome. The horizontal bars on the right represent the number of motifs in each database. The number of motifs with the same name is indicated by red numbers in Plant Cistrome and Jaspar Plants Non-Redundant 2022.

2.4.2 Redundancy Removal in Motifs

In computational biology, a motif is typically modeled as a PWM, where each row represents a nucleotide (A, C, G, and T) and each column corresponds to the nucleotide preferences (probability) within the motif (see Figure 9c). PWMs are crucial in most bioinformatics tools for tasks such as de novo motif discovery, consensus sequence recognition, sequence alignment, and motif enrichment analysis. Sequence logos offer a more intuitive representation of the motif (Figure 9d) where the height of each base is determined by IC. The IC indicates the conservation level of the motif; a higher IC value often signifies a greater degree of sequence-specific binding, while a lower IC value suggests randomness in the nucleotide occurrence at that position. The IC value for a motif is calculated as the sum of IC values from all positions within the motif (Equation 5).

$$IC = \sum_{i=1}^{4} p_i \cdot \log_2 \left(\frac{p_i}{q_i} \right) \tag{5}$$

where i represents the index of the four nucleotides (A, C, G, and T), p_i is the probability of the ith nucleotide at a certain position, and q_i is the background probability of the ith nucleotide. An IC of 0 indicates that the nucleotide occurrences are random, which occurs when p_i equals q_i.



	1	2	3	4
A	5	0	0	2
C	0	6	0	0
G	0	4	9	0
T	5	0	1	8

	1	2	3	4
A	0.5	0	0	0.2
C	0	0.6	0	0
G	0	0.4	0.9	0
T	0.5	0	0.1	0.8

Figure 9: Prediction of TF binding motifs. (a) DNA sequence with the bound TF indicate the region of interest as putative motif. (b) Frequency of nucleotides (A, C, G, and T) in the DNA sequence estimates the frequency of each base at each position. (c) Frequency counts are normalized to reflect the likelihood of nucleotide occurrence. (d) A sequence logo is a graphical representation of a motif, derived from its position weight matrix (PWM). Taller stacks of bases indicate a higher probability of occurrence of certain nucleotides. For instance, the third and fourth positions are dominantly G and T, respectively.

Examples of Motif Similarity When utilizing motif databases, an inescapable complication arises from motif similarity-whether through nearly identical PWMs or the existence of motifs with analogous names but divergent PWMs across various databases (Figure 10). For instance, *GRF6* as depicted in Jaspar Plants Non-Redundant 2022 and *GRF9* from Plant Cistrome DB, as shown in Figure 10a, share strikingly similar sequence logos, indicating similar PWMs, yet are distinguished by disparate nomenclature and database origins. Furthermore, motifs such as *DOF43* from Plant Cistrome DB and *DOF4.3* from Jaspar 2022 exhibit comparability despite differing identifiers (Figure 10d).

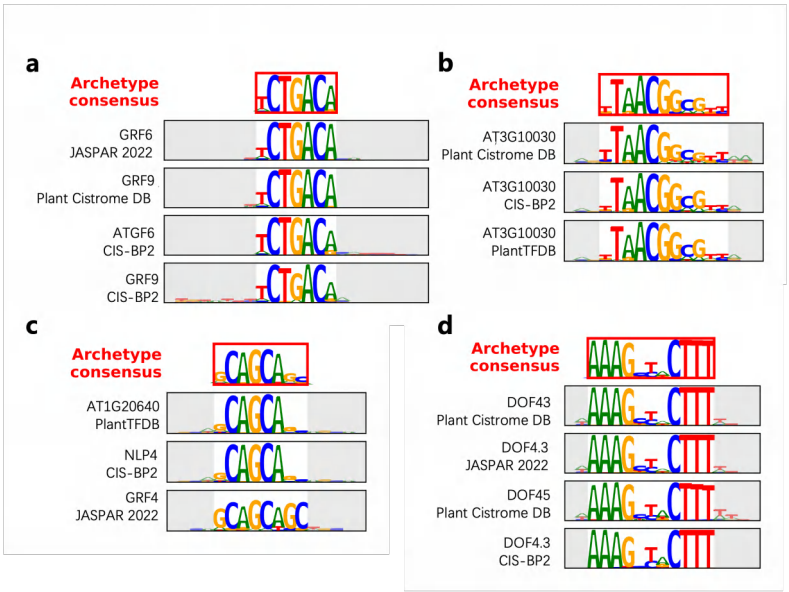


Figure 10: Comparison of motif sequence logos illustrating the similarity and potentially redundant representations across different databases. (a) Variants of *GRF6* in JASPAR 2022 and Plant Cistrome DB. (b) Representations of *AT3G10030* in Plant Cistrome DB, CIS-BP2, and PlantTFDB. (c) Diverse depictions of *AT1G20640* and *NLP4* in PlantTFDB and JASPAR 2022, (d) *DOF43* in Plant Cistrome DB and JASPAR 2022 versus *DOF4.5* in Plant Cistrome DB and CIS-BP2.

This observation underscores a potential motif redundancy. While it may be marginal during the characterization of TF DNA-binding specificities and in *de novo* motif discovery, where redundancy is often traded for exhaustive coverage—becomes significant [225]. The redundancy issue becomes especially pertinent when integrating different motif databases to construct comprehensive transcription regulatory maps or networks, as it may lead to the incorporation of superfluous motifs.

Removing motifs with low IC The distributions of individual IC values, as well as their cumulative sums, have been calculated for the 3,043 motifs in Table 1. IC values for all 3,043 motifs were calculated and then motifs with the lowest 5% of average IC values across all bases were filtered out. This filtering step was essential to exclude motifs with insufficiently informative content, thus preventing null results during motif comparisons. Figure 11a clearly illustrates the demarcation between motifs with IC values below 0.6 (5% quantile) and those above it. Generally, the average IC value of a motif is favored over the sum, as it can be disproportionately influenced by the motif's length. After the initial filtering based on the average IC values, the total number of motifs was reduced from 3,043 to 2,738.

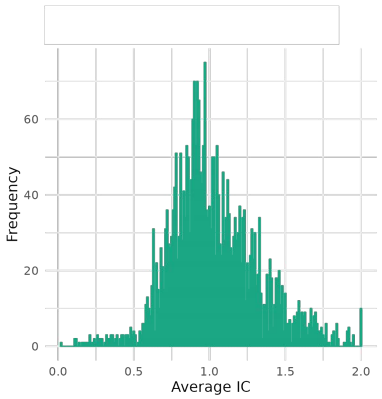


Figure 11: Distribution of information content (IC) values across 3,043 motifs. (a) A dense distribution of motifs with IC values primarily over 0.5.

Removing motifs with 'equal' PWMs In Figure 8, a substantial number of motifs across various databases share identical alternative names, suggesting potential similarity or identity. To assess this, pairs of the 2,738 motifs were compared to determine if they were (nearly) equal, considering both the length of motifs and the equality of their PWMs with a tolerance of 0.05 for each base and each nucleotide (A, C, T, and G). When motifs were deemed 'equal', only one motif was retained. This filtering process resulted in a reduction across all databases, as illustrated in Figure 12. For instance, the size of the CIS-BP2 database was reduced from 783 to 605 motifs, as indicated in the first row of the UpSet plot. Similarly, the number of shared alternative names between CIS-BP2, Jaspar 2022, Plant-Cistrome, and PlantTFDB was reduced from 12 to 3. Ultimately, this led to the creation of a consolidated database comprising 1,565 motifs.

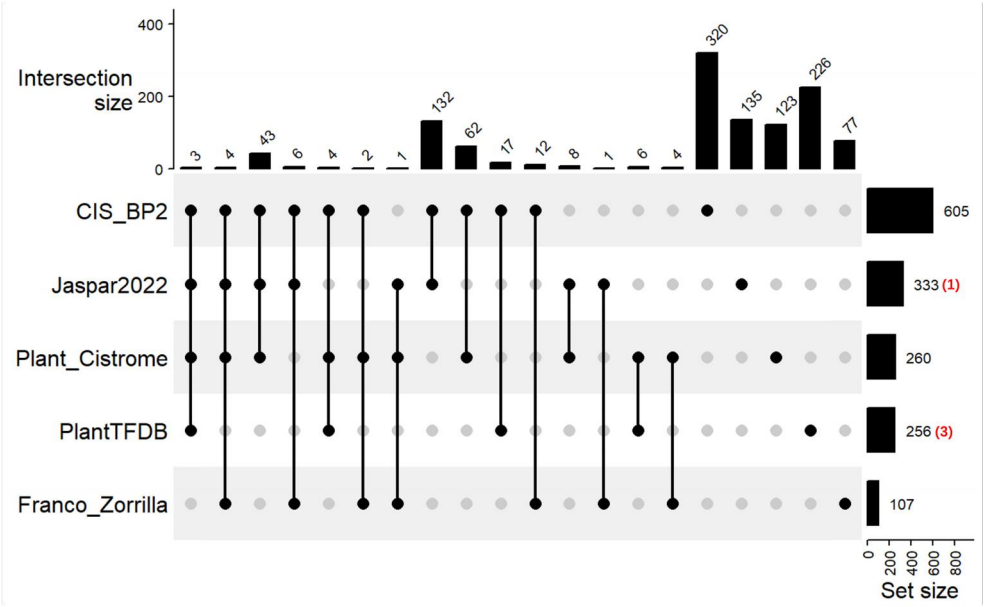


Figure 12: Intersections of motifs shared between five different motif databases after removing motifs with 'equal' PWMs. Bars with linked dots indicate unique combinations of databases sharing motifs. Intersection sizes are represented by the height of bars along the top. For instance, the first row indicates 605 motifs in CIS-BP2, with dots representing potential overlaps with other databases. First bar, marked by a line with 4 dots, denotes 3 motifs with the same names that are shared in four databases, including CIS-BP2, Jaspar Plants Non-Redundant 2022, PlantTFDB, and Plant Cistrome. Horizontal bars on the right represent the number of motifs in each database. The number of motifs with the same name is indicated by red numbers in Plant Cistrome and Jaspar Plants Non-Redundant 2022.

2.4.3 Motif Similarity Quantification

Similar or identical motifs likely correspond to the same TF. Therefore, it is essential to differentiate and correctly label these motifs to accurately reveal the underlying regulatory processes. The identification of motif similarity is crucial in this context, primarily determined by comparing the PWMs of the motifs, typically based on the alignment of nucleotide sequences[226,227,220]. Background information on nucleic acid sequences of motifs, which outlines the transition probability across a series of potential states, is vital for some comparison methodologies. Markov model is commonly used to calculate the probability of occurrence of individual nucleotide in a sequence, as well as the probability of its occurrence with a specific order.

An array of metrics is available to measure the similarity between alignments, including Euclidean distance (EUCL), Weighted Euclidean distance (WEUCL), Kullback-Leibler divergence (KL), Hellinger distance (HELL), Squared Euclidean distance (SEUCL), Manhattan distance (MAN), Pearson correlation (PCC), Weighted Pearson correlation (WPCC), Sandelin-Wasserman similarity (SW), Average log-likelihood ratio (ALLR), Lower limit average log-likelihood ratio (ALLR_LL), and Bhattacharyya coefficient (BHAT). The "TOMTOM" tool, part of MEME suite, provides functions to compute the similarity measures, facilitating a comprehensive analysis of motif similarities. Metrics such as ALLR_LL and KL require the background information of motifs because of their reliance on the statistical properties of the motif sequences. In contrast, metrics

like EUCL, PCC and SEUCL do not require extra information as they are designed to assess motifs' structural differences and compositions independently of the sequence statistics.

In this work, the "TOMTOM" tool from the MEME Suite was used to evaluate the similarity across 1,565 motifs among themselves. This comparison yielded a ranked list reflecting the significance of resemblance between each source motif and established target motifs. The EUCL metric was chosen for scoring motif alignments due to its proven efficacy and performance in matching simulated motifs with target motif databases. Notably, the EUCL metric demonstrated a remarkable success rate—up to 99% for motifs longer than 10 nucleotides[226], which aligns well with dataset in this work, where the majority of the 1,565 motifs under study are of this size or greater (Figure 13). In "TOMTOM" comparison, factors such as alignment direction, starting position (offset), and alignment sequence length contribute to the non-reciprocal similarity, meaning the similarity between motif A and motif B is not equal to the similarity between motif B and motif A[228].

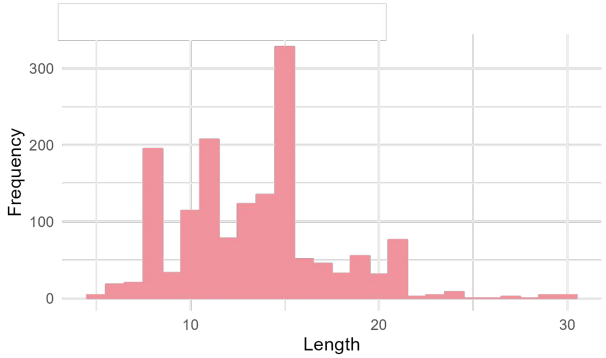


Figure 13: Distribution of motif lengths within the filtered dataset of 1,565 motifs. The majority of motifs exhibit lengths exceeding 10 nucleotides. This length distribution supports the use of Euclidean distance (EUCL) in motif comparisons.

The results of TOMTOM are presented in various forms, including HTML format with visual results, and TSV format using tables as the carrier. The information contained in the TSV format mainly consists of similarity of pairwise comparison between motifs, measured by p-value, which represents the probability that the observed similarity occurs by chance. Considering that the comparison between motifs is multiple, E-value is also used to measure similarity, which is the correction of p-value. It takes into account the multiplicity of searches in the motif database, which means that in a random database of the same size , the number of times expected to find similarities randomly. A lower E-value (closer to 0) indicates very high similarity significance, while a higher E-value (closer to 1) indicates lower significance. Other features, such as 'Overlap' or 'Query Consensus' can be queried on the MEME suite website, and further details are not provided here.

The two histograms in Figure 14 show the E-value distribution for 1,565 motifs from TOMTOM motif similarity analysis. In Figure 14a, the range of E-value is limited to above 0.01 for better presenting the data. It indicates that many of the values of E-value are greater than 0.01, which means that a considerable number of motifs are Not similar. The red bars in the figure reveal

Table 2: Output of Tomtom motif similarity

Query	Target	offset	p-value	E-value	q-value	Overlap	Orientation
MYB52	MYB111	0	3.50e-25	3.9e-23	7.8e-23	8	+
MYB52_2	MYB111	0	1.05e-25	1.1e-23	2.3e-23	8	+
MYB52_2	MYB46	0	8.57e-4	9.6e-2	9.4e-2	8	+

the distribution of E-values in the range of 0.01 to 0.5, showing a relatively flat distribution trend throughout the range. In Figure 14b, within the range of E-value less than 0.01, the blue bars emphasizes that the vast majority of motif matches have extremely low E-value, indicating that there is a very significant motif similarities in the data set. This suggests that these motifs may have high correlation and functional similarity in biological regulation.

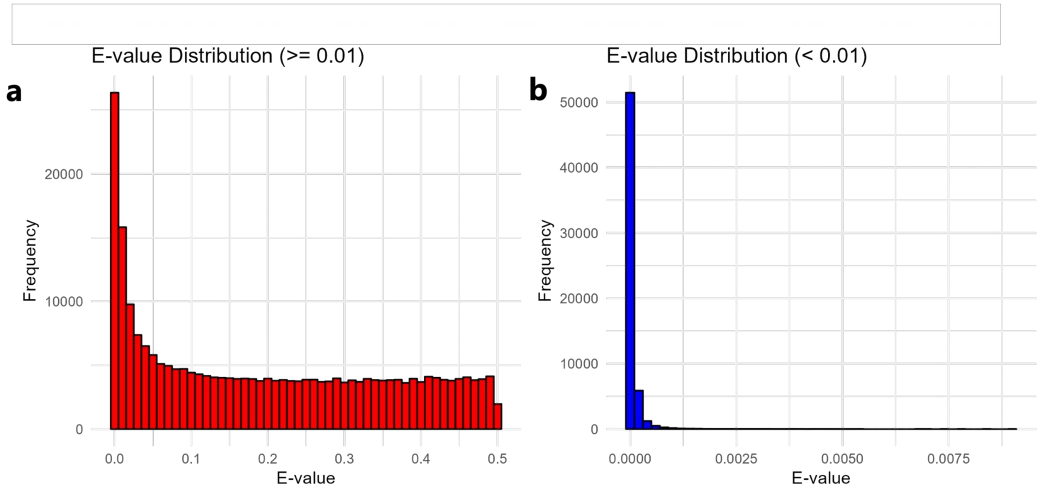


Figure 14: E-value distribution of 1,565 motifs pairs, determined by TOMTOM motif similarity. Each E-value is based on the p-value corrected for multiple motif comparisons, reflecting the probability of significant motif similarities. Lower E-value (closer to 0) indicates high significance, while a higher E-value (closer to 1) indicates lower significances. (a) The range of E-values above 0.01 (red bars) represents weakly similar motif pairs. (b) The distribution of E-values less than 0.01 (blue bars) shows that most motif matches have extremely low E-values, indicating significant aggregation and suggesting high relevance and functional similarity in gene regulation.

2.4.4 Motif Clustering Analysis

Hierarchical clustering was employed to group similar motifs based on the similarity of their TOMTOM results. Therefore, this method initiates with each motif forming an individual cluster and progressively merges the most adjacent clusters. The adjacency between two clusters is quantified by the distance, which is determined by the similarity between the motifs within them.

At each step, the hierarchical clustering algorithm identifies the two most similar clusters and merges them. This process is repeated until all clusters are amalgamated into one comprehensive cluster. The entire procedure is visualized using a dendrogram, which effectively demonstrates the relationships and distances among motifs or clusters. Each branch on the dendrogram represents a cluster, and the endpoints of these branches signify individual motifs. The lengths of the branches and their interconnections reflect the level of similarity and the sequence of merging.

The hierarchical clustering method was applied to the screened 1,565 motifs, assessing the hierarchical relationships and distances by measuring the correlation between clusters (Figure 15).

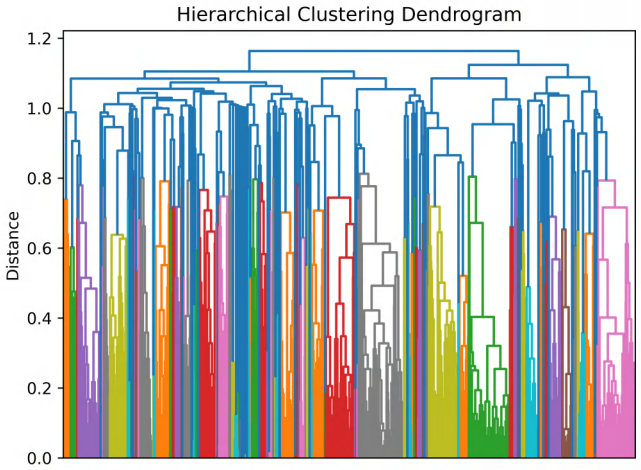


Figure 15: Hierarchical cluster dendrogram. Each node in the dendrogram contains two child nodes, indicating two similar motifs. At a distance of 0 are the "leaves," representing individual motifs. Clustering continuously groups the two nearest/most similar nodes until only one node remains.

The similarities derived from TOMTOM and hierarchical clustering are depicted in a clustered heatmap (Figure 16). Upon observing the heatmap's diagonal, several blue blocks were noted, indicative of clusters with high similarity. The most prominent block, encompassing 116 motifs, indicates a significant degree of similarity among these motifs, potentially pointing to shared biological functions or common regulatory mechanisms.

Two complementary methods, Silhouette analysis and Gap statistic were employed to determine the best threshold of the clustering distance. Silhouette analysis quantifies the tightness or closeness of clustering, yielding Silhouette scores ranging from -1 to 1, with higher scores indicating tighter clustering. Remarkably, when utilizing distance thresholds of 0.3, 0.4, and 0.5, the Silhouette scores remained consistent at approximately 0.19. Meanwhile, Gap statistic assesses clustering tightness by comparing expected values of clustering among actual and random data across different cluster sizes. Notably, a stabilization in fluctuation amplitudes in the number of clusters around 25 was noted, reaching a plateau at approximately 175 clusters, corresponding to a distance threshold of approximately 0.7 in Figure 15.

2.5 PMET Indexing of 21 Plant Species

Generally, PMET indexing entails scanning for motif matches across all promoters within a genome. Extensive promoters and multiple motifs can lead to prolonged computation times. To address this, commonly selected model plants and motif databases were employed to pre-compute motif hits using PMET's default parameters.

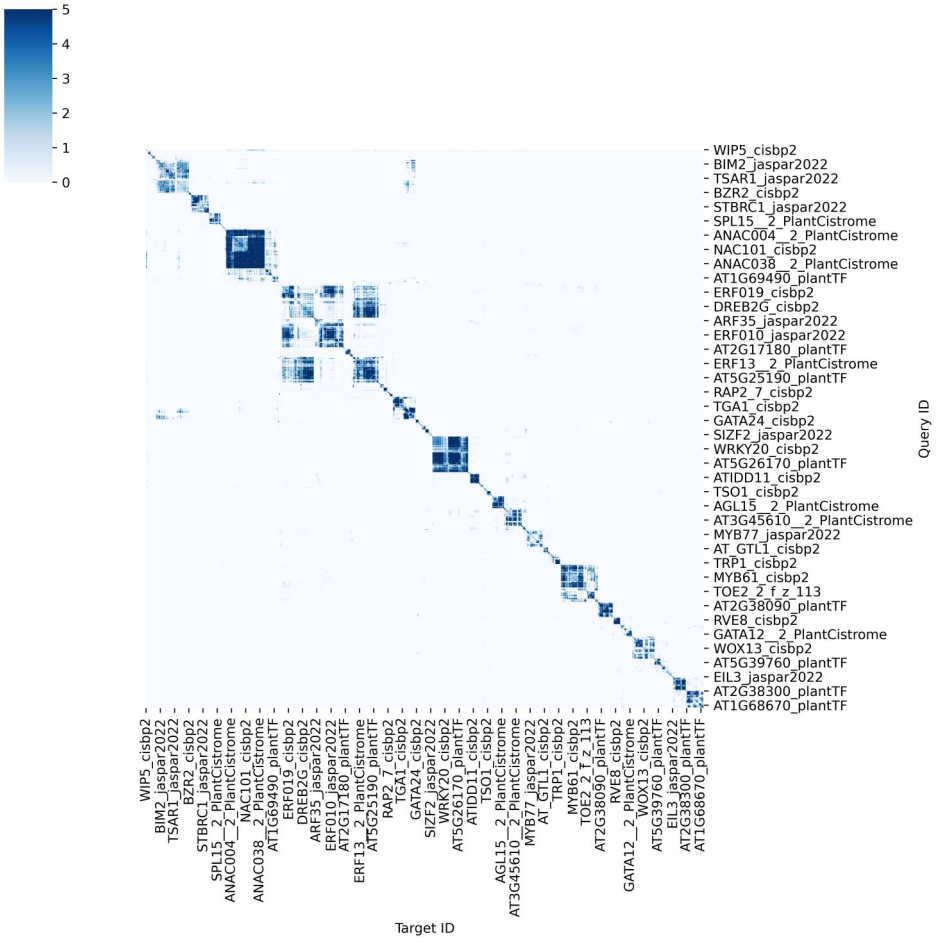


Figure 16: Heatmap illustrating similarities between motifs. Blocks with darker shades of blue in the heatmap denote clusters of motifs with greater similarity.

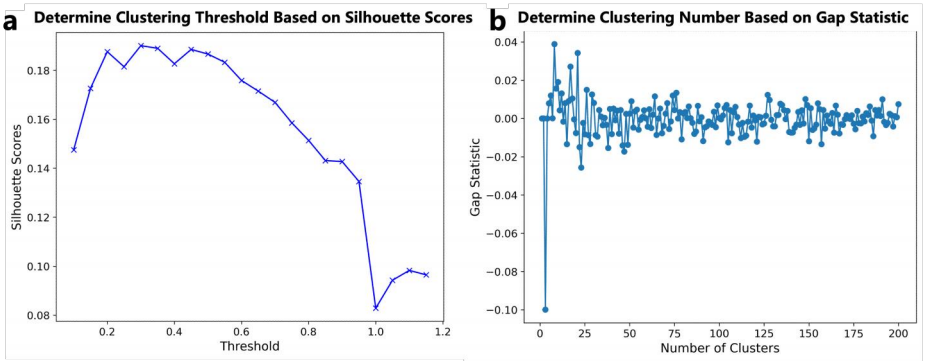


Figure 17: Silhouette analysis and gap statistic methods were used to determine optimal clustering parameters. (a) Silhouette analysis shows consistent scores around 0.19 at distance thresholds of 0.3, 0.4, and 0.5, indicating these may be optimal for motif clustering with hierarchical clustering. (b) Gap statistic reveals a plateau at 175 clusters at a distance threshold of 0.7. These methods help identify the optimal number of non-redundant motifs from a total of 1,565 motifs.

For the PMET analysis, a diverse set of 20 model plants was carefully selected to represent a wide range of genomic diversity. The curated species include a range of important crops and model organisms: *A. thaliana*, *Brachypodium distachyon*, *Brassica napus*, *Glycine max*, *Hordeum vulgare*

Morex V3, *Hordeum vulgare* R1, *Hordeum vulgare* goldenpromise, *Hordeum vulgare* v082214v1, *Medicago truncatula*, *Oryza sativa indica* 9311, *Oryza sativa indica IR8*, *Oryza sativa indica MH63*, *Oryza sativa indica* ZS97, *Oryza sativa japonica Ensembl*, *Oryza sativa japonica Kitaake*, *Oryza sativa japonica Nipponbare*, *Oryza sativa japonica V7.1*, *Solanum lycopersicum*, *Solanum tuberosum*, *Triticum aestivum*, *Vicia faba*, and *Zea mays*. The genome and gene annotation data for these species are accessible through EnsemblPlants, specifically th release version of 57.

These species were chosen to provide a comprehensive perspective of plant genetic research, enabling PMET to cover a broad spectrum of plant genomes. This approach is designed to enhance the generalizability of the motif pair enrichment analysis.

2.6 Statistical Models of PMET

Table 3: Parameters utilized in the PMET analyses for exploring motif pairs

Parameter	**Default value**
Motif Database	Franco-Zorrilla et al. (2014)[157]
Species	*A. thaliana*
Genome	TAIR10
Annotation	TAIR10
Length of promoter	1000
K of PMET	5
N of PMET	5,000
IC of PMET	4
Overlap with other gene	No
5' UTR included	No
p.adj threshold	0.05

The PMET methodology involves indexing and pairing processes, both utilizing statistical models to determine the significance of motif hit counts within promoters, ensuring that the observed motif hits are biologically meaningful rather than due to random chance. Binomial and Poisson distributions are applied to assess the number of maximal motif hits allowed per promoter (denoted as k, with a default value of 5). The PMET analysis adheres to standard settings, including a set of default parameters (Table 3).

The analysis was performed using the genome and annotations of *A. thaliana*, with a motif database from Franco-Zorrilla et al. (2014) comprising 113 motifs. Each analysis involved two distinct gene sets to determine the enrichment of paired motifs: one set specific to cell types, including 447 cortex genes, 462 epidermis genes, and 455 pericycle genes, and another set consisting of immune-responsive genes from the aforementioned cell types (Table 4 and 5). Four clusters of randomly selected genes were also used to assess the significance of motif hits. Additionally, the comparative analysis was expanded to include a dataset of 300 genes differentially expressed under heat and salt stress conditions, providing context for motif pair distribution in relation to environmental response."

The bar plots illustrate in Figure 18a, b, and c showcase the quantity of significant motif pairs identified across various gene clusters. Notably, there is the discrepancy in value ranges between

Table 4: Six clusters of immune-responsive genes used for PMET analyses

Species	Cell type	Treatment	Regulation	Number
A. thaliana	Cortex	flg22	Up	128
A. thaliana	Cortex	Pep1	Up	365
A. thaliana	Cortex	Pep1	Down	337
A. thaliana	Epidermis	flg22	Up	128
A. thaliana	Epidermis	Pep1	Up	365
A. thaliana	Epidermis	Pep1	Down	337

Table 5: Three clusters of cell type genes used for PMET analyses

Species	Cell type	Number
A. thaliana	Cortex	447
A. thaliana	Epidermis	462
A. thaliana	Pericycle	455

the three PMET analyses; the cell type-specific genes show a range of 50-250 significant motif pairs, whereas the immune-responsive genes show a narrower range of 10-50. There are very few significant results for clusters of random genes. Furthermore, when comparing the outcomes of PMET analyses for all gene clusters, the binomial and Poisson distributions yield similar results, with only minor differences in the count of significant motif pairs. For instance, within the cortex-specific gene cluster, the binomial model identifies 245 motif pairs, whereas the Poisson model reveals 246, with a difference of only one motif pair. A similar pattern is observed in the epidermis-specific gene cluster, where each model captures a unique motif pair not shared with the other. This pattern is consistent across the remaining gene clusters.

2.7 Parameter Sensitivity Analysis of PMET

2.7.1 Effect of Promoter Length Variation

The promoter is commonly referred to as the DNA sequence upstream of the transcriptional start site (TSS), yet this definition does not specify the exact DNA segment that constitutes a promoter or stipulates its length. Genome annotations typically do not explicitly label promoter regions. For *A. thaliana*, analyses of genomic layout and SNP density profiles proximal to TSS suggests an approximate promoter length of 500 bp[229]. In PMET analysis, a promoter length beyond this estimate is often used to ensure comprehensive coverage. However, increasing the promoter length beyond this estimate is not necessarily advantageous for two main reasons: the actual length of a promoter, which does not change with set parameters, and the risk of encroaching upon upstream genes, resulting in 'overlap.' Gene annotations and precise genomic coordinates facilitate the extraction of promoter sequences of defined lengths, ensuring they do not extend into adjacent genes. To quantify the impact of varying promoter lengths on PMET outcomes, promoter lengths were adjusted while other parameters remained constant, and plotted the number of motif pairs for each gene cluster against changes in promoter length. Across two distinct gene sets (three clusters of cell type-specific genes and six clusters of immune-responsive genes), it was observed

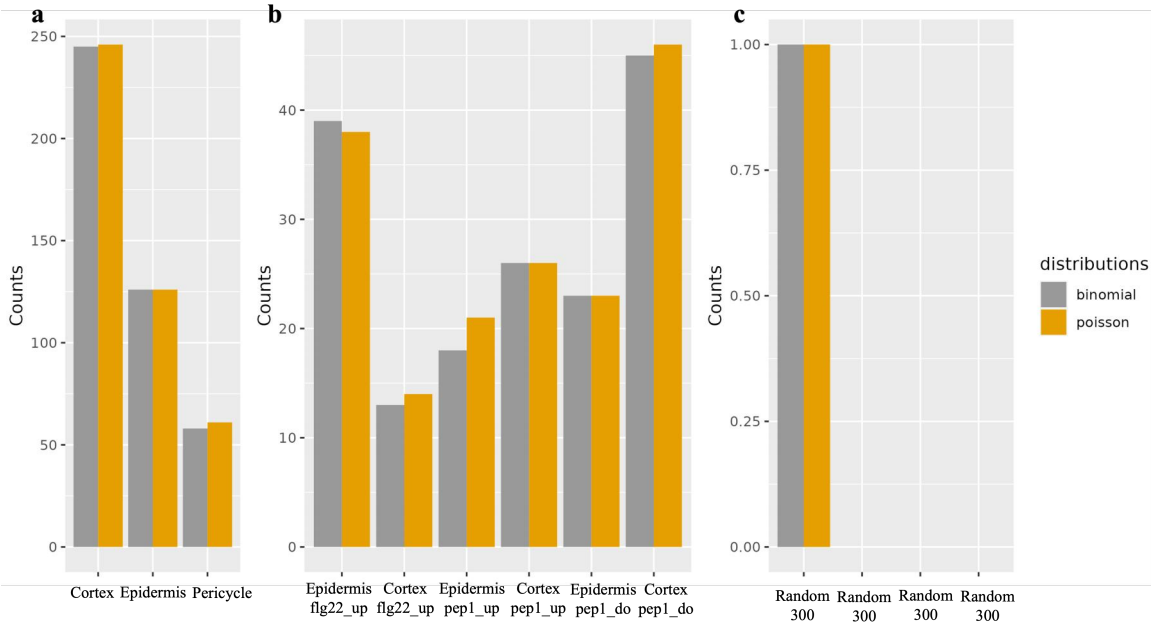


Figure 18: Distribution of significant motif pairs across various gene clusters as determined by binomial (gray) and Poisson (yellow) statistical models. Notably, (a) cell type-specific genes exhibit a higher count of significant pairs, (b) whereas immune-responsive genes show a reduced count, (c) and random gene clusters display negligible significance. Similarity between the binomial and Poisson results suggests robustness of the PMET analysis across different statistical approaches.

that promoters, excluding 5' UTRs and without overlap yielded fewer motif pairs at shorter lengths. As promoter length increased, the number of motif pairs gradually rose and plateaued or began to decrease subtly at a certain length (Figure 19 a and c). This pattern was also noted in PMET analyses of differentially expressed genes under heat and salt stress, where motif pairs enriched in up-regulated genes showed a similar trend of initial increase followed by stabilization or a slight decline (Figure 19 e and g).

The plateau observed in the number of identified motif pairs may be attributed to the inherent spatial constraints of promoter regions. Figure 20 reveals that the upstream region of the TSS for a vast majority of genes extends to an average of 2300 bp, seldom surpassing the 5,000 bp threshold. Consequently, restricting overlap inherently limits the domain searchable for motif hits in PMET, rendering extensions beyond the natural promoter boundary ineffective.

Table 6: Statistical summary of the length of TSS upstream of *A. thaliana* genes

Metric	Value
Overall Average	2303.21 bp
Filtered Average (length < 10,000bp)	1672.03 bp
Length Over 50 bp	90.00%
Length Over 100 bp	86.23%
Length Over 200 bp	78.69%
Length Over 500 bp	64.69%
Length Over 1,000 bp	48.62%
Length Over 1,500 bp	37.41%
Length Over 2,000 bp	29.59%
Length Over 10,000 bp	3.06%

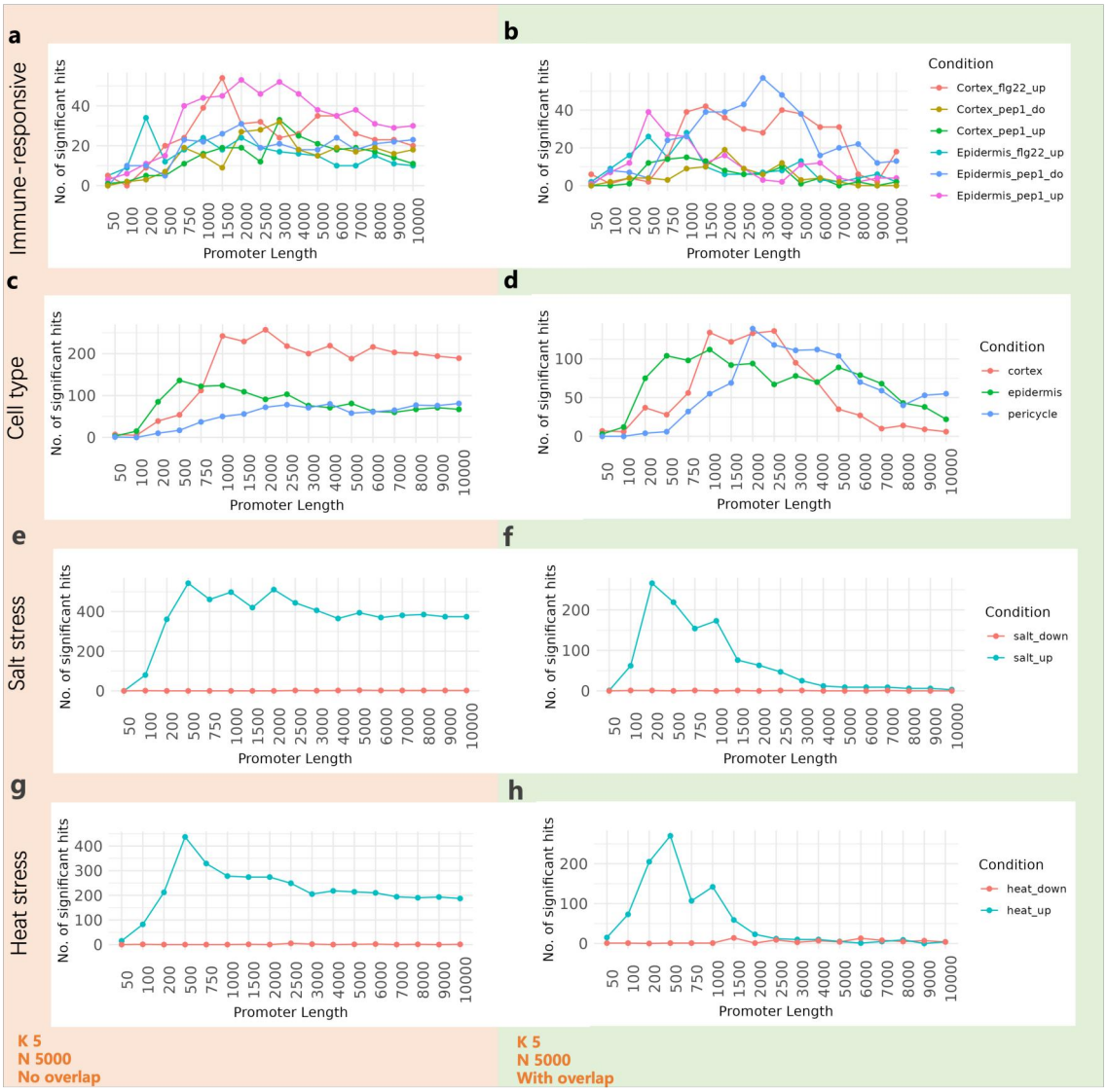


Figure 19: Impact of overlap on the detection of significant motif pairs in immune-responsive and cell type-specific genes across varying promoter lengths, with K set to 5 and$N$ to 5000. (a) and (c) show results without promoter overlap, where the number of motif pairs peaks around a promoter length of 1000 bp. In contrast, (b) and (d) show results with promoter overlap, highlighting a significant reduction in motif pair detection, emphasizing the negative influence of overlap on PMET analysis. (e) and (f) display PMET results for genes regulated under salt stress with and without overlap, respectively. (g) and (h) show PMET results for genes regulated under heat stress with and without overlap, respectively.

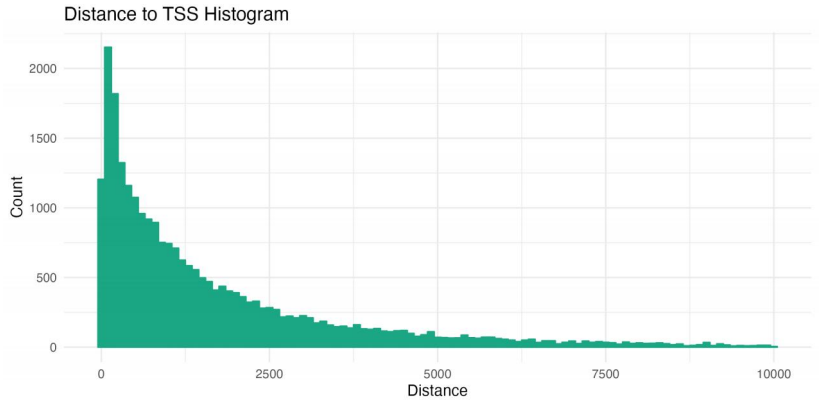


Figure 20: Distribution of gene counts at varying distances upstream from the TSS. The dataset is predominantly concentrated within 1500bp, with a minority of genes extending beyond 5000bp.

2.7.2 Effect of Sequence Overlap Thresholds

The preceding text discussed how the setting of promoter lengths and the inherent limitations of promoter regions affect PMET results. However, the potential overlap within promoters was not explored in depth. If an overlap exists, it implies that the possible search range for PMET can be arbitrary. The impact of overlap on PMET results was observed with different promoter length settings while keeping other parameters constant.

When disallowing overlap, hardly any motif pairs are detected with promoter lengths as short as 50 (or 100), for both immune-responsive and cell type-specific genes. The absence of motif pairs persists even when overlap is permitted, which aligns with expectations. As previously highlighted, the average distance upstream from the TSS to the previous gene is 2,300 bp, which reduces to 1,672 bases when excluding 847 genes extending beyond 10,000 bases in length. When the promoter length is set at 100 bases, overlap is unlikely for 86% of the genes, as their upstream regions exceed this threshold. Consequently, the number of motif pairs generally increases with longer promoter lengths, peaking around 1,000 bases. (Figure 19a-d).

Notably, the presence of overlap generally exerts a detrimental effect on the increase of motif pair counts. Without overlap, the number of motif pairs on cortex-specific genes can reach up to 300, but this number is reduced to 130 with 'overlap (Figure 19c and d).

PMET analysis on the overlap topic was extended to differentially expressed genes under salt and heat stress conditions, revealing a significant decline in the detection of motif pairs as promoter length increased. Moreover, the overall counts of motif pairs were substantially lower compared to those observed in cell type-specific genes (Figure 19e-h).

2.7.3 Effect of Promoter Set Size

In PMET, each motif's binding potential to promoters is assessed across the entire gene set. A binomial distribution-based score is calculated for each promoter, capturing the probability of observing up to k hits. This scoring allows for nuanced differentiation of promoter binding significance. Rather than setting an arbitrary significance threshold, PMET employs a ranking

system, prioritizing promoters based on their scores and retaining the top (N) as potential targets for significant motif pair interactions.

(a) and (c) show results without promoter 'overlap', where the number of motif pairs peaks around a promoter length of 1000 bp. In contrast, (b) and (d) show results with promoter 'overlap', highlighting a significant reduction in motif pair detection, underscoring the negative influence of 'overlap' on PMET analysis. (e) and (f) display PMET results for genes regulated under salt stress with and without overlap, respectively. (g) and (h) show PMET results for genes regulated under heat stress with and without overlap, respectively. To exemplify the impact of N mathematically, two motifs, *AHL12* and *TOE2*, were subjected to a 'FIMO' search to pinpoint potential hits across the spectrum of available promoters. Following the binomial score-based ranking, PMET earmarks the leading 5,000 and 25,000 promoters for the motifs in question. As depicted in Figure 21, there is a clear increase in the number of motif pairs for $N = 25,000$ where the blue bars significantly surpass the height of the yellow bars. Additionally, an increase in N notably shifts the p-value distribution towards the right, signaling a trend towards lower significance, meaning more weaker bindings are included.

