COMPARATIVE PLANT EPIGENOMICS: DRIVING INSIGHTS INTO THE EVOLUTION OF REGULATORY REGIONS

by

JOHN PAUL MENDIETA

(Under the Direction of Robert Schmitz)

ABSTRACT

The epigenome consists of proteins and DNA modifications that influence how genetic information is used and allow organisms to have varied gene expression across different cells and tissues despite identical DNA sequences. This work leverages epigenomic data to enhance our understanding of plant genomes.

First, we employed Chromatin Immunoprecipitation sequencing to identify transcriptional units in plant genomes, focusing on *Zea mays*. We identified genomic regions with two types histone modifications: one that canonically indicates active transcription throughout a gene body and another that is indicative of transcriptional start sites. We utilized these two epigenomic marks in tandem to detect transcriptional units across the genome. While many of these regions corresponded to known protein-coding genes, we also discovered new regions of transcription distant from any known gene annotations. We then leveraged this data to identify and rectify incorrectly annotated genes, which were either fractured or missing their transcription start site. We then applied this method across multiple species, revealing widespread annotation errors across numerous plant genomes.

We then explore plant epigenomics at a single-cell resolution using single-cell indexed Assay for Transposase Accessible Chromatin (sciATAC-seq), a technique that identifies open regions of a genome. These nucleosome-devoid regions often contain regulatory sequences critical for modulating gene expression. sciATAC-seq was performed on leaf tissue from five different plant species: *Zea mays, Sorghum bicolor, Panicum Miliaceum, Urochlua fusca,* and *Oryza sativa*. Using these data sets, we developed novel methods to accurately annotate cell types across diverse plant species. Furthermore, we investigate the regulatory regions associated with C4 photosynthesis—a more efficient process in hot, arid climates compared to C3 photosynthesis, which is more common. By generating genome-wide maps of chromatin regulatory regions, we identify potentially crucial transcriptional regulators of of C4 photosynthesis genes. We then use this data to explore the evolution of regulatory regions. We find differing levels of conservation, with gene regulatory networks associated with specific cell-types conserved as indicated by transcription factor binding motif enrichment. But massive turnover of sequence at genomic loci, indicating that conservation of cell-type-specificity is happening at different levels in the genome.

Together, this work develops novel tools for plant genome informatics and furthers the field of plant epigenomics.

INDEX WORDS: cis-regulatory elements, single-cell, ATAC-seq, plant genomics, functional genomics, C4 photosynthesis

COMPARATIVE PLANT EPIGENOMICS: DRIVING INSIGHTS INTO THE EVOLUTION OF REGULATORY REGIONS

by

JOHN PAUL MENDIETA

B.A, University of Colorado Boulder, 2016

A Dissertation Submitted to the Graduate Faculty of The University of Georgia in Partial Fulfillment of the Requirements for the Degree

DOCTOR OF PHILOSOPHY

ATHENS, GEORGIA

2024

COMPARATIVE PLANT EPIGENOMICS: DRIVING INSIGHTS INTO THE EVOLUTION OF REGULATORY REGIONS

by

JOHN PAUL MENDIETA

Major Professor:	Robert J. Schmitz
Committee:	Casey Bergman
	Kelly Dawe
	Zachary Lewis
	Doug Menke

Electronic Version Approved:

Ron Walcott
Vice Provost for Graduate Education and Dean of the Graduate School
The University of Georgia
May 2024

DEDICATION

To the writers who made me want to forge the future.

ACKNOWLEDGEMENTS

This PhD was an absolute journey. It took my six years, thousands of lines of code, innumerous shots of espresso, and a couple of cigars. which all felt well earned. It has been the most challenging thing I've ever done in my life. But also the thing that I am to date the most proud of. However it wouldn't have been possible without a community that supported me through it all.

To my parents Donna and Paul, I owe you everything. For fostering my love of science and supporting me through all the choices I've made in my life. My passion for this work and the world around me is in no small part due to watching you both. I'm deeply proud to be your son. Te quiero mucho.

Bob Schmitz has been a tremendous advisor who has been honest, direct, and overall transparent about how academia works. I owe him greatly for this. His thoughtfulness and care for the people he mentors and employs is remarkable. Being mentored by Bob who is as passionate as I am about hard questions and interesting results has been a delight. These days I'm grateful to consider him a mentor and a friend. Also, *zea Maize*.

The Culturrati (Noah, Sam, Kivanc) has been an amazing group of friends during a deeply long and hard process. Talking about all things both science and life related, hanging out with you guys has been a real joy during graduate school experience, and I look forward to our continued friendship.

Katie Duval has been the kind of friend that everyone dreams of making in graduate school. Someone who's passion and love for the natural world is equal to my own. I wouldn't be

the scientist or person I am today without her. Watching her grow as a scientist and person during graduate school has been one of my greatest joys. The work found in this dissertation wouldn't have been possible without her.

Thank you to the GACRC and all the research support staff UGA. Without their constant and thankless work, the scope of my research would be nowhere near it is today. I owe them all a tremendous debt.

Finally, to the innumerous friends who aren't mentioned by name. Thank you. You've all made my time here at UGA so special. I will always cherish my time here in Athens as some of the most fun in my life. Thank you all.

