Cis-Regulatory Elements in Plant Development, Adaptation, and Evolution

Inaugural-Dissertation

zur Erlangung des Doktorgrades
der Mathematisch-Naturwissenschaftlichen Fakultät

der Heinrich-Heine-Universität Düsseldorf

vorgelegt von

Duong Thi Hai Doan

aus Hai Phong, Vietnam

Düsseldorf, März 2026

aus dem Institut für Molekulare Physiologie
der Heinrich-Heine-Universität Düsseldorf

Gedruckt mit der Genehmigung der
Mathematisch-Naturwissenschaftlichen Fakultät der
Heinrich-Heine-Universität Düsseldorf

Berichterstatter:

1. **Prof. Dr. Wolf B. Frommer**

2. **Prof. Dr. Benjamin Stich**

Tag der mündlichen Prüfung: 06.11.2026

TABLE OF CONTENTS

Preamble: **7**

Summary **10**

General Introduction **12**

Chapter 1: Literature Review **14**

1.1. From Code to Function: The Genomic Foundations of Plant Traits 14

1.1.1. The regulatory genome 14

1.1.2. The importance of CREs in plant development and adaptation 15

1.2. Molecular Nature and Classification of Cis-Regulatory Elements 17

1.2.1. Chromatin Context and Spatial Regulation of CRE Activity 17

1.2.2. 3D Genome Architecture 18

1.2.3. Chromatin Accessibility 21

1.2.4. Histone Modifications and Variants 24

1.2.5. DNA Methylation 27

1.3. Functional Categories and Mechanisms of Cis-Regulatory Elements 29

1.3.1. Core promoter 30

1.3.2. Enhancers 32

1.3.3. Silencers 34

1.3.4. Insulators 35

1.4. Methods and Technologies to Identify CREs 37

1.4.1. Protein–DNA Binding Assays 38

1.4.1.1. ChIP-seq 38

1.4.1.2. DAP-seq 40

1.4.1.3. CUT&RUN 42

1.4.1.4. MOA-seq 45

1.4.2. Chromatin Accessibility Assays 46

1.4.2.1. DNase-seq 46

1.4.2.2. ATAC-seq 48

1.4.2.3. MNase-seq 51

1.4.2.4. Fiber-seq and SAM-seq..... 53
1.4.3. Single-Cell approach.....57
1.4.4. Functional Enhancer/CREs Activity Assays.....61
1.4.4.1. Classical reporter assays.....62
1.4.4.2. STARR-seq..... 62
1.4.4.3. CRISPR-based functional screens..... 64
1.5. Concluding Remarks and Outlook..... 65
Chapter 2: The evolution of Cis-regulatory elements in Embryophytes..... 68
2.1. Introduction..... 68
2.2. Material and methods..... 70
2.2.1. Plant materials..... 70
2.2.2. Data materials.....71
2.2.3. MOA-seq.....71
2.2.3.1. Protocol for monocot species.....71
2.2.3.2. Protocol for dicot species.....72
2.2.4. Raw data processing..... 73
2.2.4.1. Sequencing quality control.....73
2.2.4.2. Adaptor trimming.....73
2.2.4.3. Read merging and quality enhancement..... 73
2.2.5. MOA reads mapping.....74
2.2.6. Peak calling.....76
2.2.7. Gene annotation..... 76
2.2.8. Phylogenetic Tree Reconstruction Using Chloroplast Genes..... 77
2.2.8.1. Sequence Retrieval and Identification of Shared Chloroplast Genes..... 77
2.2.8.2. Sequence Extraction and Header Normalization..... 77
2.2.8.3. Sequence Alignment and Trimming.....78
2.2.8.4. Maximum-Likelihood Gene Tree Construction.....78
2.2.8.5. Species Tree Inference Using ASTRAL-IV..... 78
2.2.8.6. Divergence Time Estimation with TreePL..... 78

2.2.8.7. Tree Visualization....79
2.2.9. Determination of distances from MOA peaks to nearest genes....79
2.2.10. Identification of MOA peak overlap with genomic features....79
2.2.11. Quantification of genomic feature sizes....80
2.2.12. Quantification of MOA peaks....80
2.2.13. Quantification of total MOA peak width....80
2.2.14. Quantification of protein-coding genes....80
2.2.15. Calculation of actual genome size....81
2.2.16. Data visualization....81
2.3. Results....81
2.3.1. Establishment of a Multi-Species MOA-seq Dataset....81
2.3.2. Profiling TF cis-regulatory regions evolution throughout the Embryophytes clade 95
2.4. Discussion....108
Chapter 3: Pancistrome-wide characterization of cis-element evolution of the Brassinosteroid signaling pathway....112
3.1. Introduction....112
3.2. Material and methods....114
3.2.1. Plant materials....114
3.2.2. Data materials....115
3.2.3. MOA-seq....115
3.2.4. Preprocessing and alignment of MOA-seq data....115
3.2.5. MOA peak calling and identification of BR-responsive regulatory regions....116
3.2.7. BigWig track generation for genome browser visualization....117
3.2.8. MOA-seq and ChIP-seq signal visualization....118
3.2.9. Identification of Putative Orthologs of BR Biosynthesis Genes....118
3.3. Results....119
3.3.1 Comparative analysis of intronic transcription factor binding in brassinosteroid biosynthesis genes....119
3.3.2 Generation of MOA-seq profiles under brassinosteroid treatment in Arabidopsis and barley....130

3.3.3 Genome-wide and locus-specific changes in MOA-seq occupancy under altered brassinosteroid levels..... 134
3.4. Discussion..... 143
Chapter 4: Characterization of genes identified by FIND-CIS for bioengineering of drought and pathogen tolerance in maize..... 147
4.1. Introduction..... 147
4.2. Materials and Methods..... 155
4.2.1. Plant material..... 155
4.2.2. Primer design and sequences..... 155
4.2.3. Genotyping of BonnMu insertion lines..... 157
4.3. Result: Genotyping of BonnMu insertion lines..... 158
4.4. Discussion..... 162
Summary of the thesis..... 164
Appendix..... 169
5.1. Scripts and analysis pipelines..... 169
5.1.1. Adaptor trimming, read merging, and quality enhancement script..... 169
5.1.2. MOA-seq read mapping script..... 171
5.1.3. Peak-calling script..... 173
5.1.4. Helixer gene-annotation script..... 176
5.1.5. Script for calculating MOA peak to TSS distance..... 179
5.1.6. Script for intersecting MOA peaks with genome features..... 181
5.1.7. Script for calculating genome-feature coverage percentages..... 195
5.1.8. Computational pipeline for chloroplast-based phylogenetic reconstruction..... 199
5.1.8.1. Sequence retrieval..... 199
5.1.8.2. Identification of chloroplast genes shared across all species..... 199
5.1.8.3. Extraction of gene sequences and header normalization..... 199
5.1.8.4. Multiple sequence alignment..... 200
5.1.8.5. Trimming poorly aligned regions..... 201
5.1.8.5. Gene tree construction using IQ-TREE2..... 201
5.1.8.6. Species tree inference with ASTRAL-IV..... 201

5.1.8.7. Divergence time estimation with TreePL....202
5.1.8.8. Tree visualization in R....203
5.1.9. Genomic feature overlap analysis for Chapter 3....204
5.1.9.1. Analysis for Hordeum vulgare samples....204
5.1.9.2. Analysis for Arabidopsis thaliana samples....207
5.1.9.3. Script for intersecting MOA peaks with genome features for two treatment conditions....210
5.1.10. Script for BigWig track generation for genome browser visualization....217
5.1.11. MACS3 peak calling and post-processing commands for PPZ and PPZ+eBL comparisons....219
5.2. Original MOA-seq Agarose Gel Images....223
5.3. Supplement tables....229
5.4. List of Figures....230
5.5. List of Tables....234
5.6. List of Abbreviations....235
Acknowledgement....238
References....239

Preamble:

This thesis was carried out as a doctoral researcher of the Institute of Molecular Physiology, Heinrich-Heine-University Düsseldorf, in the independent research group (FIND-CIS group of Thomas Hartwig) at the Max Planck Institute for Plant Breeding Research, Cologne, Germany, and investigates the evolution of cis-regulatory architecture in plant genomes using comparative MOA-seq analyses and pathway-focused case studies.

The work was supervised by Prof. Dr. Wolf Frommer, Dr. Thomas Hartwig (both: Institute of Molecular Physiology, Heinrich-Heine-Universität Düsseldorf, Germany), and Dr. Julia Engelhorn (DIADE, University of Montpellier, CIRAD, IRD, Montpellier, France). Furthermore, guidance and advice were given by Prof. Dr. Benjamin Stich (Julius Kühn Institute, Institute for Breeding Research on Agricultural Crops, Germany), Prof. Dr. Jeffrey Ross-Ibarra (Dept. Evolution and Ecology, University of California Davis, USA), and Prof. Dr. Edward Buckler together with Dr. Cinta Romay (both: Institute for Genomic Diversity, Cornell University, USA), who provided scientific guidance, conceptual input, and feedback throughout the project. Prof. Dr. Andreas Weber (Institute of Plant Biochemistry, Heinrich Heine University Düsseldorf) as my mentor. Dr. Thomas Hartwig and Dr. Julia Engelhorn furthermore contributed to experimental design, data interpretation, and discussion of results across multiple chapters. Amelie Kok (PhD student, The Cluster of Excellence on Plant Sciences, Heinrich-Heine-Universität Düsseldorf, Germany) provided support on data analysis, interpretation, and guided me with coding, especially during the early stage of the project.

Contributions to Chapter 2

Previously unpublished MOA-seq experiments for Chapter 2 were performed collaboratively across several plant species. MOA-seq in *Oryza sativa* was carried out together with Dr. Sohini Mukherjee (laboratory manager, Institute of Molecular Physiology, Heinrich-Heine-Universität Düsseldorf, Germany). MOA-seq in *Sorghum bicolor* was performed by Thimo Trapp (former bachelor's student, Institute of Molecular Physiology, Heinrich-Heine-Universität Düsseldorf, Germany). Thimo Trapp and I performed the MOA-seq experiments in *Brassica napus*, *Vicia faba*, and *Glycine max*.

He furthermore assisted with tissue grinding for all samples used in MOA-seq experiments performed by me. In addition, Thimo Trapp contributed to the early stages of MOA-seq protocol optimization for *Physcomitrium patens* and *Betula pendula*. I completed the MOA-seq optimization for *Betula pendula*. All library preparation steps for sequencing were performed by me.

The original pipeline for MOA-seq analysis, including read preparation, read mapping, and peak calling, was developed and published by Dr. Julia Engelhorn, Max Blanck, Amelie Kok, and Dr. Thomas Hartwig. Similarly, the original idea and design for comparing genome size to MOA peak number, per gene and distance to genes, were developed and published by Dr. Julia Engelhorn, Prof. Dr. Jeffrey Ross-Ibarra, Dr. Thomas Hartwig, Felix Plotecki (Plotecki 2023), and Thimo Trapp (Trapp 2024). The original pipelines were adapted and further expanded, including adding new features by me with the support of AI tools for code development for the analysis across the meta-cistrome in chapter 2. Previously analyzed MOA-seq data for multiple species was reanalyzed using my adapted pipeline to ensure uniformity of the metacistrome data and analysis. The analysis of the data in chapter 2 was done by me, and I interpreted the results with support from Dr. Thomas Hartwig, Dr. Julia Engelhorn, and Prof. Dr. Jeffrey Ross-Ibarra.

Contributions to Chapter 3

Brassinosteroid perturbation experiments in *Hordeum vulgare* and *Arabidopsis thaliana* were performed by Dr. Jiaying Zhu at the Leibniz Institute of Plant Genetics and Crop Plant Research (IPK), Germany. MOA-seq experiments were performed by me. Initial MOA data analysis was performed by me. Comparative analysis, including figure generation, script development, and data interpretation, was performed by Dr. Thomas Hartwig, Dr. Julia Engelhorn, Amelie Kok, and me.

Contributions to Chapter 4

Functional analysis and selection of the candidate genes were performed by Dr. Thomas Hartwig, Dr. Julia Engelhorn, and Max Blank (Blank 2020). Fieldwork, plant crossing, and genotyping experiments were performed jointly by myself, supported by Thimo

Trapp, Sergius Weizel, Dr. Sohini Mukherjee, Dr. Thomas Hartwig, and Martin Krist (Field Manager, MPIPZ). Thimo Trapp provided assistance with field experiments and contributed to genotyping efforts alongside me. Viola Rajan, Nafiseh Sargheini, and Felix Plotecki supported the fieldwork during the first field season (2022).

Statement of Authorship and AI Assistance

As a non-native English speaker, I used Grammarly, QuillBot, and ChatGPT as language-support tools to improve clarity, grammar, and phrasing. Draft text was first revised using Grammarly and QuillBot. If issues in structure or flow remained, ChatGPT was used for further refinement. ChatGPT was also used to develop, review, and correct computational scripts. Any script developed by myself as well as previously developed scripts were also submitted to review by ChatGPT to adapt, identify errors, and improve them. As long as not stated otherwise – e.g. provided by other persons – experimental work, data analysis, figure preparation, and writing of this thesis were carried out by me, and the respective scientific content, data interpretation, and conclusions presented in this thesis are solely the responsibility of myself.

I hereby declare that I have prepared this dissertation independently and without unauthorized assistance, in accordance with the principles of Good Research Practice at Heinrich Heine University Düsseldorf. This dissertation has not been submitted, in whole or in part, to any other institution for academic evaluation.

Düsseldorf, 31/03/2026

Duong Thi Hai Doan

Summary

Plant genomes vary widely in size, ploidy level, and transposable element content, yet conserved biological pathways are maintained across deep evolutionary timescales. How cis-regulatory information is distributed and preserved within such divergent genomic contexts remains incompletely understood. This thesis investigates the organization and evolution of cis-regulatory landscapes in plants, with a focus on transcription factor occupancy and its relationship to genome architecture and regulatory robustness. To address this question, MOA-seq was applied across 28 plant species representing more than 500 million years of evolution and spanning a broad range of genome sizes, evolution histories, ploidy levels, and repeat content. By using a consistent experimental and analytical framework, transcription factor occupancy profiles were compared across species to assess how regulatory landscapes relate to genome expansion and genome duplication. In parallel, a conserved hormonal pathway, brassinosteroid biosynthesis, was examined across multiple species to evaluate how pathway-level regulation is maintained despite divergence in cis-regulatory architecture.

Genome-wide analyses showed that total MOA-seq peak abundance generally increased with genome size, consistent with an expansion of regulatory sequence space in larger genomes. However, regulatory element abundance did not scale proportionally with gene number, and the number and genomic distribution of peaks per gene varied substantially among species. Larger and repeat-rich genomes exhibited a shift toward more distal transcription factor occupancy, while promoter-proximal binding remained detectable even in the largest genomes examined. Ploidy history further influenced regulatory organization, with several polyploid species deviating from trends observed in diploid lineages.

Within the brassinosteroid biosynthesis pathway, intron-localized transcription factor occupancy was detected at multiple genes across species, including at the conserved DWF4 locus. While intronic occupancy was recurrent, its genomic position and extent varied among species and loci, indicating positional flexibility rather than strict conservation of cis-regulatory element location. These observations are compatible with

models in which conserved regulatory outcomes are achieved through flexible use of different genomic compartments.

Finally, a set of maize genes previously implicated in regulatory variation were examined using transposon insertion lines to explore the feasibility of gene-level functional validation for regulatory hypotheses. Although stable homozygous mutant lines could not be established for most candidates, these analyses highlight practical constraints of using gene disruption approaches to interrogate cis-regulatory variation.

Together, this work provides a comparative framework for understanding how cis-regulatory landscapes evolve in plant genomes while maintaining functional robustness in conserved biological pathways.

General Introduction

While the genetic code determining amino acid sequences has been well understood over the past decades, meanwhile, the principles specifying when and where genes are turned on and off, encoded in the so-called cis-regulatory code of the genome, are still incompletely understood, especially in plants. Although protein-coding genes make up very little of plant genomes, it is in these large regions of noncoding DNA where the regulatory elements for development and responses to environmental stimuli reside (Rodgers-Melnick et al. 2016). In the last two decades, advances in technology have greatly impacted plant biology. Hundreds of high-quality plant genomes and pangenomes have been assembled that provide comparative resources previously unavailable to researchers (Bernal-Gallardo and de Folter 2024). Simultaneously, epigenomic profiling approaches, such as chromatin accessibility assays, histone mark mapping, and 3D genome architecture studies, have produced rich datasets that have increased the granularity of our ability to examine regulatory regions. These advances have enabled researchers to begin to study the regulatory genome from many lineages, yet the functional and evolutionary contexts of these regions remain incomplete (Marand et al. 2023).

It is essential to understand CREs and their evolution since regulatory DNA variation often accounts for phenotypic diversification, domestication, and stress responses in plants (Xiang Li and Schmitz 2025). Comparative studies have primarily examined the divergence of protein-coding genes, while the evolution of regulatory landscapes across species with divergent genome size, structure, and ecological strategies has not been fully explored. The study of the evolution of CREs is critical to our understanding of how developmental pathways are rewired, how systems of stress responses are devised, and how innovations in regulation contribute to phenotypic diversity. Nevertheless, systematic cross-species comparative analyses of regulatory footprints have been lacking.

This thesis aims to understand CREs' evolution and regulatory architecture in plants. The data generated in this thesis were obtained via a method called MOA-seq (MNase-defined cistrome-Occupancy Analysis). It has been developed with an effort to overcome the obstacles from other methods, such as antibody dependence, large-scale

limitations, insufficient resolution for precise mapping of CREs, or the lack of capturing chromatin context (Savadel et al. 2021). The first chapter is a literature review providing a solid background of CRE biology, chromatin context, and current methods for regulatory element discovery. In the second chapter, MOA-seq was applied across 28 plant species representing major lineages and diverse genome architectures to compare an atlas of transcription factor binding footprints, allowing for a major study of plant CRE evolution at the genomic scale. In chapter 3, we examined how a fundamental developmental and physiological pathway, brassinosteroid biosynthesis and signaling, has been regulated and possibly rewired over evolutionary time. Finally, chapter 4 pursues functional characterization of candidate regulatory genes identified using the FIND-CIS framework (Engelhorn et al. 2025), with relevance to bioengineering.

Chapter 1: Literature Review

A Tale of Two Seedlings

Picture two corn seedlings sprouting side by side in the same field. From their first curled leaves, they look identical. Yet one quickly towers above the other, its stalks thick and ears laden with kernels, while its neighbor remains stunted. How can two plants with the same genetic blueprint diverge so dramatically? The answer lies not only in the genes themselves but also in the switches that turn them on and off: the cis-regulatory elements scattered across the genome. In this thesis, we begin by delving into the 'dark matter' of the genome, the regulatory code, and trace its evolution between species, from Arabidopsis to the silver birch tree, to reveal how fundamental pathways like hormone feedback regulation have been rewired over time.

1.1. From Code to Function: The Genomic Foundations of Plant Traits

In contrast to animals, plants are not able to flee drought, seek shade, or migrate to more benign climates. Stationary, they bear the complete force of their surroundings, scorching sun, floods, or crafty insects, with no recourse but to adapt. This immobility has created a remarkable genomic toolkit, whereby the instructions written in DNA not only provide a blueprint for protein assembly but also an adaptive system for their regulation according to situational demands.

1.1.1. The regulatory genome

The 20th century saw the first acknowledgment of DNA as the molecular basis of heredity, ending with the groundbreaking elucidation of its double-helix structure by Watson and Crick in 1953 (Watson and Crick 1953). It transformed the discipline of the biological sciences by clarifying how genetic information is copied and stored. Genes were conceived of as individual units coding for proteins in a linear sequence throughout this time, and extensive stretches of non-coding DNA were largely considered "junk," a term later popularized by Susumu Ohno in 1972 to describe their assumed lack of functional importance (Palazzo and Gregory 2014). This view began to change in the 1970s and 1980s with the discovery of the regulatory sequences in viruses and animals, which are regions of DNA that do not code for proteins but could influence the expression of the nearby genes (Banerji et al. 1981). These sequences are named as

cis-regulatory elements (CREs), which act as binding platforms for binding proteins, also known as transcription factors (TFs), encoding combinatorial logic to switch genes on or off in response to developmental or environmental stimuli (Figure 1.1).

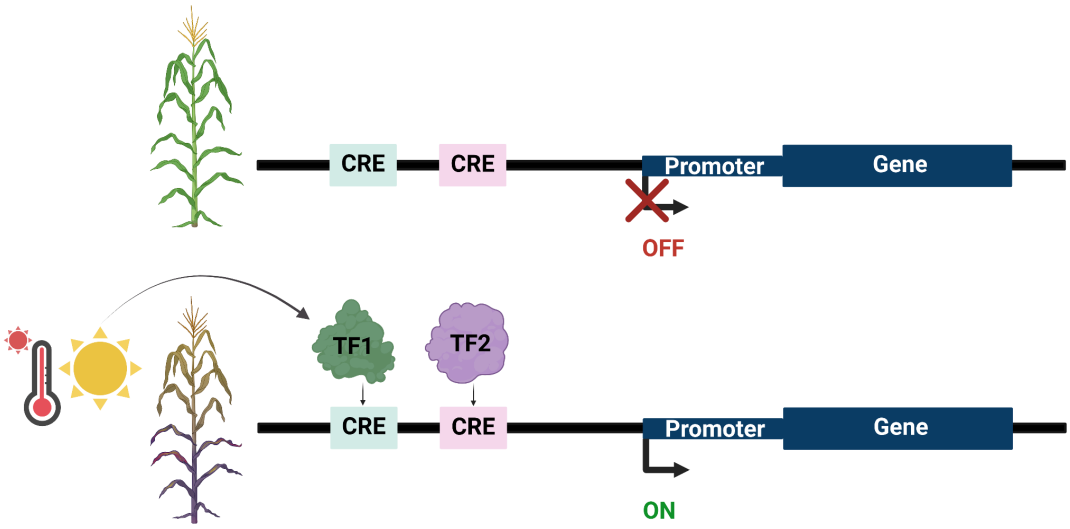


Figure 1.1: Cis-regulatory elements (CREs) act as binding platforms for transcription factors (TFs) that activate gene expression in maize in response to environmental stimuli such as drought.
Figure created by Duong Doan using BioRender.

The regulatory regions occupy the majority of the genome. In brief, estimates for the human genome suggest that over half of its sequence plays regulatory roles, far exceeding the ~2% that encodes proteins (Dunham et al. 2012). In plant genomes, it is challenging to precisely estimate the fraction of regulatory DNA owing to the wide range of variability in genome size and structure across species. For example, *Arabidopsis thaliana*, a model organism with a small genome of around 135 Mb, has relatively fewer intergenic spaces. The genomes of maize (about 2.3 Gb) and wheat (about 17 Gb), in contrast, are dominated by non-coding and repetitive DNA sequences (Schnable et al. 2009; The International Wheat Genome Sequencing Consortium (IWGSC) et al. 2018). It is believed that much of this non-coding space harbors regulatory sequences, but their precise annotation still remains challenging.

1.1.2. The importance of CREs in plant development and adaptation.

CREs act as genomic “switches” enabling the coordination of the complex plant development process and permitting plants to accommodate changing environmental

circumstances without altering protein-coding sequences. Evolutionary changes in plant morphology and physiology can often be traced through such regulatory changes: for instance, a 4.9 kb Hopscotch retrotransposon insertion upstream of the *teosinte branched1* (*tb1*) locus increases *tb1* expression, reducing branches and improving grain yield during maize domestication without altering the *tb1* protein itself (Figure 1.2 and 1.3) (Vann et al. 2015). Similarly, the concurrent selection on the *Shattering1* (*Sh1*) genes within sorghum, rice, and maize concentrated on noncoding regulatory and splice-site variants at this locus to suppress shattering, demonstrating the role for the modulation of CREs underpinning convergently evolved traits related to domestication (Z. Lin et al. 2012). Advances within functional genomics and high-throughput sequencing techniques have enabled systematic mapping of regulatory landscapes within the plant kingdom, revealing widespread patterns of open chromatin and TF binding within different tissues and environmental conditions. The rapid improvement within high-quality, chromosome-scale plant genome assemblies, together with the emergence of pangenomes, presents an unprecedented platform for comparative studies within epigenomics and sequence analysis (Marks et al. 2021). As the inventory of plant genomes and regulatory annotations continues to expand, integrative comparative studies will clarify the role for the turnover, conservation, and rewiring of CREs have driven species-specific adaptations, support predictive models for which regulatory variations lead to phenotypic outcomes, and promote precision breeding alongside synthetic modification of plant traits for future climate conditions.

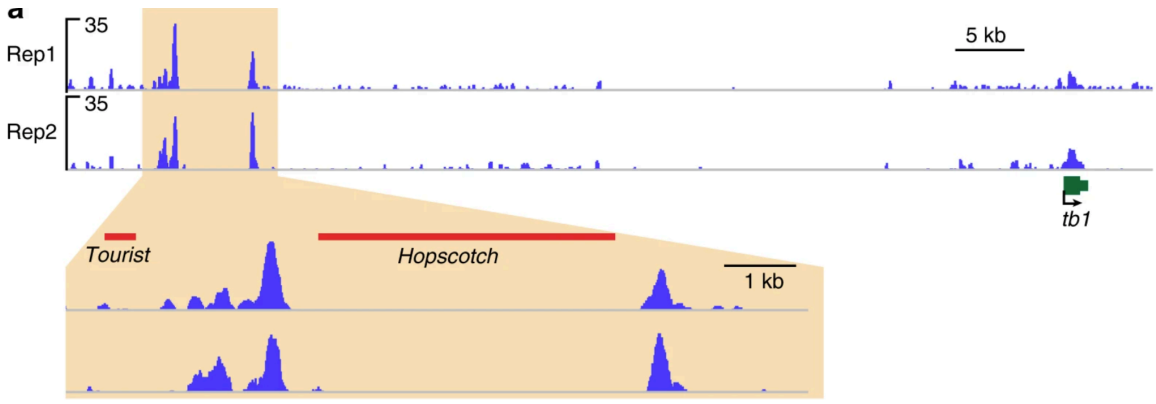


Figure 1.2: TB1 binds to a cis-regulatory element upstream of its own promoter.
ChIP-seq analysis reveals two significant TB1-binding peaks adjacent to the Hopscotch retrotransposon, a regulatory element linked to maize domestication. TB1 occupancy is

detected in two biological replicates (Rep1 and Rep2), shown in blue. The enlarged view highlights the Tourist and *Hopscotch* transposable elements (TEs) located upstream of the tb1 gene. Figure copied from Dong et al. 2019.

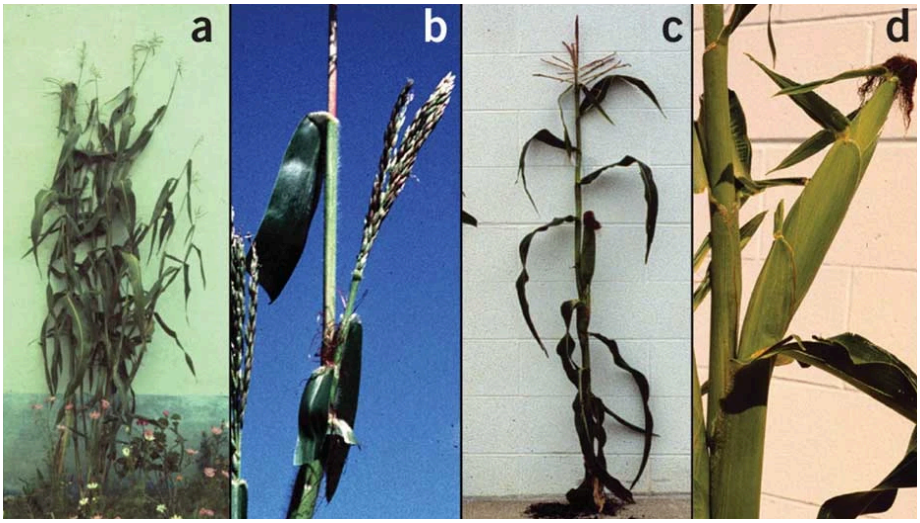


Figure 1.3: Teosinte and maize morphology.
(a) A teosinte plant exhibiting extensive branching. (b) A teosinte lateral branch ending in a tassel. (c) A maize plant with a single, unbranched main stem. (d) A maize ear shoot representing a modified lateral branch. Photos copied from Studer et al. 2011.

1.2. Molecular Nature and Classification of Cis-Regulatory Elements

1.2.1. Chromatin Context and Spatial Regulation of CRE Activity

In addition to DNA sequence, the activity of CREs is tightly controlled by the chromatin landscape and three-dimensional genome organization. Epigenetic phenomena such as DNA methylation, chromatin accessibility, and histone modifications control physical access of CREs to TFs, while genome spatial organization ensures selective enhancer–promoter interaction. Together, these layers coordinately control gene expression in development and environmental response. In this section, we review the intricate regulation of CRE activity, beginning with the 3D genome organization and followed by chromatin accessibility, histone-based control, and DNA methylation.

1.2.2. 3D Genome Architecture

In the nucleus, DNA is packaged in a complex 3D form that dictates genome functions such as DNA replication, DNA repair, and transcription. DNA is coiled around a core of histone proteins in the center to form nucleosomes. In the nucleus, the DNA is organized in a dynamic three-dimensional (3D) structure that governs the genome (Figure 1.4). An integral property of this structure is chromatin looping, which enables physical interactions between remote genomic loci, such as CREs and promoters, and thus coordinates the expression of genes (Giles et al. 2025). Analysis by high-throughput chromosome conformation capture (Hi-C) in rice has revealed extensive chromatin interactions between promoters and CREs, revealing loops ranging from tens to hundreds of kilobases that enhance the coexpression of genes that are involved in growth and stress responses (Y. Liu et al. 2017; L. Zhao et al. 2019). The loops are organized in a 3D landscape that allows specific regulation of transcription and provides a platform on which plant genomes are hierarchically arranged.

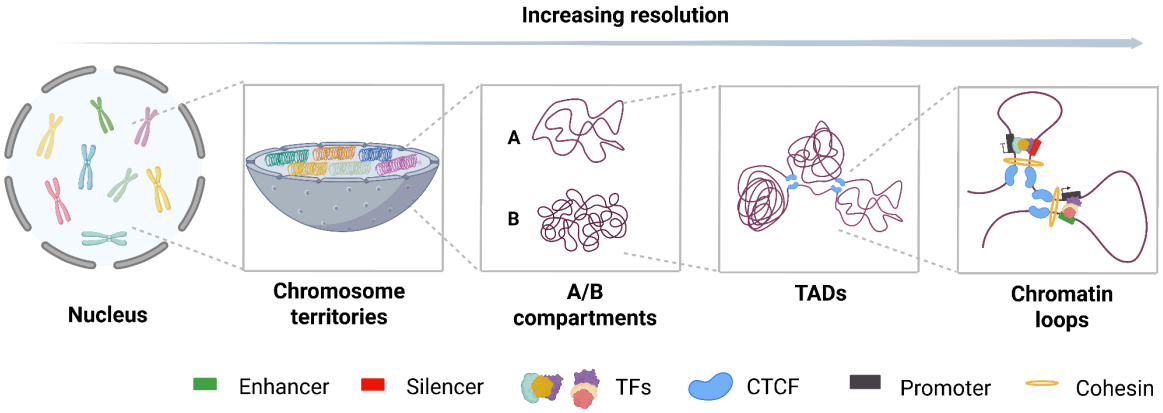


Figure 1.4: 3D genome organization at increasing resolutions.

Schematic representation of the hierarchical folding of chromatin in the nucleus. Starting from the left, there is progressively higher spatial resolution. Chromosomes are allocated specific regions in the nucleus called chromosome territories. Inside a territory, there are regions of chromatin called A and B compartments; these represent mostly transcriptionally active and inactive regions of chromatin. The latter two are split into topologically associating domains (TADs), meaning they have a high affinity within a domain rather than between domains. At the final level of organization, chromatin loops

through binding of cohesin and the architectural protein CTCF; here, distal regulatory regions like enhancers and silencers interact with the promoters of genes to control the initiation of transcription. Icons show localization of promoters, enhancers, silencers, transcription factors, cohesin, and CTCF. Figure created by Duong Doan using BioRender.

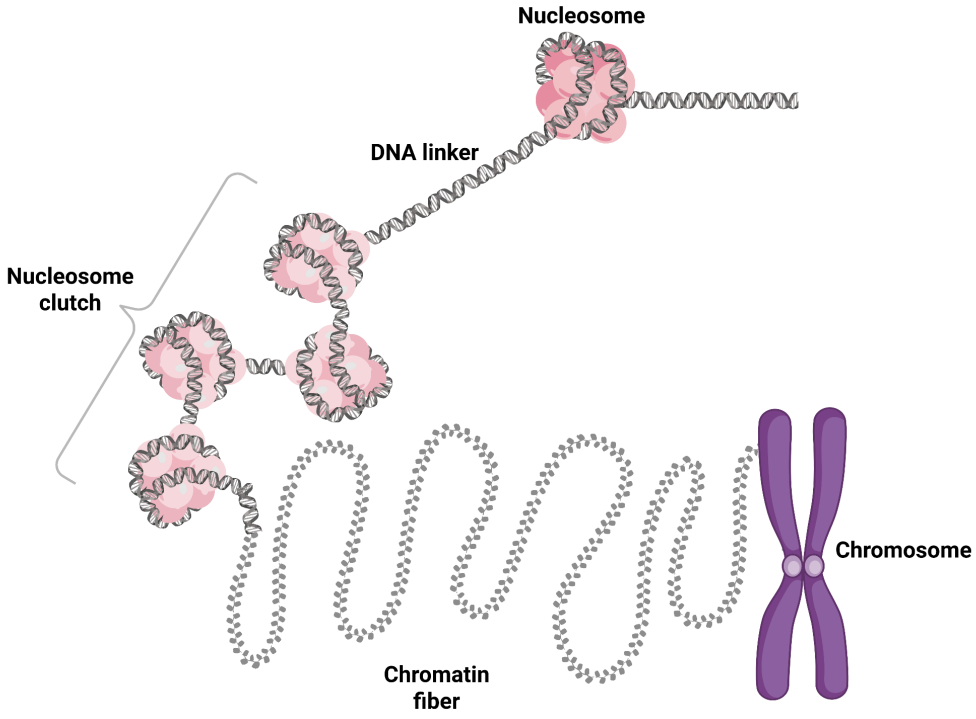


Figure 1.5: Structural organization of chromatin.
Illustration showing how chromatin is organized hierarchically, beginning with DNA coiled around histone octamers forming nucleosomes, then nucleosome clutches, followed by chromatin fibers prior to chromatin loops. Figure created by Duong Doan using BioRender.

At the foundation level, each histone octamer is wrapped around 147 bp of nucleotides to form nucleosomes, the basic structural units of chromatin, which are about 10 nanometers in diameter. Nucleosomes are connected by linker DNA to form nucleosome clutches, which are dynamic structures with size and density differences dependent on transcription activity. Nucleosome clutches contribute to the formation of chromatin

fibers, which are further compacted into loops that allow interactions between cis-regulatory elements and promoters (Figure 1.5 and 1.6) (Moraru and Schalch 2019). For example, promoter-promoter loops identified by Hi-C analysis in rice are generally associated with gene coexpression, while H3K27me3-dependent loops are linked to transcription repression found in both rice and *Arabidopsis thaliana*. This illustrates how chromatin looping can mediate both activation and silencing depending on the chromatin context (Domb et al. 2022; Feng et al. 2014). Chromatin loops are arranged in self-interacting domains known as the topologically associating domains (TADs), which have sizes from tens of kilobases to megabases, restricting enhancer-promoter interactions (Figure 1.4). TADs are present in both A and B compartments, although domain boundaries are often more pronounced in transcriptionally active regions. In animals, TADs are formed through cohesin- and CTCF-mediated loop extrusion (Szabo et al. 2019). In contrast, although cohesin complexes are conserved in plants, CTCF homologs are absent. Consequently, TAD-like structures in plants are thought to arise primarily from transcriptional and epigenetic features, such as histone modifications and the activity of specific TFs. For instance, TCP (TEOSINTE BRANCHED 1, CYCLOIDEA, and PCF1) proteins in *Marchantia polymorpha* have been shown to be enriched at TAD borders and associated with TCP1-rich domains, where they regulate gene expression but are dispensable for TAD formation, highlighting a chromatin organization mechanism distinct from CTCF-anchored loop extrusion (Karaaslan et al. 2020). TADs, or TAD-like domains, have been detected in the genomes of rice, maize, and tomato, where their boundaries are enriched in actively expressed genes and euchromatic marks. However, in *Arabidopsis thaliana,* with a small, gene-dense genome with limited repetitive DNA, canonical TADs are not observed. Instead, *Arabidopsis* displays finer-scale chromatin domains and compartmentalization, making it an outlier among plants (Domb et al. 2022; P. Dong et al. 2017).

At a larger scale, chromatin is characterized into A and B compartments, representing active and repressive states, respectively (Figure 1.4). A compartments are enriched for genes and active histone marks, span megabases, and are found to occur at telomeric regions. Meanwhile, B compartments are identified by repressed histone marks and DNA methylation, established in pericentromeric regions (Lieberman-Aiden et al. 2009). In plants, A compartments are found to exhibit greater interchromosomal contacts, such as in rice and tomato; in contrast, B compartments possess intense interchromosomal

contacts in Arabidopsis, which are mediated by H3K9me2 and AGENET DOMAIN CONTAINING PROTEIN 1 (ADCP1) through the mechanism of liquid-liquid phase separation (S. Zhao et al. 2019). The most highly organized level is the chromosome territory, one per nuclear territory occupied by one chromosome. The Rabl configuration, a polarized arrangement with centromeres and telomeres clustered at opposite nuclear poles, is adopted by large-genome species such as wheat, with centromeres and telomeres localized at opposite ends (Concia et al. 2020; Mascher et al. 2017). Whereas *Arabidopsis* has a Rosette-like organization with centromeres at the periphery of the nucleus. Polyploidy, common in plants, alters this topology (M. Wang et al. 2018). Duplication of the Arabidopsis genome in the case of the tetraploid leads to increased interchromosomal interactions and A to B compartmental switching and vice versa, which correlate with gene expression changes (Hui Zhang et al. 2019)

The 3D genome organization and dynamic alterations to the chromatin mediate CRE function during plant growth and response to stress. Potential research for future studies can further clarify 3D genome organization and its effect on CRE function. Functional analyses on TADs, particularly on enhancer specificity and DNA repair, are critical since less is known about them in plants compared to animals. The analysis of the interaction between chromatin and nuclear domains, the nuclear periphery, and the nucleolus can potentially prove the impact of dynamic nuclear shape on CRE accessibility. Single-cell Hi-C, initially established in gametes of the rice plant, has the potential to characterize cell-type-specific chromatin architecture regulating CRE function during plant development (Domb et al. 2022). TEs, which can serve as folding anchor points within genomes, also need to be explored for their impact on organizing CRE-promoter contacts, and particularly in plant genomes with repetitive elements. These advances from the perspective of high-resolution networks of chromatin contact are on the cusp of maximizing the potential to connect 3D genome architecture to transcriptional regulation in plants.

1.2.3. Chromatin Accessibility

Chromatin accessibility is central to the function of CREs, as it provides the requisite spatial context for TFs to bind to DNA and regulate gene expression in plant organisms (Klemm et al. 2019). In the genomes, there is heterogeneous chromatin composition with both euchromatin and heterochromatin, with differing influence on the accessibility of

CREs. Euchromatin is characterized by loosely condensed nucleosomes, rendering it active in transcription and rich in accessible chromatin regions (ACRs) filled with CREs, such as promoters and enhancers. Heterochromatin is characterized by densely condensed nucleosomes, making it repressive to transcription; the compaction obscures CREs and restricts TF access (Figure 1.6a) (Marand et al. 2023). Nucleosome positioning has a variable effect on the accessibility of CREs, making them either receptive to activation or protected to prevent transcription, thereby ensuring the proper regulation of genes in response to developmental and environmental cues (Figure 1.6b).

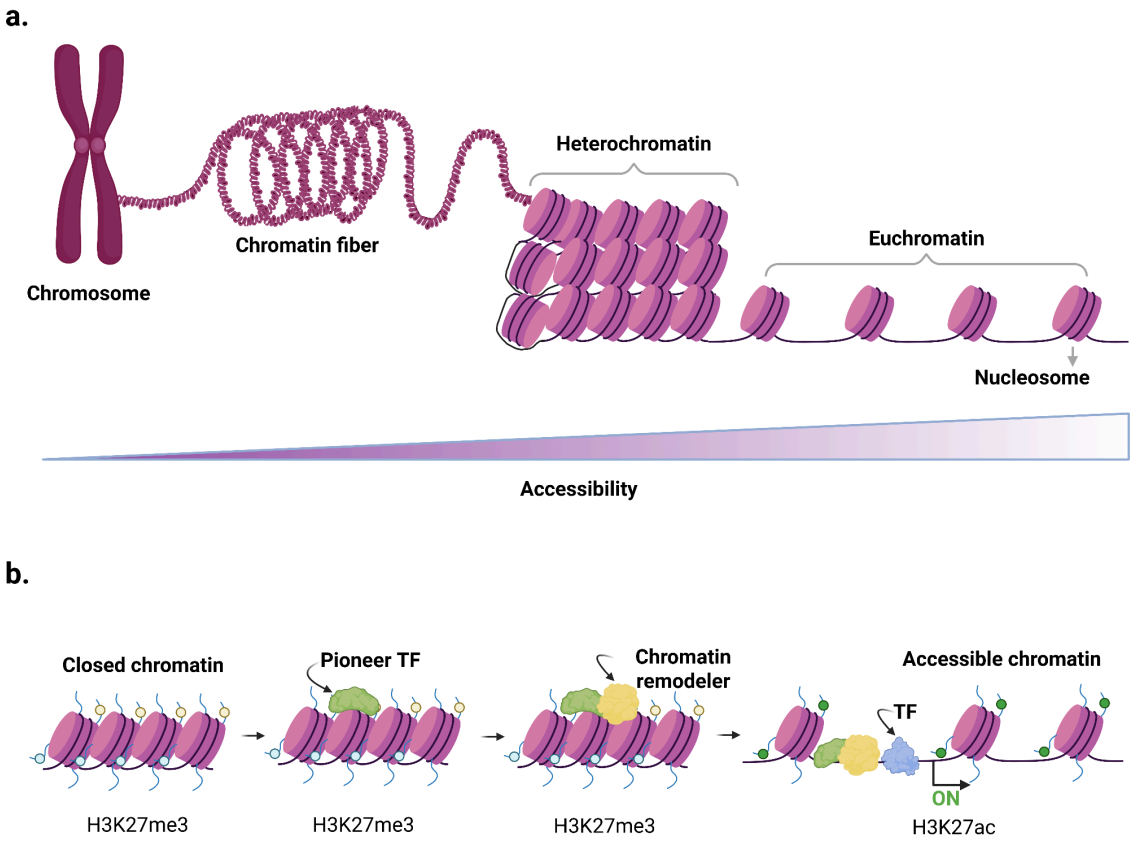


Figure 1.6: Chromatin organization and dynamic regulation of accessibility.
(a) Organizational structure of chromatin from the chromosome level down to the nucleosome level. Chromatin fibers have the ability to condense to heterochromatin regions with closely packed nucleosomes and more extended regions known as euchromatin with higher distances between the nucleosomes. Accessibility for binding factors increases from heterochromatin to euchromatin. (b) Model of chromatin opening via pioneer transcription factors. In the case of closed chromatin and repressive histone marks like H3K27me3, pioneer transcription factors can participate in the binding of DNA

to the nucleosome and the subsequent recruitment of chromatin remodeling enzymes. This results in chromatin opening and the ability of other transcription factors to interact with chromatin for gene activation. Figure created by Duong Doan using BioRender.

Nucleosome positioning impacts the accessibility of chromatin by the exposure of DNA to TF binding. Nucleosomes restrict access when tightly packed, also known as heterochromatin, occluding CREs and repressing transcription. Nucleosome-depleted regions (NDRs) or free regions constitute open chromatin, exposing CREs to TF binding (Klemm et al. 2019). The development of such accessible areas can be initiated by pioneer TFs, a unique set of TFs possessing the capability to bind to their target DNA sequences even in the context of nucleosome-associated, closed chromatin environments. Unlike most TFs requiring pre-existing accessibility, pioneer TFs can directly bind to condensed chromatin by recognizing their target DNA motifs even when partially occluded by nucleosomes. Their unique DNA-binding domains enable this nucleosomal interaction, serving as nucleation sites that recruit chromatin remodelers or histone-modifying enzymes to displace, reposition, or remove nucleosomes (Figure 1.6b) (Mayran and Drouin 2018; Zaret and Carroll 2011). This opens up the chromatin structure initially and prepares the CREs for further transcriptional activation (Chen et al. 2024). In plant development, LEAFY has functional similarity to pioneer factors in switching development by opening up inaccessible regulatory regions to other TFs to allow the initiation of transcriptional pathways (Jin et al. 2021). These pioneering TFs influence chromatin dynamics by establishing novel regulatory landscapes in a tissue- and stage-specific manner. For instance, ACRs in Arabidopsis root tissue are often located in intergenic regions, which contain distal CREs that drive tissue-specific gene expression, with the majority of ACRs positioned away from promoters (Maher et al. 2018). During drought stress, nucleosome repositioning in Arabidopsis renders CREs in stress-regulated genes available for TF binding and gene activation. Non-stressed nucleosome-bound CREs are closed, preventing unwanted gene expression from taking place (Kim et al. 2015).

Open chromatin regions, predominantly in euchromatin, constitute the center of CRE-mediated gene regulation, offering sites for the interaction between TFs and CREs. In maize, ACRs near the transcription start sites have higher concentrations of promoter CREs, whereas intergenic ACRs have distal enhancers facilitating gene control during

kernel development (Ricci et al. 2019). Also, nucleosome-free regions at the promoter CRE are located near high gene expression within rice leaf tissue, where TFs have access to regulatory regions and initiate transcriptional processes (W. Zhang et al. 2012). Taken together, the findings highlight the functional importance of ACRs as regulatory hubs instead of structural components. Relocation or occlusion of CREs due to the chromatin's dynamic mobility within nucleosomes, coupled with the elaborate topology revealed within euchromatin or heterochromatin, acts as a molecular switch to regulate spatial and temporal gene expression within different plant tissues and environments. Hence, chromatin accessibility is an indispensable mechanism component within regulatory DNA to couple genomic sequences with both the environment and development to govern transcriptional processes with great specificity.

1.2.4. Histone Modifications and Variants

While nucleosome positioning and chromatin accessibility shape the physical landscape for CRE activity, the functional condition of the chromatin is also controlled by particular histone modifications and variants that enhance or suppress transcriptional activity. The core nucleosome is made up of the histones H2A, H2B, H3, and H4, which can be replaced by non-canonical variants such as H2A.Z, H2A.X, H2A.W, and H3.3 to alter chromosome stability and gene regulation. H2A.Z, H2A.X, and H3.3 are widely conserved variants that affect nucleosome dynamics, while H2A.W is a plant-specific variant (Talbert and Henikoff 2010; Klemm et al. 2019). H2A.W exclusively marks heterochromatic regions, being absent in genic loci, and participates in the repression of TEs. H2A.Z is present in both heterochromatin and euchromatin regions, with broad implications in transcription regulation, generally promoting nucleosome mobility and transcriptional responsiveness. Meanwhile, H2A.X plays an important role in the DNA damage response. The H3 histone has three variants, including CenH3 is associated with centromeric regions, H3.1 is enriched at pericentromeric heterochromatin regions, and H3.3 is enriched at euchromatin, actively transcribed genes, enhancers, and regulatory elements. The deposition of H3.3 is a landmark of high nucleosome-turnover, transcriptionally active regions, replacing canonical H3.1 during transcription elongation and stress-induced gene activation (Figure 1.7) (Yelagandula et al. 2014; Foroozani et al. 2022).

Histone modifications are types of post-translational modifications, including acetylation, methylation, phosphorylation, ubiquitination, and SUMOylation, that occur largely to the unstructured tails of histone protein molecules. These modifications change the nucleosome structure and interaction between histone and DNA, hence influencing the transcriptional level by regulating chromatin structure and dynamics (Deal and Henikoff 2011). Of all histone modifications, acetylation reduces histone-DNA interactions, hence leading to relaxed chromatin structure and transcriptional activation, while methylation alters transcriptional levels either positively or negatively, depending on the targeted histone residue. The best-known histone modifications occur at the lysines present in histone H3, hence creating a “histone mark” indicating a molecular signal that recruits either a transcriptional activator or a repressor. Active histone marks, including H3K27ac and H3K4me1, are present in open chromatin and active CREs, while repressive histone marks, including H3K27me3 and H3K9me2, are present in compacted, transcriptionally silent chromatin (Klemm et al. 2019).

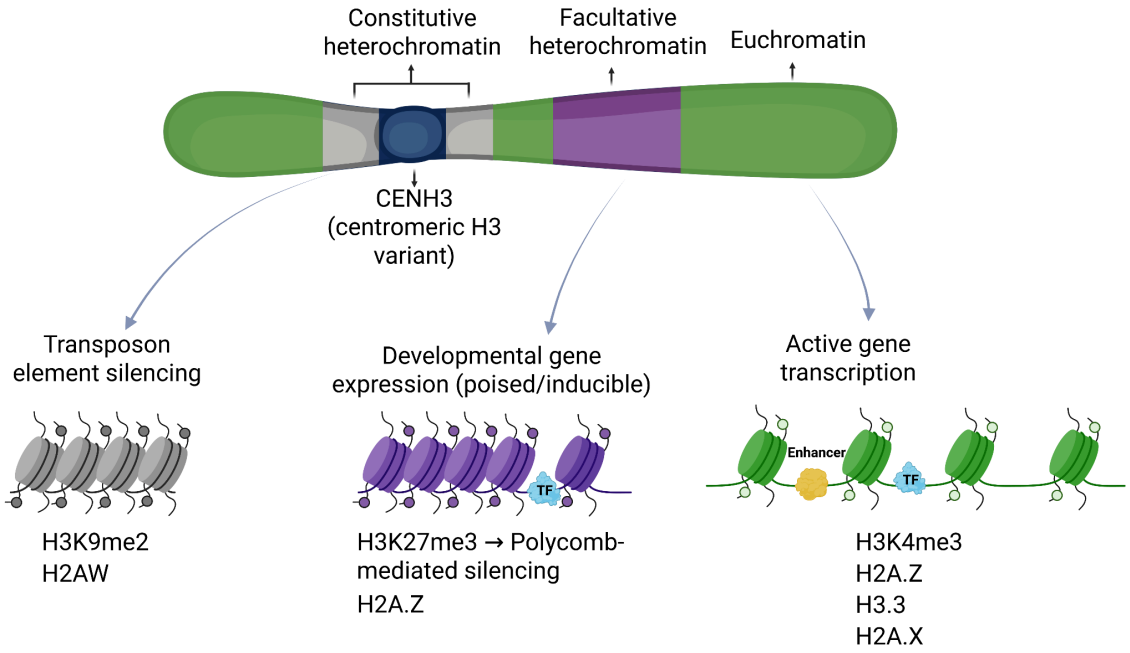


Figure 1.7. Histone variants and post-translational modifications define chromatin states and regulatory activity.

Schematic overview of a plant chromatin organization showing constitutive heterochromatin, facultative heterochromatin, and euchromatin along chromosomes. Constitutive heterochromatin is enriched for H3K9me2 and the histone variant H2A.W and is associated with TE silencing. Facultative heterochromatin is characterized by H3K27me3 and H2A.Z and marks developmentally regulated, inducible genes. Euchromatin contains active genes and enhancers and is associated with H3K4me3 and histone variants H2A.Z, H3.3, and H2A.X. The centromeric region is defined by the H3 variant CENH3. Figure created by Duong Doan using BioRender.

Histone modifications, along with histone variants, create a combinatorial "histone code" that governs the structural and functional state of chromatin. Among these modifications, methylation and acetylation of H3 lysine residues are especially well characterized for their strong correlation with gene activity. H3K4me3 is detected at the 5' ends of genes that are being actively transcribed, and labels promoters and distal CREs, which form chromatin loops that contact active genes. In contrast, H3K27me3 marks repressed genes, their promoters, and facultative heterochromatin regions and is catalyzed by the Polycomb Repressive Complexes. Other acetylation marks, such as H3K9ac and H3K56ac, are frequently detected near transcription start sites and in open chromatin regions, where they facilitate transcriptional activation and enhancer-promoter communication (Marand et al. 2023; Lafos et al. 2011; Foroozani et al. 2022).

In plants, histone variant dynamics are particularly crucial for the rapid responses to the environment and developmental transitions. In contrast to animals, where the incorporation of H2A.Z at promoters is relatively stable, plants deposit and remove H2A.Z in the gene bodies of stress-responsive loci, supporting rapid transcriptional activation under stress (Sura et al. 2017). In Arabidopsis, the co-presence of H2A.Z acetylation and H3K27ac in stress-responsive genes is an indication of a combined process mediated by the NuA4 acetyltransferase complex (Bieluszewski et al. 2022; Charron et al. 2009). Subsequently, H3K27me3, which is deposited by Polycomb Repressive Complex 2, ensures stable gene repression in development (Lafos et al. 2011). Histone variants, in conjunction with histone modifications, provide a complex

chromatin organization that carries molecular information for plants to adopt developmentally, as well as environmentally dependent transcriptional responses.

1.2.5. DNA Methylation

Arising from how nucleosome positioning and histone modification dynamically regulate CRE accessibility, DNA methylation adds a more stable, sequence context-dependent level of control that rigidifies either repressive or permissive chromatin states on both promoters and distant regulatory elements. DNA methylation is the process of adding the methyl group in the C-5 position (carbon number 5 or 5mC) of cytosine bases in the contexts of CG, CHG, and CHH (H can be A, T, or C). It represents an essential epigenetic indicator via altering the DNA structure that helps silence transposable elements and pattern gene expression through control of the availability of CREs. Each methylation context is controlled through distinct pathways. Initiation of de novo methylation is established by DOMAINS REARRANGED METHYLTRANSFERASE 2 (DRM2) through the pathway of RNA-directed DNA methylation (RdDM), which is controlled by small 24-nucleotide RNAs. Meanwhile, maintenance of methylation is carried out through METHYLTRANSFERASE 1 (MET1) for CG contexts, CHROMOMETHYLASE 3 (CMT3) in coordination with histone H3K9 dimethylation (H3K9me2) for CHG contexts, and CHROMOMETHYLASE 2 (CMT2) or DRM2 for CHH methylation (Fang et al. 2022; Huiming Zhang et al. 2018). In contrast, active DNA demethylation is mediated by several bifunctional DNA glycosylases, such as ROS1 (REPRESSOR OF SILENCING 1), DME (DEMETER), DML2, and DML3, which act by excising 5-methylcytosine via a base excision repair pathway. This action negatively regulates overall methyltransferase activity to avoid excessive gene silencing and maintain epigenetic homeostasis. These demethylases have an important role in modulating gene expression, potentially during developmental transitions (e.g., gametogenesis or embryogenesis) or in the presence of environmental and/or abiotic stresses (X. Du et al. 2023). Active demethylation of regulatory regions provides an opportunity to dynamically modulate chromatin accessibility and transcription factor binding to the transcription level, ultimately contributing to transcriptional plasticity in plants.

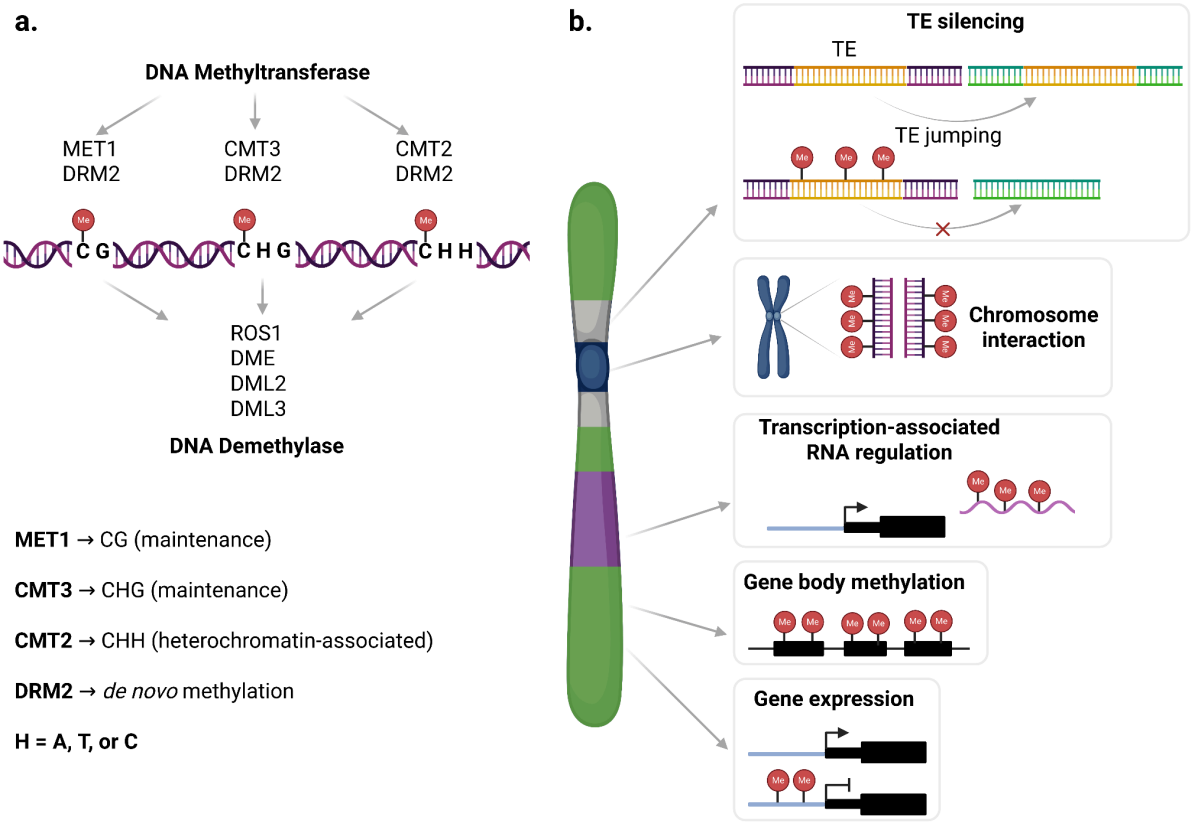


Figure 1.8. Overview of DNA methylation pathways and their regulatory functions in plants.

(a) Cytosine DNA methylation occurs in three sequence contexts: CG, CHG, and CHH (H = A, T, or C). It is established and maintained by distinct DNA methyltransferases. MET1 maintains CG methylation, CMT3 maintains CHG methylation, and CMT2 primarily catalyzes CHH methylation in heterochromatic regions. DRM2 mediates de novo DNA methylation across all sequence contexts through the RNA-directed DNA methylation (RdDM) pathway. Active DNA demethylation is carried out by the glycosylases ROS1, DME, DML2, and DML3. (b) DNA methylation contributes to multiple layers of genome regulation, including transposable element silencing, higher-order chromatin interactions, transcription-associated RNA regulation, modulation of gene expression, and gene body methylation. Together, these pathways shape chromatin organization and transcriptional outcomes across the plant genome. Figure created by Duong Doan using BioRender.

The functional roles played through DNA methylation vary depending on regions of the chromosome: actively transcribed gene promoters are likely to remain unmethylated, hence favoring transcription initiation, whereas distant CREs, which are enhancers, also commonly found in unmethylated regions that enhance their activity. Methylation in promoters is generally associated with downregulation of gene expression, whereas methylation in distal regions, particularly in CHH contexts as well as in CG and CHG sequences, may contribute to enhancer priming or repression depending on developmental and environmental conditions (Huiming Zhang et al. 2018). This methylation landscape is shown to possess a high level of plasticity. During development, site-specific demethylation of some regions allows for TFs to interact with enhancers, enabling processes such as organogenesis, and during abiotic stress, CHH methylation is controlled by regulation at stress-responsive enhancers, as observed in chili pepper, for example, with heat stress triggering hypomethylation for a surge of gene expression (Yang et al. 2025). Furthermore, DNA methylation interacts with histone modification in a complex manner, thus solidifying chromatin states, and is seen with a self-enforcing feedback loop among CHG methylation and H3K9me2, a process regulated by enzymes CMT3 (CHROMOMETHYLASE 3) and SUVH (SU(VAR)3–9 HOMOLOG) (J. Du et al. 2014; Xueqin Li et al. 2018).

In addition, DNA methylation interacts with histone modification cross-talk to stabilize chromatin states, for instance, the CHG methylation-H3K9me2 self-augmenting loop by which the coordination of enzymes CMT3 and SUVH, or the regulation of the RdDM pathway with histone marks for silencing transposable elements, whereas demethylation eliminates repressive marks, including H3K27me3, favoring open chromatin organization (Schmitz et al. 2021). In conclusion, DNA methylation is a critical part of plant genome function, with a dual role in genome integrity and adaptive gene regulation; its context dependence, coupled with interaction with histone modification and sensitivity to developmental and environmental signals, renders it a core component in the highly regulated networks that govern plant growth and adaptation.

1.3. Functional Categories and Mechanisms of Cis-Regulatory Elements

In eukaryotes, with complex genomes and dynamic regulatory requirements, CREs appear as core promoters, enhancers, silencers, and insulators, playing distinct molecular functions. Transcriptional initiation is triggered by core promoters, whereas enhancers act to amplify distal or intronic gene expression; silencers repress

transcription, and insulators mark boundaries to secure regulatory specificity. Below, we will discuss these four types of CREs and examine their mechanisms, focusing particularly on plant-specific signatures and describing comparative findings resulting from animal studies. Limitations of existing studies, particularly those concerning silencer characterization and mechanisms of plant insulators, underscore the need to pursue studies aimed at systematically understanding the richness of CREs in plant genomes.

1.3.1. Core promoter

A promoter is usually located within one to two kilobase pairs away from the transcription start site (TSS) and comprises regulatory elements such as the core promoter, enhancer, and/or silencer elements. Active promoters are usually marked by the presence of ACRs, H3K4me3 histone marks, histone acetylation (HAc), and H2A.Z, but not by methylation of the DNA (Haberle and Stark 2018; Murphy et al. 2020). The repressed promoters, on the other hand, are enriched for H3K27me3, deposited by the Polycomb Repressive Complex 2 (PRC2), and are less accessible (Ngan et al. 2020). The core promoter, whose alternate name is basal promoter, is a well-characterized set of CREs that is immensely attributed to invariant positioning around TSSs. The activity of the core promoter is essential for the initiation of transcription by assembling the transcription preinitiation apparatus comprising general TFs and RNA polymerase II. Depending on the location of TSSs, the core promoters are either focused (sharp/narrow) or dispersed (Figure 1.9). Focused promoters initiate transcription at a single or a few specific nucleotides, which are more common in simple organisms. In contrast, dispersed initiation occurs in most of the genes in eukaryotes and involves multiple start sites spread across a broader region of approximately 50 to 100 base pairs (Juven-Gershon and Kadonaga 2010; Haberle and Stark 2018). They contain some general elements such as the TATA box (position -32), CCAAT-box, TC-elements, the Initiator, and the pyrimidine (Y) patch (specifically in the plant kingdom) (Dolfini et al. 2009; Porto et al. 2014; Yamamoto et al. 2007).

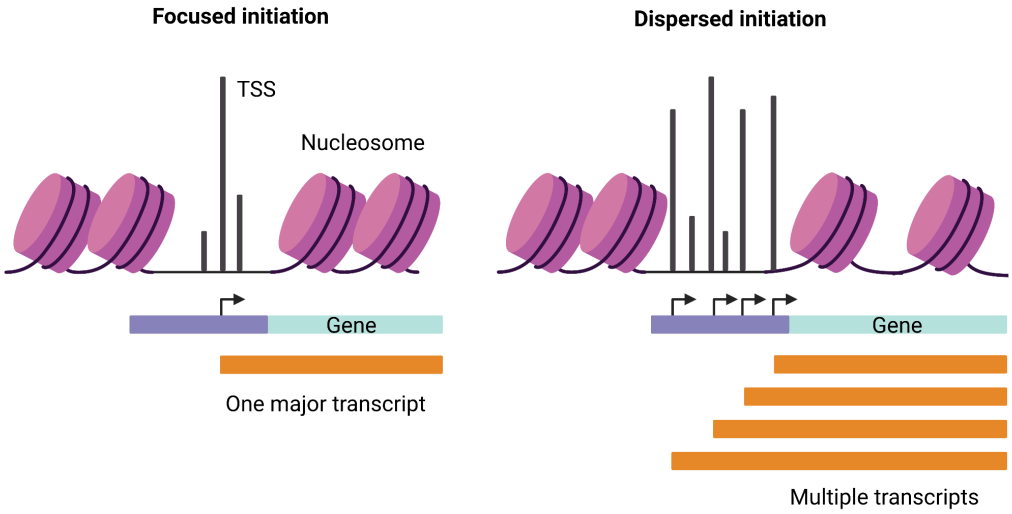


Figure 1.9: Focused and dispersed transcription initiation.

Focused promoters produce one dominant transcript from a single or narrow TSS cluster, whereas dispersed promoters initiate transcription from multiple nearby TSSs, generating diverse transcript isoforms. Figure created by Duong Doan using BioRender.

Multiple methods have been developed to investigate the core promoter and TSSs in mammals and plants, both experimentally and computationally. Cap analysis gene expression sequencing (CAGE-seq) allows detection of promoter activity and reconstruction of transcriptional regulatory networks by capturing and sequencing the 5' end of capped RNAs (Takahashi et al. 2012). RAMPAGE, short for RNA Annotation and Mapping of Promoters for the Analysis of Gene Expression, is a method that offers improvements over CAGE-seq. It helps reduce unwanted background noise from non-capped RNAs, degraded RNA fragments, and internal transcript priming, thus providing clearer detail of TSSs (Batut and Gingeras 2013). Paired-end analysis of TSSs (PEAT) sequences both the 5′ capped end and a downstream region of the same transcript. Inclusion of downstream sequence improves alignment accuracy, helps distinguish genuine TSSs from artifacts, and aids in clearer reconstruction of promoter architecture (Ni et al. 2010).

In maize, the majority of TSSs (87%) belong to the class focus initiation, with 38% of the promoters having a TATA-like motif between positions −35 and −25 relative to the TSS, which may suggest a relationship between TATA box presence and focused transcription

initiation, often associated with stress-responsive or tightly regulated genes (Mejía-Guerra et al. 2015). Also, genes that have broad TSSs tend to have elevated and more constitutive levels of expression across various tissues, implying that such promoters are linked to housekeeping functions. Genes with tissue-specific expression would likely have focused TSSs since they are suggested to perform specialized functions (Mejía-Guerra et al. 2015). However, Arabidopsis promoters exhibit a variety of initiation patterns, including focused, dispersed, and weak (or less pronounced concentration of TSS distributions). Most of them are TATA-free, associated with dispersed initiation sites (Morton et al. 2014). In conclusion, these observations emphasize that while the core promoter is positioned relatively, its sequence characteristics and initiation modes can differ significantly among species, tissues, and expression patterns. This variation reflects the variability of transcriptional control and the vital role of core promoter topology in establishing gene expression patterns.

1.3.2. Enhancers

Enhancers are CREs that enhance the rate of transcription initiation of the target gene by recruiting TFs and cofactor complexes regardless of position or orientation (Figure 1.10A). Unlike core promoters, enhancers have the inability to initiate transcription by themselves. Instead, they function to modulate the frequency of transcriptional bursts to control the transcription rate. The number of mRNA molecules synthesized in every transcriptional event is significantly determined by the core promoter (Larsson et al. 2019). Enhancers can be present in promoter regions or even 1 Mb away from the target gene, with their distribution being genome size dependent (Lu et al. 2019). The three-dimensional (3D) structure of chromatin promotes the generation of chromatin loops that place distal enhancers in close physical proximity to the promoters of their target genes, thereby establishing long-distance transcriptional control, especially in medium to large genomes (Bonev and Cavalli 2016). Consequently, enhancers can regulate non-contiguous genes or interact with several genes, or one gene can be regulated by multiple enhancers. Enhancers can also be found in the intronic regions and regulate the target genes (Meng et al. 2021). For example, the enhancer in the second intron of *AGAMOUS* controls its expression in Arabidopsis floral bud tissue (Sieburth and Meyerowitz 1997). Similarly, the Elongation Factor 1 Beta (eEF-1β) gene has the first intron carrying an enhancer-like element, which plays a crucial role in achieving high expression levels (Gidekel et al. 1996).

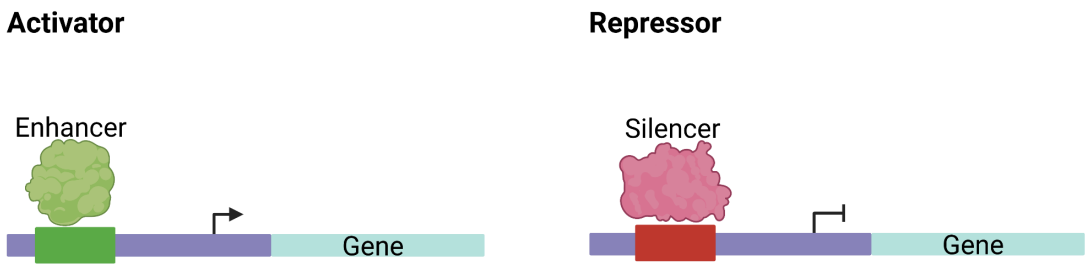


Figure 1.10: Overview of enhancer and silencer with their influence on transcription.
An enhancer recruits an activator protein, which in turn interacts with the promoter's transcription factor binding site to stimulate the associated gene's transcription. A silencer element interacts with a repressor protein that acts on the promoter region, which then reduces the transcriptional outcome. Figure created by Duong Doan using BioRender.

Enhancers can be characterized by distinctive chromatin features that set them apart from promoters and other regulatory elements. Active enhancers in animals are generally marked by histone modifications such as H3K27ac and H3K4me1, which are associated with ACRs (Calo and Wysocka 2013). In contrast, repressed enhancers are enriched with H3K27me3 and typically inaccessible. While plant enhancers share some similarity with animals, they have different signatures. For example, H3K4me1 is a hallmark of mammalian enhancers; it is nearly missing in plant ACRs, yet plant ACRs still contain H2A.Z at the flanking nucleosomes (Ricci et al. 2019). Meanwhile, H3K27ac is widely conserved across angiosperm lineages and usually marks distal enhancers, as not only it does in maize (Lu et al. 2019; Local et al. 2018; W. Yan et al. 2019) but also other species such as rice (Z. Du et al. 2013). It is, however, unknown whether specific enhancers and core promoters are compatible within plant systems. In the future, research could be directed towards the characterization of enhancer classes based on cofactor dependencies, chromatin features, and core promoter preferences. Integration of chromatin conformation tests (e.g., Hi-C) with epigenomic data could give us an understanding of how enhancer-promoter long-range interactions contribute to gene regulation during development and stress response in different plant species.

1.3.3. Silencers

Silencers share similar characteristics with enhancers in most aspects, but function reversely by decreasing or silencing the transcription of target genes (Figure 1.10B). Enhancers are critical in restricting gene activity to specific tissues or developmental stages (R. Liu et al. 2020; Pang and Snyder 2020) and are difficult to identify due to the lack of apparent sequence features and context-dependent activity. In plants, silencers tend to reside within regions of available chromatin and simultaneously contain repressive H3K27me3 modifications because Polycomb-associated regulatory regions can be depleted of nucleosomes despite being transcriptionally repressed (Xiao et al. 2017). Several genome-scale studies have begun to explore the features of chromatin that are associated with silencing activities. For example, in maize, ACRs marking H3K27me3 located upstream of genes are invariably correlated with lowered transcript levels, highlighting the importance of polycomb-mediated suppression (Ricci et al. 2019). Arabidopsis, in like manner, employs polycomb responsive elements that are involved in the recruitment of the Polycomb repressive complex 2, thus allowing targeted H3K27me3 deposition that leads to transcriptional suppression of nearby genes. This suppression requires not only the Polycomb response element sequence but also trans-acting transcription factors (Xiao et al. 2017). However, silencers in mammals are not only associated with H3K27me3 but also with H3K9me3, even at a higher rate (Doni Jayavelu et al. 2020). H3K27me3 in humans can be associated with restrictive histone marks, such as H3K27me3 and H3K9me3, functioning in conjunction, or with active marks, such as H4K20me1, H3K27ac, and H3K4me3, which signify a poised chromatin state (D. Huang et al. 2019). Some silencers are multifunctional and can also act as enhancers in other tissues or at specific developmental stages, as in the case of Drosophila (Gisselbrecht et al. 2020). However, similar multifunctional elements have not been shown to exist in plants. Although there are examples of plant regulatory regions that exhibit condition- or tissue-specific expression, there is no conclusive proof that a single sequence can serve both as a silencer and an enhancer (Schmitz et al. 2021). Such multifunctional elements interact with different sets of transcription factors, including activating and repressing proteins, which points to the complexity of their regulatory functions. As a result, it is unclear that a single chromatin signature has been found that can reliably distinguish silencers from other cis-regulatory modules, which means that chromatin marks may not be sufficient for silencer identification. This suggests that silencer activity is determined by the combinatorial action of transcription

factors together with the local cellular environment, and not by a particular set of stable epigenetic characteristics.