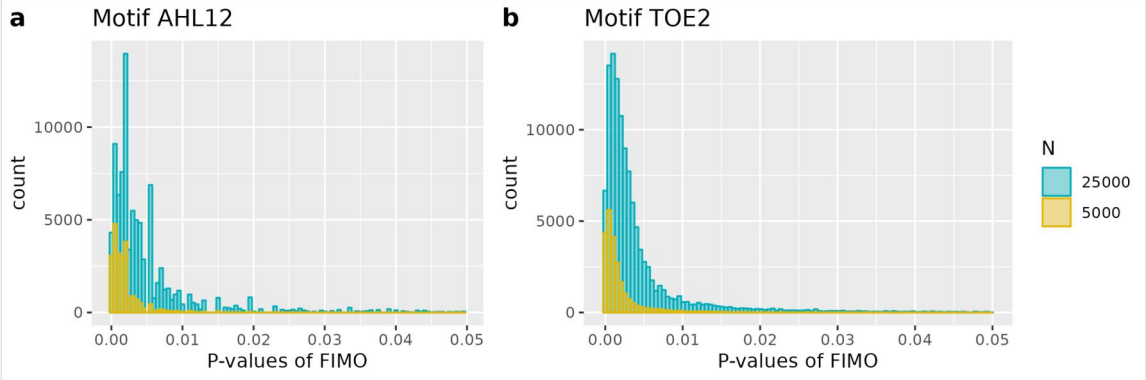


Figure 21: Distribution of p-values obtained from FIMO for motifs (a) *AHL12* and (b) *TOE2*, comparing the top 5,000 versus 25,000 promoters with hits limited to k per promoter. The analysis suggest a trend towards less significance as N increases from 500 to 27,000 in both cases.

PMET's analysis of the impact of varying promoter counts (N) on the identification of significant motif pairs reveals a direct, positive correlation: as N increases from 500 to 27,000, so does the number of significant pairs. Yet, there is an optimal range; beyond a certain N, a decline in significant pairs is observed, suggesting an upper limit to the beneficial effects of including more promoters. This trend holds across various gene sets, including immune-responsive and cell type-specific groups, as depicted in Figure S1a and b. Notably, when examining random gene sets, the occurrence of significant motif pairs remains negligible, irrespective of the N value chosen (Figure S1e).

Remarkably, for genes down-regulated during heat stress, a notable increase in the count of significant motif pairs is observed as the analyzed promoter number N rises from 10,000 to 25,000, reaching as many as 300 significant pairs. This marked increase underscores the sensitivity of PMET to promoter set size in specific stress responses. Conversely, such a trend is absent when

examining down-regulated genes under salt stress, indicating a differential impact of promoter count based on the type of stress.

In the PMET pairing analysis for motifs *AHL12* and *TOE2*, 18 cortex-specific genes were identified within the top 5,000 promoters exhibiting binding affinity for these motifs. Given the genome size and the number of cortex-specific genes as detailed in Table 7, the p-values derived from the hypergeometric test have been adjusted using the Bonferroni Correction method. It is considered that the pair of *AHL12* and *TOE2* is not significant since the adjusted p-value is 1.

When comparing the full lists of significant motif pairs for 5,000 and 25,000 promoters, Figure 23a illustrates a marked increase in the number of significant motif pairs ($p \leq 0.05$) associated with the cortex gene cluster as Nexpands from 5,000 (yellow) to 25,000 (blue). Since N is integral to the hypergeometric test, an increase in N necessitates a larger gene count to confirm motif pair enrichment within the cortex cluster. At N equal to 25,000, approximately 400 genes are implicated for a motif pair to be considered significant, despite the motifs being lower-ranked in binding strength (Figure 23b). This enrichment pattern is consistent across all cell type-specific gene categories, as depicted in Figure 23c and d.

Table 7: Parameters of hypergeometric test of motif pair *AHL12* and *TOE2*

Cell type	Cortex
Motif 1	AHL12
Motif 2	TOE2
Size of *A. thaliana* Genome	around 27,000
N	5,000
Number of cortex genes with both motifs	18
Total number of genes with both motifs	637
Number of cortex genes	439
Raw p-value	0.0217
Adjusted p-value (Bonferroni)	1
Genes with both motifs	AT1G01360;AT1G19450;AT1G47480;AT1G76490; AT2G22470;AT2G31730;AT2G33990;AT2G34070; AT3G18830;AT4G00430;AT4G01110;AT4G32480; AT4G35250;AT5G23220;AT5G47370;AT5G51460; AT5G65380;AT5G66580;

The expansion of N to include a larger set of promoters could potentially increase the number of motif pairs, incorporating those with lower binding affinities. This inclusion often leads to statistical significance within the gene sets analyzed for motif pairs. However, not all statistically significant motif pairs may reflect biologically relevant features.

2.7.4 Role of 5'UTR Regions

In this work, the PMET analysis considered the presence of motif pairs within the 5' UTR regions downstream of the TSS. In comparison to the effects induced by other parameters in PMET analysis, the inclusion or exclusion of the 5' UTR regions has a more direct impact on the quantity of motif pairs. Specifically, when the 5' UTR regions are considered, there is a general trend of decreased abundance of motif pairs across different gene clusters (Figure 24).

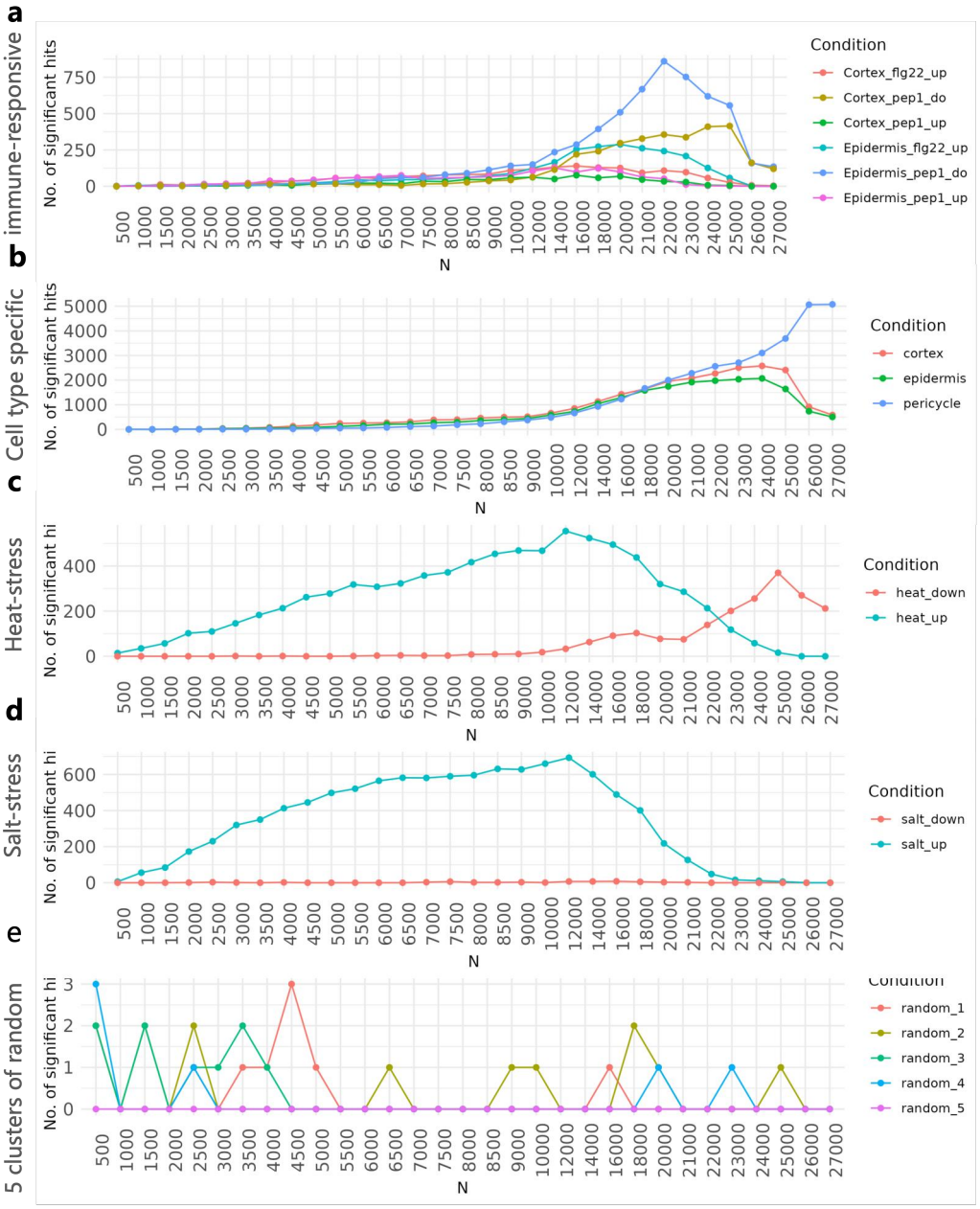


Figure 22: Relationship between the number of promoters analyzed (N) and the count of significant motif pairs identified in different gene clusters. (a) to (d) A general increase in motif pair detection with a larger promoter set size for specific gene types, followed by a subsequent plateau or decline beyond a certain threshold. (c) and (d) A fluctuant line suggests a differential impact under heat stress for down-regulated genes. (e) For random gene clusters, no clear trend, highlighting the specificity of the PMET approach to biologically relevant gene sets.

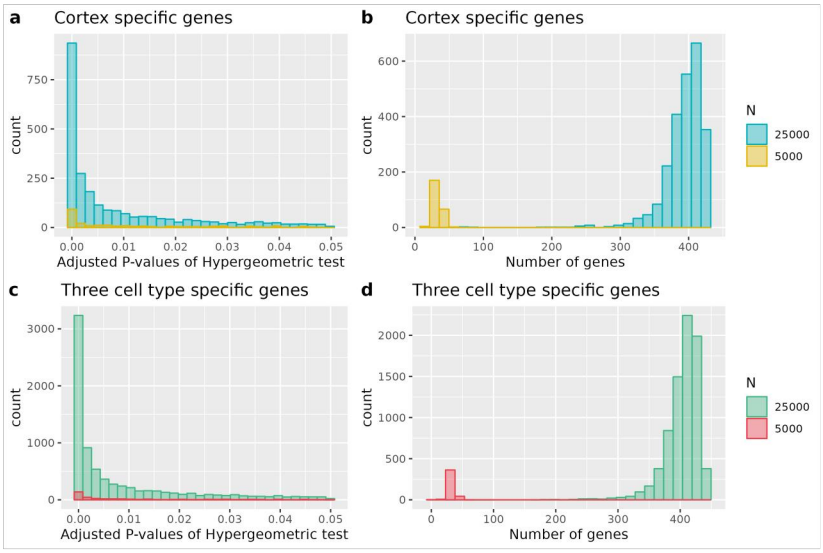


Figure 23: Distribution of significance and corresponding gene counts for significant motif pairs across the top 5,000 versus 25,000 promoters reveal a notable shift in motif enrichment among cortex-specific genes and provide a collective view of three cell type-specific genes.

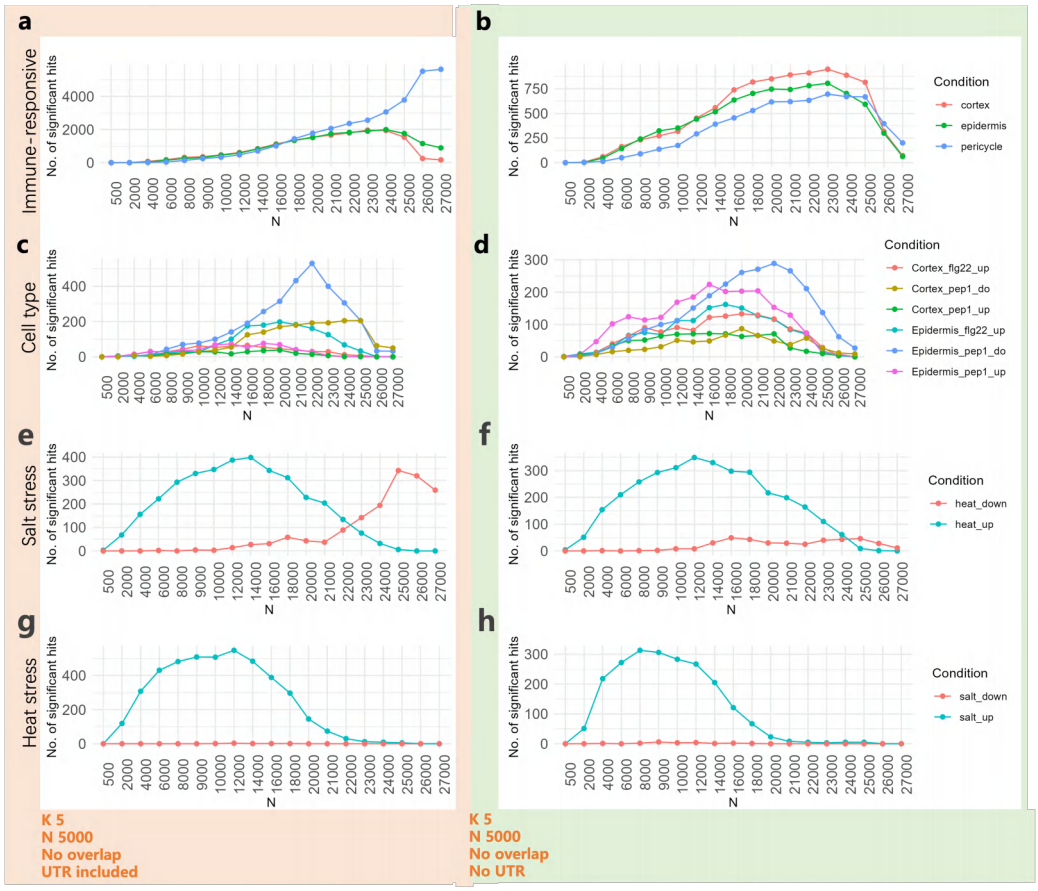


Figure 24: Impact of UTR on the detection of significant motif pairs in immune-responsive, cell type-specific, salt-, and heat stress-induced genes across varying promoter lengths, with k=5 and N=5,000. (a), (c), (e), and (g) show results without UTR, peaking around a promoter length of 1,000 bp. (b), (d), (f), and (h) show results with UTR, indicating a significant reduction in motif pair detection, highlighting the negative impact of UTR on PMET analysis of motif pairs.

2.8 Genomic Localization Patterns of Motif Pairs on Promoters

PMET distinguishes itself among motif search tools for its versatility across various configurations, particularly in terms of specifying the proximity to TSS. A new parameter, termed the TSS gap, has been introduced in this work, defined as the region extending upstream from TSS for a specified length, excluding it from PMET calculations. The introduction of this parameter helps to study the enrichment of motif pairs in different regions of the promoter. The TSS gap often set within a range of 0 to 1,000 bp in PMET analysis, allowing users to set this parameter to any value that aligns with their research requirements, enabling customized searches tailored to their unique investigational needs. Rigorous PMET analyses have been executed on diverse gene sets, encompassing immune-responsive, cell-type specific, and stress-regulated genes, employing an array of promoter lengths and TSS distances. As depicted in Figure 25, the outcomes consistently feature conspicuously dark blue blocks on the heatmap's left side, indicative of a higher concentration of significant motif pairs. This pattern is particularly pronounced within the heat-stress and salt-stress induced gene clusters. A significant reduction in the motif pair numbers is observed when the TSS gap exceeds 200 bp, as illustrated in Figure 25j and m, suggesting a predominant localization of binding sites closer to the TSS within the promoter region.

In addition to the significant influence of promoter length on motif pair enrichment, as addressed in Section 2.7.1, the distance to the TSS also exerts a crucial influence. Most of the color changes in the heat map predominantly follow a consistent pattern: an increase in the distance to TSS and promoter length correlates with a gradual decrease in the number of motif pairs, manifesting as a progressive fading of colors. Concurrently, it is important to note that the number of motif pairs in random gene sets is negligible and lacks the before-mentioned pattern, highlighting the specificity of the observed enrichment to the gene sets under investigation.

2.9 Distribution of Motif Pairs on Genomic Elements

TFs have been traditionally acknowledged for their pivotal role in gene regulation, which involves interacting with specific DNA sequences and modulating gene expression through the recruitment of coactivators or corepressors[230,231]. However, recent research has unveiled additional functionalities of TFs, demonstrating their ability to bind with RNA[232,233,234,235,236,237] but also the 5'UTR[238].

Inspired by these findings, the study in question has delved into the broader spectrum of TF activity by examining the potential motif pairs enrichment patterns of other genomic elements, such as the coding region (CDS), 3' untranslated region (UTR), and 5' UTR, during PMET analyses. This methodology is designed to reveal a more holistic view of the regulatory landscape governed by TFs, extending beyond traditional gene promoter regions to include a more extensive range of genomic elements.

To this end, analyses were conducted using PMET across different genomic element, including 3' UTR, 5' UTR, CDS, and exon, with the main parameters outlined in Table 3. The Plant Cistrome Database with 872 motifs, was chosen as the primary motif database due to its extensive

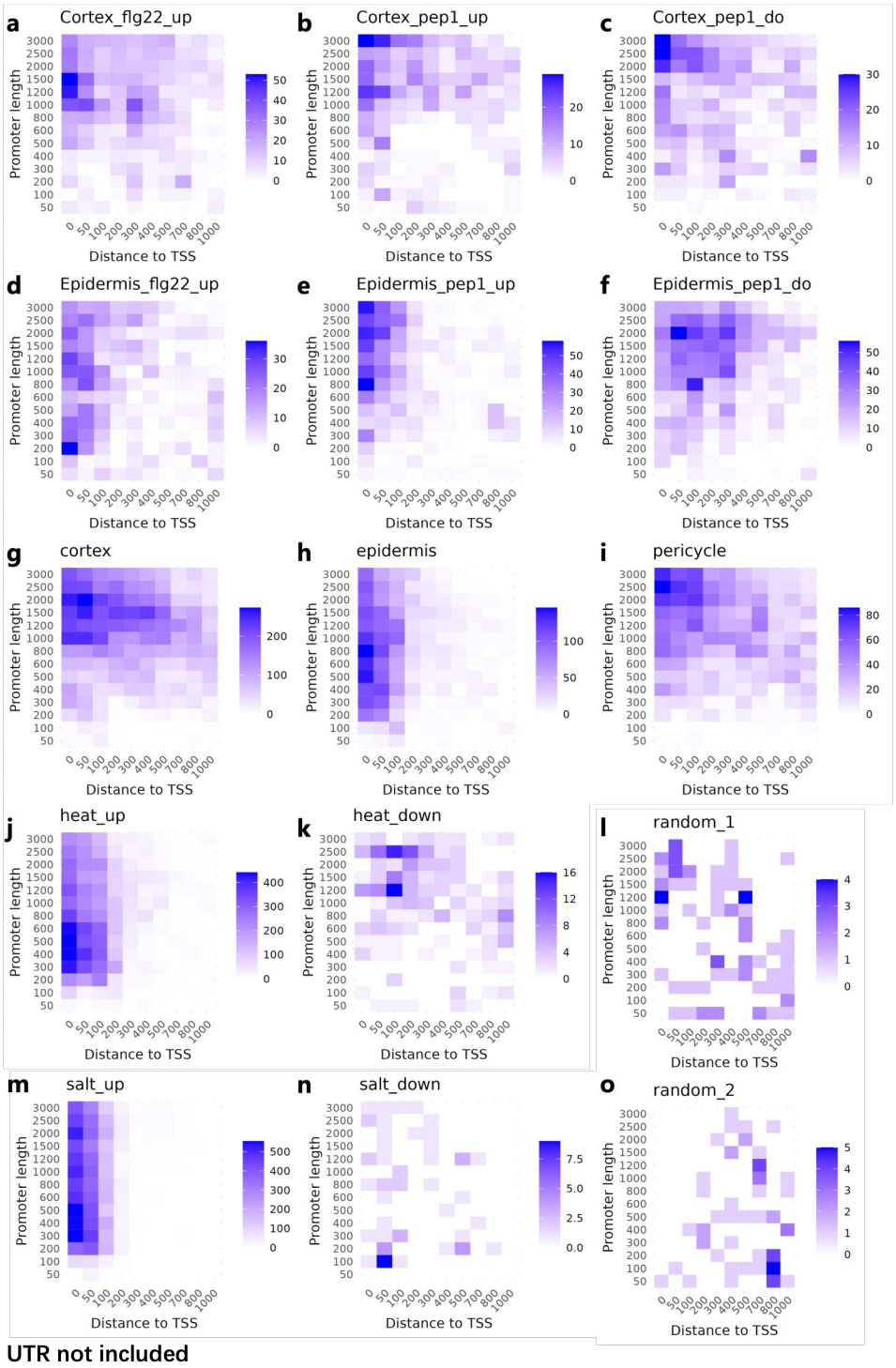


Figure 25: Comprehensive PMET analyses reveal the distribution of motif pairs in relation to promoter length and proximity to the TSS across various gene sets. (a-c) Heatmaps for the cortex show the density of motif pairs with a noticeable aggregation near the TSS, especially for genes up-regulated under flg22 treatment. (d-f) A similar trend is observed in epidermis cells. (g-i) In pericycle cells, the highest concentration of motif pairs is found closest to the TSS. (j, k, m, and n) Heatmaps display the effects of heat and salt stress on motif pair distribution, with dense clusters diminishing beyond 200 bp upstream of the TSS. (l, o) Random gene sets demonstrate a stark contrast, displaying a dispersed and sparse presence of motif pairs, which underscores the non-random nature of motif pair localization in the promoter regions of stress-responsive genes. These patterns, particularly dense pattern near the TSS, underscore the role of promoter architecture in gene regulation under stress conditions.

collection of experimentally derived motifs, detailed categorization, and comprehensive dataset. Importantly, alternative splicing leads to the generation of multiple gene models for each genomic element. For the analyses, the longest gene model was selected from each gene to represent it in PMET, to validate the hypothesis of a comprehensive TF regulatory network.

The PMET analysis of promoter regions yielded three datasets of motif pairs regarding cell types of cortex, epidermis and pericycle, each encompassing 379,756 potential motif pairs, calculated as $872 \times (872 - 1) \div 2$, derived from the Plant Cistrome Database. Focusing on the cortex as an example, applying a p-value adjustment threshold of 0.01 resulted in the retention of 7,249 motif pairs. Similar adjustments yielded 12,131 motif pairs for the epidermis and 2,710 for the pericycle, as illustrated in Figure 26a.

In investigating the regulatory potential of genomic elements beyond the conventional promoter regions, PMET analyses revealed a variable number of motif pairs across gene clusters of different cell types. The absence of a consistent pattern among these gene clusters implies the intricate and variable nature of the regulatory mechanisms at play. 25000 Significantly, the identification of motif pairs within the 3' UTR and 5' UTR regions was found to be infrequent across various gene clusters, as shown in Figure 26a. This observation can be attributed to the shorter sequence lengths characteristic of these regions, with the 3' UTRs averaging 313.24 bp and the 5' UTRs averaging 248.82 bp in length, as shown in Figure 27a,b.

Upon comparing motif pairs derived from genomic elements beyond promoters to those from promoters, a notable variable degree of overlap was observed across all gene clusters. As illustrated in Figure 26b, motif pairs originating from diverse genomic elements, including exons, exhibit significant overlap with motif pairs from promoters within three gene clusters. This overlap suggests a shared regulatory potential between these genomic regions, indicating that exon regions, traditionally not the primary focus for motif pair analysis, may also play a significant role in gene regulation.

In contrast to gene clusters associated with cell types, those involved in heat and drought stress response exhibit greater directionality and comparability. Notably, the cluster of down-regulated genes under drought stress has higher number of associated motif pairs than the up-regulated gene cluster. It is worth noting that the quantity of motif pairs in the promoter region is relatively scarce, as evidenced in Figure 28.

2.10 Heat Stress Genes

It has been introduced that the affinities of a motif to promoters are quantified by the bindings' p-value calculated by 'FIMO' in PMET indexing procedure. The parameter N is set to limit the number of promoters to which the motif binds. It is observed that the up- and down-regulated genes in *A. thaliana* under heat stress condition present an interesting pattern when studying the impacts of the varied parameter N. When the parameter N was set to a relatively small number, i.e 500 to 12,000, the motif pairs co-appearing in the promoters of up-regulated genes accumulated while there were almost no motif pairs co-appearing in promoters of down-regulated genes. A

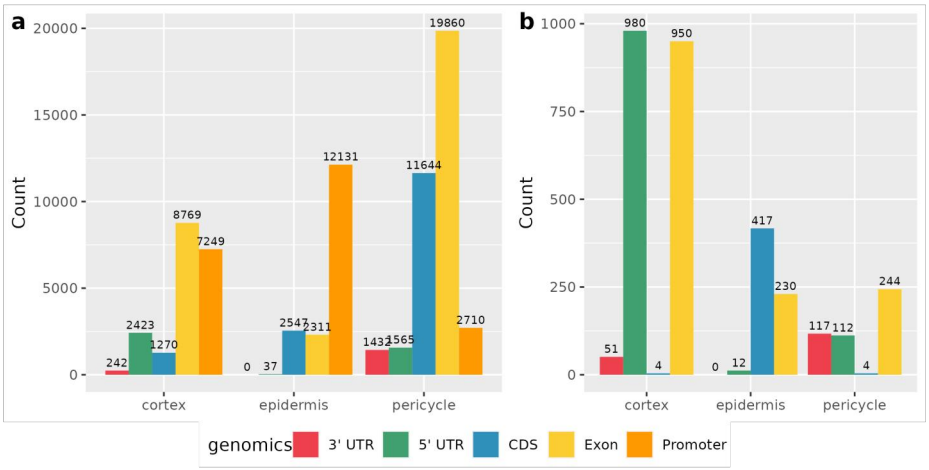


Figure 26: Comparative analyses of motif pairs across genomic elements reveal distinct distributions. The motif database, consisting of 872 motifs from the Ecker Lab (2016)[161], was utilized for PMET analyses. (a) Count of motif pairs identified within different genomic elements (promoter, 3'UTR, 5'UTR, CDS, and exons) varies across three specific cell types (cortex, epidermis, and pericycle). Notable differences in motif pair counts suggest distinct regulatory landscapes across these genomic regions. (b) Number of motif pairs shared between promoters and other genomic elements indicates potential overlaps in regulatory motifs across different genomic structures. The high prevalence of shared motif pairs, particularly in exon regions, underscores the complexity of gene regulation that extends beyond traditional promoter-centric models.

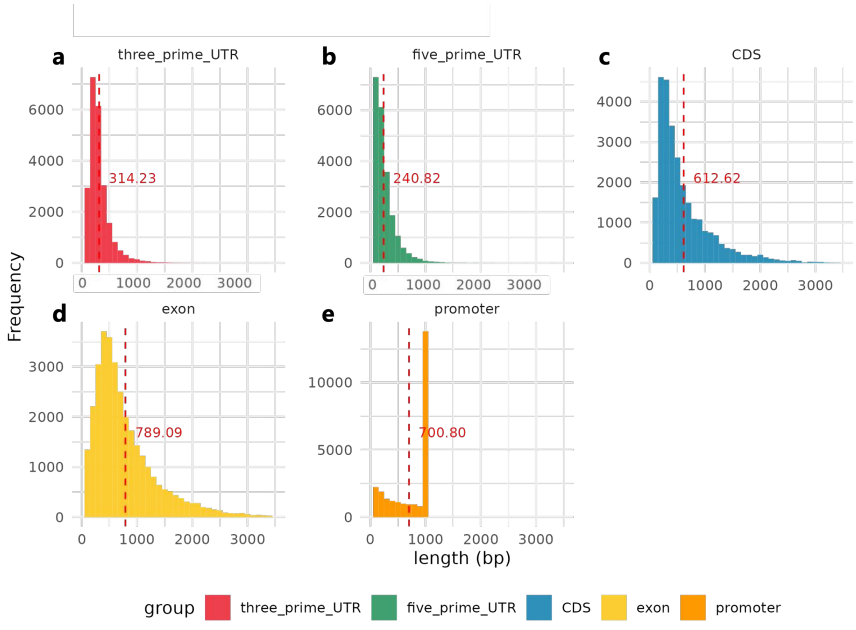


Figure 27: Length distribution of different genomic elements in *A. thaliana*.Each panel represents the frequency of genomic element lengths for promoters,3'UTRs,5'UTRs,CDS,exons,and promoters. Dashed lines on each histogram indicate the average length of the respective genomic elements,with values provided in red. Histograms highlight the variability in length across different types of genomic elements,which may have implications for the binding and regulatory potential of TFs.

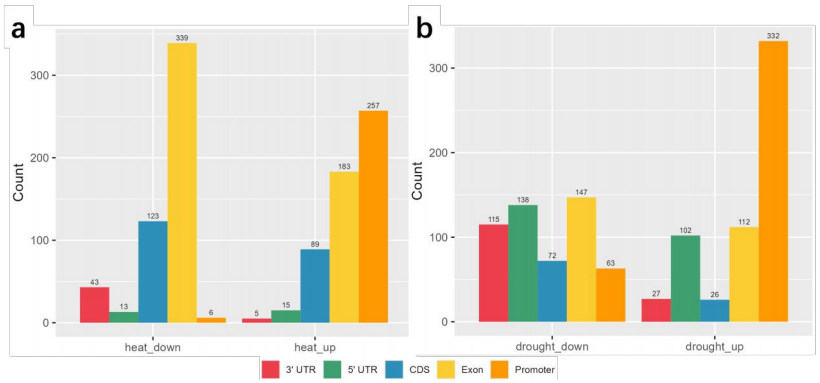


Figure 28: Motif pairs across genomic elements in genes differentially regulated under heat and drought stress. (a) Bar chart displays counts of motif pairs identified within different genomic elements (3' UTRs, 5' UTRs, CDS, and exons) in heat stress-regulated gene clusters. The motif database consists of 182 motifs derived from clustering in the previous section. (b) Bar chart illustrates the number of motif pairs shared between promoters and other genomic elements in drought stress-regulated gene clusters, indicating potential overlap in regulatory motifs across different genomic structures.

sharp increase of motif pairs co-appearing in down-regulated genes was noticed when promoters with weaker affinities were included. Especially, there is a peak of motif pairs when N was set to 25,000 (Figure 29a).

2.10.1 GO Enrichment Analysis of Heat-Stress-Induced Genes in *Arabidopsis thaliana*

This work conducted a gene ontology (GO) enrichment analysis on the up- and down-regulated genes in *A. thaliana* in response to heat stress. The primary GO terms associated with heat stress up-regulated genes predominantly relate to the cellular response to specific environmental stresses or signaling. Responses to environmental stress correspond to the plant's adaptive mechanisms to changes in the environment, such as heat, drought, salt, and hypoxia. There is also an association with abscisic acid signaling, as indicated by the GO term "abscisic acid-activated signaling pathway" (Figure 30a). The main GO terms related to heat stress down-regulated genes, as depicted in Figure 30b, are predominantly associated with biochemical and metabolic pathways. Additionally, they include functions specific to plants, such as the synthesis of glucosinolates and xyloglucans. The GO analysis results for the two gene clusters exhibit distinct differences.

2.10.2 Induced Genes under Heat Stress of Other Species

For a more comprehensive understanding, induced genes under heat stress conditions of *Solanum tuberosum* and *Triticum aestivum* were employed to conduct the PMET analyses with 113 motifs from Franco-Zorrilla et al. (2014)[157], and similar patterns of the varying number of motif pairs were observed in Figure 29b,c. Subsequently, different parameters N were selected for the up- and down-regulated gene sets of each species to maximize motif pairs. For example, in Figure 29a, when N is set to 14,000, the number of motif pairs corresponding to up-regulated gene set (green

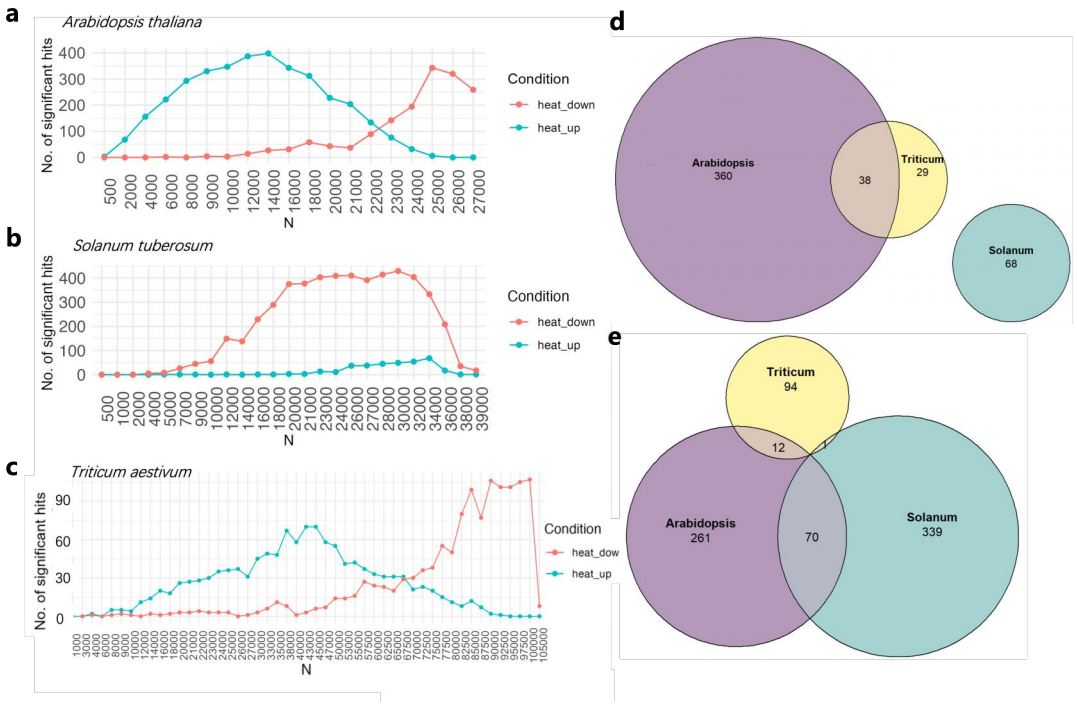


Figure 29: Effect of the changing parameter N on the number of motif pairs in up- and down-regulated genes in three plant species exposed to heat stress. (a-c) Numbers of motif pairs under Effect of parameter N on the number of motif pairs in up-(red) and down-regulated (green) genes of *A. thaliana*, *Solanum tuberosum*, and *Triticum aestivum* upon heat stress. (d) Shared motif pairs within up-regulated genes of the three species when N was set to maximize the number of motif pairs for each species. (e) Shared motif pairs within down-regulated genes of the three species when N is set to maximize the number of motif pairs for each species. All motif pairs identified through PMET analyses were filtered with a p-value threshold of < 0.01.

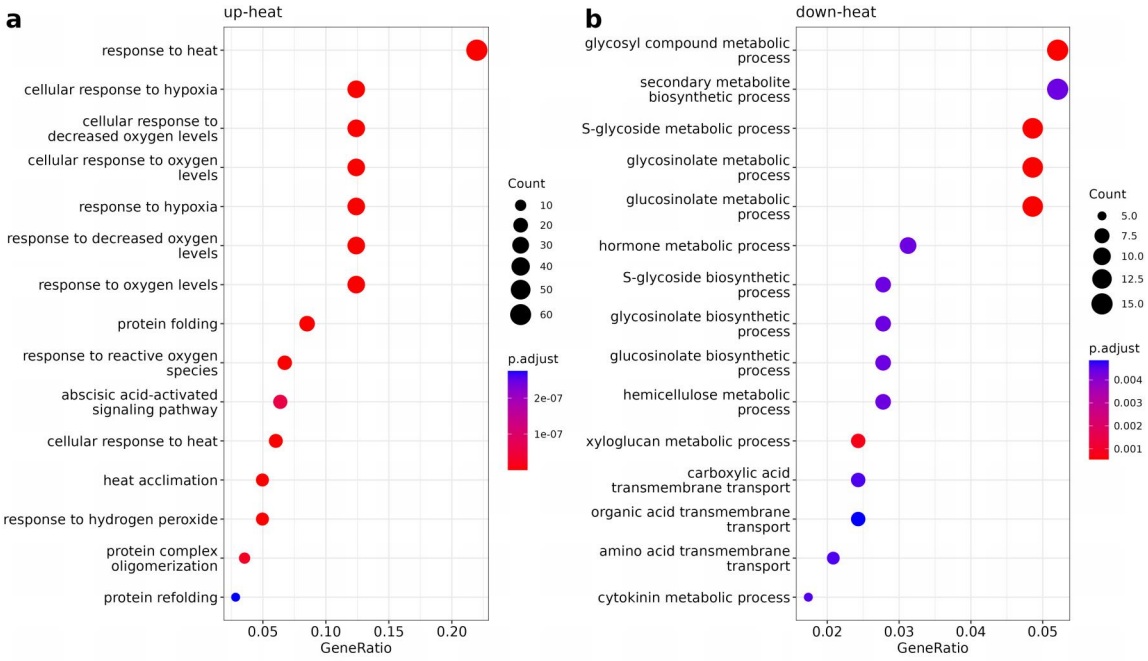


Figure 30: GO analyses for up- (a) and down-regulated (b) genes in *A. thaliana* under heat stress, revealing a significant association between the heat response in up-regulated genes and a variety of metabolic processes in down-regulated genes.

line) of *A. thaliana* is 398. For down-regulated gene set (red line), N is set to 25,000 for most motif pairs.

This study further employs a word cloud approach to visually represent the frequency of interactions between TF families and the occurrence of individual motifs within those selected motifs pairs associated with down-regulated genes under heat stress (Figure 31). The interaction between TF families is determined by extracting the family information from both members of a motif pair. For example, from the specific pair *ANAC46-MYC4*, only the TF family information is retained, resulting in the generalized pair *ANAC-MYC*. The word cloud visualization reveals that *AHL* and *MYC* are the two TF families most frequently forming motif pairs across three species. Within these motif pairs, individual motifs such as multiple members of the *AHL* family, *WOX13*, *MYC4*, and members of the *ANAC* family are particularly active. In the analysis of 410 motif pairs of *Solanum tuberosum*, it was found that 326 of these motif pairs contain at least one member belonging to the *AHL* family. This finding explains why words such as 'ahl12', 'ahl20', and 'ahl25' dominate prominently in Figure 31e. Similarly, in the work of 107 motif pairs of *Triticum aestivum*, motif pairs of the *ANAC* family appeared with a frequency of 43, which is visually represented in Figure 31d.

Figure 29d,e depict the shared motif pairs of three plant species in up- and down-regulated gene sets under heat stress with the optimal N respectively. In Figure 29d, the purple and yellow circles represent the largest numbers of motif pairs corresponding to the heat-induced up-regulated gene sets in *A. thaliana* and *Triticum aestivum*, respectively. The intersection between them consists of 38 motif pairs (Table 8), almost all of which contain heat shock factors (HSFs). Additionally, both *MYC3* and *MYC4* are frequently present in these motif pairs. An examination of 68 motif pairs from *Solanum tuberosum* (Table 9) revealed no motifs associated with HSF.

In a similar manner, the composition of 70 common motif pairs in down-regulated gene sets under heat stress for both *A. thaliana* and *Solanum tuberosum* (purple and green circles in Figure 29e) was analyzed. These motif pairs included multiple members of the *AHL* family (*AHL12*, *AHL20*, and *AHL25*) as well as other TFs, such as HSFs, *MYC3*, *MYC4*, *SPL7*, and *TCP15* (Table 10). The overlap between motif pairs from down-regulated genes in both *A. thaliana* and *Triticum aestivum* (purple and yellow circles) is detailed in Table 11. The frequent presence of *MYC4* in motif pairs may indicate its special regulatory role.

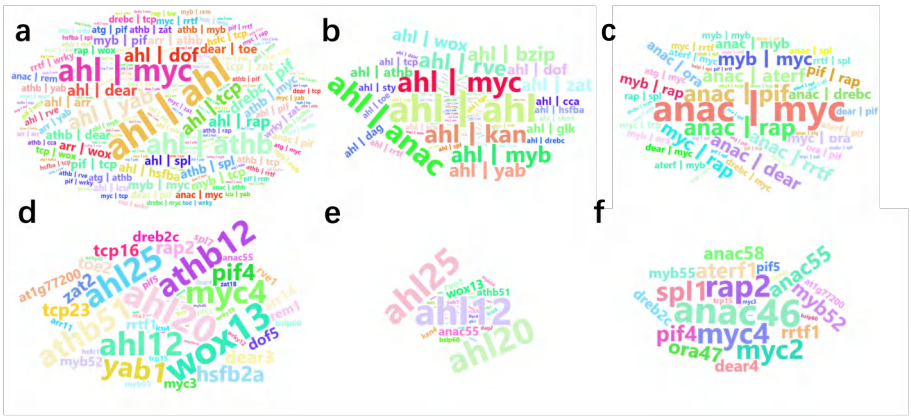


Figure 31: Word cloud of motif pairs and individual motifs associated with down-regulated genes in *A. thaliana*, *Solanum tuberosum*, and *Triticum aestivum* under heat stress. (a) Frequency of TF family pairs from motif pairs, detected in down-regulated *A. thaliana* genes under heat stress, with a notable occurrence of *AHL-AHL* and *AHL-MYC* at $N = 25,000$. (b) High frequency of *AHL-ANAC* and *AHL-MYC* motif pairs, found in down-regulated genes in *Solanum tuberosum* under heat stress at $N = 26,000$. (c) Frequent presence of *ANAC-MYC* motif pairs detected in down-regulated *Triticum aestivum* genes under heat stress at $N = 100,000$. (d)-(f) Motif derived from motif pairs with the highest counts for the corresponding species: *AHL* family members, *WOX13*, and *MYC4* for *A. thaliana*; *AHL* family members for *Solanum tuberosum*; *ANAC* family members and *MYC4* for *Triticum aestivum*.

2.11 Discussion

The regulatory functions and expression patterns of genes are often determined by multiple TF bindings within a gene region. PMET, a tool developed for this purpose, focuses on the statistical estimation of dual TF bindings. Prior research generally emphasized the binding of individual TFs, relying on the presence of these bindings for functional inference. For instance, FIMO identifies motifs in gene regions and infers regulatory functions based on the presence of these motifs. However, the statistical significance of a motif does not guarantee its functionality, as predictions of excessive significant bindings often happen and these are not confined to particular genes.

PMET addresses these issues through three key approaches:

1. **Limiting individual motif hits per promoter (K):** The number of individual motif hits per promoter is restricted to reduce the frequency of motif hits per genes. In practice, k is set to 5, based on adjusted p-values, capping the hits of each motif to a promoter at a maximum of 5.

2. **Binomial test for probability calculation:** The likelihood of k hits is computed using a binomial test, which provides a score that aids in limiting the number of hits to a gene. The default setting for N is 5,000, allowing a motif to match up to 5,000 genes.

3. **Hypergeometric distribution for shared gene clusters:** This distribution is used to describe the co-occurrence of two motifs within a specific gene cluster, enhancing the specificity of the analysis.

Table 8: 38 shared motifs pairs in promoters of up-regulated genes in *A. thaliana* and *Triticum* under heat stress (intersection of purple and yellow circles in Figure 29d).