1.3.4. Insulators

Insulator sequences are cis-regulatory components with significant roles in ensuring the proper organization of the genome. Insulator sequences are boundary components that restrict the influence of CREs on nearby genes (Figure 1.11). In contrast to enhancers and silencers, Insulator sequences are location- and orientation-dependent to efficiently confine the activity of cis-regulatory components to specific genomic regions, thus enabling proper expression of genes in different tissues and stages of development (Gaszner and Felsenfeld 2006; Jores et al. 2024). In metazoan genomes, these insulators are generally defined as DNA sequences that are bound by particular proteins, like the CTCF, that interact with the cohesin complex to enforce higher-level chromatin organization. Insulator action mediated through CTCF demarcates the genome into TADs such that CREs will only interact with genes that share the same domain. This inhibits the incorrect activation or suppression of genes, along with the sensitivity of the active genes towards the action of heterochromatin (Oki and Kamakaka 2002; Chetverina et al. 2014). However, it is not the same with plants. Although plants contain higher-level chromatin organization, no apparent CTCF-like homolog exists, which makes one question the existence of traditional insulator components (Kurbidaeva and Purugganan 2021).

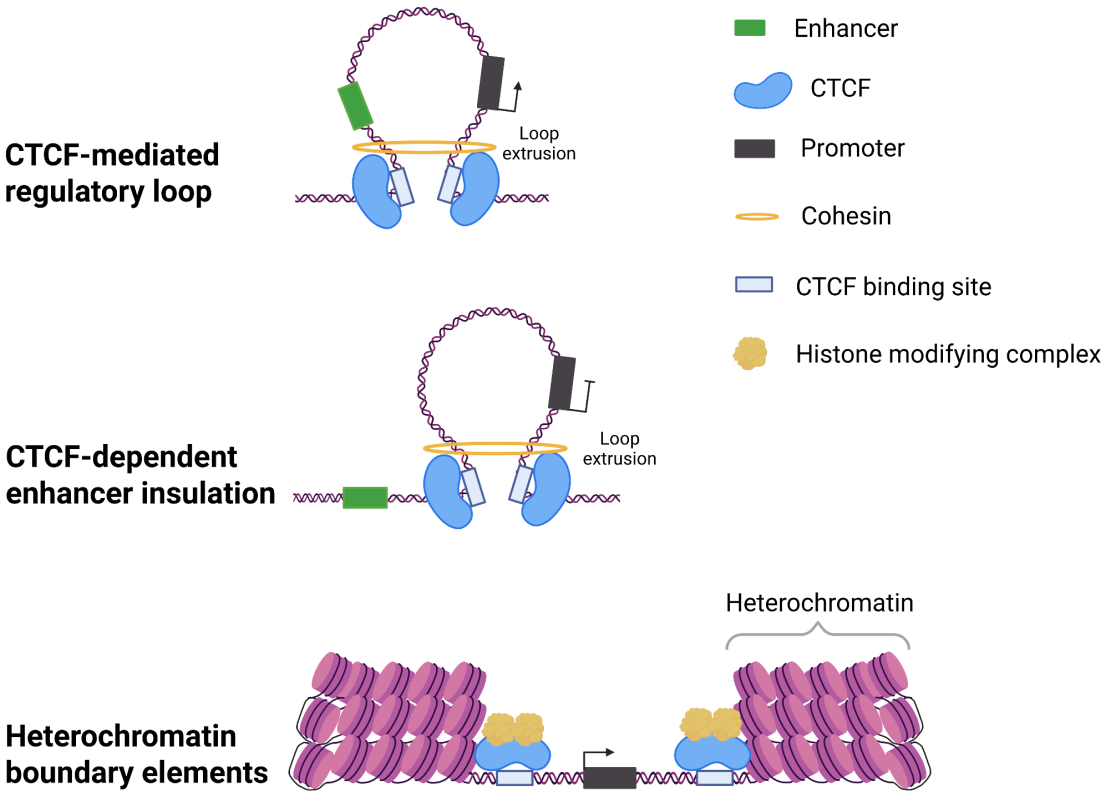


Figure 1.11: Mechanisms of insulator function.

Schematic representation of insulation mechanisms mediated through chromatin structure and chromatin-binding proteins. The top panel portrays an example of a CTCF-mediated regulatory feedback loop, where a convergent CTCF-binding pattern enables a loop anchored through a loop extrusion mediated by a cohesin complex, allowing the enhancer to interact with its target promoter, leading to controlled gene transcription. The middle panel portrays enhancer insulation mediated through CTCF binding, where a CTCF-bound boundary suppresses interaction between a non-target promoter and an enhancer, even as a loop extrusion reaction is ongoing. The bottom panel shows the heterochromatin boundary elements, where the binding sites of the protein CTCF and the histone-modifying complexes form a boundary that prevents the expansion of heterochromatin into the regions that are euchromatic. The colored symbols represent the boundary elements for the enhancers, the promoters, the proteins CTCF, the cohesin complexes, CTCF binding sites, and the histone-modifying complex. Figure created by Duong Doan using BioRender.

Nevertheless, several transcription factor families have been implicated in plant TAD boundary formation based on the motif enrichment profiles of DNA sequences at TAD borders. Although several transcription factors, like the transcription factors of the TCP family (EOSINTE BRANCHED 1 from *Zea mays*, CYCLOIDEA from *Antirrhinum majus, and* PROLIFERATING CELL FACTORS 1 and 2 from *Oryza sativa*), are known to be involved with plant TAD boundary formation, contributing to the process of chromatin organization, confirmed insulator components have not been clearly defined (Kurbidaeva and Purugganan 2021; Y. Liu et al. 2017). Other processes, like the RNA-directed DNA methylation, can mimic insulator activities that are responsible for the conservation of local boundaries between euchromatin and heterochromatin to maintain proper expression of the genes (Q. Li et al. 2015). A new study has characterized short native and synthetic plant DNA fragments with enhancer-blocking properties, which behave similarly to classical insulators. The elements suppress full and partial strong enhancer function if located between a promoter and an enhancer. Interestingly, the identical sequences behave as silencers if paired with weak enhancers, suppressing transcription position-independently (Jores et al. 2024). Context-dependent behavior is supportive of the possibility that the elements play dual regulatory functions, or that they represent a novel mode of transcription regulation unique to plants. In contrast to systems within animals, where the function of the insulators is typically mediated by proteins binding the DNA, no evidence for a requirement for binding proteins was obtained for the plant sequences reported here. Instead, inherent characteristics of the DNA, namely sequence polarity and GC content, appear to affect their behavior (Jores et al. 2024). These findings suggest a potentially dissimilar plant-specific mode of action for the insulators and confirm the need for a holistic analysis to determine if plants employ a radically different molecular scaffold to establish boundaries within the chromatin and to define interactions between cis-regulatory elements.

1.4. Methods and Technologies to Identify CREs

CREs are often found in regions of open chromatin, where transcription factors and other regulatory proteins can bind to DNA. Over the past twenty years, next-generation sequencing technologies have made it possible to genome-wide identify these regions. These techniques combine molecular enrichment with sequencing. DNA fragments that are associated with a specific feature are isolated, converted into sequencing libraries,

and aligned to a reference genome. The pattern of sequencing reads is then used to predict regulatory regions. Various assays are used to detect different regulatory features. Protein-focused assays detect transcription factor binding or histone modifications. Accessibility assays detect nucleosome-free regions. Other assays detect transcription start sites, DNA methylation, or chromatin interactions. Each assay tests a specific molecular feature. The choice of assay depends on the biological question and whether in vivo chromatin information or in vitro binding specificity is needed. In plants, these sequencing-based approaches have produced high-resolution maps of regulatory regions across tissues and species. The next sections discuss the major experimental approaches, their underlying principles, and their limitations.

1.4.1. Protein–DNA Binding Assays

1.4.1.1. ChIP-seq

Chromatin Immunoprecipitation Sequencing (ChIP-seq) is a well-known technique for mapping the genome-wide binding sites of DNA-bound proteins, including TFs and histone modifications (Figure 1.12). Originally developed in the 1980s for the examination of DNA–protein interactions (Solomon et al. 1988) and adapted for high-throughput sequencing in 2007 (Johnson et al. 2007). ChIP-seq facilitates the precise identification of genomic sites in close proximity to TFs at a resolution of about 50–200 bp (Furey 2012). The procedure involves crosslinking proteins on DNA within native chromatin, fragmenting the chromatin, performing immunoprecipitation using the antibody against the target protein, and sequencing the purified DNA fragments for identification of genome-wide binding profiles. In Drosophila, ChIP-seq was used in early embryos, the results revealed a hierarchical organization of chromatin accessibility: a pioneer TF, Zelda, opens the chromatin based on motif affinity, while patterning TFs like Dorsal and Bicoid add additional accessibility within a context-dependent fashion after the activation of enhancers (Brennan et al. 2023). As part of the ENCODE project, ChIP-seq was one of the techniques to map 208 chromatin-associated proteins (CAPs) in a human liver cancer cell line (HepG2), identifying their binding to promoters and enhancers, discovering 17 novel motif sequences, and highly occupied regions (Partridge et al. 2020).

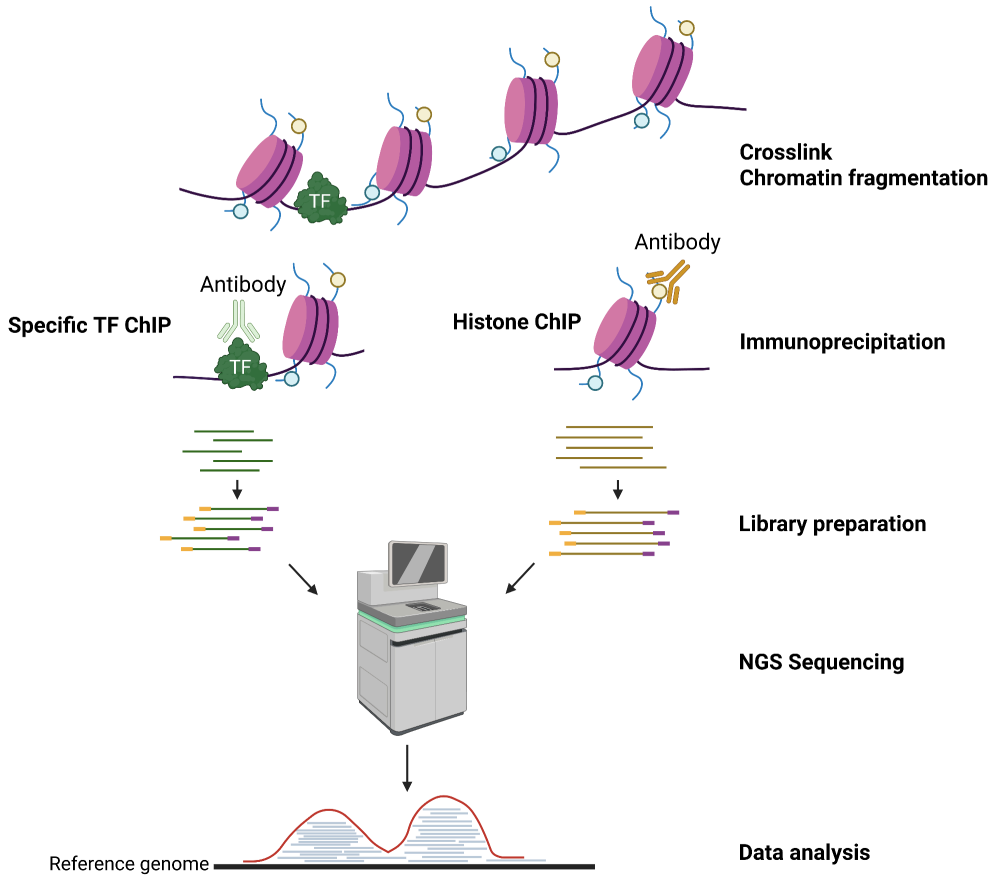


Figure 1.12: Schematic representation of ChIP-seq workflow.
Chromatin is cross-linked, fragmented, and immunoprecipitated with specific antibodies. DNA fragments are purified and undergo end repair and adapter ligation before sequencing. Finally, sequencing reads are mapped to the reference genome for further downstream analysis. Figure created by Duong Doan using BioRender.

In plants, ChIP-seq has been applied in uncovering the molecular regulation networks underlying development and stress responses. The first ChIP-seq study in maize was performed in 2012, defining the cistrome of KNOTTED1 (KN1) TF in inflorescence. KN1 binds to several thousand loci, including 643 genes modulated across various tissues, with a strong enrichment for TFs and genes involved in hormonal pathways, particularly auxin (Bolduc et al. 2012). Recently, a large-scale ChIP-seq was performed to profile 104 TFs in maize leaves, revealing 149,013 binding sites, generating a comprehensive transcription regulatory network that encompasses 77% of expressed genes (Tu et al. 2020). In Arabidopsis, the role of TF WRKY33 under submergence stress was investigated, revealing 104 WRKY33-binding genes that are involved in

oxidation-reduction, programmed cell death, and lipid biosynthesis, among others, and also discovering a new binding motif (J. Zhang et al. 2021). Despite this, ChIP-seq generally captures stable protein-DNA interactions and may miss transient or weak bindings that are crucial for rapid regulatory responses. The need for specific, high-quality antibodies also limits the use of this method within a wide range of TFs, particularly within non-model organisms.

1.4.1.2. DAP-seq

To overcome the limitations of ChIP-seq and develop a more comprehensive understanding of transcriptionally regulated mechanisms, alternative methods such as DNA Affinity Purification Sequencing (DAP-seq) have been developed. DAP-seq is a high-throughput, in vitro tool for CRE motif validation by mapping binding sites for TFs throughout plant genomes. The protocol consists of expressing TFs tagged for affinity in vitro, coating the sheared genomic library of DNA, purifying captured DNA fragments, and sequencing to determine binding sites at ~50–100 bp resolution. To overcome the limitations of antibody-dependent approaches such as ChIP-seq, the scalable design of DAP-seq enabled profiling of 529 Arabidopsis TFs to produce a complete cistrome identifying TF binding affinities and DNA methylation sensitivities (O'Malley et al. 2016), or demonstrating the control of binding site specificity of TFs and their insights towards epigenetic regulation by DNA methylation (Bartlett et al. 2017a). DAP-seq was also applied in maize to target 14 AUXIN RESPONSE FACTORS (ARFs), associating binding sites with kernel development and developmental traits important for crop breeding (Galli et al. 2018). In addition, DAP-seq was combined with RNA-seq to identify ZmMYC2 binding sites and 3,183 potential target genes involved in plant growth and stress response in maize (L. Liu et al. 2024).

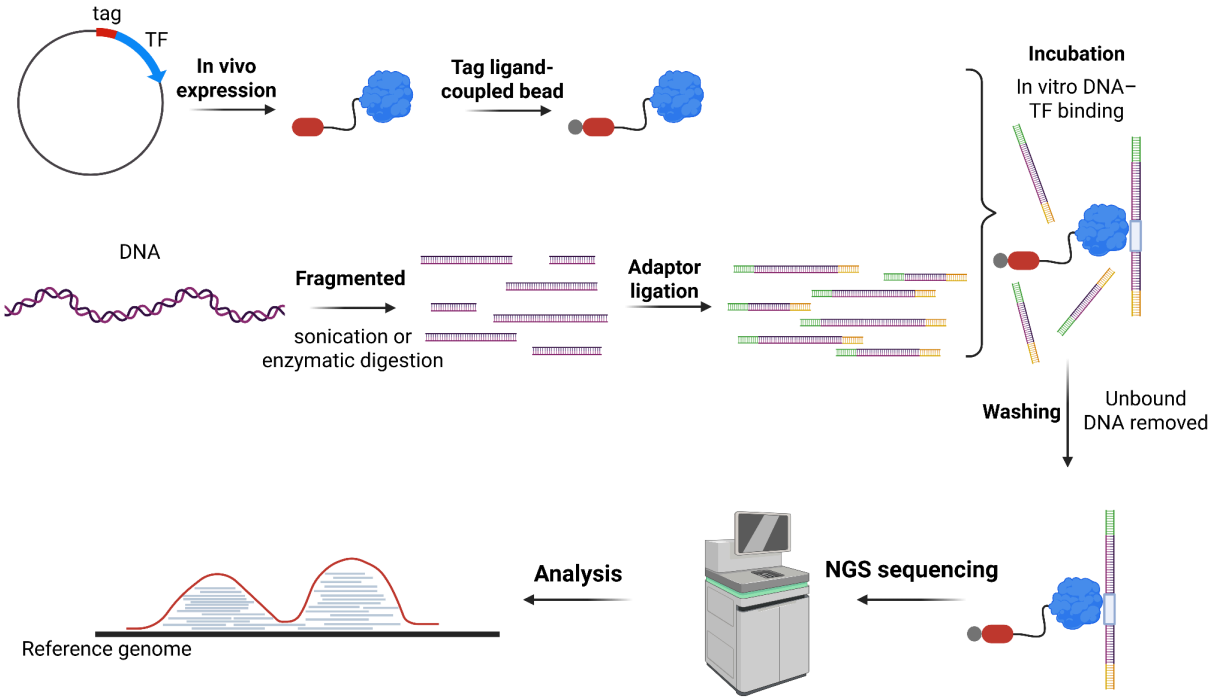


Figure 1.13: Schematic workflow of DAP-seq.

A TF is expressed in vivo as a tagged protein and captured by a tag ligand–coupled bead. Genomic DNA is fragmented either by sonication or enzymatic digestion and then ligated to sequencing adapters. Adapter-ligated DNA fragments are incubated with the bead-bound TF to allow in vitro DNA–TF binding. Unbound DNA fragments are washed away, while TF-bound DNA is recovered and subjected to next-generation sequencing. Sequencing reads are mapped to a reference genome in order to determine TF binding sites. Figure created by Duong Doan using BioRender.

Despite its benefits, DAP-seq is significantly limited by its in vitro approach. The assay neglects important chromatin contexts, including nucleosomes, histone modifications, and higher-level chromatin structures, which are vital in determining TF binding in their native environments (Bartlett et al. 2017b). As a result, DAP-seq has the potential to provide non-functional binding events. Employing the bioassay in vitro, such regions can generate confusing binding outcomes reflecting biologically irrelevant events in real chromatin contexts. The regions, when tested under in vitro conditions, can provide misleading binding activities that are unrepresentative of biologically pertinent interactions present within the natively existing chromatin context. Additionally, the

reliance of DAP-seq on individual TFs fails to capture the more complex binding mechanisms involved in multimeric TF complex interactions (M. Li et al. 2023). It may fail to identify transcription factors that bind using heteromultimeric interactions, hence limiting the range of interactions open to examination. Regardless of these limitations, nonetheless, DAP-seq remains a crucial tool for dissecting the topology of transcription factor binding within plants, providing front-row insights into gene expression regulation as well as epigenetic mechanisms. However, these deficiencies highlight the need for complementary approaches, such as in vivo footprinting, to access the intricacies of TF-DNA interactions within their endogenous contexts (Bartlett et al. 2017b).

1.4.1.3. CUT&RUN

CUT&RUN (Cleavage Under Targets and Release Using Nuclease) is an in situ chromatin profiling assay that uses an antibody to crosslink a protein A/G–MNase fusion (pAG–MNase) to intact chromatin, for effective next-to-target cleavage and release of DNA into solution for sequencing without crosslinking, with high resolution and minimal background (Figure 1.14). Cells or nuclei, which are permeabilized with digitonin, are first bound by Concanavalin A beads, incubated with target-specific antibody, and then washed to remove excess antibody, before Ca^{2+} addition for target-bound release of DNA into solution for preparation of the sequencing library (Skene and Henikoff 2017). In contrast with ChIP-seq, CUT&RUN only needs 5,000–250,000 cells and up to ten-fold less sequencing depth, is compatible with low-input and rare cell populations, and is designed for a streamlined workflow, from sample to data within less than four days to minimize experimental variability and epitope blocking, with allowance for spike-in normalization for quantitative comparison. CUT&RUN is an addition to chromatin assays such as CUT&Tag and ChIP-seq with greater signal-to-noise ratios, reduced input, and compatibility with a range of sample types, including fresh, frozen, or FACS-isolated cells (Kaya-Okur et al. 2019; Skene and Henikoff 2017).

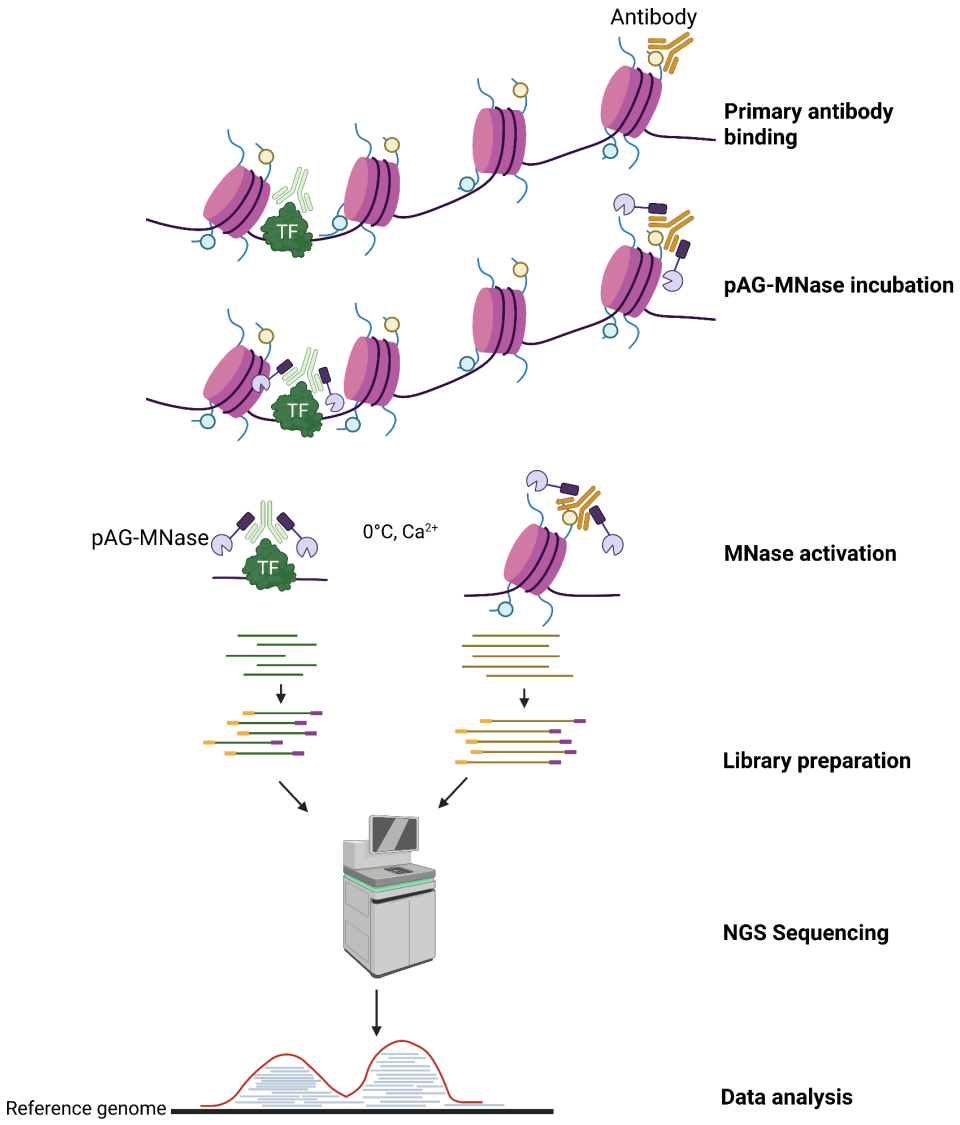


Figure 1.14: CUT&RUN process diagram.

Antibodies target a particular protein or histone modification directly within the native chromatin. A fusion protein of Protein A and micrococcal nuclease (pA-MN) attaches to the antibody, placing the MNase near the target location. Addition of Ca^{2+} triggers the cleavage of the DNA at the target region, producing short DNA fragments that can then be purified and used to create a library for sequencing. Reads are mapped to the reference genome. Figure created by Duong Doan using BioRender.

Leveraging methodology strength, CUT&RUN has been used extensively in plant and mammalian contexts for mapping chromatin features with base-pair resolution from minimal input. In mammals, CUT&RUN has mapped histone marks such as H3K4me3

and H3K27me3, or TFs such as CTCF, with 100,000 cells and with >10-fold lower sequencing depth than ChIP-seq (Womersley et al. 2025; Skene and Henikoff 2017). Versions with ultra-low input (uliCUT&RUN) enabled mapping of pluripotency factors within single mouse embryonic stem cells and pre-implantation embryos, with true single-cell sensitivity, and automated, high-throughput methods extended CUT&RUN to rare or clinical samples (Hainer et al. 2019). In plants, CUT&RUN was applied on FACS-purified Arabidopsis endosperm nuclei using only ~20,000 nuclei, compared with INTACT-ChIP's ~500,000 nuclei to generate parental-specific H3K27me3 landscapes and disclose maternal imprinting patterns (Hui Zhang et al. 2019). Subsequently, CUT&RUN has been applied to gymnosperms and monocots with success. In Norway spruce, Nakamura et al. applied CUT&RUN on 5,000–20,000 embryonic callus nuclei to generate genome-wide profiles of H3K27me3, and showed that embryogenic potential is linked with low basal H3K27me3 values, and that repressive remodelings by PRC2 are a biomarker for somatic embryo induction (Nakamura et al. 2020). In rice, Guo et al. applied CUT&RUN alongside ChIP-qPCR mapping of bHLH factor OsbHLH98, at the promoter of OsBUL1, within lamina joint tissue, demonstrating direct repression of transcription underlying leaf angle control at base-pair resolution from small nuclei preparations (Guo et al. 2021). By delivering high-resolution, low-background data from minimal input and straightforward procedures, CUT&RUN strengthens ChIP-seq and chromatin accessibility methods such as ATAC-seq, with a full toolkit for dissecting epigenomic landscapes. Despite these benefits, there are still several shortcomings of the CUT&RUN approach that should be considered. As with other antibody-based techniques, the quality and affinity of the antibody used could result in low sensitivity or the preferential cleavage of irrelevant targets (Firestone et al. 2024). Because the CUT&RUN assay directly targets native chromatin without cross-linking, maybe the detection of transient or low-affinity protein-DNA interactions could lead to biases with this assay, although this may vary based on the compactness of chromatin domains (Miller et al. 2023). The activation of pA-MNase could also require careful optimization because its preference for accessible chromatin and A/T-rich sequences could lead the assay to disproportionately assess highly accessible targets. Furthermore, the nuclear permeability and chromatin accessibility might vary greatly across different tissues or during development (Nordin et al. 2023).

1.4.1.4. MOA-seq

To improve the accuracy of CRE mapping, a new variant of MNase-seq, known as MNase-defined cistrome-Occupancy Analysis (MOA-seq), has been established as a recent method for identifying TF binding sites in ACRs at high spatial resolution (Figure 2.1). MOA-seq exploits MNase as a micrococcal nuclease with endonucleolytic and exonucleolytic activities to purify short fragments of DNA, roughly 30 base pairs long, that are shielded by protein-DNA contacts; this allows genome-wide detection of potential TF footprints without recourse to TF-specific antibodies. A systematic study of developing maize earshoot demonstrated that MOA-seq defined a cistrome consisting of 145,000 MOA footprints (MFs), making up 0.55% of the genome and bound to 215 motif families relevant to TF sites (Savadel et al. 2021). The MFs were significantly enriched at transcription start sites and termination sites, appeared in 47% of flanking genes, and 31% intergenic, thus suggesting differential patterns of TF binding. In comparative analyses, 76% of ATAC-seq peaks were found to coincide with MFs, with only 35% of MFs overlapping with ATAC-seq peaks, an indication that MOA-seq reveals several novel regulatory regions that ATAC-seq fails to identify. MOA-seq also displayed high concordance with DNS-seq (Differential Nuclease Sensitivity-sequencing), as 59% of MOA-seq peaks were discovered to concur with DNS-seq peaks, with higher accuracy because of shorter footprint lengths. Confirmation with ChIP-seq data for the TFs FASCIATED EAR4 (FEA4), KNOTTED1 (KN1), and TEOSINTE BRANCHED1 (TB1) established that 82–93% of sites for these proteins functioning as TFs were in line with MFs, with MOA-seq motifs om015 and dym33 showing strong homologies with FEA4's canonical motif and TB1's canonical motif, respectively.

Extending the application of MOA-seq, a recent study utilized haplotype-specific MOA-seq in 25 maize F1 hybrids to construct a leaf pan-cistrome under well-watered and drought conditions, identifying ~210,000 (epi)genetic variants, termed binding quantitative trait loci (bQTL), linked to TF occupancy (Engelhorn et al. 2025). Through hybrid analysis with a shared B73 progenitor, this study successfully minimized trans-effects and discovered 327,029 MOA footprints in B73xMo17 hybrids and also unveiled 48,505 allele-specific MOA polymorphisms (AMPs) with strong allelic bias. These bQTL reported overlap with previously defined regulatory loci such as VGT1, ZmTRE1, and a nearby MITE transposon lying nearby ZmNAC111, thus capturing most heritability-based variation for roughly 72% of the 143 traits studied in maize's Nested Association Mapping (NAM) panel. For drought stress conditions, MOA-seq outlined

3,629 high-confidence bQTL lying near drought-responsive genes like ZmNAC111 and ZmTINY, with haplotype-specific footprints correlating with levels of transcript abundance and drought tolerance. This methodological strategy highlights the versatility with which MOA-seq can span different genomes as well as environments, complementing conclusions previously drawn by Savadel et al. (2021) on tissue specificity. By combining high-throughput sequencing with a refined computational platform, MOA-seq enhances TF binding site predictive accuracy and combines TF-specific information with extensive chromatin accessibility profiles assembled from ATAC-seq and DNase-seq data, culminating in forming a finely detailed atlas of regulatory elements that enriches our view of gene regulation within complex genomes like maize. While MOA-seq offers advantages, a prerequisite for optimal outcome data is the species- and tissue-specific optimization of MNase digestion. Beyond this necessary step, the method has limitations: it is unable to distinguish the specific protein type (TFs, cofactors, or polymerase) or determine the functional role of identified regulatory elements, such as classifying a footprint as an enhancer, silencer, or another element. The use of bulk tissues yields an averaged signal, which can obscure cell-type-specific interactions and may cause low-affinity or transient TF binding events to be missed.

1.4.2. Chromatin Accessibility Assays

1.4.2.1. DNase-seq

DNase-seq is one of the earliest and most widely applied methods for evaluating chromatin accessibility, thus shedding valuable insight into the regulatory architectures of plant genomes. The technique is based on the idea that DNase I is preferentially active on nucleosome-depleted regions, often associated with active cis-regulatory elements. With the sequencing of the ends of DNase I-digested fragments, genome-scale representations of accessible chromatin can be generated.

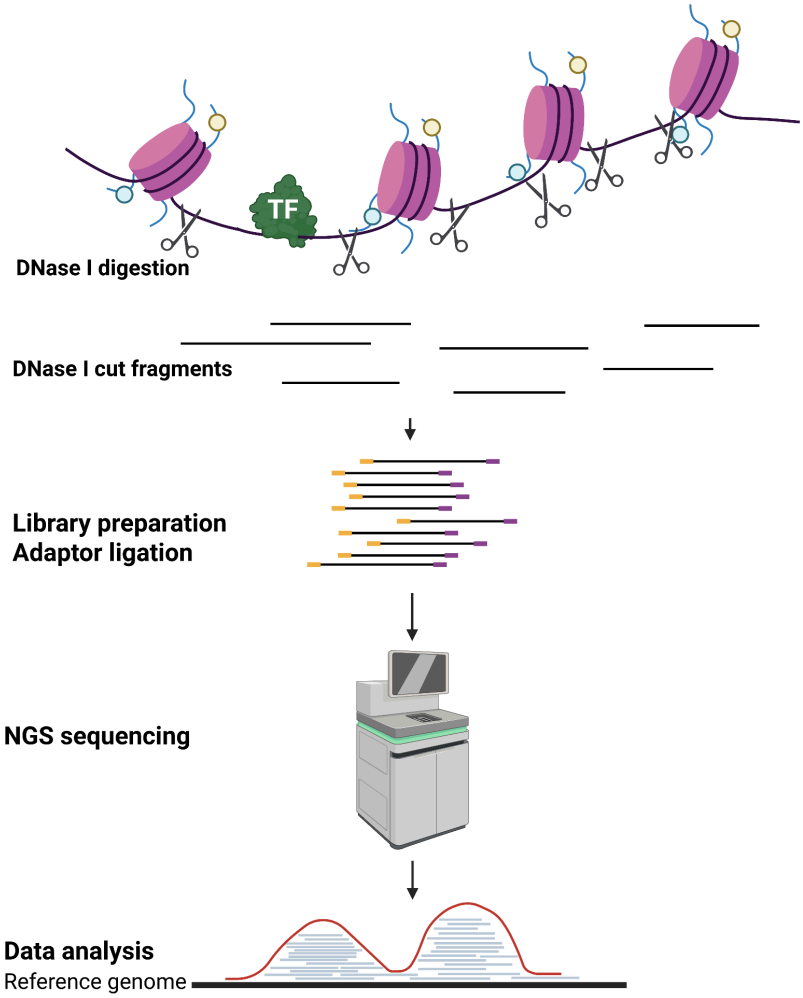


Figure 1.15: Overview of the DNase-seq protocol.
Intact nuclei were treated with DNase I to preferentially cleave linker DNA in accessible chromatin, creating short fragments. Adapters were ligated to the ends of the fragments during library preparation before sequencing. Figure created by Duong Doan using BioRender.

In plant research, DNase-seq has helped clarify regulatory elements and locate TF footprints, short regions that are resistant to DNase I cleavage due to TF binding (C. Liu et al. 2017). For example, DNase-seq has already validated reported TF binding regions on Arabidopsis, including SEP3 and AP1 association ones, while at the same time enabling the detection of many potential TF footprints (W. Zhang et al. 2012). A genome-wide analysis with DNase-seq in the rice genome has revealed drought-induced

changes in chromatin accessibility, enabling the discovery of regulatory regions and transcription factor (TF) footprints of stress-responsive genes (Rajkumar et al. 2024). In addition, in a study of cold stress response in grass species, 93 putative TF binding motifs were identified, including 60 matching known motifs such as those for CBF1, CRF4 (CYTOKININ RESPONSE FACTOR4), and ERF105. Notably, 17 TFs, including SHINE3, ERF2, and ERF9, were consistently detected across all cold-treated tissues and species, suggesting a conserved core of cold-responsive regulators. (Han et al. 2020). Nevertheless, the analysis of DNase-seq footprints requires careful accounting for the sequence bias inherent to the enzyme, since the generated artifacts can hide true binding signals. Despite these technical challenges and the large amount of material involved, DNase-seq remains one of the vital tools for characterizing the regulatory architecture within plant systems (Tsompana and Buck 2014). Recent breakthroughs in modifying this methodology for single-cell resolution may hold significant potential for advancing the understanding of chromatin organization with greater spatial and cellular resolution (Gao et al. 2021).

1.4.2.2. ATAC-seq

The Assay for Transposase-Accessible Chromatin using sequencing (ATAC-seq) is an advanced technique used to describe open chromatin areas and define CREs with high sensitivity and efficacy (Buenrostro et al. 2013). This approach utilizes an active Tn5 transposase that simultaneously cuts DNA while adding sequencing adapters to accessible chromatin, eliminating the need for enzymatic digestion, as well as downstream ligation reactions previously necessary in older techniques such as DNase-seq (Adey et al. 2010; Grandi et al. 2022). Another advantage of ATAC-seq is that it requires very little material in terms of starting material, from 500 to 50,000 nuclei, making it very useful for profiling scarce cell populations or limited tissue samples. For example, it has been used in 23 different human tumor cell types and to identify regulatory elements associated with autoimmune diseases and cancer (Corces et al. 2018). In plant tissues, ATAC-seq has been used to create high-resolution accessibility maps across development stages as well as cell populations. The application of ATAC-seq in *Arabidopsis* root epidermal cells has revealed cell-type-specific ACRs, most of which showed correlations to both transcriptional activation and repression (Maher et al. 2018). In addition, wide-scale deep ATAC-seq profiling revealed over 125,000 transcription factor footprints, with around 72% being tissue-specific (Lu et al. 2017).

These discovered motifs included biologically relevant elements such as FACTOR1/PHYTOCHROME INTERACTING FACTOR (PIF) -binding motifs in seedlings and WRKY motifs in root cells. Regardless of its advantages, data analysis in ATAC-seq is complicated by Tn5 sequence bias, but this can be resolved by applying suitable normalization methods.

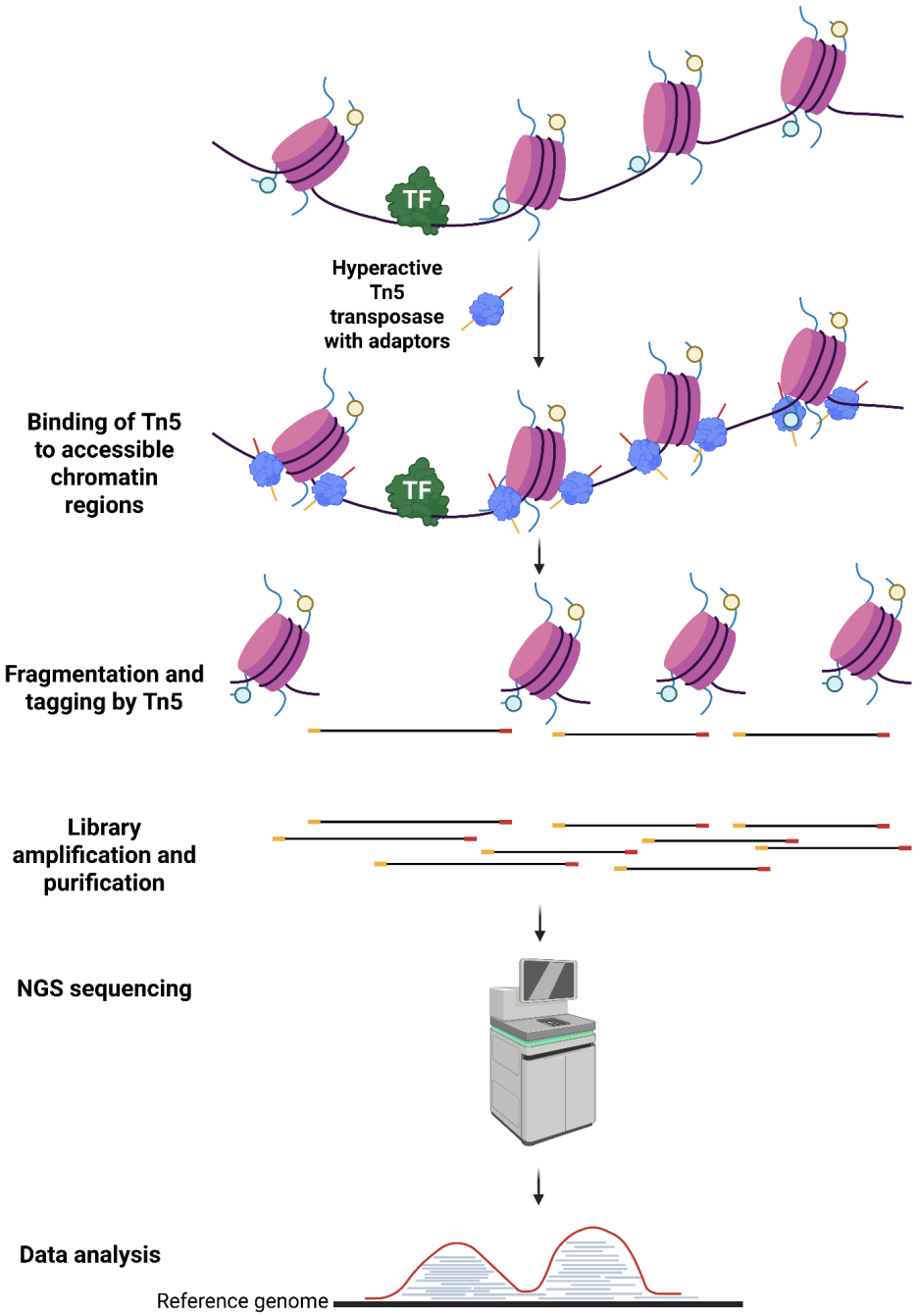


Figure 1.16: Schematic workflow of ATAC-seq.
In ATAC-seq, hyperactive Tn5 transposase pre-loaded with a pair of sequencing adapters shows preferential binding towards chromatin regions that are either depleted of nucleosomes or lie at their edges. The binding of Tn5 provokes simultaneous cutting of the DNA and attachment of adapters, referred to as tagmentation. The adapter-tagged fragments are purified and undergo PCR to generate a sequencing library. After the high-throughput sequencing, the data is analyzed by mapping the reads to a reference genome, identifying open chromatin regions and, by inference, regulatory elements like promoters and enhancers. Figure created by Duong Doan using BioRender.

While bulk ATAC-seq has provided important insight into the overall chromatin landscapes of tissues, it combines the signals from a variety of cell types, making it challenging to identify individual cell differences. To address this issue and better understand the unique regulatory feature of each cell, single-cell ATAC-seq (scATAC-seq) was first developed in mammals, enabling chromatin accessibility profiling at better resolution (Buenrostro et al. 2013). Plants with a unique cell wall and the difficulty of isolating homogeneous cell types make it challenging to adapt scATAC-seq. Two main methodologies, which are droplet-based microfluidics and combinatorial indexing (sciATAC-seq), have enabled high-resolution mapping of chromatin in diverse plant cells. In *Arabidopsis thaliana*, scATAC-seq has proven to have the capability to predict gene expression from accessibility profiles in the chromatin and can be used to reconstruct root tissue development trajectories (Dorrity et al. 2021). Comparative analysis of scATAC-seq data showed both conserved as well as species-specific dynamics in transcription factors in the phloem companion cell maturation in Arabidopsis and maize (H. Yan et al. 2025). Mitigating cost as well as scalability concerns, Tu et al. (2023) outlined an optimized sciATAC-seq protocol tailored for plant tissues, observing that it attains higher signal-to-noise ratios at lower costs compared to commercial droplet-based systems (Tu et al. 2022). Parallel to such experimental breakthroughs, recent reviews have highlighted persistent computational as well as biological challenges in the field of plant single-cell epigenomics, emphasizing the need for more efficient methodologies to define complex regulatory landscapes at the single-cell level (Mendieta et al. 2023). These efforts are improving our ability to study regulation in plants at a resolution previously thought infeasible. In conclusion, ATAC-seq and scATAC-seq are appealing tools for mapping plant regulatory landscapes to examine transcription factor

occupancy as well as accessibility to chromatin using very little material with efficient methodologies.

1.4.2.3. MNase-seq

In addition to the ongoing effort to profile ACRs at a genome-wide scale, mapping of nucleosomes is still fundamental to the understanding of the structural underpinnings of gene regulation. Nucleosome occupancy and positioning direct the accessibility of cis-regulatory factors and are able to favor or inhibit TF binding. Among the earliest and widely popular techniques to quantify the structural make-up of nucleosomes is micrococcal nuclease sequencing (MNase-seq), which leverages the low molecular mass (17 kDa) and dual nuclease activity of MNase to degrade linker DNA between nucleosomes to yield mono- and poly-nucleosomal fragments suitable to be sequenced.

In a genome-wide MNase-seq study, the occupancy and structural organization of nucleosomes were compared between Arabidopsis and rice, and evidence was presented to show that patterns of nucleosome positioning around regulatory elements are conserved and that there is a substantial correlation between depletion of nucleosomes and TSSs (T. Zhang et al. 2015). In the same fashion, MNase-seq was used to explore the reorganization of nucleosomes in cold-stressed maize. The results demonstrated that loose nucleosomes were largely associated with active regulatory regions and gene bodies. They were less likely to be found in heterochromatic regions and regions with repressive epigenetic markers. The study highlighted nucleosome sensitivity as an important factor in controlling chromatin state in response to environmental stress. (R. Huang et al. 2024). To improve the resolution as well as interpretability of data, and minimize the technical bias from MNase digestion, several studies have sought to simplify both experimental and computational procedures. A quantitative MNase-seq approach was developed by measuring nucleosome occupancy in different chromatin states using spike-in normalization and kinetic modeling, which allows for a more accurate characterization of nucleosome dynamics within genes' regulatory regions (Chereji et al. 2019). Recently, a bioinformatic MNase-seq analysis pipeline called nucleosome-based MNase accessibility (nucMACC) was developed. It can process raw sequencing data from light to moderate or heavy MNase digestions, which capture accessible and dense nucleosomes, respectively, to map genome-wide nucleosome positions and quantify their accessibility and stability. Thus, it identifies structurally altered nucleosomes with high precision. The nucMACC significantly

advances chromatin research via offering a scalable, reproducible platform for high-resolution nucleosome analysis without extensive digestion titration (Wernig-Zorc et al. 2024).

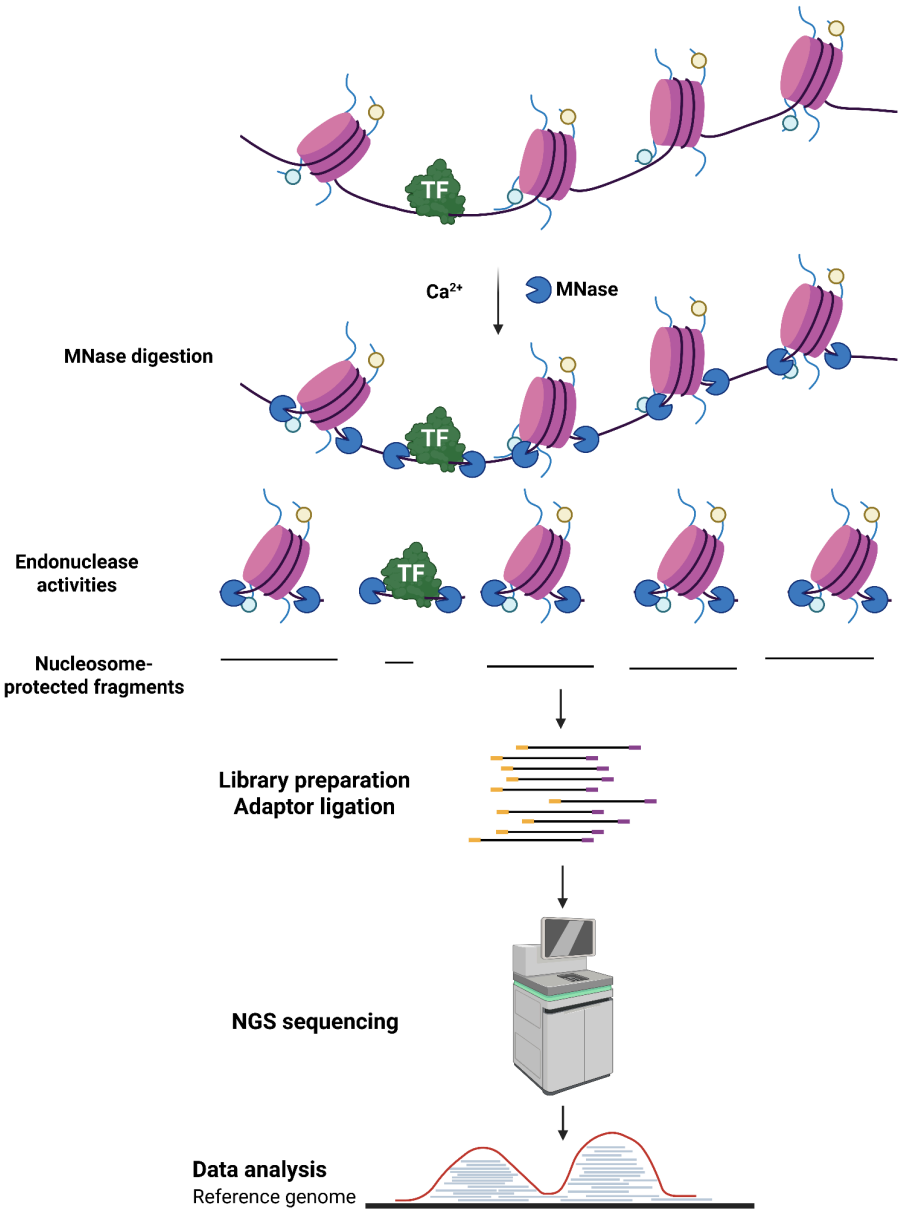


Figure 1.17: Schematic overview of MNase-seq.

Chromatin is digested with MNase in the presence of Ca^{2+}, resulting in preferential cleavage of linker DNA while DNA protected by nucleosomes or other DNA-binding proteins remains intact. Due to its endonuclease and exonuclease activities, MNase progressively digests accessible DNA, enriching for protein-bound fragments. The

protected DNA fragments are purified, ligated to sequencing adapters, and subjected to high-throughput sequencing. since TF-bound fragments are too small, they are normally destroyed during the decrosslinking process. Sequencing reads are mapped to a reference genome to infer nucleosome occupancy and chromatin organization across the genome. Figure created by Duong Doan using BioRender.

Originally designed to generate nucleosome occupancy maps, MNase-seq has subsequently also been optimized to identify regulatory regions. As an example, differential nuclease sensitivity sequencing (DNS-seq) measures lightly and highly digested chromatin to discriminate accessible regions with regulatory function (Rodgers-Melnick et al. 2016). The experimental protocol used in maize revealed base-pair resolution transcription factor footprints in inflorescence tissue as well as at sites involved with meristem regulators, like FEA4 and KN1, thus highlighting the regulatory role played by nucleosome conformation (Parvathaneni et al. 2020). These studies show that improved MNase-seq methods, such as DNS-seq, help identify important CREs needed for maize flower development. These techniques work in combination with ChIP-seq and DNase-seq by providing further regulatory regions and transcription factor interactions. By taking advantage of new experimental and computational methodologies, MNase-seq provides a further enhanced understanding of the relationship between chromatin structure and gene regulation across an array of biological and environmental contexts. However, similar to MOA-seq, MNase-seq is also impacted by a sequence bias, as MNase cleaves A/T-rich DNA much faster than G/C-rich DNA, potentially affecting apparent nucleosome occupancy. MNase digestion level also has a large effect on results: for example, under-digested samples may retain large fragments, while over-digestion may degrade nucleosome cores, making it difficult to quantitatively compare across samples. Finally, MNase-seq does not allow for distinguishing between nucleosome types such as variant histones or post-translational modifications and lacks single-molecule resolution, limiting its ability to access chromatin heterogeneity (Chereji et al. 2019).

1.4.2.4. Fiber-seq and SAM-seq

Fiber-seq (single-molecule chromatin fiber sequencing) is a revolutionary technique for mapping ACRs along individual strands of DNA, hence providing a single-molecule view of gene regulation. This technique utilizes nonspecific DNA adenine methyltransferases

(m6A-MTases), like Hia5, to methylate adenines in the accessible parts of the chromatin, such as DNase I hypersensitive sites (DHSs) and internucleosomal linkers, in intact purified nuclei, without efficiently bypassing nucleosome-bound DNA via the steric hindrance mechanism (Figure 1.18). Then, high-molecular-weight DNA is sequenced through Pacific Biosciences circular consensus sequencing, with the methylated adenines in long reads (10.9 kb average in length) marking methylase-accessible DNA sequences on DHSs (about 272 bp) and linkers (about 67 bp). This assay, without the use of antibodies, accomplishes remarkable accuracy, resolving fiber-to-fiber heterogeneity (such as 81% of fibers with open DHSs) and the cooperative action of neighboring regulators in a distance-dependent manner. In functional studies of Drosophila S2 cells, Fiber-seq has successfully determined nucleosome localization and DHS activation (Stergachis et al. 2020).

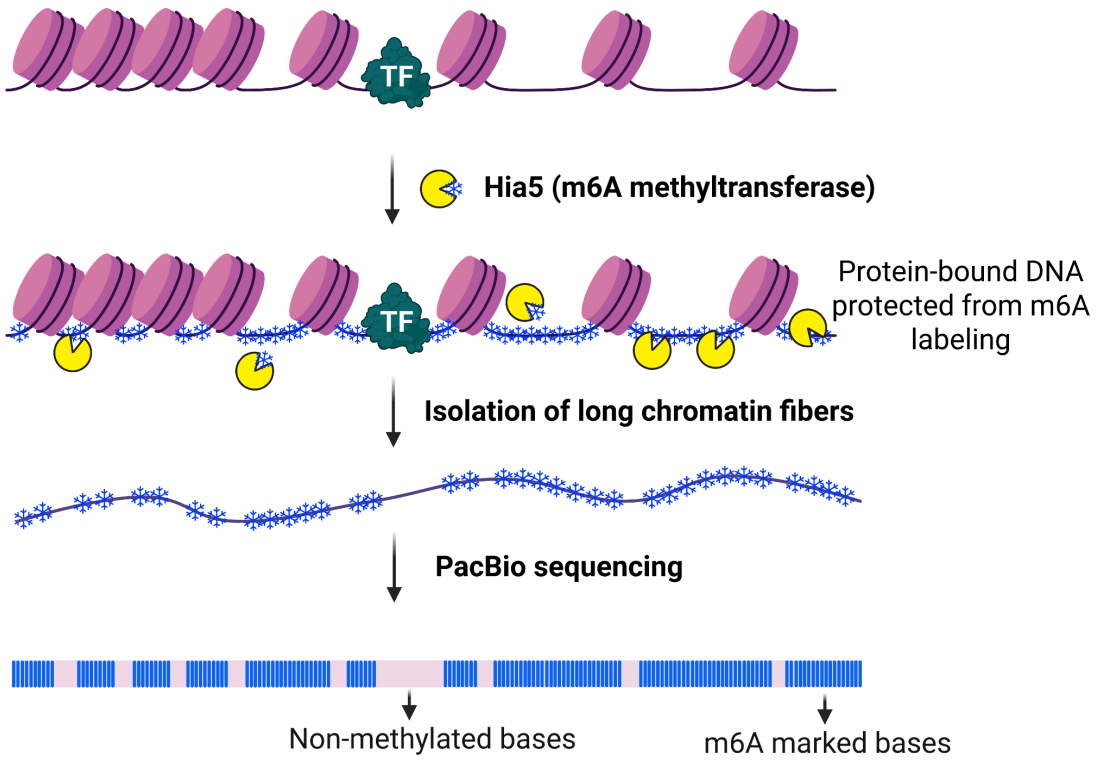


Figure 1.18: Schematic overview of Fiber-seq for single-molecule chromatin accessibility profiling.

Native chromatin fibers are incubated with the adenine-specific methyltransferase Hia5, which deposits N6-methyladenine (m6A) on exposed adenine bases along accessible DNA. DNA regions that are bound by nucleosomes or transcription factors are sterically

protected from methylation, resulting in local depletion of m6A signal. Long chromatin fibers are subsequently isolated and sequenced using PacBio long-read sequencing. At single-molecule resolution, the distribution of m6A marks along individual DNA molecules reveals chromatin accessibility patterns, nucleosome positioning, and protein occupancy across extended genomic regions. Figure created by Duong Doan using BioRender.

Recently, based on this foundation, SAM-seq (Simultaneous Accessibility and DNA Methylation Sequencing) was developed for the plant system. It combines the m6A-MTase tagging via the EcoGII enzyme with direct detection of endogenous cytosine methylation (5mC) in a single assay, and is sequenced by long-read Oxford Nanopore Technology (ONT) sequencing (Figure 1.19). As applied to nuclei of Arabidopsis and maize, SAM-seq produces high-resolution, genome-wide maps of accessibility and methylation in repetitive regions like centromeres and TEs, which can reveal nucleosome positioning, TF footprints, and cellular heterogeneity in bivalency, such as the co-occurrence of open chromatin with repressive methylation marks (Leduque et al. 2024). This ability to perform dual profiling reveals crosstalk between accessibility and methylation within nucleosomes, including decreased m6A in methylated gene bodies and TEs, thereby enabling more comprehensive views of epigenetic landscapes compared to accessibility-only methods like Fiber-seq.

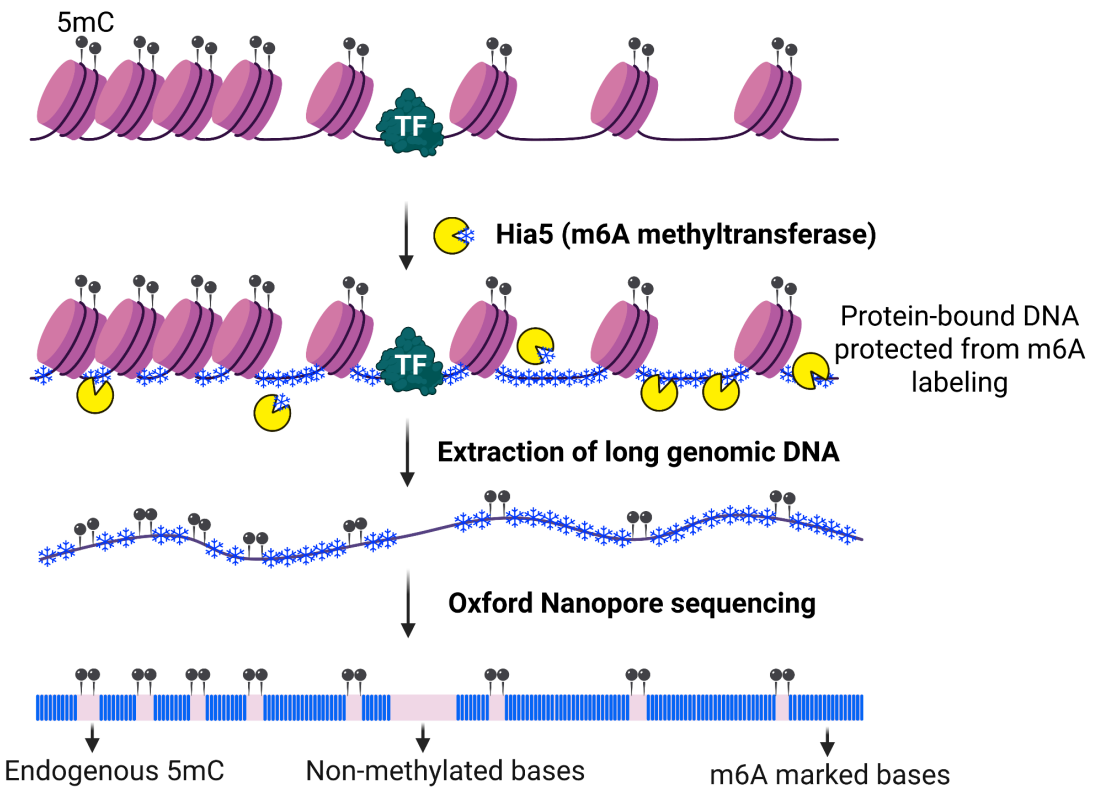


Figure 1.19: Dual profiling of chromatin accessibility and cytosine methylation with SAM-seq.

Native chromatin retains endogenous cytosine methylation (5mC) prior to in vitro treatment. Accessible DNA is labeled by the adenine methyltransferase Hia5, which deposits m6A at exposed adenine residues, while DNA bound by nucleosomes or transcription factors is protected from methylation. Long genomic DNA molecules are subsequently isolated and sequenced using Oxford Nanopore sequencing. Modified base detection enables simultaneous mapping of endogenous 5mC patterns and chromatin accessibility at single-molecule resolution. Figure created by Duong Doan using BioRender.

In the case of maize, it could shed light on haplotype-specific dynamic chromatin and hence make it especially useful in investigating epigenetic diversity in developmental and disease processes. The combination of Fiber-seq with MOA-seq, which maps TF binding sites at around 30 base pair resolution, may have the potential to improve the study of epigenomics by integrating single-molecule chromatin maps with precise TF

footprints, thus generating unprecedented information on gene regulation. The joint strategy can potentially enable researchers to identify which DNA strands have open TF binding sites, uncover co-activation of TFs together with regulatory elements, and place binding activities in the context of nucleosome organization. The strategy has the potential to make an influence in trait research by linking haplotype-specific TF binding, including bQTLs for drought tolerance (Engelhorn et al. 2025), to fiber-specific accessibility profiles, thus illuminating the molecular underpinnings of agronomic traits. By providing a single-molecule view of TF-chromatin interactions, the integrated strategy can potentially enhance our understanding of epigenetic regulation, thus enabling advances in plant breeding, with far-reaching implications for precision agriculture.

Despite the above advantages, these methods still have some drawbacks. Fiber-seq relies on PacBio sequencing, which is accurate but low-throughput and costly, thus making large plant genomes harder to scale. The methyltransferase also needs careful tuning to avoid off-target labeling inside nucleosomes. SAM-seq benefits from nanopore's long reads but inherits its higher base-calling error rates, even for improved analysis tools such as DeepSignalPlant (Leduque et al. 2024). Its use of m6A-MTases can introduce both sequence and strand biases; moreover, normalization steps may add variability when coverage is low. In general, both approaches require high-quality nuclei, which are often difficult to obtain from many plant tissues.

1.4.3. Single-Cell approach

Due to the inflexible plant cell wall and cell-type heterogeneity that preclude uniform culture, chromatin accessibility in plants previously was assessed on bulk tissues or by engineered lines carrying cell-type–specific markers (Yu et al. 2023). Single-cell ATAC-seq (scATAC-seq) avoids these limitations by microdissecting individual nuclei, marking accessible chromatin fragments with unique barcodes, and reassembling cell-resolved accessibility profiles (Figure 1.20) (Yu et al. 2023, 202).

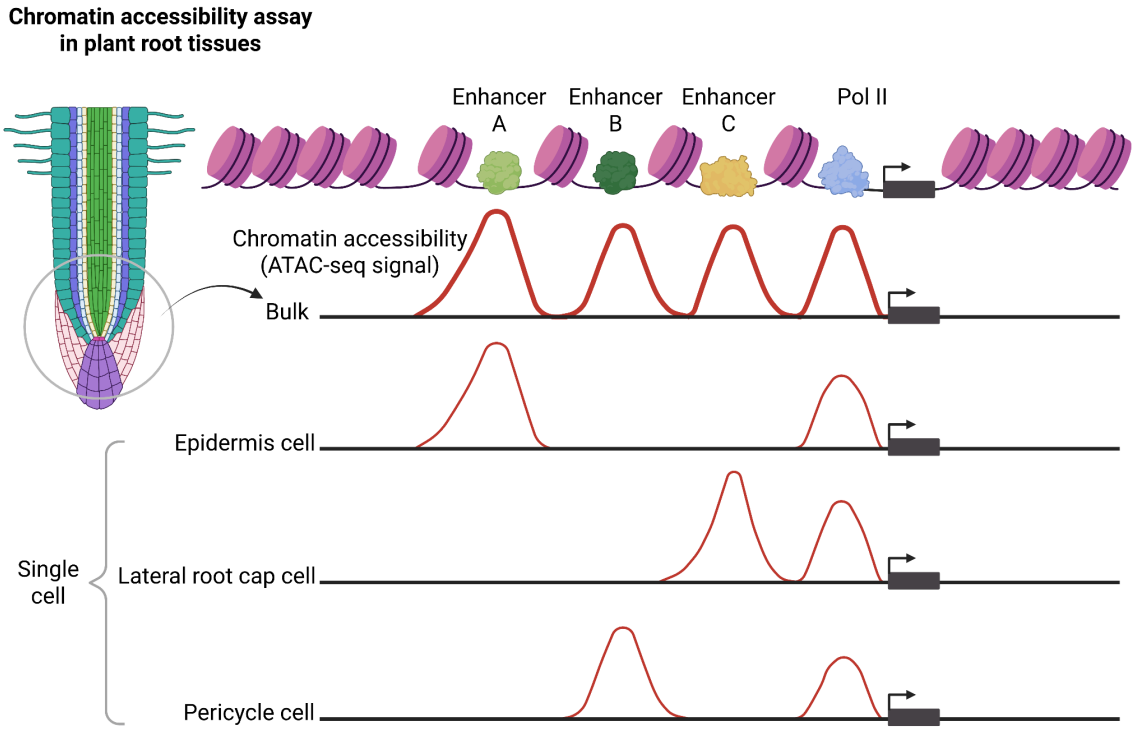


Figure 1.20: Cell-type–resolved chromatin accessibility in plant root tissues. Schematic illustration of chromatin accessibility profiles in bulk and single-cell assays derived from plant root tissue. The upper panel shows a genomic locus containing multiple regulatory elements, including distal enhancers and a promoter occupied by RNA polymerase II, together with the corresponding accessibility signal measured in bulk tissue. Bulk profiles represent the aggregate signal across heterogeneous cell types and therefore reflect combined accessibility at multiple regulatory elements. By contrast, chromatin accessibility profiling at the level of individual cells shows cell-type-specific patterns, which are shown for epidermis cells, lateral root cap cells, and pericycle cells in this figure. A particular set of regulatory regions is available in each cell type, thereby showing the differences in the usage of regulatory regions, which are not reflected when the data is measured in a bulk manner. Figure created by Duong Doan using BioRender.

Generally, there are two major approaches being employed in contemporary scATAC-seq (Figure 1.21 and 1.22). The first one is droplet-based approaches, which rely on tagmenting nuclei in bulk. It is followed by loading these nuclei together with barcodes in bead form into a microfluidic platform, in which each nucleus-bead pair is

divided into an oil droplet. Each oil droplet contains PCR reactions, which result in the attachment of barcodes to Tn5-catalyzed fragments, followed by library amplification with universal primers (Lareau et al. 2019). The second one is the combinatorial index approach (sciATAC-seq), which does not require microfluidics. These start with barcoding nuclei with Tn5 transposomes in a well-specific manner. Pooled nuclei are then redistributed in a limiting-dilution fashion, undergoing re-indexing in PCR with plate- and well-specific oligos. This two-level indexing approach facilitates profiling single cells on a large scale, requiring no droplet platforms (Cusanovich et al. 2015). In Arabidopsis root, scATAC-seq identified thousands of differentially ACRs that resolve epidermal, cortical, and endodermal lineages, demonstrating that cell-type identity can be defined solely by accessibility profiles (Tu et al. 2022). Advancements in single-cell technologies, such as microfluidic platforms (e.g., 10x Genomics) and cost-effective combinatorial indexing, have significantly improved throughput and scalability. These approaches effectively map cellular heterogeneity and enable direct peak-to-gene inference in complex tissues. However, their application in plants is limited by challenges in cell-wall digestion or nucleus isolation, inconsistent cell-type recovery, high per-cell sequencing costs, and the need for complex computational workflows.

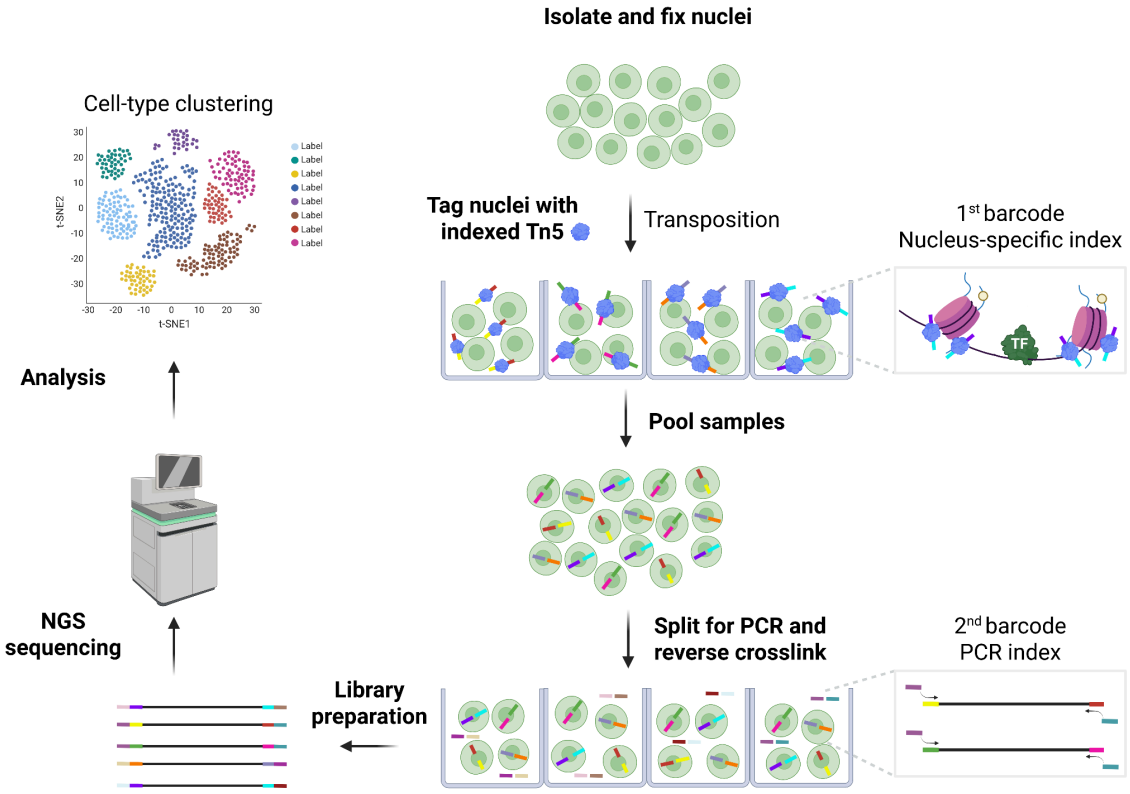


Figure 1.21: Schematic overview of single-nucleus chromatin accessibility sequencing using combinatorial indexing.

Nuclei are isolated and fixed from plant tissue, followed by tagging with indexed Tn5 transposase. During the transposition process, regions of open chromatin are fragmented and assigned an initial barcode simultaneously. Later, the nuclei are pooled and randomly distributed into new wells, and PCR amplification assigns a second barcode after the reversal of crosslinking. The combination of indexing enables the unique identification of each nucleus without requiring the need to physically separate them. The high-throughput sequencing and reassembly of chromatin regions using computational algorithms help recreate the cell type-specific landscape and clusters. Figure created by Duong Doan using BioRender.

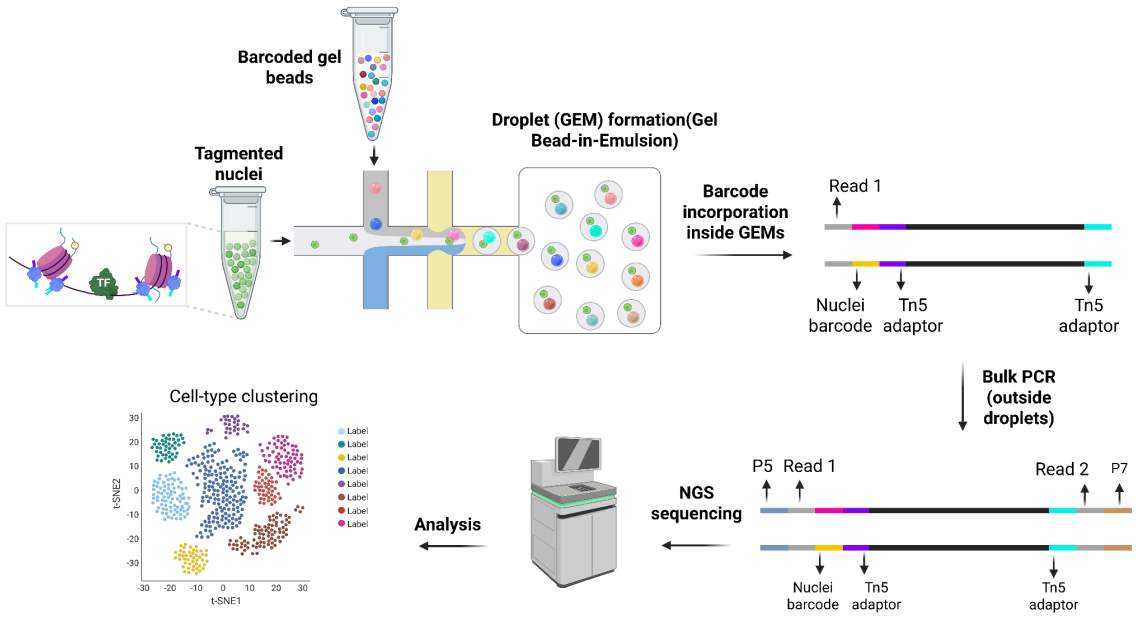


Figure 1.22: Droplet-based scATAC-seq in plants.

Intact nuclei are isolated from plant tissue and subjected to Tn5 transposase–mediated tagmentation, which fragments accessible chromatin and introduces sequencing adapters. Tagmented nuclei are then combined with barcoded gel beads in a microfluidic device to generate gel bead–in–emulsion (GEM) droplets, such that individual nuclei are co-encapsulated with a single bead carrying a unique nucleus-specific barcode. Inside the droplets, barcodes are introduced within the tagmented DNA fragments, which label all the fragments that come from the same nucleus. After the droplet breakage, bulk PCR is conducted to finish building the library as well as attaching the sequencing adapters. The libraries are then subjected to high-throughput sequencing. Paired-end reads are then mapped back to individual nuclei according to their corresponding barcodes, enabling cell-type–resolved clustering and downstream single-cell analysis. Figure created by Duong Doan using BioRender.

1.4.4. Functional Enhancer/CREs Activity Assays

While the methods mentioned above are adept at predicting CREs by mapping ACRs and TF binding sites, the validation of these elements requires targeted approaches to confirm their functional roles in gene regulation. Validation is important because it confirms that the predicted regulatory sequences not only bind TFs but also regulate gene expression, thus adding vital proof of their biological significance.

1.4.4.1. Classical reporter assays

Classic reporter assays and enhancer traps have been used to validate candidate CREs by cloning putative regulatory sequences upstream of a minimal promoter–reporter cassette and assaying expression in planta or cell-culture systems. In maize, a Hopscotch retrotransposon upstream of the domestication gene tb1 boosts expression by over tenfold in maize protoplasts, using a transient luciferase assay to confirm enhancer activity of this element in its endogenous chromatin context (Studer et al. 2011). In Arabidopsis, 14 predicted genome-wide DNase-seq intergenic DHSs were selected and cloned each in front of a minimal promoter–GUS reporter, and generated stable transformants; 71 % of them gave reproducible, tissue-specific GUS staining, verifying that open-chromatin signatures are reliably marking plant enhancers (B. Zhu et al. 2015). While stringent, steady-state integration-based assays remain laborious, qualitative, and restricted to available species, often suffering from position-effect variegation and insertional mutagenesis, potentially obscuring true enhancer function. To enable higher throughput and quantitative, real-time readouts, a high-throughput *Nicotiana benthamiana* leaf agroinfiltration system was developed in which putative enhancers cloned upstream of a mini-35S promoter–luciferase reporter are transiently assayed by bioluminescence imaging, with results being consistent with stable Arabidopsis reporters and dynamics of light-responsive activity patterns (Y. Lin et al. 2019). Nonetheless, even these temporary systems must create and filter through many constructs to achieve genome-wide coverage and may not completely recapitulate chromatin-dependent enhancers that require native genomic context to regulate correctly.

1.4.4.2. STARR-seq

Advancements in combining reporter assays with high-throughput sequencing now allow for comprehensive genome-wide analysis of regulatory functions. Self-Transcribing Active Regulatory Region sequencing (STARR-seq) is a reporter assay-based method for identifying and measuring the function of CREs, specifically enhancers, by assessing their capability to initiate transcription (Arnold et al. 2013). Originally developed for the animal system, STARR-seq has been adapted for plants and has helped discover functional enhancers in plants such as Arabidopsis and maize due to the cell wall limitation of protoplast and agroinfiltration systems (Jores et al. 2021). In STARR-seq, genomic DNA fragments are cloned downstream of a minimal promoter in a reporter

vector, transfected into cells, and sequenced to derive RNA transcripts to measure enhancer activity based on transcript level. STARR-seq provides a functional validation approach for CREs' transcriptional activity, complementing CREs or ACRs identification methods such as MOA-seq or ATAC-seq. For instance, STARR-seq was applied in maize leaf tissue using an ATAC-seq-derived library, thereby discovering distal H3K27ac-characterized enhancers enriched in TF binding sites involved in long-range chromatin interactions (Marand et al. 2023; Ricci et al. 2019). To bypass the use of plant protoplasts, STARR-seq was optimized using *Agrobacterium*-mediated transformation in *Nicotiana benthamiana* leaves. The technique screened several putative enhancers from several plants and identified small DNA segments having enhancer activity, some of which were light condition-dependent. Saturation mutagenesis was employed to localize actively working regions to a high degree of accuracy, enabling the assembly of new synthetic enhancers. This approach efficiently processed large genetic libraries and recovered numerous fragments from one leaf (Jores et al. 2020). STARR-seq can bridge sequence-based prediction and functional validation, making it a critical component in understanding the cis-regulatory landscapes in plants.

However, STARR-seq also has some limitations. The fragments are assayed outside their native chromatin environment; therefore, the method can not fully capture context-dependent regulation, such as nucleosome positioning, chromatin compaction, or 3D genome architecture (Arnold et al. 2013). Plant implementation is dependent on protoplasts or Agrobacterium infiltration, which may skew cell-type representation and trigger stress reactions that complicate enhancer activity measurements. Reporter-vector cloning also imposes size and sequence biases (Jores et al. 2021). In addition, STARR-seq selectively identifies enhancers that are active under assay conditions, possibly missing components required for particular tissue-specific or developmental cues (Marand et al. 2023). Therefore, even though STARR-seq can provide informative results, in vivo validation and complementary chromatin-based assays are typically essential.

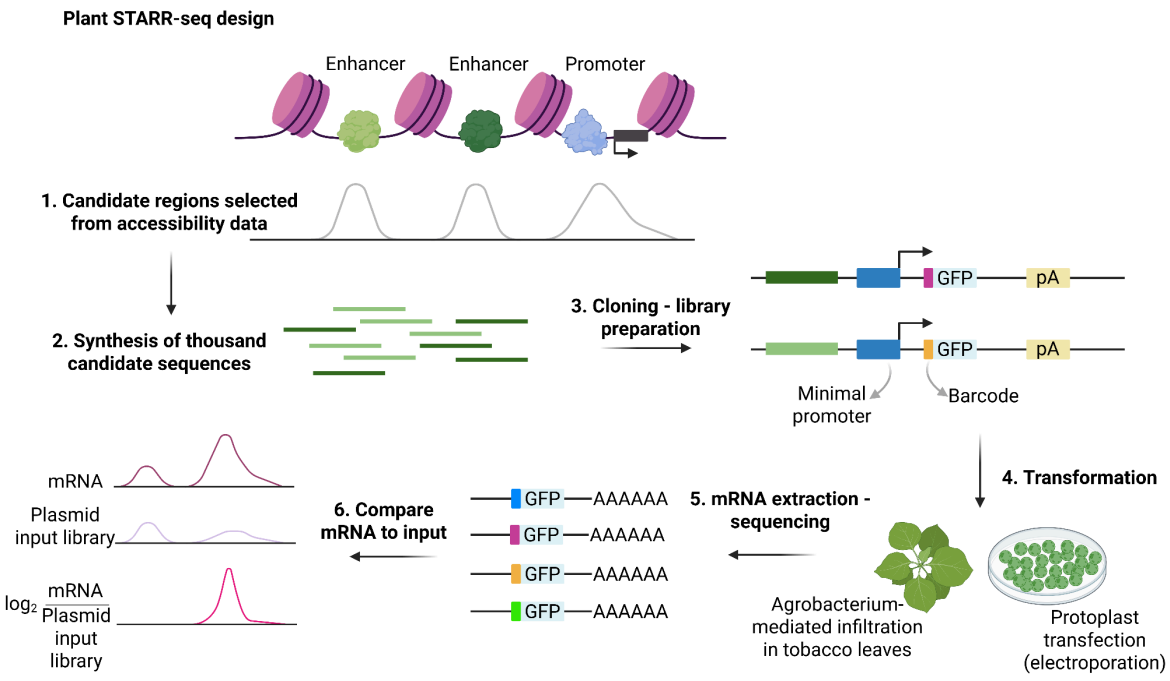


Figure 1.23: Plant STARR-seq experimental design for high-throughput enhancer identification.

Candidate cis-regulatory regions are selected based on chromatin accessibility data and synthesized as a pooled library. Each candidate sequence is cloned upstream of a minimal promoter driving a GFP reporter, together with a unique barcode that is transcribed with the reporter. The plasmid library is transiently introduced into plant cells by protoplast transfection or *Agrobacterium*-mediated infiltration. Following expression, total mRNA is extracted and sequenced to quantify barcode abundance. Enhancer activity is calculated by comparing barcode counts in the mRNA pool to those in the plasmid input library, allowing quantitative assessment of regulatory activity for thousands of candidate elements in parallel. Figure created by Duong Doan using BioRender.

1.4.4.3. CRISPR-based functional screens

In addition to high-throughput approaches like STARR-seq, CRISPR functional screens represent the most recent frontier to deconstruct plant cis-regulatory landscapes in their native genomic context. CRISPR functional screens employ sgRNA pools as libraries to interfere with candidate CREs in vivo and to measure the ensuing downstream effects on gene expression or developmental traits. High-throughput CRISPR screen systems have been developed to directly target noncoding regions by fusing catalytically dead

Cas9 (dCas9) to repression (CRISPRi) or activation (CRISPRa) domains or by utilizing wild-type Cas9 to induce deletion or base edits across densely spaced genomic targets (Bock et al. 2022). The pioneer application in the human genome of this approach was the conduct of a CRISPRi tiling screen around the GATA1 and MYC loci, identifying 470 high-confidence enhancers whose inhibition altered target gene expression (Gasperini et al. 2019). Recently, pooled CRISPRi screens mapped entire TADs at nucleotide resolution, thoroughly analyzing all CREs within the 2.8 Mb MYC TAD across various cancer cell lines (Caragine et al. 2025). In plants, CRISPR/Cas9 was used to generate in-context promoter allelic series in tomato and specific genes like SlCLV3, SlWUS, and SlYABBY, to establish a range of plant architecture and fruit size phenotypes, demonstrating how accurate cis-regulatory edits can expand breeding diversity (Rodríguez-Leal et al. 2017). Most recently, multiplexed CRISPR-Cas9 mutagenesis was applied in rice to disrupt noncoding DNA regions upstream of PSBS1, rescuing 120 transgene-free alleles from knockout through overexpression. These edits gave rise to two- to threefold boosts in OsPsbS1 protein content and related increases in non-photochemical quenching capability and water-use efficiency (Patel-Tupper et al. 2024). These promoter edits segregate the CRISPR machinery away in progeny, producing non-transgenic lines that meet stringent regulatory requirements for genome-edited crops (Rodríguez-Leal et al. 2017). Despite its great potential, plant CRISPR screening methods face a number of challenges, such as variable transformation efficiency between species and tissues, unpredictable efficacy of sgRNAs, limited accessibility to heterochromatic regions, and functional redundancy among enhancers, potentially obscuring their individual contributions.

1.5. Concluding Remarks and Outlook

The reviewed literature in this chapter presents both meaningful progress and continued challenges to understanding CREs in plants. Decades of CRE studies have established that these regulatory elements play an important role in controlling plant development, mediating environmental responses, and driving evolutionary innovation. Epigenomic profiling technologies, such as measuring chromatin accessibility, mapping histone modifications, and analyzing 3D genome architecture, have expanded our ability to annotate regulatory regions on a genome-wide basis (Rodgers-Melnick et al. 2016; Ma and Zhang 2020; P. Dong et al. 2017). Despite the accumulating wealth of data, our conceptual and mechanistic understanding of how sequence features, chromatin

context, and regulatory outcomes are related remains incomplete (Marand et al. 2023). Further, while regulatory datasets accumulate, they are concentrated in a few plant model systems; thus, the regulatory landscapes across the tremendous diversity of plant lineages remain largely unexplored.

The second fundamental theme to have evolved from these efforts relates to the methodological challenges and limitations that characterize regulatory genomics to date. All of the assays currently in use capture distinct aspects of CRE function, such as accessibility, histone signatures, TF binding, and chromatin interactions, but none offer a complete picture. Chromatin accessibility assays and histone mark profiling tend to identify large gene regulatory regions but lack the precision sufficient to define TF binding events. The TF binding footprints can be obtained using well-known protein-DNA binding assays like ChIP-seq or DAP-seq, but they have drawbacks such as antibody availability, throughput, and the difficulty of capturing native chromatin states (Klein and Hainer 2020). Numerous approaches provide only static snapshots, failing to adequately characterize regulatory dynamics in a time frame and condition-dependent manner. Equally important, functional validation remains one of the greatest bottlenecks. Even if there were thousands of candidate CREs identified either computationally or experimentally, validation methods such as reporter assays, genome editing, and complementation experiments would be time-consuming and limited to a few species (Schmitz et al. 2021; Marand et al. 2023). Thus, these methodological limitations highlight that integrative, high-resolution, and widely applicable strategies are essential, especially those that can efficiently scale across phylogenetically diverse taxa.

Beyond those technical challenges, conceptual gaps also remain. It has been revealed via comparative studies that CREs can be highly dynamic. Some regulatory elements are deeply conserved, while others undergo rapid turnover, repurposing, or complete loss across evolutionary timescales (Lu et al. 2019). However, the principles underlying this diversity when CREs are preserved, when they diverge, and how chromatin context shapes these trajectories remain unresolved. It is still unclear to what extent regulatory evolution follows predictable patterns versus lineage-specific innovations (H. Yan et al. 2025). While there are scattered examples of the rewiring of developmental, hormonal, or stress-response pathways, comprehensive cross-species comparisons are still scarce due to inconsistent datasets and limited species coverage. Consequently, our ability to discern which aspects of CREs are evolutionarily constrained versus flexible remains poorly understood.

The themes emerging from this chapter collectively emphasize the need for more systematic, cross-species, and functionally grounded studies of regulatory architecture. Fine-scale regulatory maps, comparative frameworks, and experimental validation beyond the classical model systems will contribute to addressing these questions. The subsequent chapters build upon this conceptual and methodological framework to explore these open questions using high-resolution profiling of regulatory activity across diverse plant lineages and the investigation of how key developmental and physiological pathways may have been influenced by the evolution of their regulatory elements.

Chapter 2: The evolution of Cis-regulatory elements in Embryophytes

2.1. Introduction

CREs are central to establishing the gene expression patterns that underlie plant development, physiology, and adaptation. CREs evolve through a diversity of mechanisms, including SNPs, small indels, duplication events, and structural variation that repositions regulatory sequences or alters chromatin context (Marand et al. 2023). In plants, one especially influential source of regulatory novelty is the exaptation of TEs, as TE-derived sequences frequently acquire TF binding motifs and regulatory potential (Lu et al. 2019). Comparative analyses have shown that regulatory regions can undergo rapid turnover, as gains, losses, and motif changes accumulate across evolutionary timescales (Song et al. 2021; Maher et al. 2018). While TFs themselves tend to evolve slowly, most likely due in large part to their regulatory influence on many downstream genes, the CREs to which they bind are considerably more dynamic (Weirauch and Hughes 2010). This discrepancy between the conservation of TFs and the fluidity of their target CREs suggests that regulatory evolution is driven in large part by changes in noncoding DNA rather than in TF proteins themselves. Both deep conservation and extensive divergence of regulatory sequences were evaluated across flowering plants. Some CREs, particularly those associated with developmental regulators or core physiological pathways, are conserved across tens of millions of years, indicating strong evolutionary constraint (Baumgart et al. 2025). This contrast naturally raises the following fundamental questions: when do CREs remain constrained, when do they diverge, and what are the genomic and/or ecological factors that determine whether a given regulatory site will persist or become replaced? While individual examples of regulatory divergence have been documented, like the modification of enhancer activity leading to domestication traits at the tb1 locus in maize (Studer et al. 2011), broad principles governing CRE evolution remain poorly resolved. The major conceptual challenge is the tension between predictability and plasticity in regulatory evolution. Some regulatory outputs are remarkably stable: orthologous genes often exhibit broadly similar expression patterns across species despite rapid sequence-level divergence in their CREs (Weirauch and Hughes 2010). This suggests that regulatory functions can be

preserved even as underlying sequences turn over or are reorganized. At the same time, many lineage-specific regulatory innovations appear to contribute to phenotypic diversification, ecological specialization, and domestication. The question of which aspects of the regulatory architecture are conserved versus flexible and understanding why requires comparison of regulatory data across a broad phylogenetic range.

The Andropogoneae tribe offers a uniquely informative framework for examining CRE evolution. This grass tribe includes key agricultural staples such as maize, sugarcane, sorghum, and the promising biofuel source Miscanthus. Andropogoneae species occupy 17% of the land and have a wide range of environmental adaptations (Welker et al. 2020; Vicentini et al. 2008). While these species share a common ancestor approximately 18 million years ago (Figure 2.2), they exhibit remarkable genomic diversity driven by extensive TE activity, polyploidy, and structural variation (Stitzer et al. 2025). Genome sizes vary dramatically within the tribe, yet members display broadly similar vegetative and reproductive morphologies along with conserved C4 photosynthetic pathways (Bianconi et al. 2020). This combination of conserved biological processes with highly dynamic genome structure creates a special opportunity for investigation into how regulatory sequences evolve in the context of both shared functional constraints and substantial genomic divergence. The Andropogoneae thereby offer a powerful system for assessing which CREs remain conserved within conserved pathways and which diverge in concert with either genome architecture or ecological adaptation. This extension of regulatory comparisons beyond the Andropogoneae enables the examination of CRE evolution across deeper evolutionary timescales and more varied genomic contexts. Species such as *Physcomitrium patens*, *Betula pendula*, and *Vicia faba* represent broad phylogenetic breadth, extensive variation in genome size, and contrasting TE landscapes (Jayakodi et al. 2023; Lang et al. 2018). The inclusion of species from the bryophyte lineage to large-genome angiosperms enables assessments of how genome architecture influences the evolution of regulatory sequences and provides broad comparative anchors that can help distinguish plant-wide regulatory features from lineage-specific innovations. To address these questions, this chapter presents a comparative regulatory atlas generated using MOA-seq across 27 plant species, including 2 haplotypes of *Zea mays* (B73 and Mo17), representing major evolutionary lineages, spanning over 500 million years of evolution, diverse genome

sizes, and varied ecological strategies. This dataset will provide a unified resource for assessing CRE evolution across plants.

2.2. Material and methods

2.2.1. Plant materials

Plant materials were obtained and prepared as follows:

Sixteen Andropogoneae species from the wild were grown in a greenhouse at the Danforth Plant Science Center, Cornell University, New York, USA, under the care of Prof. Ed Buckler's team. Leaf tissue was sampled in duplicate, flash-frozen in liquid nitrogen, and taken to the Max Planck Institute for Plant Breeding Research, Cologne, Germany, for MOA-seq experiments.

Oryza sativa (Kitaake) was cultivated by PhD student Laura Redzich in the Team Rice greenhouse, Institute of Molecular Physiology, Heinrich-Heine-Universität Düsseldorf, Germany. Young, expanding leaves were harvested and flash-frozen in liquid nitrogen.

Betula pendula, an inbred line from the inbred collection of the Natural Resources Institute Finland, was grown in the greenhouse of the University of Helsinki, Finland, by Dr. James Ord and Dr. Jarkko Salojärvi's laboratory. Young, developing leaves were collected in triplicate (8th and 9th leaves in separate tubes), flash-frozen in liquid nitrogen, and transported to the Max Planck Institute for Plant Breeding Research, Cologne, Germany.

Physcomitrium patens (Gransden 2004 strain) was obtained from Dr. Thomas Jacobs' laboratory at the Center for Plant Systems Biology, VIB, Ghent, Belgium. A tissue mixture of protonema and gametophore was harvested in triplicate, flash-frozen in liquid nitrogen, and transported to the Max Planck Institute for Plant Breeding Research, Cologne, Germany.

Sorghum bicolor (BTx623) and *Glycine max* (Williams 82) were cultivated by Dr. David Blasco Escámez from Dr. Nelissen's lab at the Plant Growth Dynamics, Center for Plant Systems Biology, VIB, Ghent, Belgium. In *S. bicolor*, young, expanding leaves (leaf numbers 8 and 9) were collected from plants that were 31 days old. In *G. max*, trifoliate leaves were harvested from plants 23 days old. Triplicate samples from each species were collected and flash-frozen in liquid nitrogen. *Brassica napus* (Express 617) and *Vicia faba* (Tiffany) were obtained from Prof. Dr. Agnieszka Golicz and cultivated in the greenhouse of the Department of Agrobioinformatics, Justus Liebig University, Giessen,

Germany. Young, growing leaves were collected and flash-frozen in liquid nitrogen for subsequent analysis.

2.2.2. Data materials

MOA-seq datasets for *Zea mays* (inbred lines B73 and Mo17) were obtained from a previously published study (Engelhorn et al. 2025). MOA-seq data for *Setaria viridis* were obtained from Myers et al. 2023.

Additional unpublished MOA-seq datasets generated within the group were used in this study. The *Arabidopsis thaliana* MOA-seq dataset was produced by Dr. Julia Engelhorn, Sergius Weizel (former master student, Institute of Molecular Physiology, Heinrich-Heine-Universität Düsseldorf, Germany), and Dr. Sohini Mukherjee. MOA-seq data for *Hordeum vulgare* were generated by Nafiseh Sargheini (former master student, Rheinische Friedrich-Wilhelms-Universitaet Bonn University of Bonn, Germany). The *Solanum lycopersicum* MOA-seq dataset was produced by Max Blank (former master student, Institute of Molecular Physiology, Heinrich-Heine-Universität Düsseldorf, Germany) and Dr. Julia Engelhorn.

2.2.3. MOA-seq

2.2.3.1. Protocol for monocot species

The MOA-seq protocol for monocot species was adapted from (Savadel et al. 2021) with some adjustments. The 627 µl of 16% paraformaldehyde was combined with 10 ml of fixation buffer (1.2 ml Buffer A [150 mM PIPES, 800 mM KCl, 200 mM NaCl, 20 mM EDTA, 5 mM EGTA, pH 6.8], 2.4 ml 1.6 M sorbitol, 12 µl 1 M DTT, 12 µl 1000x polyamines [0.15 M spermine, 0.5 M spermidine], 10 µl 200 mM phenanthroline, 20 µl 200 mM PMSF) in advance. This was added to 75 mg of finely powdered frozen leaf tissue. Fixation subsequently proceeded with continuous agitation (on a shaker at 40 rpm) for 10 minutes, then quenched for 5 minutes in 0.25 M glycine in continuous agitation. The combination was brought to 45 ml with MDB buffer (50 mM HEPES, 12.5% glycerol, 25 mM KCl, 4 mM $MgCl_2$, 1 mM $CaCl_2$, pH 7.6), centrifuged for 10 minutes at 1500 rcf for 4°C, then the pellet was resuspended in 5 ml MDB buffer containing 1% Triton X-100. With 20 µl of 200 mM PMSF, the nuclei were released from the cells through mechanical rupture in a polytron for 25,000 min^{-1} and 29,000 min^{-1} for three 10-second bursts for both speeds, both times at 4°C. The formed lysate was

strained through MDB pre-moistened miracloth, filled up to 25 mL with MDB buffer for washing, then centrifuged for 15 minutes at 2000 rcf at 4°C, repeated one more time for washing for purification of the nuclei. The nuclei pellet was resuspended in 70 µl MDB with 1% Triton X-100 and then treated with 7.8 µl MNase (M3: 12.5 U/mL, M4: 25 U/mL, M5: 50 U/mL, M0: only MDB buffer), followed by rapid mixing, 15-minute incubation at 37°C. To quench the digestion, 0.5 M EGTA-NaOH (pH 8.0) containing 2.5 µl 20 mg/ml of RNase A was added and kept at 30-minute room temperature incubation. The decrosslinking was in 0.5 M NaCl, 1% SDS, 0.2 mg/ml Proteinase K for 10 hours, 65°C. The DNA was purified with the NEB Monarch PCR and DNA Cleanup Kit with oligonucleosome cleanup protocol. MOA DNA was run on a 1.5 % Agarose gel to observe the MNase digestion strength. The quantification of the DNA was performed using Qubit dsDNA HS Assay Kit (Thermo Fisher Scientific). NEBNext ® Ultra II DNA Library Prep Kit for Illumina ® was used for library preparation with modification of 30-minute ligation at 20°C as well as dual size selection with NEBNext purification beads (50 µl, recovered in the supernatant, then 100 µl, beads retained). The five cycles of PCR amplification used purification in 75 µl of purification beads. The size selection used BluePippin with 3% 3-dye-free agarose gel with Q3 marker mix with 195 bp target in the 150–240 range. The final cleanup used 75 µl purification beads with elution in 15 µl 0.1x TE buffer. Libraries were sequenced on Illumina NovaSeq 6000 S4.

2.2.3.2. Protocol for dicot species

The MOA-seq protocol described above was also applied to *Vicia faba*, *Glycine max*, *Brassica napus*, and *Physcomitrium patens* with modifications: the fixation buffer was supplemented with two additional proteinase inhibitors, Aprotinin and Pepstatin A, each added at 10 µl from a 2 mg/mL stock solution. After the decrosslinking step, samples were centrifuged at 15,000 rpm for 3 minutes to remove any residual pellet before proceeding with DNA purification using the NEB Monarch PCR and DNA Cleanup Kit. This MOA-seq protocol for dicot species was previously developed in the group.

For *Betula pendula*, I optimized the dicot MOA-seq protocol with additional adjustments. The MDB buffer used during nuclei isolation contained 3 mM $CaCl_2$ instead of the standard concentration. The fixation buffer was supplemented with double the amounts of PMSF, DTT, phenanthroline, polyamines, aprotinin, and pepstatin A. Following mechanical lysis of the tissue using a polytron, the lysate was filtered through Miracloth

pre-moistened with MDB, then diluted to 25 mL with MDB containing 5% Triton X-100 for initial washing. A second wash was performed using standard MDB buffer. MNase digestion was performed with M7 (200 U/mL) and M8 (400 U/mL), and quenched with 7.5 µl of RNase A in stop solution, followed by incubation at 37 °C in a water bath. DNA purification included an additional extraction step using chloroform:isoamyl alcohol (24:1), followed by cleanup with the NEB Monarch PCR and DNA Cleanup Kit. After initial purification, 50 µl of MOA DNA was treated again with 3 µl of RNase A for 1 hour at 37 °C in a water bath. Two M7 digestions from the same sample were then pooled and subjected to another round of purification using the NEB Monarch PCR and DNA Cleanup Kit. Samples were sequenced on the same platform as the monocot samples.

2.2.4. Raw data processing

2.2.4.1. Sequencing quality control

Quality control of raw sequencing reads was performed with FastQC (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/) using the following command.

```
fastqc filename.fastq
```

Sequence duplication was attributed to sample biology rather than enrichment bias, therefore, duplicates were retained, as PCR cycles had been optimized accordingly. Adapter content was also checked to rule out contamination, with expected footprints of 30–80 bp indicating adapters should appear after 30 bp.

2.2.4.2. Adaptor trimming

Adaptor sequences of raw reads were trimmed with SeqPurge (Sturm et al. 2016) using the following code:

```
seqpurge -in1 read1.fastq -in2 read2.fastq -out1
trimmed_read1.fastq -out2 trimmed_read2.fastq -min_len 20 -qcut
0
```

2.2.4.3. Read merging and quality enhancement

Paired-end MOA-seq reads were merged into single-end reads with base quality correction using FLASH (Magoč and Salzberg 2011). The short fragment length of MOA-seq resulted in nearly complete overlap between read pairs, improving overall read quality.

```
flash read1.fastq read2.fastq -o output_prefix -M 80
--cap-mismatch-quals
```

The complete workflow used for adaptor trimming and read merging combination is provided as a script in Appendix 5.1.1.

2.2.5. MOA reads mapping

MOA-seq reads were aligned to their respective reference genomes (Table 2.1) using STAR (v2.7.11a) (Dobin et al. 2013). For each species, STAR-indexed genomes were first generated from the assembled reference `FASTA` file. MOA-seq reads were then aligned to the corresponding indexed genome. Since STAR is optimized for RNA-seq, spliced alignment was disabled by setting `--alignIntronMax 1`, ensuring DNA-appropriate mapping. The `--outFilterMultimapNmax` parameter was adjusted based on species ploidy level and genome assembly type (Table 2.1), and multimapping behavior was controlled using `--winAnchorMultimapNmax` and `--outMultimapperOrder Random`. Output BAM files were written in coordinate-sorted format. The complete command-line workflow used for indexing and mapping is provided in Appendix 5.1.2.

Table 2.1. Plant species, ploidy levels, and reference genomes used for MOA-seq mapping.

Species	Ploidy	Haploid Genome Assembly	Mapping	Reference genome
Arabidopsis thaliana	Diploid	TRUE	1x	TAIR 10
Oryza sativa	Diploid	TRUE	1x	Kitaake v3.1
Betula pendula	Diploid	TRUE	1x	*Betula pendula* (unpublished; provided by Dr. J. Salojärvi)

Physcomitrium patens	Haploid	TRUE	1x	Ppatens v6.1 (Gransden Japan)
Setaria viridis	Diploid	TRUE	1x	Setaria viridis v4.1
Sorghum bicolor	Diploid	TRUE	1x	BTx642 v1.1
Solanaceae lycopersicum	Diploid	TRUE	1x	M82 v1
Brassica napus	Allotetraploid	TRUE	2x	Express617 v1
Glycine max	Partially diploidized tetraploid	TRUE	1x	Wm82.a6.v1
Zea mays B73	PaleoTetraploid	TRUE	1x	Zm-B73-REFERENCE-NAM-5.0
Zea mays Mo17	PaleoTetraploid	TRUE	1x	Zm-Mo17-REFERENCE-CAU-2.0
Hordeum vulgare	Diploid	TRUE	1x	Morex v3
Vicia faba	Diploid	TRUE	1x	Tiffany v1
An1_*Themeda triandra*	AlloDiploid	FALSE	2x	Tt-AUB21_1-DRAFT-PanAnd-1.0
An2_*Pogonatherum paniceum*	AlloDiploid	TRUE	1x	Pi-Clark-DRAFT-PanAnd-1.0
An3_*Anatherum virginicum*	AlloDiploid	TRUE	1x	Av-Kellogg1287_8-REFERENCE-PanAnd-1.0
An4_*Heteropogon contortus*	AlloTetraploid	FALSE	4x	Hc-AUB53_1-DRAFT-PanAnd-1.0
An5_*Thelepogon elegans*	AlloDiploid	FALSE	2x	Te-Pasquet1246-DRAFT-PanAnd-1.0
An6_*Andropogon burmanicus*	AlloTetraploid	FALSE	4x	Ab-Traiperm_572-DRAFT-PanAnd-1.0
An7_*Schizachyrium scoparium*	AlloTetraploid	FALSE	4x	Ss-CAM1384-DRAFT-PanAnd-1.0
An8_*Cymbopogon citratus*	AlloTetraploid	FALSE	4x	Cc-PI314907-DRAFT-PanAnd-1.0
An9_*Zea diploperennis (Momo)*	PaleoTetraploid	TRUE	1x	Zd-Momo-REFERENCE-PanAnd-1.0

An10_*Sorghastrum nutans*	AlloTetraploid	FALSE	4x	Sn-CAM1369-DRAFT-PanAnd-1.0
An11_*Tripsacum dactyloides KS*	PaleoTetraploid	TRUE	1x	Td-KS_B6_1-REFERENCE-PanAnd-2.0
An12_*Andropogon gerardi*	AlloHexaploid	FALSE	6x	Andropogon gerardi HAP1 v1.1 and HAP2 v1.1
An13_*Hemarthria compressa*	AutoHexaploid	FALSE	6x	Hp-KelloggPI404118-DRAFT-PanAnd-1.0
An14_*Bothriochloa laguroides*	AlloHexaploid	FALSE	6x	Bl-K1279B-DRAFT-PanAnd-1.0
An15_*Elionurus tripsacoides*	AlloTetraploid	TRUE	2x	Et-Layton_Zhong168-DRAFT-PanAnd-1.0
An16_*Urelytrum digitatum*	AlloHexaploid	TRUE	3x	Ud-Pasquet1171-DRAFT-PanAnd-1.0

2.2.6. Peak calling

Peak calling was performed on MOA-seq BAM files for all 28 species using MACS3 (v3.0.0a7) (Y. Zhang et al. 2008). Prior to peak identification, BAM file processing and estimation of mean mapped fragment length were conducted using SAMtools (v1.9) (Danecek et al. 2021). For each sample, `samtools stats` was used to calculate the average fragment length, which was subsequently used as the fragment size, minimum peak length, and maximum gap parameters in MACS3. Actual genome sizes were computed from indexed FASTA files (`.fai`) and supplied to MACS3 via the `-g` parameter to enable species-specific genome-size aware peak calling. All steps, including fragment-length extraction, genome-size calculation, and execution of MACS3 with matched parameters, were automated using a custom Bash script. The complete script is provided in Appendix 5.1.3.

2.2.7. Gene annotation

Gene annotation for all species was performed using Helixer with the `land_plant` parameter (Holst et al. 2025). The Ed Buckler group generated annotations for the 16 Andropogoneae species (Stitzer et al. 2025), while the remaining 12 species were annotated as part of this study using the same Helixer workflow to maintain consistency

across datasets. All annotation steps, including input preparation and execution of Helixer, were automated using a custom script. The full script is provided in Appendix 5.1.4.

2.2.8. Phylogenetic Tree Reconstruction Using Chloroplast Genes

To construct a maximum likelihood (ML) phylogenetic tree using chloroplast genes, a pipeline was followed to process genomic data from 28 plant species, leveraging their chloroplast genomes available on NCBI. The workflow included sequence retrieval, gene identification, sequence alignment, gene tree construction, species tree inference, and estimation of divergence times. All command-line scripts and processing pipelines used in these analyses are provided in Appendix 5.1.8.

2.2.8.1. Sequence Retrieval and Identification of Shared Chloroplast Genes

Chloroplast genome sequences for 28 plant species were retrieved from the NCBI Nucleotide database with accession numbers: NC_000932.1, NC_031333.1, NC_072281.1, NC_037465.1, NC_028075.1, NC_008602.1, NC_007898.3, PP712901.1, NC_016734.1, NC_001666.2, NC_056985.1, MT120813.1, NC_035016.1, NC_029881.1, NC_086842.1, NC_035038.1, NC_035027.1, NC_042144.1, MT610089.1, NC_035032.1, NC_040111.1, MT610091.1, NC_037087.1, NC_030498.1, MT610055.1, MT610069.1, MT610053.1, NC_040131.1 (*Bothriochloa decipiens* was used instead of *Bothriochloa laguroides*). These sequences were stored in text files containing gene names and their corresponding sequences. Common genes present in all species were identified by concatenating and analyzing the gene lists. From these genomes, 56 chloroplast protein-coding genes shared by all species were identified and used for phylogenetic reconstruction:

atpA, atpB, atpE, atpF, atpH, atpI, cemA, ndhA, ndhB, ndhC, ndhD, ndhE, ndhF, ndhG, ndhH, ndhI, ndhJ, ndhK, petA, petB, petD, psaA, psaB, psaJ, psbA, psbB, psbC, psbD, psbE, psbF, psbH, psbI, psbJ, psbK, psbM, rbcL, rpl14, rpl16, rpl2, rpl20, rpl23, rpl33, rpl36, rpoB, rpoC1, rpoC2, rps11, rps12, rps14, rps15, rps18, rps2, rps3, rps4, rps7, rps8.

2.2.8.2. Sequence Extraction and Header Normalization

For each species, nucleotide sequences corresponding to the 56 core genes were extracted from the chloroplast genome annotations. FASTA headers were reformatted to

incorporate species names, gene names, and unique protein identifiers in the format species_gene_proteinID, ensuring compatibility with downstream alignment, concatenation, and phylogenetic tools. The resulting gene-specific FASTA files were then concatenated across species, producing a multi-species FASTA file for each gene.

2.2.8.3. Sequence Alignment and Trimming

Multiple sequence alignments were generated for each gene using MAFFT v7.505 (Katoh and Standley 2013) with the `--auto` setting, which selects the optimal alignment strategy based on sequence size and complexity. To remove poorly aligned or excessively gapped regions, each alignment was processed with TrimAl v1.4.rev22 (Capella-Gutiérrez et al. 2009) using the `-automated1` mode. Trimmed alignments were used for all subsequent phylogenetic analyses.

2.2.8.4. Maximum-Likelihood Gene Tree Construction

For each trimmed gene alignment, a maximum-likelihood (ML) phylogeny was inferred using IQ-TREE 2 v2.4.0 (Minh et al. 2020). Trees were estimated under the `GTR+G` substitution model with automatic parallelization (`-nt AUTO`) and the quiet execution mode. These 56 ML gene trees provided independent estimates of evolutionary history for each chloroplast gene.

2.2.8.5. Species Tree Inference Using ASTRAL-IV

To account for potential discordance among gene trees due to incomplete lineage sorting, the set of ML gene trees was used to infer a species tree with ASTRAL-IV (C. Zhang et al. 2025) from the ASTER package. A mapping file linking tip labels to species names ensured that multiple gene copies, if present, were appropriately grouped. The inferred species tree was midpoint-rooted and later explicitly rooted with *Physcomitrium patens* as the outgroup for divergence time analyses.

2.2.8.6. Divergence Time Estimation with TreePL

Divergence times were estimated using TreePL v2.6.3 (Smith and O'Meara 2012), which implements a penalized likelihood framework suitable for large phylogenies. The rooted ASTRAL species tree served as the input. The total number of sites in the concatenated trimmed sequence matrix was computed and supplied to TreePL. Calibration constraints were applied to the crown group of embryophytes using fossil-based age bounds of

470–515 million years. TreePL was first run in prime mode to optimize global parameters and assess convergence, after which an optimized configuration incorporating smoothing values and adaptive optimization settings from the priming step was used for the final dated analysis.

2.2.8.7. Tree Visualization

The time-calibrated species tree was visualized in RStudio v4.5.1 using the packages `ape v5.7-1`, `phytools v1.0-3`, `ggtree v3.8.2`, and `ggplot2 v3.4.2`. Trees were imported using `read.tree()`, rooted as needed, and assessed for ultrametricity. Final layout adjustments and text formatting were performed in Inkscape v1.4.2.

2.2.9. Determination of distances from MOA peaks to nearest genes

Distances between MOA-seq peaks and the nearest annotated genes were computed using a custom Bash workflow (Appendix 5.1.5). Gene coordinates were extracted from each species' Helixer GFF annotation file and merged into non-overlapping genic intervals. Peaks located on contigs without annotated genes or overlapping genic regions were removed to ensure that only intergenic peaks were analyzed. For each remaining peak, the nearest gene was identified using `bedtools closest`, and the resulting peak-to-gene distances were parsed to retain unique entries. The final output consisted of a tabular file listing the distance from each intergenic MOA-seq peak to its nearest gene. The graph was generated in R (v4.5.1) using ggplot2 (4.0.1).

2.2.10. Identification of MOA peak overlap with genomic features

MOA-seq peak positions were intersected with genomic features to determine their distribution across functional regions. For each species, genomic coordinates for 5′ UTRs, 3′ UTRs, exons, and genes were extracted from the Helixer GFF3 annotation and converted into merged BED intervals. Intronic regions were generated by subtracting UTR and exon intervals from merged gene models. Strand-specific upstream and downstream regions were defined at distances of 1 kb and 1–5 kb from annotated gene boundaries. Intergenic regions were identified by excluding peaks located within genes or within ±5 kb of gene boundaries. Peaks were restricted to chromosomes that contained annotated genes, and intersections between peaks and each genomic feature category were computed using bedtools intersect. Counts of overlapping peaks were summarized for downstream analyses. The full script used to generate these feature

intervals and compute overlaps is provided in Appendix 5.1.6. The graph was generated in R (v4.5.1) using ggplot2 (4.0.1).

2.2.11. Quantification of genomic feature sizes

Genome feature sizes were quantified using a standardized annotation-processing workflow. For each species, 5′ UTRs, 3′ UTRs, exons, and gene coordinates were extracted from the Helixer GFF3 annotation file and converted to BED format. All feature intervals were sorted and merged to produce non-overlapping feature sets. Genic regions were defined directly from merged gene intervals, and intronic regions were obtained by subtracting UTR and exon intervals from the genic coordinates. Strand-aware upstream and downstream regions were generated for each gene, including 1 kb flanking regions and extended 1–5 kb intervals. Intergenic regions were derived by subtracting all genic intervals from the full genome space using the chromosome sizes provided in the FASTA index (`.fai` file). For each feature class, the total number of bases encompassed by the merged intervals was calculated and recorded in a summary table. The complete Bash script used to generate feature intervals and compute feature-size totals is available in Appendix 5.1.7. The graph was generated in R (v4.5.1) using ggplot2 (4.0.1).

2.2.12. Quantification of MOA peaks

The total number of MOA peaks was determined by counting the number of entries in the MACS3-generated narrowPeak file using the command below.

```
wc -l file.narrowPeak
```

2.2.13. Quantification of total MOA peak width

The cumulative width of MOA peaks was calculated by summing the genomic lengths of all peaks in the narrowPeak file. For each peak, the width was defined as the difference between the end and start coordinates (column 3 minus column 2). The total peak width was computed using the following gawk command:

```
gawk 'BEGIN{wid=0}{wid=wid+($3-$2)}END{print wid}'
file.narrowPeak
```

2.2.14. Quantification of protein-coding genes

The total number of protein-coding genes was determined from the Helixer-generated GFF3 annotation file. Gene features were identified based on entries annotated as “gene,” and the total gene count was obtained by counting these annotations across the genome.

```
awk '$3 == "gene"' "GFFfile" | wc -l
```

2.2.15. Calculation of actual genome size

The assembled genome size was calculated from the FASTA index (.fai) files corresponding to each reference genome. Genome size was defined as the sum of the lengths of all indexed sequences. The total length of all contigs or chromosomes listed in the .fai file was summed to obtain the assembled genome size for each species.

```
gawk 'BEGIN{len=0} {len=len+$2} END{print len}' genome.fa.fai
```

2.2.16. Data visualization

Graphs were generated in RStudio v4.5.1 using the package `ggplot2` v3.4.2.

2.3. Results

2.3.1. Establishment of a Multi-Species MOA-seq Dataset

Plant regulatory landscapes vary widely across species with different genome sizes, ploidy levels, repeat compositions, and evolutionary histories. To systematically investigate these differences, a comprehensive multi-species MOA-seq dataset was established that includes both model plant species and non-model grasses, with a particular emphasis on the Andropogoneae clade (Figures 2.1 and 2.2).

This analysis integrated previously generated MOA-seq datasets from our group for *Arabidopsis thaliana* (unpublished), *Setaria viridis* (Myers et al. 2023), *Zea mays* (Engelhorn et al. 2025), and *Hordeum vulgare* (unpublished). For each species, MNase titration experiments were performed to identify comparable digestion levels, ensuring consistent data quality across datasets. However, due to limited tissue availability for Andropogoneae species, it was not feasible to perform full MNase titration series from M0 to M7 for each individual species. To address this limitation, Andropogoneae samples were divided into two groups based on genome size. Group 1 included species

An1 to An8, with genome sizes ranging from approximately 800 Mb to 3027 Mb, while Group 2 included species An9 to An16, with genome sizes ranging from approximately 2153 Mb to 6136 Mb (Table 2.2). For each group, 57 mg of tissue from each species were pooled, yielding a total of 456 mg of mixed tissue per group. MNase digestion was then performed on each pooled sample using 75 mg of tissue per digestion condition. Digestion was conducted using MNase concentrations corresponding to M0 (no enzyme), M2 (6.25 U/ml), M3 (12.5 U/ml), M4 (25 U/ml), M5 (50 U/ml), and M6 (100 U/ml). No detectable DNA degradation was observed in the M0 control samples (Figure 2.3). Previous MOA-seq data generated in *Arabidopsis thaliana* by our group indicated that stronger MNase digestion improves MOA peak resolution compared to lower digestion levels such as M1 or M2. However, excessive digestion resulted in loss of MOA signal, particularly in highly digested samples such as M6. Based on these observations, digestion levels M3, M4, and M5 were selected for MOA-seq library preparation for all 16 Andropogoneae species, with two biological replicates per species. Selected samples used for sequencing are shown in Figures 2.4, 2.5, and 2.6.

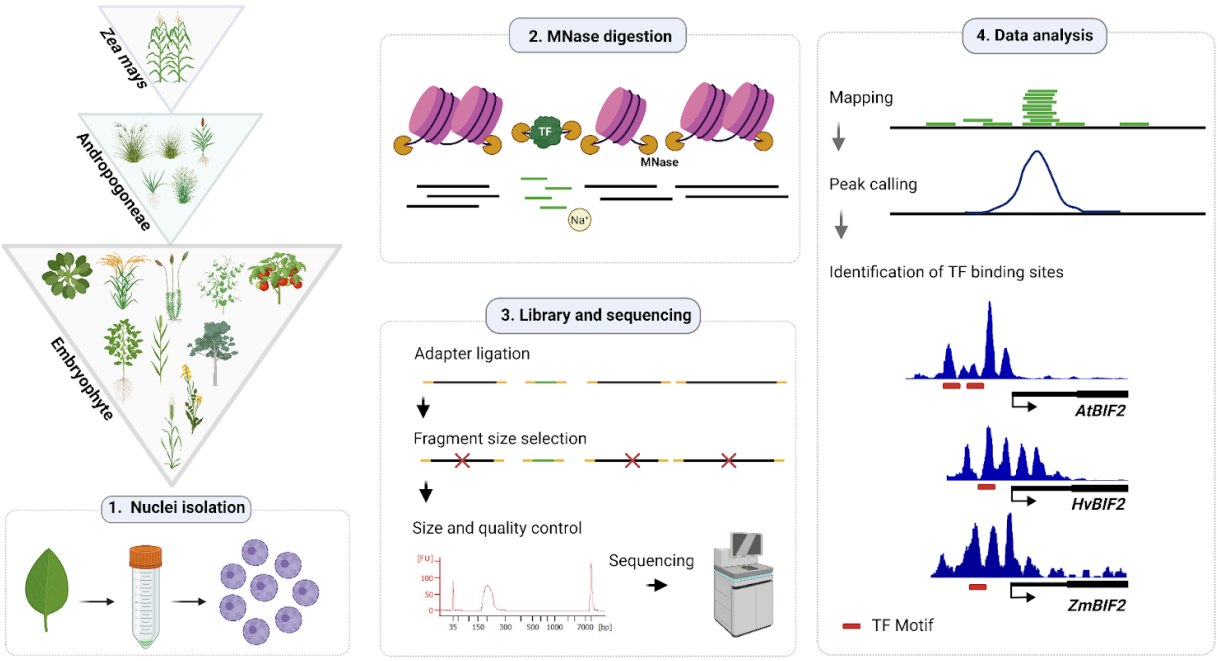


Figure 2.1: Overview of the MOA-seq procedure used to generate transcription factor footprint profiles across 27 plant species.

Leaf tissue is flash-frozen and ground in liquid nitrogen, followed by isolation of intact nuclei. Purified nuclei are subjected to micrococcal nuclease (MNase) digestion to

preferentially release TF-protected DNA fragments. After size selection, MNase-generated fragments are purified, converted into sequencing libraries, and sequenced. High-resolution MOA footprints are then identified by mapping short protected fragments to each species' genome and performing peak calling to delineate putative cis-regulatory elements. Figure created by Duong Doan using BioRender.

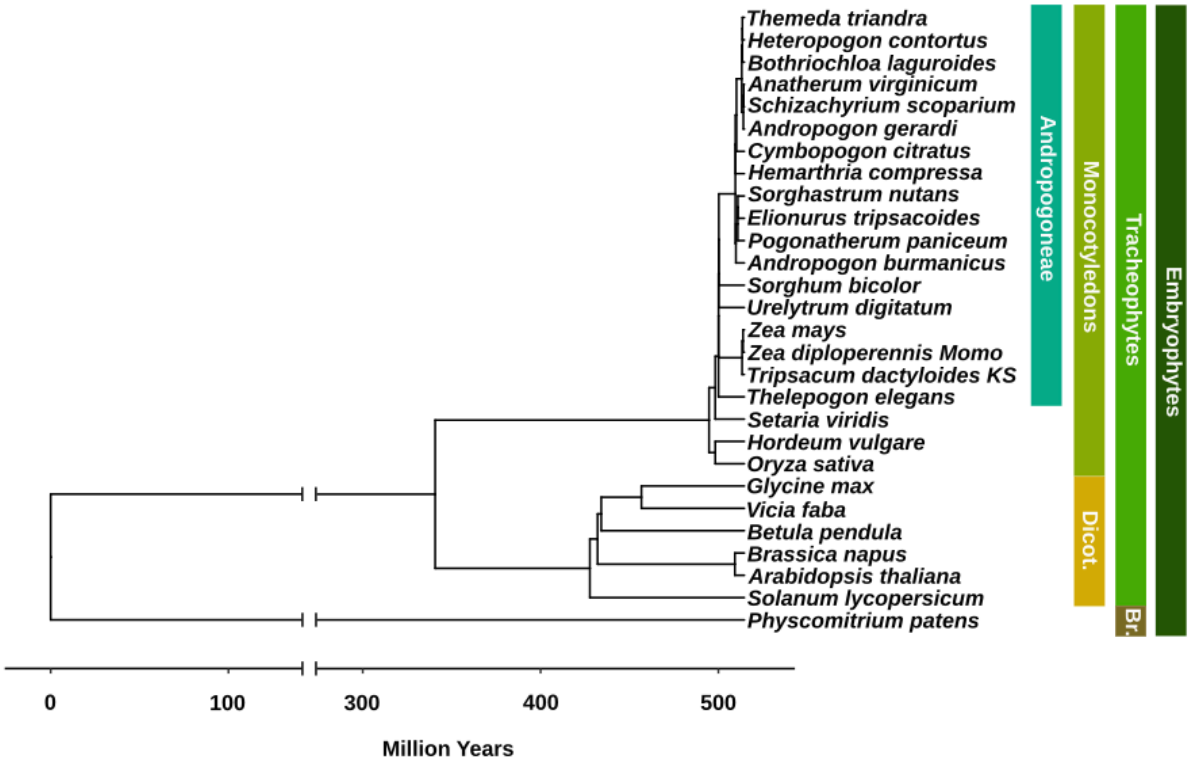


Figure 2.2: Phylogenetic relationships of the 27 plant species used for MOA-seq profiling.

A time-calibrated phylogeny illustrating evolutionary relationships among the 16 Andropogoneae species and 12 additional plant species included in this study. Andropogoneae species represent major subtribes within the clade, while the additional species span key angiosperm lineages and a bryophyte outgroup (*Physcomitrium patens*). Branch lengths reflect estimated divergence times. The species tree was inferred with ASTRAL-IV from maximum-likelihood gene trees (IQ-TREE2, GTR+G). Branch support values represent ASTRAL local posterior probabilities derived from quartet concordance among gene trees. No nonparametric bootstrap replicates were generated.

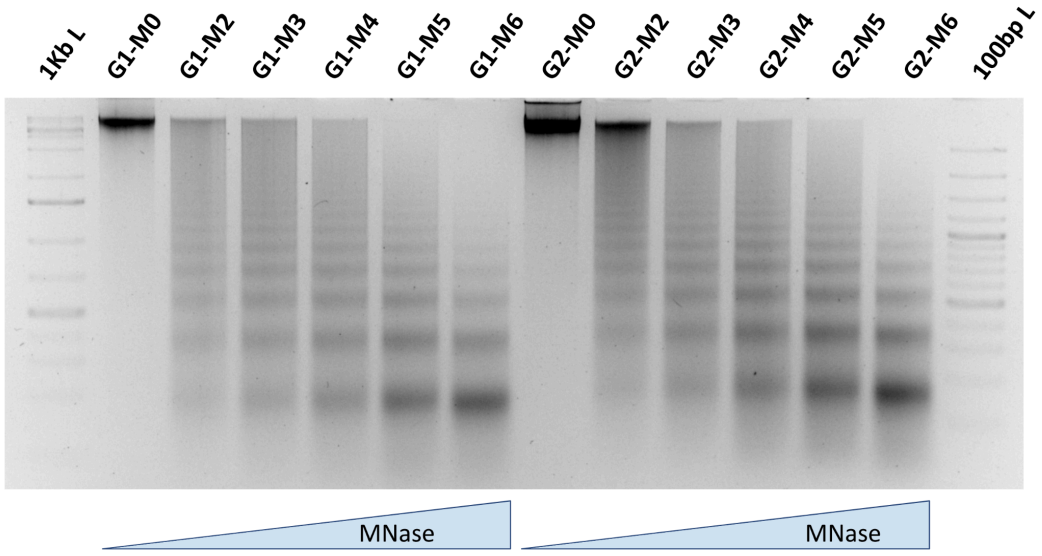


Figure 2.3: MNase partial digestion optimization for MOA-seq using pooled Andropogoneae samples.

MNase digestion profiles of pooled nuclei from two groups of Andropogoneae species (G1: An1 - An8; G2: An9 - An16) across increasing MNase concentrations (M3: 12.5 U/mL, M4: 25 U/mL, M5: 50 U/mL, M6: 100 U/mL, M0: only MDB buffer). DNA ladders flank the gel. Blue triangles indicate increasing MNase digestion strength.

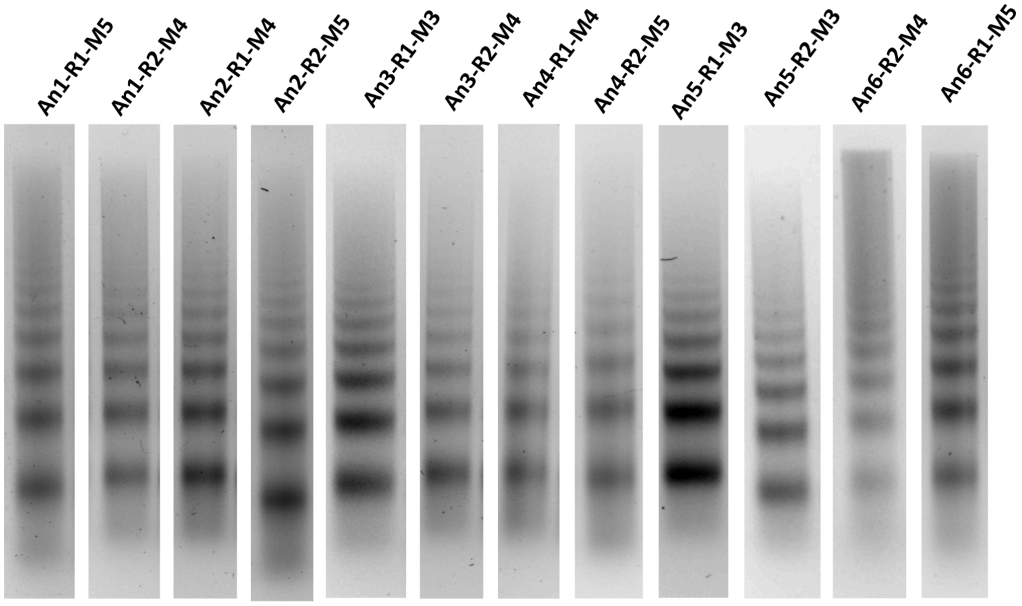


Figure 2.4: MNase partial digestion profiles of selected MOA-seq from Andropogoneae species for library preparation and sequencing, including An1 - *Themeda triandra*, An2 - *Pogonatherum paniceum*, An3 - *Anatherum virginicum*, An4 - *Heteropogon contortus*, An5 - *Thelepogon elegans, and* An6 - *Andropogon burmanicus*.

For each species, replicate samples and MNase digestion conditions are shown. Uncropped gel images for all digestion conditions are provided in Appendix 5.2.

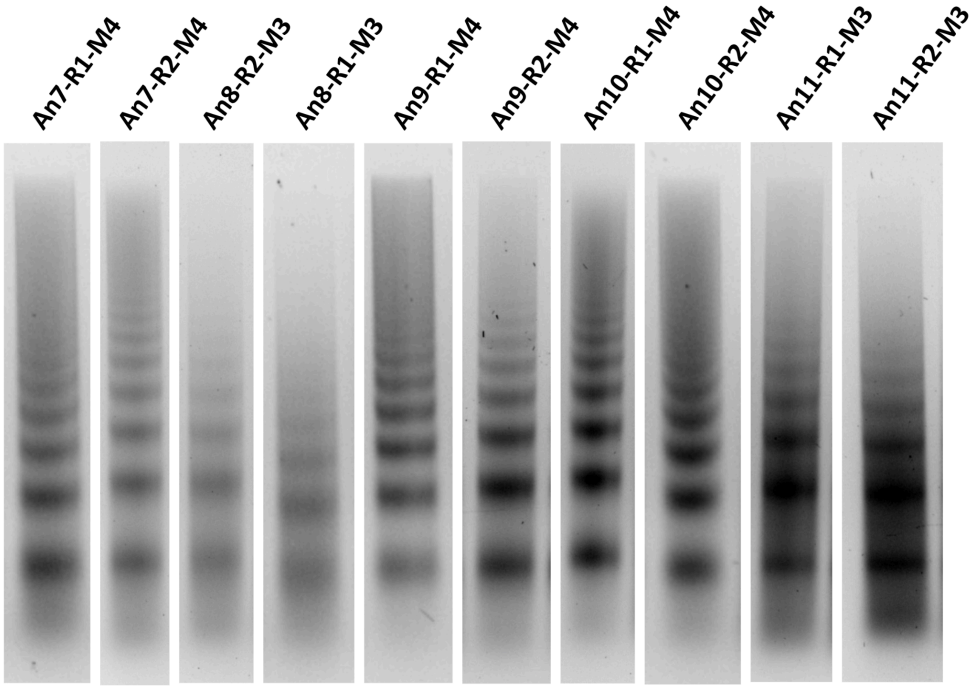


Figure 2.5: MNase partial digestion profiles of selected MOA-seq from Andropogoneae species for library preparation and sequencing, including An7 - *Schizachyrium scoparium*, An8 - *Cymbopogon citratus*, An9 - *Zea diploperennis* (Momo), An10 - *Sorghastrum nutans, and* An11 - *Tripsacum dactyloides* (KS).

For each species, replicate samples and MNase digestion conditions are M4: 25 U/mL and M5: 50 U/mL. Uncropped gel images for all digestion conditions are provided in Appendix 5.2.

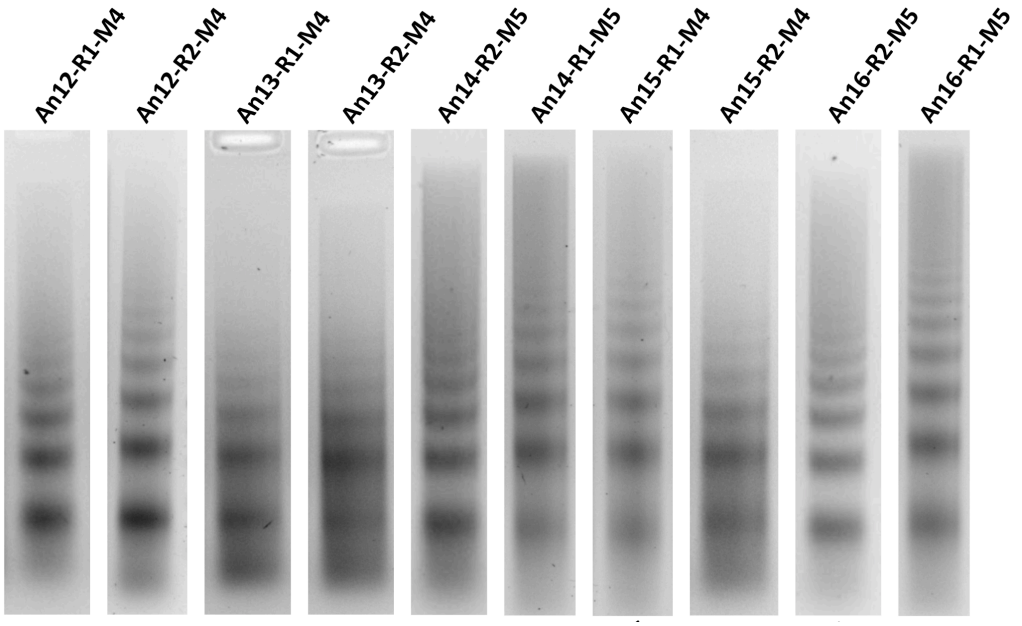


Figure 2.6: MNase partial digestion profiles of selected MOA-seq from Andropogoneae species for library preparation and sequencing, including An12 - *Andropogon gerardi*, An13 - *Hemarthria compressa*, An14 - *Bothriochloa laguroides*, An15 - *Elionurus tripsacoides, and* An16 - *Urelytrum digitatum*.
For each species, replicate samples and MNase digestion conditions are M4: 25 U/mL and M5: 50 U/mL. Uncropped gel images for all digestion conditions are provided in Appendix 5.2.

MOA-seq experiments for *Sorghum bicolor* and *Oryza sativa* were performed using the monocot protocol (Figures 2.7 and 2.8). The dicot protocol was successfully applied to *Brassica napus*, *Vicia faba*, *Glycine max*, and *Physcomitrium patens* (Figures 2.9 to 2.12). In contrast, initial attempts to perform MOA-seq on *Betula pendula* (silver birch) were unsuccessful. It was hypothesized that high levels of polyphenolic and polysaccharide compounds in woody tissues may inhibit enzymatic reactions during the MOA-seq workflow (see Appendix 5.2). To overcome this limitation, the MOA-seq protocol was further optimized for silver birch (see Methods section 2.2.3.2). Following protocol optimization, two biological replicates of *Betula pendula* were successfully generated and met quality criteria for sequencing (Figure 2.13).
Although MNase digestion was performed using a standardized protocol, digestion patterns varied between species. Digestion levels were selected based on gel inspection to obtain a comparable nucleosomal ladder with limited undigested high-molecular

weight DNA. MOA-seq optimization was initially established in maize and Arabidopsis and subsequently adapted to additional species, and efforts were made to achieve comparable digestion patterns across all samples. Residual variation in fragment distribution likely reflects differences in nuclei yield after isolation, tissue-specific chromatin properties, and species-dependent biochemical composition. These factors can influence MNase accessibility even when enzyme concentration and incubation conditions are kept constant.

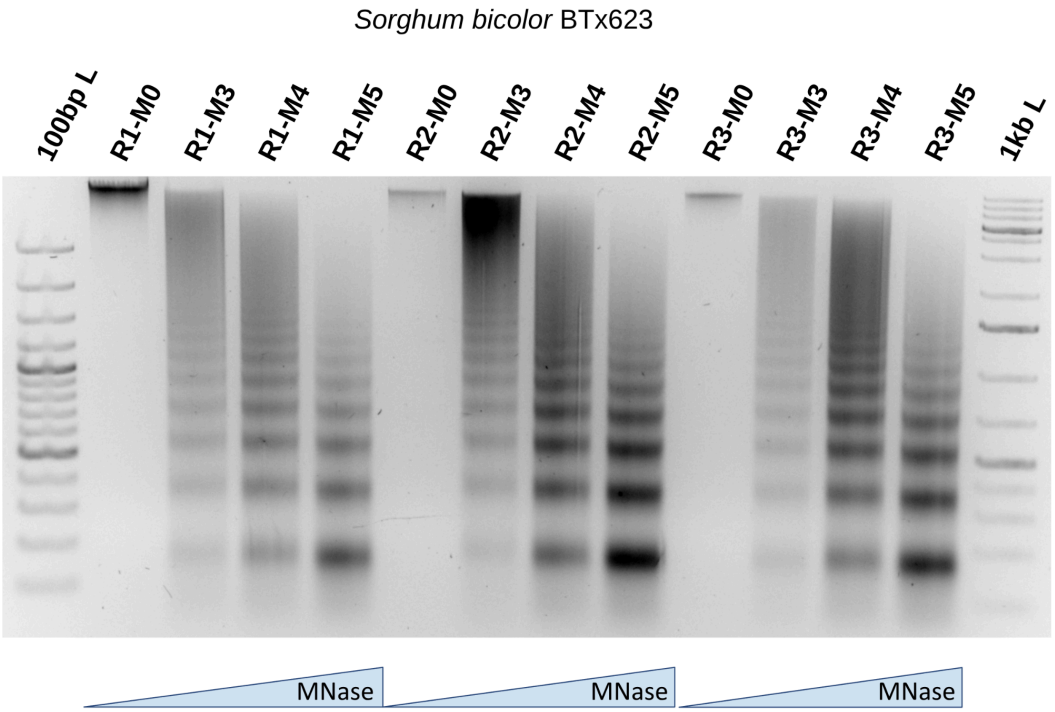


Figure 2.7: MNase partial digestion profiles of MOA-seq in *Sorghum bicolor*.
Agarose gel showing MOA-seq from three biological replicates (R1 - R3) digested with increasing MNase concentrations (M3: 12.5 U/mL, M4: 25 U/mL, M5: 50 U/mL, M0: only MDB buffer). DNA ladders are shown at both ends. Blue triangles indicate increasing MNase digestion strength. This MOA experiment was done by the bachelor student Thimo Trapp. R1-M4, R2-M4, and R3-M4 were selected for sequencing.

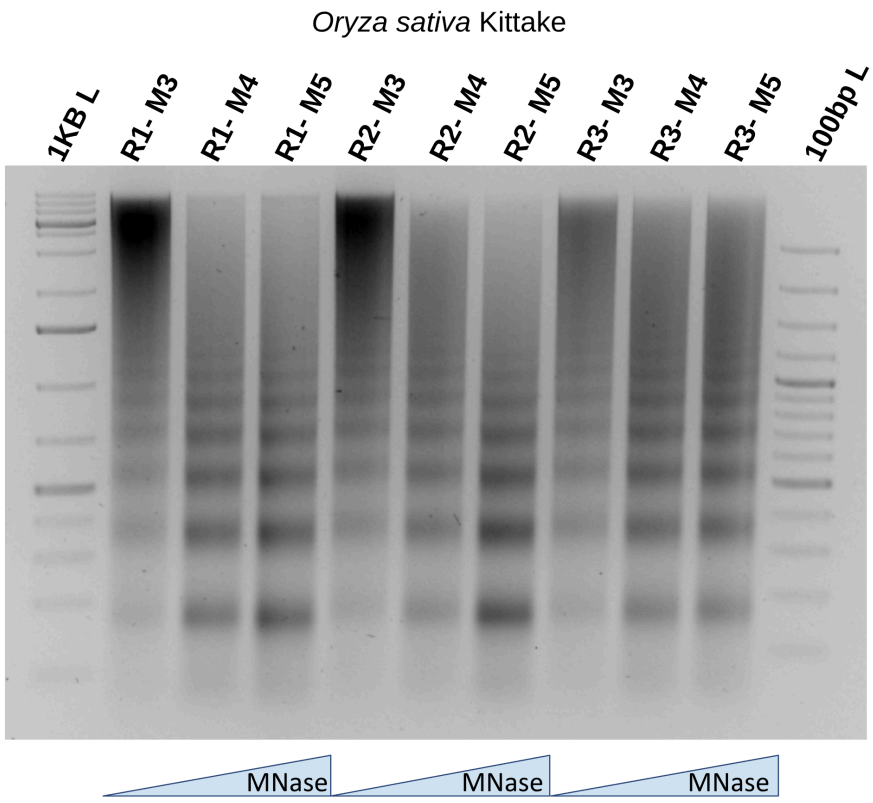


Figure 2.8: MNase partial digestion profiles of MOA-seq in *Oryza sativa*.
Agarose gel showing MOA-seq from three biological replicates (R1 - R3) digested with increasing MNase concentrations (M3: 12.5 U/mL, M4: 25 U/mL, M5: 50 U/mL). DNA ladders are shown at both ends. Blue triangles indicate increasing MNase digestion strength. This MOA experiment was performed together with Dr. Sohini Mukherjee. R1-M5, R2-M5, and R3-M5 were selected for sequencing.

Brassica Napus Express 617

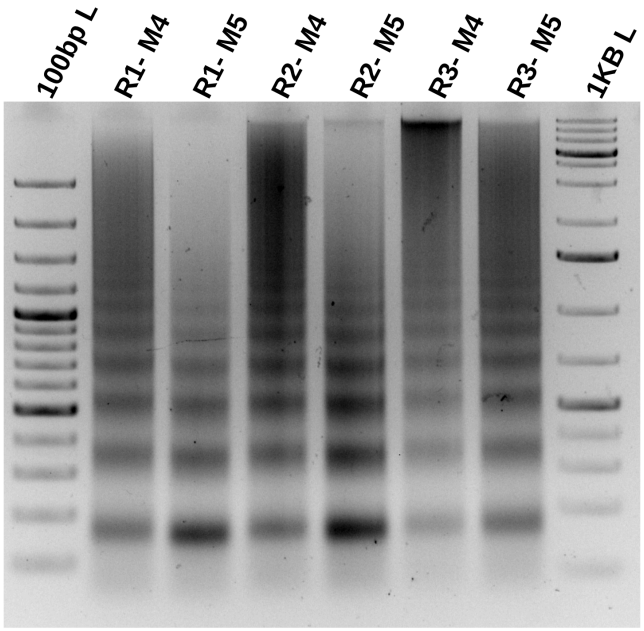


Figure 2.9: MNase partial digestion profiles of MOA-seq in *Brassica napus.*
Agarose gel showing MOA-seq from three biological replicates (R1 - R3) digested with increasing MNase concentrations (M4: 25 U/mL, M5: 50 U/mL). DNA ladders are shown at both ends. Blue triangles indicate increasing MNase digestion strength. This MOA experiment was done together with the bachelor student Thimo Trapp. R1-M5, R2-M5, and R3-M5 were selected for sequencing.

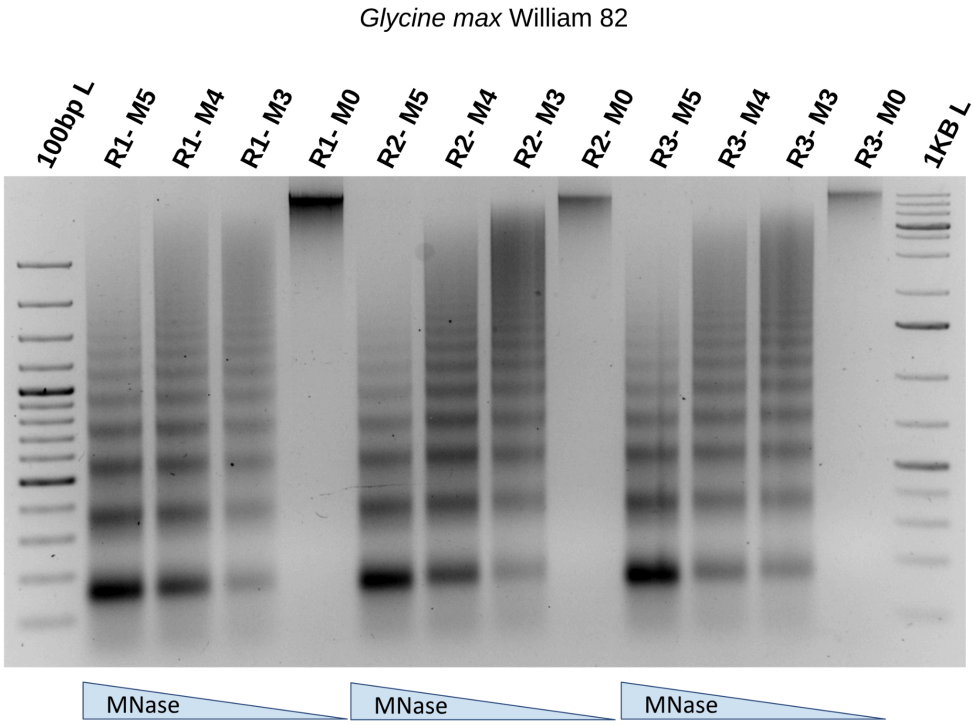


Figure 2.10: MNase partial digestion profiles of MOA-seq in *Glycine max*.
Agarose gel showing MOA-seq from three biological replicates (R1 - R3) digested with increasing MNase concentrations (M3: 12.5 U/mL, M4: 25 U/mL, M5: 50 U/mL, M0: only MDB buffer). DNA ladders are shown at both ends. Blue triangles indicate increasing MNase digestion strength. This MOA experiment was done together with the bachelor student Thimo Trapp. R1-M4, R2-M4, and R3-M4 were selected for sequencing.

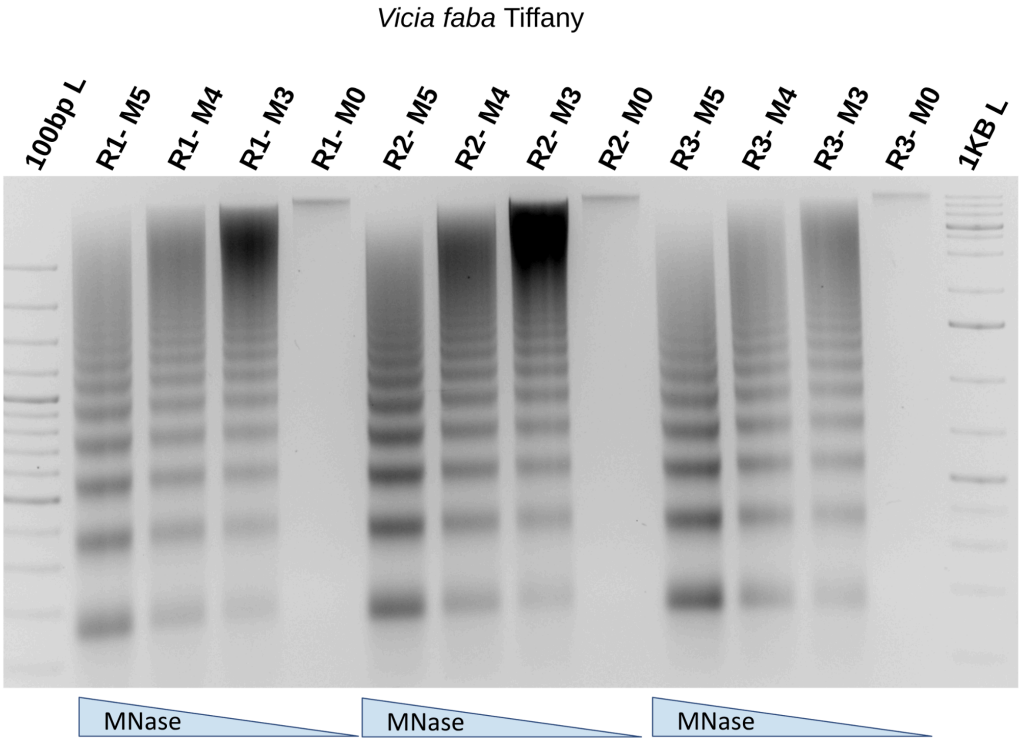


Figure 2.11: MNase partial digestion profiles of MOA-seq in *Vicia faba*.
Agarose gel showing MOA-seq from three biological replicates (R1 - R3) digested with increasing MNase concentrations (M3: 12.5 U/mL, M4: 25 U/mL, M5: 50 U/mL, M0: only MDB buffer). DNA ladders are shown at both ends. This MOA experiment was done together with the bachelor student Thimo Trapp. Blue triangles indicate increasing MNase digestion strength. R1-M5, R2-M5, and R3-M5 were selected for sequencing.