Cluster	Motif1	Motif2
heat_up	HSFB2A_2	HSFC1_2
heat_up	HSFB2A	HSFB2A_2
heat_up	HSFB2A	PIF3
heat_up	HSFC1_2	PIF3
heat_up	HSFB2A_2	PIF3
heat_up	HSFC1	MYC4
heat_up	ANAC46	HSFB2A
heat_up	BZIP60	HSFC1_2
heat_up	BZIP60_2	HSFB2A
heat_up	HSFB2A_2	MYC4
heat_up	BZIP60	HSFB2A
heat_up	BZIP60_2	HSFC1_2
heat_up	ANAC46	HSFC1_2
heat_up	BZIP60_2	HSFC1
heat_up	BZIP60_2	HSFB2A_2
heat_up	HSFB2A_2	TGA2_2
heat_up	HSFC1_2	MYC4
heat_up	HSFB2A_2	MYC3
heat_up	HSFC1_2	MYC3
heat_up	DEAR4_2	HSFB2A
heat_up	ANAC55	HSFB2A
heat_up	BZIP60	HSFB2A_2
heat_up	ANAC46	HSFB2A_2
heat_up	DREB2C	HSFB2A
heat_up	HSFB2A	MYC4
heat_up	DREB2C	HSFC1
heat_up	ANAC55	HSFC1_2
heat_up	DEAR4_2	HSFC1_2
heat_up	HSFC1_2	TGA2
heat_up	DREB2C	HSFB2A_2
heat_up	AT1G77200	HSFB2A
heat_up	AT1G77200	HSFB2A_2
heat_up	HSFB2A	REM1_2
heat_up	HSFB2A	ORA47
heat_up	HSFC1_2	TGA2_2
heat_up	HSFC1_2	TCP23
heat_up	DAG2	HSFC1_2
heat_up	DAG2	HSFB2A

2.11.1 Statistical Models

When a motif in PWM format is scanned across a promoter region using FIMO, it can yield an extensive number of hits, often exceeding 100. These hits are limited to k based on the p-values calculated by FIMO. This capping is crucial because unlimited motif hits can lead to false positives, complicating the analysis and interpretation of results. By assuming each motif hit within a promoter region as an independent and uniformly random event, a more conservative approach was adopted, focusing on the most statistically significant hits. The geometric mean of the p-values for k hits is used as the probability of random matching of that motif in the promoter region. This probability assessment serves as a critical filter, enabling us to distinguish between genuinely significant motif occurrences and random hits that are statistically less relevant.

This scenario is modeled using a binomial distribution, a statistical model well-suited to scenarios where outcomes are classified as 'success' or 'failure —for PMET analyses, a motif hit is considered a 'success'. Similarly, a Poisson distribution is used to model the occurrence of motif hits, particularly effective when events occur independently and the number of events in a fixed interval is of interest. The decision to use both binomial and Poisson distributions in PMET underscores the commitment to a rigorous statistical approach, ensuring that the results are both reliable and reproducible.

Table 9: 68 motifs pairs in promoters of up-regulated genes in *Solanum tuberosum* under heat stress. (green circle in Figure 29d)

cluster	motif1	motif2
heat_up	AHL20	DAG2
heat_up	AHL12_2	DAG2
heat_up	DAG2	RVE1
heat_up	AHL12_2	RVE1
heat_up	DAG2	ICU4
heat_up	AHL20	STY1
heat_up	AHL20_2	DAG2
heat_up	AHL20_3ARY	DAG2
heat_up	AHL25_2	DAG2
heat_up	ATHB51	RVE1
heat_up	RVE1	WOX13_2
heat_up	AHL12_2	CCA1
heat_up	AHL12	DAG2
heat_up	AHL12_2	STY1
heat_up	AHL12_3ARY	DAG2
heat_up	AHL20	ATHB51
heat_up	AHL20	SPL1_2
heat_up	AHL20_2	STY1
heat_up	AHL25	DAG2
heat_up	DAG2	YAB1
heat_up	ICU4	RVE1
heat_up	AHL12	STY1
heat_up	AHL12_2	WOX13
heat_up	AHL20	DOF5.7_2
heat_up	AHL20_3ARY	STY1
heat_up	AHL25_2	STY1
heat_up	ANAC55	RVE1
heat_up	ATHB51	CCA1
heat_up	ATHB51	DAG2
heat_up	ATHB51	STY1
heat_up	CCA1	DAG2
heat_up	AHL12_2	SPL1_2
heat_up	AHL20	AT5G28300
heat_up	AHL20_2	RVE1
heat_up	AHL20_2	SPL1_2
heat_up	AHL20_3ARY	SPL1_2
heat_up	AHL25_2	ATHB51
heat_up	AHL25_2	RVE1
heat_up	ATHB12	DAG2
heat_up	RVE1	WOX13
heat_up	AHL12	RVE1
heat_up	AHL12_2	WRKY45
heat_up	AHL12_3ARY	STY1
heat_up	AHL20	SPL1
heat_up	AHL20	WOX13
heat_up	AHL20	ZAT14
heat_up	AHL20_3ARY	RVE1
heat_up	AHL25_3ARY	STY1
heat_up	ATHB12	RVE1
heat_up	ATHB51	SPL1_2
heat_up	RVE1	SPL1_2
heat_up	AHL20	DEAR3_2
heat_up	AHL12_3ARY	RVE1
heat_up	AHL20	LBD16
heat_up	AHL20	RVE1
heat_up	AHL20	WRKY45
heat_up	AHL25	STY1
heat_up	ATHB51	SPL1
heat_up	ICU4	STY1
heat_up	RVE1	WRKY45
heat_up	AHL12_2	AHL20_2
heat_up	AHL12	WOX13_2
heat_up	AHL20_3ARY	ICU4
heat_up	CCA1_2	RVE1
heat_up	AHL20_3ARY	ATHB51
heat_up	AHL12	AHL12_2
heat_up	AHL20_2	WOX13_2
heat_up	AHL20	AHL25_3ARY

Table 10: Top 15 out of 70 motifs pairs in promoters of down-regulated genes in *A. thaliana* and *Solanum tuberosum* under heat stress. (intersection of purple and green circles in Figure 29e)

Cluster	**Motif1**	**Motif2**
heat_down	AHL12_3ARY	MYC4
heat_down	AHL12_3ARY	TCP23
heat_down	AHL12_3ARY	YAB1
heat_down	AHL25	MYC4
heat_down	AHL12_3ARY	SPL7
heat_down	AHL12_3ARY	TCP15
heat_down	AHL20_2	SPL7
heat_down	AHL25	HSFB2A
heat_down	AHL25	MYC3
heat_down	AHL25_3ARY	SPL7
heat_down	AHL25_3ARY	MYC3
heat_down	AHL25_3ARY	MYC4
heat_down	AHL12_3ARY	DOF5.7
heat_down	AHL12_3ARY	HSFB2A
heat_down	AHL20_2	MYC3
heat_down	AHL20_3ARY	MYC3
heat_down	AHL20_3ARY	MYC4
heat_down	AHL25_3ARY	HSFB2A
heat_down	AHL20	MYC4
heat_down	AHL20_2	HSFB2A
heat_down	AHL20_2	MYC4
heat_down	ATHB51	HSFB2A
heat_down	ATHB51	MYC4
heat_down	AHL20_2	GLK1
heat_down	ANAC55	YAB1
heat_down	AHL20	DOF5.7
heat_down	AHL25	YAB1
heat_down	RVE1	YAB1
heat_down	AHL25_3ARY	DOF5.7
heat_down	AHL25_3ARY	YAB1

Table 11: 12 shared motifs pairs in promoters of down-regulated genes in *A. thaliana* and *Triticum aestivum* under heat stress (intersection of purple and yellow circles in Figure 29e).

Cluster	**Motif1**	**Motif2**
heat_down	AT1G77200	MYC4
heat_down	DREB2C	MYC4
heat_down	MYB52	MYC4
heat_down	MYC4	RAP2.3
heat_down	AT1G77200	PIF4
heat_down	MYC4	RRTF1
heat_down	MYC4	RAP2.6
heat_down	MYC4	TCP15
heat_down	ANAC58	MYC4
heat_down	DREB2C	PIF4
heat_down	PIF4	TCP15
heat_down	MYB46_2	MYB52

As observed in Section 2.6, nearly identical results were obtained across different gene clusters. In genomics, the Poisson distribution is often used to model random events over entire genomes or large regions, such as gene mutations and TF binding. Such applications highlight the suitability of this model for genomic analyses where the events of interest occur infrequently. Mathematically, the Poisson distribution is derived from the binomial distribution with large trials and a small probability of success. The congruence of two statistical results is not only reassuring but also indicative of the accuracy of PMET in identifying biologically meaningful motif patterns.

The integration of these statistical models into PMET enables PMET to confidently identify and interpret motif pairs that are truly significant in gene regulation. Such specificity is crucial for unraveling the complex mechanisms of gene expression and regulation, paving the way for deeper insights into the intricate networks that govern cellular processes.

2.11.2 Threshold of Motif Clustering

In addressing the potential redundancy among 1,565 motifs, hierarchical clustering was utilized to construct a coherent hierarchical motif tree. Branches at specific levels of this tree can be selected to identify non-redundant motifs. Hierarchical clustering offers the advantage of gradually building a cluster hierarchy without the need to predefine the number of clusters, making it well-suited for aggregating and categorizing motifs from diverse databases. This method iteratively merges the most similar clusters into new clusters until all motifs have been integrated, thereby generating a dendrogram.

The flexibility of this approach lies in its ability to accommodate various types of data through adjustments to distance metrics and merging strategies. Additionally, the hierarchical structure of merged clusters provides intuitive insights into motif relationships. However, a key challenge lies in determining the optimal number of clusters. Since motif similarity, calculated using Euclidean distance, lacks inherent biological or physical significance, by which setting a specific distance threshold to control clustering outcomes is not feasible.

The maximum Silhouette score corresponds to the optimal distance threshold for hierarchical clustering, which is around 0.5. The final number of clusters is approximately 150. The Gap Statistic method confirms that the number of clusters stabilizes after reaching 150, indicating that both methods converge on the optimal number of clusters (distance threshold). In actual PMET analysis, the goal is to input as many motifs as possible to achieve a greater number of homotypic matches, which forms the foundation for identifying heterotypic matches. Taking into account the requirement for a larger set of motifs, and integrating the results from Silhouette Analysis and Gap Statistic, the hierarchical clustering distance threshold is set to 0.7, allowing a certain degree of redundancy.

2.11.3 Impact of Promoter Length and Overlap with Other Genes

An intriguing observation from the analyses, as depicted in Figure 19, is the non-linear relationship between the number of motif pairs and promoter lengths across various gene clusters. It's

noteworthy that an increase in promoter length does not consistently lead to an increase in motif pairs. An incremental quantity of motif hits is particularly evident when considering solo TF binding within a certain length of promoters; longer promoters offer more potential binding sites in predictive scenarios. However, when exploring TF-TF interactions or motif pairs, the outcomes become less predictable. In the examination of immune-responsive genes, cell type-specific genes, and up-regulated genes (both heat- and salt-stress induced), a fluctuating pattern of motif pairs was observed with increasing promoter lengths. This fluctuation may stem from the reliance on the gene cluster on the presence or absence of specific TFs.

In scenarios where promoters are short, a slight extension in length can allow for the identification of additional, more compatible TFs that synergize with already predicted TFs. As the extension continues, the number of motif pairs initially increases and then subsides, eventually reaching a plateau. This pattern is likely associated with the significance of new TFs identified in longer promoters during the PMET indexing process. For example, motif A might be considered a significant hit for a promoter of length 1000. However, for the same promoter of length 2000, motif A may no longer be considered as significant due to the emergence of a more significant motif B. This outcome is a consequence of the top k selection criterion employed in PMET indexing. Once the plateau phase is reached, the number of motif pairs stabilizes, reflecting the biological reality that promoter lengths have natural limits. As detailed in Table 6, the average length of the sequence upstream of the TSS in *A. thaliana* is approximately 1672 bp. In Figure 19, all lines demonstrate a turning point within the range of 500-1500 bp, which is consistent with the biological constraints of promoter region sizes. This observation aligns with the findings of Korkuć et al. (2014), who estimated that the average length of the promoter region upstream of gene TSS in *A. thaliana* is around 500 bp[229], based on the density distribution of single-nucleotide polymorphisms (SNPs).

When overlaps with other genes are permitted, this intrinsic limitation on promoter length is effectively removed. As depicted in Figure 19, this adjustment leads to a marked reduction in the number of identified motif pairs, which can drop to near zero in some cases. This observation prompts two hypotheses.

The first hypothesis contends that motif hits extending into neighboring genes, despite achieving statistical significance, do not necessarily impart biological relevance. These hits are unlikely to participate in cooperative interactions within specific gene clusters, suggesting that statistical significance does not guarantee biological functionality. For instance, a simplified motif pairing model (Figure S4b) illustrates that a gene, labeled as *Gene 3* exhibits highly significant hits upon extension of the promoter region. Yet, *Gene 3* does not contribute to any specific gene cluster and subsequent motif hits for this gene are also absent, suggesting taht these hits are artifacts of the analysis rather than biological context of the gene cluster.

The second hypothesis posits that motif hits extending into neighboring genes are biologically significant, potentially establishing regulatory links with these genes and thus influencing their function. For instance, if *Gene 4* is part of a different cluster, motif pairs associated with *Gene*

4 may not contribute positively to the regulatory dynamics of the *Gene 2* cluster. This reflects a complex network of interactions where the biological significance of motif hits depends on the broader gene regulatory network (Figure S4b,d). This hypothesis suggests that, in addition to the promoter, other genomic elements also possess TF binding sites that serve regulatory functions.

Both hypotheses emphasize the intricate nature of motif interactions and the importance of considering the biological context when interpreting PMET results. This underscores the necessity for careful consideration of gene overlaps in the analysis, as they can profoundly impact the interpretation of TF-TF interactions and motif pair significance within gene clusters.

2.11.4 Negligible Motif Pairs of Down-regulated Genes under Stress Conditions

In the PMET analyses of *A. thaliana* genes subjected to heat and salt stress, down-regulated genes exhibited a almost complete absence of significant motif pairs (Figure 19c,d). This observation sharply contrasts with the results for immune-responsive and cell-type-specific genes, which showed a presence of motif pairs in both up- and down-regulation (Figure 19a,b). This contrast suggests that the regulatory mechanisms under stress conditions might be differ fundamentally and could be explained by two hypothesis.

The first hypothesis pertains to genetic specificity. Genes that are up-regulated in response to stress typically function to enhance plant resilience and promote survival. In contrast, genes that are down-regulated under stress conditions are often associated with non-essential or energetically demanding biological processes, including growth, development, and metabolism. This results in a heterogeneous set of down-regulated genes, each involved in distinct biological pathways. Consequently, the underlying regulatory mechanisms are likely to be diverse and less coordinated, as evidenced by the limited number of significant motif pairs detected in the PMET analysis. The observed divergence and lack of shared regulatory patterns among down-regulated genes may reflect a broader cellular strategy of resource reallocation—shifting priorities from growth-related functions toward survival during stress conditions.

This interpretation is consistent with findings by Peredo et al. (2020), who demonstrated that while the up-regulation of genes involved in protective functions is critical, it is not solely sufficient for achieving desiccation tolerance. Coordinated down-regulation across various metabolic domains is also required[239], in agreement with results from the GO enrichment analysis. As shown in Figure 30a, more than 20% of genes up-regulated under heat stress are annotated with the GO term 'response to heat'. Additional enrichment in terms such as 'cellular response to heat' and 'heat acclimation' further supports the notion of a targeted and specific transcriptional response under heat stress conditions.

In contrast, the down-regulated gene set is associated with a wide range of metabolic processes, with individual GO categories each accounting for a relatively small fraction of genes (up to 5%), as illustrated in Figure 30b. This pattern suggests a more diffuse and less coordinated transcriptional response. Supporting this interpretation, the PMET analysis of up-regulated genes under heat stress identified motif pairs involving heat shock factors (HSFs), including *HSFB*s and *HSFBC*s,

further reinforcing the hypothesis of genetic specificity in stress-responsive gene regulation.

The second hypothesis suggests that genes down-regulated under stress conditions are primarily regulated through epigenetic mechanisms rather than direct TF binding. Epigenetic regulation plays a pivotal role in facilitating rapid and reversible responses to environmental stimuli, and numerous studies have provided evidence supporting the involvement of epigenetic mechanisms in plant stress responses[240,241,242,243,244,245,246]. For example, research on transcriptional heat stress memory—mediated by the TF *HSFA2*—demonstrates how prior exposure to heat stress can confer enhanced tolerance to subsequent stress events[247].

Despite its conceptual appeal, this hypothesis remains only partially substantiated, as it does not fully account for the preferential targeting of down-regulated genes by epigenetic regulation. Rather than serving as a definitive explanation for the observed regulatory patterns, this hypothesis highlights the potential complexity and diversity of gene regulatory strategies employed during stress responses.

2.11.5 Low-Affinity Transcription Factors and Heat-Stress-Induced Genes

The parameter N is pivotal to PMET's algorithmic performance, and determining its optimal value is challenging.In the PMET indexing process, N specifies the maximum number of promoter hits permissible for a motif. An increase in N typically leads to an increase in the count of motif pairs, as it allows for the inclusion of hits with lower affinity to the promoter. Consequently, a steady rise in motif pairs is noted as N increases. This increase plateaus at a specific threshold depending on the gene cluster, such as at 12,000 for up-regulated *A. thaliana* genes under heat stress (Figure S1c). Mathematically, a continual rise in N implies an expansion in the total sample space while the subset of effective samples, as a specific gene cluster, such as up-regulated genes under heat stress, remains constant. This expansion ultimately diminishes the probability derived from the hypergeometric distribution, which may approach zero. This aligns with the observations in this work. Intriguingly, the down-regulated genes under heat stress remain at zero up to 10,000 and then exhibit a sudden surge. This phenomenon likely results from the inclusion of weaker promoters into the target range for hits by *motif 1* and *motif 2* with the enlargement of N, intersecting with down-regulated genes as depicted in the green area of Figure S1b.

Beyond the mathematical perspective, the biological implications of of low-affinity TFBSs are also significant. It is widely acknowledged that low-affinity TFBSs play a crucial role in the transcriptional regulation[248,249,250,251,252,253], reflecting the complexity of transcriptional regulation. These TFBSs enable more refined transcriptional control by allowing for discrimination between similar TFs[251]. The lower binding efficiency due to low affinity can be offset by the concentration of TFs[251,250]. Moreover, low-affinity permits a greater diversity of TFs to bind, which significantly enhances the flexibility of gene expression regulation[248]. Additionally, the binding duration between low-affinity TFBSs and TFs tends to be shorter, suggesting a more energy-efficient, transient interaction when compared to the prolonged engagement with high-affinity TFBSs[250].

The preceding discussion has elucidated the rational existence of low-affinity TFBSs in a bio-

logical sense, yet it does not account for the observation that as the parameter N increases beyond 10,000, there is a notable increase in the number of motif pairs binding to the down-regulated genes under heat stress in *A. thaliana*, concomitant with a decrease in the number of motif pairs binding to the up-regulated genes. Motif pairs associated with down-regulation related are commonly involve TFs such as*ATHB12*, *AHL12*, *AHL20*, *AHL25*, *SPL7*, *ANNC46* (primarily for *Triticum aestivum*), and *MYC4*. Except for *SPL7*, these TFs are known to be part of the immune regulatory network of *A. thaliana*[224]. The gene *SPL7* encodes a heat stress factor (HSF) protein that confers tolerance to heat and other environmental stresses in plants[254]. Other members of the *SPL* family, particularly *SPL1* and *SPL2*, have been shown to enhance thermotolerance in *A. thaliana*[255]. Additionally, motif pairs containing *SPL1* were observed in over 10% of the down-regulated genes under heat stress in *Triticum aestivum*. The information obtained is consistent with the hypothesis that low-affinity TFBSs interacting with heat-shock or immunity-related TFs are involved in the transcriptional regulation of plants under heat stress. However, the current data are insufficient to draw more definitive conclusions. Further research, including the additional data acquisition and more in-depth analysis is necessary in the future.

Although establishing a single optimal value of N that is applicable to all species is impractical, for analyses focused on *A. thaliana*, it is recommended to set N to approximately 5,000. This recommendation is based on two observations: first, PMET results across various gene clusters showed minimal fluctuations in the number of motif pairs when N was varied from 500 to 10,000 (Figure S1a-d); second, as discussde in the following chapter, single-cell analyses of *A. thaliana* revealed that genes showing significant expression variability typically number around 2,000. Based on this empirical evidence, it is suggested that an N value of 5,000 to balance the trade-off between capturing a comprehensive set of interactions and maintaining statistical significance.

2.11.6 Genomic Localization Patterns of Motif Pairs on Promoters

Yu et al. (2006)[256] leveraged 586 experimentally verified motifs for 400 TFs, sourced from comprehensive datasets such as TRANSFAC[257], JASPAR 2014[258], Athamap[259], CIS-BP[178], and the work of Franco-Zorrilla et al.[157] to examine the positional distribution of TFBSs in the promoters of *A.thaliana*. These motifs were scrutinized across the promoter regions of all genes using the FIMO tool, with the promoter range defined as -200 to +2,000 bp upstream of the TSS. In their study, the probability density curve (PDC) for the binding of random motifs within a range from 2000 bp upstream to around 50bp upstream of the TSS essentially forms a horizontal line, indicating a random pattern of motif binding. In contrast, a small peak is observed 50bp downstream of the TSS, suggesting a preference for motif binding in that region. For the collected motifs, the PDC starts at zero at the -2000bp position and gradually rises as it approaches the TSS, culminating in a significant peak. This peak reaches its apex at -100bp upstream of the TSS, indicating a marked increase in the probability of motif binding at that location. It is notable that once past the TSS, the PDC drops sharply to near zero (Figure S3a).

Based on the aforementioned findings and observations highlighting the concentration of motif

pairs in proximity to the TSS of various gene clusters (Figure 25), it is hypothesized that omitting promoter regions near the TSS may lead PMET analyses to omit the true regulatory mechanisms. This assertion is supported by Lis et al. (2016), designating the 50 bp upstream of the TSS as the core promoter region and demonstrating the widespread distribution of 10 core promoters within this area[260]. Furthermore, Georgakopoulos-Soares et al. (2013) substantiated the concept by demonstrating that for nine TFBSs —*AHR*, *CREB1*, *CTCF*, *GABPA*, *HNF1A*, *REST*, *SP1*, *XBP1*, and *YY1* —increased expression levels are associated with closer proximity to the TSS[261]. Observations from the analyses of different gene clusters revealed that dark blocks, signifying high motif pair enrichment, were found in both extended and compressed promoter lengths. This indicates that motif binding events proximal to the TSS can interact with those at varied locations of a promoter, culminating in the formation of a statistically significant motif pair.

Individual motif binding and motif pair interactions emphasize the essential function of promoter regions in close proximity to the TSS in transcriptional regulation. Although these TSS-adjacent regions are indispensable for gene expression regulation, distal regions may be less critical under specific conditions, highlighting the paramount importance of the TSS-adjacent promoter domain in orchestrating the transcriptional landscape. Concurrently, defining an appropriate promoter length configuration is a critical for identifying valid motif pairs.

2.11.7 Distribution of Motif Pairs on Genomic Elements

Much research has confirmed that TFs bind to genomic elements that are not promoters[232,233,234,235,236,237,238], PMET analysis results have corroborated this observation as well. Typically, the 3'UTR is known for binding RNA binding proteins to rregulate translational function[262,263,264,265], with this binding occurring through motif recognition[266,267]. The motif pairs in the 3'UTR sequence of *A. thaliana* gene clusters suggest that they are involved not only in mRNA translational regulation but also in transcriptional processes. In plants, the average length of 3'UTR exceeds that of the of 5'UTR[268], as also demonstrated in this work. Contrary to the 5'UTR, the longer length of the 3'UTR does not correlate with an increased number of motif pairs. However, down-regulated gene clusters under drought and heat stress show more motif pairs in the 3'UTR region, which suggests a different regulatory mechanisms. The CDS, as a component of an exon with a relatively shorter length, interacts with smaller number of motif pairs than the exon as a whole. This observation may imply that motif pairs within the exon do not exhibit heterogeneity and positional specificity; instead, their distribution appearsmore uniform.

In Section 2.11.6, observations from this work indicate a significant increase in the binding abundance of randomly generated motifs located 50 bp downstream of the TSS of genes in *A. thaliana*. Furthermore, known meaningful motifs show minimal binding activity in the region immediately downstream of the TSS, suggesting that TF binding in this region, especially in the vicinity of the TSS within the 5' UTR, may not be biologically significant (Figure S3). Additionally, when including 5' UTR sequences, Figure 24 shows a decrease in the number of motif pairs. Based on these observations, the following scientific question arises: Do TF binding events on the 5' UTR

have biological significance?

TFs binding to genomic elements beyond the promoter region introduces a complex labyrinth to the analysis. Alternative splicing generates multiple transcript variants from a single gene, each with a distinct transcriptional region or gene model. To streamline the PMET analysis across these diverse genomic landscapes, the initial strategy involves selecting the gene model that best represents the characteristics of the gene when multiple options exist.

"The most representative" gene model can be determined by criteria such as gene model's length or other ranking methodologies. Choosing the longest gene model is the simplest and direct approach. Alternatively, a more sophisticated ranking system, akin to the one employed by The Arabidopsis Information Resource (TAIR), may be utilized, which prioritizes exons and gene models based on empirical evidence from various experiments and computational predictions, assigning a "rank=1 " to the most reliable ones. In instances where no explicit rank is specified, the default selection is the isoform denoted with a suffix of ".1", implying the primary variant as per conventional nomenclature.

The intricacies of genomic annotation pose significant challenges, often rendering simplistic approaches to gene model selection ineffective. Consider, for instance, the gene model AT1G01010.1 of *A. thaliana* (Table S1). Its CDS is an amalgamation of multiple fragments, each contributing to the composite CDS of the gene AT1G01010. This multifaceted structure defies reduction to a single, longest segment, as such a reductionist approach would neglect the regulatory interplay between the constituent fragments.

The ranking system provided by TAIR, while sophisticated, occasionally falls short in situations where multiple gene models of the same gene concurrently bear the designation of rank=1, such as with *AT1G01020.3* and *AT1G01020.5* in *A. thaliana*. This scenario illustrates the limitations of relying solely on a ranking system, as it can result in ambiguities when multiple models share the same level of presumed significance.

The limitations of relying solely on a single indicator or a potentially incomplete ranking system are evident in the aforementioned predicament, as they can lead to ambiguity or lack of uniqueness. To tackle this, a robust selection strategy must be established, that evaluates the evidence supporting each gene model within the context of its biological functions. For instance, the combination of multiple fragments of a gene model could be considered. Such a framework would prioritize determining gene models based on comprehensive understanding of biological relevance, such as the presence of alternative splicing events, the length of the gene model, and experimental validation, ensuring an accurate representation of gene regulation derived from PMET analyis.

3 High-Resolution of Single-Cell RNA Sequencing Analysis of Cell Cycle Genes in *Arabidopsis thaliana*

3.1 Introduction

Plants, as sessile organisms, encounter a multitude of biotic and abiotic stresses, for instance pathogen infections, salinity, temperature fluctuations, and drought conditions. To withstand these challenges, plants have evolved sophisticated adaptive mechanisms. Importantly, the innate immune system in plants, characterized by pattern-recognition receptors (PRRs) and pathogen-associated molecular patterns (PAMPs), orchestrates pattern-triggered immunity (PTI)[269,84]. Furthermore, there exists a profound interconnection between plant immunity and cell cycle kinetics[270]. In *A. thaliana*, for instance, the Cabbage leaf curl virus infection modulates the expression of genes linked to the cell cycle. For example, the upregulation of *CYCD3;1* or *E2FB* enhances mitotic activity and polyploidy, thereby significantly mitigating the viral infection[93]. Another pivotal response mechanism involves protein kinases, which modulate cell cycle progression in response to abiotic stressors. Under salt stress, the transcriptional levels of *CDC2a* and *CYCA2;1* are diminished, resulting in reduced lateral root formation[271]. These facts indicate that plants do not merely endure suboptimal cell cycle progression; rather, they actively modulate it through stress-responsive pathways. Furthermore, the fundamental aspects of plant development and growth are underpinned by the progression of the cell cycle.

The development of plant roots originates from the embryonic radicle, with stem cells in the apical meristematic zone engendering all cell types through mitosis and differentiation processes[272]. Studies in *A. thaliana* roots have shown that cellular proliferation and differentiation are not uniformly distributed across the root structure. Specifically, cellular division predominantly occurs within the root apical meristem (RAM)[273,274,275].

In contrast, the elongation zone (EZ) is traditionally perceived as a region where rapid cellular elongation occurs. However, contemporary perspectives propose a more nuanced understanding of the EZ: (1) Endoreduplication, a process involving DNA replication without subsequent cell division, precedes cellular expansion; (2) the application of 5-ethynyl-2'-deoxyuridine (EdU) in synchronized *BY-2* cells for brief periods has demarcated the transition between RAM and EZ, with EdU signals delineating the initiation of endoreduplication[275]. The segment spanning from 520 µm to approximately 850 µm from the root cap junction (RCJ) is categorized as the EZ in *A. thaliana*, wherein cells are capable of tripling in length in under three hours[274]. Furthermore, studies by Brady et al. (2007) and Li et al. (2016) have used the emergence of the first root hair as an indicator for the maturation zone or developmental zone (DZ)[276,277] . Brady et al. (2007)presented a high-resolution spatiotemporal map of the *A. thaliana* root, revealing the expression patterns and spatial distribution of various cell types and root zones.

scRNA-seq has emerged as a pivotal tool for elucidating cell cycle dynamics, offering unique insights into novel genes and pathways involved in stress responses and immune functions[84]. This

advanced technology enables high-resolution characterization of cellular heterogeneity and regulatory mechanisms at the individual cell level, thereby facilitating a deeper understanding of the complex interactions between cellular processes and environmental stimuli.

Using the root spatiotemporal map helps the organize transcriptional data from scRNA-seq of *A. thaliana* in according to the developmental zones. This approach enables the high-resolution analysis of cell cycle progression, overcoming the complexities introduced by the heterogeneous nature of the cell cycle across distinct developmental zones. This integrated approach not only enhances the granularity of cell cycle studies but also provides a comprehensive framework for understanding the temporal and spatial dynamics of cellular development in plant roots.

3.2 Data Acquisition

3.2.1 Cell Cycle Genes in *Arabidopsis thaliana*

In the early 2000s, multiple studies focused on the cell cycle genes in *A. thaliana* (Figure 32). Vandepoele et al. (2002) were the first to integrate experimental data; they used representative genes from the CDK family as bait to predict candidate genes in chromosome regions of interest. They employed a variety of methods, including alignment with experimental data, gene structure modifications, and multiple alignments with other family members, to ascertain whether the predicted genes were indeed part of the cell cycle gene class. Through these methods, they identified a total of 61 core cell cycle genes[38].

Furthermore, Menges et al. (2003) utilized *A. thaliana* cell suspension cultures to explore the expression patterns of cell cycle-related genes. Through data analysis and screening, they identified a total of 1082 genes that may be regulated by or associated with the cell cycle; these genes potentially play significant roles in cell cycle progression[200], with each gene assigned to a specific cell cycle phase. This work laid the foundation for subsequent research. Subsequently, Wang et al. (2004) identified 49 cell cycle proteins within the *A. thaliana* genome by employing BLAST searches and the protein families (Pfam) region analysis, which included 14 novel types not previously characterized[278]. Continuing their analysis, Beemster et al. (2005) integrated with their data on root tips to identify a set of 131 highly confident proliferation genes, which were expanded to 182 cell cycle genes[279]. Combining the information from the 1082 cell cycle genes; this work separately tabulated the numbers of cell cycle genes in the G1, S, G2, and M phases, resulting in 34, 23, 11, and 117 genes, respectively (Figure 33).

3.2.2 Single-Cell RNA Sequencing Data

In 2022, Shahan, Hsu et al. conducted a comprehensive profiling of over 96,000 *A. thaliana* root cells using the 10X Genomics scRNA-seq platform[280]. This work leveraged the acquired data; it conducted comprehensive analyses employing the Seurat R package[281]. The experimental details pertaining to each sample used in their study are meticulously presented in a comprehensive overview provided in Table 12.

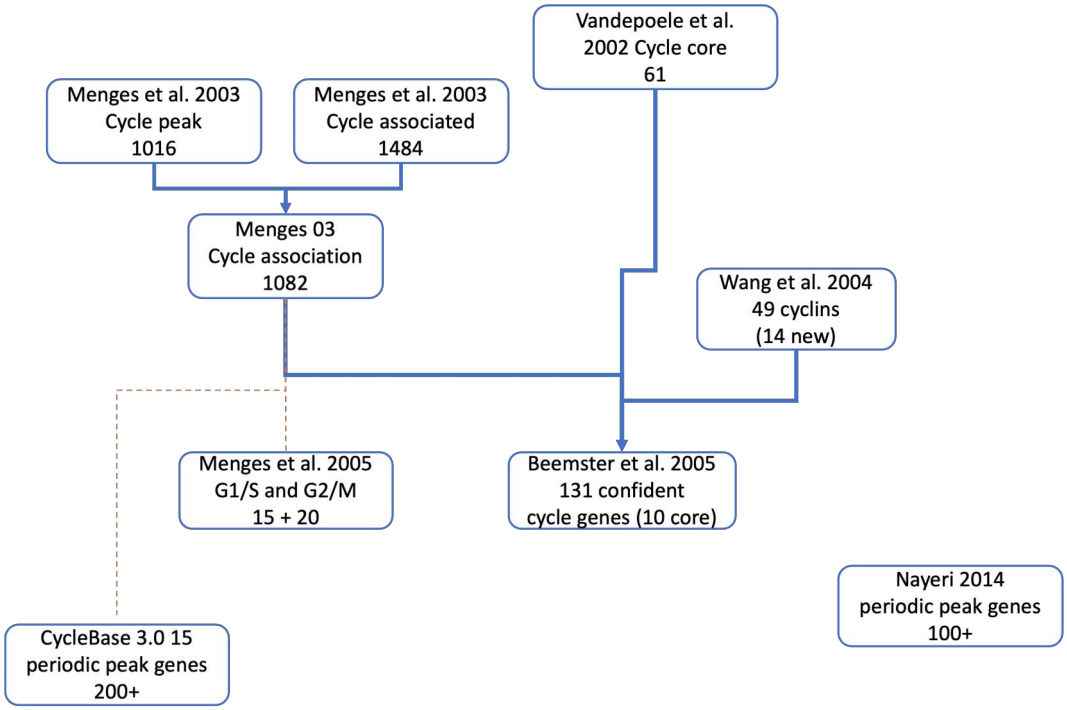


Figure 32: Overview of published sources for the identification of cell cycle genes in *A. thaliana*

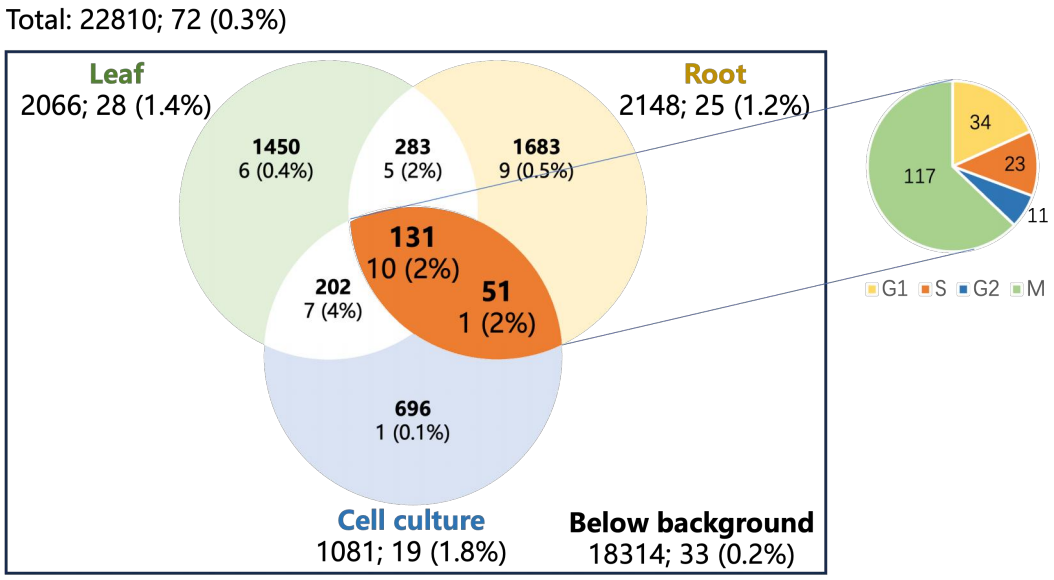


Figure 33: Enrichment of cell cycle genes among those with modulated expression during leaf development, in the root tip, and in synchronized cell cultures of *A. thaliana*. Genes were compared against the core cell cycle genes as defined by Vandepoele et al. (2002)[38] and Wang et al. (2004)[278]. The bold numbers represent the total number of genes in each population. Additional numbers denote the core cell cycle genes and their percentage within each population.

Table 12: Experimental information for each sample used in this work.

sample	Used for Atlas?	source	genotype	age	# of cells
pp1	yes	Ryu et al. 2019	Col-0	5_day	
dc1	yes	Denyer et al. 2019	Col-0	6_day	
dc2	yes	Denyer et al. 2019	Col-0	6_day	
col0	yes	Shahan, Hsu et al. 2022	Col-0	5_day	5,000
tnw1	yes	Shahan, Hsu et al. 2022	Col-0	5_day	5,000
tnw2	yes	Shahan, Hsu et al. 2022	Col-0	5_day	5,000
sc_1	yes	Shahan, Hsu et al. 2022	Col-0	7_day	10,000
sc_9	yes	Shahan, Hsu et al. 2022	Col-0	5_day	10,000
sc_10	yes	Shahan, Hsu et al. 2022	Col-0	5_day	20,000
sc_11	yes	Shahan, Hsu et al. 2022	Col-0	5_day	10,000
sc_12	yes	Shahan, Hsu et al. 2022	Col-0	5_day	10,000
sc_30	yes	Shahan, Hsu et al. 2022	Col-0	5_day	10,000
sc_31	yes	Shahan, Hsu et al. 2022	Col-0	5_day	10,000
sc_37	yes	Shahan, Hsu et al. 2022	Col-0	6_day	10,000
sc_40	yes	Shahan, Hsu et al. 2022	Col-0	6_day	10,000
sc_51	yes	Shahan, Hsu et al. 2022	Col-0	5_day	10,000
sc_20	no	Shahan, Hsu et al. 2022	Col-0	5_day	10,000
sc_21	no	Shahan, Hsu et al. 2022	Col-0	5_day	10,000
sc_25	no	Shahan, Hsu et al. 2022	scr-4	5_day	10,000
sc_36	no	Shahan, Hsu et al. 2022	scr-4	5_day	10,000
sc_52	no	Shahan, Hsu et al. 2022	scr-2	5_day	10,000
sc_53	no	Shahan, Hsu et al. 2022	scr-2	5_day	10,000

3.3 Data Preprocessing of Single-Cell RNA Sequencing Data

3.3.1 Batch Effect Correction

Data from scRNA-seq typically originated from multiple experiments, resulting in batch effects due to different capturing times, handling personnel, experimental conditions, reagent lots, technology platforms, etc. These differences lead to significant variations or batch effects in the data, which may introduce false biological signals and bias of downstream analyses. This work utilized simulation data from CellMixS package to illustrate the batch effects on scRNA-seq data[282]. Batch effects were simulated by randomly selecting 0%, 20% and 50% of gene expression values from a distribution with modified mean values (see Figure 34).

In Table 12, all *A. thaliana* scRNA-seq experiments should have the same expression distribution. To reveal the overall expression patterns of genes in different experimental batches, this work first grouped and aggregated the data by experimental batch. During the aggregation process, the sum of gene expression for all cells within the same batch was calculated. Then, simple correlation analyses was conducted, which aimed to evaluate potential gene expression changes caused by different experimental batches, thus reflecting the impact of batch effects at the gene level. This type of analysis help to understand consistency and differences between experiments and provided a solid foundation for further integration and comparison of data.

The total read counts are obtained by accumulating the read counts of all genes from all cells of the same single-cell sample. The correlation between samples is calculated and visualized in a correlation coefficient heatmap. In Figure 35, the correlations between scRNA-seq data obtained by different research groups appear relatively low. Samples from Denyer et al. (2019), specifically

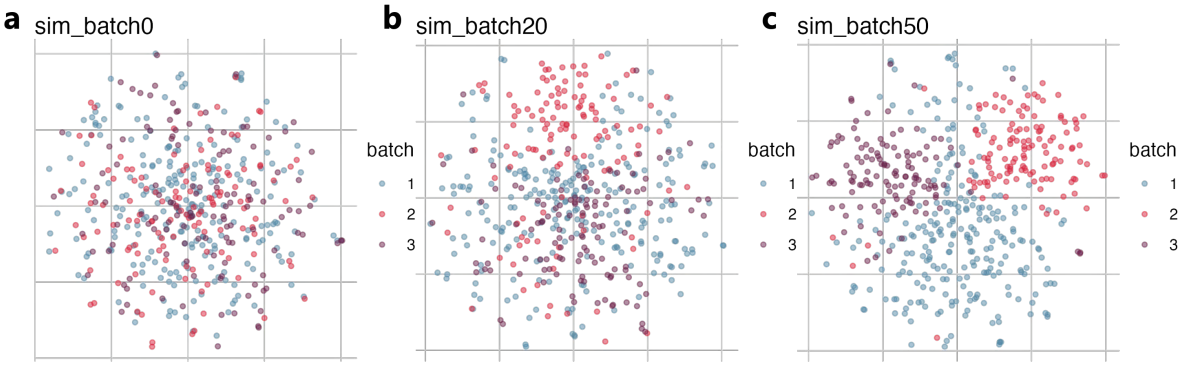


Figure 34: Demonstration of batch effects in simulated data. (a) Unaltered data with a 0% batch effect, (b) Introduction of a 20% batch effect by modifying the mean expression values, (c) A 50% batch effect, which highlights the progressive impact on the distribution of gene expression data in each scenario.

dc1 and dc2 showed significantly lower correlation with the samples from Shahan, Hsu et al. (2022)[283,280], as indicated by the blue predominance in the first two columns and rows of the figure. Samples are from the same research group are expected to exhibit higher correlations, such as those observed between dc1 and dc2, as well as most of the samples prefixed with "sc".