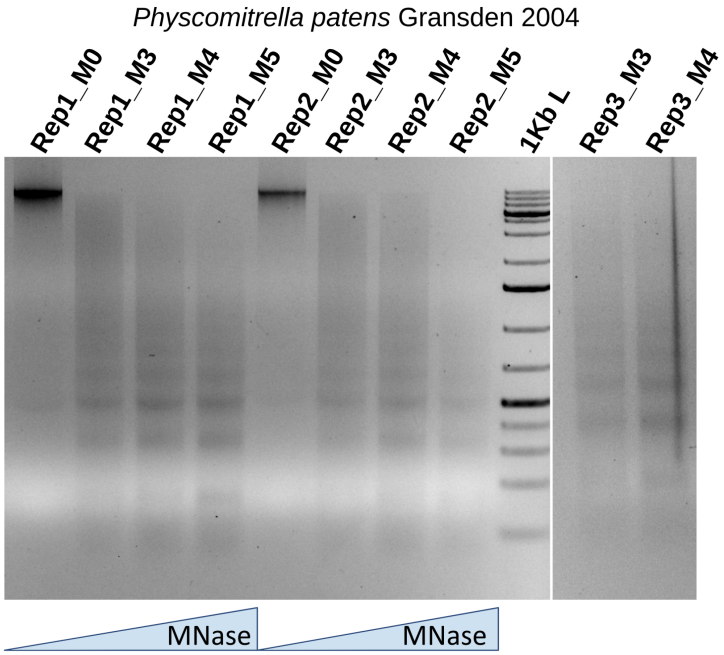


Figure 2.12: MNase partial digestion profiles of MOA-seq in *Physcomitrella patents*. Agarose gel showing MOA-seq from three biological replicates (R1 - R3) digested with increasing MNase concentrations (M3: 12.5 U/mL, M4: 25 U/mL, M5: 50 U/mL, M0: only MDB buffer). 1 Kb DNA ladders are shown in the middle. Blue triangles indicate increasing MNase digestion strength. Rep1-M4, Rep2-M4, and Rep3-M4 were selected for sequencing.

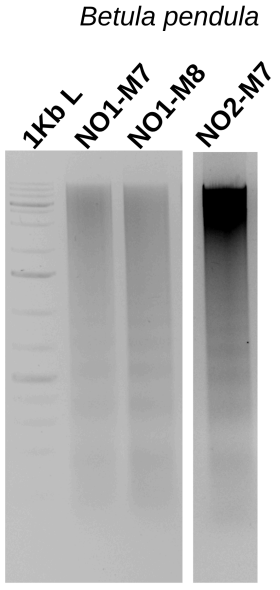


Figure 2.13: MNase partial digestion profiles of MOA-seq in *Betula pendula*.
Agarose gel showing MOA-seq from three biological replicates 1 and 2 (NO stands for sample under mock condition), digested with increasing MNase concentrations (M7: 200 U/mL and M8: 400 U/mL). Kb DNA ladders are shown in the middle. Blue triangles indicate increasing MNase digestion strength. NO1-M7 and NO2-M7 were selected for sequencing. Uncropped gel images for all digestion conditions are provided in Appendix 5.1.9.

Table 2.2: (Listed as Table 2.3.1 in final outline) Genome size and sequencing reads mapping statistics for 28 species (with two haplotypes of *Zea mays*).

Species	Raw reads	Mapped reads	Unique mapped reads	Actual genome size (bp)
Arabidopsis thaliana	127201626	124862772	106522084	119667750
Oryza sativa	332837743	327281564	282179651	381570803
Betula pendula	287926954	258101587	162333888	408610979
Physcomitrium patens	493976688	416437417	356325043	481750213
Setaria viridis	341592105	314083713	272275424	397277387
Sorghum bicolor	342852311	315558082	232721024	719894357
Solanum lycopersicum	452737535	428787752	346777601	828472512
Glycine max	448561340	423643421	354503659	925095059
Brassica napus	386815466	36365770	241680704	1011128777
Zea mays B73	292479817	252048560	148030659	2182075994
Zea mays Mo17	303357622	263088973	166914316	2178604320
Hordeum vulgare	642256389	560356338	335647354	4225577519
Vicia faba	1030121997	851267823	451669511	11366676527
An1_*Themeda triandra*	248127139	183782818	76489622	1034932660
An2_*Pogonatherum paniceum*	268164736	237218836	151086160	799622294
An3_*Anatherum virginicum*	202677650	188722645	128162323	905553168
An4_*Heteropogon contortus*	262354914	215718241	77530002	2089701397
An5_*Thelepogon elegans*	197423729	177315273	119435204	3027872127
An6_*Andropogon burmanicus*	225751964	207008012	69900934	1946409527
An7_*Schizachyrium scoparium*	235312592	215331172	105108650	2543548284

An8_*Cymbopogon citratus*	200164303	178485154	62368252	2391180077
An9_*Zea diploperennis (Momo)*	226393699	194733901	103192841	2153424507
An10_*Sorghastrum nutans*	234759910	201905449	71444630	4374693094
An11_*Tripsacum dactyloides KS*	267195765	209484482	102658937	2597784422
An12_*Andropogon gerardi*	266572896	213178886	65916777	5258294463
An13_*Hemarthria compressa*	242842483	202731922	53757027	6136823671
An14_*Bothriochloa laguroides*	224549267	205927529	121206727	3243580366
An15_*Elionurus tripsacoides*	199687185	184835268	124463946	4764026201
An16_*Urelytrum digitatum*	437499545	411860768	276897861	4703656903

After a quality investigation with FASTQC, sequencing output and mapping statistics for all 28 species included in this study are summarized in Table 2.2. Raw sequencing depth varied substantially across species, reflecting differences in genome size. Raw read counts ranged from approximately 130 million reads in *Arabidopsis thaliana* to just over 1 billion reads in *Vicia faba*. Intermediate sequencing depths were obtained for most diploid angiosperms and Andropogoneae species, typically between ~200 million and ~450 million reads.

Across species, a high proportion of raw reads aligned to their respective reference genomes, with mapping rates typically exceeding ~80% (Table 2.2) when reads were aligned to reference assemblies appropriate for each species' ploidy level and genome assembly type (Table 2.1). Small and moderately sized genomes, including *Arabidopsis thaliana*, *Oryza sativa*, *Physcomitrium patens*, and *Setaria viridis*, showed particularly high alignment efficiencies, with more than 90% of reads mapping to the reference. In contrast, mapping efficiency was reduced in large, repeat-rich genomes such as *Zea mays*, *Hordeum vulgare, Vicia faba*, and several Andropogoneae species, reflecting increased sequence redundancy, structural complexity, and incomplete assembly of highly repetitive regions. Following alignment, a substantial fraction of mapped reads remained uniquely mapped, ranging from approximately 60% to over 85% depending on genome size and complexity. Species with smaller or better-annotated genomes retained the highest proportion of uniquely mapped reads, whereas large-genome Andropogoneae species and *Vicia faba* exhibited increased multi-mapping, consistent with extensive transposable element content and recent duplication events.

Together, these results indicate that the sequencing depth and alignment performance achieved in this study are sufficient and comparable across species for downstream

analyses of MOA footprints and cis-regulatory landscapes, while appropriately reflecting known differences in genome size, repeat content, and assembly completeness.

2.3.2. Profiling TF cis-regulatory regions evolution throughout the Embryophytes clade

To examine the relationship between genome size and the abundance of regulatory sequences, a comparative analysis was conducted across 14 species with 15 high-contiguity reference genome assemblies, spanning more than two orders of magnitude in genome size, from approximately 120 Mb in *Arabidopsis thaliana* to approximately 13 Gb in *Vicia faba*. This subset includes both monocot and dicot species and represents a range of genome architectures, ploidy levels, and repeat contents. To assess whether the same trends were observable when genome assembly quality and contiguity were more variable, the analysis was subsequently extended to the full dataset of 28 species included in this study (Figure 2.14a, b and 2.15a, b).

Across all 28 species, the total number of MOA-seq peaks varied substantially between species. The smallest genome, *Arabidopsis thaliana*, contained 63,844 MOA peaks, whereas *Heteropogon contortus* (An4) exhibited the highest number of peaks at 1,042,472. Intermediate values were observed in species with moderately sized genomes, such as *Oryza sativa*, *Setaria viridis*, and *Solanum lycopersicum*. Notably, the largest diploid genome in this subset, *Vicia faba*, harbored 569,964 MOA peaks, whereas the allohexaploid *Andropogon gerardi* (An12), despite having a smaller genome size, contained 753,006 peaks (Figure 2.14b). These observations indicate that genome size alone does not fully explain variation in peak number and suggest additional contributions from ploidy and genome organization. Overall, a strong positive correlation was observed between genome size and the total number of MOA-seq peaks across the 14 species with good quality genome assemblies (Pearson's r = 0.8; Figure 2.14a). This trend indicates that, in well-assembled genomes, larger genomes tend to harbor a greater number of regulatory regions detected by MOA-seq. The relationship is consistent with the expansion of intergenic regions and distal regulatory space accompanying genome size increase (Figure 2.21). When the same analysis was extended to all 28 species, including those with more fragmented or lower-quality genome assemblies, the strength of the correlation was substantially reduced (r = 0.4; Figure 2.14b).

In addition to peak counts, it is also important to examine whether genome size also correlates with the total genomic span of regulatory regions, calculated as the summed length of all MOA-seq peaks per species (Figure 2.15a, b). Among the 14 species with high-quality genome assemblies, total peak length increased with genome size, showing a strong positive correlation (r = 0.76; Figure 2.15a). Species with larger genomes not only contained more peaks but also exhibited a greater cumulative amount of MOA peak width, suggesting an expansion of regulatory sequence space rather than a simple increase in peak number alone. However, as observed for peak counts, this relationship was markedly attenuated when all 28 species were included (r = 0.2; Figure 2.15b). The reduced correlation indicates that estimates of total regulatory space are particularly sensitive to genome assembly completeness and repetitive sequence representation. Large genomes with extensive transposable element content and fragmented assemblies tended to show greater variability in total peak length, even when peak numbers were relatively high. Taken together, Figures 2.14 and 2.15 demonstrate that genome size is positively associated with both the number and total length of MOA-seq peaks when high-quality reference genomes are available.

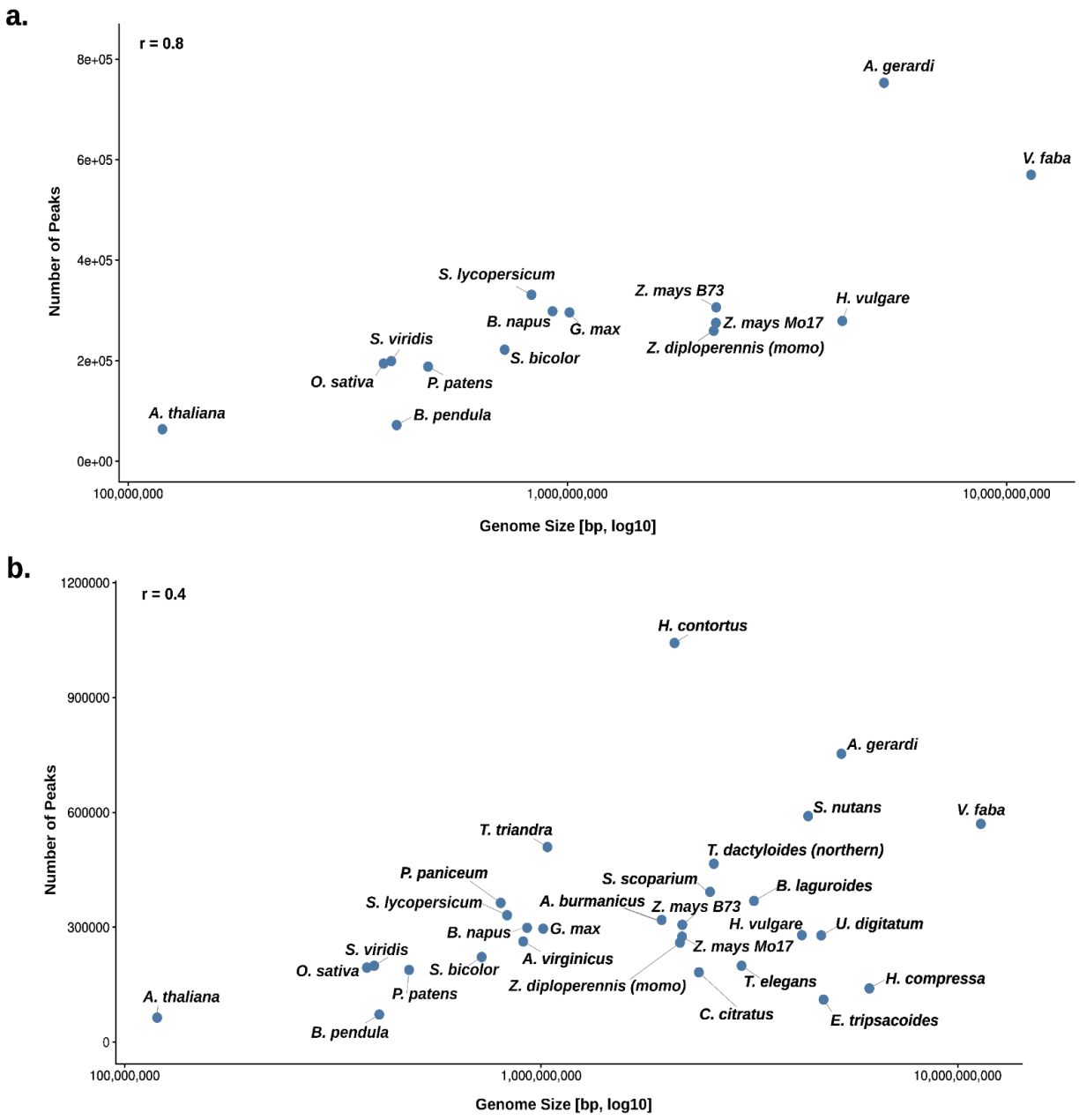


Figure 2.14: Correlation between genome size and total MOA-seq peak counts across species.

(a) Scatter plot showing genome size (x-axis, log10 scale) and the number of MOA-seq peaks (y-axis) for the subset of 14 species with well-assembled, high-contiguity reference genomes. Peaks were identified using MACS3, and the correlation coefficient was calculated in R. (b) Scatter plot including all 28 species, including those with non-chromosome-level genome assemblies. Variation in assembly contiguity may

influence peak detection sensitivity. Genome sizes are plotted on the x-axis (log10), and the corresponding MACS3 peak counts are shown on the y-axis. Each data point corresponds to one species.

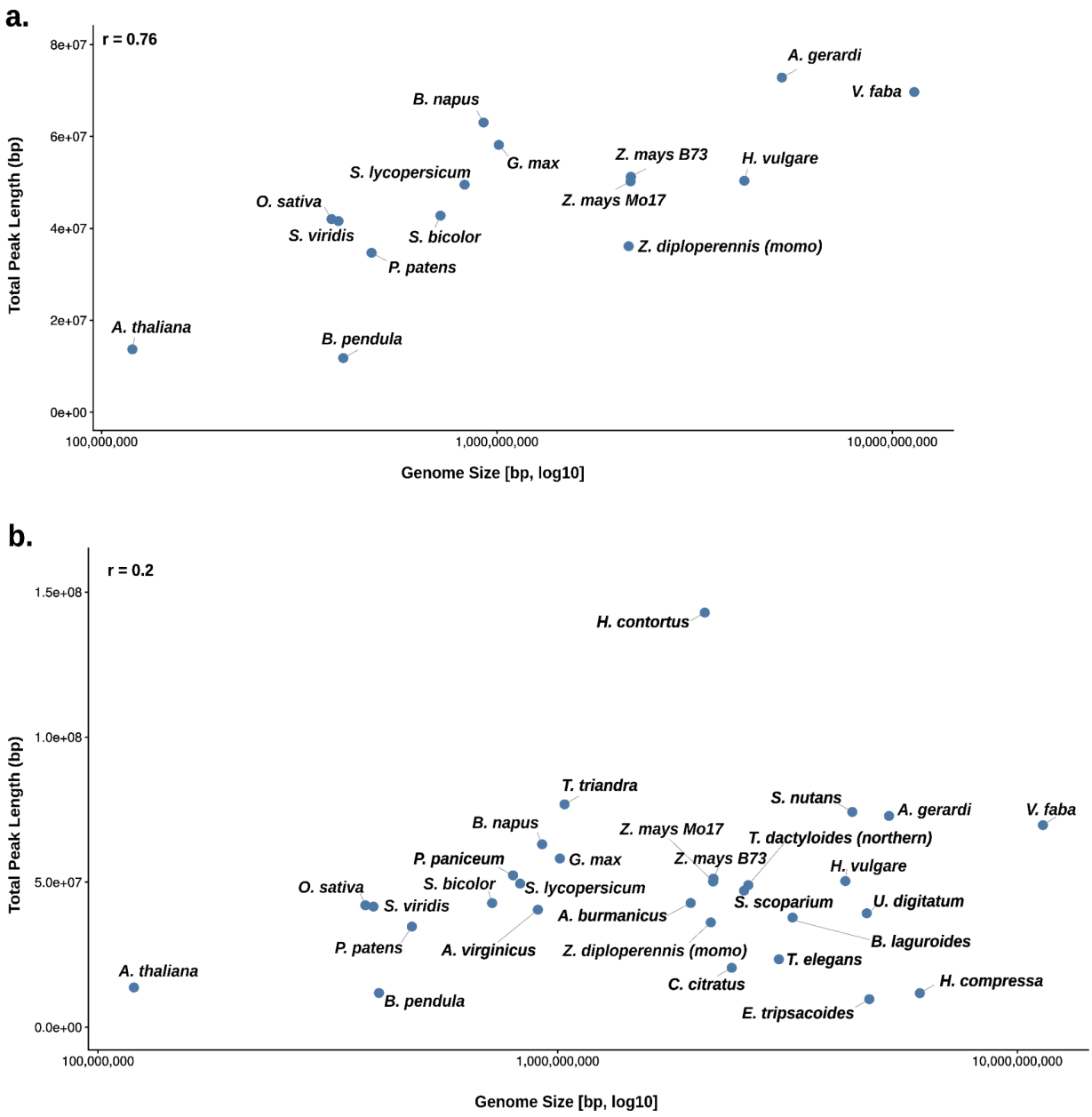


Figure 2.15. Relationship between total MOA-seq peak length and genome size across species.

(a) Scatter plot showing the relationship between genome size (log10 scale, x-axis) and the total length of MOA-seq peaks (y-axis) for the subset of species with high-contiguity

genome assemblies. (b) Same analysis including all 28 species used in this study. Total peak length was obtained by summing the genomic span of all MOA-seq peaks identified by MACS3 (v3.0.0a7). Genome sizes were derived from the reference assemblies listed in Tables 2.1 and 2.2. Pearson correlation coefficients (r) were calculated in R.

To determine whether the expansion of regulatory regions in larger genomes is associated with an increase in gene number, the relationship between genome size and the number of protein-coding genes across species was examined (Figure 2.16). Among the 14 species with high-contiguity genome assemblies, gene number increased moderately with genome size (r = 0.5; Figure 2.16a). *Andropogon gerardi* contains a markedly higher number of protein-coding genes (74,237) than *Vicia faba*, despite having a substantially smaller genome. This likely reflects the allohexaploid nature of *Andropogon gerardi*, in which whole-genome duplication has resulted in extensive retention of homoeologous gene copies. In contrast, genome expansion in *V. faba* is largely driven by repetitive sequence accumulation rather than increased gene content. Small genomes such as *Arabidopsis thaliana* encoded approximately 27000 genes, whereas large genomes such as *Vicia faba* and *Hordeum vulgare* contained on the order of 35825 and 36821 genes. However, this increase was clearly non-linear, as genome sizes varied by more than two orders of magnitude while gene numbers varied by less than threefold. When the analysis was extended to all 28 species, including those with lower-quality assemblies, a similar moderate correlation was observed (r = 0.5; Figure 2.16b), indicating that gene number alone does not scale proportionally with genome size.

The next question is whether the number of regulatory regions, as measured by MOA-seq peaks, is more directly related to gene number than to genome size. Across the 14 species with high-quality genomes, a strong positive correlation was observed between the total number of MOA peaks and the number of protein-coding genes (r = 0.7; Figure 2.18a). Species with more genes generally exhibited more MOA peaks, suggesting that a substantial fraction of accessible chromatin regions are associated with gene regulation. This relationship persisted, although with increased scatter, when all 28 species were included (r = 0.5; Figure 2.18b). The reduced correlation likely reflects increased heterogeneity in genome assembly quality and repeat content, which can affect both gene annotation and peak detection.

To assess whether regulatory complexity per gene changes with genome size, the relationship between the average number of MOA peaks per gene and the genome size was examined (Figure 2.17). Among species with high-contiguity assemblies, the average number of peaks per gene increased with genome size (r = 0.54; Figure 2.17a). Small genomes such as *Arabidopsis thaliana* showed 2.3 peaks per gene, whereas larger genomes, including *Zea mays*, *Hordeum vulgare*, and *Vicia faba*, exhibited substantially higher values, in some cases exceeding 7-10 peaks per gene. Despite its elevated gene number, *Andropogon gerardi* exhibits a relatively low average number of MOA peaks per gene (3.4), indicating that gene duplication does not translate into a proportional increase in detectable regulatory regions (Figure 2.17a). This pattern is consistent with regulatory divergence and partial silencing of duplicated genes following polyploidization, whereby only a subset of homoeologous genes remains transcriptionally active and associated with accessible chromatin or regulatory regions (Phillips et al. 2025). This comparison pattern indicates that regulatory regions accumulate disproportionately relative to gene number as genomes expand. However, when all 28 species were included, this relationship was no longer evident (r = -1; Figure 2.17b). The absence of correlation reflects increased variability among large, repeat-rich genomes, where extensive intergenic expansion and assembly fragmentation complicate the accurate assignment of regulatory regions to genes. Together, these analyses demonstrate that while gene number increases only modestly with genome size, the total number of regulatory regions and the regulatory complexity per gene increase more substantially. This decoupling of gene number from regulatory sequence expansion suggests that genome size evolution in plants is accompanied primarily by the expansion of regulatory and intergenic space rather than by proportional increases in coding capacity.

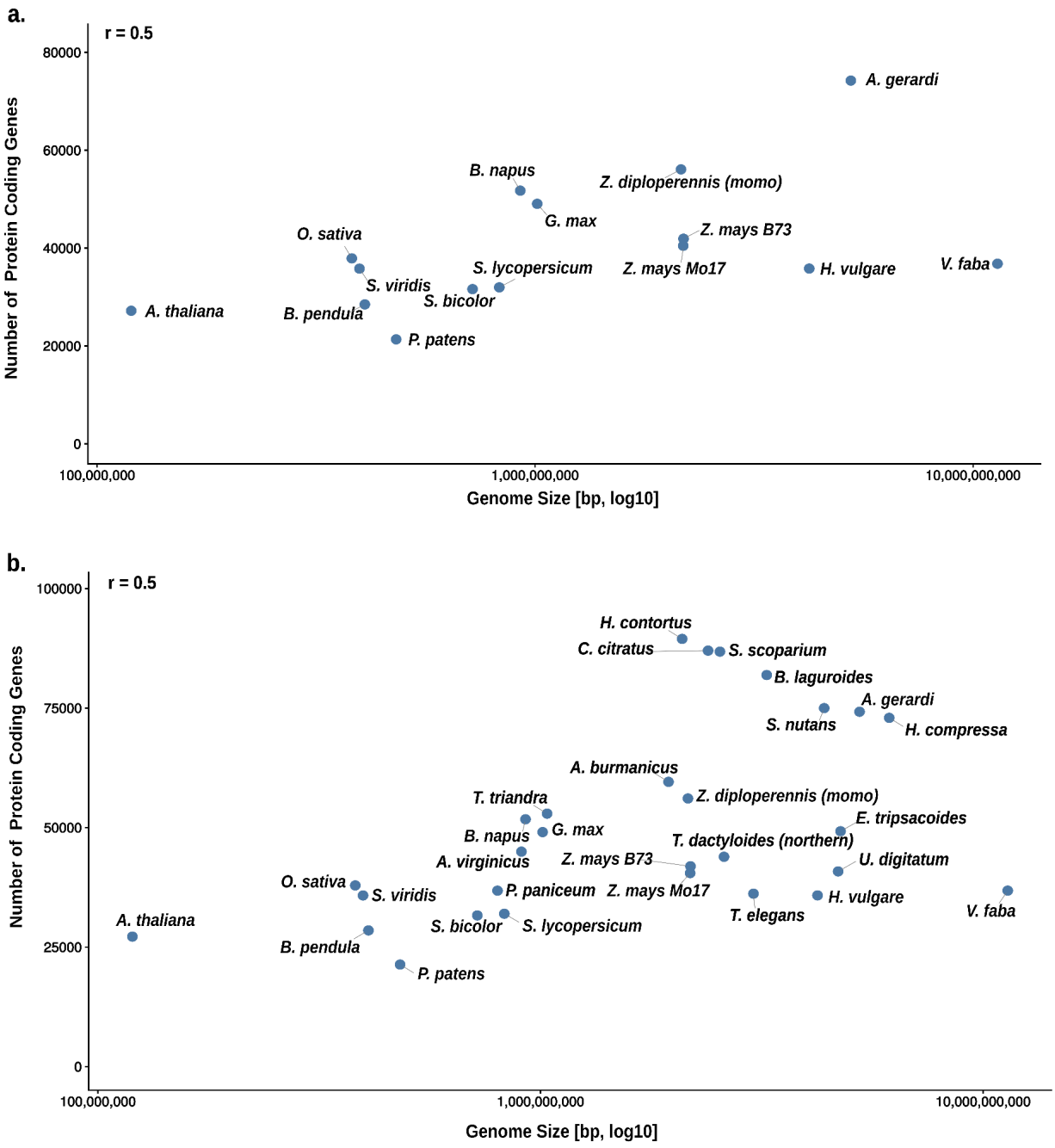


Figure 2.16: Correlation between total number of protein-coding genes and genome size across species.

(a) Scatter plot showing the relationship between genome size (log10 scale, x-axis) and the total number of protein-coding genes (y-axis) for the subset of species with high-contiguity genome assemblies. (b) Same analysis including all 28 species used in this study. Genome sizes were derived from the reference assemblies listed in Tables 2.1 and 2.2. Pearson correlation coefficients (r) were calculated in R.

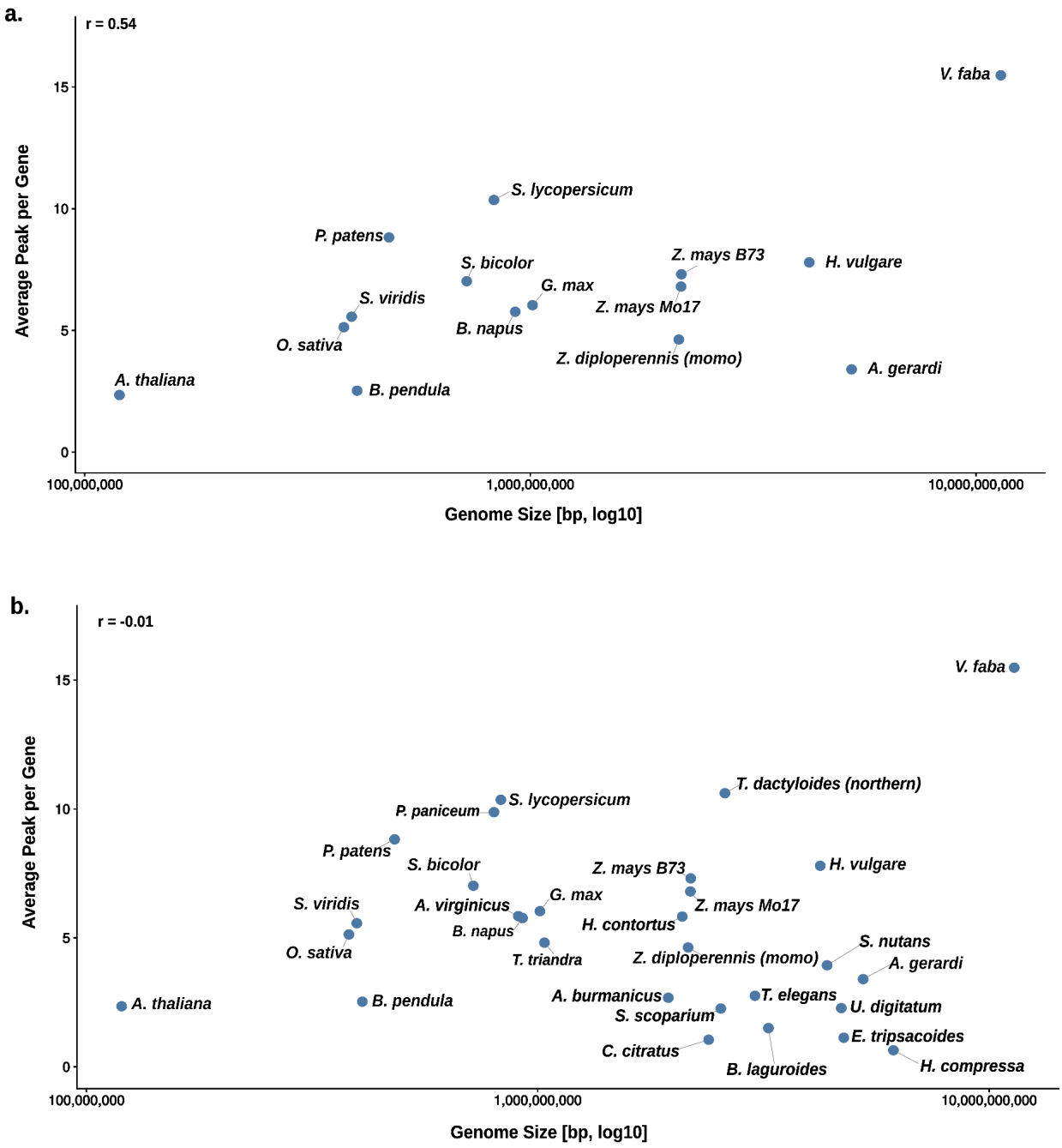


Figure 2.17: Correlation between average MOA peak per gene and genome size across species.

(a) Scatter plot showing the relationship between genome size (log10 scale, x-axis) and the average MOA peak per gene (y-axis) for the subset of species with high-contiguity genome assemblies. (b) Same analysis including all 28 species used in this study. Genome sizes were derived from the reference assemblies listed in Tables 2.1 and 2.2.

Pearson correlation coefficients (r) were calculated in R.

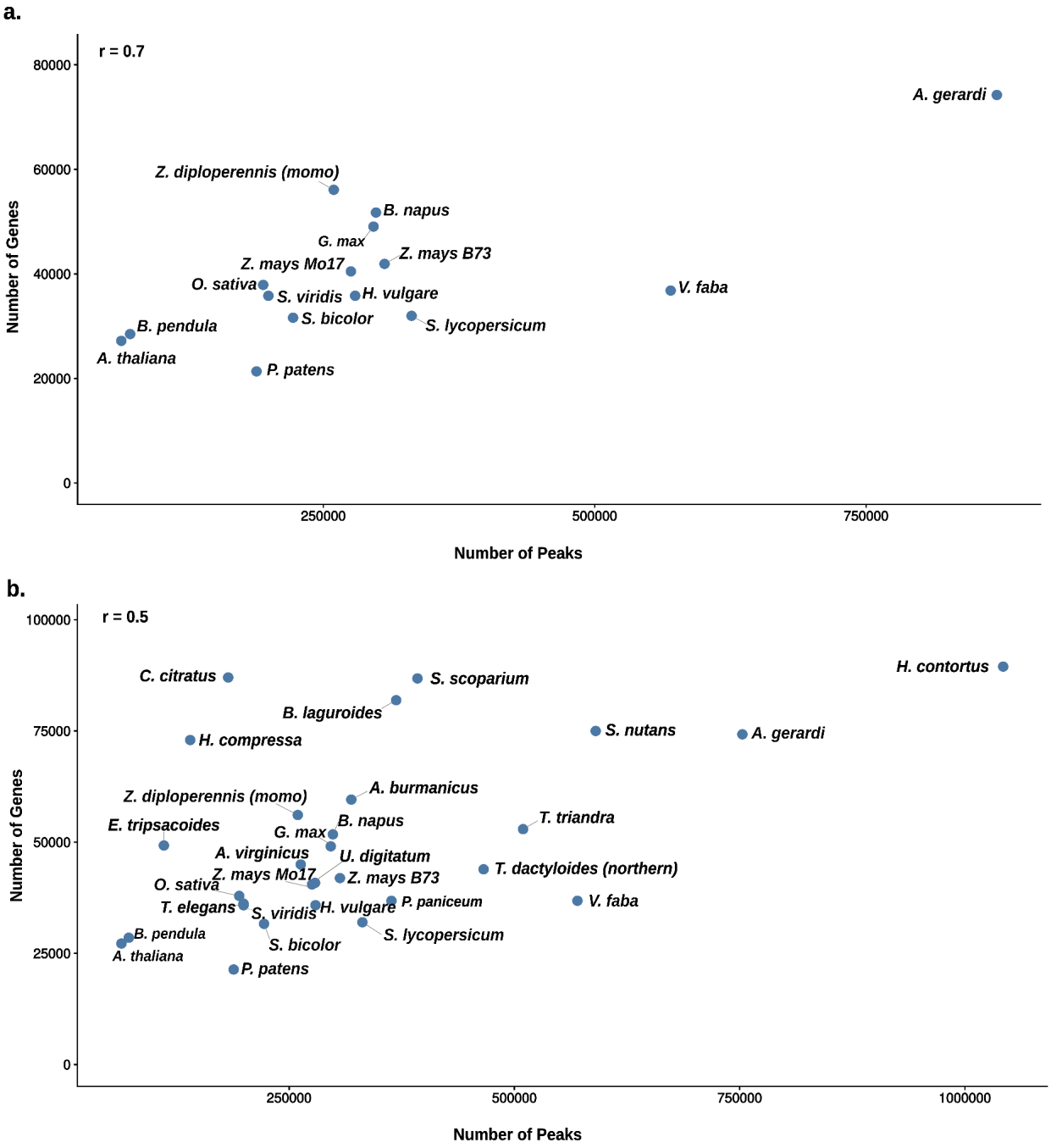


Figure 2.18: Correlation between total number of protein-coding genes and total number of MOA peaks.

(a) Scatter plot showing the relationship between total number of MOA peaks (x-axis) and the total number of protein-coding genes (y-axis) for the subset of species with high-contiguity genome assemblies. (b) Same analysis including all 28 species used in

this study. Pearson correlation coefficients (r) were calculated in R.

To assess whether genome expansion is associated with changes in the genomic positioning of regulatory elements, the distance between MOA peaks and the nearest transcription start site (TSS) was examined across 14 species with high-contiguity genome assemblies (Figure 2.19). A clear shift in MOA peak-TSS distance distributions was observed with increasing genome size. Species with compact genomes, such as *Arabidopsis thaliana* and *Oryza sativa*, exhibited a strong enrichment of MOA peaks within a few kilobases of annotated TSSs, consistent with promoter-proximal regulatory architecture. In contrast, larger and repeat-rich genomes, including *Hordeum vulgare* and *Vicia faba*, showed peak density distributions shifted toward substantially greater distances from TSSs, with many peaks located tens to hundreds of kilobases away.

Notably, *Andropogon gerardii* deviated from this general trend. Again, despite its large genome size, *Andropogon gerardii* exhibited a comparatively promoter-proximal distribution of MOA peaks, with peak to TSS distances more similar to those observed in species with smaller or more gene-dense genomes (Figure 2.19). This pattern coincides with the high gene count of *Andropogon gerardii* and its allohexaploid genome structure, suggesting that genome expansion in this lineage has been driven in part by polyploidization rather than extensive expansion of intergenic space. Consequently, regulatory elements in *Andropogon gerardii* appear to remain closely associated with genes in linear genomic distance, despite overall genome enlargement. In addition, in the largest and most repeat-rich genomes, a subset of MOA peaks remained close to transcription start sites. For example, both *Hordeum vulgare* and *Vicia faba* retained a distinct enrichment of MOA peaks within approximately 1 Kb of TSSs (Figure 2.19), consistent with accessibility at core promoters, whereas the majority of peaks shifted toward more distal positions, likely reflecting intergenic expansion driven by TEs. This pattern suggests that in most of the cases, genome size expansion is accompanied by a relative increase in distal regulatory regions, likely reflecting the accumulation of TEs and other non-coding sequences that expand intergenic space. As a consequence, regulatory elements remain functionally associated with genes but are positioned farther away in linear genomic distance.

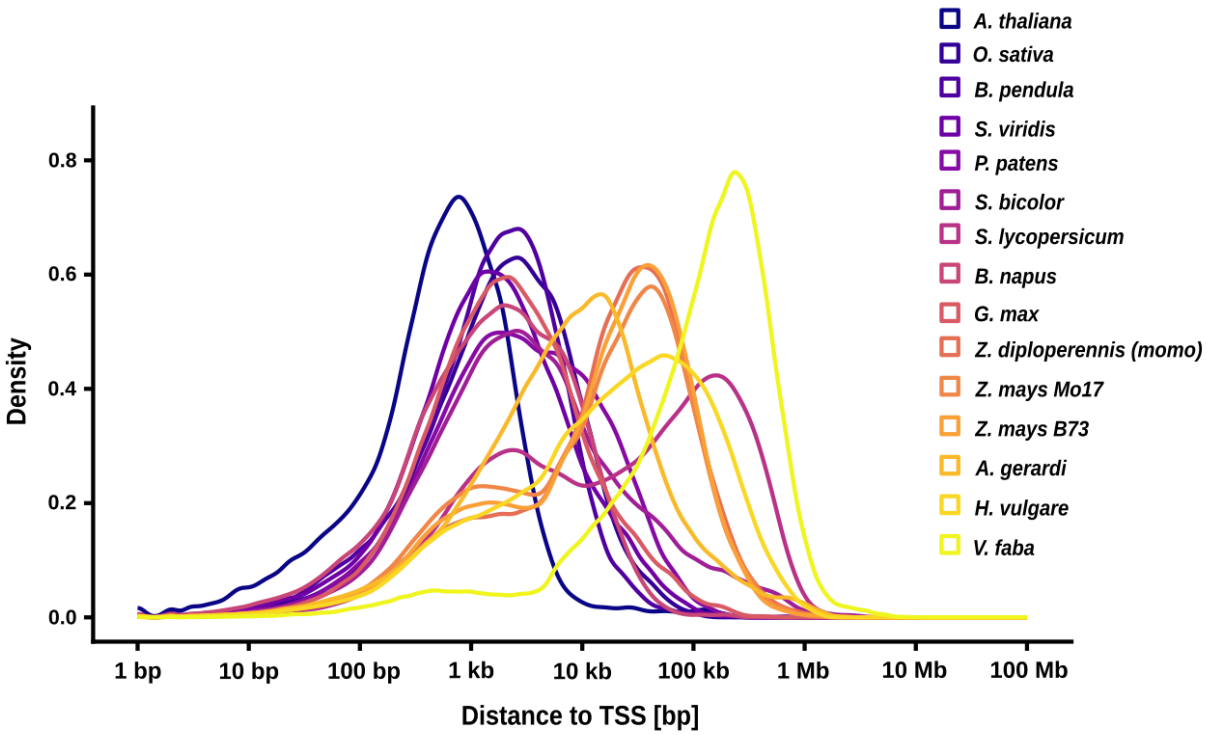


Figure 2.19: Distribution of distances between MOA peaks and the nearest transcription start site (TSS) across 14 plant species.

Kernel density plots show the genomic distance from each MOA peak to the nearest annotated TSS in 14 species with high-contiguity genome assemblies. The x-axis represents distance to the TSS in base pairs on a log10 scale, and the y-axis indicates the density of MOA peaks. The original script for calculating the distance to genes was provided and supported by Dr. Julia Engelhorn and Amelie Kok. The script was adapted and additional features added by me.

While genome expansion tends to coincide with increased intergenic sequence content in most of the species (Figure 2.21), the distribution of MOA peaks shows a similar overall trend but does not scale linearly with genome size, indicating that regulatory element placement is influenced by additional factors beyond genome size alone (Figure 2.20). While promoter-proximal and genic peaks were present in all species, large genomes such as *Hordeum vulgare* and *Vicia faba* showed a substantially higher proportion of intergenic MOA peaks. Notably, several species deviated from a simple genome size dependent trend, including *Brassica napus* (allotetraploid), *Glycine max*

(partially diploidized tetraploid), and *Andropogon gerardi* (allohexaploid). These species share histories of polyploidy and subsequent genome reorganization, suggesting that ploidy and post-polyploid diploidization may modulate regulatory element distribution independently of the genome size.

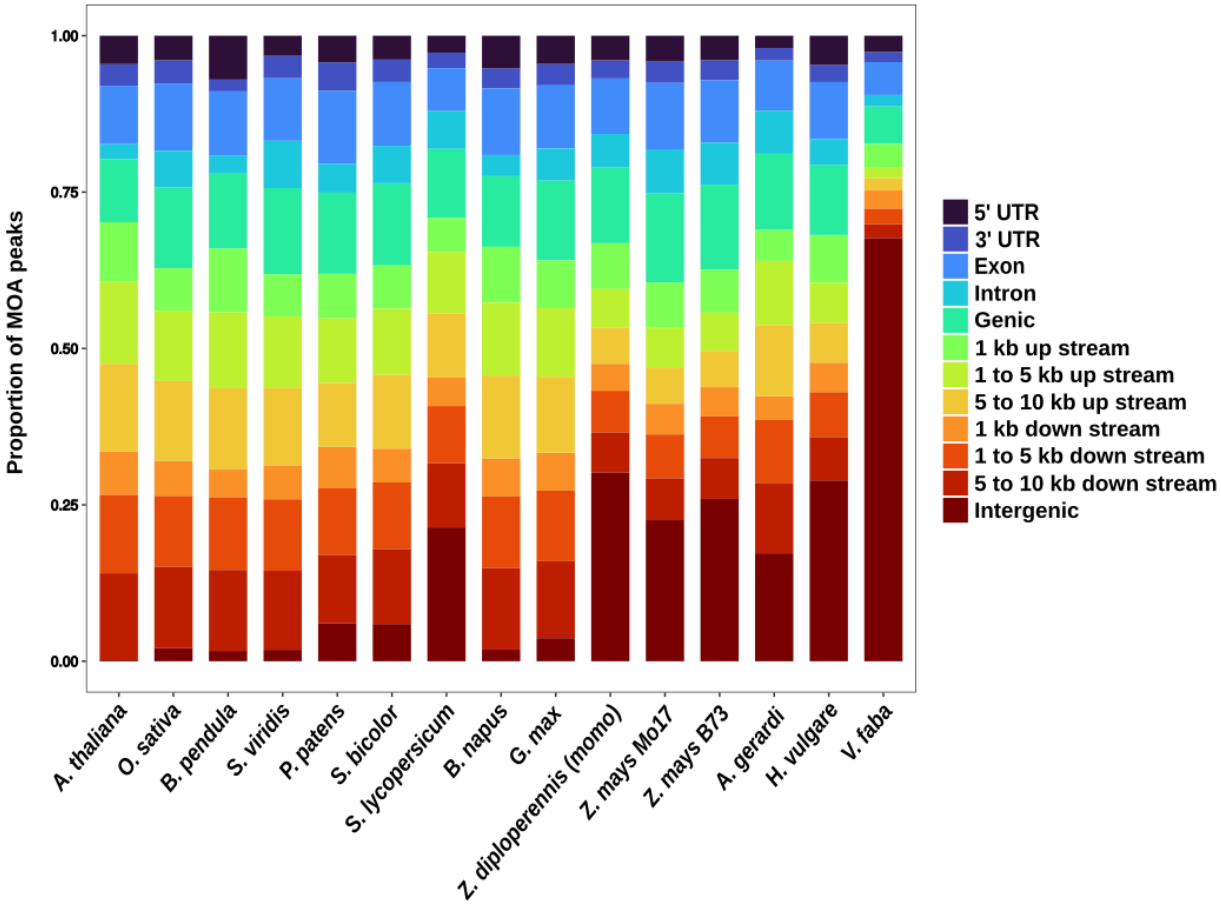


Figure 2.20: Genomic distribution of MOA peaks across 14 plant species.
Stacked bar plots show the proportion of MOA-seq peaks assigned to different genomic features in each species. Peaks were annotated relative to gene models and classified as 5′ UTR, 3′ UTR, exon, intron, genic (within annotated gene bodies), upstream (≤1 kb, 1-5 kb, and 5-10 kb from the transcription start site), downstream (≤1 kb, 1-5 kb, and 5-10 kb from the transcription end site), or intergenic. Each bar represents one species, with colors indicating the relative contribution of each genomic category, normalized to the total number of MOA peaks per species.

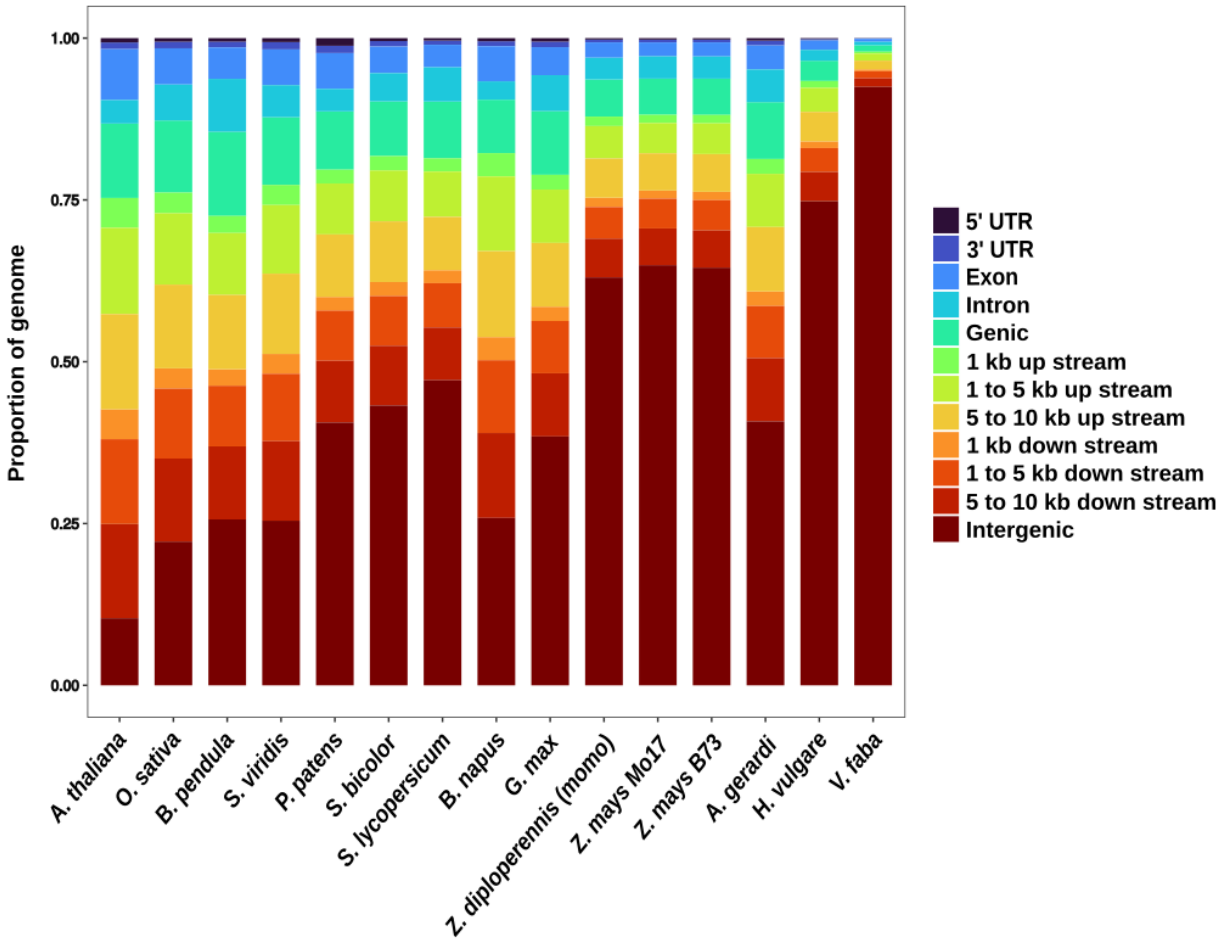


Figure 2.21: Genome-wide distribution of annotated genomic features across 14 plant species.

Stacked bar plots depict the proportion of each reference genome occupied by annotated features, including UTRs, exons, introns, genic regions, upstream and downstream regions relative to transcription start sites, and intergenic regions. Proportions are calculated based on genomic coordinates from species-specific gene annotations and reflect differences in genome organization, size, and repeat content across species. Original script to count the proportion of features originally provided and developed by Amelie Kok and Julia Engelhorn. Scripts were adapted by me to seamlessly work with Helixer.

2.4. Discussion

To date, only a limited number of studies have explicitly examined the evolution of cis-regulatory elements across multiple plant species. ATAC-seq based analyses across four and thirteen species showed that larger genomes tend to harbor fewer accessible chromatin regions near transcription start sites, consistent with an increased contribution of distal cis-regulatory elements in species with expanded intergenic space (Lu et al. 2019; Maher et al. 2018). However, these studies were largely restricted to diploid species and covered a relatively narrow evolutionary range. More recently, Baumgart et al. used DAP-seq to profile transcription factor binding across ten flowering plant species, revealing both deep conservation and extensive regulatory rewiring (Baumgart et al. 2025). While this work provided important insights into angiosperm regulatory evolution, it was limited to flowering plants and relied on an in vitro binding assay that does not capture transcription factor occupancy within native chromatin contexts.

In this chapter, a multi-species MOA-seq dataset spanning 28 species and 29 genomes was established, encompassing wide variation in genome size, ploidy level, and evolutionary history across more than 500 million years. By applying a consistent experimental and analytical framework, this dataset enables direct comparison of cis-regulatory landscapes across angiosperms and grasses, including the Andropogoneae clade. The high spatial resolution of MOA-seq provides a complementary view of regulatory architecture evolution by capturing protein-DNA interactions within native chromatin.

Across species with high-quality genome assemblies, the number of MOA peaks generally increased with genome size, suggesting that genome expansion is often accompanied by an expansion of regulatory sequence space (Figures 2.14 and 2.15). This pattern is consistent with previous observations that larger plant genomes contain increased intergenic and regulatory regions (Lu et al. 2019; Mei et al. 2018). However, the relationship was not linear, and species with similar genome sizes often differed substantially in peak number, indicating that genome size alone does not fully explain regulatory element abundance.

When MOA peak abundance was examined relative to gene number, only a moderate association was observed between peak count and total gene content (Figure 2.18). Although species with larger genomes generally contained more genes (Figure 2.16), increases in regulatory element number often exceeded changes in gene number. As a result, the average number of peaks per gene varied widely across species. This pattern suggests that the expansion of regulatory landscapes reflects broader changes in non-coding sequence space rather than simply gene duplication. As plant genomes increase in size, genes tend to be associated with more regulatory elements (Figure 2.17), but these elements are increasingly positioned farther away in linear genomic distance.

Analysis of MOA-seq peak distance to the nearest TSS revealed a systematic shift toward more distal peak positions with increasing genome size (Figure 2.19). In *Arabidopsis thaliana*, which has a compact genome, MOA-seq peaks were strongly enriched within the promoter-proximal region, consistent with previous genome-wide analyses of TF binding and chromatin accessibility in Arabidopsis (Heyndrickx et al. 2014). In contrast, species with larger and more repeat-rich genomes showed broader peak distributions, with a substantial fraction of MOA-seq peaks located tens to hundreds of kilobases away from annotated genes. This expansion of distal peak space parallels the increase in intergenic distance associated with genome enlargement driven largely by TE accumulation (Mei et al. 2018). Despite the overall shift toward distal MOA-seq peak positions in larger genomes, a promoter-proximal class of MOA-seq peaks (around 1-2 kb from TSS) remained detectable across all species analyzed, including *Vicia faba* and *Hordeum vulgare* (Figure 2.19). This pattern is consistent with genome-wide analyses of chromatin accessibility across plant species with varying genome sizes, which show that promoter-proximal accessible chromatin regions (pACRs) are retained even as genome size increases, although their relative proportion decreases compared to distal accessible chromatin regions (dACRs) (Lu et al. 2019). In that study, pACRs were defined as accessible regions within 2 kb of annotated genes and were enriched for histone modifications associated with active transcription, including H3K4me3, H3K56ac, and H3K36me3, indicating their association with core transcriptional activity. Genome expansion was instead accompanied by a marked increase in dACRs located in enlarged intergenic regions, reflecting additional regulatory capacity rather than displacement of promoter-proximal regulation. In parallel, an

example of a large genome, such as maize, harbors extensive distal regulatory landscapes capable of long-range interactions with genes, highlighting the expansion of regulatory space with genome size (Ricci et al. 2019). Together, these findings support the interpretation that promoter-proximal regulatory occupancy is maintained across plant genomes of different sizes, while genome expansion primarily increases the contribution of distal regulatory regions. The persistence of promoter-proximal MOA-seq peaks in large genomes, therefore, reflects retention of core regulatory architecture rather than a uniform redistribution of TF occupancy away from genes.

The genomic distribution of MOA peaks across functional categories showed related but not strictly proportional shifts with increasing genome size (Figures 2.20 and 2.21). Several polyploid species, including *Brassica napus*, *Glycine max*, and *Andropogon gerardi*, deviated from simple genome size-dependent trends that tend to be observed in diploid species. These lineages have experienced whole-genome duplication followed by varying degrees of diploidization, gene retention, and regulatory reorganization. Together, these observations suggest that ploidy history and post-polyploid genome restructuring influence regulatory architecture alongside genome size.

These findings highlight several avenues for future investigation. Further analyses could examine how regulatory elements are distributed relative to gene structure, addressing whether genes in larger genomes acquire additional regulatory elements or whether genome expansion primarily repositions existing regulatory elements into more distal intergenic regions. Uniform annotation of TEs across all species would enable direct testing of whether specific TE families disproportionately contribute to regulatory element expansion and whether younger TEs are associated with lineage-specific MOA peaks. In addition, synteny-based comparisons of MOA peaks could distinguish deeply conserved regulatory elements from lineage-specific innovations, helping to clarify which regulatory features are maintained across embryophytes and which arise more recently.

This study focused on a single tissue and developmental stage, and regulatory landscapes are likely to vary across cell types and environmental conditions. Differences in genome assembly contiguity and repeat annotation may also influence peak detection, particularly in the Andropogoneae clade. Future work integrating comprehensive genome assemblies, additional tissues, environmental contexts, and

functional assays will further refine our understanding of how cis-regulatory landscapes evolve across plant lineages.

Chapter 3: Pancistrome-wide characterization of cis-element evolution of the Brassinosteroid signaling pathway

3.1. Introduction

In chapter 2, the multispecies MOA-seq dataset was used to investigate the evolution of cis-regulatory landscapes across plant lineages with contrasting genome sizes, ploidy levels, and evolutionary histories. The results showed general scaling relationships between the architecture of genomes and regulatory element abundance, as well as exceptions according to ploidy and the reorganization of genomes. However, while genome-wide patterns provide important insights into regulatory architecture, they do not directly address how individual, conserved biological pathways are regulated across species or how specific cis-regulatory elements are maintained, repositioned, or rewired during evolution. Addressing these questions requires focused analysis of well-characterized pathways whose regulatory logic can be examined in a comparative context.

Brassinosteroid (BR) is a plant steroid hormone that plays a pivotal role in regulating plant development and physiological processes, including cell elongation, vascular tissue development, reproduction, and environmental responses (Planas-Riverola et al. 2019). Therefore, BR biosynthesis requires stringent transcriptional feedback regulation. Core components of the pathway are conserved across land plants, making BR signaling an ideal system in which to investigate how cis-regulatory elements evolve while maintaining pathway function. In particular, BR biosynthesis genes are subject to tight feedback regulation mediated by the TFs BZR1 and BES1, providing a well-defined regulatory context in which to study TF binding and cis-element architecture. The BR signaling cascade begins with perception of bioactive BR, such as brassinolide, by the receptor kinase BRI1 at the plasma membrane, in association with its co-receptor BAK1. Ligand binding triggers a phosphorylation cascade involving BSKs, CDG1, and the phosphatase BSU1, which in turn inhibits the GSK3-like kinase BIN2. In the absence of BRs, BIN2 phosphorylates BZR1 and BES1, promoting their cytoplasmic retention through interaction with 14-3-3 proteins and suppressing transcriptional activity. Upon BR perception, BIN2 is inactivated, allowing PP2A phosphatases to dephosphorylate

BZR1 and BES1 and enable their accumulation in the nucleus, where they regulate target gene expression (Nolan et al. 2020). Once in the nucleus, BZR1 functions both as a transcriptional activator and repressor, depending on its interaction partners. A key component of BR-mediated transcriptional repression is TOPLESS (TPL), a conserved transcriptional corepressor that directly interacts with BZR1 through its EAR repression motif. Upon BR signaling, BZR1 recruits TOPLESS to specific CREs, where TOPLESS in turn engages histone deacetylase complexes to promote local chromatin compaction and transcriptional repression (Oh, Zhu, Ryu, et al. 2014). This mechanism demonstrates that BR-responsive cis-elements function not only as BZR1 binding sites but also as platforms for assembling repressive protein complexes that modulate chromatin state.

A recent study using hybrid allele-specific BZR1 ChIP-seq identified a BR-responsive cis-regulatory region within an intron of the BR biosynthesis gene DWF4 in maize. This intronic region showed allele-specific BZR1 binding and was associated with variation in BR-responsive mRNA abundance (Hartwig et al. 2023). These observations indicate that transcription factor binding relevant to BR regulation can occur within gene bodies, raising the question of whether intronic regulatory features are more broadly employed within the BR pathway across species. Intronic CREs have been recognized as important contributors to gene regulation in plants, supported by multiple well-characterized examples. Intron-mediated enhancement (IME) has been shown to increase transcriptional output when specific introns are present near the transcription start site, even in the absence of classical enhancer motifs. For example, intronic sequences have been demonstrated to enhance expression of flowering-related genes such as AGAMOUS in *Arabidopsis thaliana* (Deyholos and Sieburth 2000), as well as developmental regulators including TRY, where intron 2 is required for proper inflorescence-specific expression through interactions with SPL9 and MADS-box transcription factors (Cui et al. 2022). Intronic enhancers have also been implicated in environmental responsiveness, such as the cold-induced expression of the vacuolar invertase gene in potato, where a discrete intronic enhancer governs temperature-dependent transcriptional activation (X. Zhu et al. 2024). In addition, intronic regulatory elements play key roles in flowering time control, as exemplified by FLOWERING LOCUS C, where intronic regions contribute to transcriptional repression during vernalization and long-term epigenetic memory (Searle et al. 2006). Also, genome

editing approaches have further demonstrated that intronic enhancers fine-tune tissue-specific gene expression, revealing their functional importance independent of promoter sequences (Meng et al. 2021). Together, these studies demonstrate that introns provide a flexible dynamic substrate for regulatory innovation. However, most previous work has focused on single species or individual genes, and the evolutionary conservation of intronic CREs within specific signaling pathways remains largely unexplored.

In this chapter, MOA-seq is applied to characterize CREs associated with brassinosteroid signaling and biosynthesis, with a particular focus on intronic regulatory regions of key pathway genes. By integrating hormone perturbation experiments in *Arabidopsis thaliana* and *Hordeum vulgare* with comparative analyses across multiple species, this work examines whether BR-responsive intronic elements, exemplified by the DWF4 locus, show conserved regulatory features, positional stability, or lineage-specific divergence. These analyses extend the genome-wide framework established in Chapter 2 to a pathway-focused context, enabling direct comparison of cis-regulatory architecture within a conserved hormonal signaling pathway.

3.2. Material and methods

3.2.1. Plant materials

Arabidopsis thaliana (Col-0) seeds were surface-sterilized and germinated on half-strength Murashige and Skoog (½ MS) medium supplemented with 2 µM propiconazole (PPZ). Plants were grown in a standard growth chamber for two weeks. Seedlings were then treated with 100 µM 24-epibrassinolide (eBL) for 4 hours. Mock control samples were treated identically but without eBL. For schematic workflow, see figure 3.1. Four biological replicates were collected for both mock and eBL-treated conditions.

Hordeum vulgare (Morex) seeds were surface-sterilized and stratified at 4°C for 10 days. Seeds were subsequently grown on germination paper in ½ MS medium containing 50 µM PPZ at 15°C for 10 days. First leaves were harvested and treated with 1 µM eBL for 4 hours, then immediately frozen in liquid nitrogen. Three biological replicates were collected for both mock and eBL-treated conditions.

These experiments were performed by Dr. Jiaying Zhu at the Leibniz-Institut für Pflanzengenetik und Kulturpflanzenforschung (IPK), Germany. Samples were shipped to the Max Planck Institute for Plant Breeding Research to perform MOA-seq and data analysis.

3.2.2. Data materials

Previously published ChIP-seq datasets for BZR1 from *Zea mays* (B73) (Hartwig et al. 2023), as well as BZR1, PIF4, and ARF6 ChIP-seq datasets from *Arabidopsis thaliana* (Col-0) (Oh, Zhu, Bai, et al. 2014; Oh et al. 2012), were used in this thesis. In addition, a published MOA-seq dataset generated from earshoot tissue of Zea mays (B73) was included (Savadel et al. 2021).

Following optimization of the MOA-seq protocol in maize leaf tissue (B73) within the group, the M4 MNase digestion condition yielded the most suitable MOA-seq signal. This dataset was therefore included together with the B73 well-watered MOA-seq data for comparative analyses presented in Table 3.3. The maize leaf M4 MOA-seq data were generated by Sergius Weizel and Dr. Sohini Mukherjee.

3.2.3. MOA-seq

MOA-seq was performed in *Arabidopsis thaliana* using the dicot protocol and in *Hordeum vulgare* using the monocot protocol as described in method section 2.2.3 of chapter 2.

3.2.4. Preprocessing and alignment of MOA-seq data

MOA-seq data analysis in this chapter followed the same preprocessing and alignment workflow described in Chapter 2. This included quality assessment with FASTQC, adapter trimming, read merging, and alignment to the appropriate reference genomes (TAIR10 for *Arabidopsis thaliana* and Morex v3 for *Hordeum vulgare*) using STAR (v2.7.11a) (Dobin et al. 2013). For both species, the `--outFilterMultimapNmax` parameter was set to 1 to retain uniquely mapped reads, reflecting their inbred diploid genomes. Detailed descriptions of these steps are provided in Methods Sections 2.2.4 and 2.2.5.

3.2.5. MOA peak calling and identification of BR-responsive regulatory regions

MOA-seq peaks were initially identified for each sample independently using MACS3 (v3.0.0a7) (Y. Zhang et al. 2008), following the same parameters and workflow described in Chapter 2. Peak calling was performed using the effective genome size for each species, the average fragment length derived from BAM file statistics, and fixed extension parameters optimized for MOA-seq data.

To identify brassinosteroid-responsive regulatory regions, differential peak calling was performed using MACS3 by comparing PPZ-treated samples with PPZ plus epibrassinolide (PPZ+eBL) samples. In this analysis, one condition was provided as the treatment (parameter `-t`) and the other as the control (parameter `-c`), allowing identification of MOA-seq peaks showing higher occupancy in one condition relative to the other. Because MACS3 detects enrichment only in the treatment relative to the control, reciprocal comparisons were performed, with PPZ-treated samples contrasted against PPZ+eBL-treated samples and vice versa. This strategy enabled detection of peaks preferentially enriched under either hormonal condition.

For *Hordeum vulgare*, differential peak calling was performed using a q-value cutoff of 0.05. For *Arabidopsis thaliana*, parameters were adjusted to account for the smaller genome size and higher read density. Local background estimation was restricted using the `--slocal` parameter, and peaks were initially called at $q \leq 0.05$ before applying an additional filtering step to retain peaks corresponding to a q-value cutoff of 0.00001. This more stringent threshold was used to reduce background signal in differential peak sets. All command-line parameters and filtering steps used for differential peak calling are provided in Appendix 5.1.11. In addition to differential peak sets, MOA-seq peaks identified independently under PPZ and PPZ+eBL conditions were retained for downstream analyses without cross-condition comparison. These peak sets were used for subsequent genomic feature overlap analyses and quantification of peak distributions across genomic regions.

3.2.6. Assignment of MOA peaks to genomic features

To examine how MOA peaks were distributed across genomic features under different conditions, peak overlap analyses were performed for *Arabidopsis thaliana* and *Hordeum vulgare* using genome annotations and MOA peak sets derived from MACS3.

Genome annotation files in GFF3 format were obtained for *Arabidopsis thaliana* (TAIR10) and *Hordeum vulgare* (Morex v3). From these annotations, genomic feature intervals were extracted and converted into BED format using custom awk-based scripts and bedtools (v2.30.0). The following feature classes were defined: 5′ UTRs, 3′ UTRs, exons, introns, and gene bodies. In addition, promoter-proximal and downstream regions were defined by extending gene coordinates to generate 1 kb upstream, 1–5 kb upstream, 1 kb downstream, and 1–5 kb downstream intervals.

Intronic regions were defined by subtracting exonic intervals from gene body intervals using the bedtools subtract function. To avoid double counting of overlapping annotations, genomic regions were assigned to a single feature class according to a fixed priority scheme: 5′ UTR > 3′ UTR > exon > intron > 1 kb upstream > 1 kb downstream > 1–5 kb upstream > 1–5 kb downstream > intergenic. Intergenic regions were defined as genomic intervals not overlapping genes or their ±5 kb flanking regions.

MOA peak sets were analyzed separately for each condition, including PPZ-treated samples, PPZ plus eBL-treated samples, and differentially bound peak sets identified by condition-specific MACS3 comparisons. For each peak set, base pair overlap with each genomic feature category was calculated using bedtools intersect, followed by summation of overlapping base pairs using awk. Feature overlap values were then normalized to the effective genome size to calculate the percentage of peak coverage within each genomic category.

For comparison with background genomic composition, the total base pair coverage of each feature class was also calculated using the same annotation-derived feature files. This allowed direct comparison between the genomic distribution of MOA peaks and the underlying genomic feature distribution.

Feature overlap statistics were summarized in tabular format and visualized using bar plots generated in R (v4.1.2) with ggplot2. Bar plots display the percentage of MOA peak coverage assigned to each genomic feature category across conditions, with consistent feature ordering to facilitate cross-condition and cross-species comparison.

All scripts used for feature extraction and overlap calculation are provided in Appendix 5.1.9.

3.2.7. BigWig track generation for genome browser visualization

For genome browser visualization, BigWig coverage tracks were generated from coordinate-sorted BAM files using `bamCoverage` from deepTools (v3.5.5). For each

BAM file, the presence of an index file was checked; if required, the BAM file was indexed using `samtools index`. Coverage was calculated at 1 bp resolution (`--binSize 1`) and written in BigWig format (`-of bigwig`). Signal was normalized using counts per million mapped reads (CPM) (`--normalizeUsing CPM`) with `--exactScaling`. Only reads meeting a minimum mapping quality threshold were included (`--minMappingQuality 255`). Coverage calculation was restricted to fragments up to 80 bp (`--maxFragmentLength 80`). No smoothing was applied (`--smoothLength 0`). The number of processing threads was specified with `-p`. BigWig files were loaded into IGV for visualization. A full script is provided in Appendix 5.1.10.

3.2.8. MOA-seq and ChIP-seq signal visualization

MOA-seq and ChIP-seq signal tracks were visualized using Integrative Genomics Viewer (IGV) v2.17.0. Normalized coverage tracks in bigWig format (CPM-normalized) were loaded together with gene annotations to inspect genomic regions of interest. Figures were exported from IGV in SVG format. Minor post-processing was performed in Inkscape v1.4.2 to add labels and a uniform white background, without altering track heights or signal representation.

3.2.9. Identification of Putative Orthologs of BR Biosynthesis Genes.

Identification of putative orthologs of BR biosynthesis genes was performed using sequence similarity-based searches. For example, for monocot species, the maize DWF4 protein sequence was used as the query, whereas for dicot species, the *Arabidopsis thaliana* DWF4 protein sequence was used. Protein BLAST searches were conducted against species-specific proteomes available in Phytozome. Candidate homologs were identified based on sequence similarity and alignment quality. In species where multiple homologous genes were detected (*Glycine max*), publicly available RNA-seq datasets (https://www.soybase.org/) were examined to assess relative expression levels. The homolog showing the highest expression in leaf tissues was selected as the representative gene for subsequent analyses. This approach does not assume strict one-to-one orthology in all cases but provides a practical framework for comparative analysis across species with differing gene family sizes.

3.3. Results

3.3.1 Comparative analysis of intronic transcription factor binding in brassinosteroid biosynthesis genes.

Brassinosteroid biosynthesis is a conserved multistep enzymatic process. Of these enzymes, DWF4, a well-studied example, encodes a cytochrome P450 monooxygenase and catalyzes an early C-6 oxidation that represents a rate-limiting step for brassinosteroid biosynthesis. As indicated in Figure 3.1, DWF4 acts upstream in both the early and late C-6 oxidation branches, influencing flux towards castasterone and brassinolide. To initiate a comparative analysis of intronic TF binding within the BR pathway, the DWF4 locus was examined as a reference case.

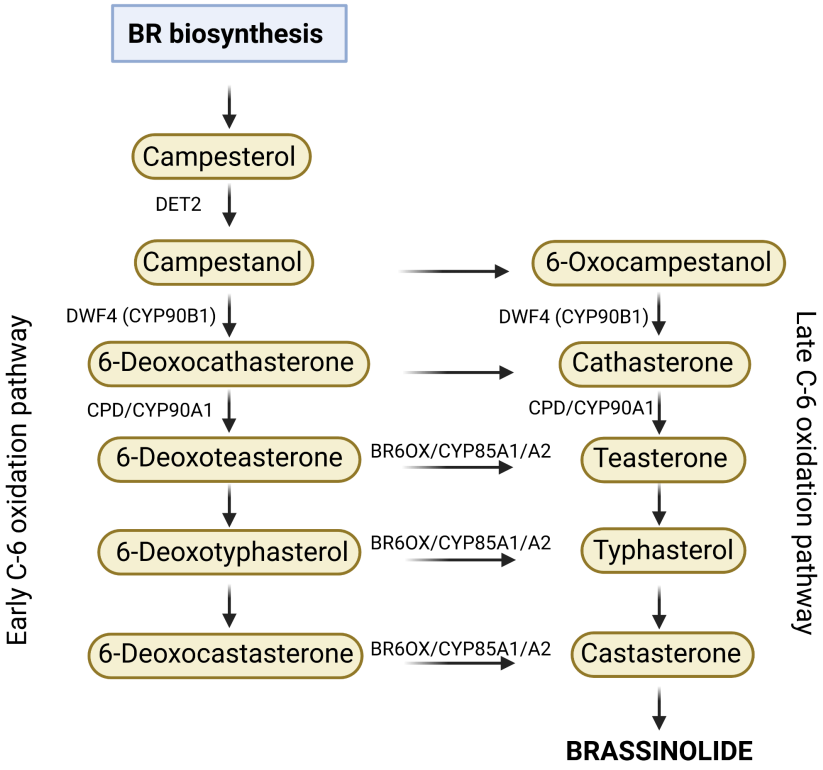


Figure 3.1: The brassinosteroid biosynthesis pathway in plants.

DET2 converts campesterol to campestanol. Downstream intermediates bifurcate into the early C-6 oxidation pathway, in which C-6 oxidation occurs at the 6-deoxocathasterone level, and the late C-6 oxidation pathway, in which C-6 oxidation occurs at the campestanol stage via 6-oxocampestanol. The names of enzymes catalyzing individual steps are shown next to the corresponding reactions and include DWF4 (CYP90B1), CPD (CYP90A1), and BR6OX (CYP85A1/A2). Both pathways again converge at castasterone, which is then converted to brassinolide. Arrows depict the

proposed enzymatic conversions between intermediates. Figure created by Duong Doan using BioRender.

Prior hybrid allele-specific ChIP-seq analyses identified intron-localized BZR1 occupancy at this gene in maize, with differences between the B73 and Mo17 haplotypes (Hartwig et al. 2023). This defined intronic binding feature provides a suitable starting point to assess whether MOA-seq detects comparable intronic occupancy and to extend the analysis to additional BR-associated genes and species. MOA-seq profiles at the maize DWF4 locus revealed a discrete footprint within the same intronic region previously reported to be bound by BZR1 in ChIP-seq analyses (Figure 3.2a). The intronic MOA-seq signal was detected in leaf tissue and localized to a defined genomic position that overlaps the region of allele-specific BZR1 binding described in maize hybrids (Hartwig et al. 2023). These observations indicate that MOA-seq captures intron-localized TF occupancy at the DWF4 locus in maize. To assess whether intron-localized transcription factor occupancy at DWF4 is restricted to a single tissue context, MOA-seq profiles from maize earshoot tissue were examined. An intronic MOA-seq footprint was also detected at the DWF4 locus in earshoot samples, at a position comparable to that observed in leaf tissue (Figure 3.4a). These data show that intron-localized MOA-seq signal at the DWF4 locus can be detected in both leaf and earshoot tissue in maize.

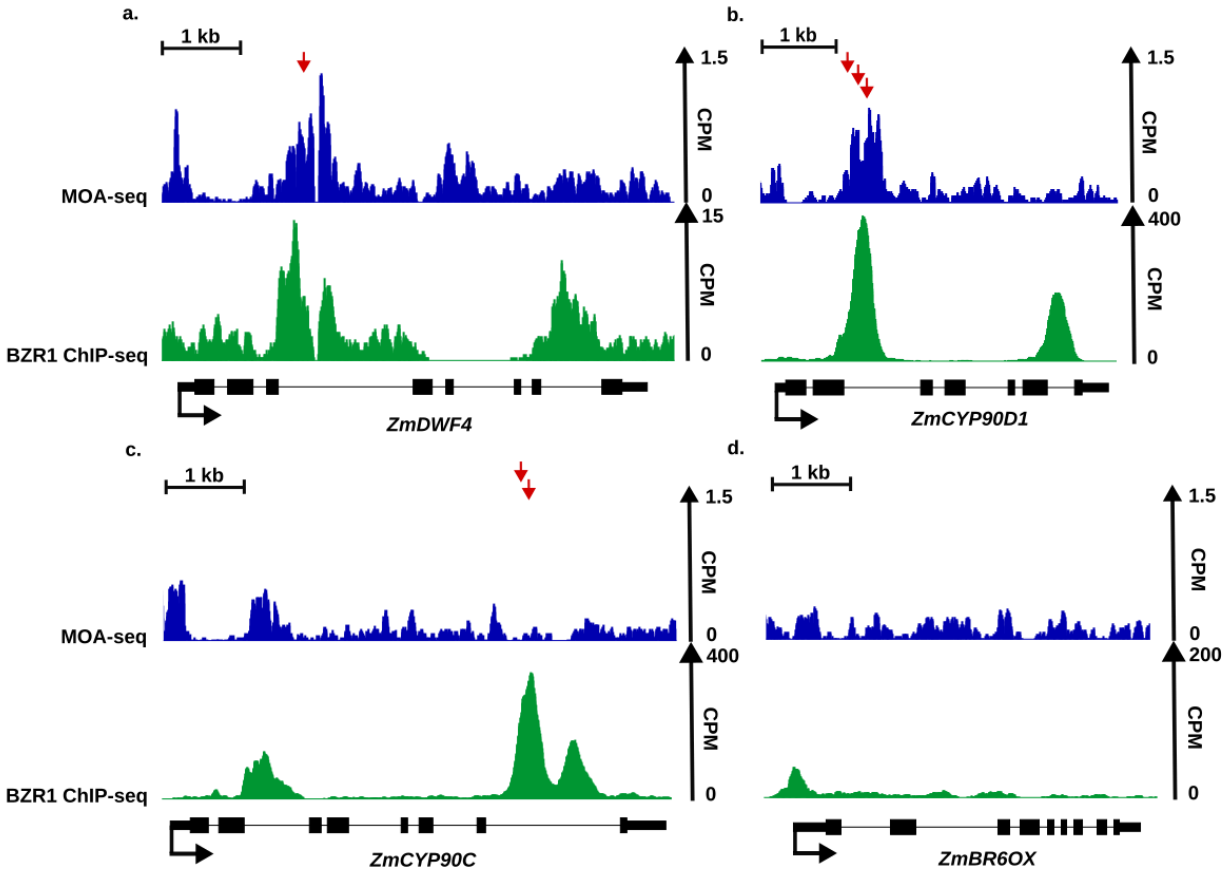


Figure 3.2: Comparison of MOA-seq and BZR1 ChIP-seq profiles at brassinosteroid biosynthesis genes in maize.

Genome browser views showing MOA-seq signal (blue) and published BZR1 ChIP-seq signal (green; Hartwig et al. 2023) at representative BR biosynthesis genes in *Zea mays*, including (a) *ZmDWF4*, (b) *ZmCYP90D1*, (c) *ZmCYP90C*, and (d) *ZmBR6OX*. Gene models are shown below each track with arrows indicating transcriptional orientation. Red arrows denote allele-specific ZmBZR1 bindings (ASBs). MOA-seq and ChIP-seq signals are displayed as counts per million (CPM). Scale bars indicate 1 kb.

To examine whether intron-localized TF binding at DWF4 is observed beyond maize, MOA-seq profiles at orthologous DWF4 loci were compared across multiple plant species. Intronic MOA-seq signal was detected at DWF4 orthologs in several species, including dicotyledonous and monocotyledonous lineages, as well as members of the Andropogoneae. The position and extent of the intronic signal varied among species. Representative examples are shown in Figure 3.3.

Given that intron-localized MOA-seq signal is observed at DWF4 across species, it was

next examined whether similar intronic binding patterns are also present at other BR biosynthesis genes. To address this, MOA-seq profiles were analyzed for additional BR-associated genes in maize. Intronic MOA-seq signal was detected at *ZmCYP90D1*, for which intron-localized BZR1 binding with allele-specific binding sites has been reported in the second intron (Figure 3.2b; Figure 3.4b). At *ZmCYP90C*, MOA-seq signal was observed in leaf tissue within the second intron, coinciding with BZR1 ChIP-seq signal, while in earshoot tissue, MOA-seq signal was detected in the last intron, overlapping a region of reported BZR1 binding with allele-specific sites (Figure 3.2c; Figure 3.4c). In contrast, *ZmBR6OX* did not show intronic BZR1 ChIP-seq or MOA-seq signal in B73 leaf tissue under well-watered conditions, but intronic MOA-seq signal was detected in earshoot tissue and leaf tissue under M4 digestion condition in the first and second introns, without corresponding allele-specific BZR1 binding sites (Figure 3.2d; Figure 3.4d; Table 3.1). Beyond those examples, intronic MOA-seq signal was detected at several additional BR biosynthesis genes in maize (Table 3.1). Intronic MOA-seq peaks were observed at *ZmDWF7*, where intronic signal was detected in at least one tissue or condition. For *ZmNa2/DWF1*, intronic MOA-seq signal was detected in earshoot tissue and in leaf tissue under the M4 MNase digestion condition, whereas no intronic signal was observed in leaf tissue under well-watered conditions. In contrast, *ZmCPD/DWF3* and *ZmDWF5* did not show detectable intronic MOA-seq signal in the maize datasets analyzed. Although BZR1 binding has been reported at other regions of these loci (Table 3.1), no intronic signal was detected under the present experimental conditions.

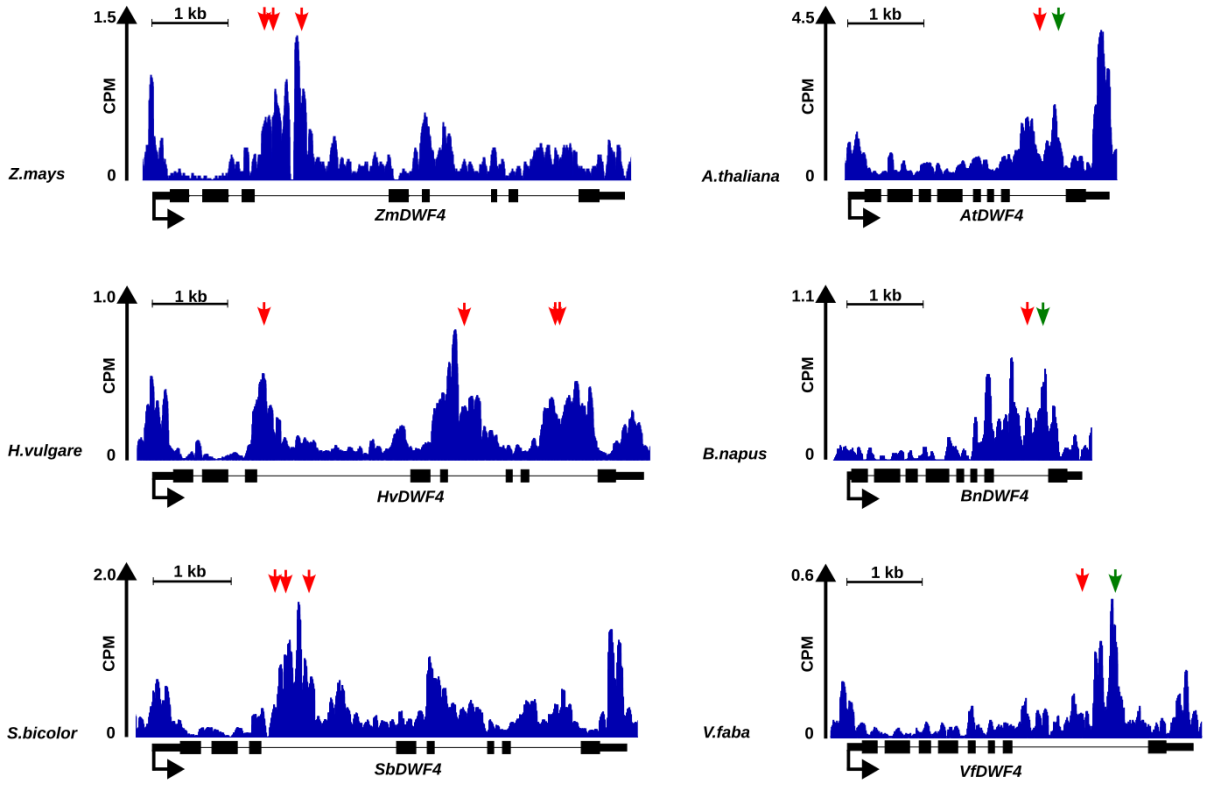


Figure 3.3: Conserved intronic regulatory elements at the *DWF4* locus across plant species.

IGV pictures show MOA-seq signal across the *DWF4* gene in *Zea mays*, *Arabidopsis thaliana*, *Hordeum vulgare*, *Brassica napus*, *Sorghum bicolor*, and *Vicia faba*. Blue tracks represent MOA-seq coverage (CPM). Gene models are shown below each track, with boxes indicating exons and arrows indicating transcriptional direction. Red arrows mark predicted BR-responsive elements (BRREs), and green arrows mark G-box motifs within or near intronic MOA peaks. Scale bars indicate 1 kb.

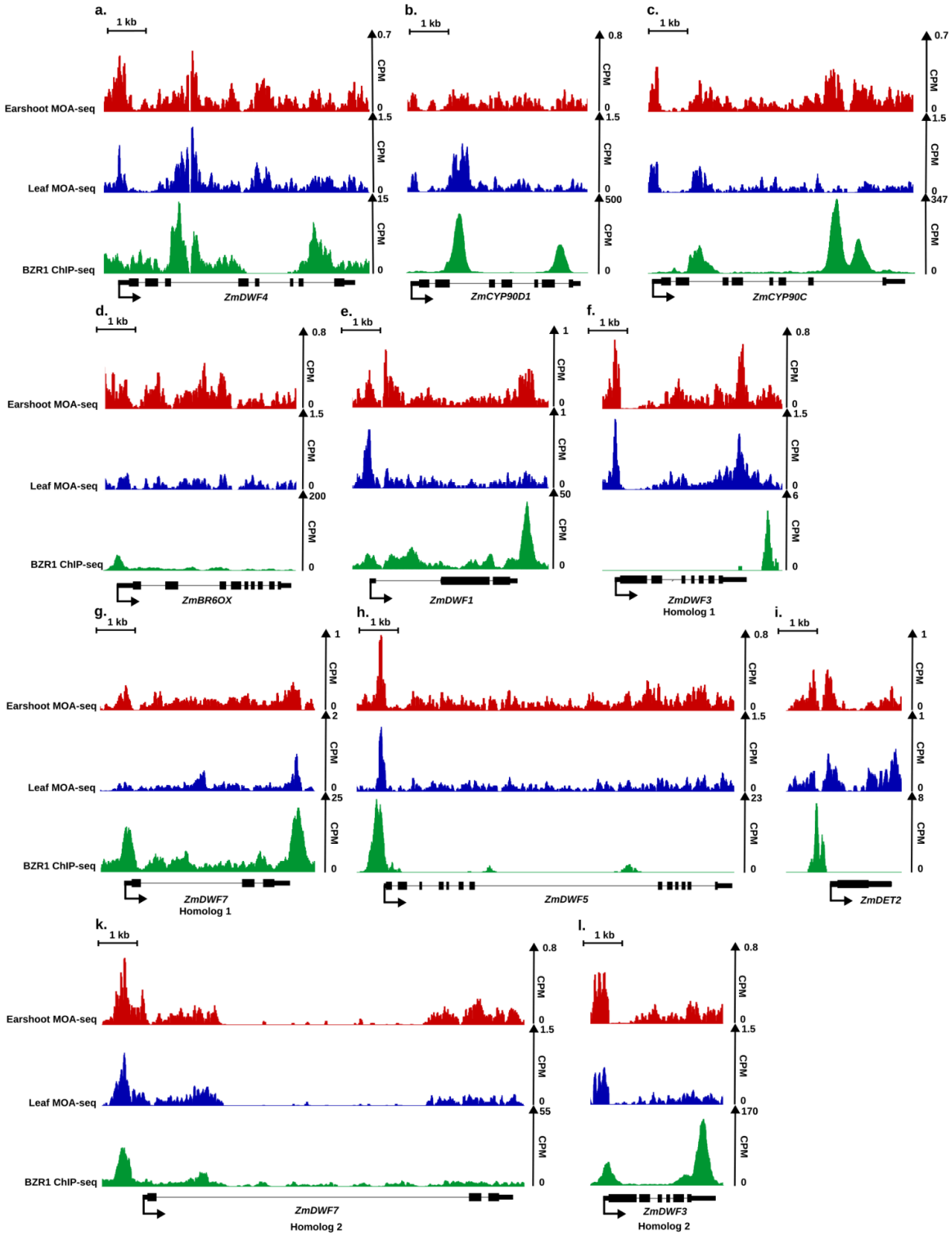

a.
b.
c.
d.
e.
f.
g.
h.
i.
k.
l.
1 kb
Earshoot MOA-seq
Leaf MOA-seq
BZR1 ChIP-seq
CPM
ZmDWF4
ZmCYP90D1
ZmCYP90C
ZmBR6OX
ZmDWF1
ZmDWF3
Homolog 1
ZmDWF7
Homolog 1
ZmDWF5
ZmDET2
ZmDWF7
Homolog 2
ZmDWF3
Homolog 2

Figure 3.4: MOA-seq profiles at brassinosteroid biosynthesis genes in maize across tissues.

Genome browser views showing MOA-seq signal in maize earshoot (red; Savadel et al. 2021) and leaf (blue), together with BZR1 ChIP-seq signal (green; Hartwig et al. 2023), across multiple genes involved in the BR biosynthesis pathway, including (a) *ZmDWF4*, (b) *ZmCYP90D1*, (c) *ZmCYP90C*, (d) *ZmBR6ox*, (e) *ZmDWF1*, (f) *ZmDWF3* homolog 1, (g) *ZmDWF7* homolog 1, (h) *ZmDWF5*, (i) *ZmDET2,* (k) *ZmDWF7* homolog 2, and (l) *ZmDWF3* homolog 2 . Gene models are shown below each panel with arrows indicating transcriptional orientation. Signal is displayed as counts per million (CPM), with independent y-axis scaling for each track. Scale bars indicate 1 kb. These profiles illustrate tissue-specific MOA accessibility patterns and their correspondence with BZR1 binding at BR pathway genes.

To place the maize observations into a broader TF binding context, MOA-seq profiles at BR biosynthesis genes were next examined in *Arabidopsis thaliana* together with published ChIP-seq datasets for BZR1, PIF4, and ARF6 (Oh, Zhu, Bai, et al. 2014). In addition to BZR1, the datasets for PIF4 and ARF6 were included to provide a broader TF binding context at BR biosynthesis genes in *Arabidopsis thaliana*. A previous study has shown that BZR1 physically and functionally interacts with PIF4 and ARF6 and that these transcription factors bind overlapping sets of genomic loci in elongating tissues. PIF4 integrates light and temperature signals, while ARF6 mediates auxin-responsive transcription, and both have been reported to act together with BZR1 at shared target genes involved in growth-related processes (Oh, Zhu, Bai, et al. 2014). The inclusion of these datasets therefore allows MOA-seq signals at BR-associated genes to be interpreted in the context of a multi-factor transcriptional network that integrates hormonal and environmental inputs, without implying direct regulatory roles for individual binding events. At *AtDWF4*, intronic MOA-seq signal was detected at a position overlapping reported ChIP-seq signal for BZR1, ARF6, and PIF4 within the gene body (Figure 3.5a; Table 3.2). For *AtCPD*, intronic MOA-seq signal was observed at a site coinciding with BZR1 ChIP-seq occupancy (Figure 3.5b; Table 3.2). At *AtDWF1*, no intronic MOA-seq signal was observed in the datasets analyzed, and ChIP-seq signal was not detected within intronic regions, whereas BZR1 ChIP-seq signal was detected primarily at the 5′ UTR (Figure 3.5c; Table 3.2). In contrast, *AtBR6OX* showed intronic

MOA-seq signal that overlaps ChIP-seq signal for BZR1, ARF6, and PIF4 within the intron region (Figure 3.5d; Table 3.2). A gene-by-gene summary of intronic MOA-seq signal and corresponding BZR1 ChIP-seq occupancy in Arabidopsis is provided in Table 3.2.

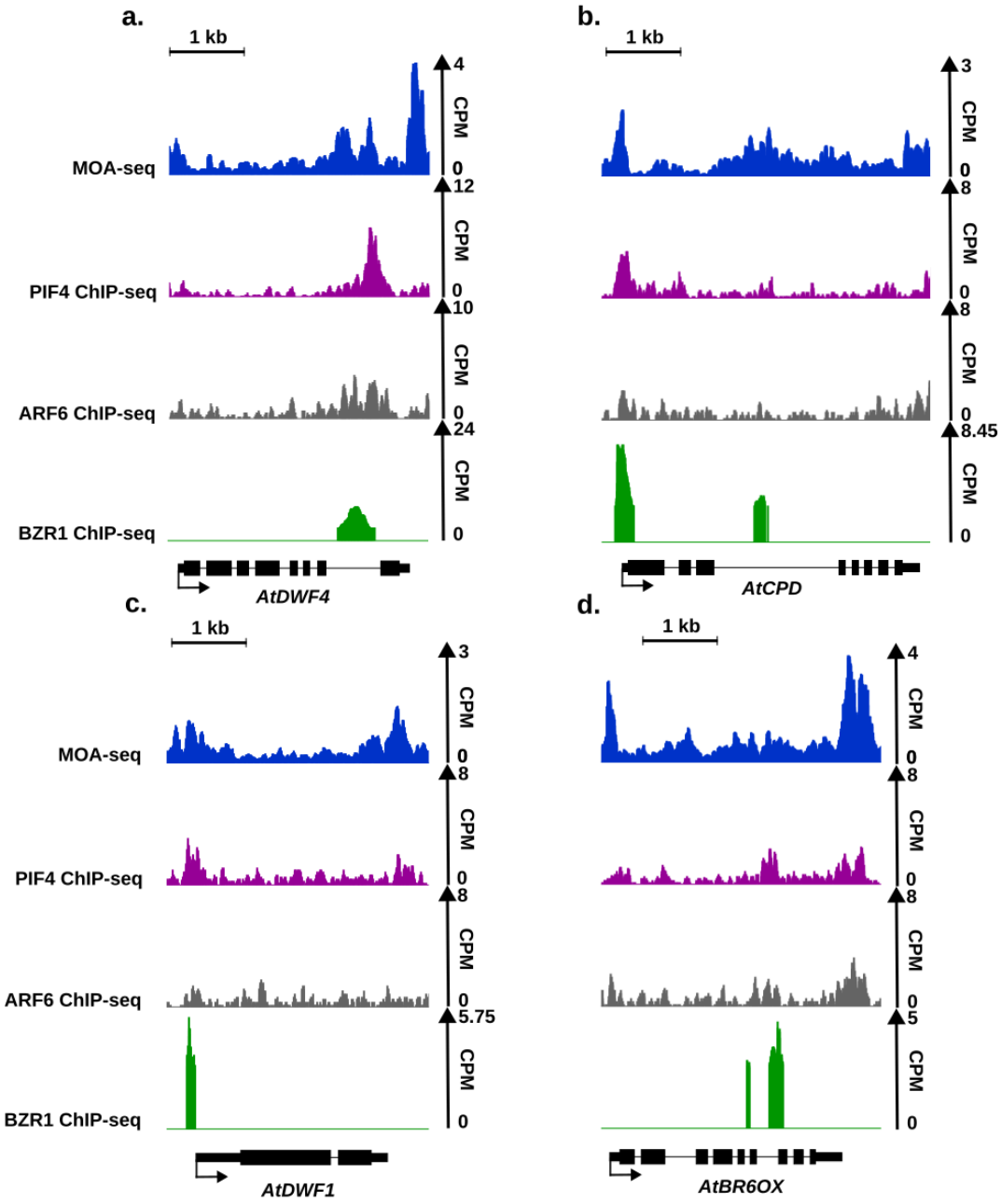


Figure 3.5: MOA-seq and transcription factor ChIP-seq profiles at brassinosteroid biosynthesis genes in *Arabidopsis thaliana*.

Genome browser views showing MOA-seq signal together with ChIP-seq profiles for PIF4 , ARF6, and BZR1 (Oh, Zhu, Bai, et al. 2014; Oh et al. 2012) at selected

brassinosteroid biosynthesis genes in *Arabidopsis thaliana*. The background level was removed in the BZR1 ChIP-seq track. Panels show (a) *AtDWF4*, (b) *AtCPD*, (c) *AtDWF1*, and (d) *AtBR6OX*. MOA-seq signal is shown in blue, PIF4 ChIP-seq in purple, ARF6 ChIP-seq in grey, and BZR1 ChIP-seq in green. Signal tracks are displayed as counts per million (CPM) and scaled independently for each track, with maximum values indicated on the right. Gene models are shown below each panel, with black boxes representing exons and arrows indicating transcriptional orientation. A scale bar of 1 kb is shown for each locus.

To extend the analysis to an additional monocot species, MOA-seq profiles at BR biosynthesis genes were examined in *Hordeum vulgare*. Intronic MOA-seq signal was detected at several BR-associated genes, including *HvDWF4*, where prominent MOA peaks were observed in the third, fifth, and seventh introns (Figure 3.6a; Table 3.3). Intronic MOA-seq signal was also detected at *HvDWF3* and *HvCYP90C*. Although the presence and position of intronic peaks varied among loci, both genes showed clear MOA peaks in the second intron (Figure 3.6b,c; Table 3.3). For *HvBR6OX*, a low-amplitude intronic MOA-seq signal was detected under the conditions analyzed (Figure 3.6d; Table 3.3). Together, these data show that intronic MOA-seq signal in barley is observed at several BR biosynthesis genes, with gene-specific differences in the number and position of intronic peaks.

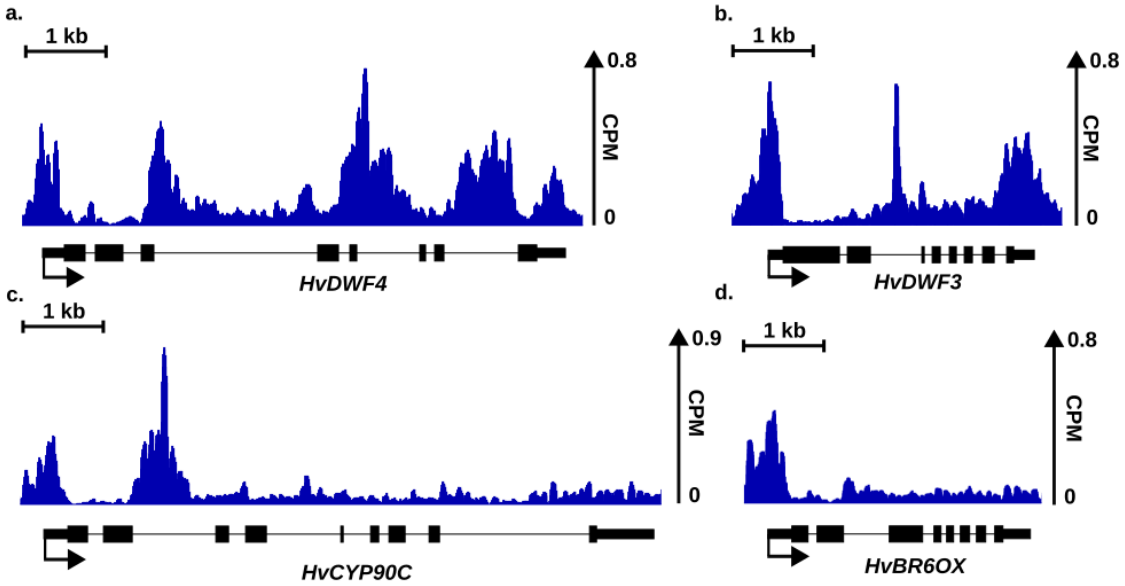


Figure 3.6: MOA-seq profiles at BR biosynthesis genes in *Hordeum vulgare*.
Genome browser tracks show MOA-seq signal (CPM) across BR biosynthesis genes in barley, including (a.) *HvDWF4*, (b.) *HvDWF3*, (c.) *HvCYP90*, and (d.) *HvBR6OX*. Gene models are shown under each track with exons indicated in black boxes, and arrows denote transcriptional orientation. Scale bars indicate 1 kb. The graph illustrates the MOA-seq signals associated with key BR pathway genes in barley.

Table 3.1: Information on the binding of ZmBZR1, allele-specific ZmBZR1 bindings (ASBs), and MOA-seq signal at the intron region of brassinosteroid-associated genes in *Zea mays*.
(This table was generated by Dr. Julia Engelhorn.)

Gene name	B73 identifier	BZR1 target	BZR1 ChIP-seq signal in intron	MOA-seq signal in intron	ASB in intron
DWF4	Zm00001eb009730	yes (only intron B73)	yes	yes, leaf (both) and ear shoot	yes
ROT/CYP90C	Zm00001eb120890	yes (only intron with exon overlap)	yes	yes, leaf (both) and ear shoot	yes
CYP90D1	Zm00001eb284090	yes (only intron with exon overlap)	yes	yes, leaf (both); not in ear shoot	yes
BRD1/BR6OX	Zm00001eb050010	yes (promoter + 5'UTR)	no	yes, ear and M4, not in B73ww	no
CYP50/DWF3 /CPD	Zm00001eb093140	yes (both UTRs)	no	yes, leaf (both); no ear	no
CPD/DWF3	Zm00001eb196530 (gene model 2)	yes (within 1 kb downstream)	no	yes, leaf (both) and ear shoot	no
DWF7 (2 genes)	Zm00001eb125130	yes (in 3' UTR)	yes (small part of 3'UTR peak also in intron in one gene model)	yes, both leaf samples, for ear shoot only at the very end of the last intron	no

				close to the non-coding exon	
DWF7 (2 genes)	Zm00001eb335380	yes (5' UTR and downstream)	no	yes, leaf (both), not in ear shoot	no
DET2/Na1	Zm00001eb147800 (as non-coding in V5, without introns in NCBI)				
Na2/DWF1	Zm00001eb228910 (intron exists in two out of 4 gene models)	yes (within 1 kb downstream)	no	yes, ear shoot and M4, not in B73ww	no
DWF5	Zm00001eb240810	yes (promoter + 5'UTR)	no	yes, leaf (both); not in ear shoot	no

Table 3.2: Information on the binding of AtBZR1 and MOA-seq signal at the intron region of brassinosteroid-associated genes in *Arabidopsis thaliana*.

(This table was generated by Dr. Julia Engelhorn.)

Gene name	Identifier	BZR1 target	BZR1 ChIP-seq signal in intron	MOA-seq signal in intron
DWF4	AT3G50660	yes (promoter and intron)	yes	yes
ROT/CYP90C	AT4G36380	yes (intron and downstream but downstream there is another gene)	yes	yes
CYP90D1	AT3G13730	yes (3' UTR)	no	no
BRD1/BR6OX	AT3G30180	yes (intron and downstream but downstream there is another gene)	yes	yes
CPD/CYP90/DWF3	AT5G05690	yes (promoter/TSS + intron)	yes	yes
DWF7	AT3G02590	yes (downstream)	no	no

DET2/Na1	AT2G38050	yes (exon)	no	yes (overlapping from an exon)
Na2/ DWF1	AT3G19820	yes (promoter/5'UTR)	no	no
DWF5	AT1G50430	yes (downstream within 1 kb)	no	yes

Table 3.3: Information on the MOA-seq signal at the intron region of brassinosteroid-associated genes in *Hordeum vulgare*.

(This table was generated by Dr. Julia Engelhorn.)

Gene name	Identifier	MOA-seq signal in intron
DWF4	HORVU.MOREX.r3.4HG0394170	yes
ROT/CYP90C/CYP90D1	HORVU.MOREX.r3.3HG0235980	yes
BRD1/BROX6	HORVU.MOREX.r3.2HG0097170	very little but significant
CPD/CYP90/DWF3	HORVU.MOREX.r3.5HG0460830	yes
DWF7	HORVU.MOREX.r3.3HG0225930	yes
DET2/ Na1	HORVU.MOREX.r3.3HG0301780	no intron
Na2/ DWF1	HORVU.MOREX.r3.7HG0749940	yes
DWF5	HORVU.MOREX.r3.3HG0320030	one small but significant peak in intron of UTR in one gene model

3.3.2 Generation of MOA-seq profiles under brassinosteroid treatment in Arabidopsis and barley.

To investigate how BR signaling influences the regulatory landscape, MOA-seq was performed in *Arabidopsis thaliana* and *Hordeum vulgare* under conditions of BR depletion and acute hormone treatment (Figure 3.7). Seeds were germinated and grown in the presence of the BR biosynthesis inhibitor propiconazole (PPZ). Subsequently, Arabidopsis seedlings and the first barley leaf were collected and treated with

epibrassinolide (eBL) for 4 hours. Mock-treated seedlings were included as controls for phenotypic comparison. This experimental design enabled the generation of genome-wide MOA footprints responsive to BR signaling in both a model dicot and a monocot crop species.

Following MNase digestion of isolated nuclei, partial digestion profiles were evaluated by agarose gel electrophoresis to assess digestion efficiency and consistency across samples (Figures 3.8 and 3.9). Samples displaying comparable digestion patterns were selected for MOA-seq library preparation and sequencing. After quality assessment with FASTQC, sequencing output and read mapping statistics are summarized in Table 3.4.

MOA-seq libraries were sequenced to high depth in both *Arabidopsis thaliana* and *Hordeum vulgare* (Table 3.4). Arabidopsis samples generated approximately 700-756 million raw reads per condition, with high mapping efficiency and 82.1% (PPZ) and 81.8% (PPZ+eBL) of reads mapping uniquely to the reference genome. In contrast, barley samples yielded approximately 490-500 million raw reads per condition, with lower unique mapping rates of 50.4% (PPZ) and 53.3% (PPZ+eBL). This difference is consistent with the larger genome size, higher repetitive content of the *Hordeum vulgare* genome, and the sequencing depth. Despite these differences, the mapping quality was sufficient to support robust identification of MOA footprints and downstream analyses within each species.

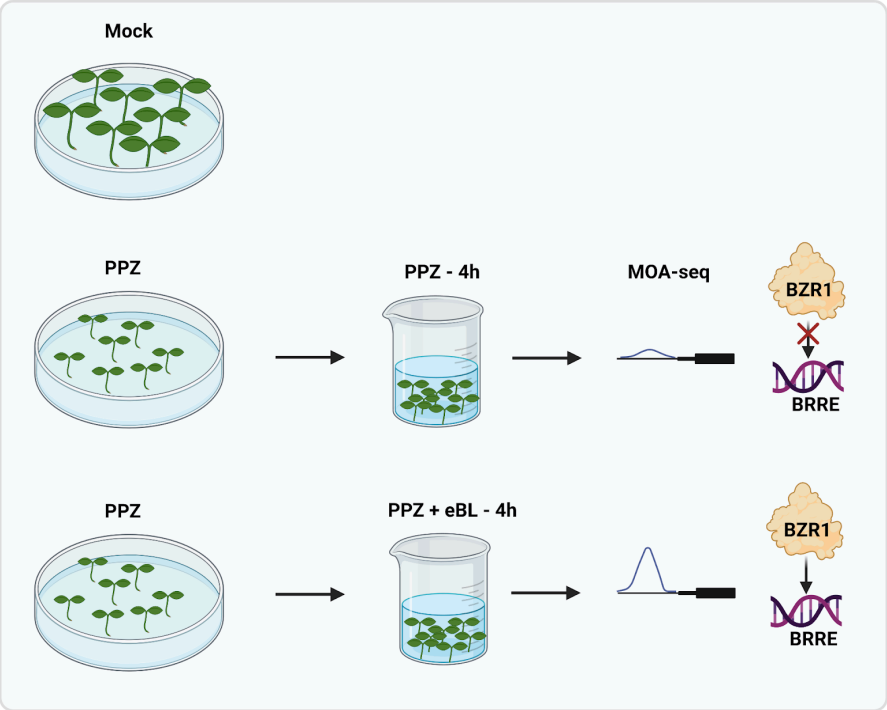


Figure 3.7: Experimental workflow for hormone perturbation and MOA-seq analysis of BR-responsive cis-regulatory elements.

Arabidopsis thaliana seedlings were grown under mock conditions or in the presence of the brassinosteroid biosynthesis inhibitor propiconazole (PPZ). PPZ-treated seedlings were treated with only PPZ or with exogenous epibrassinolide (eBL) and PPZ for 4 hours prior to harvest. MOA-seq was then performed to profile protein-DNA interaction footprints under each condition. Schematic representations illustrate reduced BZR1 association with BR-responsive elements (BRREs) under only PPZ treatment and increased accessibility and binding following eBL+PPZ treatment. The figure was created using BioRender.

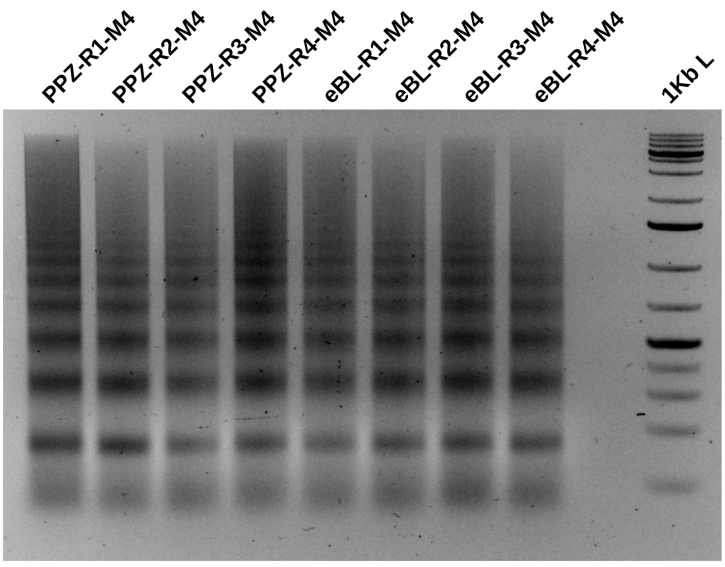


Figure 3.8: MNase partial digestion profiles of MOA-seq in *Arabidopsis thaliana.*
Agarose gel showing MNase digestion from Col0 seedlings treated with PPZ or PPZ followed by eBL for 4 hours, with four biological replicates per condition. A 1 kb DNA ladder is shown on the right. All samples shown were selected for sequencing.

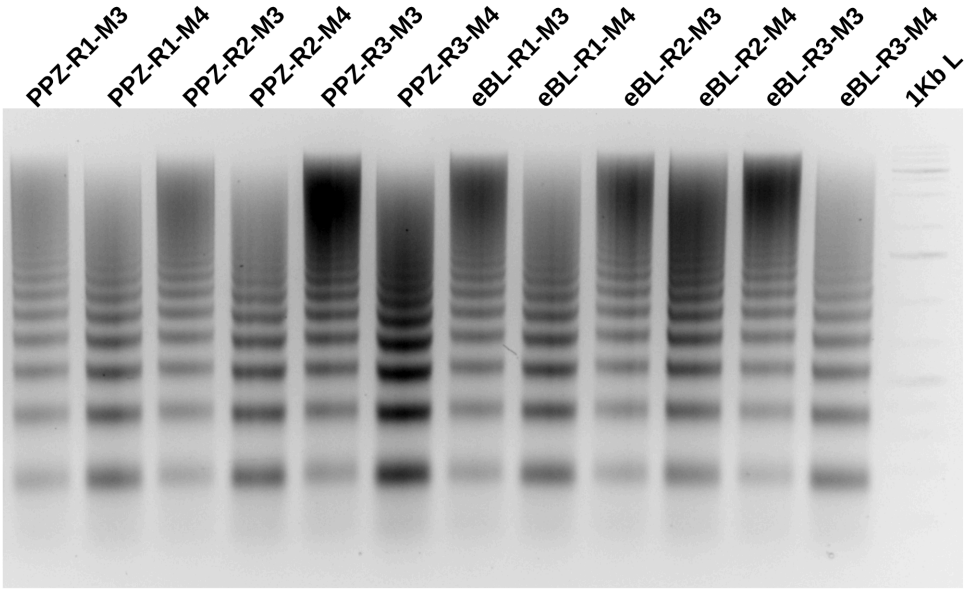


Figure 3.9: MNase partial digestion profiles of MOA-seq in *Hordeum vulgare.*
Agarose gel showing MNase digestion from Morex's first leaf treated with PPZ or PPZ followed by eBL for 4 hours, with three biological replicates per condition. Mock samples had no PPZ and eBL treatment. A 1 kb DNA ladder is shown on the right. PPZ-R1-M3, PPZ-R2-M3, PPZ-R3-M3, eBL-R1-M4, eBL-R2-M4, and eBL-R3-M4 were selected for

sequencing.

Table 3.4: Sequencing reads mapping statistics for *Arabidopsis thaliana* and *Hordeum vulgare* samples treated with PPZ and PPZ+eBL.

Sample	Treatment	Raw reads	Mapped reads	Unique mapped reads
Ara-PPZ	PPZ	755730296	744368725	620450704
Ara-eBL	PPZ + eBL	696528002	686067590	569755252
Morex-PPZ	PPZ	491221142	412869146	247828031
Morex-eBL	PPZ + eBL	499059093	424560797	265833504

3.3.3 Genome-wide and locus-specific changes in MOA-seq occupancy under altered brassinosteroid levels

MOA-seq profiles generated under PPZ and PPZ+eBL conditions were next examined to assess changes in regulatory occupancy associated with altered BR levels. MOA-seq peaks were identified using MACS3 by explicitly contrasting the two treatment conditions. Peak calling was also performed in both directions, with PPZ-treated samples used as the test condition and PPZ+eBL-treated samples as the control, and in a reciprocal comparison with PPZ+eBL-treated samples as the test condition and PPZ-treated samples as the control. For *Hordeum vulgare*, a cutoff value of 0.05 was applied, whereas a more stringent cutoff of 0.00001 was used for *Arabidopsis thaliana*, reflecting differences in sequencing depth relative to genome size (Table 3.5). This reciprocal peak-calling strategy enabled the identification of MOA-seq peaks preferentially detected under one treatment condition relative to the other. In addition to these differential peak sets, MOA-seq peaks identified independently in each treatment condition were analyzed without cross-condition comparison. The number of MOA-seq peaks identified in each reciprocal comparison differed between two species and between contrasting directions (Table 3.5). In *Hordeum vulgare*, 4,080 peaks were identified in the PPZ vs PPZ+eBL comparison, whereas 9,081 peaks were identified in the reciprocal PPZ+eBL vs PPZ comparison using a q value cutoff of 0.05. In *Arabidopsis thaliana*, 41,495 peaks were identified in the PPZ vs PPZ+eBL comparison

and 25,392 peaks in the reciprocal PPZ+eBL vs PPZ comparison using a more stringent cutoff of q = 0.00001. These differential peak sets provided the basis for both locus-specific and genome-wide analyses described below.

Table 3.5: Numbers of differentially bound MOA peaks between PPZ and PPZ+eBL treatments in *Hordeum vulgare* and *Arabidopsis thaliana.*

This table lists peaks determined by MACS3 at a q-value below 0.05 and 0.0001 for barley and Arabidopsis, respectively, using either PPZ treatment (PPZ) or PPZ plus eBL treatments (PPZ+BL) as input (first position) or as a control treatment (second position). This table was generated by Dr. Julia Engelhorn and Dr. Thomas Hartwig.

Species	PPZ vs. PPZ+eBL	PPZ+eBL vs. PPZ	Q value cutoff
Hordeum vulgare	4080	9081	0.05
Arabidopsis thaliana	41495	25392	0.00001

The analysis next focused on the DWF4 locus in *Arabidopsis thaliana* and *Hordeum vulgare* to evaluate intron-associated occupancy changes at a defined BR biosynthesis gene. Intronic MOA-seq signal was compared between these two conditions, providing a gene-level framework to assess how altered BR levels relate to intron-localized transcription factor occupancy. Although reciprocal peak calling was performed in both directions, differential peaks overlapping *HvDWF4* were detected only in the PPZ vs. PPZ+eBL comparison. No significant peaks were identified at the *HvDWF*4 locus in the reciprocal PPZ+eBL vs. PPZ analysis at the applied q value threshold (Figure 3.10). Three intron-overlapping peaks were identified within the *HvDWF4* gene body. Two peaks were located in the fifth intron at 4H:51,736,3994-51,736,4108 and 4H:51,736,4297-51,736,4389, and one peak was located in the seventh intron at 4H:51,736,5510-51,736,5624. Peak 1 exhibited a fold enrichment of 2.42 with a −log10(q value) of 13.84. Peak 2 showed a fold enrichment of 1.77 with a −log10(q value) of 2.85. Peak 3 displayed a fold enrichment of 1.93 with a −log10(q value) of 5.92 (Table 3.6). These results indicate that intron-associated MOA-seq occupancy at *HvDWF4* is detected under PPZ treatment relative to PPZ+eBL treatment.

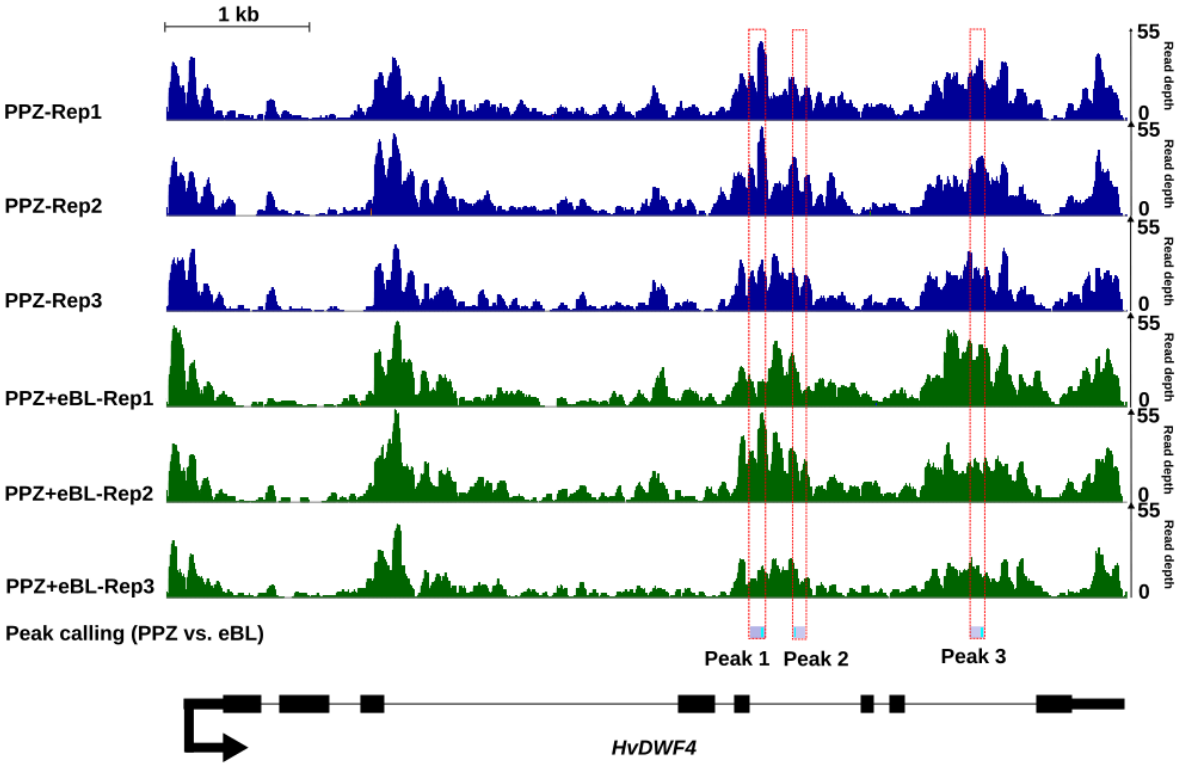


Figure 3.10: MOA-seq signal at the *HvDWF4* locus under PPZ and PPZ+eBL conditions.

Genome browser view of MOA-seq coverage across the HvDWF4 locus in barley. Three biological replicates are shown for PPZ-treated samples (blue) and PPZ+eBL-treated samples (green). Coverage tracks were generated directly from BAM files and are displayed as raw read depth with identical y-axis scaling across all tracks. Red dashed boxes indicate regions corresponding to differential peaks identified by MACS3 in the PPZ vs PPZ+eBL comparison. Three peaks (Peak 1, 2, and 3) overlap intronic regions of *HvDWF4*. The gene model is shown at the bottom, with the arrow indicating the transcription orientation, exons indicated as black boxes, and introns as connecting lines. The scale bar represents 1 kb.

Table 3.6: Differential MOA-seq peaks overlapping the intron of *the HvDWF4* in barley.

Summary of MACS3-identified differential peaks within the *HvDWF4* genomic interval for the PPZ vs. PPZ+eBL comparison. For each peak, chromosome (chr), genomic start and end positions, peak length, absolute summit position (abs_summit), pileup value,

−log10(p value), fold enrichment, and −log10(q value) are reported as provided by MACS3. Peak names correspond to the original identifiers in the MACS3 output files. All peaks shown passed the applied q-value threshold (q ≤ 0.05).

Peak No.	Peak calling	chr	start	end	length	abs_summit	pileup	-log10(p value)	fold_enrichment	-log10(q value)	name
Peak 1	PPZ vs. eBL	4H	517363994	517364108	115	517364083	132	18.9475	2.41504	13.8368	Morex_PPZ_R1-3.vs.Morex_BL_R1-3.q0.05_peak_3414
Peak 2	PPZ vs. eBL	4H	517364297	517364389	93	517364309	88	6.58727	1.76552	2.85403	Morex_PPZ_R1-3.vs.Morex_BL_R1-3.q0.05_peak_3415
Peak 3	PPZ vs. eBL	4H	517365510	517365624	115	517365608	114	10.2493	1.92524	5.92102	Morex_PPZ_R1-3.vs.Morex_BL_R1-3.q0.05_peak_3416

In contrast to barley, differential peaks overlapping *AtDWF4* were detected in both directional comparisons. All identified peaks were located within the last intron of *AtDWF4* (Figure 3.11). In the PPZ vs. PPZ+eBL comparison, two peaks were detected at Chr3:18,814,835-18,815,113 (Peak 1) and Chr3:18,814,595-18,814,667 (Peak 2.1). In the reciprocal eBL vs. PPZ comparison, a peak was identified at Chr3:18,814,593–18,814,665 (Peak 2.2) (Table 3.7). Peak 2.1 and Peak 2.2 overlap nearly completely and represent the same intronic region detected in opposite directional contrasts. Differences in enrichment statistics reflect the reciprocal peak-calling framework rather than distinct genomic locations. These results indicate that the last intron of *AtDWF4* shows condition-dependent MOA-seq occupancy under both PPZ and PPZ+eBL treatments. However, all differential peaks reported here were defined using MACS3 with the applied q-value thresholds. Visual inspection of the coverage tracks from BAM files alone does not allow a reliable distinction between random variation and statistically supported enrichment. Therefore, the interpretation of condition-specific occupancy at the DWF4 locus in both species is based on peaks that met the predefined statistical criteria of the peak calling tool.

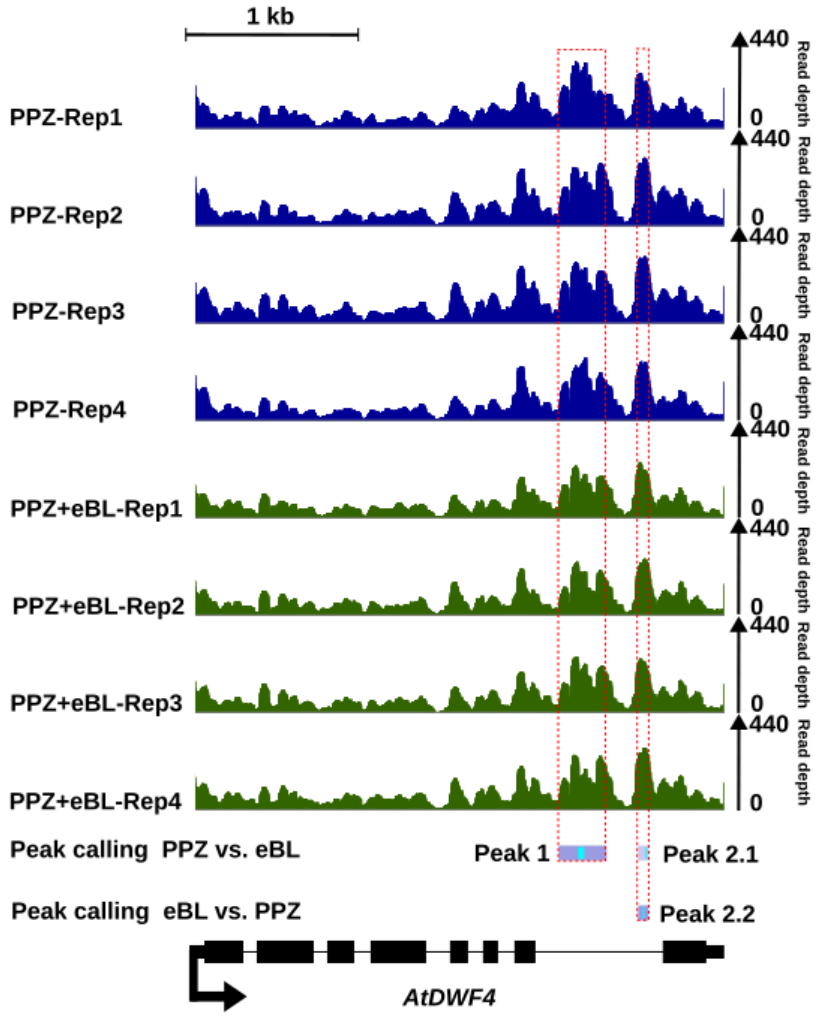


Figure 3.11: MOA-seq signal at the *AtDWF4* locus under PPZ and PPZ+eBL conditions.

Genome browser view of MOA-seq coverage across the *AtDWF4* locus in Arabidopsis. Four biological replicates are shown for PPZ-treated samples (blue) and PPZ+eBL-treated samples (green). Coverage tracks were generated directly from BAM files and are displayed as raw read depth with identical y-axis scaling across all tracks. Red dashed boxes indicate regions corresponding to differential peaks identified by MACS3 in the PPZ vs. PPZ+eBL and reciprocal comparisons. Peak 1 was identified in the PPZ vs. PPZ+eBL comparison, whereas Peak 2.1 and Peak 2.2 correspond to directional peak calls from the PPZ vs. PPZ+eBL and eBL vs. PPZ analyses, respectively. The *AtDWF4* gene model is shown at the bottom, with the arrow indicating the transcription orientation, exons indicated as black boxes, and introns as connecting lines. The scale bar represents 1 kb.

Table 3.7: Differential MOA-seq peaks overlapping the intron of the *AtDWF4* in Arabidopsis.

Summary of MACS3-identified differential peaks within the *AtDWF4* genomic interval for the PPZ vs. PPZ+eBL comparison. For each peak, chromosome (chr), genomic start and end positions, peak length, absolute summit position (abs_summit), pileup value, −log10(p value), fold enrichment, and −log10(q value) are reported as provided by MACS3. Peak names correspond to the original identifiers in the MACS3 output files. All peaks shown passed the applied q-value threshold (q ≤ 0.00001).

Peak No.	Peak calling	chr	start	end	length	abs_summit	pileup	-log10(p value)	fold_enrichment	-log10(q value)	name
Peak 1	PPZ vs. eBL	Chr3	18814835	18815113	279	18814988	1191.94	25.0332	1.37436	22.7514	Ara_Col0_PPZ_R1-4.vs.Ara_Col0_BL_R1-4.q0.05_peak_29678
Peak 2.1	PPZ vs. eBL	Chr3	18814595	18814667	73	18814612	1213.06	8.45165	1.18561	6.86676	Ara_Col0_PPZ_R1-4.vs.Ara_Col0_BL_R1-4.q0.05_peak_29677
Peak 2.2	eBL vs PPZ	Chr3	18814593	18814665	73	18814627	1222	35.3238	1.46019	32.0259	Ara_Col0_BL_R1-4.vs.Ara_Col0_PPZ_R1-4.q0.05_peak_24457

To assess whether altered BR conditions are associated with changes in the genome-wide distribution of MOA-seq signal, MOA-seq peaks were analyzed at the whole-genome level. In *Hordeum vulgare*, the genomic distribution of MOA-seq peaks differed among peak sets defined by BR treatment condition (Figure 3.12). To quantify these patterns, base pair overlap between MOA-seq peaks and annotated genomic features was assessed for peaks identified independently under PPZ and PPZ+eBL conditions, as well as for differential peak sets identified by reciprocal peak calling. Overlap was quantified as the proportion of total MOA-seq peak base pairs assigned to each genomic feature category. MOA-seq peaks identified in PPZ-treated samples and in PPZ+eBL-treated samples showed broadly similar distributions across major genomic features. In both conditions, the majority of MOA-seq peak base pairs overlapped intergenic regions, accounting for approximately 50% (~22.0 Mb) under PPZ treatment and ~51% (~22.6 Mb) under PPZ+eBL treatment (Figure 3.12, Appendix 5.3 - Table 5.1). Promoter-proximal regions within 1 kb upstream of annotated genes accounted for approximately 16% (~6.9 Mb) and ~16% (~7.2 Mb) of peak base pairs under PPZ and

PPZ+eBL treatment, respectively. In contrast, intronic regions accounted for a smaller fraction of total MOA-seq peak coverage, with approximately 4% (~1.9 Mb) under PPZ treatment and ~4% (~1.8 Mb) under PPZ+eBL treatment. Differences were more apparent in peak sets identified by reciprocal comparison between treatment conditions. When peaks were identified by contrasting PPZ-treated samples against PPZ+eBL-treated samples (PPZ vs PPZ+eBL), a higher proportion of MOA-seq peak base pairs was assigned to regions upstream of annotated genes, particularly within 1 kb of transcription start sites. Conversely, peaks identified in the reciprocal comparison (PPZ+eBL vs PPZ) showed a relatively higher contribution of intronic and distal intergenic regions (Figure 3.12, Appendix 5.3 - Table 5.1). In the differential peak sets identified by reciprocal peak calling, the relative contribution of intronic regions differed between the two contrasts. In the PPZ vs PPZ+eBL comparison, intronic regions accounted for approximately 0.016 Mb, corresponding to ~1.6% of total MOA-seq peak base pairs. In contrast, in the reciprocal PPZ+eBL vs PPZ comparison, intronic regions accounted for approximately 0.045 Mb, corresponding to ~7.1% of total MOA-seq peak base pairs (Figure 3.12, Appendix 5.3 - Table 5.1).

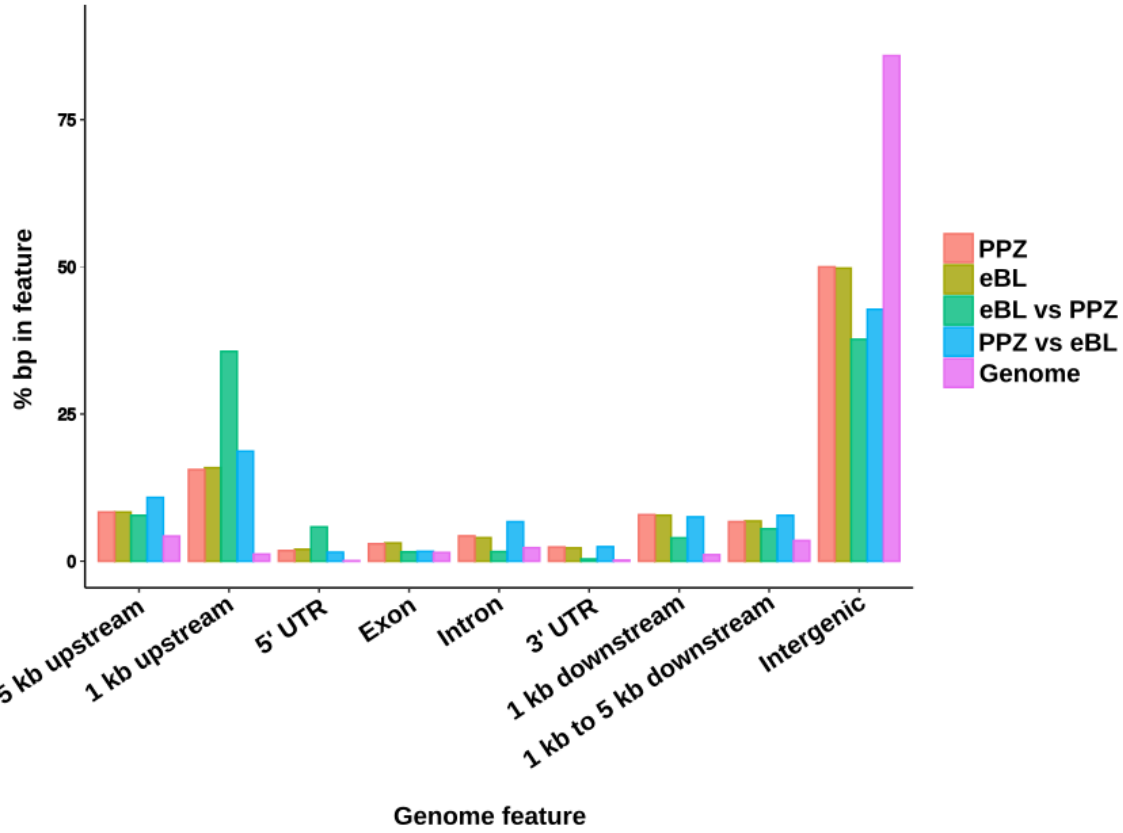


Figure 3.12: Genomic distribution of MOA peaks under altered brassinosteroid levels in *Hordeum vulgare*.

Bar plots show the proportion of base pairs assigned to different genomic features for the whole genome (purple), MOA peaks detected under propiconazole (PPZ) treatment (red), MOA peaks detected under eBL treatment (moss green), and differentially bound MOA peaks between eBL and PPZ conditions (green and blue). Genomic features include 1-5 kb upstream, 1 kb upstream, 1-5 kb downstream, 1 kb downstream, 5'UTR, 3'UTR, exon, intron, and intergenic regions relative to annotated genes. Percentages indicate the fraction of total base pairs falling within each feature category. This figure was created by Dr. Julia Engelhorn.

In *Arabidopsis thaliana*, the genomic distribution of MOA-seq peaks was examined across peak sets defined both by treatment condition and by reciprocal peak calling (Figure 3.13). MOA-seq peaks were first identified independently for PPZ- and PPZ+eBL-treated samples. In addition, differential peak sets were generated by contrasting PPZ-treated samples against PPZ+eBL-treated samples and, in a reciprocal comparison, PPZ+eBL-treated samples against PPZ-treated samples using MACS3 with a stringent cutoff (q = 0.00001; Table 3.13). Base pair overlap between MOA-seq peaks and annotated genomic features were quantified for all peak sets. MOA-seq peaks identified independently under eBL and PPZ conditions showed broadly similar distributions across genomic features. In both conditions, a large fraction of MOA-seq peak base pairs overlapped promoter-proximal regions within 1 kb upstream of annotated genes, accounting for approximately 33% (~7.4 Mb) under eBL treatment and ~36% (~7.9 Mb) under PPZ treatment (Figure 3.13, Appendix 5.3 - Table 5.2). Exonic regions accounted for a substantial fraction of peak coverage, whereas intronic regions accounted for a smaller proportion, with approximately 8% (~1.8 Mb) under eBL treatment and ~7% (~1.5 Mb) under PPZ treatment. In contrast to barley, only a small fraction of MOA-seq peak base pairs overlapped intergenic regions in Arabidopsis under both treatment conditions. Differences were more apparent in peak sets identified by reciprocal comparison between treatments. In the eBL vs PPZ comparison, intronic regions accounted for approximately 7.7% (~0.36 Mb) of total MOA-seq peak base pairs, whereas in the reciprocal PPZ vs eBL comparison, intronic regions accounted for approximately 5.0% (~0.31 Mb). In the PPZ vs eBL comparison, promoter-proximal

regions within 1 kb upstream of annotated genes accounted for the largest fraction of MOA-seq peak base pairs, representing approximately 52.5% (~3.3 Mb) of total peak coverage (Figure 3.13, Appendix 5.3 - Table 5.2).

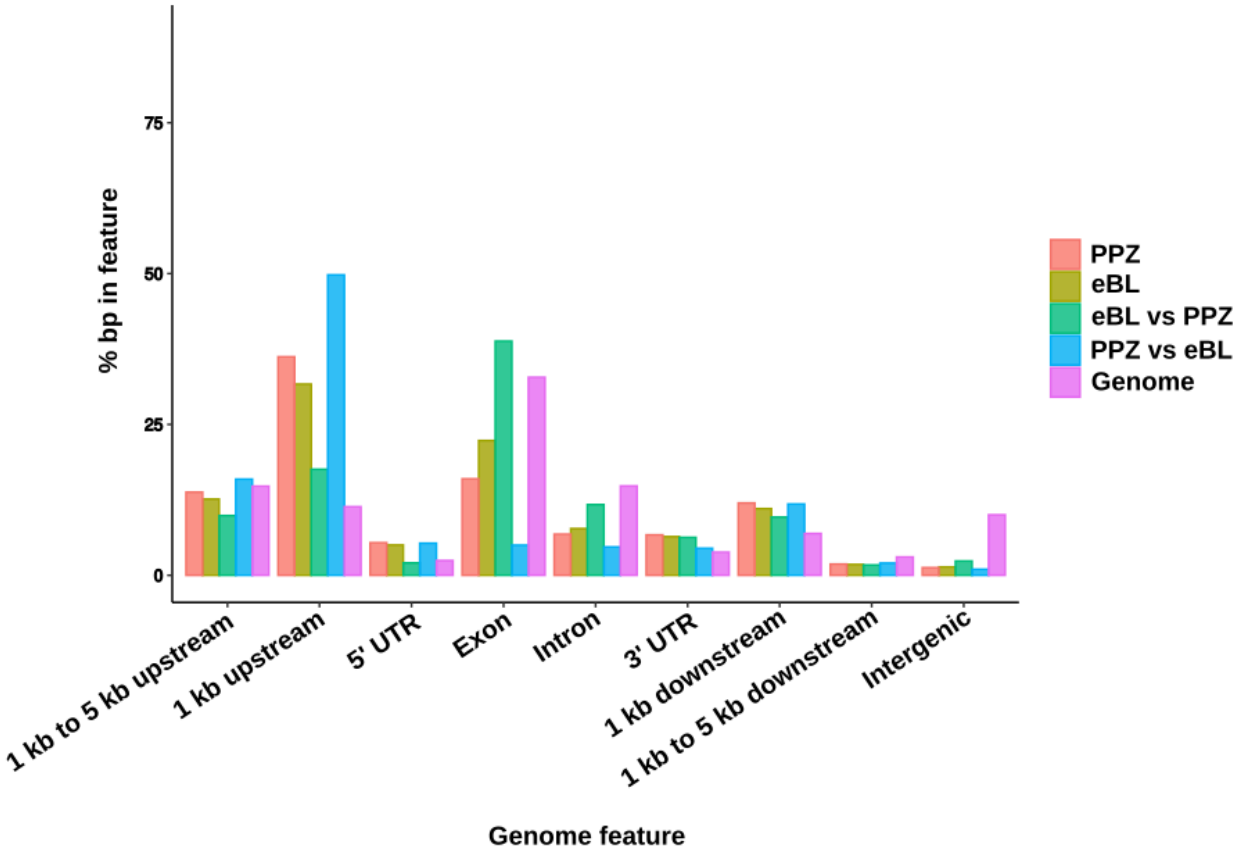


Figure 3.13: Genomic distribution of MOA peaks under altered brassinosteroid levels in *Arabidopsis thaliana*.

Bar plots show the proportion of base pairs assigned to different genomic features for the whole genome (purple), MOA peaks detected under propiconazole (PPZ) treatment (red), MOA peaks detected under eBL treatment (moss green), and differentially bound MOA peaks between eBL and PPZ conditions (green and blue). Genomic features include 1-5 kb upstream, 1 kb upstream, 1-5 kb downstream, 1 kb downstream, 5'UTR, 3'UTR, exon, intron, and intergenic regions relative to annotated genes. Percentages indicate the fraction of total base pairs falling within each feature category. This figure was created by Dr. Julia Engelhorn.

3.4. Discussion

A central aim of this chapter was to examine how a conserved signaling pathway can maintain regulatory function across species while accommodating changes in cis-regulatory architecture. The brassinosteroid biosynthesis pathway provides a suitable system for this question, as its core components and feedback regulation are conserved across land plants, whereas genome size, structure, and intron architecture differ substantially among species (Zebosi et al. 2024).

In this study, intron-localized MOA-seq signal was detected at multiple BR biosynthesis genes in *Zea mays*, *Arabidopsis thaliana*, and *Hordeum vulgare*. However, such intronic signal was not observed uniformly across all pathway genes, nor consistently across tissues or species (Figures 3.2-3.6). This pattern indicates that intronic CREs contribute to TF occupancy at a subset of BR biosynthesis loci, rather than representing a universal regulatory feature of the pathway. From an evolutionary perspective, these observations are compatible with a regulatory architecture in which conserved pathway-level outputs are maintained through flexible use of different genomic compartments, including promoters, introns, and distal regions, without requiring strict conservation of cis-element position or sequence (Ciren et al. 2024). In this context, introns represent one potential substrate for regulatory diversification within a conserved signaling pathway. Within this broader framework, the DWF4 locus provides a well-defined entry point for examining intronic regulatory features in the BR pathway. Previous work using hybrid allele-specific BZR1 ChIP-seq identified intron-localized BZR1 binding at maize DWF4, with haplotype-specific differences in occupancy and associations with BR-responsive mRNA abundance and agronomic traits (Hartwig et al. 2023). In this study, MOA-seq signal was detected at the same intronic region of maize DWF4, consistent with MOA-seq capturing TF occupancy at loci previously identified by ChIP-seq (Figure 3.2a). Intronic MOA-seq signal at DWF4 was detected in both leaf and earshoot tissue, indicating that this feature can be observed in more than one maize tissue context, although direct equivalence to the meristem tissue used for ChIP-seq should not be assumed. Together, these observations support the use of DWF4 as a reference locus for examining intronic TF occupancy within the BR pathway, without implying that this regulatory configuration is unique to this gene or invariant across conditions.

The detection of intronic occupancy at DWF4 can be placed within the known mechanistic context of BR feedback regulation. BZR1 functions in both transcriptional activation and repression depending on its interaction partners and chromatin context. In Arabidopsis, BZR1 has been shown to recruit the corepressor TOPLESS at BR-responsive loci, providing a mechanistic framework for transcriptional repression within the pathway (Oh, Zhu, Ryu, et al. 2014). BZR1 also interacts with other transcription factors involved in growth regulation, including PIF4 and ARF6, and ChIP-seq analyses have revealed overlapping binding landscapes of these factors at genes associated with hormone-responsive transcription and cell elongation (Oh, Zhu, Bai, et al. 2014). In this study, the MOA-seq signal at selected BR biosynthesis genes in Arabidopsis overlapped regions previously reported to be bound by BZR1, PIF4, and ARF6 (Figure 3.5). These overlaps place intronic and promoter-proximal MOA-seq signals within the context of known BR-associated TF networks, without implying direct regulatory interactions or functional consequences for individual binding events. Extending beyond DWF4, intronic MOA-seq signal was detected at additional BR biosynthesis genes in maize, including *ZmCYP90D1*, *ZmCYP90C*, *ZmDWF7*, *ZmDWF5*, and *ZmNa2/DWF1*, whereas other pathway genes did not show detectable intronic signal under the sequencing depth and peak-calling parameters applied (Table 3.1). Similar gene-specific patterns were observed in *Arabidopsis thaliana* and *Hordeum vulgare*, where intronic MOA-seq signal was present at some, but not all, BR biosynthesis genes (Figures 3.2 and 3.3). This non-uniform distribution indicates that intron-associated TF occupancy was detected at a subset of BR pathway genes under the examined conditions. Non-detection at individual loci can be interpreted within the limits of sequencing coverage and statistical thresholds and does not exclude low-level or condition-specific occupancy below the detection sensitivity of the present datasets. Within these analytical limits, intronic regulatory input can be detected at specific BR biosynthesis genes, consistent with locus- and condition-dependent occupancy.

The PPZ and PPZ+eBL experiments extend the locus-specific analysis at DWF4 to a genome-wide level. MOA-seq peaks identified independently under each condition showed similar overall distributions across promoters, introns, and intergenic regions in both *Arabidopsis thaliana* and *Hordeum vulgare* (Figure 3.12 and 3.13, Appendix 5.3). Differences became clearer when peaks were defined by reciprocal contrasts between

conditions. In *Hordeum vulgare*, the PPZ vs. PPZ+eBL comparison yielded a higher proportion of promoter-proximal peaks, whereas the reciprocal comparison contained a relatively larger fraction of intronic and distal intergenic peaks (Figure 3.12). In *Arabidopsis thaliana*, reciprocal peak calling also altered the feature composition of differential peaks, although the magnitude and direction of these shifts differed from barley (Figure 3.13). These observations indicate that the genomic features represented among condition-specific peaks depend on the direction of the statistical contrast and are not uniformly distributed across the genome.

At the DWF4 locus, reciprocal peak calling yielded different sets of significant peaks in the two contrast directions. In barley, intron-overlapping peaks of *HvDWF4* were detected only in the PPZ vs. PPZ+eBL comparison. In Arabidopsis, intron-overlapping peaks of *AtDWF4* were detected in both contrasts, with overlapping coordinates in the last intron identified in reciprocal analyses. These results indicate that intronic regions of DWF4 are included among the differential MOA-seq peaks detected under altered BR conditions. At both loci, peaks overlapping DWF4 introns were identified in at least one contrast direction, indicating that MOA-seq occupancy at these intronic regions differs between PPZ and PPZ+eBL treatments (Figure 3.10 and 3.11; Table 3.6 and 3.7). However, the identification of these peaks is based on the statistical criteria implemented in MACS3 and the applied q-value thresholds. Therefore, the conclusion that DWF4 intronic occupancy differs between treatments is contingent on the peak-calling framework used in this analysis.

Taken together, the gene-level and genome-wide analyses support a model in which BR biosynthesis genes are regulated through a combination of regulatory inputs distributed across multiple genomic compartments. At loci such as DWF4, intronic regions represent one location where TF occupancy can be detected and may contribute to feedback regulation, consistent with prior functional and ChIP-seq evidence in maize. At the same time, the non-uniform presence of intronic MOA-seq signal across genes, tissues, species, and treatment conditions indicate that intronic regulation is not a universal feature of BR biosynthesis genes. Within this model, intronic CREs are best viewed as flexible components of a broader regulatory architecture rather than as fixed or conserved control points. Such flexibility would allow conserved pathway-level outputs to be maintained while permitting turnover in the genomic location, number, or context of

individual cis-elements. This interpretation is compatible with previous observations that BR-responsive transcription involves context-dependent recruitment of BZR1 and associated cofactors, including TOPLESS, PIF4, and ARF6, and that regulatory outcomes depend on combinatorial interactions rather than single binding events.

Several limitations should be considered when interpreting these results. First, MOA-seq detects regions of protein-DNA occupancy but does not identify the specific TFs involved or their functional consequences. Second, comparisons between MOA-seq and published ChIP-seq datasets are necessarily indirect, as these data were generated in different tissues, developmental stages, and experimental conditions. Third, differential peak sets are sensitive to peak-calling parameters of the peak-calling tool and sequencing depth, particularly in reciprocal comparisons across treatments and species. Despite these limitations, the consistent detection of intron-localized MOA-seq signal at selected BR biosynthesis genes across species highlights introns as plausible contributors to regulatory evolution within a conserved signaling pathway. Functional testing of candidate intronic elements will be required to determine whether these regions contribute to transcriptional regulation and feedback control. Reporter-based assays, such as transient expression systems tested under PPZ and eBL treatment, could provide an initial assessment of condition-dependent regulatory activity for selected fragments. However, such assays do not capture chromatin context at the endogenous locus. Definitive evaluation of regulatory function will therefore require targeted perturbation approaches, including CRISPR-mediated deletion or repression of candidate intronic regions in planta. Together, these approaches would provide a direct link between the occupancy patterns described here and their regulatory consequences, helping to clarify how conserved hormonal pathways remain robust while their cis-regulatory architectures evolve.

Chapter 4: Characterization of genes identified by FIND-CIS for bioengineering of drought and pathogen tolerance in maize

4.1. Introduction

Zea mays is one of the world's most important food crops and an exceptionally informative system for studying cis-regulatory variation, owing to its extensive genetic diversity, high TE content, and the availability of well-characterized mapping populations. In maize, variation in CREs contributes substantially to phenotypic diversity and often explains levels of trait variation comparable to, or exceeding, those caused by coding-sequence polymorphisms. These properties make maize a powerful model for linking natural regulatory variation to agriculturally relevant traits (Nannas and Dawe 2015; Rodgers-Melnick et al. 2016). To identify functional cis-regulatory variants in maize, our laboratory recently developed the F1-mediated Identification of Non-coding Cis-element-Impacting SNPs (FIND-CIS) approach, which combines allele-specific MOA-seq with F1 hybrid analysis (Engelhorn et al. 2025). By comparing TF occupancy between alleles in a shared trans environment, FIND-CIS enables the robust identification of cis-acting sequence variants that alter TF binding. When integrated with genome-wide association studies (GWAS) from the maize Nested Association Mapping (NAM) population, FIND-CIS provides a framework to associate regulatory variation with quantitative trait loci linked to stress resilience and development. Application of FIND-CIS to 25 B73 × NAM hybrids generated a genome-wide atlas of allele-specific TF footprints and identified candidate cis-regulatory variants near genes implicated in abiotic and biotic stress responses. However, the identification of candidate CRE variants represents only the first step. Whether these variants causally affect gene regulation and contribute to stress tolerance or disease resistance requires experimental validation.