To further evaluate the impact of batch effects, this work utilized the CellMixS package to compute the cell-specific mixing score (CMS) for each cell. This score assesses whether the distance distribution between each cell and its k nearest neighbors originate from the same overall distribution. A high CMS score indicates sufficient mixing between cells and their neighbors, while a low score suggests the presence of batch-specific biases. Numerically flat distributions across different CMS values were observed, suggesting the absence of batch effects (Figure 36a). This result is consistent with the fact that the data was simulated without any batch effect. Significant batch effects were detected in the simulated data with 50% batch effects, as indicated by uniformly low CMS across all samples (Figure 36c). Upon further examination of the 22 samples delineated in Table 12, the levels of batch effects were determined to be approximately 20%, a conclusion drawn from the consistency with the flatness observed in Figure 36b. This observation is consistent with the previously identified low correlation between specific samples such as dc1, dc2, pp1, sc_37, sc_40, sc_25, and sc_36 in the upper left quadrant of the heatmap (Figure 35).

In responses to the observed batch effects due to 7 samples, the exclusion of these samples result in a compact dataset comprising 13 samples. Subsequently, quantification of batch effect using CellMixS was performed. It is evident that the lower scores was relatively reduced (Figure 36e). This suggested that the batch effects among 13 samples was negligible.

3.3.2 Integration of Single-Cell RNA Sequencing Data

Integration of scRNA-seq data is a critical step in scRNA-seq analyses, to counteract unwanted variation due to experimental batches, sample sources or experimental conditions. In general, integration methods will cluster cells of similar types or biological states and capture the underlying

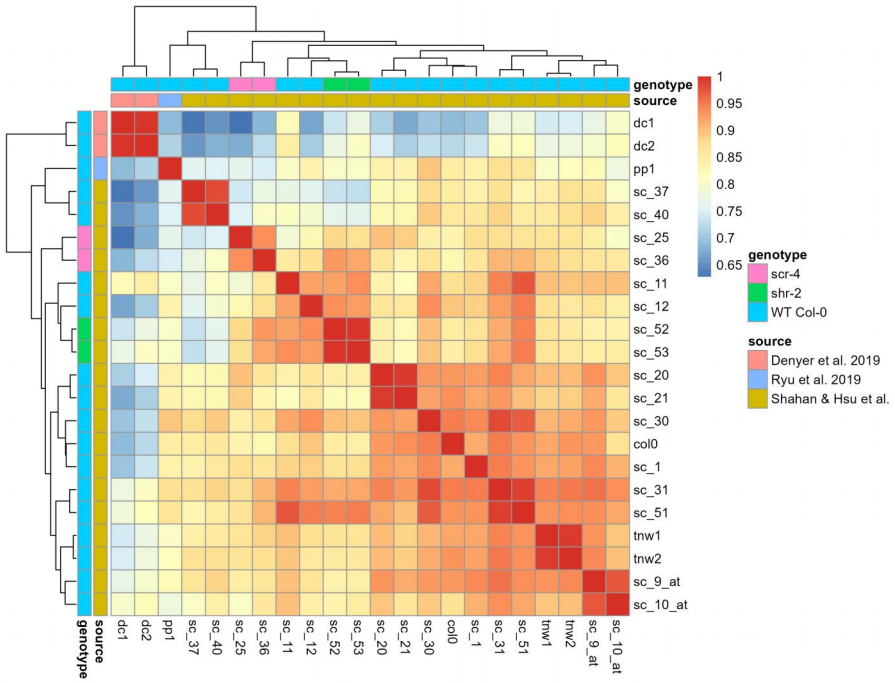


Figure 35: Heatmap illustrating correlations between single-cell samples. Correlation coefficients were calculated using a pseudo-bulk-sample strategy, aggregating read counts from all cells within the same sample. Sample IDs are marked along the x- and y-axes, with sample sources and cell lines annotated using different color schemes. Lower correlations are shown as blue blocks, especially concentrated in the first three rows and columns. This indicates weak correlations between the three samples (del, de2, and pp1) and the samples from Shahan, Hsu et al. (2022), suggesting a batch effect between samples from different laboratories. Correlation coefficients between samples from other laboratories (ppl and dcl) are around 0.7 or lower, while internal correlations within samples from Shahan, Hsu et al. can reach 0.9 or higher[280].

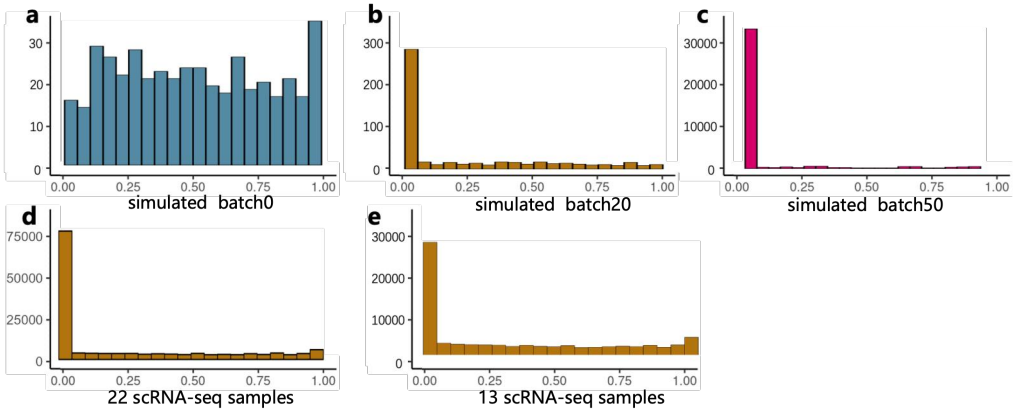


Figure 36: Quantification of batch effects in simulated and *A. thaliana* scRNA-seq data using the CellMixS mixing score. (a-c) Batch effect quantification using data simulating varying degrees of batch effects. Flatness in the bar plot suggests minimal batch effects, while less uniformity indicates more pronounced batch effects. (d) Batch effects in scRNA-seq data from 22 samples. (e) Batch effects in scRNA-seq data from 13 samples, excluding 9 samples with lower correlation. Compared to (d), the height of the first bar in (e) is significantly reduced from 75,000 to 30,000, while the values of the other bars remain relatively unchanged. This indicates a more uniform distribution and reduced batch effects.

biological variances even if they are from different experiments.

Figure 37a illustrates the application of Uniform Manifold Approximation and Projection (UMAP) to visualize human pancreatic islet cell datasets originally sequenced using two different methodologies: CelSeq (GSE81076) and SMART-Seq2 (E-MTAB-5061). This UMAP representation revealed distinct clusters that correspond not to the biological variation within the cell populations but to the technical differences between the sequencing platforms, underscoring the necessity for careful integration.

Upon implementing data integration techniques (Figures 37b,c), the UMAP plots underwent significant transformation. Cell distributions from both CelSeq and SMART-Seq2 platforms became intermingled, demonstrating a successful mitigation of technical biases. The cells were organized primarily by cell type rather than by techniques or platforms used for data generation, allowing for a more accurate interpretation of the biological differences across cell populations. The post-integration UMAP plots revealed a diverse and cohesive cellular landscape, facilitating a deeper understanding of the biological processes at play within the human pancreas. This integrated approach, as illustrated by the UMAP visualizations, underscore the power of computational techniques to overcome technical variance and to reveal the authentic biological tapestry of cellular diversity.

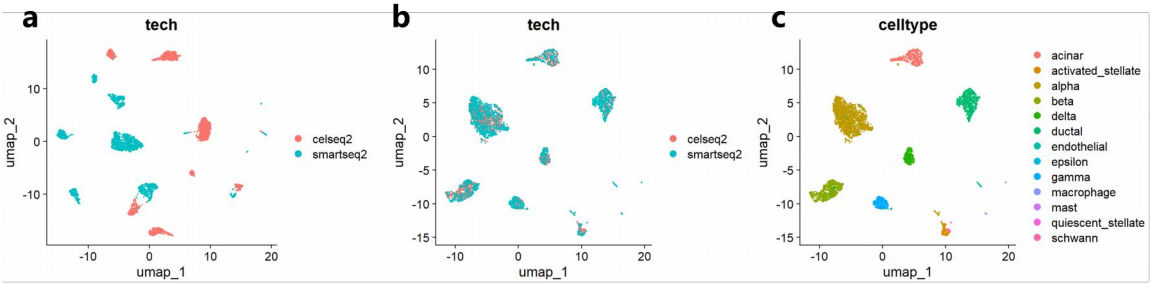


Figure 37: Demonstration of single-cell integration. (a) Two pancreatic islet cell datasets generated by distinct sequencing platforms: CelSeq (accession number GSE81076) and SMART-Seq2 (accession number E-MTAB-5061). (b) Integration of data from these two datasets. (c) Further refinement of the integrated results with cell type annotations. IMAGE: Generated from the Seurat demo.

To effectively demonstrate the process and outcomes of data integration, a focused analysis was performed on a subset of the dataset, comprising six randomly selected samples from those listed in Table 12. These samples were selected to capture the inherent diversity and complexity of the overall dataset. For visual assessment, the Uniform Manifold Approximation and Projection (UMAP) technique was employed to generate a two-dimensional representation of the high-dimensional scRNA-seq data. Within this representation, cells were annotated both by cell type and by sample of origin, providing a dual perspective on the underlying data structure.

Figures 38a-c offer a comparative view of sample dispersion patterns before and after the integration. Remarkably, these plots showed a similar degree of dispersion among the various samples, suggesting that a uniform distribution in the multi-dimensional space was achieved even in the absence of any formal integration technique. This observation might suggest that the underlying

biological variances were more pronounced and consistent than sample-specific differences.

The observation that cell populations from different samples are already well-integrated in UMAP space, even before any data integration processing, suggested that batch effects between these samples were minimal. This implies that the technical variation introduced by different batches is not expected significantly affect the underlying biological signal. As a result, cells from different samples of various cell types were already intermingling and grouping accordingly in the UMAP visualization, indicating that the cell populations were sufficiently blended based on their biological characteristics rather than their batch origins (Figure 38d-f).

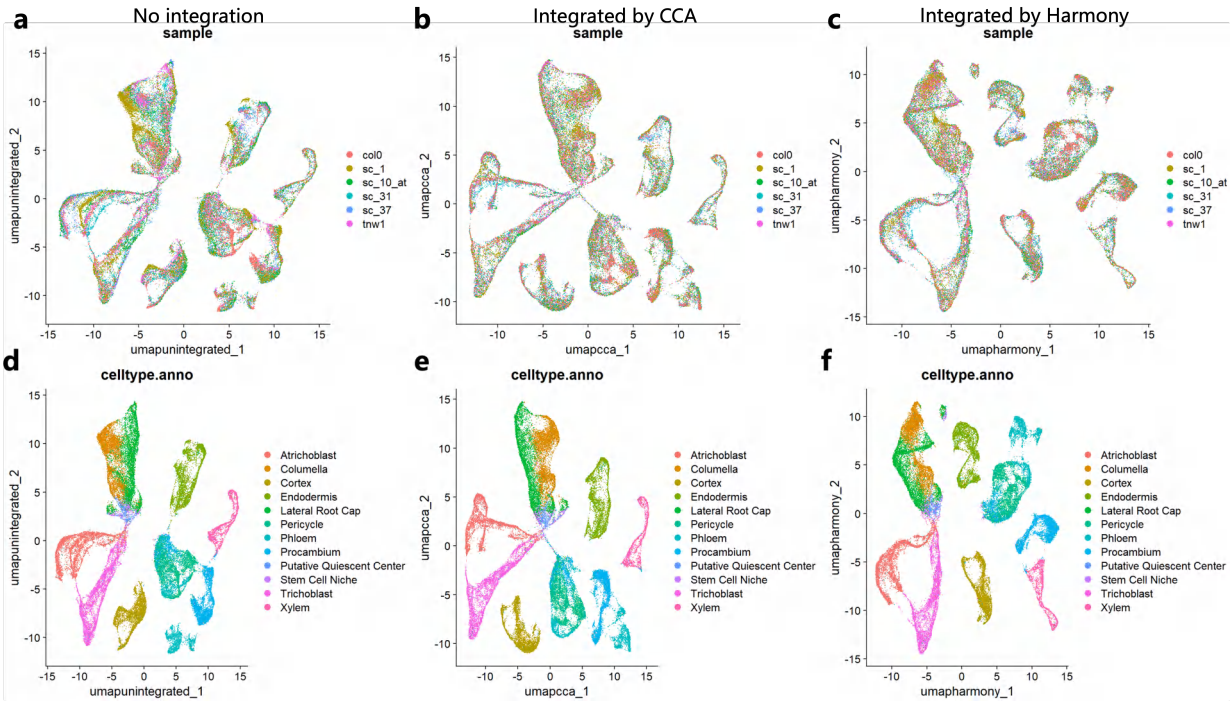


Figure 38: Quality assessment of scRNA-seq data integration for six randomly selected samples, as detailed in Table 12. (a-c) The UMAP visualizations show the distribution of different samples before and after integration, highlighting consistent dispersion across the samples. (d-f) UMAP plots display the samples before and after integration, with cell types annotated, demonstrating the alignment and uniformity of cell clusters.

3.4 Cell Cycle-related Genes in Root Developmental Zones

3.4.1 Segmentation of Root Developmental Zones

In the exploration of *A. thaliana* root development, single-cell transcriptional data can provide a compelling narrative of cellular differentiation. Cells from the RAM, EZ, and DZ were meticulously isolated by Shahan, Hsu et al. (2020), revealing that there are 50,732 cells in the RAM, 38,538 in the EZ and 21,157 in the DZ[284]. This isloation was augmented by comparing single-cell expression profiles with bulk gene expressions data of manually dissected root segments, as detailed in previous studies[276]. Moreover, the application of pseudotemporal trajectory analyses and marker-based annotations has elucidated the progression of cell development, typically initiating from the RAM

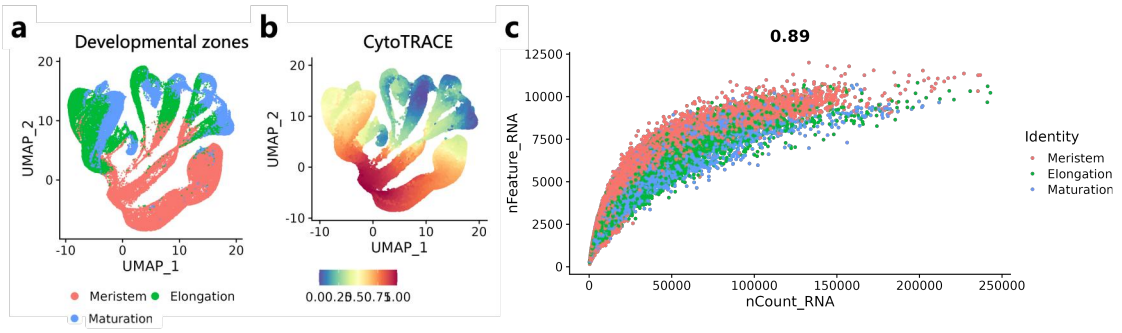


Figure 39: Development-based separation of cells. (a) Dimensionality reduction of scRNA-seq data using UMAP visualizes the distribution of cells across different developmental zones. (b) CytoTRACE provides differential profiles for each cell, displaying the aging pattern in *A. thaliana*. Higher values (red) indicate a lower degree of cellular differentiation, while lower values (blue) indicate a higher degree of differentiation. (c) nFeature_RNA represents the number of genes detected in each cell, and nCount_RNA signifies the total number of molecules detected within a cell. Cells in the root apical meristem (RAM) exhibit higher gene expression, indicated by a greater number of detectable genes compared to cells in the elongation (EZ) and differentiation zones (DZ).

and proceeding outward with EZ transitioning to DZ, as depicted in Figure 39a.

The differentiation pattern of *A. thaliana* roots mirrors the physical extension of cell types from the stem cells to differentiated root tissue. Cellular Trajectory Reconstruction Analysis using gene Counts and Expression (CytoTRACE) was used to construct cellular differentiation trajectories, corroborating the chronological progression[143]. CytoTRACE employs an unsupervised framework. It relies on the number of genes expressed per cell, with a higher number indicating younger cells, to deduce the differentiation states from scRNA-seq data. The differentiation trajectories inferred from scRNA-seq data exhibit an analogous color scheme to the cell development distribution, with warmer hues representing less differentiated cells, similar to those in the RAM as shown in Figure 39b.

This developmental pattern is further quantified by examining the number of genes detected per cell against the total number of transcripts present. As presented in Figure 39c, 'nFeature_RNA' stands for the number of genes detected in a cell, and 'nCount_RNA' represents the total RNA molecules present in a cell. The resulting plot revealed a logarithmic association, with a high correlation coefficient of 0.89, as presented in Figure 39c. The stratification of this scatter plot into three distinct layers—red for RAM, green for EZ, and blue for DZ—illustrates a gradient of gene expression, where RAM cells exhibit a higher gene count compared to those in other zones. Notably, this increase in gene number does not lead to a proportional increase in the number of RNA molecules, indicating that gene count does not confer a transcriptional advantage in terms of RNA abundance.

In summary, scRNA-seq data of *A. thaliana* effectively captured the unique expression profiles inherent to cells at different developmental zones, allowing for an unsupervised analysis of cell differentiation. This detailed view into the cellular evolution within the root provided genetic profiles and developmental patterns, accessible through advanced bioinformatics tools and analyses.

3.4.2 Two Transcriptional Programs Governing Root Development: Mitotic Cycle and Endoreduplication

To further examine the disparities in gene expression attributed to varying degrees of cell differentiation, this work scrutinized the expression patterns of canonical cell cycle genes at three distinct zones of root development and within cells at four cell cycle phases. Despite the inherent differences in expression levels among individual genes, a consistent pattern emerged: genes such as *CYCD3;3* (early G1 phase), *CYCD3;1* (late G1 phase), *CYCA3;1* (S phase), *HISTONE H4* (S phase), *CYCB1;2* (M phase), and *KNOLLE* (M phase) demonstrated very similar expression profiles across different developmental zones. A pronounced expression was observed in the RAM, with decreasing read counts corresponding to the advancement of differentiation, as illustrated in Figure 40a-f. This trend was congruent with the exclusive occurrence of cell division within the RAM, regardless of the cell cycle phase.

The uniqueness and specificity of the endocycle in EZ and DZ are analogous to that of cell division in RAM. The endoploidy level resulting from endocycle increases gradually from RAM to the DZ, with nuclei progressing from 4C (tetraploid), 8C (octaploid), and finally to 16C (hexadecaploid), indicating successive DNA duplications with cell division. Bhosale et al. (2018) identified 4378 developmentally regulated genes whose expression levels are strongly correlated with with cell endoploidy levels. Based on their *R*-values, the six most significant genes associated with the 16C level were selected in this work as *AT3G61410*, *AT4G22070*, *AT2G35730*, *AT5G24140*, *AT3G01970* and *AT1G48000*. Despite variations in gene expression levels, a distinct pattern emerges in the expression of the six 16C genes across the three developmental zones. Notably, from the RAM to the DZ, there is a discernible trend wherein the expression levels of the 16C genes progressively increase (Figure 40e-h). This observed pattern stands in stark contrast to the expression patterns depicted in Figure 40a-d, associated with S or G2 phases.

Further analyses aimed to identify genes that are prominently expressed in the EZ and DZ. From each of the endopolyploidy classes—2C, 4C, 8C, and 16C—83 genes were selected, resulting in a set of 332 marker genes specific to the different levels of endopolyploidy[285]. Subsequently, based on the expression patterns of these four groups of genes in individual cells, distinct C-value labels were assigned to the cells (Figure 41). There is a clear correspondence between the distribution of C-values and the developmental zones as delineated in Figure 40a,b, as well as with the dispersion of CytoTRACE scores. This coherence suggests a tightly regulated progression of endopolyploidy that is closely linked to cellular differentiation.

In essence, the root tip of *A. thaliana* displays diverse expression profiles that reflect its different developmental zones. Consequently, downstream single-cell analyses of the root tip should account for and seek to normalize the biological variation introduced by cell developmental states, ensuring that the data reflects true biological insights over variation related to development.

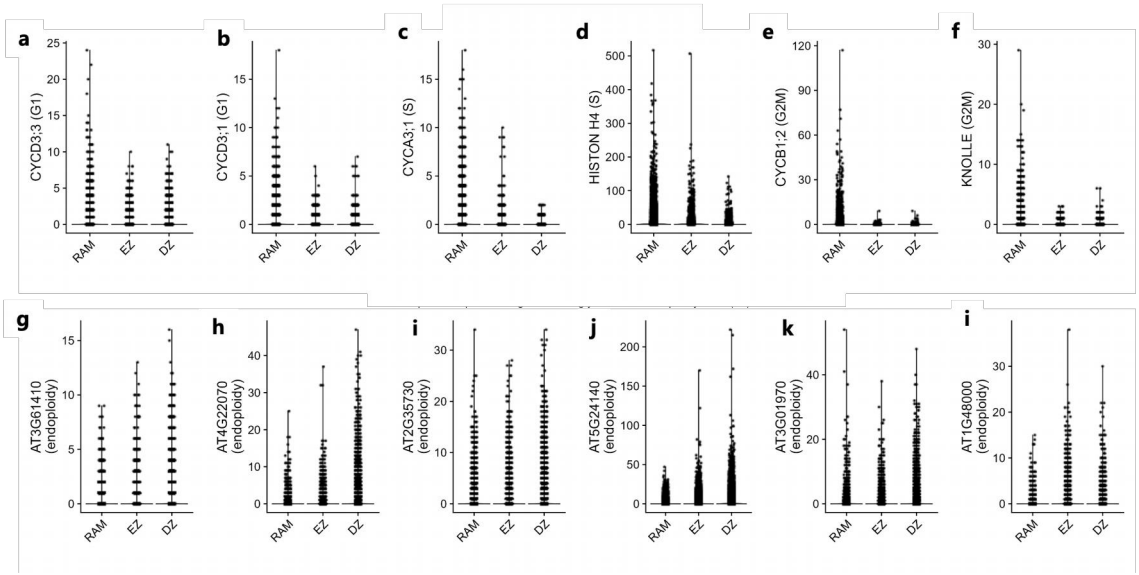


Figure 40: Expression levels of cell cycle genes decrease progressively with the degree of cell development,while the opposite trend is observed for endocycle-related genes. (a-f) Genes involved in the cell cycle, which include two representative genes from the G1 phase, two from the S phase, and two from the G2M phase, are predominantly expressed in the root apical meristem (RAM). Their expression levels diminish as cells develop and differentiate. (g-i) In contrast, the expression levels of six genes associated with endopolyploidy levels show an increasing gradient from the RAM to the differentiation zone (DZ), presenting a pattern that is in direct contrast to that of cell cycle genes.

3.4.3 Functional Ontology Landscapes Across Root Developmental Zones

The scRNA-seq data were categorized into three developmental zones, and a differential expression analysis was conducted to compare the expression levels within each zone to those of the other zones. For a gene to be classified as zone-specific, it must show expressed in at least 50% of the cells corresponding to that zone and exhibit at least a 2-fold change in expression relative to the other zones. This work identified 86 genes enriched in the RAM, 89 genes in the EZ, and 72 genes in the DZ, all with statistical significance ($p < 0.05$). As depicted in Figure 42a, each row represents a gene, and the genes are organized sequentially from RAM, through EZ, to DZ. Each column corresponds to a single cell. The intensity of the color in the heatmap indicates the level of gene expression, with the gradient ranging from purple to yellow, indicating lower and higher expression levels, respectively. The expression patterns of the genes associated with each zone were examined using scRNA-seq data, revealing that RAM-specific genes are predominantly expressed within the RAM region. However, sporadic expression of DZ-specific genes was observed in the RAM and EZ zones, which was more pronounced in the EZ (row 3, column 2 in Figure 42a). Additionally, expression was also detectable of of RAM-specific genes within EZ cells (row 1, column 2 in Figure 42a).

Temporal and spatial expression specificity of zone-specific genes are illustrated in UMAP plots. Branches indicating high intensity in Figures 42b-d represent genes that are respectively enriched in RAM, EZ, and DZ. This pattern correlates strongly with the developmental zones, as depicted

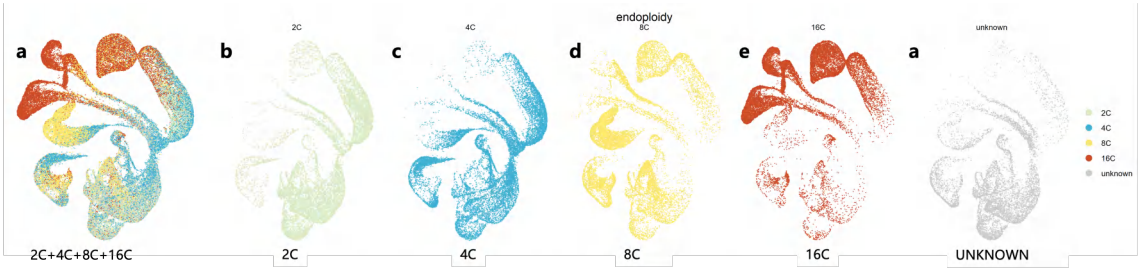


Figure 41: UMAP visualization depicting endopolyploidy across root cells in *A. thaliana*. (a-e) Cells are categorized by their C-value: 2C, 4C, 8C, and 16C, including (a) a collective representation of all classes (2C+4C+8C+16C). These clusters are delineated based on the expression patterns of 332 marker genes selected for their expression in the elongation zone (EZ) and differentiation zone (DZ), with colors indicating specific endopolyploidy levels.

in Figure 39a, demonstrating a clear association between gene expression and the respective root zones.

GO analyses were conducted to elucidate the functional profiles of zone-specific genes identified in the root zones. Notably, *trichoblast differentiation* and *root hair cell differentiation*, along with other similar GO terms, were highlighted, indicating a specialization for root hair formation in the EZ (Figure 43b). In the DZ-specific GO analysis, there was a conspicuous association with terms related to oxygen stress, including *response to hypoxia* and *response to oxidative stress* (Figure 43c). Intriguingly, similar terms pertaining to hypoxia were also noted in the RAM-specific dataset, suggesting possible roles in oxygen sensing or signaling across these developmental zones (Figure 43a). For RAM-specific genes, significant enrichment was observed in GO categories pertinent to ribosomal structure and biogenesis, exemplified by terms such as *ribosome biogenesis*, *ribosomal small subunit assembly*, and *ribonucleoprotein complex biogenesis*. This reflects the high demand for protein synthesis machinery in the actively dividing cells of the RAM (Figure 43a). Within the EZ-specific gene set, GO terms such as *cell development* and *developmental growth* were prominent, reflecting the pivotal role of these genes in cellular expansion and elongation processes.

3.5 Cell Cycle Genes of Root Developmental Zones

3.5.1 Split Classification of Cells by Cell Cycle Phase

In the previous sections, this work delved into the single-cell expression data of *A. thaliana* across different developmental zones, as well as gene expression profiles during four phases of the cell cycle. For instance, within the scRNA-seq dataset of the RAM, comprising 50,732 cells, expression patterns of genes of cell cycle phases were analyzed. This analysis identified cells that exhibited higher expression levels of these gene sets during their respective cell cycle phases compared to cells in other phases. The AUCell method quantifies the enrichment of target gene sets in specific cells by calculating the area under the curve (AUC)[286]. VISION constructs a K-nearest neighbors (KNN) graph to depict cell-to-cell similarities using principal component analysis (PCA) as the latent space and employs a Wilcoxon rank-sum test to identify signaling pathways with statistically significant expression differences in specific cell populations[287]. scMiko builds shared nearest

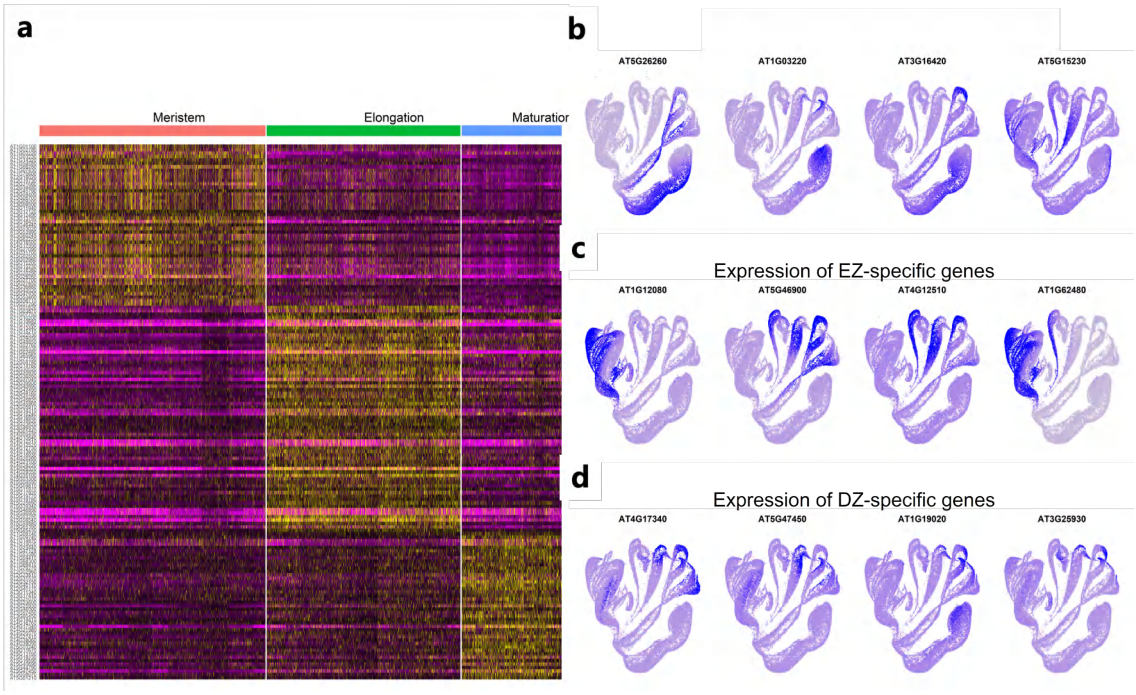


Figure 42: Expression patterns of zone-specific genes in corresponding cells are visualized. The enrichment of selected zone-specific genes is revealed in distinct UMAP atlases. (a) Expression levels of genes specific to the three developmental zones within root cells are enriched along the diagonal of the heatmap. These zone-specific gene lists were derived from differential expression analyses of scRNA-seq data, using varied selection criteria to ensure a comparable number of differentially expressed genes across zones. (b-d) Four representative genes from each of the three developmental zones exhibit high expression levels, and their spatial distribution on UMAP plots aligns closely with the delineated regions within the root developmental profile (as shown in Figure 39).

neighbor (SNN) graphs between cells and their k-th nearest neighbors using the Jaccard index and clusters these graphs using the Louvain algorithm[288]. Single-sample GSEA (ssGSEA) extends gene set enrichment analysis (GSEA) to the single-sample level, calculating enrichment scores for each sample against specific gene sets, thereby determining the expression trends within those gene sets[289].

The four methods were employed to assess the enrichment of cell cycle gene expression within cells across three distinct developmental zones, retaining only the top 1,500 cells that demonstrated the highest enrichment scores for each respective cell cycle gene set. Illustrated with data from the RAM, Figure 44 presents an UpSet plot showing the intersecting distribution of the top 1,500 cells identified across the four methods. In the G1 phase (Figure 44a), 650 intersecting cells were identified; in the S phase, G2 phase, and M phase, there are 697, 347, and 1,144 intersecting cells, respectively (Figure 44b-d). Reasons for the high number of intersecting cells in the M phase could include the larger quantity of M phase genes, in contrast to the G2 phase, which has fewer than one-tenth the number of genes (11 to 117).

The process of selecting cells with the highest enrichment in specific gene sets has the inherent potential for a single cell to be enriched in multiple gene sets, which introduces ambiguity into the delineation of cycles and complicates subsequent differential analysis. This issue is inherent to

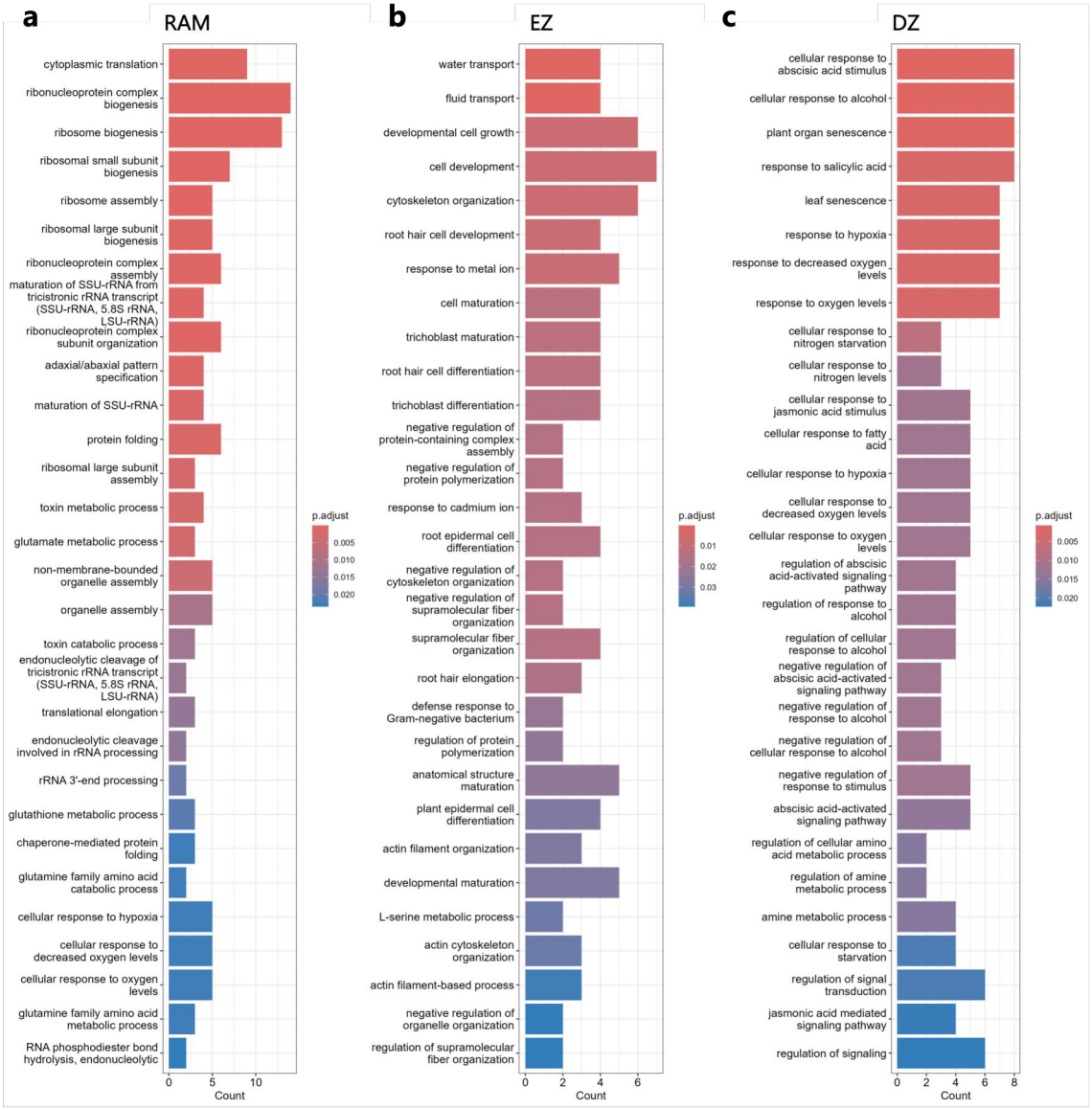


Figure 43: GO analyses indicate zone-specific genes enriched in distinct biological functions. (a) Genes exclusive to the root apical meristem (RAM) were significantly enriched in categories related to ribosome structure and biogenesis, including terms such as *ribosome biogenesis*, *ribosomal small subunit assembly*, and *ribonucleoprotein complex biogenesis*. (b) Genes exclusive to the elongation zone (EZ) were prominently associated with GO terms for *cell development* and *developmental growth*. (c) Genes specific to the differentiation zone (DZ) revealed enrichment in terms related to abscisic acid and oxygen stress.

all four methods. Addressing this, each method was applied with a focus on identifying the top 1,500 cells by enrichment scores for specific cycle phases, excluding cells enriched in multiple gene sets. Ensuring exclusivity for each cell cycle, the cell counts identified by all methods are no longer fixed to 1,500. Accordingly, the number of cells identified by all four methods has significantly decreased, which is shown in green bars of Figures 45a-d. The most dramatic change is observed in the G1 phase, with the count dropping from 636 cells (orange bar in Figure 44a) to 78 cells (orange bar in Figure 45a).

Selecting a specific number of cells from those identified by these methods is necessary to facil-

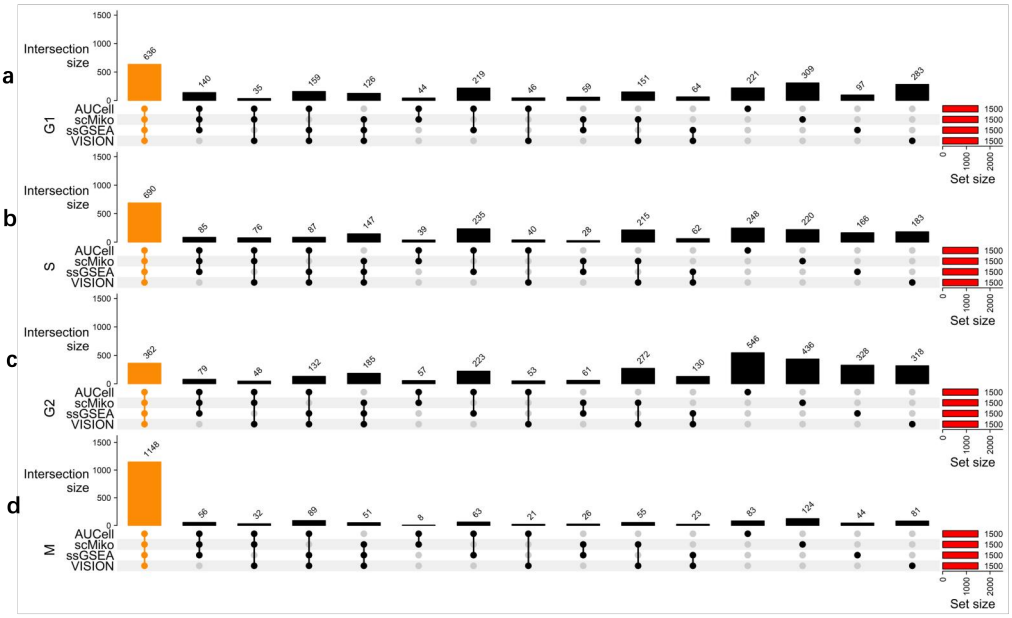


Figure 44: UpSet plot depicting the intersection of cell identification among four analytical methods (AUCell, scMiko, ssGSEA, VISION) in the root apical meristem (RAM). The top 1,500 cells with the most enriched cell cycle gene expression for each phase were selected. Red bars represent the unique cell counts identified by each method, while orange bars show cells identified by all methods. (a) G1 phase with 650 intersecting cells, (b) S phase with 697 cells, (c) G2 phase with 347 cells, and (d) M phase with 1,144 cells, highlighting the methodological consensus and variability in identifying cell phases.

itate single-cell level differential analyses of cell cycles. The ideal scenario involves choosing four equally numbered subsets of cells from those commonly identified by all four methods. Figure 45 shows that the distribution of cells commonly identified is uneven when phase exclusivity is considered. The selection criteria have been adjusted to include cells identified as enriched by at least two methods, matching those represented by the orange and purple bars. To ensure a balanced number of cells across different cycles, a threshold of 500 cells per cell cycle subset has been set. Similar selection procedures were applied to the EZ and DZ to assemble comparable subsets of cells. However, the M phase was excluded from the selection in both zones due to the absence of mitotic division, maintaining contextual relevance and accuracy in light of the unique cellular activities characteristic of the EZ and DZ.

Refined criteria were applied, prioritizing exclusivity in cell cycle phases and balanced quantities, facilitating the curation of cell subsets from three distinct developmental zones. The distributions for each cell cycle phase, subsequent to these criteria, are detailed in Table 13.

Table 13: Number of cells per cell cycle phase and root developmental zone.

Cell	G1	S	G2	M
RAM	392	383	355	497
EZ	277	401	414	/
DZ	364	484	481	/

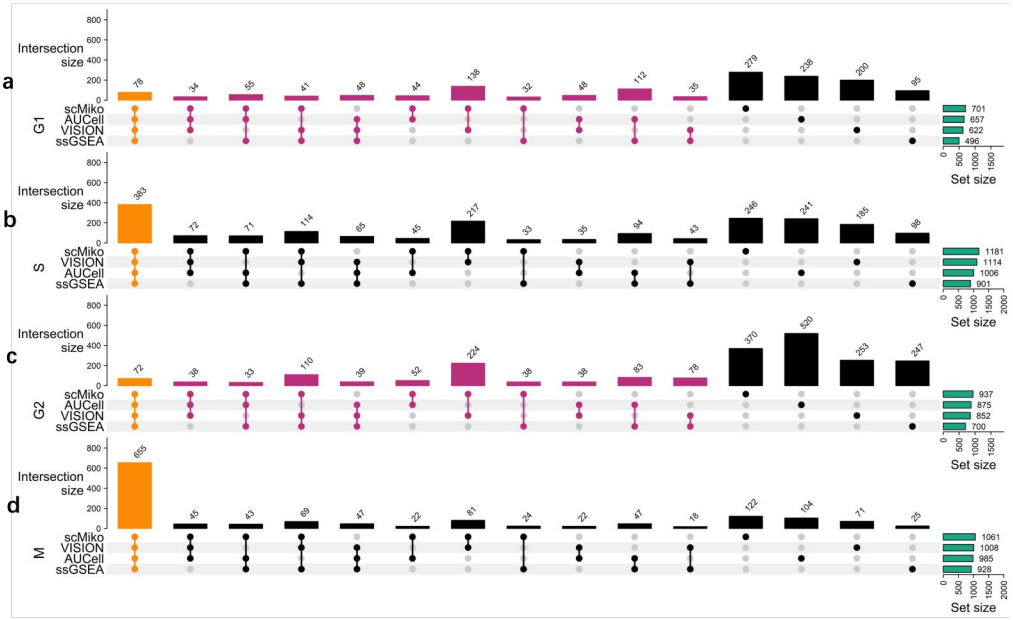


Figure 45: UpSet plot illustrating the unique and intersecting identifications of cell populations by four analytical methods (AUCell, scMiko, ssGSEA, VISION) across the cell cycle phases G1 (a), S (b), G2 (c), and M (d). The enforcement of cell cycle exclusivity in the analysis reduced the number of cells commonly identified across all methods. The plots reveal how stringent criteria alter the intersection sizes, particularly evident in the G1 phase, where the common identification count drops from 650 to 86 cells when exclusivity is imposed. This approach mitigates the ambiguity of cycle delineation and supports more precise differential analyses. Green bars indicate the total number of cells identified by all methods post-exclusivity enforcement, which differ from the initial fixed count of 1,500, reflecting a significant decrease in cells commonly identified by all methods.