In this chapter, candidate genes identified by FIND-CIS are experimentally evaluated for their potential roles in drought tolerance and pathogen resistance. Seven genes were prioritized based on four criteria: (i) the presence of a haplotype-specific MOA footprint difference within ±5 kb of the gene, (ii) proximity to GWAS signals for drought- or disease-related traits, (iii) allele-specific expression differences in hybrid transcriptome data, and (iv) prior literature linking the gene to relevant biological processes. The selected candidates span multiple stress-related pathways.

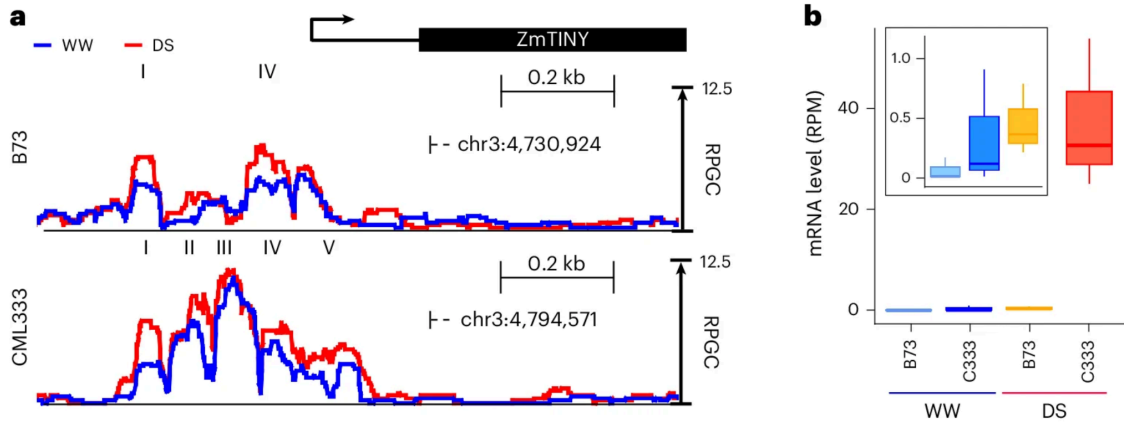


Figure 4.1: Haplotype-specific MOA-seq signal and expression at *ZmTINY*.
(a) Average MOA-seq coverage (three biological replicates) at the *ZmTINY* locus in B73 × CML333 hybrids under well-watered (WW) and drought stress (DS) conditions. Allele-specific profiles are shown for B73 and CML333. Roman numerals mark drought-associated MOA-seq peak regions identified by MACS3 ($P < 0.05$). (b) Allele-resolved transcript levels of *ZmTINY* in B73 × CML333 hybrids under WW and DS conditions (three biological replicates). The inset highlights expression differences among WW-B73, WW-CML333, and DS-B73 samples. This figure was copied from Engelhorn et al. 2025.

ZmTINY is a maize homolog of the Arabidopsis AP2/ERF transcription factor TINY, which has been implicated in the integration of brassinosteroid signaling and drought responses in Arabidopsis (Xie et al. 2019). In the FIND-CIS dataset, the *ZmTINY* locus showed pronounced haplotype-specific MOA-seq footprints in the promoter region in B73 × CML333 hybrids under both well-watered and drought-stress conditions (Figure 4.1a). Several MOA peaks were detected in both haplotypes, but their relative intensity and distribution differed between B73 and CML333, indicating allele-dependent differences in TF occupancy. Consistent with these occupancy patterns, allele-specific mRNA abundance differed between the two haplotypes, with higher transcript accumulation observed for the CML333 allele under drought conditions (Figure 4.1b). Together, these data place *ZmTINY* among loci showing coordinated haplotype-dependent differences in TF occupancy and transcript abundance under stress-related conditions. Although *ZmTINY* represents a strong candidate based on regulatory and expression data, no corresponding BonnMu insertion lines were available

for this gene in the BonnMu collection. As a result, *ZmTINY* could not be included in the genotyping and mutant-based analyses performed in this chapter.

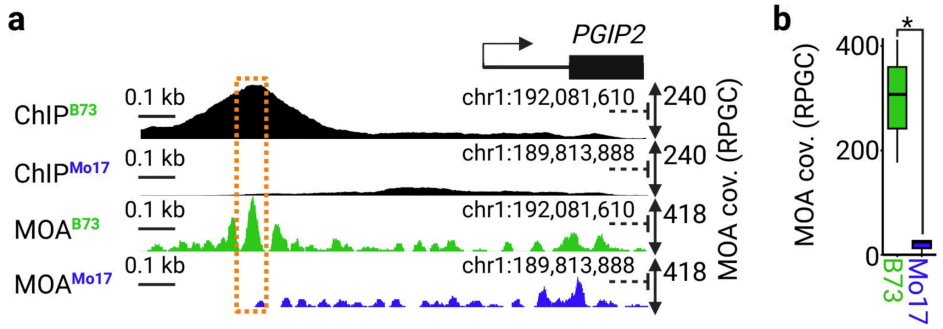


Figure 4.2: Allele-specific transcription factor occupancy at the *ZmPGIP2* locus. (a) Genome browser view showing averaged, normalized allele-specific ZmBZR1 ChIP-seq signal (top black tracks) together with MOA-seq signal (bottom tracks; B73 shown in green and Mo17 in blue) at the region upstream of *ZmPGIP2*. MOA-seq tracks represent fragment-center signal averaged across three biological replicates. The dashed box marks the region used for quantitative analysis. (b) Normalized MOA-seq coverage within the highlighted upstream region of *ZmPGIP2*, averaged across three biological replicates. Differences between alleles were assessed using a two-sided t-test (P = 0.015). This figure was copied from Engelhorn et al. 2025.

ZmPGIP2 encodes a polygalacturonase-inhibiting protein that has been implicated in cell wall-associated defense responses in maize and other plant species (Ferrari et al. 2003; Kalunke et al. 2015; Kump et al. 2011). This locus was prioritized based on pronounced allele-specific differences in TF occupancy detected upstream of the gene in hybrid backgrounds. In B73 × Mo17 hybrids, MOA-seq signal showed substantially higher occupancy on the B73 allele relative to the Mo17 allele within a defined upstream region that overlaps a previously reported BZR1 ChIP-seq peak (Figure 4.2a). This allele-dependent difference was consistent across biological replicates and was quantified as a marked contrast in average MOA coverage between the two alleles (Figure 4.2b). The colocalization of allele-specific MOA-seq signal with TF binding detected by ChIP-seq supports the interpretation that this upstream region represents a site of differential TF occupancy between haplotypes. PGIP genes have been associated with quantitative variation in resistance to fungal pathogens, and the observed regulatory

differences at *ZmPGIP2* coincide with genomic regions previously linked to disease resistance traits. Together, these observations motivated the inclusion of *ZmPGIP2* for functional interrogation, without implying a direct or exclusive regulatory mechanism.

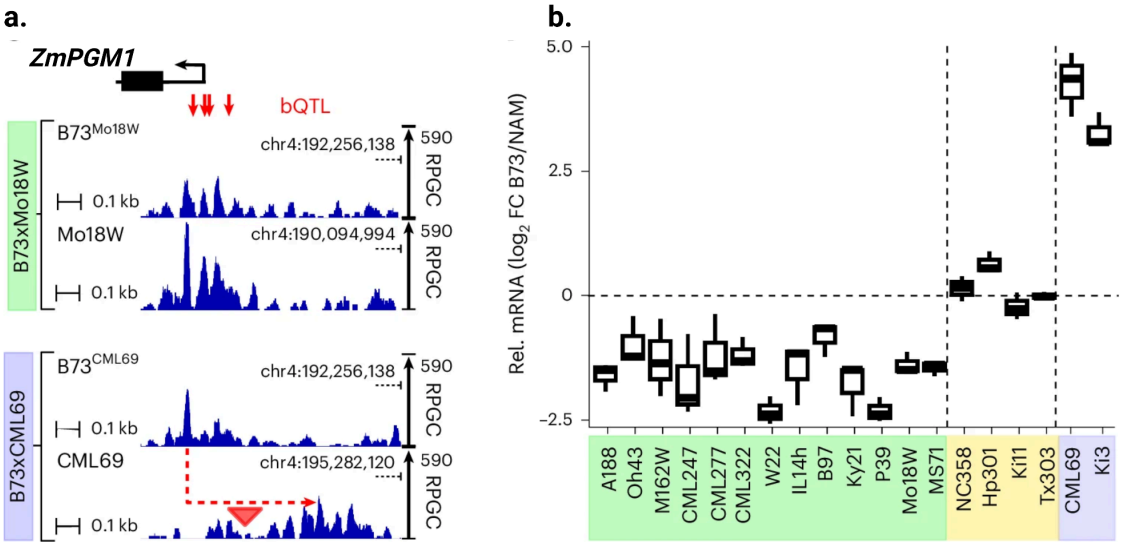


Figure 4.3: Allele-specific transcription factor occupancy and mRNA abundance variation at the *ZmPGM1* locus.

(a) Genome browser view showing averaged, normalized MOA-seq signal upstream of *ZmPGM1* in F1 hybrids B73 × Mo18W and B73 × CML69. Tracks represent allele-specific MOA-seq fragment-center signal averaged across biological replicates. Red arrows indicate the position of a previously identified bQTL, and the red triangle marks a MITE transposon insertion present in the CML69 haplotype, associated with a shift in the position of MOA-seq peaks. (b) Relative allele-specific *ZmPGM1* mRNA abundance across B73 × NAM F1 hybrids, shown as $\log_2$ fold change of B73 relative to the NAM allele. Hybrids are grouped based on MOA-seq occupancy patterns at the upstream regulatory region, highlighting associations between allele-specific TF occupancy and transcriptional output. This figure was copied from Engelhorn et al. 2025.

ZmPGM1 (PHOSPHOGLYCERATE MUTASE1) encodes phosphoglucomutase, a central enzyme in carbohydrate metabolism that links glycolysis, starch biosynthesis, and broader metabolic fluxes. *ZmPGM1* was prioritized based on allele-specific differences in

TF occupancy detected upstream of the locus. In B73 × Mo18W hybrids, MOA-seq profiles showed higher occupancy on the Mo18W allele relative to B73 at a defined upstream region that overlaps a previously identified binding QTL (bQTL) (Figure 4.3a). In contrast, in B73 × CML69 hybrids, the MOA-seq signal upstream of *ZmPGM1* was similar in overall magnitude between alleles but exhibited a shift in genomic position in the CML69 haplotype. This positional shift coincided with the presence of a transposable element insertion upstream of the transcription start site, consistent with altered local regulatory architecture rather than changes in overall occupancy strength. These allele-specific differences in TF occupancy were accompanied by corresponding differences in mRNA abundance across a panel of B73 × NAM hybrids (Figure 4.3b). Hybrids carrying NAM alleles associated with higher MOA-seq occupancy generally showed increased relative mRNA abundance compared to B73, whereas haplotypes with shifted or reduced occupancy showed smaller or negative expression differences. Together, these patterns indicate that natural sequence variation upstream of *ZmPGM1* is associated with both altered TF occupancy and quantitative differences in mRNA abundance across maize haplotypes, motivating its inclusion for functional interrogation without assuming a direct causal mechanism.

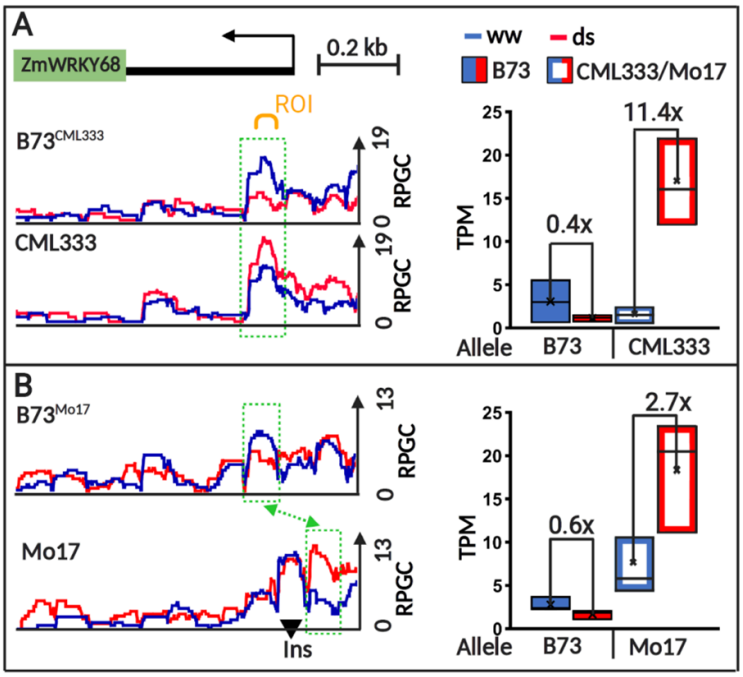


Figure 4.4: Allele-specific regulatory variation at *ZmWRKY68*.
Allele-specific MOA-seq profiles at the *ZmWRKY68* locus in B73 and CML333/Mo17 backgrounds reveal haplotype-dependent differences in TF occupancy near the promoter region. Boxplots show corresponding allele-specific mRNA abundance differences under drought-stressed (DS) and well-watered (WW) conditions. Regions of interest (ROI) indicate candidate quantification of MOA-Seq coverage in the three replicates. This figure was copied from Max Blank's thesis.

ZmWRKY68 was selected as a candidate gene based on haplotype-specific differences in MOA-seq signal detected upstream of the gene and corresponding allele-specific differences in mRNA abundance in hybrid backgrounds. WRKY TFs have been reported to participate in transcriptional responses to both biotic and abiotic stresses, including dehydration-responsive genes and ABA-associated drought response, providing a functional context for examining regulatory variation at this locus (Rushton et al. 2010; Y. Wang et al. 2023; Tang and Xia 2025). In F1 hybrids between B73 and CML333 or Mo17, a defined upstream region showed differential MOA-seq signal between alleles, with higher signal detected at the non-B73 allele in both hybrid combinations (Figure 4.4). This region coincided with differences in mRNA abundance between alleles. Under drought stress conditions, the CML333 and Mo17 alleles showed higher mRNA abundance compared to the B73 allele, whereas under well-watered conditions, the allelic differences were reduced. Together, these observations motivated the inclusion of *ZmWRKY68* for further functional analysis.

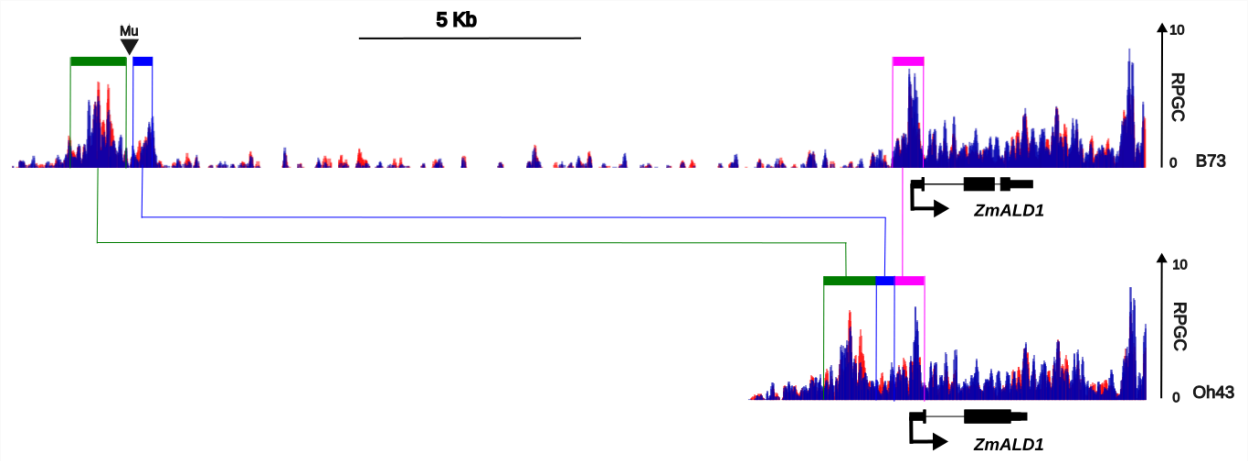


Figure 4.5: Allele-specific MOA-seq signal at the *ZmALD1* (Aldolase) locus.
Average, normalized MOA-seq coverage at the *ZmALD1* locus for the B73 and Oh43 alleles. Blue and red tracks indicate well-watered (WW) and drought-stressed (DS) conditions, respectively. Shaded regions highlight MOA-seq peaks detected at corresponding genomic positions in both alleles. The position of a Mu transposon insertion in the B73 haplotype is indicated by a black triangle.

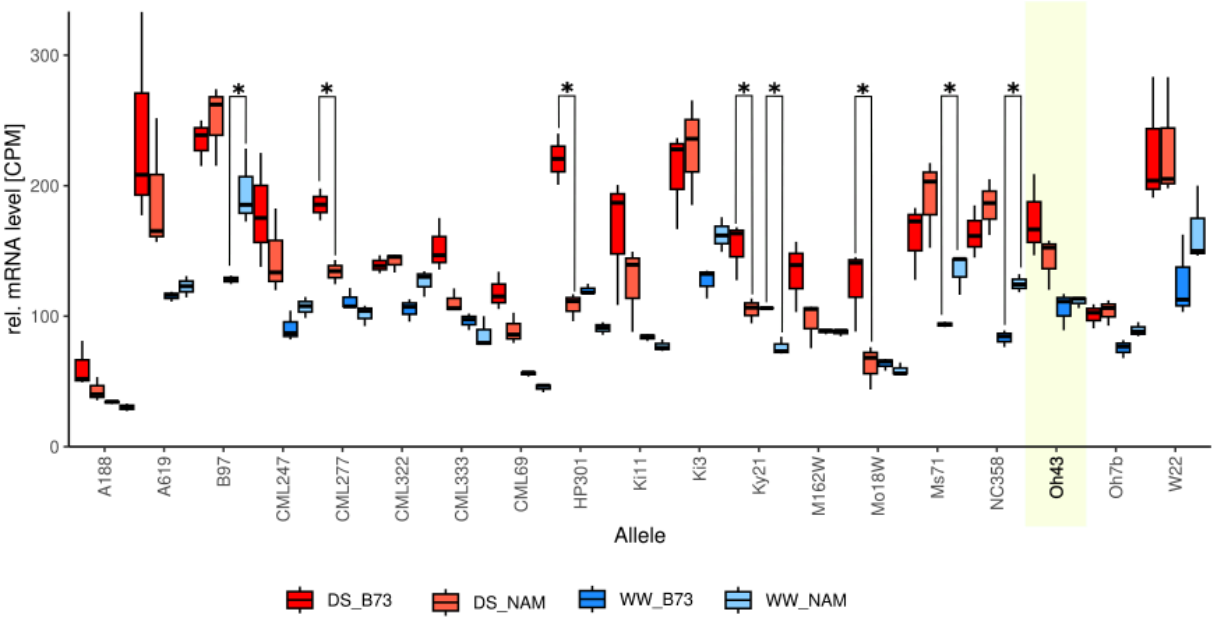


Figure 4.6: mRNA abundance at the Aldolase1 (ZmADL1) locus in 25 maize hybrids.
Relative mRNA abundance of *ZmADL1* in maize NAM founder alleles under well-watered (WW) and drought stress (DS) conditions. Boxplots represent biological replicates, and colors represent treatment and genetic background. The B73 × Oh43 hybrid is marked in yellow. Data presented here relate to the dataset described in Engelhorn et al. (2025), where the experimental design, RNA-seq library preparation, and read alignment are described. For this figure, normalized read counts were pulled per allele and treatment for the Aldolase1 locus using awk. Statistical analysis of allele-specific mRNA abundance between the NAM allele and the B73 allele in each hybrid and treatment was conducted using DESeq2, as described in genome-wide differential expression analyses in Engelhorn et al. (2025). A false discovery rate-corrected p-value threshold of 0.05 was used. Asterisks denote significant

allele-specific differences for the respective hybrid and treatment. The figure was created using R (v4.1.1). This figure was originally created by Dr. Julia Engelhorn.

ZmALD1, encoding a fructose-bisphosphate aldolase involved in central carbon metabolism, was included as a candidate based on haplotype-associated differences in MOA-seq signal and corresponding variation in mRNA abundance across maize lines. Also, it was identified as a candidate gene associated with drought-related traits in maize (C. Li et al. 2016). MOA-seq profiles at the *ZmALD1* locus revealed reproducible peaks in both the B73 and Oh43 haplotypes under well-watered and drought-stressed conditions; however, the positions of these peaks differed between alleles, indicating a positional shift in TF occupancy relative to gene structure (Figure 4.5). A Mu transposon insertion is present in the B73 haplotype at this locus and coincides with the altered genomic positioning of MOA-seq peaks. Despite the detection of MOA-seq signal in both alleles, *ZmALD1* mRNA abundance varied across NAM founder haplotypes and treatment conditions, with the B73 × Oh43 hybrid showing intermediate responses compared to other crosses (Figure 4.6). Together, these observations indicate that regulatory variation at *ZmALD1* is associated with haplotype-specific differences in local genomic context and TF occupancy, motivating its inclusion for functional interrogation without assuming a direct regulatory role for individual MOA-seq peaks.

Collectively, these candidates represent loci with converging genomic evidence for regulatory divergence, providing a focused set of targets for experimental evaluation, while acknowledging that direct phenotypic effects remain to be established. Mutant-based analyses were used to test whether candidate genes identified by FIND-CIS contribute to drought- or pathogen-related phenotypes when disrupted. Using BonnMu transposon insertion lines (Marcon et al. 2020), loss-of-function alleles were identified, genotyped, and validated for a subset of candidates. While comprehensive phenotypic characterization was not completed for all lines, these experiments establish the feasibility of functionally interrogating FIND-CIS candidates in vivo and highlight key experimental constraints in linking cis-regulatory variation to complex traits.

4.2. Materials and Methods

4.2.1. Plant material.

BonnMu transposon insertion lines targeting candidate genes were derived from the BonnMu population in the B73 genetic background (Marcon et al. 2020). For the candidate genes analyzed in this chapter, segregating F2-family seed stocks were available from prior propagation and crossing efforts within the laboratory. These seed stocks were not genotypically fixed at the time they were received and therefore contained a mixture of wild-type, heterozygous, and insertion-carrying individuals. As a result, all lines were subjected to renewed genotypic characterization prior to downstream analyses to identify plants carrying the predicted Mu insertions and to determine zygosity. No assumptions regarding genotype were made based on seed provenance, and all experimental material used in this study was validated by PCR-based genotyping as described below.

4.2.2. Primer design and sequences.

Insertion lines from the BonnMu population were genotyped using a PCR-based strategy combining gene-specific primers with a Mu transposon-specific primer. Gene-specific primers were designed to flank the predicted Mu insertion sites based on the B73 reference genome annotation, allowing amplification of the wild-type allele. To detect the presence of Mu insertions, the Mu terminal inverted repeat primer MuInt19 (5'GCCTCYATTTCGTCGAATC3') was used in combination with one gene-specific primer. For each candidate gene, primer pairs were selected to amplify regions spanning the annotated insertion site. Successful amplification with gene-specific primer pairs indicated the presence of the wild-type allele, whereas amplification using a gene-specific primer in combination with MuInt19 indicated the presence of a Mu insertion. This strategy allowed discrimination between wild-type, heterozygous, and homozygous insertion genotypes. Primer sequences were designed by previous colleagues and available in the group primer stocks. All primer sequences used in this study are listed in Table 4.1.

Table 4.1: Genotyping primers used for BonnMu insertion lines.

Candidate gene	Mu-seq F2-family stock	Mu Insertion identifier	Insertion position	Primer	Primer sequence (5'-3')

ZmPGM1 (Zm00001eb196320)	BonnMu-6-B-0874	BonnMu0117235	Exon 4	Mutase_mu_2_fw	AATCCACTTGCCCTCAATGTT
				Mutase_mu_2_rev	CTGTGCACTCTCGGTTTGATTA
	BonnMu-5-B-0292	BonnMu0117236	5' UTR	Mutase_mu_3_fw	CCACCTTCGTCACTCGTTACA
				Mutase_mu_3_rev	CTGTCTAAAATGAGTTTGGCACATT
	BonnMu-6-B-0871	BonnMu0117239	5' UTR	Mutase_mu_3_fw	CCACCTTCGTCACTCGTTACA
				Mutase_mu_3_rev	CTGTCTAAAATGAGTTTGGCACATT
	BonnMu-6-B-0766	BonnMu0117238	5' UTR	Mutase_mu_3_fw	CCACCTTCGTCACTCGTTACA
				Mutase_mu_3_rev	CTGTCTAAAATGAGTTTGGCACATT
ZmWRKY68 (Zm00001eb356430)	BonnMu-7-C-0338	BonnMu0217791	Exon 3	WRKY68_mu_fw	GCGGAGTCTGCAAACATGG
				WRKY68_mu_rev	GCCGCGCACTTGGACATA
ZmPGIP2 (Zm00001eb034880)	BonnMu-1-A-0263	BonnMu0020980	Exon	PGIP2_mu_fw	GCATTAACCACACATCTTAATTTTT
				PGIP2_mu_rev	CGTGCCATCATTCCTCGGTA
	BonnMu-1-A-0269	BonnMu0020981	Exon	PGIP2_mu_fw	GCATTAACCACACATCTTAATTTTT
				PGIP2_mu_rev	CGTGCCATCATTCCTCGGTA
	BonnMu-6-B-0857	BonnMu0020982	Exon	PGIP2_mu_fw	GCATTAACCACACATCTTAATTTTT
				PGIP2_mu_rev	CGTGCCATCATTCCTCGGTA
ZmAldolase 1 (Zm00001eb144240)	BonnMu-6-B-0815	BonnMu0086334	5’UTR	Aldolase_mu1_fw	CATGCTGTATCCAACTGCACTC
				Aldolase_mu1_rev	CATTTCAAACTCCG

					ATCTGTGGA
	BonnMu-1-A-0474	BonnMu0086329	5'UTR	Aldolase_mu1_fw	CATGCTGTATCCAACTGCACTC
				Aldolase_mu1_rev	CATTTCAAACTCCGATCTGTGGA
	BonnMu-5-B-0618	BonnMu0086333	5'UTR	Aldolase_mu1_fw	CATGCTGTATCCAACTGCACTC
				Aldolase_mu1_rev	CATTTCAAACTCCGATCTGTGGA
	BonnMu-4-A-1336	BonnMu0086331	5'UTR	Aldolase_mu1_fw	CATGCTGTATCCAACTGCACTC
				Aldolase_mu1_rev	CATTTCAAACTCCGATCTGTGGA
	BonnMu-7-C-0571	BonnMu0086336	Intron 1	Aldolase_mu2_fw	GCTCGTCCTGTTTCCAGTTG
				Aldolase_mu2_rev	GTAGTACTTGGCGCAACGCT
	BonnMu-7-C-0421	BonnMu0086335	Exon 2	Aldolase_mu3_fw	ACATTGATCGCTGCGCTTAC
				Aldolase_mu3_rev	CCAAACAAGTTGCTGTCTCC

4.2.3. Genotyping of BonnMu insertion lines

BonnMu F2-family stocks carrying predicted Mu insertions in candidate genes were obtained from the BonnMu population. Plants were grown in the open field of the Max Planck Institute for Plant Breeding Research, Cologne, Germany. Genomic DNA was extracted from 100 mg of young leaf tissue of two-week-old plants using a standard CTAB-based protocol. PCR reactions were performed using gene-specific primers and the MuInt19 primer with GoTaq Green Master Mix from Promega.

PCR reactions were carried out in a total volume of 10 µl containing 2× GoTaq Green Master Mix (Promega), gene-specific primers with or without MuInt19, genomic DNA, and nuclease-free water. Amplification was performed using an initial denaturation at 95°C for 2 min, followed by 35 cycles of 95°C for 30 s, 60°C for 30 s, and 72°C for 1 min,

with a final extension at 72°C for 5 min. PCR products were analyzed by 1% agarose gel electrophoresis.

For each line, PCR assays were conducted to test for (i) amplification of the wild-type allele using gene-specific primer pairs and (ii) amplification of the insertion allele using MuInt19 in combination with a gene-specific primer. Lines lacking amplification of the wild-type allele but showing Mu-specific amplification were classified as putative homozygous insertion mutants. Lines showing both amplification products were classified as heterozygous. Insertion identifiers, genomic positions, BonnMu stock IDs, and associated primer sequences are summarized in Table 4.1.

4.3. Result: Genotyping of BonnMu insertion lines.

BonnMu F2-family seed stocks corresponding to candidate genes were obtained from existing laboratory material generated in prior experiments. These seed stocks originated from crosses performed prior to the start of this project and lacked definitive genotype annotation for Mu insertion status. Therefore, systematic re-genotyping was required to establish insertion presence and zygosity before further genetic analysis. Genotyping was performed using gene-specific primers in combination with the Mu terminal inverted repeat primer MuInt19. PCR assays were designed to detect both the wild-type allele and the insertion allele at each candidate locus. Amplification with gene-specific primer pairs indicated the presence of the wild-type allele, whereas amplification using MuInt19 in combination with a gene-specific primer indicated the presence of a Mu insertion. Lines showing both amplification products were classified as heterozygous, while lines lacking wild-type amplification but showing Mu-specific amplification were classified as putative homozygous insertion mutants (Figure 4.7, 4.8, and 4.9).

Genotyping and propagation were conducted over two consecutive field seasons at the Max Planck Institute for Plant Breeding Research (MPIPZ). During the 2022 field season, a total of 593 seeds were sown, and 260 individual plants were genotyped. Mu insertions were confirmed for several candidate genes, including *ZmPGM1*, *ZmWRKY68*, and *ZmPGIP2*. Specifically, 11 heterozygous insertion-positive plants were identified for *ZmPGM1*, 1 heterozygous plant for *ZmWRKY68*, and 15 heterozygous plants for *ZmPGIP2*. No homozygous insertion mutants were detected for any of the

tested loci (Table 4.2). To increase the likelihood of recovering homozygous mutants, selfing and backcrossing strategies were initiated based on these genotyping results. During the 2023 field season, 240 seeds were sown from selfed and backcross-derived material, and 172 plants were genotyped. This included progeny from second backcrosses and selfed heterozygous lines for *ZmWRKY68* and *ZmPGIP2*. Genotyping identified 18 heterozygous *ZmWRKY68* plants from selfed progeny, 13 heterozygous *ZmWRKY68* plants from backcross-derived material, and 33 heterozygous *ZmPGIP2* plants from selfed progeny (Table 4.3). As in the previous season, no homozygous insertion mutants were recovered.

Across both field seasons, Mu insertions were reproducibly detected at multiple candidate loci, including WRKY68, PGIP2, ZmALD1, and PGM1, confirming segregation of insertion alleles within the populations. However, genotyping consistently identified heterozygous individuals, and the observed segregation ratios did not follow simple Mendelian expectations (Table 4.3). One possible explanation for this behavior is continued activity of the Mu transposon system, which is known to exhibit somatic and germline mobility in maize. Ongoing transposition could alter insertion presence or zygosity between generations, complicating stable inheritance and recovery of homozygous insertion lines. While this interpretation is consistent with the observed data, transposon activity was not directly assayed in this study.

Despite extensive genotyping, selfing, and backcrossing efforts over two field seasons, stable homozygous insertion mutants could not be obtained for most candidate genes within the timeframe of this project. In addition, for some candidates, such as *ZmTINY*, no suitable BonnMu insertion lines were available, precluding genotyping and downstream functional analysis altogether. Nevertheless, the genotyping results establish the presence and segregation of Mu insertions at multiple candidate loci and document the practical challenges associated with functional interrogation of cis-regulatory candidates using inherited, unlabeled transposon populations. These results provide a framework for future studies using stabilized insertion lines or alternative genetic approaches.

Table 4.2: Genotyping of BonnMu mutants in the summer 2022 field season.

Candidate gene	Mu Insertion identifier	Insertion position	No. of seed sowing	No. of genotyped plants	No. of mutant plants
ZmPGM1	BonnMu0117235	Exon 4	40	4	0
	BonnMu0117236	5' UTR	40	17	0
	BonnMu0117239	5' UTR	40	12	0
	BonnMu0117238	5' UTR	40	17	11
ZMWRKY68	BonnMu0217791	Exon 3	40	8	1
ZmPGIP2	BonnMu0020980	Exon	40	16	0
	BonnMu0020981	Exon	40	21	15
	BonnMu0020982	Exon	40	17	0
ZmALD1	BonnMu0086334	5’UTR	40	22	0
	BonnMu0086329	5’UTR	40	11	0
	BonnMu0086333	5’UTR	40	15	0
	BonnMu0086331	5’UTR	55	45	0
	BonnMu0086336	Intron 1	40	25	0
	BonnMu0086335	Exon 2	58	30	0

Table 4.3: Genotyping of BonnMu mutants in the summer 2023 field season.

Candidate gene	Mu Insertion identifier	Insertion position	Number of mutant plant	Number of WT plant plant	Type of cross
ZmWRKY68	BonnMu0217791	Exon 3	18	49	Hetero selfed
ZmWRKY68	BonnMu0217791	Exon 3	13	48	Backcross
ZmPGIP2	BonnMu0020981	Exon	33	67	Hetero selfed

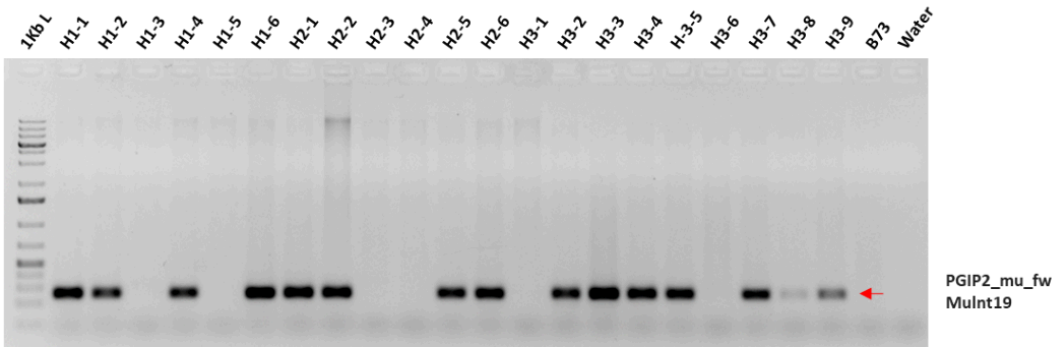


Figure 4.7: PCR-based detection of a Mu transposon insertion in *ZmPGIP2*.

1% agarose gel electrophoresis of PCR products generated using the Mu-TIR primer MuInt19 in combination with a PGIP2-specific forward primer (PGIP2_mu_fw). Lanes correspond to individual plants in the field, with a 1 kb DNA ladder shown on the left. Amplification of a band of the expected size indicates the presence of the Mu1 transposon insertion BonnMu0020981 located in an exon of *ZmPGIP2*. The B73 wild-type control and a no-template (water) control are included on the right. The red arrow highlights a representative Mu-specific amplification product.

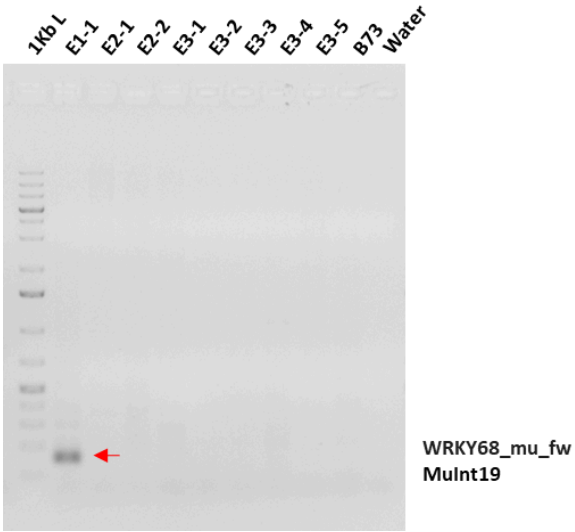


Figure 4.8: PCR-based detection of a Mu transposon insertion in *ZmWRKY68*.

1% agarose gel electrophoresis of PCR products generated using the Mu-TIR primer MuInt19 in combination with a WRKY68-specific forward primer (WRKY68_mu_fw). Lanes correspond to individual plants in the field, with a 1 kb DNA ladder shown on the left. Amplification of a band of the expected size indicates the presence of the Mu1 transposon insertion BonnMu0217791 located in an exon of *ZmWRKY68*. The B73

wild-type control and a no-template (water) control are included on the right. The red arrow highlights a representative Mu-specific amplification product.

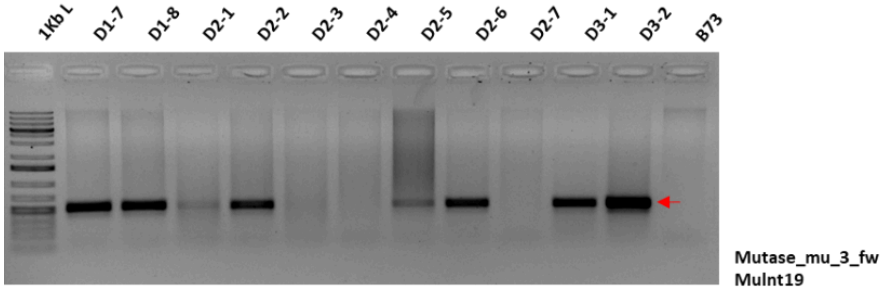


Figure 4.9: PCR-based detection of a Mu transposon insertion in *ZmPGM1*.
1% agarose gel electrophoresis of PCR products generated using the Mu-TIR primer MuInt19 in combination with a PGM1-specific forward primer (Mutase_mu_3_fw). Lanes correspond to individual plants in the field, with a 1 kb DNA ladder shown on the left. Amplification of a band of the expected size indicates the presence of the Mu1 transposon insertion BonnMu0117238 located in the 5'UTR of *ZmPGM1*. The B73 wild-type control and a no-template (water) control are included on the right. The red arrow highlights a representative Mu-specific amplification product.

4.4. Discussion

The aim of this chapter was to explore whether candidate cis-regulatory variants identified by FIND-CIS could be functionally interrogated using available genetic resources in maize. FIND-CIS integrates allele-specific TF occupancy, allele-specific mRNA abundance, and association with quantitative trait loci, thereby providing a strong basis for prioritizing candidate regulatory loci. However, translating these predictions into genetic validation requires appropriate perturbation strategies, which in this study relied on Mu transposon insertion lines from the BonnMu population.

Extensive genotyping across two field seasons confirmed the presence of Mu insertions in several candidate genes, including WRKY68, PGIP2, ZmPGM1, and ZmALD1. In all cases, however, insertion alleles were detected predominantly in the heterozygous state, and stable homozygous insertion mutants could not be obtained within the timeframe of this project. Segregation patterns deviated from simple Mendelian expectations, consistent with previous observations that Mu elements can remain active and continue to transpose, complicating inheritance and fixation of insertion alleles. An additional

constraint was the provenance of the seed material. The BonnMu-derived stocks used in this study originated from crosses performed prior to the start of this project, and detailed genotype information was not available at the time the work began. Consequently, extensive re-genotyping was required to establish the presence and zygosity of Mu transposon insertions before downstream analyses could be pursued. Although multiple rounds of genotyping, selfing, and backcrossing were carried out, these efforts did not result in the recovery of stable homozygous insertion lines suitable for phenotypic analysis within the project timeframe. For some candidate genes, including *ZmTINY*, no corresponding BonnMu insertion lines were available, which precluded genotyping and further functional assessment.

Together, these results highlight a central challenge in regulatory genomics. Genome-wide approaches such as FIND-CIS are effective at identifying candidate regulatory variants associated with phenotypic traits, but functional validation of these predictions often requires experimental tools that are not ideally matched to the underlying regulatory mechanisms. Random transposon insertion populations such as BonnMu are well suited for disrupting gene function, yet they are less effective for testing hypotheses involving noncoding cis-regulatory elements, allele-specific effects, or subtle quantitative changes in transcription. In this chapter, transposon insertion lines were used to assess whether genes identified through FIND-CIS represent functionally relevant components of drought- and pathogen-associated pathways at the gene level, rather than to directly interrogate specific cis-regulatory variants. Although this approach did not allow causal links between individual regulatory elements and phenotypes to be established, it provides an empirical evaluation of the extent to which gene disruption strategies can be used to follow up regulatory predictions. The outcomes underscore both practical and conceptual limitations of gene-level perturbations for validating cis-regulatory hypotheses and point to targeted approaches, such as genome editing of candidate regulatory sequences using CRISPR/Cas-based methods, as more appropriate strategies for directly testing regulatory variation identified by FIND-CIS in maize.

Summary of the thesis

Research Goals and Accomplishments

The central aim of this thesis was to understand how cis-regulatory architecture evolves in plant genomes and how regulatory information is distributed and maintained as genome size, structure, and sequence composition diverge. Rather than focusing on individual regulatory elements in isolation, this work examined transcription factor occupancy at a genome-wide scale and within a conserved biological pathway to assess how regulatory outcomes are preserved over evolutionary time. Plant genomes provide an informative system for addressing this question, as they vary widely in size, ploidy levels, and transposable element content while retaining deeply conserved fundamental pathways. This raises the question of how regulatory information is organized within expanding and structurally diverse genomes and whether conserved pathways rely on fixed cis-regulatory element positions or instead exploit flexible genomic compartments. By combining MOA-seq–based profiling across species with focused analyses of the brassinosteroid biosynthesis pathway, this thesis sought to link regulatory genome organization to evolutionary constraint and functional robustness.

A major component of this thesis was the generation and analysis of a comparative MOA-seq dataset spanning 28 plant species and 29 genomes with wide variation in genome size, ploidy level, and evolutionary history. This dataset provided a unified framework to examine how transcription factor occupancy is distributed across plant genomes that differ substantially in structure, repeat content, and gene density. Across species, larger genomes generally contained higher numbers of MOA-seq peaks, indicating that genome expansion is associated with an increase in regulatory sequence space. However, the relationship between genome size and peak number was variable rather than proportional, and species with quite similar genome sizes often differed substantially in regulatory element abundance. The peak number also showed only a moderate association with gene number, resulting in wide variation in the number of regulatory elements per gene across species. Despite these differences, a consistent organizational feature was observed across genomes. In all species analyzed, including large diploid genomes such as *Vicia faba* and large grass genomes such as *Hordeum vulgare*, a distinct class of promoter-proximal MOA peaks was retained. At the same time, increasing genome size was associated with a progressive shift toward more distal

regulatory elements, reflecting the expansion of intergenic space. Ploidy history further shaped regulatory landscapes. Polyploid species frequently deviated from trends observed in diploids, indicating that whole-genome duplication and subsequent genome restructuring influence the distribution of regulatory elements beyond the effects of genome size alone. In particular, species within the Andropogoneae tribe combined large genome sizes with complex ploidy histories, underscoring their value as systems for studying cis-regulatory evolution. Together, these analyses establish that plant regulatory genomes expand and reorganize as genome size and ploidy change, while maintaining a core association between transcription factor occupancy and gene-proximal regions. This provides a genome-wide context for the pathway-focused analyses presented in chapter 3.

Building on the genome-wide analyses in Chapter 2, this thesis next examined how cis-regulatory architecture is deployed within a deeply conserved biological pathway, using BR biosynthesis as a case study. BR signaling and feedback regulation are maintained across land plants, providing an opportunity to assess how regulatory outcomes are preserved despite substantial divergence in genome size, structure, and intron architecture. By integrating MOA-seq with comparative analyses across maize, *Arabidopsis thaliana*, and *Hordeum vulgare*, Chapter 3 focused on TF occupancy within BR-associated genes, with particular attention to intronic regulatory regions. Intronic MOA-seq signal was detected at the BR biosynthesis gene DWF4 in multiple species, as well as at additional pathway genes, although not uniformly across loci, tissues, or species. The presence of intron-localized occupancy across angiosperms, combined with variation in peak position and number, is consistent with a model in which conserved pathway-level regulation is achieved through flexible use of genomic compartments rather than strict conservation of individual cis-element positions. These findings extend genome-wide observations from Chapter 2 to a pathway-focused context, suggesting that introns represent one of several regulatory substrates that can be utilized and repositioned during evolution while maintaining functional robustness. Perturbation of BR homeostasis using the BR biosynthesis inhibitor propiconazole (PPZ) and PPZ supplemented with epibrassinolide (eBL) further revealed that TF occupancy landscapes respond to altered hormonal conditions at both gene-specific and genome-wide levels. While overall distributions of MOA-seq peaks across genomic features remained broadly similar between treatments, differential peak sets identified by

reciprocal comparisons showed shifts in feature composition, including changes in the relative contribution of intronic regions. These patterns varied between species, highlighting lineage-specific responses within a shared regulatory framework. Together, the results of Chapter 3 illustrate how a fundamental signaling pathway can retain conserved regulatory logic while accommodating evolutionary turnover in cis-regulatory architecture.

The final part of this thesis is to explore whether candidate genes identified by FIND-CIS could be functionally evaluated using available genetic resources. This thesis included a mutant-based analysis using BonnMu transposon insertion lines in maize. Several candidate genes associated with drought tolerance or pathogen resistance were selected based on allele-specific MOA-seq footprints, GWAS proximity, and allele-specific mRNA abundance in hybrids. Genotyping of BonnMu-derived seed stocks confirmed the presence of Mu insertions at multiple candidate loci, including *ZmPGIP2*, *ZmWRKY68*, *ZmPGM1*, and *ZmALD1*. However, across two field seasons, all confirmed insertion carriers were heterozygous, and no stable homozygous insertion lines were obtained within the timeframe of the project. In addition, segregation ratios did not consistently follow Mendelian expectations, likely reflecting ongoing Mu transposon activity and the complex provenance of the seed material. As a result, downstream phenotypic analyses of homozygous loss-of-function mutants could not be completed. Although this project did not yield definitive functional validation of individual candidate genes, it provided a practical assessment of the feasibility and limitations of using random transposon insertion populations to test regulatory predictions derived from genome-wide cis-regulatory analyses.

Challenges and future directions

This thesis highlights both conceptual and practical challenges inherent in studying CREs' evolution at a genome-wide scale and across evolutionary distances. A central challenge arises from the complexity of plant genomes themselves. Large differences in genome size, ploidy history, repeat composition, and assembly quality complicate direct comparisons of regulatory landscapes across species. Although a consistent MOA-seq framework enabled cross-species analyses, variation in genome contiguity and annotation quality, particularly in repeat-rich and polyploid genomes, likely influences

peak detection and positional assignment. Continued improvements in long-read genome assemblies, uniform repeat annotation, and gene models will be essential to refine comparative analyses of regulatory architecture. These advances will facilitate future analyses of how transposable elements intersect with TF occupancy, including their potential roles in expanding regulatory sequence space and shaping lineage-specific binding landscapes. The second challenge concerns the interpretation of TF occupancy data. MOA-seq provides high-resolution maps of protein-DNA interactions in native chromatin, but occupancy alone does not establish regulatory function. Future work integrating MOA-seq with complementary approaches such as transcriptional perturbation, chromatin modification profiling, and reporter assays will be required to distinguish functionally relevant CREs from non-functional or context-dependent binding events. Extending MOA-seq analyses across tissues, developmental stages, and environmental conditions will further clarify how regulatory landscapes change dynamically and how stable regulatory features are maintained.

In the context of conserved pathway regulation, the BR biosynthesis pathway provided a tractable system to examine how regulatory logic is preserved while cis-regulatory architecture varies. However, hormone perturbations and reciprocal peak calling capture associations rather than direct regulatory mechanisms. Future studies combining targeted perturbation of candidate cis-regulatory elements with controlled hormone treatments and transcriptional readouts will be needed to test specific feedback models and to dissect the contribution of intronic versus promoter-proximal regulatory elements.

The genotyping-based functional validation efforts underscored a further limitation: random transposon insertion populations, while powerful for gene disruption, are not very suitable for testing hypotheses centered on subtle cis-regulatory variation, allele-specific effects, or quantitative modulation of gene activity. The difficulty in obtaining stable homozygous insertion lines and the inability to target noncoding regions directly limited phenotypic validation. These challenges point to targeted genome editing approaches, such as CRISPR/Cas-mediated modification of candidate CREs and gene knockout, as a more appropriate strategy for future work aimed at testing regulatory predictions derived from FIND-CIS and MOA-seq.

Looking forward, this thesis establishes a foundation for integrating comparative regulatory genomics with functional experimentation. Combining cross-species occupancy maps, synteny-aware analyses, and targeted cis-regulatory perturbation will allow future studies to directly test how conserved biological functions are maintained despite extensive genome evolution. Such approaches will be particularly powerful in large, repeat-rich crop genomes, where regulatory variation plays a major role in phenotypic diversity and adaptation.

Appendix

5.1. Scripts and analysis pipelines

5.1.1. Adaptor trimming, read merging, and quality enhancement script.

This script was developed by me using the same SeqPurge (adapter trimming) and FLASH (read merging and quality improvement) parameter settings as those implemented in the Hartwig Group's processing pipelines. Original scripts were provided by Amelie Kok.

```
#!/bin/bash
#run by ./Seqpurge_FLASH.sh infile1 infile2 thread
SeqpurgeOutDir FlashOutDir outputFileNamePrefix
#This script combines SeqPurge and FLASH with separate output
directories

EXPECTED_ARGS=6
E_BADARGS=6
if [ $# -ne $EXPECTED_ARGS ]; then
    echo "Usage: ./Seqpurge_FLASH.sh infile1 infile2 thread
SeqpurgeOutDir FlashOutDir outputFileNamePrefix"
    echo "This script combines SeqPurge and FLASH with separate
output directories"
    exit $E_BADARGS
fi

infile1=$1
infile2=$2
threads=$3
SeqpurgeOutDir=$4
FlashOutDir=$5
outputFileNamePrefix=$6

# Properly initialize Conda in a non-interactive script
eval "$($HOME/miniconda3/bin/conda shell.bash hook)"
```

```
# Activate SeqPurge environment and run SeqPurge
echo "Activating Seqpurge environment..."
conda activate ngsbits_env

echo "Trimming adapters for files: ${infile1} and ${infile2}
with output prefix ${outputFileNamePrefix} in directory
${SeqpurgeOutDir}"

if [ ! -e ${SeqpurgeOutDir}/${outputFileNamePrefix}.trim.qc ];
then
  SeqPurge \
  -min_len 30 \
  -threads ${threads} \
  -in1 ${infile1} \
  -in2 ${infile2} \
  -out1 ${SeqpurgeOutDir}/${outputFileNamePrefix}_1.trim.gz \
  -out2 ${SeqpurgeOutDir}/${outputFileNamePrefix}_2.trim.gz \
  -summary ${SeqpurgeOutDir}/${outputFileNamePrefix}.trim.stats
\
  -qc ${SeqpurgeOutDir}/${outputFileNamePrefix}.trim.qc \
  -compression_level 5
fi

if [ $? -ne 0 ]; then
  echo "SeqPurge failed. Check the log for details."
  exit 1
else
  echo "SeqPurge completed successfully. Output is in
${SeqpurgeOutDir}/${outputFileNamePrefix}_1.trim.gz and
${SeqpurgeOutDir}/${outputFileNamePrefix}_2.trim.gz"
fi

# Activate FLASH environment and run FLASH
echo "Activating FLASH environment..."
conda activate flash
```

```
echo "Merging the reads
${SeqpurgeOutDir}/${outputFileNamePrefix}_1.trim.gz and
${SeqpurgeOutDir}/${outputFileNamePrefix}_2.trim.gz"

flash \
${SeqpurgeOutDir}/${outputFileNamePrefix}_1.trim.gz \
${SeqpurgeOutDir}/${outputFileNamePrefix}_2.trim.gz \
-t ${threads} \
--cap-mismatch-quals \
-M 80 \
-z \
-o ${outputFileNamePrefix} \
-d ${FlashOutDir}

if [ $? -ne 0 ]; then
  echo "FLASH failed. Check the log for details."
  exit 1
else
  echo "FLASH completed successfully. Output is in
${FlashOutDir}"
fi
```

5.1.2. MOA-seq read mapping script.

This script was developed using the same STAR parameter settings as those implemented in the Hartwig Group's processing pipelines. At first, a script was provided by Amelie Kok, named "`STAR_align_MOA.sh`" which is located in "`/biodata/dep_psl/grp_frommer/Duong/Documentation/Scripts`". After adapting the parameters from Amelie Kok and Dr. Thomas Hartwig, I generated this script with modified papameter `--limitBAMsortRAM 8978484371`. The `--outFilterMultimapNmax` is changed based on the ploidy level of each species.

```
#!/bin/bash
```

```
#run by ./STAR_align_MOA_1xmap.sh GenomeDir threads
inputfilesInSTARformat outFileNamePrefix
EXPECTED_ARGS=4
E_BADARGS=4
if [ $# -ne $EXPECTED_ARGS ]
then
  echo "Usage: ./STAR_align_MOA_1xmap.sh /STARGenomeDir threads
inputfilesInSTARformat outFileNamePrefix"
  echo "This script runs STAR to align MOA reads."
  exit $E_BADARGS

fi

Genome=$1
threads=$2
infiles=$3
Output=$4

bin=/netscratch/dep_psl/grp_frommer/Thomas/bin

ulimit -n 2000

echo "Aligning ${infiles} to Genome ${Genome}"

${bin}/STAR/source/STAR \
--genomeDir ${Genome} \
--readFilesCommand zcat \
--runThreadN ${threads} \
--readFilesIn ${infiles} \
--outSAMmultNmax 1 \
--outFilterMultimapNmax 1 \
--winAnchorMultimapNmax 100 \
--outMultimapperOrder Random \
--runRNGseed 100 \
--outFileNamePrefix ${Output} \
```

```
--outBAMsortingBinsN 10 \
--limitBAMsortRAM 8978484371 \
--alignIntronMax 1 \
--outSAMtype BAM SortedByCoordinate \
--genomeLoad NoSharedMemory
```

5.1.3. Peak-calling script.

A parameter optimization analysis was performed to evaluate the impact of peak-calling parameters controlling peak width and merging behavior. Specifically, the MACS3 parameters `--min-length` and `--max-gap` were iteratively varied across a range of values derived from the fragment length distribution. For each parameter combination, peak calls were visually compared to the corresponding coverage tracks to assess concordance between called peaks and signal enrichment. Parameter settings were selected based on agreement between peak boundaries and the underlying read coverage.

The final optimised script (`MACS3_maxgap-is-1x-minlength.sh`) differs from the original Hartwig group script in several aspects. First, the fragment length is extracted using the "`average length`" field from the BAM statistics file, instead of the "`first fragment`" estimate used previously . This value is then consistently used to define multiple MACS3 parameters.

Second, the relationship between peak width parameters was modified. In the original script, the maximum gap was set to twice the fragment length (`--max-gap = 2 × fragment length`) . In the optimised version, both the minimum peak length and maximum gap are set equal to the fragment length (`--min-length = fragment length, --max-gap = fragment length`) . This constrains peak merging and enforces a consistent scale between peak extension and merging distance.

Third, the optimized script introduces the use of the fragment length as the smoothing window via the `-s` parameter, which was retained but is now derived from the revised fragment length estimate . In addition, the q-value threshold is provided as an input

parameter, allowing flexible adjustment across runs, whereas the original implementation used a fixed threshold of 0.05.

Other parameters, including the use of the actual genome size, `--nomodel`, `--extsize`, duplicate retention, and buffer size, were retained to maintain comparability with the original workflow.

```
#!/bin/bash
#run by ./MACS3_maxgap-is-1x-minlength.sh bamPathAndFile
sampleName q genomePathAndFasta outputDirectory
EXPECTED_ARGS=5
E_BADARGS=5

if [ $# -ne $EXPECTED_ARGS ]; then
  echo "Usage: ./MACS3_maxgap-is-1x-minlength.sh bamPathAndFile
sampleName q genomePathAndFasta outputDirectory"
  echo "This script runs MACS3 to do peak calling with actual
genome size. Min_length is average fragment, and max_gap is
equal min_length"
  exit $E_BADARGS
fi

bam="$1"
sampleName="$2"
q="$3"
genome="$4"
outDir="$5"

# Loading the environment module system (if required)
source /netscratch/common/Saurabh/envMod/init/bash

# Activate Conda environment
```

```
.
/netscratch/dep_psl/grp_frommer/Thomas/bin/miniconda/etc/profil
e.d/conda.sh
conda activate macs3

echo "BAM file: ${bam}"
echo "Sample name: ${sampleName}"
echo "q-value: ${q}"
echo "Genome: ${genome}"
echo "Output directory: ${outDir}"

## Check if fasta index is present to calculate actual genome
size
if [ ! -f "${genome}.fai" ]; then
    echo "Indexing genome fasta..."
    samtools faidx "${genome}"
fi

## Create the stats file if non-existent
if [ ! -f "${bam}Stats.txt" ]; then
    echo "Generating stats for BAM file..."
    samtools stats -@ 40 "${bam}" > "${bam}Stats.txt"
fi

## Calculate fragment length and maximum gap size
frg_length=$(gawk -v OFS='\t' '{if($2=="average" &&
$3=="length:"){print $4}}' "${bam}Stats.txt")

if [ -z "$frg_length" ]; then
    echo "Error: Could not determine fragment length from stats
file."
    exit 1
fi
```

```
echo "The average fragment length for ${sampleName} is
${frg_length} bp"

## Calculate actual genome size
ags=$(gawk 'BEGIN{len=0} {len=len+$2} END{print len}'
"${genome}.fai")
echo "Actual genome size was measured as ${ags} bp"

## Running the peakcalling
echo "Running MACS3 peak calling..."
macs3 callpeak \
    -t "${bam}" \
    -n "${sampleName}_q${q}_actGS" \
    --outdir "${outDir}" \
    -f BAM \
    -q "${q}" \
    -g "${ags}" \
    -s "${frg_length}" \
    --min-length "${frg_length}" \
    --max-gap "${frg_length}" \
    --nomodel \
    --extsize "${frg_length}" \
    --keep-dup all \
    --buffer-size 100000

echo "Peak calling completed."
```

5.1.4. Helixer gene-annotation script.

A Helixer annotation script provided by Aaryan Bhatia, a PhD student at the Department of Chromosome Biology, Max Planck Institute for Plant Breeding Research, was used as a conceptual starting point. However, the original version contained hard-coded parameters and inflexible settings (e.g., fixed chromosome number and file-handling assumptions). To accommodate different genomes and improve reproducibility, I redeveloped the script by introducing command-line arguments, automated directory

handling, flexible chromosome splitting, and robust error checking, while retaining the overall structure of running Helixer per chromosome and merging the resulting GFF3 files.

```
#!/bin/bash

# This script is used for running Helixer (gene annotation) in
large genomes in a memory-efficient way.
# Usage: ./Helixer.sh genome split_dir result_dir chr_number

# Input arguments
genome="$1"
split_dir="$2"
result_dir="$3"
chr_number="$4"



# Check if the correct number of arguments is given
if [ "$#" -ne 4 ]; then
    echo "Usage: ./Helixer.sh genome split_dir result_dir
chr_number"
    exit 1
fi

# Load necessary modules
module load singularity

# Create directories
mkdir -p "$split_dir" "$result_dir"

# Split genome into chromosomes; all contigs and scaffolds will
be in the debris file
echo "Splitting genome into chromosomes and debris files..."
```

```
declare -i count=0
awk -v chr_num="$chr_number" '/^>/ {count++; if (count <=
chr_num) {file=sprintf("%s/chr_%d.fa", "'$split_dir'", count)}
else {file="'$split_dir'/debris.fa"}} {print > file}' "$genome"

echo "Genome is successfully split."

# Run Helixer on each chromosome
echo "Starting Helixer annotation"

for chr_fasta in "$split_dir"/chr_*.fa; do
    [ -e "$chr_fasta" ] || continue  # Skip if no files exist
    chr_name=$(basename "$chr_fasta" .fa)
    echo "Processing $chr_name"
    singularity run --nv \
        -B "$PWD:$PWD" \
        -B /netscratch:/netscratch \

/opt/share/software/scs/appStore/agnosticApps/annotation/helixe
r/v0.3.3/helixer-docker_helixer_v0.3.3_cuda_11.8.0-cudnn8.sif
Helixer.py \
        --fasta-path "$chr_fasta" \
        --lineage land_plant \
        --gff-output-path
"$result_dir/${chr_name}_helixer.gff3"
done

# Run Helixer on debris sequences
echo "Processing debris sequences"
if [ -f "$split_dir/debris.fa" ]; then
singularity run --nv \
        -B "$PWD:$PWD" \
        -B /netscratch:/netscratch \

/opt/share/software/scs/appStore/agnosticApps/annotation/helixe
```

```
r/v0.3.3/helixer-docker_helixer_v0.3.3_cuda_11.8.0-cudnn8.sif
Helixer.py \
        --fasta-path "$split_dir/debris.fa" \
        --lineage land_plant \
        --gff-output-path "$result_dir/debris_helixer.gff3"
else
    echo "No debris sequences found."
fi

echo "Helixer annotation completed. Merging GFF3 files..."

# Merge GFF3 files
genome_name=$(basename "$genome" .fa)
echo "##gff-version 3" >
"$result_dir/Helixer_${genome_name}_merged.gff3"

cat "$result_dir"/chr_*_helixer.gff3
"$result_dir/debris_helixer.gff3" >>
"$result_dir/Helixer_${genome_name}_merged.gff3"

echo "Annotation is completed for $genome. Result is in
$result_dir"
```

5.1.5. Script for calculating MOA peak to TSS distance.

This script was created by Dr. Julia Engelhorn and Amelie Kok.

```
#!/bin/bash
#run by ./getPeakDistToGenes.bash gffFile narrowPeaksFile
outDirAndPrefix
EXPECTED_ARGS=3
E_BADARGS=3
if [ $# -ne $EXPECTED_ARGS ]
then
```

```
  echo "Usage: ./getPeakDistToGenes.bash gffFile
narrowPeaksFile outDirAndPrefix"
  echo "This script calculates the distance of peaks to genes
given a narrowPeak and a gff file."
  exit $E_BADARGS

fi

gff=$1
peaks=$2
outDirAndPrefix=$3


echo "Input gff: ${gff}"

if [ "${gff##*.}" = "gz" ];then
    gzBase=${gff%.*}
    gffBase=${gzBase%.*}
else
    gffBase=${gff%.*}
fi

echo "Gff base-name is: ${gffBase}"


echo "Get genic regions from gff..."

if [ "${gff##*.}" = "gz" ];then
    zcat ${gff} | awk -v OFS='\t' '{if($3=="gene") {print
$1,$4-1,$5} }' | sortBed | mergeBed >
${gffBase}.only_genic_merged.bed
else
    cat ${gff} | awk -v OFS='\t' '{if($3=="gene") {print
$1,$4-1,$5} }' | sortBed | mergeBed >
${gffBase}.only_genic_merged.bed
```

```
fi

echo "Excluding peaks in contigs that have no gene annotation
in the gff..."
gawk '{if(NR==FNR){contigs[$1]=$1; next} if($1 in
contigs){print $0}}' ${gffBase}.only_genic_merged.bed ${peaks}
> ${outDirAndPrefix}.bed


echo "Excluding peaks within genes..."
bedtools subtract -a ${outDirAndPrefix}.bed -b
${gffBase}.only_genic_merged.bed -A | sortBed >
${outDirAndPrefix}.peaks_no_genes.bed

echo "Calculating the distance of peaks to genes..."
bedtools closest -a ${outDirAndPrefix}.peaks_no_genes.bed -b
${gffBase}.only_genic_merged.bed -d >
${outDirAndPrefix}.peaks_no_genes.dist.bed

echo "Remove duplicate entries..."
awk -v OFS='\t' '{print $1,$2,$3,$(NF)}'
${outDirAndPrefix}.peaks_no_genes.dist.bed | sort | uniq >
${outDirAndPrefix}.peaks_no_genes.dist.noDup.txt

echo "Creating file with only distances..."
# get sample name
sName=$(basename ${outDirAndPrefix})

# print file
awk -v OFS='\t' -v sname=${sName} '{print $4, sname}'
${outDirAndPrefix}.peaks_no_genes.dist.noDup.txt >
${outDirAndPrefix}.peaks_no_genes.dist.noDup.only_dist.txt

echo "--------- DONE ---------"
```

5.1.6. Script for intersecting MOA peaks with genome features.

The feature-overlap workflow used in this study was built upon an initial script provided by Amelie Kok and Dr. Julia Engelhorn. The original script was used for the chapter 3 analyses. Because the original script was written for GFF3 files that were published alongside the genomes. These annotations were performed using various tools, but in Chapter 2, Helixer annotation was used. As a result, for Chapter 2, I revised the workflow and added 5 kb to 10 kb of upstream and downstream features.

- Rewrote script:

```
#!/bin/bash
# Usage: ./getPeakFeatureOverlap_Helixer.sh gffFile
narrowPeakFile outDir name
# Output: BED files for genome features + peak overlaps per
feature, including 5–10 kb upstream/downstream

set -euo pipefail

### ------------------------
### 1. Parse input
### ------------------------
if [ $# -ne 4 ]; then
    echo "Usage: $0 gffFile narrowPeakFile outDir name"
    exit 4
fi

gff=$1
peaks=$2
outDir=$3
name=$4
featDir=${outDir}/${name}_featureBeds

mkdir -p "$featDir"
```

```
cd "$featDir"

### ------------------------
### 2. Choose input opener
### ------------------------
if [[ $gff == *.gz ]]; then
    OPEN="zcat $gff"
else
    OPEN="cat $gff"
fi

echo "Processing GFF: $gff"
echo "Creating feature BED files in: $featDir"

### ------------------------
### 3. Extract features (clean 6-column BED)
### ------------------------

$OPEN | gawk -v OFS="\t" '$3=="five_prime_UTR" {print $1, $4-1,
$5, $9, ".", $7}' \
| bedtools sort -i - | bedtools merge -i - \
> ${name}_5UTR.bed

$OPEN | gawk -v OFS="\t" '$3=="three_prime_UTR" {print $1,
$4-1, $5, $9, ".", $7}' \
| bedtools sort -i - | bedtools merge -i - \
> ${name}_3UTR.bed

$OPEN | gawk -v OFS="\t" '$3=="exon" {print $1, $4-1, $5, $9,
".", $7}' \
| bedtools sort -i - | bedtools merge -i - \
> ${name}_exon.bed

# gene BED (6-column), then merge
```

```
$OPEN | gawk -v OFS="\t" '$3=="gene" {print $1, $4-1, $5, $9,
".", $7}' \
> ${name}_gene_raw.bed

bedtools sort -i ${name}_gene_raw.bed | bedtools merge -i - \
> ${name}_gene.bed

rm -f ${name}_gene_raw.bed

### -------------------------
### 4. Build genic and intronic regions
### -------------------------

cp ${name}_gene.bed ${name}_genic.bed

# Intron = genic − UTR − exon
bedtools subtract -a ${name}_genic.bed -b ${name}_5UTR.bed \
  | bedtools subtract -a - -b ${name}_3UTR.bed \
  | bedtools subtract -a - -b ${name}_exon.bed \
> ${name}_intron.bed

### -------------------------
### 5. Strand-aware flanking regions
### -------------------------

# Upstream 1 kb
$OPEN | gawk -v OFS="\t" '
    $3=="gene" {
        if ($7=="+") { s=($4-1)-1000; if(s<0)s=0; e=($4-1); }
        else         { s=$5; e=$5+1000; }
        if (e>s) print $1,s,e,$9,".",$7
    }' \
| bedtools sort -i - | bedtools merge -i - \
> ${name}_up1kb.bed
```

```
# Downstream 1 kb
$OPEN | gawk -v OFS="\t" '
    $3=="gene" {
        if ($7=="+") { s=$5; e=$5+1000; }
        else         { s=($4-1)-1000; if(s<0)s=0; e=($4-1); }
        if (e>s) print $1,s,e,$9,".",$7
    }' \
| bedtools sort -i - | bedtools merge -i - \
> ${name}_down1kb.bed

# Upstream 1-5 kb
$OPEN | gawk -v OFS="\t" '
    $3=="gene" {
        if ($7=="+") {
            s=($4-1)-5000; if(s<0)s=0;
            e=($4-1)-1000; if(e<0)e=0;
        } else {
            s=$5+1000; e=$5+5000;
        }
        if (e>s) print $1,s,e,$9,".",$7
    }' \
| bedtools sort -i - | bedtools merge -i - \
> ${name}_up1to5kb.bed

# Downstream 1-5 kb
$OPEN | gawk -v OFS="\t" '
    $3=="gene" {
        if ($7=="+") {
            s=$5+1000; e=$5+5000;
        } else {
            s=($4-1)-5000; if(s<0)s=0;
            e=($4-1)-1000; if(e<0)e=0;
        }
        if (e>s) print $1,s,e,$9,".",$7
    }' \
```

```
| bedtools sort -i - | bedtools merge -i - \
> ${name}_down1to5kb.bed

# Upstream 5–10 kb
$OPEN | gawk -v OFS="\t" '
    $3=="gene" {
        if ($7=="+") {
            s=($4-1)-10000; if(s<0)s=0;
            e=($4-1)-5000;  if(e<0)e=0;
        } else {
            s=$5+5000; e=$5+10000;
        }
        if (e>s) print $1,s,e,$9,".",$7
    }' \
| bedtools sort -i - | bedtools merge -i - \
> ${name}_up5to10kb.bed

# Downstream 5–10 kb
$OPEN | gawk -v OFS="\t" '
    $3=="gene" {
        if ($7=="+") {
            s=$5+5000; e=$5+10000;
        } else {
            s=($4-1)-10000; if(s<0)s=0;
            e=($4-1)-5000;  if(e<0)e=0;
        }
        if (e>s) print $1,s,e,$9,".",$7
    }' \
| bedtools sort -i - | bedtools merge -i - \
> ${name}_down5to10kb.bed

### ------------------------
### 6. Intergenic = peaks not in gene ± 10 kb
### ------------------------
```

```
# Build genome sizes from GFF coordinates (robust to
header/name mismatches)
$OPEN | awk -v OFS='\t' '
    $0 !~ /^#/ {
        if(($5+0) > (max[$1]+0)) max[$1]=$5+0
    }
    END { for (chr in max) print chr, max[chr] }
' | sort -k1,1 > genome.txt

bedtools slop -i ${name}_gene.bed -g genome.txt -b 10000 >
${name}_gene10kb.bed

### -------------------------
### 7. Filter peaks to chromosomes that have genes
### -------------------------

bedtools sort -i "$peaks" > peaks.sorted
gawk 'NR==FNR {c[$1]; next} ($1 in c)' ${name}_genic.bed
peaks.sorted \
> ${name}_peaks.filtered.bed

### -------------------------
### 8. Peak intersections
### -------------------------

declare -A FEATS=(
   ["5UTR"]="${name}_5UTR.bed"
   ["3UTR"]="${name}_3UTR.bed"
   ["exon"]="${name}_exon.bed"
   ["intron"]="${name}_intron.bed"
   ["genic"]="${name}_genic.bed"
   ["up1kb"]="${name}_up1kb.bed"
   ["down1kb"]="${name}_down1kb.bed"
   ["up1to5kb"]="${name}_up1to5kb.bed"
   ["down1to5kb"]="${name}_down1to5kb.bed"
```

```
    ["up5to10kb"]="${name}_up5to10kb.bed"
    ["down5to10kb"]="${name}_down5to10kb.bed"
)

base=$(basename "$peaks" .narrowPeak)

for k in "${!FEATS[@]}"; do
    bedtools intersect -sorted \
        -a ${name}_peaks.filtered.bed \
        -b ${FEATS[$k]} \
    > ${base}.${k}.bp.bed
done

# Intergenic peaks (outside gene ± 10 kb)
bedtools subtract -sorted \
    -a ${name}_peaks.filtered.bed \
    -b ${name}_gene10kb.bed \
> ${base}.intergenic.bp.bed

### -------------------------
### 9. Summary
### -------------------------

wc -l ${base}.*.bp.bed > ${name}_results.csv

echo "-------- DONE --------"
echo "Results: ${featDir}/${name}_results.csv"
```

- Original script from Dr. Julia Engelhorn and Amelie Kok:

```
#!/bin/bash
#run by ./getPeaksOvelFeatures.bash gffFile narrowPeaksFile
outDir name
EXPECTED_ARGS=4
```

```
E_BADARGS=4
if [ $# -ne $EXPECTED_ARGS ]
then
  echo "Usage: ./getPeaksOvelFeatures.bash gffFile
narrowPeaksFile outDir name"
  echo "This script determines how many peaks overlap with
different features in the genome."
  exit $E_BADARGS

fi

gff=$1
peaks=$2
outDir=$3
name=$4


if [ ! -d ${outDir} ]
then
    mkdir -p ${outDir}
fi


if [ ! -d ${outDir}/${name}_featureBeds ]
then
    mkdir -p ${outDir}/${name}_featureBeds
fi


cd ${outDir}/${name}_featureBeds

echo "${name}"
echo "Create bed files of genome features..."

if [ "${gff##*.}" = "gz" ];then
```

```
    echo "5'UTR"
    zcat ${gff} | gawk -v OFS='\t' '{if($3=="five_prime_UTR")
{print $1,$4-1,$5,$2,$3,$6,$7,$8,$9} }' | sortBed | mergeBed >
${name}_5P_merged.bed
    echo "3'UTR"
    zcat ${gff} | gawk -v OFS='\t' '{if($3=="three_prime_UTR")
{print $1,$4-1,$5,$2,$3,$6,$7,$8,$9} }' | sortBed | mergeBed >
${name}_3P_merged.bed
    echo "Exons"
    zcat ${gff} | gawk -v OFS='\t' '{if($3=="exon") {print
$1,$4-1,$5,$2,$3,$6,$7,$8,$9} }' | sortBed | mergeBed >
${name}_exon_merged.bed
    echo "Make gene gff"
    zcat ${gff} | gawk -v OFS='\t' '{if($3=="gene") print $0}'
> ${name}_genes.gff
else
    echo "5'UTR"
    cat ${gff} | gawk -v OFS='\t' '{if($3=="five_prime_UTR")
{print $1,$4-1,$5,$2,$3,$6,$7,$8,$9} }' | sortBed | mergeBed >
${name}_5P_merged.bed
    echo "3'UTR"
    cat ${gff} | gawk -v OFS='\t' '{if($3=="three_prime_UTR")
{print $1,$4-1,$5,$2,$3,$6,$7,$8,$9} }' | sortBed | mergeBed >
${name}_3P_merged.bed
    echo "Exons"
    cat ${gff} | gawk -v OFS='\t' '{if($3=="exon") {print
$1,$4-1,$5,$2,$3,$6,$7,$8,$9} }' | sortBed | mergeBed >
${name}_exon_merged.bed
    echo "Make gene gff"
    cat ${gff} | gawk -v OFS='\t' '{if($3=="gene") print $0}' >
${name}_genes.gff
fi

# There is no "intron" annotation. To get those, get the ones
in genes which are not in exons
```

```
echo "Genic regions"
gawk -v OFS='\t' '{if($3=="gene") {print
$1,$4-1,$5,$2,$3,$6,$7,$8,$9} }' ${name}_genes.gff | sortBed |
mergeBed > ${name}_genic_merged.bed

echo "1kb upstream"
gawk -v OFS='\t' '{if($7=="+" && $4>=1001) {print
$1,$4-1001,$4-1,$2,$3,$6,$7,$8,$9} else {if($7=="+" &&
$4<1001){print $1,0,$4-1,$2,$3,$6,$7,$8,$9} else {print
$1,$5,$5+1000,$2,$3,$6,$7,$8,$9}}}' ${name}_genes.gff | sortBed
| mergeBed > ${name}_1kb_up_merged.bed


echo "1-5kb upstream"
gawk -v OFS='\t' '{if($7=="+" && $4>=5001) {print
$1,$4-5001,$4-1001,$2,$3,$6,$7,$8,$9} else {if($7=="+" &&
$4<5001 && $4>1001){print $1,0,$4-1001,$2,$3,$6,$7,$8,$9}
else{if($7=="+" && $4<=1001){} else {print
$1,$5+1000,$5+5000,$2,$3,$6,$7,$8,$9}}}}' ${name}_genes.gff |
sortBed | mergeBed > ${name}_1kb_to_5kb_up_merged.bed


echo "1kb downstream"
gawk -v OFS='\t' '{if($7=="-" && $4>=1001) {print
$1,$4-1001,$4-1,$2,$3,$6,$7,$8,$9} else {if($7=="-" &&
$4<1001){print $1,0,$4-1,$2,$3,$6,$7,$8,$9} else {print
$1,$5,$5+1000,$2,$3,$6,$7,$8,$9}}}' ${name}_genes.gff | sortBed
| mergeBed > ${name}_1kb_down_merged.bed


echo "1-5kb upstream"
gawk -v OFS='\t' '{if($7=="-" && $4>=5001) {print
$1,$4-5001,$4-1001,$2,$3,$6,$7,$8,$9} else {if($7=="-" &&
$4<5001 && $4>1001){print $1,0,$4-1001,$2,$3,$6,$7,$8,$9}
else{if($7=="-"n && $4<=1001){} else {print
```

```
$1,$5+1000,$5+5000,$2,$3,$6,$7,$8,$9}}}}' ${name}_genes.gff |
sortBed | mergeBed > ${name}_1kb_to_5kb_down_merged.bed


## To get intergenic make file with all genes plus 5kb upstream
and 5 kb downstream (to be subtracted)
echo "all genes plus 5kb upstream and 5 kb downstream"
gawk -v OFS='\t' '{if($4>=5001) {print
$1,$4-5001,$5+5000,$2,$3,$6,$7,$8,$9} else {print
$1,0,$5+5000,$2,$3,$6,$7,$8,$9}}' ${name}_genes.gff | sortBed |
mergeBed > ${name}_genes_5kb_up_5kb_down_merged.bed



echo "Account for possible scaffold entries that have 0 0"
for i in 5P 3P exon genic 1kb_up 1kb_to_5kb_up 1kb_down
1kb_to_5kb_down genes_5kb_up_5kb_down; do echo "${i}"; gawk -v
OFS='\t' '{if($2==(-1)){print $1, 0, $3} else {print $0}}'
${name}_${i}_merged.bed > ${name}_${i}_merged_noNeg1.bed; done



# Priority
# 5'UTR
# 3'UTR
# Exons
# Intron
# 1kbup
# 1kbdown
# 1-5kb up
# 1-5kb down
# intergenic
```

```
### Have to add -sorted option for very large chromosomes

# To enable this, substract the higher priorities from lower

echo "Subtract according to priority..."

echo "3P"
bedtools subtract -sorted -a ${name}_3P_merged_noNeg1.bed -b
${name}_5P_merged_noNeg1.bed > ${name}_3P_merged_sub.bed

echo "exon"
bedtools subtract -sorted -a ${name}_exon_merged_noNeg1.bed -b
${name}_5P_merged_noNeg1.bed | bedtools subtract -sorted -a -
-b ${name}_3P_merged_noNeg1.bed > ${name}_exon_merged_sub.bed

echo "intron"
bedtools subtract -sorted -a ${name}_genic_merged_noNeg1.bed -b
${name}_5P_merged_noNeg1.bed | bedtools subtract -sorted -a -
-b ${name}_3P_merged_noNeg1.bed | bedtools subtract -sorted -a
- -b ${name}_exon_merged_noNeg1.bed >
${name}_intron_merged_sub.bed

echo "1kp up"
bedtools subtract -sorted -a ${name}_1kb_up_merged_noNeg1.bed
-b ${name}_genic_merged_noNeg1.bed >
${name}_1kb_up_merged_sub.bed

echo "1kb down"
bedtools subtract -sorted -a ${name}_1kb_down_merged_noNeg1.bed
-b ${name}_genic_merged_noNeg1.bed | bedtools subtract -sorted
-a - -b ${name}_1kb_up_merged_noNeg1.bed >
${name}_1kb_down_merged_sub.bed

echo "1kb to 5kb up"
```

```
bedtools subtract -sorted -a
${name}_1kb_to_5kb_up_merged_noNeg1.bed -b
${name}_genic_merged_noNeg1.bed | bedtools subtract -sorted -a
- -b ${name}_1kb_up_merged_noNeg1.bed | bedtools subtract
-sorted -a - -b ${name}_1kb_down_merged_noNeg1.bed >
${name}_1kb_to_5kb_up_merged_sub.bed

echo "1kb to 5kb down"
bedtools subtract -sorted -a
${name}_1kb_to_5kb_down_merged_noNeg1.bed -b
${name}_genic_merged_noNeg1.bed | bedtools subtract -sorted -a
- -b ${name}_1kb_up_merged_noNeg1.bed | bedtools subtract
-sorted -a - -b ${name}_1kb_down_merged_noNeg1.bed | bedtools
subtract -sorted -a - -b
${name}_1kb_to_5kb_up_merged_noNeg1.bed >
${name}_1kb_to_5kb_down_merged_sub.bed




echo "Get overlapping peaks..."

cd ${outDir}

echo "Excluding peaks in contigs that have no gene annotation
in the gff..."
gawk '{if(NR==FNR){contigs[$1]=$1; next} if($1 in
contigs){print $0}}'
${outDir}/${name}_featureBeds/${name}_genic_merged_noNeg1.bed
${peaks} > ${name}_peaks.bed

for i in 5P genic; do echo "${i}"; intersectBed -sorted -a
${name}_peaks.bed -b
${outDir}/${name}_featureBeds/${name}_${i}_merged_noNeg1.bed  >
`basename ${peaks} .narrowPeak`.${i}.bp.bed; done
```

```
for i in 3P exon intron 1kb_up 1kb_to_5kb_up 1kb_down
1kb_to_5kb_down; do echo "${i}"; intersectBed -sorted -a
${name}_peaks.bed -b
${outDir}/${name}_featureBeds/${name}_${i}_merged_sub.bed >
`basename ${peaks} .narrowPeak`.${i}.bp.bed; done

echo "intergenic"
bedtools subtract -sorted -a ${name}_peaks.bed -b
${outDir}/${name}_featureBeds/${name}_genes_5kb_up_5kb_down_mer
ged_noNeg1.bed > `basename ${peaks}
.narrowPeak`.intergenic.bp.bed


wc -l *.bp.bed > ${name}_results.csv



echo "--------- DONE ---------"
```

5.1.7. Script for calculating genome-feature coverage percentages.

The script used to extract the genome feature size in this study was written by me for this thesis. The workflow, implemented in Bash using BEDTools (v2.30.0) and gawk, includes automatic extraction, sorting, merging, and length quantification of the major genomic feature categories: UTRs, exons, introns, upstream/downstream regions, and intergenic space. The script was designed to provide strand-aware definitions, consistent BED formatting, and reproducible integration with genome index (.fai) files.

```
#!/bin/bash
# Usage: ./genome_feature_sizes_full.sh Species GFF_file
FAI_file outDir

set -euo pipefail

if [ $# -ne 4 ]; then
    echo "Usage: $0 Species GFF_file FAI_file outDir"
```

```
    exit 1
fi

species=$1
gff=$2
fai=$3
outDir=$4

mkdir -p "$outDir"
cd "$outDir"

featureDir="${species}_featureBeds"
mkdir -p "$featureDir"

echo "Processing $species..."

# Choose correct opener
if [[ $gff == *.gz ]]; then
    OPEN="zcat $gff"
else
    OPEN="cat $gff"
fi

###############################################
# 1. Extract base features (raw BED)
###############################################

$OPEN | gawk -v OFS="\t" '$3=="five_prime_UTR"  {print
$1,$4-1,$5,$7}' > "$featureDir/5UTR.raw.bed"
$OPEN | gawk -v OFS="\t" '$3=="three_prime_UTR" {print
$1,$4-1,$5,$7}' > "$featureDir/3UTR.raw.bed"
$OPEN | gawk -v OFS="\t" '$3=="exon"            {print
$1,$4-1,$5,$7}' > "$featureDir/exon.raw.bed"
$OPEN | gawk -v OFS="\t" '$3=="gene"            {print
$1,$4-1,$5,$7}' > "$featureDir/gene.raw.bed"

###############################################
# 2. Sort + merge base feature files
###############################################

for f in "$featureDir"/*.raw.bed; do
    out="${f%.raw.bed}.bed"
    bedtools sort -faidx "$fai" -i "$f" | bedtools merge -i - >
"$out"
    rm "$f"
```

```
done

################################################
# 3. Build GENIC
################################################
cp "$featureDir/gene.bed" "$featureDir/genic.bed"

################################################
# 4. Build INTRON = genic - UTR - exon
################################################

bedtools subtract -a "$featureDir/genic.bed" -b
"$featureDir/5UTR.bed" \
  | bedtools subtract -a - -b "$featureDir/3UTR.bed" \
  | bedtools subtract -a - -b "$featureDir/exon.bed" \
  | bedtools sort -faidx "$fai" -i - \
  > "$featureDir/intron.bed"

################################################
# 5. Strand-aware upstream/downstream 1 kb
################################################

# Upstream 1 kb
$OPEN | gawk -v OFS="\t" '
    $3=="gene" {
        if ($7=="+") {
            s = ($4-1)-1000; if (s<0) s=0;
            e = $4-1;
        } else {
            s = $5; e = $5+1000;
        }
        if (e>s) print $1,s,e
    }
' | bedtools sort -faidx "$fai" -i - | bedtools merge -i - \
> "$featureDir/up1kb.bed"

# Downstream 1 kb
$OPEN | gawk -v OFS="\t" '
    $3=="gene" {
        if ($7=="+") {
            s = $5; e = $5+1000;
        } else {
            s = ($4-1)-1000; if (s<0) s=0;
            e = $4-1;
        }
```

```
        if (e>s) print $1,s,e
    }
' | bedtools sort -faidx "$fai" -i - | bedtools merge -i - \
> "$featureDir/down1kb.bed"

################################################
# 6. Upstream/downstream 1–5 kb (strand-aware)
################################################

$OPEN | gawk -v OFS="\t" '
    $3=="gene" {
        if ($7=="+") {
            s = ($4-1)-5000; if (s<0) s=0;
            e = ($4-1)-1000; if (e<0) e=0;
        } else {
            s = $5+1000; e = $5+5000;
        }
        if (e>s) print $1,s,e
    }
' | bedtools sort -faidx "$fai" -i - | bedtools merge -i - \
> "$featureDir/up1to5kb.bed"

$OPEN | gawk -v OFS="\t" '
    $3=="gene" {
        if ($7=="+") {
            s = $5+1000; e = $5+5000;
        } else {
            s = ($4-1)-5000; if (s<0) s=0;
            e = ($4-1)-1000; if (e<0) e=0;
        }
        if (e>s) print $1,s,e
    }
' | bedtools sort -faidx "$fai" -i - | bedtools merge -i - \
> "$featureDir/down1to5kb.bed"

################################################
# 7. Intergenic = genome − genic
################################################

bedtools complement \
    -i "$featureDir/genic.bed" \
    -g "$fai" \
    | bedtools sort -faidx "$fai" -i - \
    > "$featureDir/intergenic.bed"
```

```
##############################################
# 8. Output base counts
##############################################

outFile="${outDir}/genome_feature_sizes.tsv"

if [ ! -f "$outFile" ]; then
    echo -e "Species\tFeature\tBases" > "$outFile"
fi

for bed in "$featureDir"/*.bed; do
    feature=$(basename "$bed" .bed)
    bases=$(awk '{sum += $3-$2} END{print sum}' "$bed")
    echo -e "$species\t$feature\t$bases" >> "$outFile"
done

echo "Done for $species!"
```

5.1.8. Computational pipeline for chloroplast-based phylogenetic reconstruction

For this study, I developed a custom computational pipeline to reconstruct the phylogenetic tree from chloroplast gene sequences.

5.1.8.1. Sequence retrieval

Chloroplast genomes were downloaded from the NCBI database using the accession numbers listed in the Methods section. Each genome was stored in a text file containing gene annotations and nucleotide sequences.

5.1.8.2. Identification of chloroplast genes shared across all species

Extract gene names from annotated FASTA files.

```
for file in *sequence.txt; do
  species=$(basename "$file" -sequence.txt)
  grep "^>" "$file" | sed -n 's/.*\[gene=\([^]]*\)\].*/\1/p' |
sort -u > "${species}_genes.txt"
done
```

Identify common genes present in all species. A total of 56 genes were identified as shared across all 28 species.

```
cat *_genes.txt | sort | uniq -c | awk -v n=$(ls *_genes.txt |
wc -l) '$1 == n {print $2}' > common_genes.txt
```

5.1.8.3. Extraction of gene sequences and header normalization

Extract nucleotide sequences for each shared gene.

```
while read gene; do
  for file in *chloroplast-sequence.txt; do
    species=$(basename "$file" -chloroplast-sequence.txt)
    awk -v g="$gene" '/^>/ {p=($0 ~ "\\[gene=" g "\\]")} p'
"$file" >> "${species}_${gene}.fasta"
  done
done < common_genes.txt
```

Standardize FASTA headers to include species name, gene name, and a unique protein identifier.

```
for f in *_*.fasta; do
  species=$(echo $f | cut -d'_' -f1)
  gene=$(echo $f | cut -d'_' -f2 | sed 's/.fasta//')
  count=0
  awk -v sp="$species" -v g="$gene" -v c="$count" '
    /^>/ {
      match($0, /\[protein_id=([^]]+)\]/, a)
      pid = (a[1] != "" ? a[1] : "copy"++c)
      print ">"sp"_"g"_"pid
      next
    }
    {print}
  ' $f > tmp && mv tmp $f
done
```

Concatenate sequences for alignment.

```
for gene in $(cat common_genes.txt); do
  cat *_${gene}.fasta > ${gene}_all.fasta
done
```

5.1.8.4. Multiple sequence alignment

Align each gene using MAFFT (v7.505)

```
for file in *_all.fasta; do
  mafft --auto "$file" > "${file%.fasta}_aligned.fasta" &
done
wait
```

5.1.8.5. Trimming poorly aligned regions

Trim alignments with TrimAl (v1.4.rev22).

```
for f in *_aligned.fasta; do
  trimal -in "$f" -out "${f%.fasta}_trimmed.fasta" -automated1
done
```

5.1.8.5. Gene tree construction using IQ-TREE2

Maximum-likelihood gene trees were constructed with IQ-TREE2 (v2.4.0)

```
mkdir -p gene_tree_IQ-tree2

for aln in *_trimmed.fasta; do
  base=$(basename "$aln" _trimmed.fasta)
  iqtree2 -s "$aln" -m GTR+G -nt AUTO -quiet -pre
gene_tree_IQ-tree2/"$base"
done
```

5.1.8.6. Species tree inference with ASTRAL-IV

Concatenate tree files.

```
cat *_trimmed.treefile > all_genes.tre
```

Create a mapping file linking sequence IDs to species names.

```
cat all_genes.tre | tr '(),:;' '\n' | grep -vE '^[0-9.eE+-]+$'
| grep -v '^$' \
| sort -u | while read tip; do
    species=$(echo "$tip" | cut -d'_' -f1)
    echo -e "$tip\t$species"
done > map.txt
```

Run ASTRAL-IV.

```
astral4 -i all_genes.tre -a map.txt -o
28species_tree_with_name.tre -t 80
```

5.1.8.7. Divergence time estimation with TreePL

Root the species tree using *Physcomitrium patens.*

```
nw_reroot 28species_tree_with_name.tre moss >
rooted_28species_tree_with_name.tre
```

The total number of sites in the concatenated trimmed alignment (`numsites`) was calculated using the following command.

```
grep -v "^>" supermatrix_trimmed.fasta | tr -d '\n' | wc -c
```

TreePL configuration (treepl.config):

```
treefile = rooted_28species_tree_with_name.tre
outfile = 28species_with_name_MYA.tre
smooth = 100
numsites = 79587840
mrca = EMBRYOPHYTES moss maize arabidopsis
min = EMBRYOPHYTES 470
```

```
max = EMBRYOPHYTES 515
thorough
prime
```

Run priming:

```
treepl treepl.config
```

Optimized configuration (treepl_optimized.config)

```
treefile = rooted_28species_tree_with_name.tre
outfile = 28species_with_name_MYA.tre
smooth = 100
numsites = 79587840
mrca = EMBRYOPHYTES moss maize arabidopsis
min = EMBRYOPHYTES 470
max = EMBRYOPHYTES 515
thorough
opt = 1
moredetail
optad = 4
moredetailad
optcvad = 2
```

Run final dating:

treepl treepl_optimized.config

5.1.8.8. Tree visualization in R

```
library(ape)
library(phytools)
library(ggtree)
library(ggplot2)
```

```
tree <- read.tree("28species_with_name_MYA.tre")

tree_mid <- midpoint.root(tree)
is.ultrametric(tree_mid)

p <- ggtree(tree_mid) + theme_tree2() + geom_tiplab()
print(p)
```

Final aesthetic modifications were performed in Inkscape (v1.4.2)

5.1.9. Genomic feature overlap analysis for Chapter 3.

This analysis was performed by Dr. Thomas Hartwig, Dr. Julia Engelhorn (DIADE, University of Montpellier, CIRAD, IRD, Montpellier, France), and Amelie Kok (PhD student).

5.1.9.1. Analysis for *Hordeum vulgare* samples

Copy peak files into working directory

```
/netscratch/dep_psl/grp_frommer/Thomas/Results/MOA_barley/BR/MA
CS3/pies

cp
/netscratch/dep_psl/grp_frommer/Thomas/Results/MOA_barley/BR/MA
CS3/Morex_BL_R1-3vsMorex_PPZ_R1-3/Morex_BL_R1-3.vs.Morex_PPZ_R1
-3.q0.05_peaks.narrowPeak ./Morex_BL_vs_ppz_for_pie.bed

cp
/netscratch/dep_psl/grp_frommer/Thomas/Results/MOA_barley/BR/MA
CS3/Morex_BL_R1-3vsMorex_PPZ_R1-3/Morex_PPZ_R1-3.vs.Morex_BL_R1
-3.q0.05_peaks.narrowPeak ./Morex_ppz_vs_BL_for_pie.bed

cp
/netscratch/dep_psl/grp_frommer/Duong/results/MOA-seq/brassinos
```

```
teroid/Barley-morex/MACS3/Morex_PPZ_R1-3_1x_q0.02_actGS_peaks.n
arrowPeak ./Morex_ppz_for_pie.bed

cp
/netscratch/dep_psl/grp_frommer/Duong/results/MOA-seq/brassinos
teroid/Barley-morex/MACS3/Morex_BL_R1-3_1x_q0.02_actGS_peaks.na
rrowPeak ./Morex_BL_for_pie.bed
```

Decompress GFF3 annotation

```
zcat
/netscratch/dep_psl/grp_frommer/Michael_Thomas/Genomes/Hor_vul/
morex/v3/Hordeum_vulgare.MorexV3_pseudomolecules_assembly.54.gf
f3.gz >
Hordeum_vulgare.MorexV3_pseudomolecules_assembly.54.gff3
```

Run overlap script for each peak set

```
/biodata/dep_psl/grp_frommer/Julia/scripts/getPeaksOvelFeatures
_bp.bash
#run by ./getPeaksOvelFeatures.bash gffFile narrowPeaksFile
outDir name

/biodata/dep_psl/grp_frommer/Julia/scripts/getPeaksOvelFeatures
_bp.bash
/netscratch/dep_psl/grp_frommer/Thomas/Results/MOA_barley/BR/MA
CS3/pies/Hordeum_vulgare.MorexV3_pseudomolecules_assembly.54.gf
f3                                          Morex_BL_for_pie.bed
/netscratch/dep_psl/grp_frommer/Thomas/Results/MOA_barley/BR/MA
CS3/pies BL

/biodata/dep_psl/grp_frommer/Julia/scripts/getPeaksOvelFeatures
_bp.bash
/netscratch/dep_psl/grp_frommer/Thomas/Results/MOA_barley/BR/MA
CS3/pies/Hordeum_vulgare.MorexV3_pseudomolecules_assembly.54.gf
f3                                        Morex_ppz_for_pie.bed
```

```
/netscratch/dep_psl/grp_frommer/Thomas/Results/MOA_barley/BR/MA
CS3/pies ppz

/biodata/dep_psl/grp_frommer/Julia/scripts/getPeaksOvelFeatures
_bp.bash
/netscratch/dep_psl/grp_frommer/Thomas/Results/MOA_barley/BR/MA
CS3/pies/Hordeum_vulgare.MorexV3_pseudomolecules_assembly.54.gf
f3                                  Morex_BL_vs_ppz_for_pie.bed
/netscratch/dep_psl/grp_frommer/Thomas/Results/MOA_barley/BR/MA
CS3/pies BL_vs_ppz

/biodata/dep_psl/grp_frommer/Julia/scripts/getPeaksOvelFeatures
_bp.bash
/netscratch/dep_psl/grp_frommer/Thomas/Results/MOA_barley/BR/MA
CS3/pies/Hordeum_vulgare.MorexV3_pseudomolecules_assembly.54.gf
f3                                  Morex_ppz_vs_BL_for_pie.bed
/netscratch/dep_psl/grp_frommer/Thomas/Results/MOA_barley/BR/MA
CS3/pies ppz_vs_BL
```

Whole-genome feature bp counting from intermediate feature beds

```
/netscratch/dep_psl/grp_frommer/Thomas/Results/MOA_barley/BR/MA
CS3/pies/BL_featureBeds

for i in *sub.bed; do gawk -v OFS='\t' -v i=${i}
'BEGIN{a=0}{a=a+($3-$2)}END{print i, a}' $i; done

BL_1kb_down_merged_sub.bed      31587026
BL_1kb_to_5kb_down_merged_sub.bed       100094904
BL_1kb_to_5kb_up_merged_sub.bed 121368901
BL_1kb_up_merged_sub.bed        34347180
BL_3P_merged_sub.bed    5129838
BL_exon_merged_sub.bed  42132427
BL_intron_merged_sub.bed        65300430
```

5' UTR does not have a subtracted version:

```
for i in BL_5P_merged_noNeg1.bed; do gawk -v OFS='\t' -v i=${i}
'BEGIN{a=0}{a=a+($3-$2)}END{print i, a}' $i; done

BL_5P_merged_noNeg1.bed 2502174
```

To get inter genic regions, /netscratch/dep_psl/grp_frommer/Julia/Barley_annotation/feature_bed_files/MorexV3_genes_5kb_up_5kb_down_merged.bed needs to be subtracted from the whole genome size or better the effective genome size

```
for i in BL_genes_5kb_up_5kb_down_merged_noNeg1.bed ; do gawk
-v OFS='\t' -v i=${i} 'BEGIN{a=0}{a=a+($3-$2)}END{print i, a}'
$i; done

BL_genes_5kb_up_5kb_down_merged_noNeg1.bed      402462880
```

5.1.9.2. Analysis for *Arabidopsis thaliana* samples

Copy peak files into working directory /netscratch/dep_psl/grp_frommer/Thomas/Results/MOA_Ara_BR/features

```
cp
/netscratch/dep_psl/grp_frommer/Duong/results/MOA-seq/brassinosteroid/Arabidopsis-Col0/MACS3/Arabidopsis_BL_R1-4_1x_q0.02_actGS_peaks.narrowPeak ./Ara_Col0_BL_for_pie.bed

cp
/netscratch/dep_psl/grp_frommer/Duong/results/MOA-seq/brassinosteroid/Arabidopsis-Col0/MACS3/Arabidopsis_PPZ_R1-4_1x_q0.02_actGS_peaks.narrowPeak ./Ara_Col0_ppz_for_pie.bed
```

```
cp
/netscratch/dep_psl/grp_frommer/Thomas/Results/MOA_Ara_BR/MACS3
/Ara_Col0_BL_R1-4vsAra_Col0_PPZ_R1-4_small_lambda/Ara_Col0_BL_R
1-4.vs.Ara_Col0_PPZ_R1-4.q0.00001_peaks_small_lambda.narrowPeak
./Ara_Col0_BLvsppz_0.00001_small_lambda.bed

cp
/netscratch/dep_psl/grp_frommer/Thomas/Results/MOA_Ara_BR/MACS3
/Ara_Col0_PPZ_R1-4vsAra_Col0_BL_R1-4_small_lambda/Ara_Col0_PPZ_
R1-4.vs.Ara_Col0_BL_R1-4.q0.00001_peaks_small_lambda.narrowPeak
./Ara_Col0_ppzvsBL_0.00001_small_lambda.bed
```

Run overlap script for each peak set

```
/biodata/dep_psl/grp_frommer/Julia/scripts/getPeaksOvelFeatures
_bp.bash
#run by ./getPeaksOvelFeatures.bash gffFile narrowPeaksFile
outDir name

/netscratch/dep_psl/grp_frommer/Michael_Thomas/Genomes/Athalian
a/TAIR10/TAIR10_GFF3_genes.gff

/biodata/dep_psl/grp_frommer/Julia/scripts/getPeaksOvelFeatures
_bp.bash
/netscratch/dep_psl/grp_frommer/Michael_Thomas/Genomes/Athalian
a/TAIR10/TAIR10_GFF3_genes.gff Ara_Col0_BL_for_pie.bed
/netscratch/dep_psl/grp_frommer/Thomas/Results/MOA_Ara_BR/featu
res BL

/biodata/dep_psl/grp_frommer/Julia/scripts/getPeaksOvelFeatures
_bp.bash
/netscratch/dep_psl/grp_frommer/Michael_Thomas/Genomes/Athalian
a/TAIR10/TAIR10_GFF3_genes.gff Ara_Col0_ppz_for_pie.bed
/netscratch/dep_psl/grp_frommer/Thomas/Results/MOA_Ara_BR/featu
res ppz
```

```
/biodata/dep_psl/grp_frommer/Julia/scripts/getPeaksOvelFeatures
_bp.bash
/netscratch/dep_psl/grp_frommer/Michael_Thomas/Genomes/Athalian
a/TAIR10/TAIR10_GFF3_genes.gff
Ara_Col0_BLvsppz_0.00001_small_lambda.bed
/netscratch/dep_psl/grp_frommer/Thomas/Results/MOA_Ara_BR/featu
res BLvsppz_SL_q5

/biodata/dep_psl/grp_frommer/Julia/scripts/getPeaksOvelFeatures
_bp.bash
/netscratch/dep_psl/grp_frommer/Michael_Thomas/Genomes/Athalian
a/TAIR10/TAIR10_GFF3_genes.gff
Ara_Col0_ppzvsBL_0.00001_small_lambda.bed
/netscratch/dep_psl/grp_frommer/Thomas/Results/MOA_Ara_BR/featu
res ppzvsBL_SL_q5
```

Whole-genome feature bp counting from intermediate feature beds

```
/netscratch/dep_psl/grp_frommer/Thomas/Results/MOA_Ara_BR/featu
res/BL_featureBeds

for i in *sub.bed; do gawk -v OFS='\t' -v i=${i}
'BEGIN{a=0}{a=a+($3-$2)}END{print i, a}' $i; done


BL_1kb_down_merged_sub.bed      8910449
BL_1kb_to_5kb_down_merged_sub.bed       3894313
BL_1kb_to_5kb_up_merged_sub.bed 14621337
BL_1kb_up_merged_sub.bed        18999162
BL_3P_merged_sub.bed    4920112
BL_exon_merged_sub.bed  42225800
BL_intron_merged_sub.bed        19051932
```

```
5' UTR does not have a subtracted version:

for i in BL_5P_merged_noNeg1.bed; do gawk -v OFS='\t' -v i=${i}
'BEGIN{a=0}{a=a+($3-$2)}END{print i, a}' $i; done


BL_5P_merged_noNeg1.bed 3149194


To get inter genic regions,
/netscratch/dep_psl/grp_frommer/Julia/Barley_annotation/feature
_bed_files/MorexV3_genes_5kb_up_5kb_down_merged.bed needs to be
subtracted from the whole genome size or better the effective
genome size

for i in BL_genes_5kb_up_5kb_down_merged_noNeg1.bed ; do gawk
-v OFS='\t' -v i=${i} 'BEGIN{a=0}{a=a+($3-$2)}END{print i, a}'
$i; done

BL_genes_5kb_up_5kb_down_merged_noNeg1.bed      106765443

EG: 119667750
```

5.1.9.3. Script for intersecting MOA peaks with genome features for two treatment conditions.

This script for intersecting MOA peaks with genome features for two treatment conditions was developed by Amelie Kok and Dr. Julia Engelhorn.

```
#!/bin/bash
#run by ./getPeaksOvelFeatures.bash gffFile narrowPeaksFile
outDir name
EXPECTED_ARGS=4
E_BADARGS=4
if [ $# -ne $EXPECTED_ARGS ]
```

```
then
  echo "Usage: ./getPeaksOvelFeatures.bash gffFile
narrowPeaksFile outDir name"
  echo "This script determines how many peaks overlap with
different features in the genome."
  exit $E_BADARGS

fi

gff=$1
peaks=$2
outDir=$3
name=$4


if [ ! -d ${outDir} ]
then
    mkdir -p ${outDir}
fi


if [ ! -d ${outDir}/${name}_featureBeds ]
then
    mkdir -p ${outDir}/${name}_featureBeds
fi


cd ${outDir}/${name}_featureBeds

echo "${name}"
echo "Create bed files of genome features..."

if [ "${gff##*.}" = "gz" ];then
    echo "5'UTR"
```

```
    zcat ${gff} | gawk -v OFS='\t' '{if($3=="five_prime_UTR")
{print $1,$4-1,$5,$2,$3,$6,$7,$8,$9} }' | sortBed | mergeBed >
${name}_5P_merged.bed
    echo "3'UTR"
    zcat ${gff} | gawk -v OFS='\t' '{if($3=="three_prime_UTR")
{print $1,$4-1,$5,$2,$3,$6,$7,$8,$9} }' | sortBed | mergeBed >
${name}_3P_merged.bed
    echo "Exons"
    zcat ${gff} | gawk -v OFS='\t' '{if($3=="exon") {print
$1,$4-1,$5,$2,$3,$6,$7,$8,$9} }' | sortBed | mergeBed >
${name}_exon_merged.bed
    echo "Make gene gff"
    zcat ${gff} | gawk -v OFS='\t' '{if($3=="gene") print $0}'
> ${name}_genes.gff
else
    echo "5'UTR"
    cat ${gff} | gawk -v OFS='\t' '{if($3=="five_prime_UTR")
{print $1,$4-1,$5,$2,$3,$6,$7,$8,$9} }' | sortBed | mergeBed >
${name}_5P_merged.bed
    echo "3'UTR"
    cat ${gff} | gawk -v OFS='\t' '{if($3=="three_prime_UTR")
{print $1,$4-1,$5,$2,$3,$6,$7,$8,$9} }' | sortBed | mergeBed >
${name}_3P_merged.bed
    echo "Exons"
    cat ${gff} | gawk -v OFS='\t' '{if($3=="exon") {print
$1,$4-1,$5,$2,$3,$6,$7,$8,$9} }' | sortBed | mergeBed >
${name}_exon_merged.bed
    echo "Make gene gff"
    cat ${gff} | gawk -v OFS='\t' '{if($3=="gene") print $0}' >
${name}_genes.gff
fi

# There is no "intron" annotation. To get those, get the ones
in genes which are not in exons
echo "Genic regions"
```

```
gawk -v OFS='\t' '{if($3=="gene") {print
$1,$4-1,$5,$2,$3,$6,$7,$8,$9} }' ${name}_genes.gff | sortBed |
mergeBed > ${name}_genic_merged.bed

echo "1kb upstream"
gawk -v OFS='\t' '{if($7=="+" && $4>=1001) {print
$1,$4-1001,$4-1,$2,$3,$6,$7,$8,$9} else {if($7=="+" &&
$4<1001){print $1,0,$4-1,$2,$3,$6,$7,$8,$9} else {print
$1,$5,$5+1000,$2,$3,$6,$7,$8,$9}}}' ${name}_genes.gff | sortBed
| mergeBed > ${name}_1kb_up_merged.bed


echo "1-5kb upstream"
gawk -v OFS='\t' '{if($7=="+" && $4>=5001) {print
$1,$4-5001,$4-1001,$2,$3,$6,$7,$8,$9} else {if($7=="+" &&
$4<5001 && $4>1001){print $1,0,$4-1001,$2,$3,$6,$7,$8,$9}
else{if($7=="+" && $4<=1001){} else {print
$1,$5+1000,$5+5000,$2,$3,$6,$7,$8,$9}}}}' ${name}_genes.gff |
sortBed | mergeBed > ${name}_1kb_to_5kb_up_merged.bed


echo "1kb downstream"
gawk -v OFS='\t' '{if($7=="-" && $4>=1001) {print
$1,$4-1001,$4-1,$2,$3,$6,$7,$8,$9} else {if($7=="-" && $4<1001
&& $4>1){print $1,0,$4-1,$2,$3,$6,$7,$8,$9} else{if($7=="-" &&
$4==1){} else {print $1,$5,$5+1000,$2,$3,$6,$7,$8,$9}}}}'
${name}_genes.gff | sortBed | mergeBed >
${name}_1kb_down_merged.bed


echo "1-5kb upstream"
gawk -v OFS='\t' '{if($7=="-" && $4>=5001) {print
$1,$4-5001,$4-1001,$2,$3,$6,$7,$8,$9} else {if($7=="-" &&
$4<5001 && $4>1001){print $1,0,$4-1001,$2,$3,$6,$7,$8,$9}
else{if($7=="-" && $4<=1001){} else {print
```

```
$1,$5+1000,$5+5000,$2,$3,$6,$7,$8,$9}}}}' ${name}_genes.gff |
sortBed | mergeBed > ${name}_1kb_to_5kb_down_merged.bed


## To get intergenic make file with all genes plus 5kb upstream
and 5 kb downstream (to be subtracted)
echo "all genes plus 5kb upstream and 5 kb downstream"
gawk -v OFS='\t' '{if($4>=5001) {print
$1,$4-5001,$5+5000,$2,$3,$6,$7,$8,$9} else {print
$1,0,$5+5000,$2,$3,$6,$7,$8,$9}}' ${name}_genes.gff | sortBed |
mergeBed > ${name}_genes_5kb_up_5kb_down_merged.bed


echo "Account for possible scaffold entries that have 0 0"
for i in 5P 3P exon genic 1kb_up 1kb_to_5kb_up 1kb_down
1kb_to_5kb_down genes_5kb_up_5kb_down; do echo "${i}"; gawk -v
OFS='\t' '{if($2==(-1)){print $1, 0, $3} else {print $0}}'
${name}_${i}_merged.bed > ${name}_${i}_merged_noNeg1.bed; done



# Priority
# 5'UTR
# 3'UTR
# Exons
# Intron
# 1kbup
# 1kbdown
# 1-5kb up
# 1-5kb down
# intergenic
```

```
### Have to add -sorted option for very large chromosomes

# To enable this, substract the higher priorities from lower

echo "Subtract according to priority..."

echo "3P"
bedtools subtract -sorted -a ${name}_3P_merged_noNeg1.bed -b
${name}_5P_merged_noNeg1.bed > ${name}_3P_merged_sub.bed

echo "exon"
bedtools subtract -sorted -a ${name}_exon_merged_noNeg1.bed -b
${name}_5P_merged_noNeg1.bed | bedtools subtract -sorted -a -
-b ${name}_3P_merged_noNeg1.bed > ${name}_exon_merged_sub.bed

echo "intron"
bedtools subtract -sorted -a ${name}_genic_merged_noNeg1.bed -b
${name}_5P_merged_noNeg1.bed | bedtools subtract -sorted -a -
-b ${name}_3P_merged_noNeg1.bed | bedtools subtract -sorted -a
- -b ${name}_exon_merged_noNeg1.bed >
${name}_intron_merged_sub.bed

echo "1kp up"
bedtools subtract -sorted -a ${name}_1kb_up_merged_noNeg1.bed
-b ${name}_genic_merged_noNeg1.bed >
${name}_1kb_up_merged_sub.bed

echo "1kb down"
bedtools subtract -sorted -a ${name}_1kb_down_merged_noNeg1.bed
-b ${name}_genic_merged_noNeg1.bed | bedtools subtract -sorted
-a - -b ${name}_1kb_up_merged_noNeg1.bed >
${name}_1kb_down_merged_sub.bed

echo "1kb to 5kb up"
```

```
bedtools subtract -sorted -a
${name}_1kb_to_5kb_up_merged_noNeg1.bed -b
${name}_genic_merged_noNeg1.bed | bedtools subtract -sorted -a
- -b ${name}_1kb_up_merged_noNeg1.bed | bedtools subtract
-sorted -a - -b ${name}_1kb_down_merged_noNeg1.bed >
${name}_1kb_to_5kb_up_merged_sub.bed

echo "1kb to 5kb down"
bedtools subtract -sorted -a
${name}_1kb_to_5kb_down_merged_noNeg1.bed -b
${name}_genic_merged_noNeg1.bed | bedtools subtract -sorted -a
- -b ${name}_1kb_up_merged_noNeg1.bed | bedtools subtract
-sorted -a - -b ${name}_1kb_down_merged_noNeg1.bed | bedtools
subtract -sorted -a - -b
${name}_1kb_to_5kb_up_merged_noNeg1.bed >
${name}_1kb_to_5kb_down_merged_sub.bed




echo "Get overlapping peaks..."

cd ${outDir}

echo "Excluding peaks in contigs that have no gene annotation
in the gff..."
gawk '{if(NR==FNR){contigs[$1]=$1; next} if($1 in
contigs){print $0}}'
${outDir}/${name}_featureBeds/${name}_genic_merged_noNeg1.bed
${peaks} > ${name}_peaks.bed

for i in 5P genic; do echo "${i}"; intersectBed -sorted -a
${name}_peaks.bed -b
${outDir}/${name}_featureBeds/${name}_${i}_merged_noNeg1.bed  >
`basename ${peaks} .narrowPeak`.${i}.bp.bed; done
```

```
for i in 3P exon intron 1kb_up 1kb_to_5kb_up 1kb_down
1kb_to_5kb_down; do echo "${i}"; intersectBed -sorted -a
${name}_peaks.bed -b
${outDir}/${name}_featureBeds/${name}_${i}_merged_sub.bed >
`basename ${peaks} .narrowPeak`.${i}.bp.bed; done

echo "intergenic"
bedtools subtract -sorted -a ${name}_peaks.bed -b
${outDir}/${name}_featureBeds/${name}_genes_5kb_up_5kb_down_mer
ged_noNeg1.bed > `basename ${peaks}
.narrowPeak`.intergenic.bp.bed


wc -l *.bp.bed > ${name}_results_peak_numbers.csv



if [ -f ${name}_results_added_bp.csv ]
then
echo "removing previous output file"
   rm ${name}_results_added_bp.csv
fi


for i in `basename ${peaks} .narrowPeak`*.bp.bed; do awk -v
OFS='\t' -v i=${i} 'BEGIN{a=0}{a=a+($3-$2)}END{print i,a}' ${i}
>> ${name}_results_added_bp.csv;done



echo "--------- DONE ---------"
```

5.1.10. Script for BigWig track generation for genome browser visualization.

The script used in this thesis to generate BigWig coverage tracks from BAM files was written by me based on an existing bamCoverage workflow originally developed by Dr. Thomas Hartwig for MOA-seq analyses. The original script implemented a parameterized framework for systematic testing of bamCoverage settings across multiple samples and conditions. For the analyses presented here, I implemented an independent bash script that retained the relevant bamCoverage parameters from the original workflow while simplifying the structure for use with individual BAM files and being able to run on a high-performance computer (HPC) server. No changes were made to the normalization strategy or core bamCoverage settings relative to those used in the original workflow. The adapted script was used to generate CPM-normalized BigWig files for genome browser visualization and downstream comparison.

```
#!/bin/bash
# Run bamCoverage to make a BigWig file from a BAM file
# Usage: ./bamcoverage_CPM_bigwig_hpc.sh threads inputBamFile
outputPathandName

EXPECTED_ARGS=3
E_BADARGS=3
if [ $# -ne $EXPECTED_ARGS ]; then
         echo "Usage: ./bamcoverage.sh threads inputBamFile
outputPathandName"
    exit $E_BADARGS
fi

threads=$1
inputBamFile=$2
outputPathandName=$3

# Load modules
source /opt/share/software/scs/appStore/modules/init/profile.sh
module load deeptools/v3.5.5

# Index BAM file if index doesn't exist
```

```
echo "Checking for BAM index for ${inputBamFile}"
if [ -f "${inputBamFile}.bai" ]; then
     echo "Index already exists: ${inputBamFile}.bai — skipping
indexing."
else
    echo "Indexing BAM file: ${inputBamFile}"
    samtools index -c "${inputBamFile}"
fi

# Run bamCoverage
echo "Running bamCoverage on ${inputBamFile}"
bamCoverage \
    -b "${inputBamFile}" \
    -o "${outputPathandName}" \
    -of bigwig \
    -p "${threads}" \
    --normalizeUsing CPM \
    --minMappingQuality 255 \
    --exactScaling \
    --smoothLength 0 \
    --maxFragmentLength 80 \
    --binSize 1

echo "bamCoverage complete. Output file: ${outputPathandName}"
```

5.1.11. MACS3 peak calling and post-processing commands for PPZ and PPZ+eBL comparisons.

```
# PPZ vs PPZ+eBL (and reciprocal) in barley (Hordeum vulgare) and
Arabidopsis
########################################
# Barley (Hordeum vulgare, Morex V3)
########################################
```

```
# Effective genome size (EG)
# EG=2848066973

# PPZ vs PPZ+eBL (PPZ as treatment, PPZ+eBL as control)
macs3 callpeak \
  -t
/netscratch/dep_psl/grp_frommer/Duong/results/MOA-seq/brassinoste
roid/Barley-morex/STAR/Morex_PPZ_R1-3Aligned.sortedByCoord.out.ba
m \
  -c
/netscratch/dep_psl/grp_frommer/Duong/results/MOA-seq/brassinoste
roid/Barley-morex/STAR/Morex_BL_R1-3Aligned.sortedByCoord.out.bam
\
  -n Morex_PPZ_R1-3.vs.Morex_BL_R1-3.q0.05 \
  --outdir
/netscratch/dep_psl/grp_frommer/Thomas/Results/MOA_barley/BR/MACS
3/Morex_PPZ_R1-3vsMorex_BL_R1-3 \
  -f BAM -q 0.05 -g 2848066973 -s 60 --min-length 60 --max-gap 60
\
  --nomodel --extsize 60 --keep-dup all --buffer-size 10000000

# PPZ+eBL vs PPZ (PPZ+eBL as treatment, PPZ as control)
macs3 callpeak \
  -t
/netscratch/dep_psl/grp_frommer/Duong/results/MOA-seq/brassinoste
roid/Barley-morex/STAR/Morex_BL_R1-3Aligned.sortedByCoord.out.bam
\
  -c
/netscratch/dep_psl/grp_frommer/Duong/results/MOA-seq/brassinoste
roid/Barley-morex/STAR/Morex_PPZ_R1-3Aligned.sortedByCoord.out.ba
m \
  -n Morex_BL_R1-3.vs.Morex_PPZ_R1-3.q0.05 \
  --outdir
/netscratch/dep_psl/grp_frommer/Thomas/Results/MOA_barley/BR/MACS
3/Morex_BL_R1-3vsMorex_PPZ_R1-3 \
```

```
  -f BAM -q 0.05 -g 2848066973 -s 60 --min-length 60 --max-gap 60
\
  --nomodel --extsize 60 --keep-dup all --buffer-size 10000000

########################################
# Arabidopsis thaliana (TAIR10)
########################################

# Effective genome size (EG)
# EG=119667750

# PPZ vs PPZ+eBL (PPZ as treatment, PPZ+eBL as control), local
lambda restricted
macs3 callpeak \
  -t
/netscratch/dep_psl/grp_frommer/Duong/results/MOA-seq/brassinoste
roid/Arabidopsis-Col0/STAR/Arabidopsis_PPZ_R1-4Aligned.sortedByCo
ord.out.bam \
  -c
/netscratch/dep_psl/grp_frommer/Duong/results/MOA-seq/brassinoste
roid/Arabidopsis-Col0/STAR/Arabidopsis_BL_R1-4Aligned.sortedByCoo
rd.out.bam \
  -n Ara_Col0_PPZ_R1-4.vs.Ara_Col0_BL_R1-4.q0.05 \
  --outdir
/netscratch/dep_psl/grp_frommer/Thomas/Results/MOA_Ara_BR/MACS3/A
ra_Col0_PPZ_R1-4vsAra_Col0_BL_R1-4_small_lambda \
  -f BAM -q 0.05 -g 119667750 -s 60 --min-length 60 --max-gap 60
\
  --nomodel --extsize 60 --keep-dup all --buffer-size 10000000 \
  --slocal 120

# PPZ+eBL vs PPZ (PPZ+eBL as treatment, PPZ as control), local
lambda restricted
macs3 callpeak \
```

```
  -t
/netscratch/dep_psl/grp_frommer/Duong/results/MOA-seq/brassinoste
roid/Arabidopsis-Col0/STAR/Arabidopsis_BL_R1-4Aligned.sortedByCoo
rd.out.bam \
  -c
/netscratch/dep_psl/grp_frommer/Duong/results/MOA-seq/brassinoste
roid/Arabidopsis-Col0/STAR/Arabidopsis_PPZ_R1-4Aligned.sortedByCo
ord.out.bam \
  -n Ara_Col0_BL_R1-4.vs.Ara_Col0_PPZ_R1-4.q0.05 \
  --outdir
/netscratch/dep_psl/grp_frommer/Thomas/Results/MOA_Ara_BR/MACS3/A
ra_Col0_BL_R1-4vsAra_Col0_PPZ_R1-4_small_lambda \
  -f BAM -q 0.05 -g 119667750 -s 60 --min-length 60 --max-gap 60
\
  --nomodel --extsize 60 --keep-dup all --buffer-size 10000000 \
  --slocal 120

########################################
# Post-processing: filter to q = 1e-5 (0.00001) for Arabidopsis
# narrowPeak column 9 is -log10(qvalue); q <= 1e-5 corresponds to
-log10(q) >= 5
########################################

gawk -v OFS='\t' '{if($9>5) print $0}' \

Ara_Col0_BL_R1-4.vs.Ara_Col0_PPZ_R1-4.q0.05_peaks_small_lambda.na
rrowPeak \
  >
Ara_Col0_BL_R1-4.vs.Ara_Col0_PPZ_R1-4.q0.00001_peaks_small_lambda
.narrowPeak

gawk -v OFS='\t' '{if($9>5) print $0}' \

Ara_Col0_PPZ_R1-4.vs.Ara_Col0_BL_R1-4.q0.05_peaks_small_lambda.na
rrowPeak \
```

> Ara_Col0_PPZ_R1-4.vs.Ara_Col0_BL_R1-4.q0.00001_peaks_small_lambda.narrowPeak

5.2. Original MOA-seq Agarose Gel Images

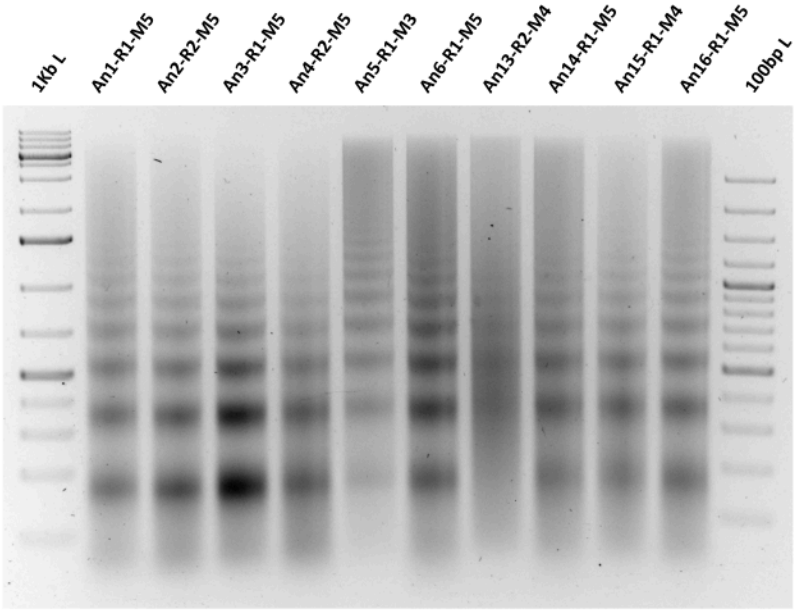


Figure 5.1: 1.5% agarose gel image showing the MOA-seq partial digestion of Andropogoneae species. An1 - *Themeda triandra*, An2 - *Pogonatherum paniceum*, An3 - *Anatherum virginicus*, An4 - *Heteropogon contortus*, An5 - *Thelepogon elegans*, An6 - *Andropogon burmanicus*, An13 - *Hemarthria compressa*, An14 - *Bothriochloa laguroides*, An15 - *Elionurus tripsacoides*, An16 - *Urelytrum digitatum*.

The 1 Kb ladder is on the left, and the 100 bp ladder is on the right.

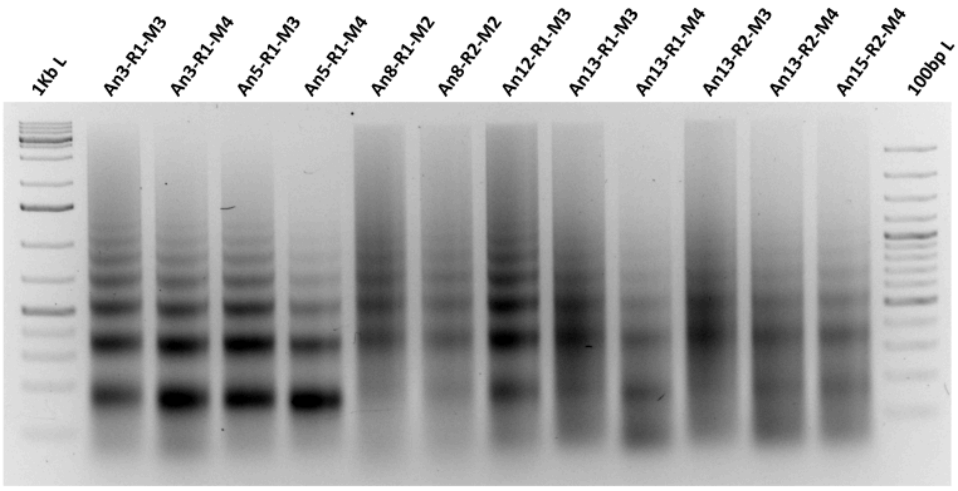


Figure 5.2: 1.5% agarose gel image showing the MOA-seq partial digestion of Andropogoneae species. An3 - *Anatherum virginicus*, An5 - *Thelepogon elegans*, An8 - *Cymbopogon citratus*, An12 - *Andropogon gerardi*, An13 - *Hemarthria compressa*, An15 - *Elionurus tripsacoides*.

The 1 Kb ladder is on the left, and the 100 bp ladder is on the right.

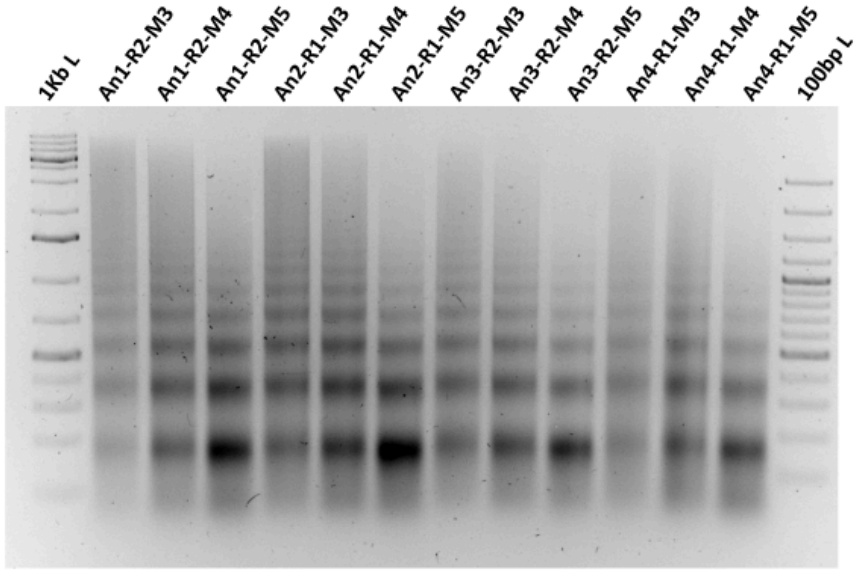


Figure 5.3: 1.5% agarose gel image showing the MOA-seq partial digestion of Andropogoneae species. An1 - *Themeda triandra*, An2 - *Pogonatherum paniceum*, An3 - *Anatherum virginicus*, An4 - *Heteropogon contortus*.

The 1 Kb ladder is on the left, and the 100 bp ladder is on the right.

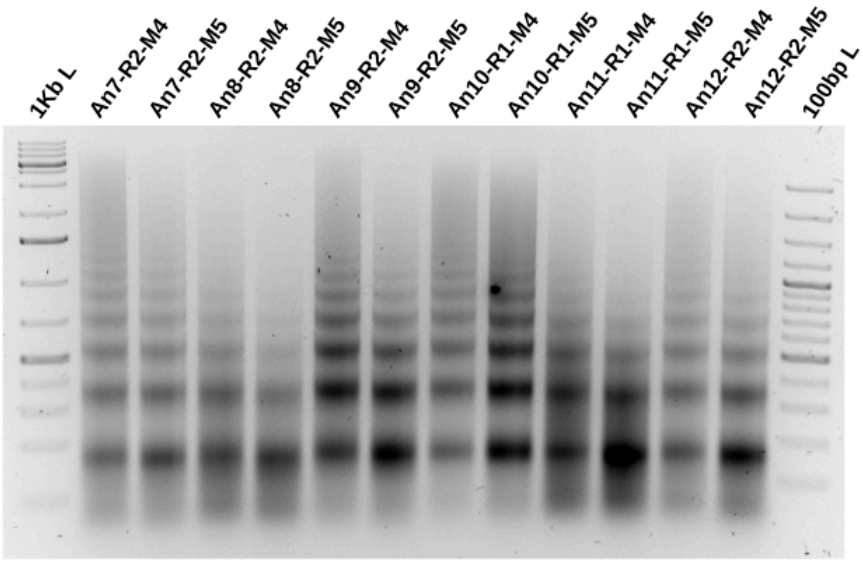


Figure 5.4: 1.5% agarose gel image showing the MOA-seq partial digestion of Andropogoneae species. An7 - *Schizachyrium scoparium*, An8 - *Cymbopogon citratus*, An9 - *Zea diploperennis Momo*, An10 - *Sorghastrum nutans*, An11 - T*ripsacum dactyloides KS*, An12 - *Andropogon gerardi*.

The 1 Kb ladder is on the left, and the 100 bp ladder is on the right.

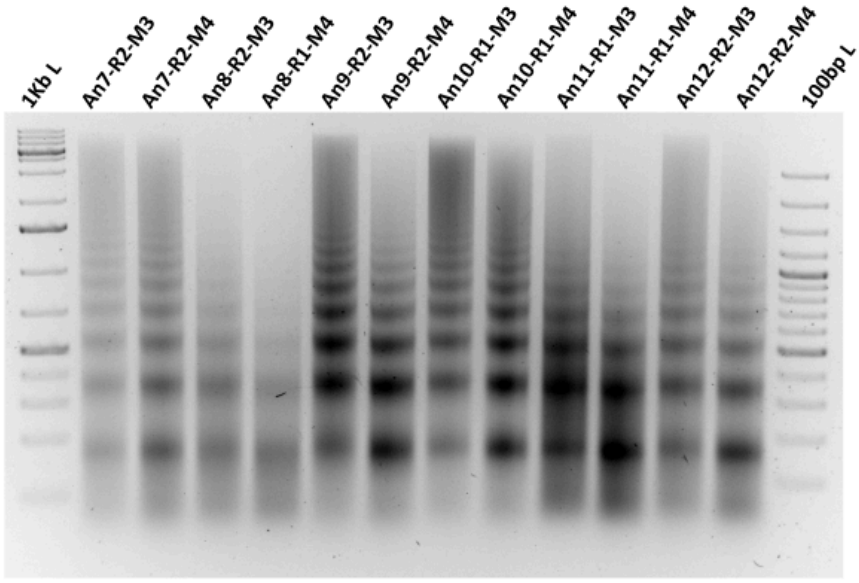


Figure 5.5: 1.5% agarose gel image showing the MOA-seq partial digestion of Andropogoneae species. An7 - *Schizachyrium scoparium*, An8 - *Cymbopogon citratus*, An9 - *Zea diploperennis Momo*, An10 - *Sorghastrum nutans*, An11 - T*ripsacum dactyloides KS*, An12 - *Andropogon gerardi*.

The 1 Kb ladder is on the left, and the 100 bp ladder is on the right.

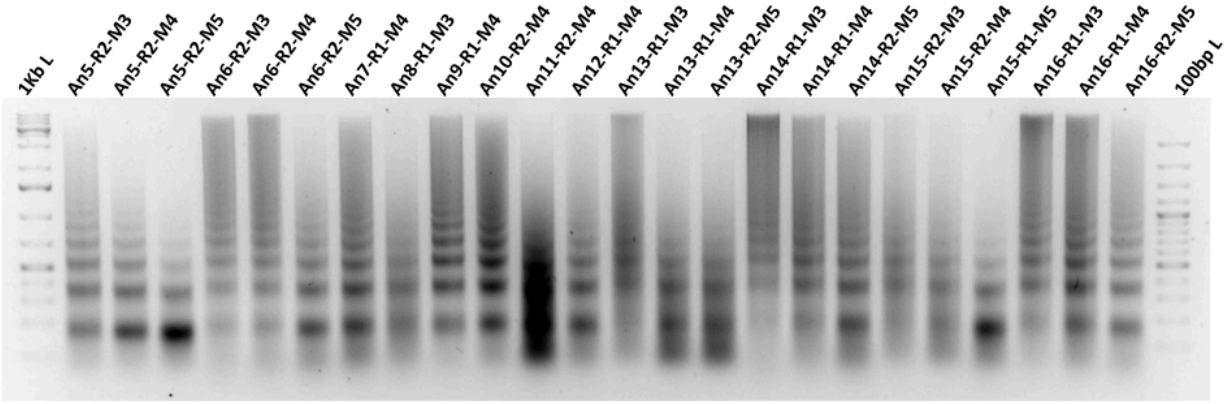


Figure 5.6: 1.5% agarose gel image showing the MOA-seq partial digestion of Andropogoneae species. An5 - *Thelepogon elegans*, An6 - *Andropogon burmanicus*, An7 - *Schizachyrium scoparium*, An8 - *Cymbopogon citratus*, An9 - *Zea diploperennis Momo*, An10 - *Sorghastrum nutans*, An11 - T*ripsacum dactyloides KS*, An12 - *Andropogon gerardi,* An13 - *Hemarthria compressa*, An14 - *Bothriochloa laguroides,* An15 - *Elionurus tripsacoides,* An16 - *Urelytrum digitatum.*

The 1 Kb ladder is on the left, and the 100 bp ladder is on the right.

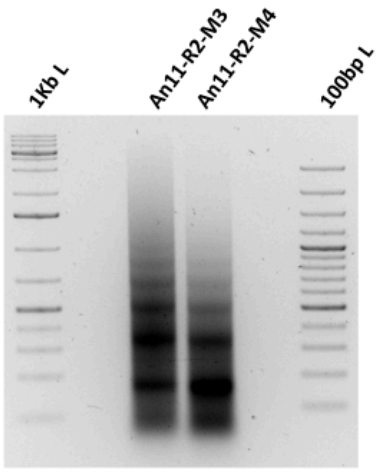


Figure 5.7: 1.5% agarose gel image showing the MOA-seq partial digestion of Andropogoneae species. An11 - T*ripsacum dactyloides KS.*

The 1 Kb ladder is on the left, and the 100 bp ladder is on the right.

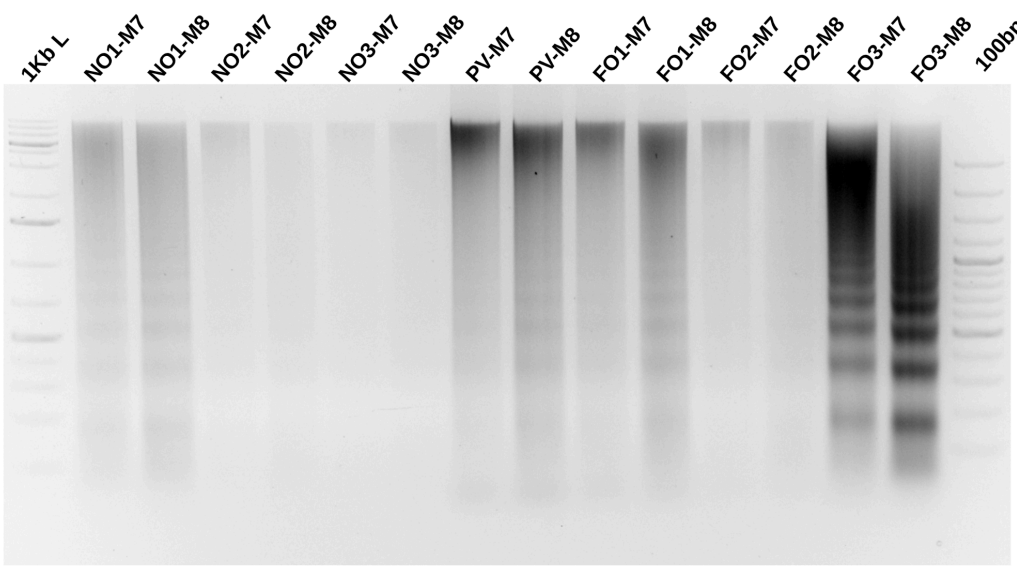


Figure 5.8: 1.5% agarose gel image showing the MOA-seq partial digestion of *Betula pendula* (NO: mock condition, FFO: flowering treatment) samples.

The 1 Kb ladder is on the left, and the 100 bp ladder is on the right.

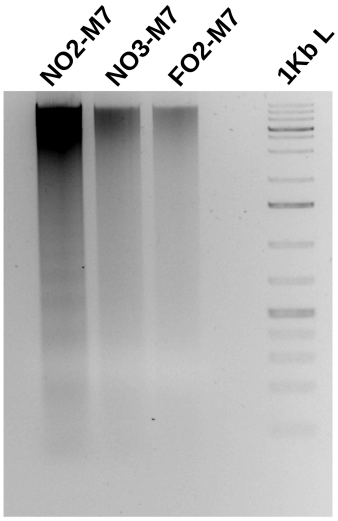


Figure 5.9: 1.5% agarose gel image showing the MOA-seq partial digestion of *Betula pendula* (NO: mock condition, FFO: flowering treatment) samples.

The 1 Kb ladder is on the right.

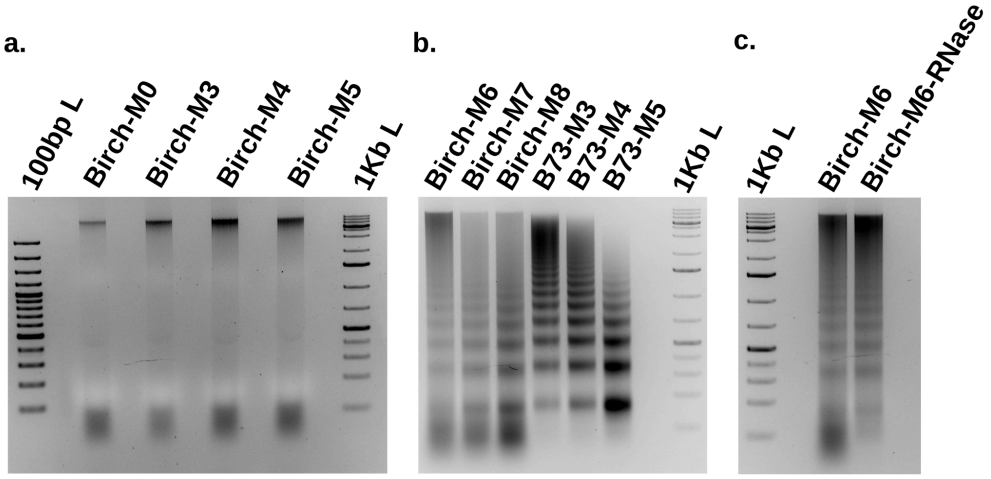


Figure 5.10: 1.5% agarose gel image showing the MOA-seq partial digestion of *Betula pendula* (FO1) samples.

(a) Samples of the first trial experiment with the standard protocol were loaded from left to right with the digestion conditions M0, M3, M4, and M5. (b) Samples were conducted with the optimized protocol but without the RNase I treatment step, and a comparison with Zea mays (B73) samples was performed with the standard protocol. (c) Sample before and after RNase I treatment with an extra cleanup step.

5.3. Supplement tables

Table 5.1: Genomic feature distribution of MOA-seq peak base pairs under PPZ and PPZ+eBL treatments in *Hordeum vulgare.*

Peak set	Intergenic	1 kb upstream	Intron	Exon	5′ UTR	3′ UTR
PPZ	22.03 Mb (50.0%)	6.86 Mb (15.6%)	1.89 Mb (4.3%)	1.30 Mb (3.0%)	0.80 Mb (1.8%)	1.06 Mb (2.4%)
PPZ+eBL	22.61 Mb (50.7%)	7.21 Mb (16.2%)	1.81 Mb (4.1%)	1.41 Mb (3.2%)	0.92 Mb (2.1%)	1.02 Mb (2.3%)
PPZ vs PPZ+eBL	0.37 Mb (36.6%)	0.35 Mb (34.6%)	0.016 Mb (1.6%)	0.015 Mb (1.5%)	0.057 Mb (5.7%)	0.004 Mb (0.4%)
PPZ+eBL vs PPZ	0.28 Mb (44.7%)	0.12 Mb (19.5%)	0.045 Mb (7.1%)	0.011 Mb (1.8%)	0.010 Mb (1.6%)	0.016 Mb (2.6%)

Table 5.2: Genomic feature distribution of MOA-seq peak base pairs under PPZ and PPZ+eBL treatments in *Arabidopsis thaliana*.

Peak set	Intergenic	1 kb upstream	Intron	Exon	5′ UTR	3′ UTR
BL	0.32 Mb (1.4%)	7.44 Mb (33.4%)	1.82 Mb (8.1%)	5.25 Mb (23.5%)	1.18 Mb (5.3%)	1.51 Mb (6.7%)
PPZ	0.27 Mb (1.3%)	7.86 Mb (36.1%)	1.48 Mb (6.8%)	3.47 Mb (16.0%)	1.17 Mb (5.4%)	1.46 Mb (6.7%)
PPZ+eBL vs PPZ	0.073 Mb (1.6%)	0.54 Mb (11.6%)	0.36 Mb (7.7%)	1.20 Mb (25.5%)	0.063 Mb (1.3%)	0.19 Mb (4.1%)
PPZ vs PPZ+eBL	0.066 Mb (1.0%)	3.32 Mb (52.5%)	0.31 Mb (5.0%)	0.33 Mb (5.3%)	0.35 Mb (5.6%)	0.30 Mb (4.7%)

For both tables, MOA-seq peaks identified under PPZ and PPZ+eBL treatment conditions, as well as differential peak sets defined by reciprocal comparisons (PPZ vs PPZ+eBL and PPZ+eBL vs PPZ), were assigned to annotated genomic features. For each peak set, the total number of base pairs (Mb) overlapping each feature category is shown together with the corresponding percentage of total peak base pairs within that set. Genomic features were defined as 1 kb upstream of annotated transcription start sites, introns, exons, 5′ UTR, 3′ UTR, and intergenic regions. Percentages represent the fraction of total MOA-seq peak base pairs within each peak set assigned to the respective feature category. These tables were generated based on the initial analysis from Dr. Julia Engelhorn.

5.4. List of Figures

Figure 1.1: Cis-regulatory elements (CREs) act as binding platforms for transcription factors (TFs) that activate gene expression in maize in response to environmental stimuli such as drought....14

Figure 1.2: TB1 binds to a cis-regulatory element upstream of its own promoter....15

Figure 1.3: Teosinte and maize morphology....16

Figure 1.4: 3D genome organization at increasing resolutions....17

Figure 1.5: Structural organization of chromatin....18

Figure 1.6: Chromatin organization and dynamic regulation of accessibility....21

Figure 1.7. Histone variants and post-translational modifications define chromatin states and regulatory activity....25

Figure 1.8. Overview of DNA methylation pathways and their regulatory functions in plants....27

Figure 1.9: Focused and dispersed transcription initiation....30

Figure 1.10: Overview of enhancer and silencer with their influence on transcription....32

Figure 1.11: Mechanisms of insulator function....35

Figure 1.12: Schematic representation of ChIP-seq workflow....38

Figure 1.13: Schematic workflow of DAP-seq....40

Figure 1.14: CUT&RUN process diagram....42

Figure 1.15: Overview of the DNase-seq protocol....46

Figure 1.16: Schematic workflow of ATAC-seq....49

Figure 1.17: Schematic overview of MNase-seq....51

Figure 1.18: Schematic overview of Fiber-seq for single-molecule chromatin accessibility profiling....53

Figure 1.19: Dual profiling of chromatin accessibility and cytosine methylation with SAM-seq...55

Figure 1.20: Cell-type–resolved chromatin accessibility in plant root tissues....57

Figure 1.21: Schematic overview of single-nucleus chromatin accessibility sequencing using combinatorial indexing....59

Figure 1.22: Droplet-based scATAC-seq in plants....60

Figure 1.23: Plant STARR-seq experimental design for high-throughput enhancer identification...