3.5.2 Genes Involved in the Cell Cycle

Afterward, the identification of cell cycles was established across three developmental zones. Cell cycle-specific expression matrices were subset from the comprehensive scRNA-seq dataset, applying normalization and scaling to enhance data comparability (Figure 46). Subsequent differential expression analyses identified genes significantly enriched in each cell cycle phase. Seurat's FindMarkers function identified differentially expressed marker genes between two groups of cells.

To ensure that the identified marker genes of a certain cycle do not exhibit differential expression across other cell cycle phases, the FindMarkers function compares the target cycle cells against all other cells. This comparison evaluates gene expression levels between the two groups using various statistical methods. The output typically includes a list of differentially expressed genes, and the clustering of individual marker genes, the average log fold change of marker genes (avg_log2FC), the percentage of cells within the target group expressing individual marker genes ($pct.1$, the percentage of cells within the comparison group expressing the same marker genes ($pct.2$), along with statistical measures including mean expression difference, p-value., and adjusted p-value.

- The average logarithmic fold change (avg_log2FC) of the gene during the target cell cycle must be greater than 1 to ensure a significant change in its expression level.
- The proportion of gene expression in cells of the corresponding cycle ($pct.1$) must exceed

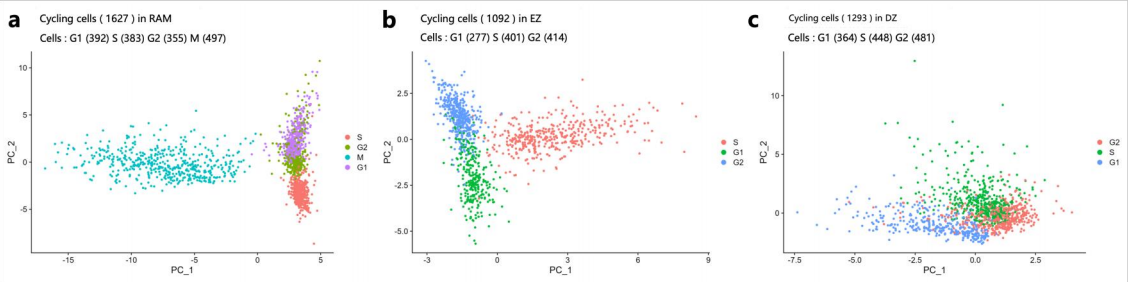


Figure 46: Principal component analysis (PCA) of cycling cells in individual developmental zones. Following the isolation of cell cycle-specific expression matrices from the comprehensive scRNA-seq dataset, the data underwent normalization and scaling to ensure comparability. PCA plots illustrate the resulting distributions: in the root apical meristem (RAM) (a) with 1,627 cells, in the elongation zone (EZ) (b) with 1,092 cells, and in the differential zone (DZ) (c) with 1,293 cells. Each dot represents one cell. The plots demonstrate clear separation of data after normalization and scaling, revealing distinct clustering patterns by cell cycle phase and phase-specific gene expression profiles.

40%, that is, $pct.1 > 0.4$, to ensure cell cycle specificity.

- Genes were sorted in descending order according to avg_log2FC, and in ascending order according to $pct.2$, and in descending order according to $pct.1$.

- Finally, the top 30 genes were selected as positive marker genes (positive markers) of the target cell cycle.

For negative markers, the same filtering conditions were used, but avg_log2FC was set to be less than -1 as the criterion to ensure that the expression levels of these genes were significantly reduced in target cell cycles. Selected positive and negative markers are stored in Table S6, S7, S8, and S9.

Heatmap plots were used to examine marker gene expression across cells assigned to different cell cycle phases in the RAM. Figure 47a shows that cells predominantly express positive marker genes during their respective cell cycle phase, with some G2 phase genes also expressed in S phase cells (as seen in row 3, column 2). In contrast, negative marker genes typically show minimal expression during their associated cell cycle phase, as indicated by the consistently low expression levels from the upper left to the lower right in Figure 47b.

This work compared newly identified marker genes with reference genes (32 for G1, 23 for S, 11 for G2, and 117 for M). Additionally, it included an analysis of 1966 root-associated genes that were not assigned to any cell cycle phase. The results were depicted in Figure 33. For instance, of the 30 positive marker genes selected for the G1 phase presented in Table S6, 6 overlap with the reference genes, as shown by the red and blue areas in Figure 48a. Furthermore, it was noted that the newly identified marker genes with lower constraints ($avg_log2FC > 0.8$) show greater enrichment in the root overall compared to their respective target cell cycle phases, as evidenced by the overlap of blue and green in Figure 48b.

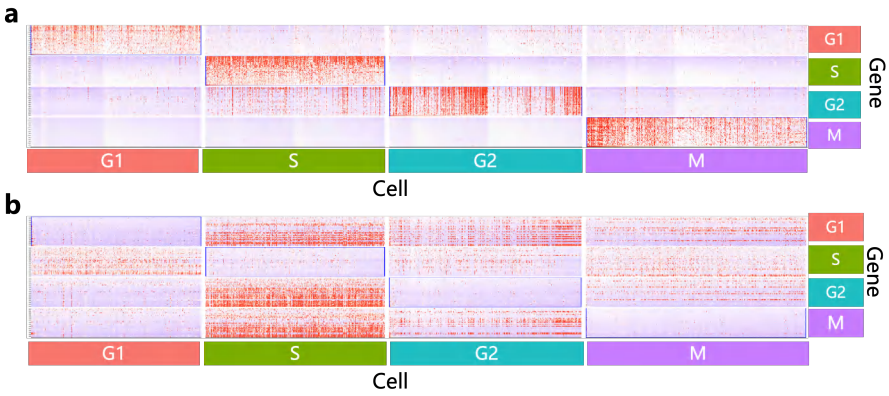


Figure 47: Expression profiles of up-regulated and down-regulated marker genes across different cell cycle phases in the root apical meristem (RAM). (a) Expression of up-regulated marker genes within cells undergoing cell cycle phases G1 (red), S (green), G2 (blue), and M (purple). The expression of these markers is primarily restricted to their respective phases, with a slight expression of G2 genes in S phase cells. (b) Expression of down-regulated marker genes, which are minimally expressed in their target cell cycle phases, illustrated by continuous areas of low expression along the diagonal of the heatmap from upper left to lower right, indicating a distinct separation of genes. This expression pattern underscores the cell cycle phase-specific expression of these genes.

3.5.3 Conserved Cell Cycle Characteristics across Developmental Zones

Within the RAM and EZ, expression profiles of cycling cells were characterized using 30 identified positive markers in the RAM's G1, S, and G2 phases. Observations revealed that in both zones, contiguous blocks of high expression aligned along the diagonals of the heatmaps, with the left representing the RAM and the right the EZ, as shown in Figure 49. This pattern suggests that genes from the same cycle phase but different developmental zones may share overlapping regulatory functions or co-expressed gene sets.

To determine if the observed gene expression patterns are non-random, genes differentially expressed within the same phase but across different zones were compared. The selection of definitive cell cycle genes followed strict criteria, including: initially, only genes with a p-value less than 0.01 were considered to guarantee statistical significance; secondly, genes had to be present in at least 70% of the target cells to affirm their biological relevance; and lastly, genes were ranked by the fold change in expression, with further sorting of expression levels in the target ($pct.1$) and non-target (($pct.2$) cells in descending and ascending orders, respectively. These ranking procedures aimed to evaluate the expression differences and specificity of genes within the cell cycle comprehensively. Table 14 enumerates the cell cycle genes across three developmental zones.

Table 14: Number of selected cells cycle genes per cell cycle phase and developmental zones.

Cycle genes	G1	S	G2	M
RAM	1130	757	357	579
EZ	278	492	273	/
DZ	412	325	336	/

Genes associated with the G1, S, and G2 phases of the cell cycle were identified by differentially

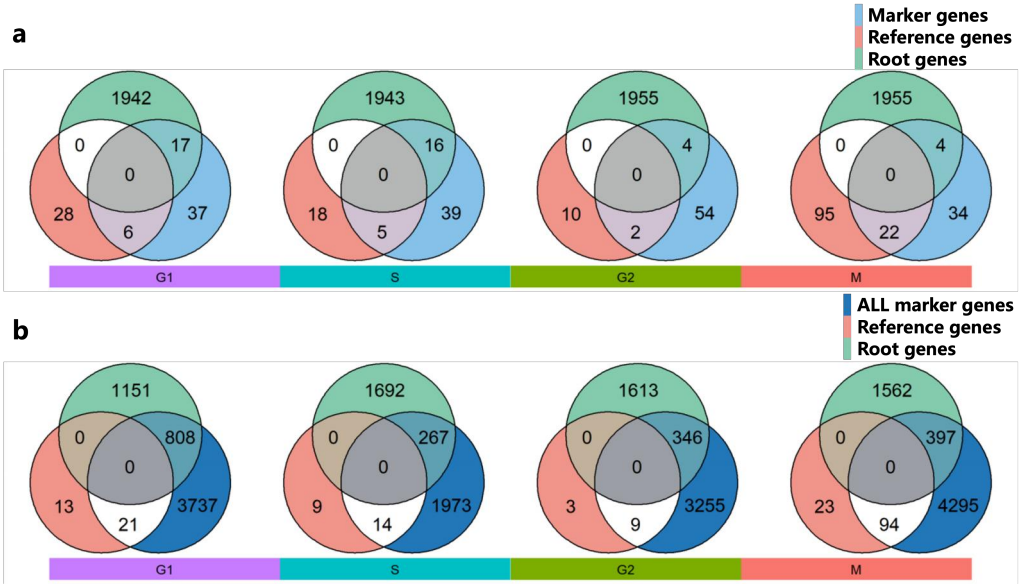


Figure 48: Relationship between newly identified marker genes and published reference genes for each cell cycle phase (G1, S, G2, and M) as well as genes expressed in roots but not assigned to any specific cell cycle phase. (a) Intersections of up-regulated markers for each phase. (b) Pronounced enrichment of newly identified marker genes within the general root transcriptome from Beemster et al. (2005)[279] compared to their presence within designated cell cycle phases.

analyzing cycling cells within the RAM, EZ, and DZ. These genes were then compared between the developmental zones. Figures 50a-d illustrate the overlap in gene expression among the G1, S, and G2 phases across the analyzed developmental zones.

3.6 Applying PMET to Cell Cycle Genes

This work has detailed the process of identifying specific motif pairs within gene sets using PMET in the last chapter. It was employed to probe the enrichment of motif pairs within transcriptional regulatory networks of cell cycle gene sets. The gene sets employed in the PMET analyses were the cell cycle genes specific to the RAM. Table 14 describes how to obtain these gene sets. In an effort to ensure comparability across the PMET analyses of the four gene sets representing distinct cell cycle phases, the quantity of genes in each set was standardized to 357. This standardization was predicated on the gene set for the G2 phase, which contained the smallest number of genes, as shown in Table 14.

Figure 51a provides a heatmap showing motif pairs enriched across four sets of newly found cell cycle genes, each containing 357 genes, within the RAM. These enriched pairs have been filtered based on their p-values and exclusivity to particular cell cycle phases, and are differentiated by a color-coded scheme. The axis of the heatmap displays motifs from the top 30 combinations of motif pairs, with each cell cycle gene set represented, organized alphabetically. The aggregation of colored blocks within the heatmap offers insights into the enrichment patterns, revealing potential regulatory synergies between motifs that could be pivotal in governing cell cycle progression. This visualization underscores the nuanced interplay of motif pairs and highlights PMET's utility in

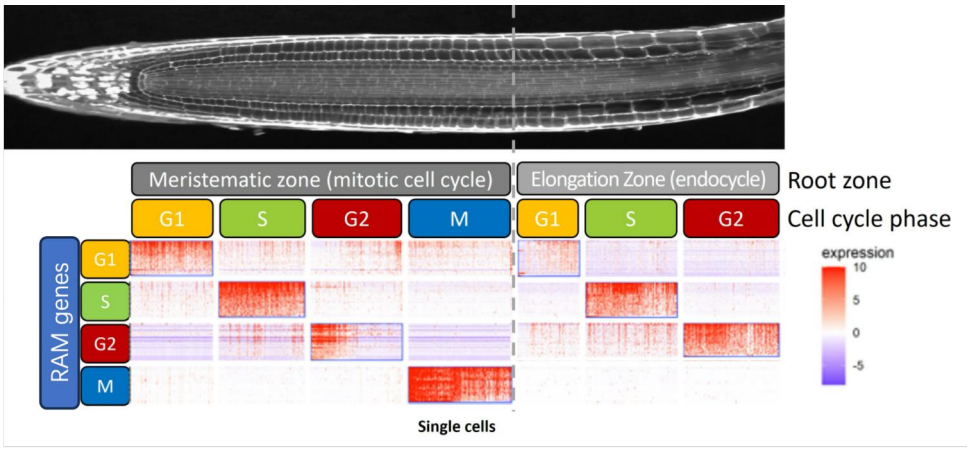


Figure 49: Expression patterns of cycling cells across G1, S, and G2 cycle phases within the root apical meristem (RAM) and elongation zone (EZ), using gene markers for G1, S, and G2 phases derived from differential analyses of cycling cells in the RAM. In both the RAM and EZ, the diagonal exhibits elevated levels of gene expression. These two heatmaps suggest the general regulatory functions of cell cycle genes in different zones.

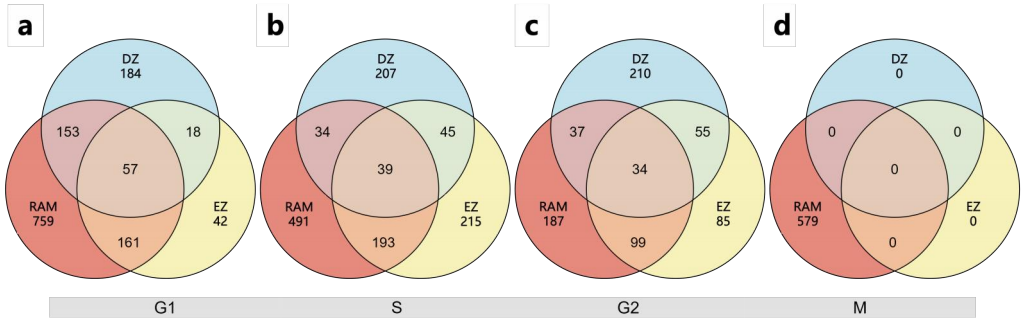


Figure 50: Intersections of selected cell cycle genes in developmental zones of *A. thaliana*. The criteria for selecting cell cycle genes in each developmental zone are detailed in Table 14. (a-c) G1, S, and G2 phase cell cycle genes obtained from three different developmental regions (red represents the root apical meristem (RAM), yellow the elongation zone (EZ), and blue the differentiation zone (DZ)), with their similarities and differences depicted through Venn diagrams. (d) Mitosis occurs exclusively in RAM and EZ; hence, in the representation of M phase genes, the quantities indicated by the circles for DZ are zero.

uncovering complex regulatory relationships.

Figure 51b-e presents the enrichment of motif pairs across distinct sets of cell cycle genes, each analyzed independently. With the assistance of Dr. Ruth Schäfer, this work achieved detailed annotation for a subset of clustered motif pairs.

To ensure the cycle-specificity of motif pairs, the output from the PMET was meticulously filtered, ensuring that each motif pair exists only in one gene set. Subsequently, the top 100 most significant motif pairs were selected as candidates from each cycle phase (52a). An analysis was then conducted to summarize the occurrence of TF family members binding to site motifs within each candidate set. This comprehensive summary is depicted in Figure 52b. Apparently, the TF family showed enrichment across different cell cycle phases. Taking the G1 cycle as an example, among the 100 motif pairs, 87 motifs are derived from the *MYB* (MYB-histone) family.

In the top 100 lists for each cell cycle phase, for each TF binding motif, the respective

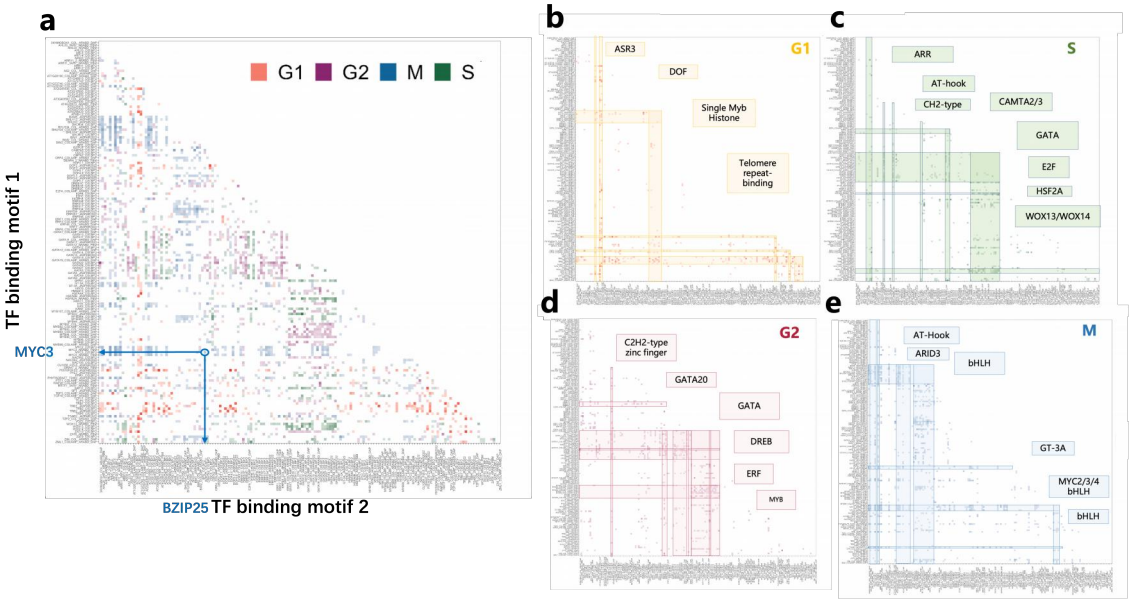
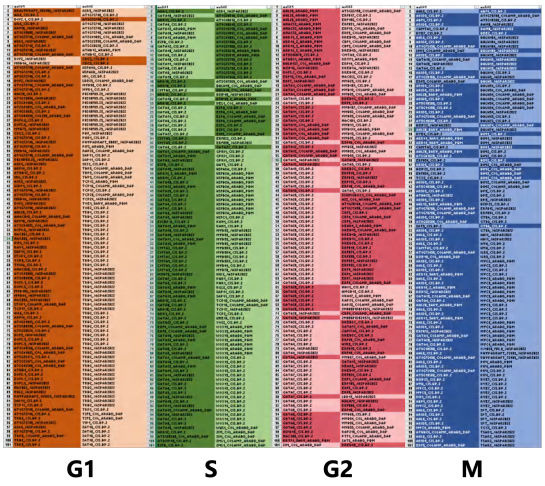


Figure 51: Distribution of TF binding motifs in promoters of cell cycle phase-specific genes (G1, S, G2, M) in the RAM highlights the transcriptional regulatory landscape of gene expression. (a) Distribution of motif pairs in four cell cycle gene sets with a color-coded scheme: red for G1, green for S, purple for G2, and blue for M. Each pixel in the heatmap represents the co-occurrence of one motif pair. This visual arrangement enables the comparison and identification of identical and unique motif pairs, as well as individual motifs associated with each gene set corresponding to different cell cycle phases. For example, BZIP25 (x-axis) and MYC3 (y-axis) pair up in promoters of genes expressed in the M phase. MYC3 only appears in blue cells associated with the M phase. (b-e) Distribution of motif pairs associated with each cell cycle gene set to identify specific motif pairs and motifs. For example, within the gene set related to the G1 phase, motifs associated with the ASR3 family co-occur with multiple other motifs.

TF was identified and attributed to a TF family[290,291,292]. In some cases, further distinctions were made: *AHL*s (AT-hook motif nuclear-localized) are members of the *AT-hook* TF family[293]. The APETALA2/ethylene responsive element binding protein (*AP2/EREBP*) family was subdivided into the *AP2*, dehydration-responsive element-binding (*DREB*), and ethylene response factor (*EREBP-ERF*) subfamilies[294]. Homeobox (*HB*)-containing TFs are subdivided into those containing a leucine zipper (*LZ*) domain (*HB-LZ* TFs) and those without a *LZ* domain (*HB* (no *LZ*) TFs). WUSCHEL-RELATED HOMEOBOX (*WOX*) proteins 13 and 14 belong to the latter group[295]. *MYB* (myoblastosis) domain-containing TFs share *MYB* DNA-binding domains with the Avian myeloblastosis virus protein (*v-MYB*) and *c-MYB* from vertebrates. Plant *MYB* TFs can be divided into four classes: 1R-, 2R-, 3R-, and 4R- *MYB* TFs, based on the number of *MYB* domain repeats[296]. This work distinguished between *MYB-R2R3* proteins, which contain an R2 and an R3 *MYB* domain, *MYB*-Histone proteins, which contain a single N-terminal *MYB* domain and a central globular histone domain, and other *MYB*-related TFs[296,297]. It should be noted that B-type *A. thaliana* response regulators (*ARR*s) and other *GAPR* TFs also contain a *MYB*-related DNA-binding motif[290,298].

In light of the insights gleaned from the prominent TF families labeled in the PMET heatmaps (Figure 51) and top 100 lists of motif pairs (Figure 52), certain TFs were earmarked for in-depth

a



b

	G1	S	G2	M
AHL	2	0	3	18
AP2/EREBP (AP2)	6	0	0	0
AP2/EREBP (DREB)	0	2	34	8
AP2/EREBP (EREBP-ERF)	1	2	32	5
ARF	7	0	0	0
ARID	0	1	0	25
BHLH/MYC	3	2	2	82
bZIP	0	3	1	13
C2C2 (Zn)-DOF	21	0	1	0
C2C2 (Zn)-GATA	0	75	85	8
C2C2 (Zn)-YABBY	2	0	0	0
C2H2 (Zn)	14	6	2	3
C3H-type (Zn)	5	0	0	1
CAMTA	0	10	2	3
E2F/DP	0	10	0	0
EIL	0	1	0	0
GARP	0	14	5	5
GRAS	0	1	0	0
HB	4	1	1	2
HB (no LZ) (WOX13/14)	0	24	0	0
HMG-box	1	0	2	0
HSF	2	12	1	0
MYB (MYB-Histone)	87	0	0	0
MYB (MYB-related)	13	18	7	16
MYB (R2R3)	6	3	12	0
NAC	6	0	6	0
TCP	9	4	1	0
Trihelix	3	0	1	4
OTHER	8	11	2	7

Figure 52: Top 100 TF binding motif pairs in promoters of genes of each cell cycle phase, alongside the occurrence of TF family members within the top 100 motif pairs of each phase. (a) The 100 most significant motif pairs identified by PMET, with filter settings ensuring each motif pair appears only once per gene set. (b) Occurrence of TF family members in the top 100 motif pairs of each gene set.

functional investigations. The decision was based on the exclusive or predominant presence of the respective binding motif within promoters of phase-specifically regulated genes. Table 15 provides a comprehensive enumeration of the selected TFs, identified using arabidopsis genome initiative (AGI) code, alongside their expression profiles in the RAM based on the single-cell seq expression data. For example, *ASR3*, *SMH*, and *TRB* were listed in Figure 51b, owing to their recurrent presence in the top 100 G1-specific motif pairs. This work conducted further analyses on these TFs, in order to identify cell cycle genes that they co-regulate. For this purpose, distinct color schemes were employed to highlight the gene set with the most genes involved. Apparently, cycle-specific TFs invariably exhibit more interactions with genes pertinent to their respective cell cycle phases.

3.7 Discussion

Throughout the lifecycle of plants, the cell cycle plays a critical role in driving growth and development. In the case of *A. thaliana*, extensive research has yielded numerous lists of genes related to the cell cycle, such as 61 core cell cycle genes[38], 49 cyclin genes (with 14 newly identified)[278], 1,082 cycle-associated genes[200], and 2,148 genes from root tips[279]. However, these genes were identified either through cell culture methods, which may not fully replicate the authentic biochemical environment of *A. thaliana* root tips, or they focus on cyclins and CDKs. There are significantly fewer G2 genes compared to other phases. In this work, an integrative approach was adopted, combining existing cell cycle gene sets with scRNA-seq data enriched with root developmental information, to construct a comprehensive compendium of cell cycle genes. This investigation revealed that,

across different developmental zones, cell cycle genes exhibit shared and distinct characteristics, providing new insights into the regulation and function of the cell cycle in plant development.

3.7.1 Minimal Batch Effects in Single-Cell RNA Sequencing Data from Standardized Experimental Settings

In order to accurately present biological information without being affected by differences in time, technology platforms, processing personnel, or reagent batches, scRNA-seq data from different sources usually undergo batch effect correction before downstream analyses. Many studies focus on developing new algorithms or tools to remove batch effects, or benchmarking multiple tools. One frequently overlooked question is whether batch effect correction is necessary. For data from different laboratories, it is obvious that batch effect correction is necessary before further investigation. However, what about samples from a single cell line or from the same laboratory?

This work commenced with an analysis of correlations among single-cell samples, using a pseudo-bulk sample strategy. This initial investigation revealed a distinct pattern: samples from diverse sources exhibited lower correlation coefficients than those from similar sources, whereas samples from identical sources demonstrated a high degree of correlation with one another.

Lütge et al. (2021) introduced CellMixS[282], a bioinformatic tool specifically designed to quantify the batch effect among different batches. The application of CellMixS to datasets by Shahan, Hsu et al. (2020)[284] revealed that the level of the batch effect was not as significant as initially anticipated. This is consistent with what has been revealed by the correlations between samples. In the data integration section, the lack of stark separation by batch in the UMAP plots before integration (Figure 38a) also implies that the batch effect between samples was minimal, and the cells are already well-mixed.

Based on these findings, this work posits that an experimental batch should be defined as a collection of a set of multiple samples subjected to identical experimental conditions, rather than considering each individual sample as an independent batch. Combined with the consideration of the potential overcorrection and loss of biological signal, this work utilized 13 single-cell samples from Shahan, Hsu et al. (2020)[284] and applied batch effect correction to these samples for downstream analyses.

3.7.2 Developmental Zone-Invariant Cell Cycle Signatures

Leveraging the scRNA-seq data provided by Shahan, Hsu et al. (2020)[284], this work sought to expand the catalog of cell cycle genes specific to *A. thaliana* roots. This was achieved by integrating cell cycle-related genes from the cell culture of *A. thaliana* cells[200] and *A. thaliana* root-specific genes[279], followed by a differential expression analysis at the single-cell level[281]. The stratified scRNA-seq data enabled the distinction of cell cycle genes across three zones of root development. Subsequent analyses within this study uncovered notable enrichment patterns of an identical set of cell cycle genes across different developmental zones. For instance, genes associated with the G1 phase, identified through RAM single-cell differential analysis, exhibited pronounced expression in

G1 phase cells within both the EZ and DZ (Figure 49). This observation highlights the conservation of cell cycle genes during root development. Moreover, the disparities in the expression of genes across different developmental zones within the same cell cycle may also mirror their distinct roles in growth and development. Variations in the expression or suppression patterns of G2 phase genes in the EZ and DZ compared with RAM may prevent cells from entering the M phase. The morphological and functional differences between cells in the EZ and DZ, such as root hair formation and cell differentiation[299], may be related to differences in gene expression from the G1 to the G2 phase.

3.7.3 Cell Cycle-Phased Transcription Factor Cooperation Revealed by Paired Motif Analysis in *Arabidopsis thaliana*

This work used PMET to find paired motifs within a set of periodic genes, and then identified the regulatory processes of TFs. In particular, after the motif pairs were filtered and visualized in a heatmap, information about enriched TF families was obtained.

Previous methods for studying TF regulation of biological processes included co-expression analysis[300,301]. The A. thaliana co-expression tool (ACT)[302] revealed that the expression of *OBP1* from the *DOF* family positively correlated with 48 genes involved in DNA/RNA metabolism and cell division, with the experiment verifying that the *OBP1* gene reached its peak in the G1 phase[303]. With the help of PMET, this work found that G1 phase genes were frequently found in combinations of *DOF* family TFs with other TFs. The involved *DOF* family members include *DOF2.1* (CISBP2), *DOF2* (JASPAR 2022), *DOF3* (JASPAR 2022), *DOF4.1* (CISBP2), *DOF4.6* (CISBP2), *DOF5.3* (JASPAR 2022), *DOF5.6* (JASPAR 2022), and *DOF5.7* (CISBP2).

Although other members of the *DOF* family, appearing in the PMET results above along with *OBP1*, belong to the same TF family, indicating they have similar DNA-binding domains, their regulatory networks and functions may differ. This involves different expression patterns, target genes, and regulatory mechanisms. The specific functions, expression patterns, and target genes of TFs need to be determined through experimental studies. Clearly, this is a massive project that is too arduous to complete, highlighting the value of PMET.

It has been confirmed that S phase genes are regulated by the *E2F* TF family. Overexpression of *E2F1*, *E2F2*, and *E2F3* causes quiescent immortalized rodent fibroblasts to enter the S phase, and *E2F3a* is a transcriptional activator that mainly exists in the S phase[304]. Given the conserved behavior of *E2F* TFs as regulators of the G1-S phase transition in animals and plants[305,306,307], combined with the results of PMET analysis (Figure 51), it is suggested that some *E2F* members have S phase-specific properties.

In summary, PMET, as a bioinformatics method, focuses on searching for paired TF motif binding within gene sets, exploring the interactions and synergistic effects of TFs. This helps to uncover potential transcriptional regulatory networks and offers a theoretical foundation for subsequent experimental validation.

CC phase	TF ID	AGI	expressed in RAM	# of G1 genes	# of S genes	# of G2 genes	# of M genes
G1	ASR3	AT2G33550	weak	171	93	139	78
G1	SMH	AT1G72740	yes	236	107	147	61
NA	SMH	AT1G17520	yes	NA	NA	NA	NA
G1	TRB1	AT4G39530	yes	224	75	130	20
G1	TRB2	AT5G67580	yes	217	114	128	98
NA	TRB3	AT3G49850	yes	NA	NA	NA	NA
S	ARR10	AT3G16857	yes	71	100	70	88
S	ARR11	AT1G67710	wea	83	139	131	109
S	ARR12	AT2G25180	yes	62	104	94	18
S	CAMTA2	AT5G64220	yes	21	114	109	106
S	CAMTA3	AT2G22300	yes	56	105	100	100
S	E2FA	AT2G36010	yes	85	91	83	0
S	E2FB	AT5G22220	yes	97	108	102	0
S	E2FC	AT1G47870	yes	26	66	0	0
S	GATA10	AT1G08000	yes	75	132	123	116
S	HSFB2A	AT5G62020	yes	52	94	97	59
NA	HSFB2B	AT4G11660	yes	NA	NA	NA	NA
S	KAN1	AT5G16560	yes	58	97	25	20
S	WOX13	AT4G35550	yes	69	103	56	0
S	WOX14	AT1G20700	wea	58	78	67	0
G2	ERF48(DREB2C)	AT2G40340	yes	50	49	109	65
G2	GATA11	AT1G08010	yes	67	136	131	114
G2	GATA12	AT5G25830	yes	55	141	137	123
G2	GATA20	AT2G18380	wead	71	144	136	123
G2	MYB93	AT1G34670	no	48	65	102	92
M	AHL20	AT4G14465	yes	39	74	91	93
M	AHL6	AT5G62260	yes	47	0	62	84
M	ARID3	AT1G20910	yes	32	40	50	83
M	AT1G75490	AT1G75490	weak	67	94	94	99
M	AT3G10580	AT3G10580	no	20	19	25	75
M	BAM8	AT5G45300	yes	0	54	38	94
M	BHLH13	AT1G01260	yes	26	31	26	110
M	BHLH34	AT3G23210	yes	80	79	96	143
M	BHLH104	AT4G14410	yes	23	0	73	109
M	BIM1	AT5G08130	yes	53	48	74	109
M	BIM2	AT1G69010	yes	78	85	118	149
M	GT3A	AT5G01380	yes	0	0	0	0
M	MYC2	AT1G32640	no	57	65	54	125
M	MYC3	AT5G46760	yes	75	71	39	131
M	MYC4	AT4G17880	yes	28	29	33	114
M	PIF4	AT2G43010	weak	89	74	94	134
M	RVE7	AT1G18330	yes	65	65	55	89

Table 15: Selected TFs from the most abundant TF families highlighted in the PMET heatmaps and top 100 lists of motif pairs. These TFs were confirmed to be expressed in the RAM and show a significant correlation with their respective cell cycle genes. The correlation is visually accentuated through the use of distinctive color schemes: orange for G1 phase, green for S phase, red for G2 phase, and blue for M phase.

Discussion and Future Work

4 Conclusions

Published scRNA-seq data and cell cycle genes in *A. thaliana* were integrated to study the comprehensive profile of cell cycle genes in different developmental zones of the root. The batch effect was characterized by qualitative correlations and quantitative analyses using CellMixS, enhancing data quality control. Meanwhile, PMET was employed to investigate the transcriptional regulation network of cell cycle genes.

The main findings of the thesis are as follows:

1. Batch effects among samples from the same laboratory are relatively minor, a consistent finding observed with both the correlation method and the CellMixS approach[282]. Even when genotypic differences exist between samples, the impact of these differences on the correlation coefficient is significantly less than the variations observed between samples from different laboratories. Therefore, focusing on 13 samples from the same laboratory with the *WT Col-0* genotype will allow for a more controlled analysis, minimizing batch effects and providing clearer insights into the genetic and molecular mechanisms.

2. Genes associated with the cell cycle exhibit substantial quantitative differences in expression levels, with read counts potentially varying. However, within the three developmental zones of the root (RAM, EZ, and DZ), these genes show a consistent trend of gradual decline in expression from RAM to EZ to DZ. In contrast, the expression of genes highly related to the endocycle increases with cell age. This observation, validated at the single-cell level, suggests that as cells mature, their level of endocycle activity rises while cell division activity diminishes.

3. An updated examination of cell cycle genes reveals both similarities and differences in gene composition across the three developmental zones of the root.

4. Using PMET to search for motif pairs among cell cycle genes, it was found that members of specific TF families, along with other TFs, are present in the promoter regions. Most of these TF family members exhibit periodic activity, regulating gene expression only during specific phases of the cell cycle.

4.1 From Suspension Cultures to Single-Cell Sequencing: Cell Cycle Gene Expression in *Arabidopsis thaliana*

Derived from the cell suspension culture of *A. thaliana* ecotype Landsberg erecta, the MM1 and MM2d cell lines, despite differences in morphology and growth characteristics, both exhibit distinct partial synchronization in the cell cycle. By removing and then re-adding a carbon source, both MM1 and MM2d cell lines can achieve partial synchronization in the G1/S phase. The further

developed MM2d cell line, for the first time, the separation of cells in the S and M phases using sucrose starvation. Additionally, a more precise synchronization of the S-G2-M phases was achieved using the Aphidicolin block/release method. After adding Aphidicolin, approximately 80% of the cells were arrested in the S phase, and within the subsequent few hours, 92% of the cells progressed to the G2 phase[198]. In subsequent studies, synchronized *A. thaliana* MM2d suspension-cultured cells, achieved through treatment with Aphidicolin and sucrose starvation, were used for transcriptome analysis on the Affymetrix ATH1 microarray. This analysis identified a 1082 genes, including cell ccle regulated genes and cell cycle associated genes. The former's expression patterns exhibit periodic changes corresponding to cell cycle phases, indicating a key component in cell cycle regulation. The latter show significant changes in expression levels at specific stages of the cell cycle but lack typical periodic patterns, suggesting potential roles in the cell cycle process. Their expression may be influenced by experimental conditions such as synchronization treatment or be related to the loss of cell-to-cell interactions. These early studies have laid the foundation for the study of the cell cycle in *A. thaliana*.

However, the gene list derived from cell culture has limitations regarding its applicability to whole-plant systems. The root apex, a primary growth point, is the core organ for cell cycle activity, with an internal microenvironment and cellular characteristics that are difficult to fully mimic in cell culture. To overcome these limitations, this study integrated the gene set from cell culture with the *A. thaliana* root tip gene expression atlas by Birnbaum et al. (2003)[308], identifying 182 key cell cycle regulatory genes in the context of the root tip. This integrated analysis helps to reveal the tissue specificity of plants and characteristics of the cell cycle, laying the foundation for further exploration of a more comprehensive and detailed gene set for the plant cell cycle. In addition, the intersection of these two gene lists with the cell cycle genes during leaf development[279] included 10 out of the 80 core genes defined by Vandepoele et al. (2002)[38] and Wang et al. (2004)[278]. This finding not only validates the conservation of cell cycle regulatory mechanisms but also indicates that comparative analysis across different datasets can enhance the confidence level of biological discoveries.

Applying the 182 root-specific and cell cycle regulatory genes, meticulous analyses were conducted on single-cell root samples of *A. thaliana*. Samples were categorized by developmental zones, enabling the identification of cells at different cell cycle phases within each zones. Additionally, differential expression analysis identified gens associated with specific cycle phases across developmental zones. The resulting list of cell cycle-related genes not only includes genes related to CDK and cyclins but also encompasses a broader set, contributing to a deeper understanding of cell proliferation, development, immune function, and tolerance mechanisms[309]. Section 3.5.3 discusses the similarities and differences among the periodic genes in developmental zones. The similarities highlight the conservation of cell cycle regulatory mechanisms[310,311], while the differences may relate to the spatiotemporal specificity of cell cycle regulation[312] and biological differences between cell types[313,314].

4.2 PMET

TF cooperative binding is a prevalent mechanism of transcriptional regulation mechanism, manifesting in various forms, including protein-protein interactions, DNA-mediated cooperative binding without direct protein contact, DNA cooperation that enhances protein affinity, and nucleosome-mediated cooperation[179,315,196]. PMET is designed to identify TF binding site pairs in the promoter regions of a gene set. In this work, PMET has been enhanced and expanded from its original foundation, with web deployment implemented through R-Shiny, and an in-depth investigation into the impact of PMET key parameters on motif pairs identification across gene sets.

PMET's core components consist of Bash scripts, Python scripts, and compiled C/C++ binaries, offer flexibility for computational tasks in any Debian-based Linux environment and enable computational resource allocation based on specific needs. Equipped with comprehensive tutorials, this an open-source project allows users to modify source code for specific particular research aims. Additionally, users can optimize the use of computational resources by selectively running specific parts of the program. PMET-R-shiny provides an interactive web interface (https://www.pmet.online/), allowing users to conveniently access PMET's computational and visualization features through a web page, enhancing user experience and lowering the barrier to entry.

During the PMET indexing, a motif represented by a PWM model matches a gene's promoter and retains the k most significant hits. The geometric mean of the p-values for these k hits serves as an estimate for the probability of random motif matching in the promoter region, with authenticity of these hits assessed using statistical distributions. This work has compared PMET results based on binomial and Poisson distributions and obtained nearly identical outcomes. When the event probability is very low, the binomial distribution approximates to the Poisson distribution. Therefore, it can be concluded that motif hits during the PMET indexing process are significant, and the filtering operation through statistical distributions further ensures that the results of PMET are statistically meaningful.

The initial design of PMET aimed to streamline the user experience by minimizing necessary inputs or parameters. Despite this simplicity, PMET remains sensitive to input parameters, often yielding outcomes that are linear to parameter adjustments. This highlights the need for careful consideration of PMET parameters to ensure results with biological meanings.

The search region length is a key parameter in PMET analysis, crucial for identifying motif pairs. However, increasing this length does not necessarily identify more motif pairs due to biological constraints on the actual lengths of promoter regions that do not extend with the increased parameter. Extending the search region excessively may cause sequence overlap with other genes, potentially introducing statistically significant but biologically irrelevant motif bindings and potentially reducing the number of identified motif pairs. Alternatively, overlapping regions may contain statistically significant and biologically meaningful motif bindings unrelated to the transcriptional regulation of the original gene cluster but pertain to the regulatory elements of another gene. Investigating the distribution of promoter region lengths within the genome of a specific

species and setting an appropriate search region length based on this information is essential. This step is critical for ensuring the accuracy and biological relevance of the analysis results.

Yu et al.(2006) found that in *A. thaliana*, hotspot matching regions for individual motifs are predominantly concentrated within a range of 500 to 50 bp upstream of the TSS[256]. However, PMET analysis reveals that motif pairs involve motif binding events both near and at varying distances from the TSS. Significantly, excluding the 100bp sequence immediately upstream of the TSS from the analysis greatly reduces PMET's ability to detect statistically significant motif pairs. This phenomenon is consistent across gene clusters regardless of the set of the promoter length. These observations suggest that the motif pairs in the promoter region are such that one motif is located close to the TSS, while the other motif may be situated elsewhere within the promoter often have one motif close to the TSS and the other elsewhere, including regions adjacent to and further from the TSS. Therefore, setting an appropriate promoter length is crucial for maximizing the detection of all potential motif pairs and ensuring the accuracy and biological relevance of PMET analysis.

PMET can identify statistical significant motif pairs on non-promoter genomic elements like 5'UTR and 3'UTR, highlighting the complexity of gene transcriptional regulation. Yu et al.'s study (2006)[256] also observed a significant number of non-functional random motif bindings in the region immediately downstream of the TSS, namely the 5'UTR (Figure S3a). Figure 24 shows that when the promoter sequence includes the 5' UTR, the number of motif pairs identified by PMET analysis is reduced compared to when the sequence contains only the promoter region without the 5' UTR. However, is it sufficient to conclude that TF binding events on the 5' UTR lack biological significance based solely on these phenomena? The potential role of the 5' UTR region in gene expression regulation requires further research to deeply understand the biological functions of the 5' UTR region and the specific roles of TF binding events.

4.3 Occurrence Patterns of Motif Pairs on Genes under Stress

In Sections 2.7.1 and 2.10, this work investigates how varying PMET parameters affect the number of identified motif pairs in the gene set induced by heat stress, revealing several meaningful phenomena. Specifically, when examining the parameter N (the number of genes or promoters a motif can match), it was observed across three species that increasing N allows more genes to participate in PMET indexing, the number of motif pairs identified in the set of up-regulated genes under heat stress initially increases and then decreases. Notably, this pattern is also observed in the set of down-regulated genes under heat stress, only when the value of N is very large. Additionally, a similar pattern is observed in induce gene sets under drought stress of other species, suggesting that this phenomenon is not unique to *A. thaliana* and may be a common feature across various species and stress conditions. This pattern indicates that in the response to heat stress, there may be a synergistic action of low-affinity motifs involved in transcriptional regulation.

GO analysis of heat stress-up-regulated genes shows that they are associated with stress response processes, and identified motif pairs correspond to the binding sites of heat shock factors

(HSFs), validating the functionality of PMET in studying transcriptional regulation. In contrast, GO terms for heat stress-down-regulated genes are predominantly involved in metabolic and synthetic biological processes. Among the motif pairs of these down-regulated genes, several are associated with immune-related TFs, such as *ATHB12*, *AHL12*, *AHL20*, *AHL25*, *ANNC46*, and *MYC4*. Currently, it is not possible to logically connect the relationship between heat stress, down-regulated genes, and immune-related TFs into a coherent narrative. This suggests that future research needs to delve deeper into the mechanisms of transcriptional regulation to understand the roles of these factors in the heat stress response.

5 Outlook

By integrating scRNA-seq data with the known list of A. thaliana cell cycle genes for cell cycle clustering and differential expression analysis at the single-cell level, this work successfully expanded the atlas of cell cycle genes, particularly those associated with the G2 phase. The further refined PMET and its expanded functionalities were employed to gain deeper insights into the transcriptional regulation mechanisms of cell cycle-related genes, thereby validating the newly identified candidates.

However, PMET's performance remains computationally constrained, primarily due to high memory demands. Future developments should focus on optimizing PMET' s search for heterotypic motif pairs (the pairing procedure) by avoiding the simultaneous loading of all homotypic motif files into memory. Additionally, as highlighted in the discussion, optimizing PMET' s search region length would be beneficial. Ideally, PMET should analyze the entire promoter region—from upstream of the TSS to the adjacent gene boundary—rather than relying on a fixed threshold. Refining the pairing procedure for heterotypic motif pairs would not only reduce computational overhead but also enhance the detection of cooperative TF binding events, a critical aspect of understanding combinatorial gene regulation.

Beyond software optimization, PMET could be expanded in the future to support searches for multiple TF binding events, such as ternary motif combinations. This enhancement would better align with actual biological processes, enabling a more realistic simulation of regulatory networks. Compared to single or paired motifs, multiple TF binding events exhibit greater specificity and selectivity.

To further validate the newly identified genes *in silico*, future work will include *in vivo* experiments. The COLORFUL-Circuit platform will be employed to monitor cell cycle progression in real-time using fluorescent markers, each corresponding to a specific phase of the cell cycle. This approach will allow direct observation of the expression patterns of the newly identified cycle genes. Additionally, cell cycle-specific motif pairs identified by PMET highlight key TFs involved in cell cycle regulation. Introducing T-DNA insertions into these TF genes will disrupt their regulatory functions, impeding the cell cycle process and enabling direct experimental observation of regulatory dynamics — ultimately validating the roles of these cell cycle genes.

By combining experimental validation using the COLORFUL-Circuit platform with T-DNA insertion technology, this approach will confirm the accuracy of the newly identified cell cycle genes and empirically validate computational predictions from PMET analysis.

Bibliography

[1] Inzé, D.; De Veylder, L. Cell cycle regulation in plant development. *Annu. Rev. Genet.* **2006**, *40*, 77–105.

[2] Gutierrez, C. The Arabidopsis cell division cycle. *The Arabidopsis Book/American Society of Plant Biologists* **2009**, *7*.

[3] Francis, D. What's New in the Plant Cell Cycle? *Progress in botany* **2009**, 33–49.

[4] Alberts, B.; Johnson, A.; Lewis, J.; Raff, M.; Roberts, K.; Walter, P. An overview of the cell cycle. *Molecular Biology of the Cell. 4th edition* **2002**,

[5] Kalucka, J.; Missiaen, R.; Georgiadou, M.; Schoors, S.; Lange, C.; De Bock, K.; Dewerchin, M.; Carmeliet, P. Metabolic control of the cell cycle. *Cell cycle* **2015**, *14*, 3379–3388.

[6] Icard, P.; Simula, L. Metabolic oscillations during cell-cycle progression. *Trends in Endocrinology & Metabolism* **2022**, *33*, 447–450.

[7] Willis, N. A.; Zhou, C.; Elia, A. E.; Murray, J. M.; Carr, A. M.; Elledge, S. J.; Rhind, N. Identification of S-phase DNA damage-response targets in fission yeast reveals conservation of damage-response networks. *Proceedings of the National Academy of Sciences* **2016**, *113*, E3676–E3685.

[8] Morgan, D. O. *The cell cycle: principles of control*; New science press, 2007.

[9] Das-Bradoo, S.; Bielinsky, A. DNA replication and checkpoint control in S phase. *Nature* **2010**, *9*, 74–79.

[10] Giotti, B.; Joshi, A.; Freeman, T. C. Meta-analysis reveals conserved cell cycle transcriptional network across multiple human cell types. *Bmc Genomics* **2017**, *18*, 1–12.

[11] Harvey, S. L.; Kellogg, D. R. Conservation of mechanisms controlling entry into mitosis: budding yeast wee1 delays entry into mitosis and is required for cell size control. *Current Biology* **2003**, *13*, 264–275.

[12] Blomen, V.; Boonstra, J. Cell fate determination during G1 phase progression. *Cellular and Molecular Life Sciences* **2007**, *64*, 3084–3104.

[13] Barnum, K. J.; O' Connell, M. J. Cell cycle regulation by checkpoints. *Cell cycle control: mechanisms and protocols* **2014**, 29–40.

[14] Jacobs, T. Control of the cell cycle. *Developmental biology* **1992**, *153*, 1–15.

[15] De Veylder, L.; Beeckman, T.; Inzé, D. The ins and outs of the plant cell cycle. *Nature Reviews Molecular Cell Biology* **2007**, *8*, 655–665.

[16] García-Gómez, M. L.; Ornelas-Ayala, D.; Garay-Arroyo, A.; García-Ponce, B.; Sánchez, M. d. l. P.; Álvarez-Buylla, E. R. A system-level mechanistic explanation for asymmetric stem cell fates: Arabidopsis thaliana root niche as a study system. *Scientific Reports* **2020**, *10*, 3525.

[17] Menke, F. L.; Scheres, B. Plant asymmetric cell division, vive la différence! *Cell* **2009**, *137*, 1189–1192.

[18] Edgar, B. A.; Orr-Weaver, T. L. Endoreplication cell cycles: more for less. *Cell* **2001**, *105*, 297–306.

[19] Breuer, C.; Braidwood, L.; Sugimoto, K. Endocycling in the path of plant development. *Current opinion in plant biology* **2014**, *17*, 78–85.

[20] Kondorosi, E.; Roudier, F.; Gendreau, E. Plant cell-size control: growing by ploidy? *Current opinion in plant biology* **2000**, *3*, 488–492.

[21] Larkins, B. A.; Dilkes, B. P.; Dante, R. A.; Coelho, C. M.; Woo, Y.-m.; Liu, Y. Investigating the hows and whys of DNA endoreduplication. *Journal of Experimental Botany* **2001**, *52*, 183–192.

[22] Caro, E.; Desvoyes, B.; Ramirez-Parra, E.; Sanchez, M.; Gutierrez, C. Endoreduplication control during plant development. *SEB Experimental Biology Series* **2008**, *59*, 167–187.

[23] De Veylder, L.; Larkin, J. C.; Schnittger, A. Molecular control and function of endoreplication in development and physiology. *Trends in plant science* **2011**, *16*, 624–634.

[24] Kasili, R.; Huang, C.-C.; Walker, J. D.; Simmons, L. A.; Zhou, J.; Faulk, C.; Hülskamp, M.; Larkin, J. C. BRANCHLESS TRICHOMES links cell shape and cell cycle control in Arabidopsis trichomes. *Development* **2011**, *138*, 2379–2388.

[25] Ramirez-Parra, E.; Gutierrez, C. The many faces of chromatin assembly factor 1. *Trends in plant science* **2007**, *12*, 570–576.

[26] Castellano, M. d. M.; Boniotti, M. B.; Caro, E.; Schnittger, A.; Gutierrez, C. DNA replication licensing affects cell proliferation or endoreplication in a cell type–specific manner. *The Plant Cell* **2004**, *16*, 2380–2393.

[27] Breuer, C.; Stacey, N. J.; West, C. E.; Zhao, Y.; Chory, J.; Tsukaya, H.; Azumi, Y.; Maxwell, A.; Roberts, K.; Sugimoto-Shirasu, K. BIN4, a novel component of the plant DNA topoisomerase VI complex, is required for endoreduplication in Arabidopsis. *The Plant Cell* **2007**, *19*, 3655–3668.

[28] Shimotohno, A.; Aki, S. S.; Takahashi, N.; Umeda, M. Regulation of the plant cell cycle in response to hormones and the environment. *Annual review of plant biology* **2021**, *72*, 273–296.

[29] Francis, D. The plant cell cycle- 15 years on. *New Phytologist* **2007**, *174*, 261–278.

[30] Bramsiepe, J.; Wester, K.; Weinl, C.; Roodbarkelari, F.; Kasili, R.; Larkin, J. C.; Hülskamp, M.; Schnittger, A. Endoreplication controls cell fate maintenance. *PLoS genetics* **2010**, *6*, e1000996.

[31] Hulskamp, M.; Schnittger, A.; Folkers, U. Pattern formation and cell differentiation: trichomes in Arabidopsis as a genetic model system. *International review of cytology* **1998**, *186*, 147–178.

[32] Boudolf, V.; Vlieghe, K.; Beemster, G. T.; Magyar, Z.; Acosta, J. A. T.; Maes, S.; Van Der Schueren, E.; Inzeé, D.; De Veylder, L. The plant-specific cyclin-dependent kinase CDKB1; 1 and transcription factor E2Fa-DPa control the balance of mitotically dividing and endoreduplicating cells in Arabidopsis. *The Plant Cell* **2004**, *16*, 2683–2692.

[33] Melaragno, J. E.; Mehrotra, B.; Coleman, A. W. Relationship between endopolyploidy and cell size in epidermal tissue of Arabidopsis. *The Plant Cell* **1993**, *5*, 1661–1668.

[34] Traas, J.; Hülskamp, M.; Gendreau, E.; Höfte, H. Endoreduplication and development: rule without dividing? *Current opinion in plant biology* **1998**, *1*, 498–503.

[35] Beemster, G. T.; De Vusser, K.; De Tavernier, E.; De Bock, K.; Inzé, D. Variation in growth rate between Arabidopsis ecotypes is correlated with cell division and A-type cyclin-dependent kinase activity. *Plant physiology* **2002**, *129*, 854–864.

[36] Sugimoto-Shirasu, K.; Roberts, K. "Big it up" : endoreduplication and cell-size control in plants. *Current opinion in plant biology* **2003**, *6*, 544–553.

[37] Harashima, H.; Schnittger, A. The integration of cell division, growth and differentiation. *Current opinion in plant biology* **2010**, *13*, 66–74.

[38] Vandepoele, K.; Raes, J.; De Veylder, L.; Rouzé, P.; Rombauts, S.; Inzé, D. Genome-wide analysis of core cell cycle genes in Arabidopsis. *The Plant Cell* **2002**, *14*, 903–916.

[39] Menges, M.; De Jager, S. M.; Gruissem, W.; Murray, J. A. Global analysis of the core cell cycle regulators of Arabidopsis identifies novel genes, reveals multiple and highly specific profiles of expression and provides a coherent model for plant cell cycle control. *The Plant Journal* **2005**, *41*, 546–566.

[40] Inzé, D. *Annual Plant Reviews, Cell Cycle Control and Plant Development*; John Wiley & Sons, 2007; Vol. 32.

[41] Joubès, J.; Chevalier, C.; Dudits, D.; Heberle-Bors, E.; Inzé, D.; Umeda, M.; Renaudin, J.-P. CDK-related protein kinases in plants. *The plant cell cycle* **2000**, 63–76.

[42] Porceddu, A.; Stals, H.; Reichheld, J.-P.; Segers, G.; De Veylder, L.; de Pinho Barrôco, R.; Casteels, P.; Van Montagu, M.; Inzé, D.; Mironov, V. A plant-specific cyclin-dependent kinase

is involved in the control of G2/M progression in plants. *Journal of Biological Chemistry* **2001**, *276*, 36354–36360.

[43] Shimotohno, A.; Umeda, M. CDK phosphorylation. *Annual Plant Reviews Volume 32: Cell Cycle Control and Plant Development* **2007**, 114–137.

[44] Kumar, N.; Harashima, H.; Kalve, S.; Bramsiepe, J.; Wang, K.; Sizani, B. L.; Bertrand, L. L.; Johnson, M. C.; Faulk, C.; Dale, R.; others Functional conservation in the SIAMESE-RELATED family of cyclin-dependent kinase inhibitors in land plants. *The Plant Cell* **2015**, *27*, 3065–3080.

[45] Evans, T.; Rosenthal, E. T.; Youngblom, J.; Distel, D.; Hunt, T. Cyclin: a protein specified by maternal mRNA in sea urchin eggs that is destroyed at each cleavage division. *Cell* **1983**, *33*, 389–396.

[46] Renaudin, J.-P.; Doonan, J. H.; Freeman, D.; Hashimoto, J.; Hirt, H.; Inzé, D.; Jacobs, T.; Kouchi, H.; Rouzé, P.; Sauter, M.; others Plant cyclins: a unified nomenclature for plant A-, B-and D-type cyclins based on sequence organization. *Plant molecular biology* **1996**, *32*, 1003–1018.

[47] Wei, W.; Ayad, N. G.; Wan, Y.; Zhang, G.-J.; Kirschner, M. W.; Kaelin Jr, W. G. Degradation of the SCF component Skp2 in cell-cycle phase G1 by the anaphase-promoting complex. *Nature* **2004**, *428*, 194–198.

[48] Yamaguchi, M.; Fabian, T.; Sauter, M.; Bhalerao, R. P.; Schrader, J.; Sandberg, G.; Umeda, M.; Uchimiya, H. Activation of CDK-activating kinase is dependent on interaction with H-type cyclins in plants. *The Plant Journal* **2000**, *24*, 11–20.

[49] La, H.; Li, J.; Ji, Z.; Cheng, Y.; Li, X.; Jiang, S.; Venkatesh, P. N.; Ramachandran, S. Genome-wide analysis of cyclin family in rice (Oryza Sativa L.). *Molecular Genetics and Genomics* **2006**, *275*, 374–386.

[50] Abrahams, S.; Cavet, G.; Oakenfull, E. A.; Carmichael, J. P.; Shah, Z. H.; Soni, R.; Murray, J. A. A novel and highly divergent Arabidopsis cyclin isolated by complementation in budding yeast. *Biochimica et Biophysica Acta (BBA)-Molecular Cell Research* **2001**, *1539*, 1–6.

[51] Rossignol, P.; Stevens, R.; Perennes, C.; Jasinski, S.; Cella, R.; Tremousaygue, D.; Bergounioux, C. AtE2F-a and AtDP-a, members of the E2F family of transcription factors, induce Arabidopsis leaf cells to re-enter S phase. *Molecular Genetics and Genomics* **2002**, *266*, 995–1003.

[52] del Pozo, J. C.; Diaz-Trivino, S.; Cisneros, N.; Gutierrez, C. The balance between cell division and endoreplication depends on E2FC-DPB, transcription factors regulated by the ubiquitin-SCFSKP2A pathway in Arabidopsis. *The Plant Cell* **2006**, *18*, 2224–2235.

[53] Sozzani, R.; Maggio, C.; Varotto, S.; Canova, S.; Bergounioux, C.; Albani, D.; Cella, R. Interplay between Arabidopsis activating factors E2Fb and E2Fa in cell cycle progression and development. *Plant physiology* **2006**, *140*, 1355–1366.

[54] Mariconti, L.; Pellegrini, B.; Cantoni, R.; Stevens, R.; Bergounioux, C.; Cella, R.; Albani, D. The E2F family of transcription factors from Arabidopsis thaliana: novel and conserved components of the retinoblastoma/E2F pathway in plants. *Journal of Biological Chemistry* **2002**, *277*, 9911–9919.

[55] Gutierrez, C. Coupling cell proliferation and development in plants. *Nature cell biology* **2005**, *7*, 535–541.

[56] Nakagami, H.; Kawamura, K.; Sugisaka, K.; Sekine, M.; Shinmyo, A. Phosphorylation of retinoblastoma-related protein by the cyclin D/cyclin-dependent kinase complex is activated at the G1/S-phase transition in tobacco. *The Plant Cell* **2002**, *14*, 1847–1857.

[57] Boniotti, M. B.; Gutierrez, C. A cell-cycle-regulated kinase activity phosphorylates plant retinoblastoma protein and contains, in Arabidopsis, a CDKA/cyclin D complex. *The Plant Journal* **2001**, *28*, 341–350.

[58] Kosugi, S.; Ohashi, Y. Interaction of the Arabidopsis E2F and DP proteins confers their concomitant nuclear translocation and transactivation. *Plant physiology* **2002**, *128*, 833–843.

[59] Del Pozo, J. C.; Boniotti, M. B.; Gutierrez, C. Arabidopsis E2Fc functions in cell division and is degraded by the ubiquitin-SCFAtSKP2 pathway in response to light. *The Plant Cell* **2002**, *14*, 3057–3071.

[60] Chen, P.; Takatsuka, H.; Takahashi, N.; Kurata, R.; Fukao, Y.; Kobayashi, K.; Ito, M.; Umeda, M. Arabidopsis R1R2R3-Myb proteins are essential for inhibiting cell division in response to DNA damage. *Nature communications* **2017**, *8*, 635.

[61] Sun, Y.; Dilkes, B. P.; Zhang, C.; Dante, R. A.; Carneiro, N. P.; Lowe, K. S.; Jung, R.; Gordon-Kamm, W. J.; Larkins, B. A. Characterization of maize (Zea mays L.) Wee1 and its activity in developing endosperm. *Proceedings of the National Academy of Sciences* **1999**, *96*, 4180–4185.

[62] Sorrell, D. A.; Marchbank, A.; McMahon, K.; Dickinson, R. J.; Rogers, H. J.; Francis, D. A WEE1 homologue from Arabidopsis thaliana. *Planta* **2002**, *215*, 518–522.

[63] McGowan, C. H.; Russell, P. Cell cycle regulation of human WEE1. *The EMBO journal* **1995**, *14*, 2166–2175.

[64] Alfieri, C.; Zhang, S.; Barford, D. Visualizing the complex functions and mechanisms of the anaphase promoting complex/cyclosome (APC/C). *Open biology* **2017**, *7*, 170204.

[65] Capron, A.; Ökrész, L.; Genschik, P. First glance at the plant APC/C, a highly conserved ubiquitin–protein ligase. *Trends in plant science* **2003**, *8*, 83–89.

[66] Jasinski, S.; Riou-Khamlichi, C.; Roche, O.; Perennes, C.; Bergounioux, C.; Glab, N. The CDK inhibitor NtKIS1a is involved in plant development, endoreduplication and restores normal development of cyclin D3; 1-overexpressing plants. *Journal of Cell Science* **2002**, *115*, 973–982.

[67] Churchman, M. L.; Brown, M. L.; Kato, N.; Kirik, V.; Hulskamp, M.; Inze, D.; De Veylder, L.; Walker, J. D.; Zheng, Z.; Oppenheimer, D. G.; others SIAMESE, a plant-specific cell cycle regulator, controls endoreplication onset in Arabidopsis thaliana. *The Plant Cell* **2006**, *18*, 3145–3157.

[68] Wang, K.; Ndathe, R. W.; Kumar, N.; Zeringue, E. A.; Kato, N.; Larkin, J. C. The CDK inhibitor SIAMESE targets both CDKA; 1 and CDKB1 complexes to establish endoreplication in trichomes. *Plant physiology* **2020**, *184*, 165–175.

[69] Liang, J.-W.; Tian, F.-L.; Lan, Z.-R.; Huang, B.; Zhuang, W.-Z. Selection characterization on overlapping reading frame of multiple-protein-encoding P gene in Newcastle disease virus. *Veterinary Microbiology* **2010**, *144*, 257–263.

[70] Huang, Y.; Wang, X.; Zhang, F.; Huo, X.; Fu, R.; Liu, J.; Sun, W.; Kang, D.; Jing, X. The identification of a bacterial strain BGI-1 isolated from the intestinal flora of Blattella germanica, and its anti-entomopathogenic fungi activity. *Journal of economic entomology* **2013**, *106*, 43–49.

[71] Zhang, F.; Wang, X.; Huang, Y.; Zhao, Z.; Zhang, S.; Gong, X.; Xie, L.; Kang, D.; Jing, X. Differential expression of hemolymph proteins between susceptible and insecticide-resistant Blattella germanica (Blattodea: Blattellidae). *Environmental entomology* **2014**, *43*, 1117–1123.

[72] Zhang, F.; Liu, D.; Wang, L.; Li, T.; Chang, Q.; An, L.; Yang, G. Characterization of IgM-binding protein: A pIgR-like molecule expressed by intestinal epithelial cells in the common carp (Cyprinus carpio L.). *Veterinary immunology and immunopathology* **2015**, *167*, 30–35.

[73] Xing, N.; Ji, L.; Song, J.; Ma, J.; Li, S.; Ren, Z.; Xu, F.; Zhu, J. Cadmium stress assessment based on the electrocardiogram characteristics of zebra fish (Danio rerio): QRS complex could play an important role. *Aquatic Toxicology* **2017**, *191*, 236–244.

[74] Zhao, S.; Jiang, Y.; Zhao, Y.; Huang, S.; Yuan, M.; Zhao, Y.; Guo, Y. CASEIN KINASE1-LIKE PROTEIN2 regulates actin filament stability and stomatal closure via phosphorylation of actin depolymerizing factor. *The Plant Cell* **2016**, *28*, 1422–1439.

[75] Tang, G.; Shao, F.; Xu, P.; Shan, L.; Liu, Z. Overexpression of a peanut NAC gene, AhNAC4, confers enhanced drought tolerance in tobacco. *Russian Journal of Plant Physiology* **2017**, *64*, 525–535.

[76] Sui, N. Photoinhibition of Suaeda salsa to chilling stress is related to energy dissipation and water-water cycle. *Photosynthetica* **2015**, *53*, 207–212.

[77] Qu, A.-L.; Ding, Y.-F.; Jiang, Q.; Zhu, C. Molecular mechanisms of the plant heat stress response. *Biochemical and biophysical research communications* **2013**, *432*, 203–207.

[78] Jagadish, S. K.; Way, D. A.; Sharkey, T. D. Plant heat stress: Concepts directing future research. *Plant, cell & environment* **2021**, *44*, 1992–2005.

[79] Cui, F.; Sui, N.; Duan, G.; Liu, Y.; Han, Y.; Liu, S.; Wan, S.; Li, G. Identification of metabolites and transcripts involved in salt stress and recovery in peanut. *Frontiers in Plant Science* **2018**, *9*, 217.

[80] Shavrukov, Y.; Hirai, Y. Good and bad protons: genetic aspects of acidity stress responses in plants. *Journal of experimental botany* **2016**, *67*, 15–30.

[81] Sharma, S.; Chatterjee, S.; Kataria, S.; Joshi, J.; Datta, S.; Vairale, M. G.; Veer, V. A review on responses of plants to UV-B radiation related stress. *UV-B Radiation: From Environmental Stressor to Regulator of Plant Growth* **2017**, 75–97.

[82] Ghori, N.-H.; Ghori, T.; Hayat, M.; Imadi, S.; Gul, A.; Altay, V.; Ozturk, M. Heavy metal stress and responses in plants. *International journal of environmental science and technology* **2019**, *16*, 1807–1828.

[83] Lichtenthaler, H. K. The stress concept in plants: an introduction. *Annals of the new York Academy of sciences* **1998**, *851*, 187–198.

[84] Qi, F.; Zhang, F. Cell cycle regulation in the plant response to stress. *Frontiers in plant science* **2020**, *10*, 1765.

[85] Takahashi, N.; Ogita, N.; Takahashi, T.; Taniguchi, S.; Tanaka, M.; Seki, M.; Umeda, M. A regulatory module controlling stress-induced cell cycle arrest in Arabidopsis. *Elife* **2019**, *8*, e43944.

[86] Meguro, A.; Sato, Y. Salicylic acid antagonizes abscisic acid inhibition of shoot growth and cell cycle progression in rice. *Scientific reports* **2014**, *4*, 4555.

[87] Li, F.; Wang, L.; Zhang, Z.; Li, T.; Feng, J.; Xu, S.; Zhang, R.; Guo, D.; Xue, J. ZmSMR4, a novel cyclin-dependent kinase inhibitor (CKI) gene in maize (Zea mays L.), functions as a key player in plant growth, development and tolerance to abiotic stress. *Plant Science* **2019**, *280*, 120–131.

[88] Peres, A.; Churchman, M. L.; Hariharan, S.; Himanen, K.; Verkest, A.; Vandepoele, K.; Magyar, Z.; Hatzfeld, Y.; Van Der Schueren, E.; Beemster, G. T.; others Novel plant-specific cyclin-dependent kinase inhibitors induced by biotic and abiotic stresses. *Journal of Biological Chemistry* **2007**, *282*, 25588–25596.

[89] Yi, D.; Alvim Kamei, C. L.; Cools, T.; Vanderauwera, S.; Takahashi, N.; Okushima, Y.; Eekhout, T.; Yoshiyama, K. O.; Larkin, J.; Van den Daele, H.; others The Arabidopsis SIAMESE-RELATED cyclin-dependent kinase inhibitors SMR5 and SMR7 regulate the DNA damage checkpoint in response to reactive oxygen species. *The Plant Cell* **2014**, *26*, 296–309.

[90] Jones, J. D.; Dangl, J. L. The plant immune system. *nature* **2006**, *444*, 323–329.

[91] Wang, S.; Gu, Y.; Zebell, S. G.; Anderson, L. K.; Wang, W.; Mohan, R.; Dong, X. A noncanonical role for the CKI-RB-E2F cell-cycle signaling pathway in plant effector-triggered immunity. *Cell host & microbe* **2014**, *16*, 787–794.

[92] Chandran, D.; Rickert, J.; Huang, Y.; Steinwand, M. A.; Marr, S. K.; Wildermuth, M. C. Atypical E2F transcriptional repressor DEL1 acts at the intersection of plant growth and immunity by controlling the hormone salicylic acid. *Cell host & microbe* **2014**, *15*, 506–513.

[93] Ascencio-Ibánez, J. T.; Sozzani, R.; Lee, T.-J.; Chu, T.-M.; Wolfinger, R. D.; Cella, R.; Hanley-Bowdoin, L. Global analysis of Arabidopsis gene expression uncovers a complex array of changes impacting pathogen response and cell cycle during geminivirus infection. *Plant physiology* **2008**, *148*, 436–454.

[94] Chandran, D.; Tai, Y. C.; Hather, G.; Dewdney, J.; Denoux, C.; Burgess, D. G.; Ausubel, F. M.; Speed, T. P.; Wildermuth, M. C. Temporal global expression data reveal known and novel salicylate-impacted processes and regulators mediating powdery mildew growth and reproduction on Arabidopsis. *Plant physiology* **2009**, *149*, 1435–1451.

[95] Buenrostro, J. D.; Giresi, P. G.; Zaba, L. C.; Chang, H. Y.; Greenleaf, W. J. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nature methods* **2013**, *10*, 1213–1218.

[96] Belton, J.-M.; McCord, R. P.; Gibcus, J. H.; Naumova, N.; Zhan, Y.; Dekker, J. Hi–C: a comprehensive technique to capture the conformation of genomes. *Methods* **2012**, *58*, 268–276.

[97] Crawford, G. E.; Holt, I. E.; Whittle, J.; Webb, B. D.; Tai, D.; Davis, S.; Margulies, E. H.; Chen, Y.; Bernat, J. A.; Ginsburg, D.; others Genome-wide mapping of DNase hypersensitive sites using massively parallel signature sequencing (MPSS). *Genome research* **2006**, *16*, 123–131.

[98] Garavello, M.; Cuenca, J.; Dreissig, S.; Fuchs, J.; Houben, A.; Aleza, P. Assessing ploidy level analysis and single pollen genotyping of diploid and euploid citrus genotypes by fluorescence-activated cell sorting and whole-genome amplification. *Frontiers in Plant Science* **2019**, *10*, 483186.

[99] Chen, X.; Xu, H.; Shu, X.; Song, C.-X. Mapping epigenetic modifications by sequencing technologies. *Cell Death & Differentiation* **2023**, 1–10.

[100] Zatopek, K. M.; Potapov, V.; Maduzia, L. L.; Alpaslan, E.; Chen, L.; Evans Jr, T. C.; Ong, J. L.; Ettwiller, L. M.; Gardner, A. F. RADAR-seq: A RAre DAmage and Repair sequencing method for detecting DNA damage on a genome-wide scale. *DNA repair* **2019**, *80*, 36–44.

[101] Yang, C.-C.; Chen, M.-H.; Lin, S.-Y.; Andrews, E. H.; Cheng, C.; Liu, C.-C.; Chen, J. J. Inferring condition-specific targets of human TF-TF complexes using ChIP-seq data. *BMC genomics* **2017**, *18*, 1–10.

[102] Vogel, M. J.; Peric-Hupkes, D.; Van Steensel, B. Detection of in vivo protein–DNA interactions using DamID in mammalian cells. *Nature protocols* **2007**, *2*, 1467–1478.

[103] Pott, S.; Lieb, J. D. Single-cell ATAC-seq: strength in numbers. *Genome biology* **2015**, *16*, 1–4.

[104] Nagano, T.; Lubling, Y.; Stevens, T. J.; Schoenfelder, S.; Yaffe, E.; Dean, W.; Laue, E. D.; Tanay, A.; Fraser, P. Single-cell Hi-C reveals cell-to-cell variability in chromosome structure. *Nature* **2013**, *502*, 59–64.

[105] Jin, W.; Tang, Q.; Wan, M.; Cui, K.; Zhang, Y.; Ren, G.; Ni, B.; Sklar, J.; Przytycka, T. M.; Childs, R.; others Genome-wide detection of DNase I hypersensitive sites in single cells and FFPE tissue samples. *Nature* **2015**, *528*, 142–146.

[106] Grosselin, K.; Durand, A.; Marsolier, J.; Poitou, A.; Marangoni, E.; Nemati, F.; Dahmani, A.; Lameiras, S.; Reyal, F.; Frenoy, O.; others High-throughput single-cell ChIP-seq identifies heterogeneity of chromatin states in breast cancer. *Nature genetics* **2019**, *51*, 1060–1066.

[107] Goldman, S. L.; MacKay, M.; Afshinnekoo, E.; Melnick, A. M.; Wu, S.; Mason, C. E. The impact of heterogeneity on single-cell sequencing. *Frontiers in genetics* **2019**, *10*, 8.

[108] Altschuler, S. J.; Wu, L. F. Cellular heterogeneity: do differences make a difference? *Cell* **2010**, *141*, 559–563.

[109] Elsasser, W. M. Outline of a theory of cellular heterogeneity. *Proceedings of the National Academy of Sciences* **1984**, *81*, 5126–5129.

[110] Hadjantonakis, A.-K.; Arias, A. M. Single-cell approaches: pandora' s box of developmental mechanisms. *Developmental Cell* **2016**, *38*, 574–578.

[111] Tirosh, I.; Izar, B.; Prakadan, S. M.; Wadsworth, M. H.; Treacy, D.; Trombetta, J. J.; Rotem, A.; Rodman, C.; Lian, C.; Murphy, G.; others Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-Seq. *Science* **2016**, *352*, 189–196.

[112] Jovic, D.; Liang, X.; Zeng, H.; Lin, L.; Xu, F.; Luo, Y. Single-cell RNA sequencing technologies and applications: A brief overview. *Clinical and translational medicine* **2022**, *12*, e694.

[113] Miyamoto, D. T.; Zheng, Y.; Wittner, B. S.; Lee, R. J.; Zhu, H.; Broderick, K. T.; Desai, R.; Fox, D. B.; Brannigan, B. W.; Trautwein, J.; others RNA-Seq of single prostate CTCs implicates noncanonical Wnt signaling in antiandrogen resistance. *Science* **2015**, *349*, 1351–1356.

[114] Angermueller, C.; Clark, S. J.; Lee, H. J.; Macaulay, I. C.; Teng, M. J.; Hu, T. X.; Krueger, F.; Smallwood, S. A.; Ponting, C. P.; Voet, T.; others Parallel single-cell sequencing links transcriptional and epigenetic heterogeneity. *Nature methods* **2016**, *13*, 229–232.

[115] Rooijers, K.; Markodimitraki, C. M.; Rang, F. J.; de Vries, S. S.; Chialastri, A.; de Luca, K. L.; Mooijman, D.; Dey, S. S.; Kind, J. Simultaneous quantification of protein–DNA contacts and transcriptomes in single cells. *Nature biotechnology* **2019**, *37*, 766–772.

[116] Markodimitraki, C. M.; Rang, F. J.; Rooijers, K.; de Vries, S. S.; Chialastri, A.; de Luca, K. L.; Lochs, S. J.; Mooijman, D.; Dey, S. S.; Kind, J. Simultaneous quantification of protein–DNA interactions and transcriptomes in single cells with scDam&T-seq. *Nature protocols* **2020**, *15*, 1922–1953.

[117] Hashimshony, T.; Wagner, F.; Sher, N.; Yanai, I. CEL-Seq: single-cell RNA-Seq by multiplexed linear amplification. *Cell reports* **2012**, *2*, 666–673.

[118] Cusanovich, D. A.; Daza, R.; Adey, A.; Pliner, H. A.; Christiansen, L.; Gunderson, K. L.; Steemers, F. J.; Trapnell, C.; Shendure, J. Multiplex single-cell profiling of chromatin accessibility by combinatorial cellular indexing. *Science* **2015**, *348*, 910–914.

[119] Tang, F.; Barbacioru, C.; Wang, Y.; Nordman, E.; Lee, C.; Xu, N.; Wang, X.; Bodeau, J.; Tuch, B. B.; Siddiqui, A.; others mRNA-Seq whole-transcriptome analysis of a single cell. *Nature methods* **2009**, *6*, 377–382.

[120] Islam, S.; Kjällquist, U.; Moliner, A.; Zajac, P.; Fan, J.-B.; Lönnerberg, P.; Linnarsson, S. Characterization of the single-cell transcriptional landscape by highly multiplex RNA-seq. *Genome research* **2011**, *21*, 1160–1167.

[121] Herzenberg, L. A.; Sweet, R. G.; Herzenberg, L. A. Fluorescence-activated cell sorting. *Scientific American* **1976**, *234*, 108–118.

[122] Espina, V.; Wulfkuhle, J. D.; Calvert, V. S.; VanMeter, A.; Zhou, W.; Coukos, G.; Geho, D. H.; Petricoin III, E. F.; Liotta, L. A. Laser-capture microdissection. *Nature protocols* **2006**, *1*, 586–603.

[123] Whitesides, G. M. The origins and the future of microfluidics. *nature* **2006**, *442*, 368–373.

[124] Wang, Z.; Huang, A. S.; Tang, L.; Wang, J.; Wang, G. Microfluidic-assisted single-cell RNA sequencing facilitates the development of neutralizing monoclonal antibodies against SARS-CoV-2. *Lab on a Chip* **2024**, *24*, 642–657.

[125] Diacumakos, E. G. *Methods in cell biology*; Elsevier, 1974; Vol. 7; pp 287–311.

[126] Zhang, Z.; Ferenczi, M.; Thomas, C. A micromanipulation technique with a theoretical cell model for determining mechanical properties of single mammalian cells. *Chemical engineering science* **1992**, *47*, 1347–1354.

[127] Goetz, J. J.; Trimarchi, J. M. Transcriptome sequencing of single cells with Smart-Seq. *Nature biotechnology* **2012**, *30*, 763–765.

[128] Picelli, S.; Björklund, Å. K.; Faridani, O. R.; Sagasser, S.; Winberg, G.; Sandberg, R. Smart-seq2 for sensitive full-length transcriptome profiling in single cells. *Nature methods* **2013**, *10*, 1096–1098.

[129] Sheng, K.; Cao, W.; Niu, Y.; Deng, Q.; Zong, C. Effective detection of variation in single-cell transcriptomes using MATQ-seq. *Nature methods* **2017**, *14*, 267–270.

[130] Jaitin, D. A.; Kenigsberg, E.; Keren-Shaul, H.; Elefant, N.; Paul, F.; Zaretsky, I.; Mildner, A.; Cohen, N.; Jung, S.; Tanay, A.; others Massively parallel single-cell RNA-seq for marker-free decomposition of tissues into cell types. *Science* **2014**, *343*, 776–779.

[131] Islam, S.; Zeisel, A.; Joost, S.; La Manno, G.; Zajac, P.; Kasper, M.; Lönnerberg, P.; Linnarsson, S. Quantitative single-cell RNA-seq with unique molecular identifiers. *Nature methods* **2014**, *11*, 163–166.

[132] Danielski, K. *Single Cell Transcriptomics: Methods and Protocols*; Springer, 2022; pp 1–28.

[133] Macosko, E. Z.; Basu, A.; Satija, R.; Nemesh, J.; Shekhar, K.; Goldman, M.; Tirosh, I.; Bialas, A. R.; Kamitaki, N.; Martersteck, E. M.; others Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell* **2015**, *161*, 1202–1214.

[134] Zheng, G. X.; Terry, J. M.; Belgrader, P.; Ryvkin, P.; Bent, Z. W.; Wilson, R.; Ziraldo, S. B.; Wheeler, T. D.; McDermott, G. P.; Zhu, J.; others Massively parallel digital transcriptional profiling of single cells. *Nature communications* **2017**, *8*, 14049.

[135] Chen, X.; Teichmann, S. A.; Meyer, K. B. From tissues to cell types and back: single-cell gene expression analysis of tissue architecture. *Annual Review of Biomedical Data Science* **2018**, *1*, 29–51.

[136] Kolodziejczyk, A. A.; Kim, J. K.; Svensson, V.; Marioni, J. C.; Teichmann, S. A. The technology and biology of single-cell RNA sequencing. *Molecular cell* **2015**, *58*, 610–620.

[137] Tanay, A.; Regev, A. Scaling single-cell genomics from phenomenology to mechanism. *Nature* **2017**, *541*, 331–338.

[138] Wagner, A.; Regev, A.; Yosef, N. Revealing the vectors of cellular identity with single-cell genomics. *Nature biotechnology* **2016**, *34*, 1145–1160.

[139] Qian, Z.; Bao, L. *Single-cell omics*; Elsevier, 2019; pp 35–44.

[140] Ianevski, A.; Giri, A. K.; Aittokallio, T. Fully-automated and ultra-fast cell-type identification using specific marker combinations from single-cell transcriptomic data. *Nature communications* **2022**, *13*, 1246.

[141] Shekhar, K.; Menon, V. *Identification of cell types from single-cell transcriptomic data*; Springer, 2019.

[142] Cuomo, A. S.; Seaton, D. D.; McCarthy, D. J.; Martinez, I.; Bonder, M. J.; Garcia-Bernardo, J.; Amatya, S.; Madrigal, P.; Isaacson, A.; Buettner, F.; others Single-cell RNA-sequencing of differentiating iPS cells reveals dynamic genetic effects on gene expression. *Nature communications* **2020**, *11*, 810.

[143] Gulati, G. S.; Sikandar, S. S.; Wesche, D. J.; Manjunath, A.; Bharadwaj, A.; Berger, M. J.; Ilagan, F.; Kuo, A. H.; Hsieh, R. W.; Cai, S.; others Single-cell transcriptional diversity is a hallmark of developmental potential. *Science* **2020**, *367*, 405–411.

[144] Latchman, D. S. Transcription factors: an overview. *The international journal of biochemistry & cell biology* **1997**, *29*, 1305–1312.

[145] Lambert, S. A.; Jolma, A.; Campitelli, L. F.; Das, P. K.; Yin, Y.; Albu, M.; Chen, X.; Taipale, J.; Hughes, T. R.; Weirauch, M. T. The human transcription factors. *Cell* **2018**, *172*, 650–665.

[146] Lee, T. I.; Young, R. A. Transcriptional regulation and its misregulation in disease. *Cell* **2013**, *152*, 1237–1251.

[147] Young, R. A. Control of the embryonic stem cell state. *Cell* **2011**, *144*, 940–954.

[148] Damante, G.; Fabbro, D.; Pellizzari, L.; Civitareale, D.; Guazzi, S.; Polycarpou-Schwartz, M.; Cauci, S.; Quadrifoglio, F.; Formisano, S.; Lauro, R. D. Sequence-specific DNA recognition by the thyroid transcription factor-1 homeodomain. *Nucleic acids research* **1994**, *22*, 3075–3083.

[149] Geertz, M.; Shore, D.; Maerkl, S. J. Massively parallel measurements of molecular interaction kinetics on a microfluidic platform. *Proceedings of the National Academy of Sciences* **2012**, *109*, 16540–16545.

[150] Boyle, P.; Després, C. Dual-function transcription factors and their entourage: unique and unifying themes governing two pathogenesis-related genes. *Plant signaling & behavior* **2010**, *5*, 629–634.

[151] Mo, X.; Kowenz-Leutz, E.; Xu, H.; Leutz, A. Ras induces mediator complex exchange on C/EBPβ. *Molecular cell* **2004**, *13*, 241–250.

[152] Valin, A.; Gill, G. Regulation of the dual-function transcription factor Sp3 by SUMO. *Biochemical Society Transactions* **2007**, *35*, 1393–1396.

[153] Collins, F. S.; Green, E. D.; Guttmacher, A. E.; Guyer, M. S.; Institute, U. N. H. G. R. A vision for the future of genomics research. *nature* **2003**, *422*, 835–847.

[154] Bickhart, D. M.; Liu, G. E. Identification of candidate transcription factor binding sites in the cattle genome. *Genomics, proteomics & bioinformatics* **2013**, *11*, 195–198.

[155] Stewart, A. J.; Hannenhalli, S.; Plotkin, J. B. Why transcription factor binding sites are ten nucleotides long. *Genetics* **2012**, *192*, 973–985.

[156] Badis, G.; Berger, M. F.; Philippakis, A. A.; Talukder, S.; Gehrke, A. R.; Jaeger, S. A.; Chan, E. T.; Metzler, G.; Vedenko, A.; Chen, X.; others Diversity and complexity in DNA recognition by transcription factors. *Science* **2009**, *324*, 1720–1723.

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

[158] Bailly, C.; Kluza, J.; Martin, C.; Ellis, T.; Waring, M. J. DNase I footprinting of small molecule binding sites on DNA. *Oligonucleotide synthesis* **2005**, 319–342.

[159] Chodosh, L. A. Mobility Shift DNA-Binding Assay Using Gel Electrophoresis: DNA-Protein Interactions. *Current protocols in molecular biology* **1988**, *3*, 12–2.

[160] Hellman, L. M.; Fried, M. G. Electrophoretic mobility shift assay (EMSA) for detecting protein–nucleic acid interactions. *Nature protocols* **2007**, *2*, 1849–1861.

[161] O'Malley, R. C.; Huang, S.-s. C.; Song, L.; Lewsey, M. G.; Bartlett, A.; Nery, J. R.; Galli, M.; Gallavotti, A.; Ecker, J. R. Cistrome and epicistrome features shape the regulatory DNA landscape. *Cell* **2016**, *165*, 1280–1292.