63

Figure 2.1: Overview of the MOA-seq procedure used to generate transcription factor footprint profiles across 27 plant species.....81

Figure 2.2: Phylogenetic relationships of the 27 plant species used for MOA-seq profiling.....82

Figure 2.3: MNase partial digestion optimization for MOA-seq using pooled Andropogoneae samples..... 83

Figure 2.4: MNase partial digestion profiles of selected MOA-seq from Andropogoneae species for library preparation and sequencing, including An1 - Themeda triandra, An2 - Pogonatherum paniceum, An3 - Anatherum virginicum, An4 - Heteropogon contortus, An5 - Thelepogon elegans, and An6 - Andropogon burmanicus..... 84

Figure 2.5: MNase partial digestion profiles of selected MOA-seq from Andropogoneae species for library preparation and sequencing, including An7 - Schizachyrium scoparium, An8 - Cymbopogon citratus, An9 - Zea diploperennis (Momo), An10 - Sorghastrum nutans, and An11 - Tripsacum dactyloides (KS).....84

Figure 2.6: MNase partial digestion profiles of selected MOA-seq from Andropogoneae species for library preparation and sequencing, including An12 - Andropogon gerardi, An13 - Hemarthria compressa, An14 - Bothriochloa laguroides, An15 - Elionurus tripsacoides, and An16 - Urelytrum digitatum.....85

Figure 2.7: MNase partial digestion profiles of MOA-seq in Sorghum bicolor..... 86

Figure 2.8: MNase partial digestion profiles of MOA-seq in Oryza sativa..... 87

Figure 2.9: MNase partial digestion profiles of MOA-seq in Brassica napus.....88

Figure 2.10: MNase partial digestion profiles of MOA-seq in Glycine max.....89

Figure 2.11: MNase partial digestion profiles of MOA-seq in Vicia faba.....90

Figure 2.12: MNase partial digestion profiles of MOA-seq in Physcomitrella patents..... 91

Figure 2.13: MNase partial digestion profiles of MOA-seq in Betula pendula..... 92

Figure 2.14: Correlation between genome size and total MOA-seq peak counts across species... 96

Figure 2.15. Relationship between total MOA-seq peak length and genome size across species. 97

Figure 2.16: Correlation between total number of protein-coding genes and genome size across species.....100

Figure 2.17: Correlation between average MOA peak per gene and genome size across species....101
Figure 2.18: Correlation between total number of protein-coding genes and total number of MOA peaks....102
Figure 2.19: Distribution of distances between MOA peaks and the nearest transcription start site (TSS) across 14 plant species....104
Figure 2.20: Genomic distribution of MOA peaks across 14 plant species....105
Figure 2.21: Genome-wide distribution of annotated genomic features across 14 plant species...106
Figure 3.1: The brassinosteroid biosynthesis pathway in plants....118
Figure 3.2: Comparison of MOA-seq and BZR1 ChIP-seq profiles at brassinosteroid biosynthesis genes in maize....120
Figure 3.3: Conserved intronic regulatory elements at the DWF4 locus across plant species..122
Figure 3.4: MOA-seq profiles at brassinosteroid biosynthesis genes in maize across tissues..124
Figure 3.5: MOA-seq and transcription factor ChIP-seq profiles at brassinosteroid biosynthesis genes in Arabidopsis thaliana....125
Figure 3.6: MOA-seq profiles at BR biosynthesis genes in Hordeum vulgare....127
Figure 3.7: Experimental workflow for hormone perturbation and MOA-seq analysis of BR-responsive cis-regulatory elements....131
Figure 3.8: MNase partial digestion profiles of MOA-seq in Arabidopsis thaliana....132
Figure 3.9: MNase partial digestion profiles of MOA-seq in Hordeum vulgare....132
Figure 3.10: MOA-seq signal at the HvDWF4 locus under PPZ and PPZ+eBL conditions....135
Figure 3.11: MOA-seq signal at the AtDWF4 locus under PPZ and PPZ+eBL conditions....137
Figure 3.12: Genomic distribution of MOA peaks under altered brassinosteroid levels in Hordeum vulgare....140
Figure 3.13: Genomic distribution of MOA peaks under altered brassinosteroid levels in Arabidopsis thaliana....141
Figure 4.1: Haplotype-specific MOA-seq signal and expression at ZmTINY....147
Figure 4.2: Allele-specific transcription factor occupancy at the ZmPGIP2 locus....148
Figure 4.3: Allele-specific transcription factor occupancy and mRNA abundance variation at the ZmPGM1 locus....149

Figure 4.4: Allele-specific regulatory variation at ZmWRKY68..... 151
Figure 4.5: Allele-specific MOA-seq signal at the ZmALD1 (Aldolase) locus..... 152
Figure 4.6: mRNA abundance at the Aldolase1 (ZmADL1) locus in 25 maize hybrids..... 152
Figure 4.7: PCR-based detection of a Mu transposon insertion in ZmPGIP2..... 160
Figure 4.8: PCR-based detection of a Mu transposon insertion in ZmWRKY68..... 160
Figure 4.9: PCR-based detection of a Mu transposon insertion in ZmPGM1.....161
Figure 5.1: 1.5% agarose gel image showing the MOA-seq partial digestion of Andropogoneae species. An1 - Themeda triandra, An2 - Pogonatherum paniceum,.....222
An3 - Anatherum virginicus, An4 - Heteropogon contortus, An5 - Thelepogon elegans, An6 - Andropogon burmanicus, An13 - Hemarthria compressa, An14 - Bothriochloa laguroides, An15 - Elionurus tripsacoides, An16 - Urelytrum digitatum.....222
Figure 5.2: 1.5% agarose gel image showing the MOA-seq partial digestion of Andropogoneae species. An3 - Anatherum virginicus, An5 - Thelepogon elegans, An8 - Cymbopogon citratus, An12 - Andropogon gerardi, An13 - Hemarthria compressa, An15 - Elionurus tripsacoides.... 223
Figure 5.3: 1.5% agarose gel image showing the MOA-seq partial digestion of Andropogoneae species. An1 - Themeda triandra, An2 - Pogonatherum paniceum,.....223
An3 - Anatherum virginicus, An4 - Heteropogon contortus..... 223
Figure 5.4: 1.5% agarose gel image showing the MOA-seq partial digestion of Andropogoneae species. An7 - Schizachyrium scoparium, An8 - Cymbopogon citratus, An9 - Zea diploperennis Momo, An10 - Sorghastrum nutans, An11 - Tripsacum dactyloides KS, An12 - Andropogon gerardi.....224
Figure 5.5: 1.5% agarose gel image showing the MOA-seq partial digestion of Andropogoneae species. An7 - Schizachyrium scoparium, An8 - Cymbopogon citratus, An9 - Zea diploperennis Momo, An10 - Sorghastrum nutans, An11 - Tripsacum dactyloides KS, An12 - Andropogon gerardi.....225
Figure 5.6: 1.5% agarose gel image showing the MOA-seq partial digestion of Andropogoneae species. An5 - Thelepogon elegans, An6 - Andropogon burmanicus, An7 - Schizachyrium scoparium, An8 - Cymbopogon citratus, An9 - Zea diploperennis Momo, An10 - Sorghastrum nutans, An11 - Tripsacum dactyloides KS, An12 - Andropogon gerardi, An13 - Hemarthria compressa, An14 - Bothriochloa laguroides, An15 - Elionurus tripsacoides, An16 - Urelytrum digitatum..... 225
Figure 5.7: 1.5% agarose gel image showing the MOA-seq partial digestion of Andropogoneae

species. An11 - Tripsacum dactyloides KS.....226
Figure 5.8: 1.5% agarose gel image showing the MOA-seq partial digestion of Betula pendula (NO: mock condition, FFO: flowering treatment) samples.....226
Figure 5.9: 1.5% agarose gel image showing the MOA-seq partial digestion of Betula pendula (NO: mock condition, FFO: flowering treatment) samples.....227
Figure 5.10: 1.5% agarose gel image showing the MOA-seq partial digestion of Betula pendula (FO1) samples.....227

5.5. List of Tables

Table 2.1. Plant species, ploidy levels, and reference genomes used for MOA-seq mapping.... 73
Table 2.2: (Listed as Table 2.3.1 in final outline) Genome size and sequencing reads mapping statistics for 28 species (with two haplotypes of Zea mays).....92
Table 3.1: Information on the binding of ZmBZR1, allele-specific ZmBZR1 bindings (ASBs), and MOA-seq signal at the intron region of brassinosteroid-associated genes in Zea mays.....127
Table 3.2: Information on the binding of AtBZR1 and MOA-seq signal at the intron region of brassinosteroid-associated genes in Arabidopsis thaliana.....128
Table 3.3: Information on the MOA-seq signal at the intron region of brassinosteroid-associated genes in Hordeum vulgare.....129
Table 3.4: Sequencing reads mapping statistics for Arabidopsis thaliana and Hordeum vulgare samples treated with PPZ and PPZ+eBL..... 133
Table 3.5: Numbers of differentially bound MOA peaks between PPZ and PPZ+eBL treatments in Hordeum vulgare and Arabidopsis thaliana.....134
Table 3.6: Differential MOA-seq peaks overlapping the intron of the HvDWF4 in barley.....135
Table 3.7: Differential MOA-seq peaks overlapping the intron of the AtDWF4 in Arabidopsis.. 138
Table 4.1: Genotyping primers used for BonnMu insertion lines..... 154
Table 4.2: Genotyping of BonnMu mutants in the summer 2022 field season..... 159
Table 4.3: Genotyping of BonnMu mutants in the summer 2023 field season..... 159
Table 5.1: Genomic feature distribution of MOA-seq peak base pairs under PPZ and PPZ+eBL treatments in Hordeum vulgare..... 228
Table 5.2: Genomic feature distribution of MOA-seq peak base pairs under PPZ and PPZ+eBL treatments in Arabidopsis thaliana.....228

5.6. List of Abbreviations

Abbreviation	Full Term / Description
ABA	Abscisic acid
ACR	Accessible chromatin region
AP2/ERF	APETALA2 / Ethylene Response Factor
ASB	Allele-specific binding
ASTRAL	Accurate Species TRee ALgorithm
ATAC-seq	Assay for Transposase-Accessible Chromatin sequencing
BAM	Binary Alignment/Map format
BonnMu	Bonn Mutator transposon population
bp	Base pair
BR	Brassinosteroid
BR6OX	Brassinosteroid-6-oxidase
BRRE	Brassinosteroid-responsive element
BZR1	Brassinazole Resistant 1
CAGE-seq	Cap analysis gene expression sequencing
CENH3	Centromeric histone H3 variant
CG	Cytosine-guanine context
CHG	Cytosine-H-guanine context
CHH	Cytosine-H-H context
ChIP-seq	Chromatin immunoprecipitation sequencing
CMT2/3	Chromomethylase 2 / Chromomethylase 3
CPD	Constitutive Photomorphogenesis and Dwarfism
CPM	Counts per million mapped reads
CRE	Cis-regulatory element
CRISPR	Clustered Regularly Interspaced Short Palindromic Repeats
CTCF	CCCTC-binding factor
CUT&RUN	Cleavage Under Targets and Release Using Nuclease
DAP-seq	DNA affinity purification sequencing
DET2	De-etiolated 2
DME	Demeter
DML2/3	Demeter-like 2 / Demeter-like 3

DNase-seq	DNase I hypersensitive site sequencing
DRM2	Domains Rearranged Methyltransferase 2
DS	Drought stress
DWF	Dwarf (BR biosynthesis genes such as DWF4)
eBL	Epibrassinolide
eEF-1β	Elongation factor 1 beta
F1 / F2	First / Second filial generation
Fiber-seq	Fiber sequencing
FIND-CIS	F1-mediated Identification of Non-coding Cis-element-Impacting SNPs
FO1	Flowering treatment 1 (Betula pendula experiment)
GWAS	Genome-wide association study
H2A.Z/W/X	Histone H2A variants (Z, Plant-specific W, X)
H3.1 / H3.3	Canonical histone H3 variant / Histone H3 variant 3
H3K27ac	Histone H3 lysine 27 acetylation
H3K27me3	Histone H3 lysine 27 trimethylation
H3K4me1/3	Histone H3 lysine 4 mono/trimethylation
H3K56ac	Histone H3 lysine 56 acetylation
H3K9ac	Histone H3 lysine 9 acetylation
H3K9me2/3	Histone H3 lysine 9 di/trimethylation
H4K20me1	Histone H4 lysine 20 monomethylation
Hi-C	High-throughput chromosome conformation capture
IGV	Integrative Genomics Viewer
kb	Kilobase
MACS3	Model-based Analysis of ChIP-Seq (version 3)
MET1	Methyltransferase 1
MNase-seq	Micrococcal nuclease sequencing
MOA-seq	MNase-defined cistrome-Occupancy Analysis sequencing
Mu	Mutator transposon
NAM	Nested Association Mapping
NO	Mock condition (Betula pendula experiment)
NuA4	Nucleosome acetyltransferase of H4 complex
PEAT	Paired-End Analysis of Transcription Start sites
PCR	Polymerase chain reaction
PGIP	Polygalacturonase-inhibiting protein
PIF4	Phytochrome Interacting Factor 4

PPZ	Propiconazole
PRC2	Polycomb Repressive Complex 2
RAMPAGE	RNA Annotation and Mapping of Promoters for Analysis of Gene Expression
RdDM	RNA-directed DNA methylation
RNase I	Ribonuclease I
ROS1	Repressor of Silencing 1
SAM-seq	Single-molecule accessibility mapping sequencing
SNP	Single nucleotide polymorphism
STARR-seq	Self-Transcribing Active Regulatory Region sequencing
SUVH	SU(VAR)3–9 Homolog
TAD	Topologically associating domain
TCP	TEOSINTE BRANCHED1, CYCLOIDEA, PCF1
TE	Transposable element
TF	Transcription factor
TIR	Terminal inverted repeat
TSS	Transcription start site
UTR	Untranslated region
WW	Well-watered

Acknowledgement

I sincerely thank Prof. Dr. Wolf Frommer, Prof. Dr. Benjamin Stich, and Dr. Thomas Hartwig for their supervision, scientific rigor, and continued support throughout this project. Their critical perspective and commitment to precise scientific thinking strongly shaped this work. I also thank Dr. Julia Engelhorn for her contributions to experimental design, data interpretation, and constructive discussions across multiple chapters of this thesis.

I am grateful to Prof. Dr. Jeffrey Ross-Ibarra, Prof. Dr. Edward Buckler, and Prof. Dr. Hank Bass for insightful discussions and for providing an evolutionary perspective that strengthened the comparative aspects of this project. I thank Dr. Cinta Romay and Dr. Michelle Stitzer for their support and for their contributions to the conceptual development and analysis of the project. I also thank Dr. Jiaying Zhu for her advice and support with the hormone treatment experiments described in Chapter 3.

My sincere thanks go to Prof. Dr. Andreas Weber for serving as my mentor and for his guidance and encouragement.

I gratefully acknowledge all collaborators who provided tissue samples and supported the experimental work that enabled the generation of this large dataset.

I am very thankful to Prof. Dr. Raphaël Mercier (Managing Director, Max Planck Institute for Plant Breeding Research), Dr. Christian Dumpitak (iGRAD, Heinrich-Heine-Universität Düsseldorf), Dr. Monika Schlosser (PhD Coordinator, International Max Planck Research Schools, MPIPZ), and Dr. Magdalena Marek (Gender Equality Officer, MPIPZ) for their continuous support and encouragement.

I thank Dr. Sohini Mukherjee, Nafiseh Sargheini, Sergius Weizel, Roman Hensel, Felix Plotecki, Hengqi Ji, Viola Rajan, Dominik Schueller, Natalie Bajaczyk, and Lisa Heller from the Crop Yield in Maize team for their support during my PhD. I am especially grateful to Thimo Trapp, my first junior student, for his dedication, reliability, and encouragement. I also thank Amelie Kok for her assistance with data analysis and interpretation and for her support in coding. Many thanks go to all members of AG Frommer for their support and for fostering a constructive scientific environment. I further acknowledge colleagues and collaborators at the Max Planck Institute for Plant Breeding Research who contributed through discussions and technical assistance.

Finally, I thank my family and friends for their unwavering support throughout this journey, particularly during challenging times.

References

Adey, Andrew, Hilary G. Morrison, Asan, et al. 2010. “Rapid, Low-Input, Low-Bias Construction of Shotgun Fragment Libraries by High-Density in Vitro Transposition.” *Genome Biology* 11 (12): R119. https://doi.org/10.1186/gb-2010-11-12-r119.

Arnold, Cosmas D., Daniel Gerlach, Christoph Stelzer, Łukasz M. Boryń, Martina Rath, and Alexander Stark. 2013. “Genome-Wide Quantitative Enhancer Activity Maps Identified by STARR-Seq.” *Science*, ahead of print, March 1. World. https://doi.org/10.1126/science.1232542.

Banerji, Julian, Sandro Rusconi, and Walter Schaffner. 1981. “Expression of a β-Globin Gene Is Enhanced by Remote SV40 DNA Sequences.” *Cell* 27 (2): 299–308. https://doi.org/10.1016/0092-8674(81)90413-X.

Bartlett, Anna, Ronan C. O’Malley, Shao-shan Carol Huang, et al. 2017a. “Mapping Genome-Wide Transcription-Factor Binding Sites Using DAP-Seq.” *Nature Protocols* 12 (8): 1659–72. https://doi.org/10.1038/nprot.2017.055.

Bartlett, Anna, Ronan C. O’Malley, Shao-shan Carol Huang, et al. 2017b. “Mapping Genome-Wide Transcription-Factor Binding Sites Using DAP-Seq.” *Nature Protocols* 12 (8): 1659–72. https://doi.org/10.1038/nprot.2017.055.

Batut, Philippe, and Thomas R. Gingeras. 2013. “Rampage: Promoter Activity Profiling by Paired-End Sequencing of 5′-Complete cDNAs.” *Current Protocols in Molecular Biology / Edited by Frederick M. Ausubel ... [et Al.]* 104 (November): Unit-25B.11. https://doi.org/10.1002/0471142727.mb25b11s104.

Baumgart, Leo A., Sharon I. Greenblum, Abraham Morales-Cruz, et al. 2025. “Recruitment, Rewiring and Deep Conservation in Flowering Plant Gene Regulation.” *Nature Plants* 11 (8): 1514–27. https://doi.org/10.1038/s41477-025-02047-0.

Bernal-Gallardo, Judith Jazmin, and Stefan de Folter. 2024. “Plant Genome Information Facilitates Plant Functional Genomics.” *Planta* 259 (5): 117. https://doi.org/10.1007/s00425-024-04397-z.

Bianconi, Matheus E., Jan Hackel, Maria S. Vorontsova, et al. 2020. “Continued Adaptation of C4 Photosynthesis After an Initial Burst of Changes in the Andropogoneae Grasses.” *Systematic Biology* 69 (3): 445–61. https://doi.org/10.1093/sysbio/syz066.

Bieluszewski, Tomasz, Weronika Sura, Wojciech Dziegielewski, et al. 2022. “NuA4 and H2A.Z Control Environmental Responses and Autotrophic Growth in Arabidopsis.” *Nature Communications* 13 (1): 277. https://doi.org/10.1038/s41467-021-27882-5.

Blank, Max. 2020. “High-Throughput Identification of Functional Cistrome and Transcriptome Variation Underlying Drought Response in Crops.” Heinrich-Heine University Düsseldorf.

Bock, Christoph, Paul Datlinger, Florence Chardon, et al. 2022. “High-Content CRISPR Screening.” *Nature Reviews Methods Primers* 2 (1): 1–23. https://doi.org/10.1038/s43586-021-00093-4.

Bolduc, Nathalie, Alper Yilmaz, Maria Katherine Mejia-Guerra, et al. 2012. “Unraveling the KNOTTED1 Regulatory Network in Maize Meristems.” *Genes & Development* 26 (15): 1685–90. https://doi.org/10.1101/gad.193433.112.

Bonev, Boyan, and Giacomo Cavalli. 2016. “Organization and Function of the 3D Genome.” *Nature Reviews Genetics* 17 (11): 661–78. https://doi.org/10.1038/nrg.2016.112.

Brennan, Kaelan J., Melanie Weilert, Sabrina Krueger, et al. 2023. “Chromatin Accessibility in the Drosophila Embryo Is Determined by Transcription Factor Pioneering and Enhancer Activation.” *Developmental Cell* 58 (19): 1898-1916.e9. https://doi.org/10.1016/j.devcel.2023.07.007.

Buenrostro, Jason D., Paul G. Giresi, Lisa C. Zaba, Howard Y. Chang, and William J. Greenleaf. 2013. “Transposition of Native Chromatin for Fast and Sensitive Epigenomic Profiling of Open Chromatin, DNA-Binding Proteins and Nucleosome Position.” *Nature Methods* 10 (12): 1213–18. https://doi.org/10.1038/nmeth.2688.

Calo, Eliezer, and Joanna Wysocka. 2013. “Modification of Enhancer Chromatin: What, How, and Why?” *Molecular Cell* 49 (5): 825–37. https://doi.org/10.1016/j.molcel.2013.01.038.

Capella-Gutiérrez, Salvador, José M. Silla-Martínez, and Toni Gabaldón. 2009. “trimAl: A Tool for Automated Alignment Trimming in Large-Scale Phylogenetic Analyses.” *Bioinformatics* 25 (15): 1972–73. https://doi.org/10.1093/bioinformatics/btp348.

Caragine, Christina M., Victoria T. Le, Meer Mustafa, et al. 2025. “Comprehensive Dissection of Cis-Regulatory Elements in a 2.8 Mb Topologically Associated Domain in Six Human Cancers.” *Nature Communications* 16 (1): 1611. https://doi.org/10.1038/s41467-025-56568-5.

Charron, Jean-Benoit F., Hang He, Axel A. Elling, and Xing Wang Deng. 2009. “Dynamic Landscapes of Four Histone Modifications during Deetiolation in Arabidopsis.” *The Plant Cell* 21 (12): 3732–48. https://doi.org/10.1105/tpc.109.066845.

Chen, Yang, Rui Liang, Yong Li, et al. 2024. “Chromatin Accessibility: Biological Functions, Molecular Mechanisms and Therapeutic Application.” *Signal Transduction and Targeted Therapy* 9 (1): 340. https://doi.org/10.1038/s41392-024-02030-9.

Chereji, Răzvan V., Terri D. Bryson, and Steven Henikoff. 2019. “Quantitative MNase-Seq Accurately Maps Nucleosome Occupancy Levels.” *Genome Biology* 20 (1): 198. https://doi.org/10.1186/s13059-019-1815-z.

Chetverina, Darya, Tsutomu Aoki, Maksim Erokhin, Pavel Georgiev, and Paul Schedl. 2014. “Making Connections: Insulators Organize Eukaryotic Chromosomes into Independent Cis-Regulatory Networks.” *BioEssays : News and Reviews in Molecular, Cellular and Developmental Biology* 36 (2): 163–72. https://doi.org/10.1002/bies.201300125.

Ciren, Danielle, Sophia Zebell, and Zachary B. Lippman. 2024. “Extreme Restructuring of Cis-Regulatory Regions Controlling a Deeply Conserved Plant Stem Cell Regulator.” *PLOS Genetics* 20 (3): e1011174. https://doi.org/10.1371/journal.pgen.1011174.

Concia, Lorenzo, Alaguraj Veluchamy, Juan S. Ramirez-Prado, et al. 2020. “Wheat Chromatin Architecture Is Organized in Genome Territories and Transcription Factories.” *Genome Biology* 21 (1): 104. https://doi.org/10.1186/s13059-020-01998-1.

Corces, M. Ryan, Jeffrey M. Granja, Shadi Shams, et al. 2018. “The Chromatin Accessibility Landscape of Primary Human Cancers.” *Science (New York, N.Y.)* 362 (6413): eaav1898. https://doi.org/10.1126/science.aav1898.

Cui, Wenrui, Zhiwen Chen, Xiaoxia Shangguan, et al. 2022. “*TRY* Intron2 Determined Its Expression in Inflorescence Activated by SPL9 and MADS-Box Genes in Arabidopsis.” *Plant Science* 321 (August): 111311. https://doi.org/10.1016/j.plantsci.2022.111311.

Cusanovich, Darren A., Riza Daza, Andrew Adey, et al. 2015. “Multiplex Single-Cell Profiling of Chromatin Accessibility by Combinatorial Cellular Indexing.” *Science* 348 (6237): 910–14. https://doi.org/10.1126/science.aab1601.

Deal, Roger B., and Steven Henikoff. 2011. “Histone Variants and Modifications in Plant Gene Regulation.” *Current Opinion in Plant Biology* 14 (2): 116–22. https://doi.org/10.1016/j.pbi.2010.11.005.

Deyholos, Michael K., and Leslie E. Sieburth. 2000. “Separable Whorl-Specific Expression and Negative Regulation by Enhancer Elements within the AGAMOUS Second Intron.” *The Plant Cell* 12 (10): 1799–810. https://doi.org/10.1105/tpc.12.10.1799.

Dobin, Alexander, Carrie A. Davis, Felix Schlesinger, et al. 2013. “STAR: Ultrafast Universal RNA-Seq Aligner.” *Bioinformatics* 29 (1): 15–21. https://doi.org/10.1093/bioinformatics/bts635.

Dolfini, Diletta, Zambelli ,Federico, Pavesi ,Giulio, and Roberto and Mantovani. 2009. “A Perspective of Promoter Architecture from the CCAAT Box.” *Cell Cycle* 8 (24): 4127–37. https://doi.org/10.4161/cc.8.24.10240.

Domb, Katherine, Nan Wang, Guillaume Hummel, and Chang Liu. 2022. “Spatial Features and Functional Implications of Plant 3D Genome Organization.” *Annual Review of Plant Biology* 73 (1): 173–200. https://doi.org/10.1146/annurev-arplant-102720-022810.

Dong, Pengfei, Xiaoyu Tu, Po-Yu Chu, et al. 2017. “3D Chromatin Architecture of Large Plant Genomes Determined by Local A/B Compartments.” *Molecular Plant* 10 (12): 1497–509. https://doi.org/10.1016/j.molp.2017.11.005.

Dong, Zhaobin, Yuguo Xiao, Rajanikanth Govindarajulu, et al. 2019. “The Regulatory Landscape of a Core Maize Domestication Module Controlling Bud Dormancy and Growth Repression.” *Nature Communications* 10 (1): 3810. https://doi.org/10.1038/s41467-019-11774-w.

Doni Jayavelu, Naresh, Ajay Jajodia, Arpit Mishra, and R. David Hawkins. 2020. “Candidate Silencer Elements for the Human and Mouse Genomes.” *Nature Communications* 11 (1): 1061. https://doi.org/10.1038/s41467-020-14853-5.

Dorrity, Michael W., Cristina M. Alexandre, Morgan O. Hamm, et al. 2021. “The Regulatory Landscape of Arabidopsis Thaliana Roots at Single-Cell Resolution.” *Nature Communications* 12 (1): 3334. https://doi.org/10.1038/s41467-021-23675-y.

Du, Jiamu, Lianna M. Johnson, Martin Groth, et al. 2014. “Mechanism of DNA Methylation-Directed Histone Methylation by KRYPTONITE.” *Molecular Cell* 55 (3): 495–504. https://doi.org/10.1016/j.molcel.2014.06.009.

Du, Xuan, Zhenlin Yang, Guohui Xie, et al. 2023. “Molecular Basis of the Plant ROS1-Mediated Active DNA Demethylation.” *Nature Plants* 9 (2): 271–79. https://doi.org/10.1038/s41477-022-01322-8.

Du, Zhou, Hui Li, Qiang Wei, et al. 2013. “Genome-Wide Analysis of Histone Modifications: H3K4me2, H3K4me3, H3K9ac, and H3K27ac in Oryza Sativa L. Japonica.” *Molecular Plant* 6 (5): 1463–72. https://doi.org/10.1093/mp/sst018.

Dunham, Ian, Anshul Kundaje, Shelley F. Aldred, et al. 2012. “An Integrated Encyclopedia of DNA Elements in the Human Genome.” *Nature* 489 (7414): 57–74. https://doi.org/10.1038/nature11247.

Engelhorn, Julia, Samantha J. Snodgrass, Amelie Kok, et al. 2025. “Genetic Variation at Transcription Factor Binding Sites Largely Explains Phenotypic Heritability in Maize.” *Nature Genetics* 57 (9): 2313–22. https://doi.org/10.1038/s41588-025-02246-7.

Fang, Jian, Jianjun Jiang, Sarah M. Leichter, et al. 2022. “Mechanistic Basis for Maintenance of CHG DNA Methylation in Plants.” *Nature Communications* 13 (1): 3877. https://doi.org/10.1038/s41467-022-31627-3.

Feng, Suhua, Shawn J. Cokus, Veit Schubert, Jixian Zhai, Matteo Pellegrini, and Steven E. Jacobsen. 2014. “Genome-Wide Hi-C Analyses in Wild-Type and Mutants Reveal High-Resolution Chromatin Interactions in *Arabidopsis*.” *Molecular Cell* 55 (5): 694–707. https://doi.org/10.1016/j.molcel.2014.07.008.

Ferrari, Simone, Donatella Vairo, Frederick M. Ausubel, Felice Cervone, and Giulia De Lorenzo. 2003. “Tandemly Duplicated Arabidopsis Genes That Encode Polygalacturonase-Inhibiting Proteins Are Regulated Coordinately by Different Signal Transduction Pathways in Response to Fungal Infection.” *The Plant Cell* 15 (1): 93–106. https://doi.org/10.1105/tpc.005165.

Firestone, Tessa M., Bryan J. Venters, Katherine Novitzky, et al. 2024. “High-Efficiency Genomic Mapping of Chromatin-Associated Targets with CUT&RUN.” Preprint, bioRxiv, December 6. https://doi.org/10.1101/2024.12.03.626419.

Foroozani, Maryam, Dylan H. Holder, and Roger B. Deal. 2022. “Histone Variants in the Specialization of Plant Chromatin.” *Annual Review of Plant Biology* 73 (Volume 73, 2022): 149–72. https://doi.org/10.1146/annurev-arplant-070221-050044.

Furey, Terrence S. 2012. “ChIP–Seq and beyond: New and Improved Methodologies to Detect and Characterize Protein–DNA Interactions.” *Nature Reviews Genetics* 13 (12): 840–52. https://doi.org/10.1038/nrg3306.

Galli, Mary, Arjun Khakhar, Zefu Lu, et al. 2018. “The DNA Binding Landscape of the Maize AUXIN RESPONSE FACTOR Family.” *Nature Communications* 9 (1): 4526. https://doi.org/10.1038/s41467-018-06977-6.

Gao, Weiwu, Wai Lim Ku, Lixia Pan, et al. 2021. “Multiplex Indexing Approach for the Detection of DNase I Hypersensitive Sites in Single Cells.” *Nucleic Acids Research* 49 (10): e56. https://doi.org/10.1093/nar/gkab102.

Gasperini, Molly, Andrew J. Hill, José L. McFaline-Figueroa, et al. 2019. “A Genome-Wide Framework for Mapping Gene Regulation via Cellular Genetic Screens.” *Cell* 176 (1): 377-390.e19. https://doi.org/10.1016/j.cell.2018.11.029.

Gaszner, Miklos, and Gary Felsenfeld. 2006. “Insulators: Exploiting Transcriptional and Epigenetic Mechanisms.” *Nature Reviews Genetics* 7 (9): 703–13. https://doi.org/10.1038/nrg1925.

Gidekel, Manuel, Beatriz Jimenez, and Luis Herrera-Estrella. 1996. “The First Intron of the *Arabidopsis Thaliana* Gene Coding for Elongation Factor 1β Contains an Enhancer-like Element.” *Gene* 170 (2): 201–6. https://doi.org/10.1016/0378-1119(95)00837-3.

Giles, Katherine A., Phillippa C. Taberlay, Anthony J. Cesare, and Mathew J. K. Jones. 2025. “Roles for the 3D Genome in the Cell Cycle, DNA Replication, and Double Strand Break Repair.” *Frontiers in Cell and Developmental Biology* 13 (February): 1548946. https://doi.org/10.3389/fcell.2025.1548946.

Gisselbrecht, Stephen S., Alexandre Palagi, Jesse V. Kurland, et al. 2020. “Transcriptional Silencers in Drosophila Serve a Dual Role as Transcriptional Enhancers in Alternate Cellular Contexts.” *Molecular Cell* 77 (2): 324-337.e8. https://doi.org/10.1016/j.molcel.2019.10.004.

Grandi, Fiorella C., Hailey Modi, Lucas Kampman, and M. Ryan Corces. 2022. “Chromatin Accessibility Profiling by ATAC-Seq.” *Nature Protocols* 17 (6): 1518–52. https://doi.org/10.1038/s41596-022-00692-9.

Guo, Jiangfan, Wei Li, Lianguang Shang, et al. 2021. “OsbHLH98 Regulates Leaf Angle in Rice through Transcriptional Repression of OsBUL1.” *New Phytologist* 230 (5): 1953–66. https://doi.org/10.1111/nph.17303.

Haberle, Vanja, and Alexander Stark. 2018. “Eukaryotic Core Promoters and the Functional Basis of Transcription Initiation.” *Nature Reviews Molecular Cell Biology* 19 (10): 621–37. https://doi.org/10.1038/s41580-018-0028-8.

Hainer, Sarah J., Ana Bošković, Kurtis N. McCannell, Oliver J. Rando, and Thomas G. Fazzio. 2019. “Profiling of Pluripotency Factors in Single Cells and Early Embryos.” *Cell* 177 (5): 1319-1329.e11. https://doi.org/10.1016/j.cell.2019.03.014.

Han, Jinlei, Pengxi Wang, Qiongli Wang, et al. 2020. “Genome-Wide Characterization of DNase I-Hypersensitive Sites and Cold Response Regulatory Landscapes in Grasses.” *The Plant Cell* 32 (8): 2457–73. https://doi.org/10.1105/tpc.19.00716.

Hartwig, Thomas, Michael Banf, Gisele Passaia Prietsch, et al. 2023. “Hybrid Allele-Specific ChIP-Seq Analysis Identifies Variation in Brassinosteroid-Responsive Transcription Factor Binding Linked to Traits in Maize.” *Genome Biology* 24 (1): 108. https://doi.org/10.1186/s13059-023-02909-w.

Heyndrickx, Ken S., Jan Van de Velde, Congmao Wang, Detlef Weigel, and Klaas Vandepoele. 2014. “A Functional and Evolutionary Perspective on Transcription Factor Binding in Arabidopsis Thaliana[C][W].” *The Plant Cell* 26 (10): 3894–910. https://doi.org/10.1105/tpc.114.130591.

Holst, Felix, Anthony M. Bolger, Felicitas Kindel, et al. 2025. “Helixer: Ab Initio Prediction of Primary Eukaryotic Gene Models Combining Deep Learning and a Hidden Markov Model.” *Nature Methods*, November 24, 1–8. https://doi.org/10.1038/s41592-025-02939-1.

Huang, Di, Hanna M. Petrykowska, Brendan F. Miller, Laura Elnitski, and Ivan Ovcharenko. 2019. “Identification of Human Silencers by Correlating Cross-Tissue Epigenetic Profiles and Gene Expression.” *Genome Research* 29 (4): 657–67. https://doi.org/10.1101/gr.247007.118.

Huang, Ranran, Qingquan Zhu, Shentong Tao, et al. 2024. “Genome-Wide Characterization of Fragile and Resistant Nucleosomes in Response to Cold Stress in Maize.” *Environmental and Experimental Botany* 222 (June): 105760. https://doi.org/10.1016/j.envexpbot.2024.105760.

Jayakodi, Murukarthick, Agnieszka A. Golicz, Jonathan Kreplak, et al. 2023. “The Giant Diploid Faba Genome Unlocks Variation in a Global Protein Crop.” *Nature* 615 (7953): 652–59. https://doi.org/10.1038/s41586-023-05791-5.

Jin, Run, Samantha Klasfeld, Yang Zhu, et al. 2021. “LEAFY Is a Pioneer Transcription Factor and Licenses Cell Reprogramming to Floral Fate.” *Nature Communications* 12 (1): 626. https://doi.org/10.1038/s41467-020-20883-w.

Johnson, David S., Ali Mortazavi, Richard M. Myers, and Barbara Wold. 2007. “Genome-Wide Mapping of in Vivo Protein-DNA Interactions.” *Science* 316 (5830): 1497–502. https://doi.org/10.1126/science.1141319.

Jores, Tobias, Nicholas A. Mueth, Jackson Tonnies, et al. 2024. “Small DNA Elements That Act as Both Insulators and Silencers in Plants.” *bioRxiv*, September 19, 2024.09.13.612883. https://doi.org/10.1101/2024.09.13.612883.

Jores, Tobias, Jackson Tonnies, Michael W. Dorrity, Josh T. Cuperus, Stanley Fields, and Christine Queitsch. 2020. “Identification of Plant Enhancers and Their Constituent Elements by STARR-Seq in Tobacco Leaves[OPEN].” *The Plant Cell* 32 (7): 2120–31. https://doi.org/10.1105/tpc.20.00155.

Jores, Tobias, Jackson Tonnies, Travis Wrightsman, et al. 2021. “Synthetic Promoter Designs Enabled by a Comprehensive Analysis of Plant Core Promoters.” *Nature Plants* 7 (6): 842–55. https://doi.org/10.1038/s41477-021-00932-y.

Juven-Gershon, Tamar, and James T. Kadonaga. 2010. “Regulation of Gene Expression via the Core Promoter and the Basal Transcriptional Machinery.” *Developmental Biology*, Special Section: Gene Expression and Development, vol. 339 (2): 225–29. https://doi.org/10.1016/j.ydbio.2009.08.009.

Kalunke, Raviraj M., Silvio Tundo, Manuel Benedetti, Felice Cervone, Giulia De Lorenzo, and Renato D’Ovidio. 2015. “An Update on Polygalacturonase-Inhibiting Protein (PGIP), a Leucine-Rich Repeat Protein That Protects Crop Plants against Pathogens.” *Frontiers in Plant Science* 6 (March). https://doi.org/10.3389/fpls.2015.00146.

Karaaslan, Ezgi Süheyla, Nan Wang, Natalie Faiß, et al. 2020. “Marchantia TCP Transcription Factor Activity Correlates with Three-Dimensional Chromatin Structure.” *Nature Plants* 6 (10): 1250–61. https://doi.org/10.1038/s41477-020-00766-0.

Katoh, Kazutaka, and Daron M. Standley. 2013. “MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability.” *Molecular Biology and Evolution* 30 (4): 772–80. https://doi.org/10.1093/molbev/mst010.

Kaya-Okur, Hatice S., Steven J. Wu, Christine A. Codomo, et al. 2019. “CUT&Tag for Efficient Epigenomic Profiling of Small Samples and Single Cells.” *Nature Communications* 10 (1): 1930. https://doi.org/10.1038/s41467-019-09982-5.

Kim, Jong-Myong, Taku Sasaki, Minoru Ueda, Kaori Sako, and Motoaki Seki. 2015. “Chromatin Changes in Response to Drought, Salinity, Heat, and Cold Stresses in Plants.” *Frontiers in Plant Science* 6 (March). https://doi.org/10.3389/fpls.2015.00114.

Klein, David C., and Sarah J. Hainer. 2020. “Genomic Methods in Profiling DNA Accessibility and Factor Localization.” *Chromosome Research* 28 (1): 69–85. https://doi.org/10.1007/s10577-019-09619-9.

Klemm, Sandy L., Zohar Shipony, and William J. Greenleaf. 2019. “Chromatin Accessibility and the Regulatory Epigenome.” *Nature Reviews Genetics* 20 (4): 207–20. https://doi.org/10.1038/s41576-018-0089-8.

Kump, Kristen L., Peter J. Bradbury, Randall J. Wisser, et al. 2011. “Genome-Wide Association Study of Quantitative Resistance to Southern Leaf Blight in the Maize Nested Association Mapping Population.” *Nature Genetics* 43 (2): 163–68. https://doi.org/10.1038/ng.747.

Kurbidaeva, Amina, and Michael Purugganan. 2021. “Insulators in Plants: Progress and Open Questions.” *Genes* 12 (9): 1422. https://doi.org/10.3390/genes12091422.

Lafos, Marcel, Phillip Kroll, Mareike L. Hohenstatt, Frazer L. Thorpe, Oliver Clarenz, and Daniel Schubert. 2011. “Dynamic Regulation of H3K27 Trimethylation during Arabidopsis Differentiation.” *PLOS Genetics* 7 (4): e1002040. https://doi.org/10.1371/journal.pgen.1002040.

Lang, Daniel, Kristian K. Ullrich, Florent Murat, et al. 2018. “The Physcomitrella Patens Chromosome-Scale Assembly Reveals Moss Genome Structure and Evolution.” *The Plant Journal* 93 (3): 515–33. https://doi.org/10.1111/tpj.13801.

Lareau, Caleb A., Fabiana M. Duarte, Jennifer G. Chew, et al. 2019. “Droplet-Based Combinatorial Indexing for Massive-Scale Single-Cell Chromatin Accessibility.” *Nature Biotechnology* 37 (8): 916–24. https://doi.org/10.1038/s41587-019-0147-6.

Larsson, Anton J. M., Per Johnsson, Michael Hagemann-Jensen, et al. 2019. “Genomic Encoding of Transcriptional Burst Kinetics.” *Nature* 565 (7738): 251–54. https://doi.org/10.1038/s41586-018-0836-1.

Leduque, Basile, Alejandro Edera, Clémentine Vitte, and Leandro Quadrana. 2024. “Simultaneous Profiling of Chromatin Accessibility and DNA Methylation in Complete Plant Genomes Using Long-Read Sequencing.” *Nucleic Acids Research* 52 (11): 6285–97. https://doi.org/10.1093/nar/gkae306.

Li, Chunhui, Baocheng Sun, Yongxiang Li, et al. 2016. “Numerous Genetic Loci Identified for Drought Tolerance in the Maize Nested Association Mapping Populations.” *BMC Genomics* 17 (1): 894. https://doi.org/10.1186/s12864-016-3170-8.

Li, Miaomiao, Tao Yao, Wanru Lin, et al. 2023. “Double DAP-Seq Uncovered Synergistic DNA Binding of Interacting bZIP Transcription Factors.” *Nature Communications* 14 (1): 2600. https://doi.org/10.1038/s41467-023-38096-2.

Li, Qing, Jonathan I. Gent, Greg Zynda, et al. 2015. “RNA-Directed DNA Methylation Enforces Boundaries between Heterochromatin and Euchromatin in the Maize Genome.” *Proceedings of the National Academy of Sciences* 112 (47): 14728–33. https://doi.org/10.1073/pnas.1514680112.

Li, Xiang, and Robert J. Schmitz. 2025. “Cis-Regulatory Dynamics in Plant Domestication.” *Trends in Genetics* 41 (11): 984–94. https://doi.org/10.1016/j.tig.2025.02.005.

Li, Xueqin, C. Jake Harris, Zhenhui Zhong, et al. 2018. “Mechanistic Insights into Plant SUVH Family H3K9 Methyltransferases and Their Binding to Context-Biased Non-CG DNA Methylation.” *Proceedings of the National Academy of Sciences* 115 (37): E8793–802. https://doi.org/10.1073/pnas.1809841115.

Lieberman-Aiden, Erez, Nynke L. van Berkum, Louise Williams, et al. 2009. “Comprehensive Mapping of Long-Range Interactions Reveals Folding Principles of the Human Genome.” *Science* 326 (5950): 289–93. https://doi.org/10.1126/science.1181369.

Lin, Yuan, Fanli Meng, Chao Fang, Bo Zhu, and Jiming Jiang. 2019. “Rapid Validation of Transcriptional Enhancers Using Agrobacterium-Mediated Transient Assay.” *Plant Methods* 15 (March): 21. https://doi.org/10.1186/s13007-019-0407-y.

Lin, Zhongwei, Xianran Li, Laura M. Shannon, et al. 2012. “Parallel Domestication of the Shattering1 Genes in Cereals.” *Nature Genetics* 44 (6): 720–24. https://doi.org/10.1038/ng.2281.

Liu, Chang, Ying-Juan Cheng, Jia-Wei Wang, and Detlef Weigel. 2017. “Prominent Topologically Associated Domains Differentiate Global Chromatin Packing in Rice from Arabidopsis.” *Nature Plants* 3 (9): 742–48. https://doi.org/10.1038/s41477-017-0005-9.

Liu, Lijun, Yuhan Zhang, Chen Tang, Jine Wu, Jingye Fu, and Qiang Wang. 2024. “Genome-Wide Identification of ZmMYC2 Binding Sites and Target Genes in Maize.” *BMC Genomics* 25 (1): 397. https://doi.org/10.1186/s12864-024-10297-z.

Liu, Ruochen, Xiuping Zou, You Wang, Qin Long, and Yan Pei. 2020. “A 100 Bp GAGA Motif-Containing Sequence in AGAMOUS Second Intron Is Able to Suppress the Activity of CaMV35S Enhancer in Vegetative Tissues.” *PLoS ONE* 15 (3): e0230203. https://doi.org/10.1371/journal.pone.0230203.

Liu, Yue, Wenli Zhang, Kang Zhang, et al. 2017. “Genome-Wide Mapping of DNase I Hypersensitive Sites Reveals Chromatin Accessibility Changes in Arabidopsis Euchromatin and Heterochromatin Regions under Extended Darkness.” *Scientific Reports* 7 (1): 4093. https://doi.org/10.1038/s41598-017-04524-9.

Local, Andrea, Hui Huang, Claudio P. Albuquerque, et al. 2018. “Identification of H3K4me1-Associated Proteins at Mammalian Enhancers.” *Nature Genetics* 50 (1): 73–82. https://doi.org/10.1038/s41588-017-0015-6.

Lu, Zefu, Brigitte T. Hofmeister, Christopher Vollmers, Rebecca M. DuBois, and Robert J. Schmitz. 2017. “Combining ATAC-Seq with Nuclei Sorting for Discovery of Cis-Regulatory Regions in Plant Genomes.” *Nucleic Acids Research* 45 (6): e41. https://doi.org/10.1093/nar/gkw1179.

Lu, Zefu, Alexandre P. Marand, William A. Ricci, Christina L. Ethridge, Xiaoyu Zhang, and Robert J. Schmitz. 2019. “The Prevalence, Evolution and Chromatin Signatures of Plant Regulatory Elements.” *Nature Plants* 5 (12): 1250–59. https://doi.org/10.1038/s41477-019-0548-z.

Ma, Shaoqian, and Yongyou Zhang. 2020. “Profiling Chromatin Regulatory Landscape: Insights into the Development of ChIP-Seq and ATAC-Seq.” *Molecular Biomedicine* 1 (1): 9. https://doi.org/10.1186/s43556-020-00009-w.

Magoč, Tanja, and Steven L. Salzberg. 2011. “FLASH: Fast Length Adjustment of Short Reads to Improve Genome Assemblies.” *Bioinformatics* 27 (21): 2957–63. https://doi.org/10.1093/bioinformatics/btr507.

Maher, Kelsey A., Marko Bajic, Kaisa Kajala, et al. 2018. “Profiling of Accessible Chromatin Regions across Multiple Plant Species and Cell Types Reveals Common Gene Regulatory Principles and New Control Modules.” *The Plant Cell* 30 (1): 15–36. https://doi.org/10.1105/tpc.17.00581.

Marand, Alexandre P., Andrea L. Eveland, Kerstin Kaufmann, and Nathan M. Springer. 2023. “Cis-Regulatory Elements in Plant Development, Adaptation, and Evolution.” *Annual Review of Plant Biology* 74 (Volume 74, 2023): 111–37. https://doi.org/10.1146/annurev-arplant-070122-030236.

Marcon, Caroline, Lena Altrogge, Yan Naing Win, et al. 2020. “BonnMu: A Sequence-Indexed Resource of Transposon-Induced Maize Mutations for Functional Genomics Studies.” *Plant Physiology* 184 (2): 620–31. https://doi.org/10.1104/pp.20.00478.

Marks, Rose A., Scott Hotaling, Paul B. Frandsen, and Robert VanBuren. 2021. “Representation and Participation across 20 Years of Plant Genome Sequencing.” *Nature Plants* 7 (12): 1571–78. https://doi.org/10.1038/s41477-021-01031-8.

Mascher, Martin, Heidrun Gundlach, Axel Himmelbach, et al. 2017. “A Chromosome Conformation Capture Ordered Sequence of the Barley Genome.” *Nature* 544 (7651): 427–33. https://doi.org/10.1038/nature22043.

Mayran, Alexandre, and Jacques Drouin. 2018. “Pioneer Transcription Factors Shape the Epigenetic Landscape.” *The Journal of Biological Chemistry* 293 (36): 13795–804. https://doi.org/10.1074/jbc.R117.001232.

Mei, Wenbin, Markus G. Stetter, Daniel J. Gates, Michelle C. Stitzer, and Jeffrey Ross-Ibarra. 2018. “Adaptation in Plant Genomes: Bigger Is Different.” *American Journal of Botany* 105 (1): 16–19. https://doi.org/10.1002/ajb2.1002.

Mejía-Guerra, María Katherine, Wei Li, Narmer F. Galeano, et al. 2015. “Core Promoter Plasticity Between Maize Tissues and Genotypes Contrasts with Predominance of Sharp Transcription Initiation Sites.” *The Plant Cell* 27 (12): 3309–20. https://doi.org/10.1105/tpc.15.00630.

Mendieta, John Pablo, Ankush Sangra, Haidong Yan, Mark A. A. Minow, and Robert J. Schmitz. 2023. “Exploring Plant -Regulatory Elements at Single-Cell Resolution: Overcoming Biological and Computational Challenges to Advance Plant Research.” *The Plant Journal* 115 (6): 1486–99. https://doi.org/10.1111/tpj.16351.

Meng, Fanli, Hainan Zhao, Bo Zhu, et al. 2021. “Genomic Editing of Intronic Enhancers Unveils Their Role in Fine-Tuning Tissue-Specific Gene Expression in Arabidopsis Thaliana.” *The Plant Cell* 33 (6): 1997–2014. https://doi.org/10.1093/plcell/koab093.

Miller, Geneva, Lindsey M. Rollosson, Carrie Saada, Serenity J. Wade, and Danae Schulz. 2023. “Adaptation of CUT&RUN for Use in African Trypanosomes.” *PLOS ONE* 18 (11): e0292784. https://doi.org/10.1371/journal.pone.0292784.

Minh, Bui Quang, Heiko A. Schmidt, Olga Chernomor, et al. 2020. “IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era.” *Molecular Biology and Evolution* 37 (5): 1530–34. https://doi.org/10.1093/molbev/msaa015.

Moraru, Manuela, and Thomas Schalch. 2019. “Chromatin Fiber Structural Motifs as Regulatory Hubs of Genome Function?” *Essays in Biochemistry* 63 (1): 123–32. https://doi.org/10.1042/EBC20180065.

Morton, Taj, Jalean Petricka, David L. Corcoran, et al. 2014. “Paired-End Analysis of Transcription Start Sites in Arabidopsis Reveals Plant-Specific Promoter Signatures[C][W].” *The Plant Cell* 26 (7): 2746–60. https://doi.org/10.1105/tpc.114.125617.

Murphy, Kristin E., Fanju W. Meng, Claire E. Makowski, and Patrick J. Murphy. 2020. “Genome-Wide Chromatin Accessibility Is Restricted by ANP32E.” *Nature Communications* 11 (October): 5063. https://doi.org/10.1038/s41467-020-18821-x.

Myers, Zachary A., Clair M. Wootan, Zhikai Liang, et al. 2023. “Conserved and Variable Heat Stress Responses of the Heat Shock Factor Transcription Factor Family in Maize and Setaria Viridis.” *Plant Direct* 7 (4): e489. https://doi.org/10.1002/pld3.489.

Nakamura, Miyuki, Rita A. Batista, Claudia Köhler, and Lars Hennig. 2020. “Polycomb Repressive Complex 2-Mediated Histone Modification H3K27me3 Is Associated with Embryogenic Potential in Norway Spruce.” *Journal of Experimental Botany* 71 (20): 6366–78. https://doi.org/10.1093/jxb/eraa365.

Nannas, Natalie J., and R. Kelly Dawe. 2015. “Genetic and Genomic Toolbox of Zea Mays.” *Genetics* 199 (3): 655–69. https://doi.org/10.1534/genetics.114.165183.

Ngan, Chew Yee, Chee Hong Wong, Harianto Tjong, et al. 2020. “Chromatin Interaction Analyses Elucidate the Roles of PRC2-Bound Silencers in Mouse Development.” *Nature Genetics* 52 (3): 264–72. https://doi.org/10.1038/s41588-020-0581-x.

Ni, Ting, David L. Corcoran, Elizabeth A. Rach, et al. 2010. “A Paired-End Sequencing Strategy to Map the Complex Landscape of Transcription Initiation.” *Nature Methods* 7 (7): 521–27. https://doi.org/10.1038/nmeth.1464.

Nolan, Trevor M., Nemanja Vukašinović, Derui Liu, Eugenia Russinova, and Yanhai Yin. 2020. “Brassinosteroids: Multidimensional Regulators of Plant Growth, Development, and Stress Responses[OPEN].” *The Plant Cell* 32 (2): 295–318. https://doi.org/10.1105/tpc.19.00335.

Nordin, Anna, Gianluca Zambanini, Pierfrancesco Pagella, and Claudio Cantù. 2023. “The CUT&RUN Suspect List of Problematic Regions of the Genome.” *Genome Biology* 24 (August): 185. https://doi.org/10.1186/s13059-023-03027-3.

Oh, Eunkyoo, Jia-Ying Zhu, Ming-Yi Bai, Rafael Augusto Arenhart, Yu Sun, and Zhi-Yong Wang. 2014. “Cell Elongation Is Regulated through a Central Circuit of Interacting Transcription Factors in the Arabidopsis Hypocotyl.” *eLife* 3 (May): e03031. https://doi.org/10.7554/eLife.03031.

Oh, Eunkyoo, Jia-Ying Zhu, Hojin Ryu, Ildoo Hwang, and Zhi-Yong Wang. 2014. “TOPLESS Mediates Brassinosteroid-Induced Transcriptional Repression through Interaction with BZR1.” *Nature Communications* 5 (1): 4140. https://doi.org/10.1038/ncomms5140.

Oh, Eunkyoo, Jia-Ying Zhu, and Zhi-Yong Wang. 2012. “Interaction between BZR1 and PIF4 Integrates Brassinosteroid and Environmental Responses.” *Nature Cell Biology* 14 (8): 802–9. https://doi.org/10.1038/ncb2545.

Oki, Masaya, and Rohinton T. Kamakaka. 2002. “Blockers and Barriers to Transcription: Competing Activities?” *Current Opinion in Cell Biology* 14 (3): 299–304. https://doi.org/10.1016/S0955-0674(02)00327-7.

O'Malley, Ronan C., Shao-shan Carol Huang, Liang Song, et al. 2016. "Cistrome and Epicistrome Features Shape the Regulatory DNA Landscape." *Cell* 165 (5): 1280–92. https://doi.org/10.1016/j.cell.2016.04.038.

Palazzo, Alexander F., and T. Ryan Gregory. 2014. "The Case for Junk DNA." *PLoS Genetics* 10 (5): e1004351. https://doi.org/10.1371/journal.pgen.1004351.

Pang, Baoxu, and Michael P. Snyder. 2020. "Systematic Identification of Silencers in Human Cells." *Nature Genetics* 52 (3): 254–63. https://doi.org/10.1038/s41588-020-0578-5.

Partridge, E. Christopher, Surya B. Chhetri, Jeremy W. Prokop, et al. 2020. "Occupancy Maps of 208 Chromatin-Associated Proteins in One Human Cell Type." *Nature* 583 (7818): 720–28. https://doi.org/10.1038/s41586-020-2023-4.

Parvathaneni, Rajiv K., Edoardo Bertolini, Md Shamimuzzaman, et al. 2020. "The Regulatory Landscape of Early Maize Inflorescence Development." *Genome Biology* 21 (1): 165. https://doi.org/10.1186/s13059-020-02070-8.

Patel-Tupper, Dhruv, Armen Kelikian, Anna Leipertz, et al. 2024. "Multiplexed CRISPR-Cas9 Mutagenesis of Rice PSBS1 Noncoding Sequences for Transgene-Free Overexpression." *Science Advances* 10 (23): eadm7452. https://doi.org/10.1126/sciadv.adm7452.

Phillips, Alyssa R., Taylor AuBuchon-Elder, Kerrie Barry, et al. 2025. "The Origins and Adaptive Consequences of Polyploidy in a Dominant Prairie Grass." Preprint, bioRxiv, November 25. https://doi.org/10.1101/2025.11.25.690567.

Planas-Riverola, Ainoa, Aditi Gupta, Isabel Betegón-Putze, Nadja Bosch, Marta Ibañes, and Ana I. Caño-Delgado. 2019. "Brassinosteroid Signaling in Plant Development and Adaptation to Stress." *Development (Cambridge, England)* 146 (5): dev151894. https://doi.org/10.1242/dev.151894.

Plotecki, Felix. 2023. "Diversity of Cis-Regulatory Elements in Arabidopsis Thaliana and Zea Mays L." Heinrich-Heine-Universität Düsseldorf.

Porto, Milena Silva, Morganna Pollynne Nóbrega Pinheiro, Vandré Guevara Lyra Batista, Roseane Cavalcanti dos Santos, Péricles de Albuquerque Melo Filho, and Liziane Maria de Lima. 2014. "Plant Promoters: An Approach of Structure and Function." *Molecular Biotechnology* 56 (1): 38–49. https://doi.org/10.1007/s12033-013-9713-1.

Rajkumar, Mohan Singh, Kunal Tembhare, Rohini Garg, and Mukesh Jain. 2024. "Genome-Wide Mapping of DNase I Hypersensitive Sites Revealed Differential Chromatin Accessibility and Regulatory DNA Elements under Drought Stress in Rice Cultivars." *The Plant Journal* 119 (4): 2063–79. https://doi.org/10.1111/tpj.16864.

Ricci, William A., Zefu Lu, Lexiang Ji, et al. 2019. "Widespread Long-Range Cis-Regulatory Elements in the Maize Genome." *Nature Plants* 5 (12): 1237–49. https://doi.org/10.1038/s41477-019-0547-0.

Rodgers-Melnick, Eli, Daniel L. Vera, Hank W. Bass, and Edward S. Buckler. 2016. “Open Chromatin Reveals the Functional Maize Genome.” *Proceedings of the National Academy of Sciences* 113 (22): E3177–84. https://doi.org/10.1073/pnas.1525244113.

Rodríguez-Leal, Daniel, Zachary H. Lemmon, Jarrett Man, Madelaine E. Bartlett, and Zachary B. Lippman. 2017. “Engineering Quantitative Trait Variation for Crop Improvement by Genome Editing.” *Cell* 171 (2): 470-480.e8. https://doi.org/10.1016/j.cell.2017.08.030.

Rushton, Paul J., Imre E. Somssich, Patricia Ringler, and Qingxi J. Shen. 2010. “WRKY Transcription Factors.” *Trends in Plant Science* 15 (5): 247–58. https://doi.org/10.1016/j.tplants.2010.02.006.

Savadel, Savannah D., Thomas Hartwig, Zachary M. Turpin, et al. 2021. “The Native Cistrome and Sequence Motif Families of the Maize Ear.” *PLOS Genetics* 17 (8): e1009689. https://doi.org/10.1371/journal.pgen.1009689.

Schmitz, Robert J., Erich Grotewold, and Maike Stam. 2021. “Cis-Regulatory Sequences in Plants: Their Importance, Discovery, and Future Challenges.” *The Plant Cell* 34 (2): 718–41. https://doi.org/10.1093/plcell/koab281.

Schnable, Patrick S., Doreen Ware, Robert S. Fulton, et al. 2009. “The B73 Maize Genome: Complexity, Diversity, and Dynamics.” *Science* 326 (5956): 1112–15. https://doi.org/10.1126/science.1178534.

Searle, Iain, Yuehui He, Franziska Turck, et al. 2006. “The Transcription Factor FLC Confers a Flowering Response to Vernalization by Repressing Meristem Competence and Systemic Signaling in Arabidopsis.” *Genes & Development* 20 (7): 898–912. https://doi.org/10.1101/gad.373506.

Sieburth, L. E., and E. M. Meyerowitz. 1997. “Molecular Dissection of the AGAMOUS Control Region Shows That Cis Elements for Spatial Regulation Are Located Intragenically.” *The Plant Cell* 9 (3): 355–65. https://doi.org/10.1105/tpc.9.3.355.

Skene, Peter J., and Steven Henikoff. 2017. “An Efficient Targeted Nuclease Strategy for High-Resolution Mapping of DNA Binding Sites.” *eLife* 6 (January): e21856. https://doi.org/10.7554/eLife.21856.

Smith, Stephen A., and Brian C. O’Meara. 2012. “treePL: Divergence Time Estimation Using Penalized Likelihood for Large Phylogenies.” *Bioinformatics* 28 (20): 2689–90. https://doi.org/10.1093/bioinformatics/bts492.

Solomon, Mark J., Pamela L. Larsen, and Alexander Varshavsky. 1988. “Mapping proteinDNA Interactions in Vivo with Formaldehyde: Evidence That Histone H4 Is Retained on a Highly Transcribed Gene.” *Cell* 53 (6): 937–47. https://doi.org/10.1016/S0092-8674(88)90469-2.

Song, Baoxing, Edward S. Buckler, Hai Wang, et al. 2021. “Conserved Noncoding Sequences Provide Insights into Regulatory Sequence and Loss of Gene Expression in Maize.” *Genome Research* 31 (7): 1245–57. https://doi.org/10.1101/gr.266528.120.

Stergachis, Andrew B., Brian M. Debo, Eric Haugen, L. Stirling Churchman, and John A. Stamatoyannopoulos. 2020. “Single-Molecule Regulatory Architectures Captured by Chromatin Fiber Sequencing.” *Science* 368 (6498): 1449–54. https://doi.org/10.1126/science.aaz1646.

Stitzer, Michelle C., Arun S. Seetharam, Armin Scheben, et al. 2025. “Extensive Genome Evolution Distinguishes Maize within a Stable Tribe of Grasses.” Preprint, bioRxiv, January 24. https://doi.org/10.1101/2025.01.22.633974.

Studer, Anthony, Qiong Zhao, Jeffrey Ross-Ibarra, and John Doebley. 2011. “Identification of a Functional Transposon Insertion in the Maize Domestication Gene Tb1.” *Nature Genetics* 43 (11): 1160–63. https://doi.org/10.1038/ng.942.

Sturm, Marc, Christopher Schroeder, and Peter Bauer. 2016. “SeqPurge: Highly-Sensitive Adapter Trimming for Paired-End NGS Data.” *BMC Bioinformatics* 17 (1): 208. https://doi.org/10.1186/s12859-016-1069-7.

Sura, Weronika, Michał Kabza, Wojciech M. Karlowski, et al. 2017. “Dual Role of the Histone Variant H2A.Z in Transcriptional Regulation of Stress-Response Genes.” *The Plant Cell* 29 (4): 791–807. https://doi.org/10.1105/tpc.16.00573.

Szabo, Quentin, Frédéric Bantignies, and Giacomo Cavalli. 2019. “Principles of Genome Folding into Topologically Associating Domains.” *Science Advances* 5 (4): eaaw1668. https://doi.org/10.1126/sciadv.aaw1668.

Takahashi, Hazuki, Sachiko Kato, Mitsuyoshi Murata, and Piero Carninci. 2012. “CAGE-Cap Analysis Gene Expression: A Protocol for the Detection of Promoter and Transcriptional Networks.” *Methods in Molecular Biology (Clifton, N.J.)* 786: 181–200. https://doi.org/10.1007/978-1-61779-292-2_11.

Talbert, Paul B., and Steven Henikoff. 2010. “Histone Variants — Ancient Wrap Artists of the Epigenome.” *Nature Reviews Molecular Cell Biology* 11 (4): 264–75. https://doi.org/10.1038/nrm2861.

Tang, Yao, and Pengguo Xia. 2025. “WRKY Transcription Factors: Key Regulators in Plant Drought Tolerance.” *Plant Science* 359 (October): 112647. https://doi.org/10.1016/j.plantsci.2025.112647.

The International Wheat Genome Sequencing Consortium (IWGSC), Rudi Appels, Kellye Eversole, et al. 2018. “Shifting the Limits in Wheat Research and Breeding Using a Fully Annotated Reference Genome.” *Science* 361 (6403): eaar7191. https://doi.org/10.1126/science.aar7191.

Trapp, Thimo. 2024. “Comparative Analysis of Transcription Factor Binding Sites across Plant Genomes with Different Levels of Complexity.” Heinrich-Heine-Universität Düsseldorf.

Tsompana, Maria, and Michael J. Buck. 2014. “Chromatin Accessibility: A Window into the Genome.” *Epigenetics & Chromatin* 7 (1): 33. https://doi.org/10.1186/1756-8935-7-33.

Tu, Xiaoyu, Alexandre P. Marand, Robert J. Schmitz, and Silin Zhong. 2022. “A Combinatorial Indexing Strategy for Low-Cost Epigenomic Profiling of Plant Single Cells.” *Plant Communications*, Focus Issue on Bioinformatic Tools & Databases, vol. 3 (4): 100308. https://doi.org/10.1016/j.xplc.2022.100308.

Tu, Xiaoyu, María Katherine Mejía-Guerra, Jose A. Valdes Franco, et al. 2020. “Reconstructing the Maize Leaf Regulatory Network Using ChIP-Seq Data of 104 Transcription Factors.” *Nature Communications* 11 (1): 5089. https://doi.org/10.1038/s41467-020-18832-8.

Vann, Laura, Thomas Kono, Tanja Pyhäjärvi, Matthew B. Hufford, and Jeffrey Ross-Ibarra. 2015. “Natural Variation in Teosinte at the Domestication Locus Teosinte Branched1 (Tb1).” *PeerJ* 3 (April): e900. https://doi.org/10.7717/peerj.900.

Vicentini, Alberto, Janet C. Barber, Sandra S. Aliscioni, Liliana M. Giussani, and Elizabeth A. Kellogg. 2008. “The Age of the Grasses and Clusters of Origins of C4 Photosynthesis.” *Global Change Biology* 14 (12): 2963–77. https://doi.org/10.1111/j.1365-2486.2008.01688.x.

Wang, Maojun, Pengcheng Wang, Min Lin, et al. 2018. “Evolutionary Dynamics of 3D Genome Architecture Following Polyploidization in Cotton.” *Nature Plants* 4 (2): 90–97. https://doi.org/10.1038/s41477-017-0096-3.

Wang, Yang, Wangshu Li, Jianzhou Qu, Fenghai Li, Wanli Du, and Jianfeng Weng. 2023. “Genome-Wide Characterization of the Maize (Zea Mays L.) WRKY Transcription Factor Family and Their Responses to Ustilago Maydis.” *International Journal of Molecular Sciences* 24 (19): 14916. https://doi.org/10.3390/ijms241914916.

Watson, J. D., and F. H. C. Crick. 1953. “Molecular Structure of Nucleic Acids: A Structure for Deoxyribose Nucleic Acid.” *Nature* 171 (4356): 737–38. https://doi.org/10.1038/171737a0.

Weirauch, Matthew T., and Timothy R. Hughes. 2010. “Conserved Expression without Conserved Regulatory Sequence: The More Things Change, the More They Stay the Same.” *Trends in Genetics* 26 (2): 66–74. https://doi.org/10.1016/j.tig.2009.12.002.

Welker, Cassiano A. D., Michael R. McKain, Matt C. Estep, et al. 2020. “Phylogenomics Enables Biogeographic Analysis and a New Subtribal Classification of Andropogoneae (Poaceae—Panicoideae).” *Journal of Systematics and Evolution* 58 (6): 1003–30. https://doi.org/10.1111/jse.12691.

Wernig-Zorc, Sara, Fabian Kugler, Leo Schmutterer, et al. 2024. “nucMACC: An MNase-Seq Pipeline to Identify Structurally Altered Nucleosomes in the Genome.” *Science Advances*, ahead of print, July 5. World. https://doi.org/10.1126/sciadv.adm9740.

Womersley, Howard J., Daniel Muliaditan, Ramanuj DasGupta, and Lih Feng Cheow. 2025. “Single-Nucleus CUT&RUN Elucidates the Function of Intrinsic and Genomics-Driven Epigenetic Heterogeneity in Head and Neck Cancer Progression.” *Genome Research* 35 (1): 162–77. https://doi.org/10.1101/gr.279105.124.

Xiao, Jun, Run Jin, Xiang Yu, et al. 2017. “Cis and Trans Determinants of Epigenetic Silencing by Polycomb Repressive Complex 2 in Arabidopsis.” *Nature Genetics* 49 (10): 1546–52. https://doi.org/10.1038/ng.3937.

Xie, Zhouli, Trevor Nolan, Hao Jiang, et al. 2019. “The AP2/ERF Transcription Factor TINY Modulates Brassinosteroid-Regulated Plant Growth and Drought Responses in Arabidopsis.” *The Plant Cell* 31 (8): 1788–806. https://doi.org/10.1105/tpc.18.00918.

Yamamoto, Yoshiharu Y., Hiroyuki Ichida, Minami Matsui, et al. 2007. “Identification of Plant Promoter Constituents by Analysis of Local Distribution of Short Sequences.” *BMC Genomics* 8 (1): 67. https://doi.org/10.1186/1471-2164-8-67.

Yan, Haidong, John P. Mendieta, Xuan Zhang, et al. 2025. “A Single-Cell Rice Atlas Integrates Multi-Species Data to Reveal Cis-Regulatory Evolution.” *Nature Plants* 11 (10): 2050–71. https://doi.org/10.1038/s41477-025-02106-6.

Yan, Wenhao, Dijun Chen, Julia Schumacher, et al. 2019. “Dynamic Control of Enhancer Activity Drives Stage-Specific Gene Expression during Flower Morphogenesis.” *Nature Communications* 10 (1): 1705. https://doi.org/10.1038/s41467-019-09513-2.

Yang, Hongji, Guorong Yu, Zhidong Lv, et al. 2025. “Epigenome Profiling Reveals Distinctive Regulatory Features and Cis-Regulatory Elements in Pepper.” *Genome Biology* 26 (1): 121. https://doi.org/10.1186/s13059-025-03595-6.

Yelagandula, Ramesh, Hume Stroud, Sarah Holec, et al. 2014. “The Histone Variant H2A.W Defines Heterochromatin and Promotes Chromatin Condensation in Arabidopsis.” *Cell* 158 (1): 98–109. https://doi.org/10.1016/j.cell.2014.06.006.

Yu, Xiaole, Zhixin Liu, and Xuwu Sun. 2023. “Single-Cell and Spatial Multi-Omics in the Plant Sciences: Technical Advances, Applications, and Perspectives.” *Plant Communications* 4 (3): 100508. https://doi.org/10.1016/j.xplc.2022.100508.

Zaret, Kenneth S., and Jason S. Carroll. 2011. “Pioneer Transcription Factors: Establishing Competence for Gene Expression.” *Genes & Development* 25 (21): 2227–41. https://doi.org/10.1101/gad.176826.111.

Zebosi, Brian, Erik Vollbrecht, and Norman B. Best. 2024. “Brassinosteroid Biosynthesis and Signaling: Conserved and Diversified Functions of Core Genes across Multiple Plant Species.” *Plant Communications* 5 (9): 100982. https://doi.org/10.1016/j.xplc.2024.100982.

Zhang, Chao, Rasmus Nielsen, and Siavash Mirarab. 2025. “ASTER: A Package for Large-Scale Phylogenomic Reconstructions.” *Molecular Biology and Evolution*, July 16, msaf172. https://doi.org/10.1093/molbev/msaf172.

Zhang, Hui, Ruiqin Zheng, Yunlong Wang, et al. 2019. “The Effects of Arabidopsis Genome Duplication on the Chromatin Organization and Transcriptional Regulation.” *Nucleic Acids Research* 47 (15): 7857–69. https://doi.org/10.1093/nar/gkz511.

Zhang, Huiming, Zhaobo Lang, and Jian-Kang Zhu. 2018. “Dynamics and Function of DNA Methylation in Plants.” *Nature Reviews Molecular Cell Biology* 19 (8): 489–506. https://doi.org/10.1038/s41580-018-0016-z.

Zhang, Junlin, Bao Liu, Yan Song, et al. 2021. “Genome-Wide (ChIP-Seq) Identification of Target Genes Regulated by WRKY33 during Submergence Stress in Arabidopsis.” *BMC Genomic Data* 22 (1): 16. https://doi.org/10.1186/s12863-021-00972-5.

Zhang, Tao, Wenli Zhang, and Jiming Jiang. 2015. “Genome-Wide Nucleosome Occupancy and Positioning and Their Impact on Gene Expression and Evolution in Plants.” *Plant Physiology* 168 (4): 1406–16. https://doi.org/10.1104/pp.15.00125.

Zhang, Wenli, Tao Zhang, Yufeng Wu, and Jiming Jiang. 2012. “Genome-Wide Identification of Regulatory DNA Elements and Protein-Binding Footprints Using Signatures of Open Chromatin in Arabidopsis.” *The Plant Cell* 24 (7): 2719–31. https://doi.org/10.1105/tpc.112.098061.

Zhang, Yong, Tao Liu, Clifford A. Meyer, et al. 2008. “Model-Based Analysis of ChIP-Seq (MACS).” *Genome Biology* 9 (9): R137. https://doi.org/10.1186/gb-2008-9-9-r137.

Zhao, Lun, Shuangqi Wang, Zhilin Cao, et al. 2019. “Chromatin Loops Associated with Active Genes and Heterochromatin Shape Rice Genome Architecture for Transcriptional Regulation.” *Nature Communications* 10 (1): 3640. https://doi.org/10.1038/s41467-019-11535-9.

Zhao, Shuai, Lingling Cheng, Yifei Gao, et al. 2019. “Plant HP1 Protein ADCP1 Links Multivalent H3K9 Methylation Readout to Heterochromatin Formation.” *Cell Research* 29 (1): 54–66. https://doi.org/10.1038/s41422-018-0104-9.

Zhu, Bo, Wenli Zhang, Tao Zhang, Bao Liu, and Jiming Jiang. 2015. “Genome-Wide Prediction and Validation of Intergenic Enhancers in Arabidopsis Using Open Chromatin Signatures.” *The Plant Cell* 27 (9): 2415–26. https://doi.org/10.1105/tpc.15.00537.

Zhu, Xiaobiao, Airu Chen, Nathaniel M. Butler, et al. 2024. “Molecular Dissection of an Intronic Enhancer Governing Cold-Induced Expression of the Vacuolar Invertase Gene in Potato.” *The Plant Cell* 36 (5): 1985–99. https://doi.org/10.1093/plcell/koae050.