[162] Berger, M. F.; Bulyk, M. L. Protein binding microarrays (PBMs) for rapid, high-throughput characterization of the sequence specificities of DNA binding proteins. *Gene Mapping, Discovery, and Expression: Methods and Protocols* **2006**, 245–260.

[163] Riley, T. R.; Slattery, M.; Abe, N.; Rastogi, C.; Liu, D.; Mann, R. S.; Bussemaker, H. J. SELEX-seq: a method for characterizing the complete repertoire of binding site preferences for transcription factor complexes. *Hox Genes: Methods and Protocols* **2014**, 255–278.

[164] Zykovich, A.; Korf, I.; Segal, D. J. Bind-n-Seq: high-throughput analysis of in vitro protein–DNA interactions using massively parallel sequencing. *Nucleic acids research* **2009**, *37*, e151–e151.

[165] Nutiu, R.; Friedman, R. C.; Luo, S.; Khrebtukova, I.; Silva, D.; Li, R.; Zhang, L.; Schroth, G. P.; Burge, C. B. Direct measurement of DNA affinity landscapes on a high-throughput sequencing instrument. *Nature biotechnology* **2011**, *29*, 659–664.

[166] Stormo, G. D.; Zuo, Z.; Chang, Y. K. Spec-seq: determining protein–DNA-binding specificity by sequencing. *Briefings in functional genomics* **2015**, *14*, 30–38.

[167] Maerkl, S. J.; Quake, S. R. A systems approach to measuring the binding energy landscapes of transcription factors. *Science* **2007**, *315*, 233–237.

[168] Christensen, R. G.; Gupta, A.; Zuo, Z.; Schriefer, L. A.; Wolfe, S. A.; Stormo, G. D. A modified bacterial one-hybrid system yields improved quantitative models of transcription factor specificity. *Nucleic acids research* **2011**, *39*, e83–e83.

[169] Meng, X.; Brodsky, M. H.; Wolfe, S. A. A bacterial one-hybrid system for determining the DNA-binding specificity of transcription factors. *Nature biotechnology* **2005**, *23*, 988–994.

[170] Robertson, G.; Hirst, M.; Bainbridge, M.; Bilenky, M.; Zhao, Y.; Zeng, T.; Euskirchen, G.; Bernier, B.; Varhol, R.; Delaney, A.; others Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing. *Nature methods* **2007**, *4*, 651–657.

[171] Johnson, D. S.; Mortazavi, A.; Myers, R. M.; Wold, B. Genome-wide mapping of in vivo protein-DNA interactions. *Science* **2007**, *316*, 1497–1502.

[172] Rhee, H. S.; Pugh, B. F. Comprehensive genome-wide protein-DNA interactions detected at single-nucleotide resolution. *Cell* **2011**, *147*, 1408–1419.

[173] Bailey, T. L.; Johnson, J.; Grant, C. E.; Noble, W. S. The MEME suite. *Nucleic acids research* **2015**, *43*, W39–W49.

[174] Hashim, F. A.; Mabrouk, M. S.; Al-Atabany, W. Review of different sequence motif finding algorithms. *Avicenna journal of medical biotechnology* **2019**, *11*, 130.

[175] Castellana, S.; Biagini, T.; Parca, L.; Petrizzelli, F.; Bianco, S. D.; Vescovi, A. L.; Carella, M.; Mazza, T. A comparative benchmark of classic DNA motif discovery tools on synthetic data. *Briefings in Bioinformatics* **2021**, *22*, bbab303.

[176] Castro-Mondragon, J. A.; Riudavets-Puig, R.; Rauluseviciute, I.; Berhanu Lemma, R.; Turchi, L.; Blanc-Mathieu, R.; Lucas, J.; Boddie, P.; Khan, A.; Manosalva Pérez, N.; others JASPAR 2022: the 9th release of the open-access database of transcription factor binding profiles. *Nucleic acids research* **2022**, *50*, D165–D173.

[177] Tian, F.; Yang, D.-C.; Meng, Y.-Q.; Jin, J.; Gao, G. PlantRegMap: charting functional regulatory maps in plants. *Nucleic acids research* **2020**, *48*, D1104–D1113.

[178] 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.; others Determination and inference of eukaryotic transcription factor sequence specificity. *Cell* **2014**, *158*, 1431–1443.

[179] Morgunova, E.; Taipale, J. Structural perspective of cooperative transcription factor binding. *Current opinion in structural biology* **2017**, *47*, 1–8.

[180] Smith, N. C.; Matthews, J. M. Mechanisms of DNA-binding specificity and functional gene regulation by transcription factors. *Current opinion in structural biology* **2016**, *38*, 68–74.

[181] Wunderlich, Z.; Mirny, L. A. Different gene regulation strategies revealed by analysis of binding motifs. *Trends in genetics* **2009**, *25*, 434–440.

[182] Zheng, N.; Fraenkel, E.; Pabo, C. O.; Pavletich, N. P. Structural basis of DNA recognition by the heterodimeric cell cycle transcription factor E2F–DP. *Genes & development* **1999**, *13*, 666–674.

[183] LaRonde-LeBlanc, N. A.; Wolberger, C. Structure of HoxA9 and Pbx1 bound to DNA: Hox hexapeptide and DNA recognition anterior to posterior. *Genes & development* **2003**, *17*, 2060–2072.

[184] Piper, D. E.; Batchelor, A. H.; Chang, C.-P.; Cleary, M. L.; Wolberger, C. Structure of a HoxB1–Pbx1 heterodimer bound to DNA: role of the hexapeptide and a fourth homeodomain helix in complex formation. *Cell* **1999**, *96*, 587–597.

[185] Merabet, S.; Saadaoui, M.; Sambrani, N.; Hudry, B.; Pradel, J.; Affolter, M.; Graba, Y. A unique Extradenticle recruitment mode in the Drosophila Hox protein Ultrabithorax. *Proceedings of the National Academy of Sciences* **2007**, *104*, 16946–16951.

[186] Merabet, S.; Mann, R. S. To be specific or not: the critical relationship between Hox and TALE proteins. *Trends in Genetics* **2016**, *32*, 334–347.

[187] Bobola, N.; Merabet, S. Homeodomain proteins in action: similar DNA binding preferences, highly variable connectivity. *Current opinion in genetics & development* **2017**, *43*, 1–8.

[188] Sánchez, M.; Jennings, P. A.; Murre, C. Conformational changes induced in Hoxb-8/Pbx-1 heterodimers in solution and upon interaction with specific DNA. *Molecular and cellular biology* **1997**, *17*, 5369–5376.

[189] Tahirov, T. H.; Inoue-Bungo, T.; Morii, H.; Fujikawa, A.; Sasaki, M.; Kimura, K.; Shiina, M.; Sato, K.; Kumasaka, T.; Yamamoto, M.; others Structural analyses of DNA recognition by the AML1/Runx-1 Runt domain and its allosteric control by CBFβ. *Cell* **2001**, *104*, 755–767.

[190] Tahirov, T. H.; Inoue-Bungo, T.; Sasaki, M.; Shiina, M.; Kimura, K.; Sato, K.; Kumasaka, T.; Yamamoto, M.; Kamiya, N.; Ogata, K. Crystallization and preliminary X-ray analyses of

quaternary, ternary and binary protein–DNA complexes with involvement of AML1/Runx-1/CBFα Runt domain, CBFβ and the C/EBPβ bZip region. *Acta Crystallographica Section D: Biological Crystallography* **2001**, *57*, 850–853.

[191] Pettersson, M.; Schaffner, W. Synergistic activation of transcription by multiple binding sites for NF-κB even in absence of co-operative factor binding to DNA. *Journal of molecular biology* **1990**, *214*, 373–380.

[192] Vashee, S.; Melcher, K.; Ding, W. V.; Johnston, S. A.; Kodadek, T. Evidence for two modes of cooperative DNA binding in vivo that do not involve direct protein–protein interactions. *Current biology* **1998**, *8*, 452–458.

[193] Jolma, A.; Yin, Y.; Nitta, K. R.; Dave, K.; Popov, A.; Taipale, M.; Enge, M.; Kivioja, T.; Morgunova, E.; Taipale, J. DNA-dependent formation of transcription factor pairs alters their binding specificity. *Nature* **2015**, *527*, 384–388.

[194] Adams, C. C.; Workman, J. L. Binding of disparate transcriptional activators to nucleosomal DNA is inherently cooperative. *Molecular and cellular biology* **1995**,

[195] Miller, J. A.; Widom, J. Collaborative competition mechanism for gene activation in vivo. *Molecular and cellular biology* **2003**, *23*, 1623–1632.

[196] Mirny, L. A. Nucleosome-mediated cooperativity between transcription factors. *Proceedings of the National Academy of Sciences* **2010**, *107*, 22534–22539.

[197] Kato, K.; Gális, I.; Suzuki, S.; Araki, S.; Demura, T.; Criqui, M.-C.; Potuschak, T.; Genschik, P.; Fukuda, H.; Matsuoka, K.; others Preferential up-regulation of G2/M phase-specific genes by overexpression of the hyperactive form of NtmybA2 lacking its negative regulation domain in tobacco BY-2 cells. *Plant physiology* **2009**, *149*, 1945–1957.

[198] Menges, M.; Murray, J. A. Synchronous Arabidopsis suspension cultures for analysis of cell-cycle gene activity. *The Plant Journal* **2002**, *30*, 203–212.

[199] Trolet, A.; Baldrich, P.; Criqui, M.-C.; Dubois, M.; Clavel, M.; Meyers, B. C.; Genschik, P. Cell cycle–dependent regulation and function of ARGONAUTE1 in plants. *The Plant Cell* **2019**, *31*, 1734–1750.

[200] Menges, M.; Hennig, L.; Gruissem, W.; Murray, J. A. Genome-wide gene expression in an Arabidopsis cell suspension. *Plant molecular biology* **2003**, *53*, 423–442.

[201] Donohue, L. K.; Guo, M. G.; Zhao, Y.; Jung, N.; Bussat, R. T.; Kim, D. S.; Neela, P. H.; Kellman, L. N.; Garcia, O. S.; Meyers, R. M.; others A cis-regulatory lexicon of DNA motif combinations mediating cell-type-specific gene regulation. *Cell genomics* **2022**, *2*.

[202] Günesdogan, U.; Surani, M. A. Developmental competence for primordial germ cell fate. *Current topics in developmental biology* **2016**, *117*, 471–496.

[203] Stormo, G. D. Modeling the specificity of protein-DNA interactions. *Quantitative biology* **2013**, *1*, 115–130.

[204] Boeva, V.; Clément, J.; Régnier, M.; Roytberg, M. A.; Makeev, V. J. Exact p-value calculation for heterotypic clusters of regulatory motifs and its application in computational annotation of cis-regulatory modules. *Algorithms for molecular biology* **2007**, *2*, 1–15.

[205] Dror, I.; Rohs, R.; Mandel-Gutfreund, Y. How motif environment influences transcription factor search dynamics: Finding a needle in a haystack. *BioEssays* **2016**, *38*, 605–612.

[206] Levo, M.; Zalckvar, E.; Sharon, E.; Machado, A. C. D.; Kalma, Y.; Lotam-Pompan, M.; Weinberger, A.; Yakhini, Z.; Rohs, R.; Segal, E. Unraveling determinants of transcription factor binding outside the core binding site. *Genome research* **2015**, *25*, 1018–1029.

[207] Sharon, E.; van Dijk, D.; Kalma, Y.; Keren, L.; Manor, O.; Yakhini, Z.; Segal, E. Probing the effect of promoters on noise in gene expression using thousands of designed sequences. *Genome research* **2014**, *24*, 1698–1706.

[208] Sharon, E.; Kalma, Y.; Sharp, A.; Raveh-Sadka, T.; Levo, M.; Zeevi, D.; Keren, L.; Yakhini, Z.; Weinberger, A.; Segal, E. Inferring gene regulatory logic from high-throughput measurements of thousands of systematically designed promoters. *Nature biotechnology* **2012**, *30*, 521–530.

[209] Smith, R. P.; Taher, L.; Patwardhan, R. P.; Kim, M. J.; Inoue, F.; Shendure, J.; Ovcharenko, I.; Ahituv, N. Massively parallel decoding of mammalian regulatory sequences supports a flexible organizational model. *Nature genetics* **2013**, *45*, 1021–1028.

[210] Manosalva Perez, N.; Ferrari, C.; Engelhorn, J.; Depuydt, T.; Nelissen, H.; Hartwig, T.; Vandepoele, K. MINI-AC: Inference of plant gene regulatory networks using bulk or single-cell accessible chromatin profiles. *bioRxiv* **2023**, 2023–05.

[211] Frith, M. C.; Hansen, U.; Weng, Z. Detection of cis-element clusters in higher eukaryotic DNA. *Bioinformatics* **2001**, *17*, 878–889.

[212] Cartharius, K.; Frech, K.; Grote, K.; Klocke, B.; Haltmeier, M.; Klingenhoff, A.; Frisch, M.; Bayerlein, M.; Werner, T. MatInspector and beyond: promoter analysis based on transcription factor binding sites. *Bioinformatics* **2005**, *21*, 2933–2942.

[213] Frith, M. C.; Li, M. C.; Weng, Z. Cluster-Buster: Finding dense clusters of motifs in DNA sequences. *Nucleic acids research* **2003**, *31*, 3666–3668.

[214] GuhaThakurta, D.; Stormo, G. D. Identifying target sites for cooperatively binding factors. *Bioinformatics* **2001**, *17*, 608–621.

[215] Xing, E. P.; Wu, W.; Jordan, M. I.; Karp, R. M. LOGOS: a modular Bayesian model for de novo motif detection. *Journal of Bioinformatics and Computational Biology* **2004**, *2*, 127–154.

[216] Aerts, S.; Van Loo, P.; Thijs, G.; Moreau, Y.; De Moor, B. Computational detection of cis-regulatory modules. *Bioinformatics* **2003**, *19*, ii5–ii14.

[217] Zhou, Q.; Wong, W. H. CisModule: de novo discovery of cis-regulatory modules by hierarchical mixture modeling. *Proceedings of the National Academy of Sciences* **2004**, *101*, 12114–12119.

[218] Kirchner, P.; Bourdenx, M.; Madrigal-Matute, J.; Tiano, S.; Diaz, A.; Bartholdy, B. A.; Will, B.; Cuervo, A. M. Proteome-wide analysis of chaperone-mediated autophagy targeting motifs. *PLoS biology* **2019**, *17*, e3000301.

[219] Wang, S.; Hu, H.; Li, X. A systematic study of motif pairs that may facilitate enhancer–promoter interactions. *Journal of integrative bioinformatics* **2022**, *19*.

[220] Jiang, P.; Singh, M. CCAT: combinatorial code analysis tool for transcriptional regulation. *Nucleic Acids Research* **2014**, *42*, 2833–2847.

[221] Vandel, J.; Cassan, O.; Lèbre, S.; Lecellier, C.-H.; Bréhélin, L. Probing transcription factor combinatorics in different promoter classes and in enhancers. *BMC genomics* **2019**, *20*, 1–19.

[222] Rich, C. Cell type-specific transcriptomic analyses of immunity in Arabidopsis thaliana roots. PhD thesis, Massachusetts Institute of Technology, Coventry, England, UK, 2018.

[223] Joshi, A.; Van Parys, T.; Van de Peer, Y.; Michoel, T. Characterizing regulatory path motifs in integrated networks using perturbational data. *Genome biology* **2010**, *11*, 1–14.

[224] Rich-Griffin, C.; Eichmann, R.; Reitz, M. U.; Hermann, S.; Woolley-Allen, K.; Brown, P. E.; Wiwatdirekkul, K.; Esteban, E.; Pasha, A.; Kogel, K.-H.; others Regulation of cell type-specific immunity networks in Arabidopsis roots. *Plant Cell* **2020**, *32*, 2742–2762.

[225] Habib, N.; Kaplan, T.; Margalit, H.; Friedman, N. A novel Bayesian DNA motif comparison method for clustering and retrieval. *PLoS computational biology* **2008**, *4*, e1000010.

[226] Gupta, S.; Stamatoyannopoulos, J. A.; Bailey, T. L.; Noble, W. S. Quantifying similarity between motifs. *Genome biology* **2007**, *8*, 1–9.

[227] Mahony, S.; Benos, P. V. STAMP: a web tool for exploring DNA-binding motif similarities. *Nucleic acids research* **2007**, *35*, W253–W258.

[228] Li, S.; Zheng, E. B.; Zhao, L.; Liu, S. Nonreciprocal and conditional cooperativity directs the pioneer activity of pluripotency transcription factors. *Cell Reports* **2019**, *28*, 2689–2703.

[229] Korkuć, P.; Schippers, J. H.; Walther, D. Characterization and identification of cis-regulatory elements in Arabidopsis based on single-nucleotide polymorphism information. *Plant Physiology* **2014**, *164*, 181–200.

[230] Lelli, K. M.; Slattery, M.; Mann, R. S. Disentangling the many layers of eukaryotic transcriptional regulation. *Annual review of genetics* **2012**, *46*, 43–68.

[231] Spitz, F.; Furlong, E. E. Transcription factors: from enhancer binding to developmental control. *Nature reviews genetics* **2012**, *13*, 613–626.

[232] Cassiday, L. A.; Maher III, L. J. Having it both ways: transcription factors that bind DNA and RNA. *Nucleic acids research* **2002**, *30*, 4118–4126.

[233] Xu, Y.; Huangyang, P.; Wang, Y.; Xue, L.; Devericks, E.; Nguyen, H. G.; Yu, X.; Oses-Prieto, J. A.; Burlingame, A. L.; Miglani, S.; others ERα is an RNA-binding protein sustaining tumor cell survival and drug resistance. *Cell* **2021**, *184*, 5215–5229.

[234] Theunissen, O.; Rudt, F.; Guddat, U.; Mentzel, H.; Pieler, T. RNA and DNA binding zinc fingers in Xenopus TFIIIA. *Cell* **1992**, *71*, 679–690.

[235] Sigova, A. A.; Abraham, B. J.; Ji, X.; Molinie, B.; Hannett, N. M.; Guo, Y. E.; Jangi, M.; Giallourakis, C. C.; Sharp, P. A.; Young, R. A. Transcription factor trapping by RNA in gene regulatory elements. *Science* **2015**, *350*, 978–981.

[236] Saldaña-Meyer, R.; Rodriguez-Hernaez, J.; Escobar, T.; Nishana, M.; Jácome-López, K.; Nora, E. P.; Bruneau, B. G.; Tsirigos, A.; Furlan-Magaril, M.; Skok, J.; others RNA interactions are essential for CTCF-mediated genome organization. *Molecular cell* **2019**, *76*, 412–422.

[237] Holmes, Z. E.; Hamilton, D. J.; Hwang, T.; Parsonnet, N. V.; Rinn, J. L.; Wuttke, D. S.; Batey, R. T. The Sox2 transcription factor binds RNA. *Nature communications* **2020**, *11*, 1805.

[238] Sega, P.; Kruszka, K.; Szewc, Ł.; Szweykowska-Kulińska, Z.; Pacak, A. Identification of transcription factors that bind to the 5 -UTR of the barley PHO2 gene. *Plant molecular biology* **2020**, *102*, 73–88.

[239] Peredo, E. L.; Cardon, Z. G. Shared up-regulation and contrasting down-regulation of gene expression distinguish desiccation-tolerant from intolerant green algae. *Proceedings of the National Academy of Sciences* **2020**, *117*, 17438–17445.

[240] Liu, J.; Feng, L.; Li, J.; He, Z. Genetic and epigenetic control of plant heat responses. *Frontiers in plant science* **2015**, *6*, 267.

[241] Hou, Y.; Yan, Y.; Cao, X. Epigenetic regulation of thermomorphogenesis in Arabidopsis thaliana. *Abiotech* **2022**, *3*, 12–24.

[242] Casal, J. J.; Balasubramanian, S. Thermomorphogenesis. *Annual review of plant biology* **2019**, *70*, 321–346.

[243] Chang, Y.-N.; Zhu, C.; Jiang, J.; Zhang, H.; Zhu, J.-K.; Duan, C.-G. Epigenetic regulation in plant abiotic stress responses. *Journal of integrative plant biology* **2020**, *62*, 563–580.

[244] He, K.; Cao, X.; Deng, X. Histone methylation in epigenetic regulation and temperature responses. *Current Opinion in Plant Biology* **2021**, *61*, 102001.

[245] Zhao, J.; Lu, Z.; Wang, L.; Jin, B. Plant responses to heat stress: physiology, transcription, noncoding RNAs, and epigenetics. *International journal of molecular sciences* **2020**, *22*, 117.

[246] Nishio, H.; Kawakatsu, T.; Yamaguchi, N. Beyond heat waves: Unlocking epigenetic heat stress memory in Arabidopsis. *Plant Physiology* **2023**, kiad558.

[247] Lämke, J.; Brzezinka, K.; Altmann, S.; Bäurle, I. A hit-and-run heat shock factor governs sustained histone methylation and transcriptional stress memory. *The EMBO journal* **2016**, *35*, 162–175.

[248] Crocker, J.; Noon, E. P.-B.; Stern, D. L. The soft touch: low-affinity transcription factor binding sites in development and evolution. *Current topics in developmental biology* **2016**, *117*, 455–469.

[249] Shahein, A.; López-Malo, M.; Istomin, I.; Olson, E. J.; Cheng, S.; Maerkl, S. J. Systematic analysis of low-affinity transcription factor binding site clusters in vitro and in vivo establishes their functional relevance. *Nature Communications* **2022**, *13*, 5273.

[250] de Boer, C. G. The continuum of transcription factor affinities. *Nature Reviews Genetics* **2024**, 1–1.

[251] Kribelbauer, J. F.; Rastogi, C.; Bussemaker, H. J.; Mann, R. S. Low-affinity binding sites and the transcription factor specificity paradox in eukaryotes. *Annual review of cell and developmental biology* **2019**, *35*, 357–379.

[252] Bosdriesz, E.; Wortel, M. T.; Haanstra, J. R.; Wagner, M. J.; De La Torre Cortes, P.; Teusink, B. Low affinity uniporter carrier proteins can increase net substrate uptake rate by reducing efflux. *Scientific reports* **2018**, *8*, 1–9.

[253] Tsai, A.; Muthusamy, A. K.; Alves, M. R.; Lavis, L. D.; Singer, R. H.; Stern, D. L.; Crocker, J. Nuclear microenvironments modulate transcription from low-affinity enhancers. *Elife* **2017**, *6*, e28975.

[254] Yamanouchi, U.; Yano, M.; Lin, H.; Ashikari, M.; Yamada, K. A rice spotted leaf gene, Spl7, encodes a heat stress transcription factor protein. *Proceedings of the National Academy of Sciences* **2002**, *99*, 7530–7535.

[255] Chao, L.-M.; Liu, Y.-Q.; Chen, D.-Y.; Xue, X.-Y.; Mao, Y.-B.; Chen, X.-Y. Arabidopsis transcription factors SPL1 and SPL12 confer plant thermotolerance at reproductive stage. *Molecular plant* **2017**, *10*, 735–748.

[256] Yu, C.-P.; Lin, J.-J.; Li, W.-H. $Positional distribution of transcription factor binding$ sites in Arabidopsis thaliana. *Scientific reports* **2016**, *6*, 25164.

[257] Matys, V.; Kel-Margoulis, O. V.; Fricke, E.; Liebich, I.; Land, S.; Barre-Dirrie, A.; Reuter, I.; Chekmenev, D.; Krull, M.; Hornischer, K.; others TRANSFAC® and its module TRANSCompel®: transcriptional gene regulation in eukaryotes. *Nucleic acids research* **2006**, *34*, D108–D110.

[258] Mathelier, A.; Zhao, X.; Zhang, A. W.; Parcy, F.; Worsley-Hunt, R.; Arenillas, D. J.; Buchman, S.; Chen, C.-y.; Chou, A.; Ienasescu, H.; others JASPAR 2014: an extensively expanded and updated open-access database of transcription factor binding profiles. *Nucleic acids research* **2014**, *42*, D142–D147.

[259] Bülow, L.; Steffens, N. O.; Galuschka, C.; Schindler, M.; Hehl, R. AthaMap: from in silico data to real transcription factor binding sites. *In silico biology* **2006**, *6*, 243–252.

[260] Lis, M.; Walther, D. The orientation of transcription factor binding site motifs in gene promoter regions: does it matter? *BMC genomics* **2016**, *17*, 1–21.

[261] Georgakopoulos-Soares, I.; Deng, C.; Agarwal, V.; Chan, C. S.; Zhao, J.; Inoue, F.; Ahituv, N. Transcription factor binding site orientation and order are major drivers of gene regulatory activity. *Nature communications* **2023**, *14*, 2333.

[262] Chen, C.-Y.; Gherzi, R.; Ong, S.-E.; Chan, E. L.; Raijmakers, R.; Pruijn, G. J.; Stoecklin, G.; Moroni, C.; Mann, M.; Karin, M. AU binding proteins recruit the exosome to degrade ARE-containing mRNAs. *Cell* **2001**, *107*, 451–464.

[263] Baltz, A. G.; Munschauer, M.; Schwanhäusser, B.; Vasile, A.; Murakawa, Y.; Schueler, M.; Youngs, N.; Penfold-Brown, D.; Drew, K.; Milek, M.; others The mRNA-bound proteome and its global occupancy profile on protein-coding transcripts. *Molecular cell* **2012**, *46*, 674–690.

[264] Barreau, C.; Paillard, L.; Osborne, H. B. AU-rich elements and associated factors: are there unifying principles? *Nucleic acids research* **2005**, *33*, 7138–7150.

[265] Lebedeva, S.; Jens, M.; Theil, K.; Schwanhäusser, B.; Selbach, M.; Landthaler, M.; Rajewsky, N. Transcriptome-wide analysis of regulatory interactions of the RNA-binding protein HuR. *Molecular cell* **2011**, *43*, 340–352.

[266] Besse, F.; de Quinto, S. L.; Marchand, V.; Trucco, A.; Ephrussi, A. Drosophila PTB promotes formation of high-order RNP particles and represses oskar translation. *Genes & development* **2009**, *23*, 195–207.

[267] Kristjánsdóttir, K.; Fogarty, E. A.; Grimson, A. Systematic analysis of the Hmga2 3 UTR identifies many independent regulatory sequences and a novel interaction between distal sites. *Rna* **2015**, *21*, 1346–1360.

[268] Pesole, G.; Mignone, F.; Gissi, C.; Grillo, G.; Licciulli, F.; Liuni, S. Structural and functional features of eukaryotic mRNA untranslated regions. *Gene* **2001**, *276*, 73–81.

[269] Yuan, M.; Jiang, Z.; Bi, G.; Nomura, K.; Liu, M.; Wang, Y.; Cai, B.; Zhou, J.-M.; He, S. Y.; Xin, X.-F. Pattern-recognition receptors are required for NLR-mediated plant immunity. *Nature* **2021**, *592*, 105–109.

[270] Bao, Z.; Hua, J. Linking the cell cycle with innate immunity in Arabidopsis. *Molecular plant* **2015**, *8*, 980–982.

[271] West, G.; Inzé, D.; Beemster, G. T. Cell cycle modulation in the response of the primary root of Arabidopsis to salt stress. *Plant physiology* **2004**, *135*, 1050–1058.

[272] Delay, C.; Imin, N.; Djordjevic, M. A. Regulation of Arabidopsis root development by small signaling peptides. *Frontiers in plant science* **2013**, *4*, 352.

[273] Scheres, B.; Benfey, P.; Dolan, L. Root development. *The Arabidopsis book/American Society of Plant Biologists* **2002**, *1*.

[274] Verbelen, J.-P.; Cnodder, T. D.; Le, J.; Vissenberg, K.; Baluška, F. The root apex of Arabidopsis thaliana consists of four distinct zones of growth activities: meristematic zone, transition zone, fast elongation zone and growth terminating zone. *Plant signaling & behavior* **2006**, *1*, 296–304.

[275] Hayashi, K.; Hasegawa, J.; Matsunaga, S. The boundary of the meristematic and elongation zones in roots: endoreduplication precedes rapid cell expansion. *Scientific reports* **2013**, *3*, 1–8.

[276] Brady, S. M.; Orlando, D. A.; Lee, J.-Y.; Wang, J. Y.; Koch, J.; Dinneny, J. R.; Mace, D.; Ohler, U.; Benfey, P. N. A high-resolution root spatiotemporal map reveals dominant expression patterns. *Science* **2007**, *318*, 801–806.

[277] Li, S.; Yamada, M.; Han, X.; Ohler, U.; Benfey, P. N. High-resolution expression map of the Arabidopsis root reveals alternative splicing and lincRNA regulation. *Developmental cell* **2016**, *39*, 508–522.

[278] Wang, G.; Kong, H.; Sun, Y.; Zhang, X.; Zhang, W.; Altman, N.; DePamphilis, C. W.; Ma, H. Genome-wide analysis of the cyclin family in Arabidopsis and comparative phylogenetic analysis of plant cyclin-like proteins. *Plant physiology* **2004**, *135*, 1084–1099.

[279] Beemster, G. T.; De Veylder, L.; Vercruysse, S.; West, G.; Rombaut, D.; Van Hummelen, P.; Galichet, A.; Gruissem, W.; Inzé, D.; Vuylsteke, M. Genome-wide analysis of gene expression profiles associated with cell cycle transitions in growing organs of Arabidopsis. *Plant physiology* **2005**, *138*, 734–743.

[280] Shahan, R.; Hsu, C.-W.; Nolan, T. M.; Cole, B. J.; Taylor, I. W.; Greenstreet, L.; Zhang, S.; Afanassiev, A.; Vlot, A. H. C.; Schiebinger, G.; others A single-cell Arabidopsis root atlas reveals developmental trajectories in wild-type and cell identity mutants. *Developmental cell* **2022**, *57*, 543–560.

[281] Satija, R.; Farrell, J. A.; Gennert, D.; Schier, A. F.; Regev, A. Spatial reconstruction of single-cell gene expression data. *Nature biotechnology* **2015**, *33*, 495–502.

[282] Lütge, A.; Zyprych-Walczak, J.; Kunzmann, U. B.; Crowell, H. L.; Calini, D.; Malhotra, D.; Soneson, C.; Robinson, M. D. CellMixS: quantifying and visualizing batch effects in single-cell RNA-seq data. *Life science alliance* **2021**, *4*.

[283] Denyer, T.; Ma, X.; Klesen, S.; Scacchi, E.; Nieselt, K.; Timmermans, M. C. Spatiotemporal developmental trajectories in the Arabidopsis root revealed using high-throughput single-cell RNA sequencing. *Developmental cell* **2019**, *48*, 840–852.

[284] Shahan, R.; Hsu, C.-W.; Nolan, T. M.; Cole, B. J.; Taylor, I. W.; Vlot, A. H. C.; Benfey, P. N.; Ohler, U. A single cell Arabidopsis root atlas reveals developmental trajectories in wild type and cell identity mutants. *bioRxiv* **2020**, *3*, 1147.

[285] Bhosale, R.; Boudolf, V.; Cuevas, F.; Lu, R.; Eekhout, T.; Hu, Z.; Van Isterdael, G.; Lambert, G. M.; Xu, F.; Nowack, M. K.; others A spatiotemporal DNA endoploidy map of the Arabidopsis root reveals roles for the endocycle in root development and stress adaptation. *The Plant Cell* **2018**, *30*, 2330–2351.

[286] Aibar, S.; González-Blas, C. B.; Moerman, T.; Huynh-Thu, V. A.; Imrichova, H.; Hulselmans, G.; Rambow, F.; Marine, J.-C.; Geurts, P.; Aerts, J.; others SCENIC: single-cell regulatory network inference and clustering. *Nature methods* **2017**, *14*, 1083–1086.

[287] DeTomaso, D.; Jones, M. G.; Subramaniam, M.; Ashuach, T.; Chun, J. Y.; Yosef, N. Functional interpretation of single cell similarity maps. *Nature communications* **2019**, *10*, 1–11.

[288] Mikolajewicz, N.; Gacesa, R.; Aguilera-Uribe, M.; Brown, K. R.; Moffat, J.; Han, H. Multilevel cellular and functional annotation of single-cell transcriptomes using scPipeline. *Communications Biology* **2022**, *5*, 1142.

[289] Barbie, D. A.; Tamayo, P.; Boehm, J. S.; Kim, S. Y.; Moody, S. E.; Dunn, I. F.; Schinzel, A. C.; Sandy, P.; Meylan, E.; Scholl, C.; others Systematic RNA interference reveals that oncogenic KRAS-driven cancers require TBK1. *Nature* **2009**, *462*, 108–112.

[290] Riechmann, J. L.; Ratcliffe, O. J. A genomic perspective on plant transcription factors. *Current opinion in plant biology* **2000**, *3*, 423–434.

[291] Qu, L.-J.; Zhu, Y.-X. Transcription factor families in Arabidopsis: major progress and outstanding issues for future research. *Current Opinion in Plant Biology* **2006**, *9*, 544–549.

[292] Mitsuda, N.; Ohme-Takagi, M. Functional analysis of transcription factors in Arabidopsis. *Plant and Cell Physiology* **2009**, *50*, 1232–1248.

[293] Favero, D. S.; Kawamura, A.; Shibata, M.; Takebayashi, A.; Jung, J.-H.; Suzuki, T.; Jaeger, K. E.; Ishida, T.; Iwase, A.; Wigge, P. A.; others AT-hook transcription factors restrict petiole growth by antagonizing PIFs. *Current Biology* **2020**, *30*, 1454–1466.

[294] Mizoi, J.; Shinozaki, K.; Yamaguchi-Shinozaki, K. AP2/ERF family transcription factors in plant abiotic stress responses. *Biochimica et Biophysica Acta (BBA)-Gene Regulatory Mechanisms* **2012**, *1819*, 86–96.

[295] Ariel, F. D.; Manavella, P. A.; Dezar, C. A.; Chan, R. L. The true story of the HD-Zip family. *Trends in plant science* **2007**, *12*, 419–426.

[296] Cao, Y.; Li, K.; Li, Y.; Zhao, X.; Wang, L. MYB transcription factors as regulators of secondary metabolism in plants. *Biology* **2020**, *9*, 61.

[297] Hofr, C.; Šultesová, P.; Zimmermann, M.; Mozgová, I.; Procházková Schrumpfová, P.; Wimmerová, M.; Fajkus, J. Single-Myb-histone proteins from Arabidopsis thaliana: a quantitative study of telomere-binding specificity and kinetics. *Biochemical Journal* **2009**, *419*, 221–230.

[298] Safi, A.; Medici, A.; Szponarski, W.; Ruffel, S.; Lacombe, B.; Krouk, G. The world according to GARP transcription factors. *Current opinion in plant biology* **2017**, *39*, 159–167.

[299] Grierson, C.; Nielsen, E.; Ketelaarc, T.; Schiefelbein, J. Root hairs. *The Arabidopsis Book/American Society of Plant Biologists* **2014**, *12*.

[300] Gachon, C. M.; Langlois-Meurinne, M.; Henry, Y.; Saindrenan, P. Transcriptional co-regulation of secondary metabolism enzymes in Arabidopsis: functional and evolutionary implications. *Plant molecular biology* **2005**, *58*, 229–245.

[301] Persson, S.; Wei, H.; Milne, J.; Page, G. P.; Somerville, C. R. Identification of genes required for cellulose synthesis by regression analysis of public microarray data sets. *Proceedings of the National Academy of Sciences* **2005**, *102*, 8633–8638.

[302] Manfield, I. W.; Jen, C.-H.; Pinney, J. W.; Michalopoulos, I.; Bradford, J. R.; Gilmartin, P. M.; Westhead, D. R. Arabidopsis Co-expression Tool (ACT): web server tools for microarray-based gene expression analysis. *Nucleic acids research* **2006**, *34*, W504–W509.

[303] Skirycz, A.; Radziejwoski, A.; Busch, W.; Hannah, M. A.; Czeszejko, J.; Kwaśniewski, M.; Zanor, M.-I.; Lohmann, J. U.; De Veylder, L.; Witt, I.; others The DOF transcription factor OBP1 is involved in cell cycle regulation in Arabidopsis thaliana. *The Plant Journal* **2008**, *56*, 779–792.

[304] Attwooll, C.; Denchi, E. L.; Helin, K. The E2F family: specific functions and overlapping interests. *The EMBO journal* **2004**, *23*, 4709–4716.

[305] Dyson, N. The regulation of E2F by pRB-family proteins. *Genes & development* **1998**, *12*, 2245–2262.

[306] Trimarchi, J. M.; Lees, J. A. Sibling rivalry in the E2F family. *Nature reviews Molecular cell biology* **2002**, *3*, 11–20.

[307] De Veylder, L.; Joubès, J.; Inzé, D. Plant cell cycle transitions. *Current opinion in plant biology* **2003**, *6*, 536–543.

[308] Birnbaum, K.; Shasha, D. E.; Wang, J. Y.; Jung, J. W.; Lambert, G. M.; Galbraith, D. W.; Benfey, P. N. A gene expression map of the Arabidopsis root. *Science* **2003**, *302*, 1956–1960.

[309] Balomenos, D.; Martínez-A, C. Cell-cycle regulation in immunity, tolerance and autoimmunity. *Immunology today* **2000**, *21*, 551–555.

[310] van den Heuvel, S. Cell-cycle regulation. *WormBook: The Online Review of C. elegans Biology [Internet]* **2005**,

[311] Elledge, S. J. Cell cycle checkpoints: preventing an identity crisis. *Science* **1996**, *274*, 1664–1672.

[312] Fox, S.; Southam, P.; Pantin, F.; Kennaway, R.; Robinson, S.; Castorina, G.; Sánchez-Corrales, Y. E.; Sablowski, R.; Chan, J.; Grieneisen, V.; others Spatiotemporal coordination of cell division and growth during organ morphogenesis. *PLoS Biology* **2018**, *16*, e2005952.

[313] Sozzani, R.; Cui, H.; Moreno-Risueno, M.; Busch, W.; Van Norman, J.; Vernoux, T.; Brady, S.; Dewitte, W.; Murray, J. A. H.; Benfey, P. Spatiotemporal regulation of cell-cycle genes by SHORTROOT links patterning and growth. *Nature* **2010**, *466*, 128–132.

[314] Roeder, A. H.; Cunha, A.; Ohno, C. K.; Meyerowitz, E. M. Cell cycle regulates cell type in the Arabidopsis sepal. *Development* **2012**, *139*, 4416–4427.

[315] Siersbæk, R.; Rabiee, A.; Nielsen, R.; Sidoli, S.; Traynor, S.; Loft, A.; Poulsen, L. L. C.; Rogowska-Wrzesinska, A.; Jensen, O. N.; Mandrup, S. Transcription factor cooperativity in early adipogenic hotspots and super-enhancers. *Cell reports* **2014**, *7*, 1443–1455.

[316] Arnold, C. D.; Gerlach, D.; Stelzer, C.; Boryń, Ł. M.; Rath, M.; Stark, A. Genome-wide quantitative enhancer activity maps identified by STARR-seq. *Science* **2013**, *339*, 1074–1077.

6 Supplementary Material

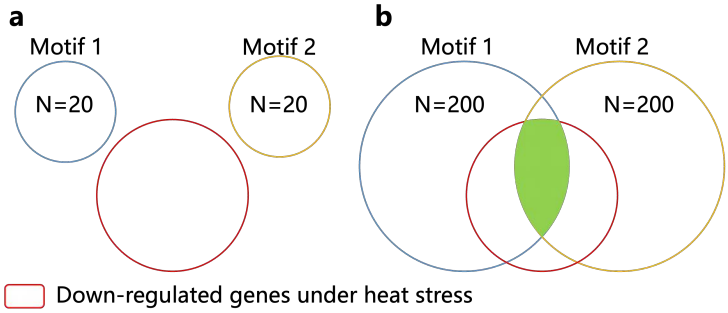


Figure S1: Models of how weaker promoters can cause an increase of motif pairs. Blue and yellow circles represent sets of promoters bound by motifs respectively, with the overlap indicating shared promoters. (a) When N is small, no promoters are shared between *motif 1* and *motif 2*, then there is surely no down-regulated genes included. (b) When N is bigger, the green shaded area highlights promoters where two motifs co-occur significantly, suggesting potential regulatory interactions between the corresponding TFs.

Table S1: Gene model with multiple fragments of *A. thaliana* genes

chromosome	element	start	stop	gene model
1	CDS	3760	3913	AT1G01010.1
1	CDS	3996	4276	AT1G01010.1
1	CDS	4486	4605	AT1G01010.1
1	CDS	4706	5095	AT1G01010.1
1	CDS	5174	5326	AT1G01010.1

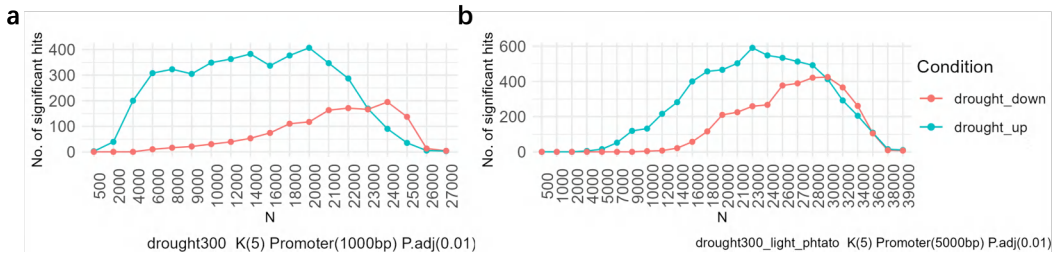


Figure S2: Effect on number of motif pairs under changing parameter N for up- and down-regulated genes under drought stress of three plant species. (a) depicts numbers of motif pairs under changing parameter N for up- (red) and down-regulated (green) genes in *A. thaliana* and *Solanum tuberosum* under drought stress. All motif pairs identified through PMET analysis are filtered with a p-value threshold of < 0.01 for statistical significance.

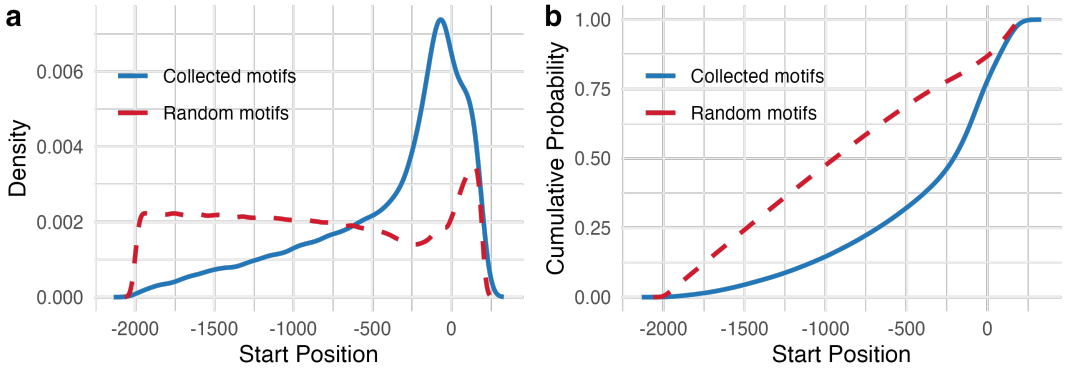


Figure S3: Positional distribution and cumulative probabilities of motifs in the promoter region of *A. thaliana*. (a) The probability density function (blue line) reflects the localized abundance of predicted motifs within the -200 to +2,000 bp region upstream of the TSS, with the peak indicating a higher prevalence of motifs proximal (50 bp) to the TSS. The dashed red line depicts the uniform distribution of random motifs for comparison, illustrating the non-random nature of the random motif distribution. (b) Solid blue line indicates the cumulative probability of the experimentally verified motifs, offering a visual representation of motif accumulation along the promoter region.

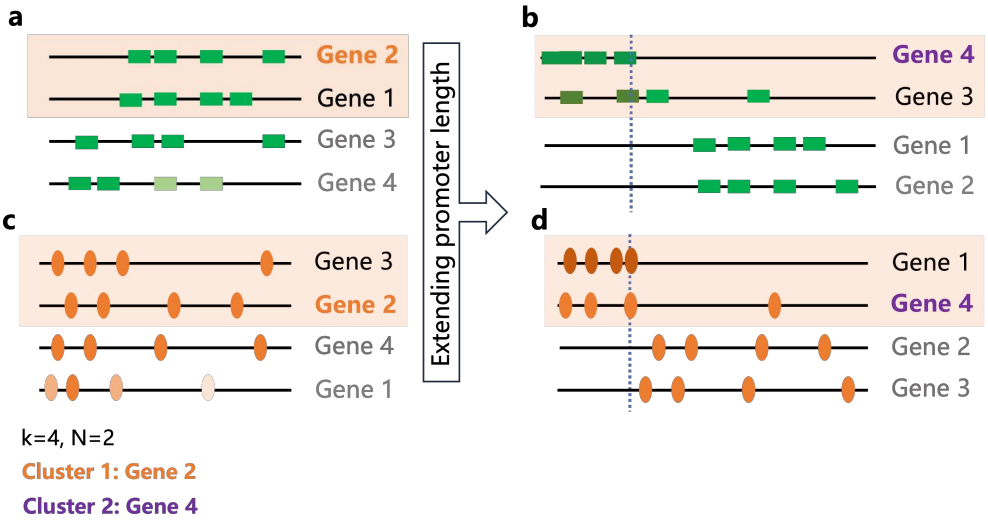


Figure S4: Motif hit distribution on gene promoters before and after promoter extension. (a, c) Initial promoter regions of four genes (*Gene 1*, *Gene 2*, *Gene 3*, and *Gene 4*), with two motifs (represented by green and orange blocks) matched to each promoter. (b, d) Effect of promoter extension: (a and c) *Gene 2*, belonging to a specific gene cluster, initially pairs with two motifs; (b) Extension reveals additional green motif affinity in new genes; it is hypothesized that *Gene 3* is a random gene without cluster specificity, while *Gene 4* is a part of a cluster, other than the cluster of *Gene 2*. (d) Post-extension, newly included genes show increased affinity for the orange motif.

Figure S5: Word cloud of motif pairs and individual motifs involved in the highest number of motif pairs associated with down-regulated genes in *A. thaliana*, and *Solanum tuberosum* under drought stress. (a) displays the frequency of TF family pairs from motif pairs bound to the down-regulated *A. thaliana* genes under drought stress, with a notable occurrence of *AHL-MB* at N=24,000. (b) Illustrates the high frequency of *AHL-MYB* motif pairs associated with drought-stressed, down-regulated genes in *Solanum tuberosum* at N=28,000. (c)-(d) show motifs derived from the motif pairs with the highest counts for the corresponding species: *ICU4* for *A. thaliana*; *AHL* family members for *Solanum tuberosum*.

Table S2: GO analyis of up-regulated 300 heat stress genes

GO Description	Gene Count	p.adj
response to heat	62	7.39e-50
cellular response to hypoxia	35	1.81e-26
cellular response to decreased oxygen levels	35	1.81e-26
cellular response to oxygen levels	35	1.81e-26
response to hypoxia	35	1.92e-21
response to decreased oxygen levels	35	3.96e-21
response to oxygen levels	35	3.96e-21
protein folding	26	3.60e-19
protein maturation	26	1.25e-14
response to reactive oxygen species	20	2.27e-13
cellular response to abscisic acid stimulus	19	2.36e-07
cellular response to alcohol	19	2.36e-07
abscisic acid-activated signaling pathway	18	4.11e-08
regulation of post-embryonic development	18	2.21e-04
cellular response to heat	17	1.25e-14
heat acclimation	14	2.15e-12
response to hydrogen peroxide	14	5.38e-12
regulation of signal transduction	14	8.91e-03
regulation of signaling	14	1.01e-02
regulation of cell communication	14	1.12e-02

Table S3: GO analyis of down-regulated 300 heat stress genes

GO Description	Gene Count	p.adj
glycosyl compound metabolic process	16	1.61e-04
S-glycoside metabolic process	15	1.61e-04
glycosinolate metabolic process	15	1.61e-04
glucosinolate metabolic process	15	1.61e-04
secondary metabolite biosynthetic process	15	4.08e-03
response to red or far red light	15	5.37e-03
cell wall biogenesis	15	1.48e-02
regulation of hormone levels	11	2.34e-02
cell wall organization	10	7.47e-02
meiotic cell cycle	10	8.23e-02
external encapsulating structure organization	10	1.34e-01
carbohydrate biosynthetic process	10	1.72e-01
sexual reproduction	10	3.18e-01
hormone metabolic process	9	4.28e-03
cell wall polysaccharide metabolic process	9	9.40e-03
cell wall macromolecule metabolic process	9	2.45e-02
glucan metabolic process	9	3.61e-02
sulfur compound biosynthetic process	9	1.17e-01
anatomical structure formation involved in morphogenesis	9	1.21e-01
response to light intensity	9	2.40e-01

Table S4: GO analyis of up-regulated 300 drought stress genes

GO Description	Gene Count	p.adj
cellular response to abscisic acid stimulus	22	1.52e-09
cellular response to alcohol	22	1.52e-09
flavonoid metabolic process	20	3.20e-11
flavonoid biosynthetic process	19	2.34e-14
abscisic acid-activated signaling pathway	19	7.54e-09
pigment metabolic process	18	3.66e-05
pigment biosynthetic process	14	1.50e-05
secondary metabolite biosynthetic process	13	1.01e-02
response to heat	13	1.20e-02
anthocyanin-containing compound metabolic process	12	1.07e-09
hexosyltransferase activity	12	3.59e-02
response to extracellular stimulus	11	3.07e-02
cellular response to abiotic stimulus	10	2.58e-02
cellular response to environmental stimulus	10	2.58e-02
response to nutrient levels	10	2.73e-02
UDP-glycosyltransferase activity	10	3.59e-02
anthocyanin-containing compound biosynthetic process	9	2.39e-08
regulation of flavonoid biosynthetic process	9	3.40e-08
regulation of abscisic acid-activated signaling pathway	8	4.40e-03
regulation of response to alcohol	8	4.40e-03

Table S5: GO analyis of down-regulated 300 drought stress genes

GO Description	Gene Count	p.adj
response to auxin	26	5.74e-09
response to light intensity	20	4.86e-06
unidimensional cell growth	20	8.87e-05
cell wall biogenesis	20	1.04e-04
regulation of hormone levels	18	4.86e-06
hydrolase activity, acting on glycosyl bonds	18	4.17e-05
hydrolase activity, hydrolyzing O-glycosyl compounds	17	3.25e-05
glycosyl compound metabolic process	14	9.70e-04
secretory vesicle	14	4.11e-07
apoplast	14	4.98e-04
S-glycoside metabolic process	13	9.70e-04
glycosinolate metabolic process	13	9.70e-04
glucosinolate metabolic process	13	9.70e-04
response to hypoxia	13	3.57e-03
response to decreased oxygen levels	13	4.30e-03
response to oxygen levels	13	4.30e-03
secondary metabolite biosynthetic process	13	9.53e-03
response to red or far red light	13	2.07e-02
cellular response to hypoxia	12	9.70e-04
cellular response to decreased oxygen levels	12	9.70e-04

Table S6: GO analyis of up-regulated 300 salt stress genes

GO Description	Gene Count	p.adj
plant organ senescence	16	1.48e-04
response to heat	16	2.15e-04
leaf senescence	14	2.54e-04
cellular response to abscisic acid stimulus	13	7.46e-04
cellular response to alcohol	13	7.46e-04
response to hypoxia	13	8.27e-04
response to decreased oxygen levels	13	9.28e-04
response to oxygen levels	13	9.28e-04
cellular response to hypoxia	12	2.54e-04
cellular response to decreased oxygen levels	12	2.54e-04
cellular response to oxygen levels	12	2.54e-04
abscisic acid-activated signaling pathway	12	4.29e-04
regulation of signal transduction	12	1.60e-02
regulation of signaling	12	1.95e-02
regulation of cell communication	12	2.11e-02
response to reactive oxygen species	10	2.71e-04
multicellular organismal reproductive process	10	3.88e-03
multicellular organism reproduction	10	4.39e-03
negative regulation of nucleobase-containing compound metabolic process	10	1.01e-02
seed maturation	9	1.48e-04

Table S7: GO analyis of down-regulated 300 salt stress genes

GO Description	Gene Count	p.adj
monoatomic ion transport	21	1.40e-07
oxidoreductase activity, acting on paired donors, with incorporation or reduction of molecular oxygen	18	7.88e-05
monoatomic ion transmembrane transport	15	3.85e-06
tetrapyrrole binding	14	1.89e-03
monoatomic ion transmembrane transporter activity	14	2.55e-03
monoatomic cation transport	13	1.38e-03
salt transmembrane transporter activity	13	8.48e-04
heme binding	13	2.49e-03
monoatomic cation transmembrane transport	12	2.68e-04
inorganic ion transmembrane transport	12	1.38e-03
response to red or far red light	12	2.93e-02
monooxygenase activity	12	6.79e-04
monoatomic cation transmembrane transporter activity	12	9.82e-03
inorganic cation transmembrane transport	11	1.38e-03
metal ion transport	11	1.75e-03
metal ion transmembrane transporter activity	11	1.61e-03
inorganic cation transmembrane transporter activity	11	1.91e-02
monoatomic ion homeostasis	10	1.70e-02
regulation of hormone levels	10	2.87e-02
cell wall organization	10	3.09e-02

gene	p_val	avg_log2FC	pct.1	pct.2	p_val_adj	cluster	gene_symbol	highlight	core
AT3G23830	1.91E-164	2.79	0.972	0.362	5.46E-160	G1	RBG4	Reference	Core
AT5G49560	2.67E-77	2.71	0.497	0.093	7.67E-73	G1		FALSE	FALSE
AT3G15357	8.18E-96	2.56	0.68	0.178	2.34E-91	G1		Unsigned Reference	FALSE
AT4G25630	1.55E-149	2.53	0.957	0.337	4.43E-145	G1	MED36A	Reference	FALSE
AT2G30860	1.14E-36	2.40	0.614	0.293	3.27E-32	G1	GSTF9	FALSE	FALSE
AT3G55510	2.19E-81	2.38	0.536	0.096	6.28E-77	G1	RBL	Reference	FALSE
AT1G76405	3.67E-88	2.38	0.609	0.126	1.05E-83	G1	OEP21B	Reference	Core
AT2G03780	2.02E-73	2.35	0.497	0.092	5.79E-69	G1		Reference	FALSE
AT3G12860	1.00E-76	2.25	0.477	0.077	2.87E-72	G1		Unsigned Reference	FALSE
AT4G12600	2.77E-115	2.24	0.931	0.479	7.94E-111	G1		FALSE	FALSE
AT3G15240	7.69E-81	2.21	0.607	0.141	2.20E-76	G1		Unsigned Reference	FALSE
AT1G15250	2.07E-113	2.20	0.81	0.25	5.94E-109	G1	RPL37A	Unsigned Reference	FALSE
AT2G12646	8.00E-94	2.18	0.652	0.139	2.29E-89	G1		FALSE	FALSE
AT1G53542	2.80E-48	2.07	0.444	0.123	8.01E-44	G1		FALSE	FALSE
AT5G06210	1.46E-65	2.06	0.561	0.158	4.18E-61	G1	S-RBP11	FALSE	FALSE
AT1G23100	1.38E-66	2.05	0.652	0.233	3.96E-62	G1		Unsigned Reference	FALSE
AT2G27840	5.13E-78	2.04	0.736	0.272	1.47E-73	G1	HDT4	Unsigned Reference	FALSE
AT3G57490	5.80E-81	2.04	0.744	0.269	1.66E-76	G1	RPS2D	FALSE	FALSE
AT3G22660	2.61E-90	2.02	0.787	0.299	7.49E-86	G1	EBP2	Unsigned Reference	FALSE
AT1G50110	4.77E-59	2.01	0.454	0.096	1.37E-54	G1	BCAT6	Reference	Core
AT3G13160	2.31E-64	2.00	0.576	0.164	6.62E-60	G1		Unsigned Reference	FALSE
AT2G34260	1.36E-70	2.00	0.602	0.168	3.89E-66	G1	WDR55	Unsigned Reference	FALSE
AT1G52930	5.04E-99	1.95	0.83	0.295	1.44E-94	G1	BRIX1-2	Unsigned Reference	FALSE
AT2G23040	6.12E-50	1.95	0.452	0.12	1.75E-45	G1		FALSE	FALSE
AT3G44750	1.55E-105	1.95	0.843	0.258	4.45E-101	G1	HDT1	Unsigned Reference	FALSE
AT4G15248	5.30E-41	1.94	0.424	0.124	1.52E-36	G1	MIP1A	FALSE	FALSE
AT1G74560	1.58E-89	1.94	0.843	0.367	4.53E-85	G1	NRP1	Unsigned Reference	FALSE
AT3G10050	2.68E-64	1.94	0.49	0.104	7.67E-60	G1	OMR1	Unsigned Reference	FALSE
AT2G19385	2.56E-60	1.92	0.614	0.21	7.34E-56	G1		Unsigned Reference	FALSE
AT2G20490	6.46E-116	1.91	0.964	0.536	1.85E-111	G1	NOP10	Unsigned Reference	FALSE
AT1G62480	7.80E-17	-4.17	0.561	0.624	2.24E-12	G1		FALSE	FALSE
AT1G72150	2.94E-21	-4.08	0.221	0.436	8.42E-17	G1	PATL1	FALSE	FALSE
AT4G26320	1.95E-12	-3.87	0.325	0.432	5.58E-08	G1	AGP13	FALSE	FALSE
AT5G53250	6.97E-25	-3.74	0.36	0.548	2.00E-20	G1	AGP22	FALSE	FALSE
AT1G11260	1.44E-34	-3.54	0.239	0.527	4.13E-30	G1	STP1	FALSE	FALSE
AT1G52070	8.01E-17	-3.38	0.284	0.451	2.30E-12	G1	JAL10	Unsigned Reference	FALSE
AT2G47270	2.62E-21	-3.37	0.19	0.404	7.52E-17	G1	UPB1	FALSE	FALSE
AT5G10400	5.83E-05	-3.36	0.424	0.414	1.00E+00	G1	HTR2	FALSE	FALSE
AT5G56540	4.24E-25	-3.23	0.652	0.718	1.21E-20	G1	AGP14	FALSE	FALSE
AT2G27970	2.85E-21	-3.23	0.272	0.467	8.18E-17	G1	CKS2	Unsigned Reference	FALSE
AT3G53650	5.78E-10	-3.23	0.421	0.482	1.66E-05	G1		FALSE	FALSE
AT1G50490	1.64E-04	-3.21	0.482	0.467	1.00E+00	G1	UBC20	FALSE	Core
AT1G19835	4.10E-24	-3.19	0.175	0.413	1.17E-19	G1	FPP4	FALSE	FALSE
AT3G13520	3.47E-26	-3.11	0.688	0.748	9.95E-22	G1	AGP12	FALSE	FALSE
AT5G59870	2.57E-16	-3.08	0.317	0.46	7.37E-12	G1	HTA6	FALSE	FALSE
AT3G16240	1.97E-24	-2.98	0.497	0.633	5.65E-20	G1	TIP2-1	FALSE	FALSE
AT3G15540	3.83E-18	-2.93	0.228	0.416	1.10E-13	G1	IAA19	FALSE	FALSE
AT5G59690	2.18E-14	-2.83	0.467	0.54	6.26E-10	G1		FALSE	FALSE
AT5G22880	7.29E-16	-2.79	0.261	0.418	2.09E-11	G1	H2B	FALSE	FALSE
AT3G04320	5.41E-19	-2.78	0.244	0.436	1.55E-14	G1		FALSE	FALSE
AT3G52920	3.49E-20	-2.77	0.305	0.485	1.00E-15	G1		FALSE	FALSE
AT2G30930	8.91E-15	-2.76	0.457	0.556	2.55E-10	G1		FALSE	FALSE
AT5G08130	2.74E-25	-2.72	0.165	0.418	7.87E-21	G1	BIM1	FALSE	FALSE
AT5G65360	7.80E-06	-2.72	0.726	0.627	2.24E-01	G1	HTR2	FALSE	FALSE
AT3G27360	6.34E-14	-2.71	0.543	0.577	1.82E-09	G1	HTR2	FALSE	Core
AT3G12110	7.04E-13	-2.66	0.371	0.48	2.02E-08	G1	ACT11	FALSE	FALSE
AT3G48340	5.12E-21	-2.60	0.302	0.499	1.47E-16	G1	CEP2	FALSE	FALSE
AT3G58120	7.04E-19	-2.52	0.272	0.446	2.02E-14	G1	BZIP61	FALSE	FALSE
AT5G54640	3.10E-17	-2.50	0.431	0.537	8.90E-13	G1	RAT5	FALSE	FALSE
AT2G28740	1.06E-09	-2.49	0.561	0.569	3.03E-05	G1	HIS4	FALSE	FALSE

Figure S6: The analyses identified G1 phase markers genes exhibiting a twofold increase in expression ($avg\_log2FC > 1$) and presence in at least 40% ($pct.1 > 0.40$) of the cells in the target phase. These genes were primarily ranked by fold change. In cases of ties, genes were further ranked by ascending expression in non-target phases ($pct.2$) and descending expression in the target phase ($pct.1$). The top 30 genes were selected as potential positive markers. Negative markers were determined by negative avg_log2FC ($avg\_log2FC > 1$) and other similar selection criteria to those used for positive markers. The positive markers marked in red indicate that they originate from the list of 1080 cycle related genes proposed by Menges et al. (2003)[316], while those marked in green correspond to the negative markers. The markers labeled with the term "Core" in yellow are included in the core cell cycle genes summarized by Beemster et al. (2005)[279].

gene	p_val	avg_log2FC	pct.1	pct.2	p_val_adj	cluster	gene_symbol	highlight	core
AT5G13060	7.90E-190	6.03	0.62	0.02	2.26E-185	S	ABAP1	FALSE	FALSE
AT5G10400	8.15E-241	5.87	0.98	0.24	2.34E-236	S	HTR2	FALSE	FALSE
AT3G14740	2.93E-183	5.52	0.64	0.03	8.39E-179	S		FALSE	FALSE
AT5G10390	9.74E-229	5.51	0.88	0.12	2.79E-224	S	HTR2	Reference	Core
AT1G04020	3.25E-241	5.41	0.78	0.03	9.33E-237	S	ATBARD1	FALSE	FALSE
AT1G47210	3.79E-238	5.30	0.87	0.09	1.09E-233	S	CYCA3-2	FALSE	FALSE
AT5G43080	9.92E-171	5.12	0.59	0.02	2.84E-166	S	CYCA3-1	FALSE	FALSE
AT3G27640	5.58E-138	5.07	0.49	0.02	1.60E-133	S		FALSE	FALSE
AT5G23420	1.91E-247	4.99	0.89	0.08	5.48E-243	S	HMGB7	Reference	Core
AT3G53650	1.32E-173	4.94	0.92	0.33	3.79E-169	S		FALSE	FALSE
AT3G46320	1.45E-240	4.91	0.96	0.18	4.15E-236	S		FALSE	FALSE
AT3G25100	9.97E-187	4.89	0.64	0.02	2.86E-182	S	CDC45	FALSE	FALSE
AT5G59870	3.38E-222	4.81	0.97	0.26	9.69E-218	S	HTA6	FALSE	FALSE
AT4G26960	4.50E-200	4.78	0.75	0.06	1.29E-195	S		FALSE	FALSE
AT3G45930	1.30E-223	4.70	0.95	0.20	3.73E-219	S		FALSE	FALSE
AT5G61000	5.00E-197	4.69	0.73	0.05	1.43E-192	S	RPA1D	FALSE	FALSE
AT5G65360	9.04E-209	4.62	1.00	0.54	2.59E-204	S	HTR2	FALSE	FALSE
AT2G28720	1.75E-150	4.56	0.60	0.05	5.02E-146	S		FALSE	FALSE
AT4G39380	2.75E-138	4.55	0.51	0.02	7.89E-134	S		FALSE	FALSE
AT4G00020	1.15E-184	4.51	0.65	0.03	3.31E-180	S	MEE43	FALSE	FALSE
AT2G42260	1.08E-137	4.45	0.52	0.03	3.10E-133	S	PYM	FALSE	FALSE
AT5G59690	1.02E-217	4.42	1.00	0.38	2.91E-213	S		FALSE	FALSE
AT3G27360	3.16E-190	4.32	0.97	0.45	9.05E-186	S	HTR2	Reference	Core
AT3G02820	6.17E-207	4.29	0.82	0.09	1.77E-202	S		Reference	Core
AT5G08020	8.04E-160	4.27	0.60	0.04	2.30E-155	S	RPA1B	FALSE	FALSE
AT4G21070	3.39E-175	4.25	0.62	0.03	9.73E-171	S	BRCA1	FALSE	FALSE
AT2G28740	4.52E-191	4.19	0.98	0.44	1.29E-186	S	HIS4	FALSE	FALSE
AT5G63550	4.22E-226	4.06	0.95	0.18	1.21E-221	S		FALSE	FALSE
AT4G27230	1.57E-196	3.99	0.89	0.17	4.49E-192	S	HTA2	FALSE	FALSE
AT3G27060	1.94E-230	3.99	0.96	0.17	5.56E-226	S	TSO2	FALSE	FALSE
AT1G50490	4.88E-21	-4.12	0.34	0.51	1.40E-16	S	UBC20	FALSE	Core
AT1G02690	7.04E-24	-3.32	0.22	0.45	2.02E-19	S	IMPA-6	FALSE	Core
AT3G60900	1.56E-27	-2.32	0.18	0.46	4.48E-23	S	FLA10	Unsigned Reference	FALSE
AT3G57150	2.41E-17	-1.96	0.33	0.49	6.90E-13	S	CBF5	FALSE	FALSE
AT3G23830	3.80E-16	-1.89	0.40	0.53	1.09E-11	S	RBG4	FALSE	Core
AT5G05080	2.21E-08	-1.87	0.68	0.65	6.33E-04	S	UBC22	FALSE	FALSE
AT4G25630	1.48E-18	-1.78	0.34	0.52	4.25E-14	S	MED36A	FALSE	FALSE
AT2G18230	3.53E-14	-1.71	0.29	0.46	1.01E-09	S	PPA2	Unsigned Reference	FALSE
AT4G36180	4.51E-12	-1.66	0.41	0.53	1.29E-07	S		FALSE	FALSE
AT3G44750	1.84E-10	-1.65	0.29	0.42	5.27E-06	S	HDT1	Unsigned Reference	FALSE
AT1G48920	3.02E-29	-1.62	0.61	0.73	8.64E-25	S	NUCL1	Unsigned Reference	FALSE
AT1G10760	2.28E-07	-1.55	0.31	0.41	6.52E-03	S	GWD1	Unsigned Reference	FALSE
AT1G56110	5.14E-13	-1.41	0.50	0.57	1.47E-08	S	NOP56	Unsigned Reference	FALSE
AT3G22310	8.33E-08	-1.40	0.32	0.41	2.39E-03	S	RH9	Unsigned Reference	FALSE
AT5G27120	8.40E-11	-1.39	0.43	0.51	2.41E-06	S	NOP5-1	FALSE	FALSE
AT5G14640	2.69E-09	-1.37	0.60	0.60	7.71E-05	S	ASK5	Unsigned Reference	FALSE
AT5G43700	3.68E-05	-1.33	0.38	0.43	1.00E+00	S	IAA4	FALSE	FALSE
AT1G02810	9.61E-18	-1.33	0.48	0.66	2.75E-13	S	PME7	FALSE	FALSE
AT3G13080	1.18E-09	-1.30	0.34	0.47	3.37E-05	S	ABCC3	FALSE	FALSE
AT4G25730	9.09E-09	-1.29	0.37	0.46	2.61E-04	S		Unsigned Reference	FALSE
AT2G43480	1.13E-08	-1.29	0.35	0.46	3.23E-04	S	PER26	Unsigned Reference	FALSE
AT2G28900	3.83E-05	-1.29	0.36	0.42	1.00E+00	S	OEP161	Unsigned Reference	FALSE
AT2G41790	2.48E-05	-1.28	0.53	0.54	7.12E-01	S	PXM16	FALSE	FALSE
AT1G56680	1.20E-06	-1.28	0.32	0.43	3.44E-02	S		Unsigned Reference	FALSE
AT3G05060	1.72E-10	-1.26	0.46	0.53	4.94E-06	S	NOP5-2	Unsigned Reference	FALSE
AT3G18230	5.54E-06	-1.26	0.48	0.54	1.59E-01	S		FALSE	FALSE
AT5G62190	8.60E-09	-1.26	0.54	0.57	2.47E-04	S	RH7	Unsigned Reference	FALSE
AT5G14520	7.75E-08	-1.24	0.44	0.50	2.22E-03	S	PES	Unsigned Reference	FALSE
AT1G52930	2.39E-08	-1.24	0.35	0.44	6.85E-04	S	BRIX1-2	Unsigned Reference	FALSE
AT1G11260	2.30E-03	-1.22	0.45	0.47	1.00E+00	S	STP1	FALSE	FALSE

Figure S7: The analyses identified S phase markers genes exhibiting a twofold increase in expression ($avg\_log2FC > 1$) and presence in at least 40% ($pct.1 > 0.40$) of the cells in the target phase. These genes were primarily ranked by fold change. In cases of ties, genes were further ranked by ascending expression in non-target phases ($pct.2$) and descending expression in the target phase ($pct.1$). The top 30 genes were selected as potential positive markers. Negative markers were determined by negative avg_log2FC ($avg\_log2FC > 1$) and other similar selection criteria to those used for positive markers. The positive markers marked in red indicate that they originate from the list of 1080 cycle related genes proposed by Menges et al. (2003)[316], while those marked in green correspond to the negative markers. The markers labeled with the term "Core" in yellow are included in the core cell cycle genes summarized by Beemster et al. (2005)[279].

gene	p_val	avg_log2FC	pct.1	pct.2	p_val_adj	cluster	gene_symbol	highlight	core
AT3G59370	8.22E-47	4.31	0.43	0.14	2.35E-42	G2		FALSE	FALSE
AT2G13820	3.47E-72	4.01	0.48	0.11	9.94E-68	G2		FALSE	FALSE
AT1G12080	3.40E-57	3.83	0.55	0.23	9.74E-53	G2		FALSE	FALSE
AT1G23080	1.27E-69	3.82	0.43	0.09	3.64E-65	G2	PIN7	FALSE	FALSE
AT4G11210	2.03E-60	3.73	0.45	0.12	5.81E-56	G2	DIR14	FALSE	FALSE
AT2G01420	1.19E-75	3.49	0.51	0.12	3.42E-71	G2	PIN4	Unsigned Reference	FALSE
AT5G59010	2.28E-65	3.48	0.55	0.21	6.54E-61	G2		Reference	FALSE
AT1G72150	2.45E-41	3.38	0.56	0.33	7.02E-37	G2	PATL1	FALSE	FALSE
AT1G80690	1.62E-63	3.37	0.42	0.09	4.64E-59	G2		FALSE	FALSE
AT1G22530	1.75E-54	3.36	0.58	0.28	5.02E-50	G2	PATL2	FALSE	FALSE
AT2G46890	1.48E-12	3.34	0.43	0.36	4.23E-08	G2		FALSE	FALSE
AT1G65960	3.66E-56	3.28	0.46	0.15	1.05E-51	G2	GAD2	FALSE	FALSE
AT1G70940	1.24E-69	3.23	0.47	0.11	3.54E-65	G2	PIN3	FALSE	FALSE
AT5G03610	4.21E-61	3.16	0.45	0.11	1.21E-56	G2		FALSE	FALSE
AT2G37050	7.60E-54	3.16	0.42	0.12	2.18E-49	G2		FALSE	FALSE
AT2G26730	1.52E-52	3.12	0.46	0.15	4.36E-48	G2		FALSE	FALSE
AT5G53250	1.20E-49	3.05	0.67	0.45	3.45E-45	G2	AGP22	FALSE	FALSE
AT1G55330	1.84E-44	2.98	0.47	0.18	5.28E-40	G2	AGP21	FALSE	FALSE
AT2G27230	6.13E-58	2.97	0.53	0.19	1.76E-53	G2	LHW	Reference	FALSE
AT1G11185	2.42E-32	2.97	0.41	0.17	6.94E-28	G2		FALSE	FALSE
AT1G78260	2.94E-45	2.94	0.42	0.13	8.44E-41	G2		FALSE	FALSE
AT1G33470	8.42E-37	2.93	0.45	0.21	2.41E-32	G2		FALSE	FALSE
AT4G26320	9.30E-46	2.82	0.60	0.34	2.66E-41	G2	AGP13	FALSE	FALSE
AT4G02290	2.69E-59	2.80	0.59	0.25	7.71E-55	G2	AtGH9B13	FALSE	FALSE
AT1G62480	3.09E-41	2.80	0.71	0.58	8.86E-37	G2		FALSE	FALSE
AT5G51780	8.18E-45	2.77	0.43	0.15	2.34E-40	G2		FALSE	FALSE
AT5G19190	1.17E-29	2.74	0.41	0.18	3.35E-25	G2		FALSE	FALSE
AT2G30930	2.85E-42	2.68	0.67	0.49	8.16E-38	G2		FALSE	FALSE
AT4G37250	3.69E-44	2.64	0.50	0.21	1.06E-39	G2		FALSE	FALSE
AT1G22330	1.09E-42	2.62	0.41	0.14	3.11E-38	G2		FALSE	FALSE
AT1G50490	4.16E-56	-4.69	0.16	0.58	1.19E-51	G2	UBC20	FALSE	Core
AT5G10400	1.78E-30	-4.24	0.22	0.49	5.09E-26	G2	HTR2	FALSE	FALSE
AT3G53650	7.03E-44	-3.84	0.20	0.56	2.01E-39	G2		FALSE	FALSE
AT1G02690	6.78E-45	-3.56	0.12	0.49	1.94E-40	G2	IMPA-6	FALSE	Core
AT1G54690	4.41E-48	-3.44	0.24	0.59	1.26E-43	G2	HTA3	FALSE	FALSE
AT3G46320	1.75E-36	-3.35	0.13	0.44	5.00E-32	G2		FALSE	FALSE
AT3G27060	2.49E-35	-3.10	0.12	0.43	7.14E-31	G2	TSO2	FALSE	FALSE
AT3G45930	2.10E-33	-3.04	0.15	0.45	6.03E-29	G2		FALSE	FALSE
AT5G59690	4.36E-40	-2.98	0.28	0.61	1.25E-35	G2		FALSE	FALSE
AT1G69770	1.52E-37	-2.96	0.08	0.41	4.36E-33	G2	CMT3	Unsigned Reference	FALSE
AT5G22880	2.36E-37	-2.90	0.14	0.47	6.76E-33	G2	H2B	FALSE	FALSE
AT5G65360	2.03E-27	-2.87	0.49	0.71	5.81E-23	G2	HTR2	FALSE	FALSE
AT1G09200	1.59E-55	-2.79	0.34	0.72	4.55E-51	G2	HTR2	FALSE	FALSE
AT2G28740	9.43E-30	-2.73	0.39	0.63	2.70E-25	G2	HIS4	FALSE	FALSE
AT3G54560	2.61E-52	-2.71	0.21	0.61	7.49E-48	G2	H2AV	FALSE	FALSE
AT2G33790	4.42E-31	-2.69	0.15	0.44	1.27E-26	G2	AGP30	FALSE	FALSE
AT2G37470	8.17E-45	-2.68	0.39	0.72	2.34E-40	G2		FALSE	FALSE
AT5G63550	1.81E-28	-2.63	0.16	0.43	5.18E-24	G2		FALSE	FALSE
AT3G27360	8.32E-24	-2.59	0.40	0.63	2.38E-19	G2	HTR2	FALSE	Core
AT2G29570	4.09E-35	-2.50	0.11	0.42	1.17E-30	G2	PCNA2	Unsigned Reference	FALSE
AT1G07820	9.32E-38	-2.48	0.50	0.75	2.67E-33	G2		FALSE	FALSE
AT4G22230	2.19E-48	-2.46	0.11	0.52	6.28E-44	G2		FALSE	FALSE
AT5G59870	1.07E-10	-2.44	0.35	0.46	3.06E-06	G2	HTA6	FALSE	FALSE
AT1G07790	1.50E-46	-2.32	0.26	0.65	4.31E-42	G2	HTB1	FALSE	FALSE
AT3G46940	4.03E-32	-2.30	0.11	0.41	1.16E-27	G2	DUT1	FALSE	FALSE
AT3G45980	2.69E-29	-2.26	0.63	0.79	7.71E-25	G2	H2B	FALSE	FALSE
AT2G27970	5.97E-22	-2.22	0.24	0.49	1.71E-17	G2	CKS2	Unsigned Reference	FALSE
AT5G59030	5.56E-40	-2.19	0.08	0.44	1.59E-35	G2	COPT1	FALSE	FALSE
AT3G04320	1.01E-47	-2.11	0.09	0.50	2.90E-43	G2		FALSE	FALSE
AT5G54640	1.69E-29	-2.10	0.31	0.59	4.84E-25	G2	RAT5	FALSE	FALSE

Figure S8: The analyses identified G2 phase markers genes exhibiting a twofold increase in expression ($avg\_log2FC > 1$) and presence in at least 40% ($pct.1 > 0.40$) of the cells in the target phase. These genes were primarily ranked by fold change. In cases of ties, genes were further ranked by ascending expression in non-target phases ($pct.2$) and descending expression in the target phase ($pct.1$). The top 30 genes were selected as potential positive markers. Negative markers were determined by negative avg_log2FC ($avg\_log2FC > 1$) and other similar selection criteria to those used for positive markers. The positive markers marked in red indicate that they originate from the list of 1080 cycle related genes proposed by Menges et al. (2003)[316], while those marked in green correspond to the negative markers. The markers labeled with the term "Core" in yellow are included in the core cell cycle genes summarized by Beemster et al. (2005)[279].s

gene	p_val	avg_log2FC	pct.1	pct.2	p_val_adj	cluster	gene_symbol	highlight	core
AT1G02730	1.07E-216	13.42	0.65	0.00	3.07E-212	M	CSLD5	Reference	Core
AT4G14330	1.32E-130	13.03	0.42	0.00	3.78E-126	M	KIN10A	FALSE	FALSE
AT3G51280	1.15E-183	12.35	0.57	0.00	3.29E-179	M		Reference	Core
AT4G38062	3.53E-185	12.34	0.57	0.00	1.01E-180	M		FALSE	FALSE
AT5G48310	2.59E-194	11.68	0.60	0.00	7.41E-190	M		Reference	Core
AT5G51600	1.77E-172	10.91	0.54	0.00	5.07E-168	M	MAP65-3	Reference	Core
AT1G34355	1.10E-122	10.69	0.40	0.00	3.14E-118	M	PS1	Reference	Core
AT1G16630	2.81E-139	10.47	0.45	0.00	8.05E-135	M		FALSE	FALSE
AT5G55520	4.71E-138	10.44	0.44	0.00	1.35E-133	M		Reference	Core
AT5G06150	1.83E-128	10.43	0.43	0.01	5.25E-124	M	CYCB1-2	FALSE	FALSE
AT4G05190	1.85E-239	10.36	0.72	0.01	5.30E-235	M	KIN14D	Reference	Core
AT3G02640	5.89E-118	10.33	0.40	0.01	1.69E-113	M		Reference	Core
AT5G16250	2.75E-175	10.08	0.56	0.01	7.88E-171	M		Reference	Core
AT2G38160	1.18E-180	9.97	0.57	0.00	3.37E-176	M		FALSE	FALSE
AT4G09060	1.21E-160	9.47	0.52	0.01	3.47E-156	M		FALSE	FALSE
AT4G01730	2.64E-169	9.46	0.54	0.01	7.55E-165	M	PAT18	Reference	Core
AT1G03780	1.59E-147	9.40	0.48	0.00	4.55E-143	M	TPX2	FALSE	FALSE
AT1G18370	4.52E-205	9.33	0.63	0.01	1.30E-200	M	KIN7A	Reference	Core
AT4G31840	5.22E-110	9.28	0.40	0.02	1.50E-105	M	ENODL15	Reference	Core
AT5G15510	1.08E-144	9.11	0.47	0.00	3.09E-140	M		Reference	Core
AT2G22610	6.18E-179	8.72	0.57	0.01	1.77E-174	M		Reference	Core
AT1G76310	2.70E-145	8.65	0.47	0.00	7.74E-141	M	CYCB2;4	Reference	Core
AT3G44050	4.42E-213	8.49	0.67	0.02	1.27E-208	M	KIN12E	Reference	Core
AT1G20610	2.75E-172	8.34	0.56	0.01	7.88E-168	M	CYCB2-3	FALSE	FALSE
AT1G44110	2.66E-125	8.32	0.43	0.01	7.62E-121	M	CYCA1-1	Reference	Core
AT5G17160	2.64E-192	8.31	0.65	0.03	7.57E-188	M		Reference	Core
AT3G23670	1.96E-230	8.15	0.71	0.01	5.62E-226	M	KIN12B	Reference	Core
AT3G58650	1.54E-153	8.04	0.51	0.01	4.40E-149	M		Reference	Core
AT1G72670	6.24E-165	7.96	0.54	0.01	1.79E-160	M	iqd8	Unsigned Reference	FALSE
AT1G53140	1.74E-146	7.93	0.49	0.01	4.98E-142	M	DRP5A	Reference	Core
AT5G10400	5.23E-59	-5.66	0.13	0.53	1.50E-54	M	HTR2	FALSE	FALSE
AT5G59870	2.22E-57	-5.01	0.15	0.54	6.35E-53	M	HTA6	FALSE	FALSE
AT1G62480	7.48E-34	-4.50	0.44	0.68	2.14E-29	M		FALSE	FALSE
AT4G02060	2.72E-63	-4.50	0.02	0.45	7.78E-59	M	MCM7	Unsigned Reference	FALSE
AT3G46940	1.51E-57	-4.13	0.05	0.45	4.32E-53	M	DUT1	FALSE	FALSE
AT1G07370	1.70E-65	-4.06	0.04	0.48	4.86E-61	M	PCNA	Unsigned Reference	FALSE
AT1G54690	2.00E-70	-3.91	0.19	0.63	5.72E-66	M	HTA3	FALSE	FALSE
AT4G26320	3.01E-26	-3.86	0.24	0.48	8.61E-22	M	AGP13	FALSE	FALSE
AT3G45930	2.47E-45	-3.83	0.13	0.47	7.08E-41	M		FALSE	FALSE
AT5G65360	5.76E-46	-3.66	0.45	0.73	1.65E-41	M	HTR2	FALSE	FALSE
AT4G39800	5.86E-58	-3.54	0.12	0.54	1.68E-53	M	IPS1	FALSE	FALSE
AT5G63550	2.12E-52	-3.53	0.09	0.47	6.07E-48	M		FALSE	FALSE
AT4G32460	2.70E-47	-3.52	0.07	0.43	7.74E-43	M		FALSE	FALSE
AT3G46320	1.59E-32	-3.52	0.17	0.44	4.56E-28	M		FALSE	FALSE
AT2G29570	8.65E-62	-3.33	0.04	0.46	2.48E-57	M	PCNA2	Unsigned Reference	FALSE
AT2G21790	4.04E-51	-3.18	0.04	0.42	1.16E-46	M	RNR1	Unsigned Reference	FALSE
AT3G53650	9.39E-13	-3.15	0.38	0.50	2.69E-08	M		FALSE	FALSE
AT4G02290	2.28E-37	-3.10	0.12	0.43	6.54E-33	M	AtGH9B13	FALSE	FALSE
AT4G27230	2.96E-40	-3.10	0.11	0.43	8.47E-36	M	HTA2	FALSE	FALSE
AT3G16570	4.77E-44	-3.10	0.06	0.40	1.37E-39	M	RALF23	FALSE	FALSE
AT3G27360	1.44E-27	-3.08	0.42	0.63	4.12E-23	M	HTR2	FALSE	Core
AT2G30520	3.33E-43	-2.98	0.11	0.46	9.54E-39	M	RPT2	FALSE	FALSE
AT3G05910	2.64E-52	-2.92	0.09	0.48	7.58E-48	M	PAE12	FALSE	FALSE
AT2G46890	4.61E-20	-2.89	0.21	0.44	1.32E-15	M		FALSE	FALSE
AT5G59690	7.17E-22	-2.86	0.40	0.57	2.05E-17	M		FALSE	FALSE
AT5G53250	1.35E-23	-2.81	0.35	0.57	3.87E-19	M	AGP22	FALSE	FALSE
AT2G28740	3.56E-28	-2.80	0.40	0.63	1.02E-23	M	HIS4	FALSE	FALSE
AT5G09960	5.31E-38	-2.67	0.18	0.52	1.52E-33	M		FALSE	FALSE
AT1G72150	1.19E-11	-2.66	0.29	0.43	3.41E-07	M	PATL1	FALSE	FALSE
AT5G47370	1.45E-35	-2.55	0.23	0.53	4.17E-31	M	HAT2	FALSE	FALSE

Figure S9: The analyses identified M phase markers genes exhibiting a twofold increase in expression ($avg\_log2FC > 1$) and presence in at least 40% ($pct.1 > 0.40$) of the cells in the target phase. These genes were primarily ranked by fold change. In cases of ties, genes were further ranked by ascending expression in non-target phases ($pct.2$) and descending expression in the target phase ($pct.1$). The top 30 genes were selected as potential positive markers. Negative markers were determined by negative avg_log2FC ($avg\_log2FC > 1$) and other similar selection criteria to those used for positive markers. The positive markers marked in red indicate that they originate from the list of 1080 cycle related genes proposed by Menges et al. (2003)[316], while those marked in green correspond to the negative markers. The markers labeled with the term "Core" in yellow are included in the core cell cycle genes summarized by Beemster et al. (2005)[279].