TABLE OF CONTENTS

LIST OF FIGURES

Page

Chapter 1

Introduction and Literature Review

Exploring Plant Cis-Regulatory Elements at Single-Cell Resolution: Overcoming Biological and Computational Challenges to Advance Plant Research[1]

[1] Mendieta, John Pablo. Accepted by the plant journal. Reprinted here with the permission of publisher, 3/14/2024

Abstract:

Cis-regulatory elements (CREs) are important sequences for gene expression and for plant biological processes such as development, evolution, domestication, and stress response. However, studying CREs in plant genomes has been challenging. The totipotent nature of plant cells, coupled with inability to maintain plant cell types in culture and the inherent technical challenges posed by the cell wall have limited our understanding of how plant cell types acquire and maintain their identities and respond to the environment via CRE usage. Advances in single-cell epigenomics have revolutionized the field identifying cell-type-specific CREs. These new technologies have the potential to significantly advance our understanding of plant CRE biology, and shed light on how the regulatory genome gives rise to diverse plant phenomena. However, there are significant biological and computational challenges associated with analyzing single-cell epigenomic datasets. In this review, we discuss the historical and foundational underpinnings of plant single-cell research, challenges and common pitfalls in analysis of plant single-cell epigenomic data, and highlight biological challenges unique to plants. Additionally, we discuss how the application of single-cell epigenomic data in various contexts stands to transform our understanding of the importance of CREs in plant genomes.

Introduction:

Multicellular eukaryotes arose due to evolutionary pressures driving the sub-functionalization of cells into dedicated roles, allowing organisms to have functions that are more advanced than its cellular components. Cellular specialization results from

differential use of the genome between cell types, which is partly driven by variable use of *cis*-regulatory elements (CREs) that are important for gene transcription and silencing. In plants, cell types have evolved specialized metabolisms and unique cell wall morphologies that link form to function enabling cells to fill their structural and physiological roles *in planta* (Alberts et al., 2002). Plant cell structures fascinated early plant scientists; the observation of microscopic cell wall 'cages' within onion leaves led to Robert Hooke to develop the term 'cell' and variable cell wall morphologies were first used to classify plant cell types (Hooke, 1665). However, plant cell-type definitions have been continually refined by sequential scientific breakthroughs such as the increased resolution of microscopy and advances in molecular genetic techniques. Advances in single-cell genomics allow measurement of cell-type-specific transcripts and CRE chromatin accessibility, which is bettering understanding of the gene regulatory networks present in cells, and how they impact all manner of phenomena in *planta*. However, there are numerous technical and biological challenges associated with single-cell genomics data that must be overcome before these questions can be addressed.

In this perspective, we discuss the historical ways plant cell types have been described and how cell-type definitions have evolved. We examine how cell types have been defined genetically, and how this identified marker genes critical for cell-type function. We discuss current biological and technical challenges associated with the single-cell genomics identification of cell-type-specific regulatory sequences. Lastly, we highlight how emerging technologies will overcome some of these challenges, improving

the ability to study the cellular context in which molecular processes affect plant phenotypes.

Plant Cell Types - A Historical Perspective:

Scientists have been describing the cellular make-up of plants for centuries. Plant cell biology began with the advent of microscopy and histology, with early descriptions of stomata and guard cells dating back to 1671 (*Anatomie des plantes*, n.d.). This research laid the groundwork of plant anatomy and established early models of plant cell types. Cells were classified based on the structure of their cell walls, with parenchyma having thin non-lignified cell walls, collenchyma having thick non-lignified cell walls, and sclerenchyma having lignified cell walls (Imperatorskai͡a akademīi͡a nauk (Russia) et al., 1868). Although critical, these early descriptions of plant “cell types" had limited resolution and overlooked cells with unique structure and function. Advances in microscopy in the following centuries facilitated more accurate descriptions of plant cell types. Increasing microscopy resolution produced descriptions of cell-type subclasses within the classical definitions of parenchyma, collenchyma and sclerenchyma (Leroux, 2012). This led to the first described companion cells, sieve tube elements, and bundle sheath cells (Strasburger, 1888; Wilhelm, 1880). These newly described cell types were not just categorized but were described in their developmental and gross anatomical contexts within the plant. Foundational work by Esau and Sharman combined microscopy with serial sectioning experiments, describing vascular development in multiple plant species (Esau, 1939, 1954; Sharman, 1942). This combination of techniques revealed the cellular patterns in mature tissue, and how these arrangements emerge from their cellular

precursors (Esau, 1943; Sharman, 1942). Further work focused on the meristem, a collection of plant stem cells that divide to produce new growth. Tracking cellular division and maturation from meristems provided an early understanding of plant cell-type differentiation, revealing how anatomical patterns are established by development (Evert et al., 2006).

Early on it was understood that DNA encoded the genetic instructions which give rise to plant form, but our understanding the genetic processes that controlled cell-fate decisions were limited. Initial genetic analysis exploited the clonal development of mutant sectors with visible phenotypes. In brief, these studies used mutagens, like X-rays, to induce somatic mutations in progenitor cells to determine the cells contributions to organismal phenotype. In plants, mutant based studies demonstrated that manipulation of DNA sequence could radically change plant phenotypes and cell fates(Hake & Freeling, 1986; Sinha & Hake, 1990). For instance, stable mutagenesis gave rise to *liguleless-1* mutants that have radically different leaf morphology with a misplaced ligule on the margin of the leaf blade (Becraft & Freeling, 1991). However, these studies were limited in their capacity to identify the sequence causing these morphological alterations. This changed with advances in molecular genetic techniques that allowed for pinpointed manipulation of plant DNA.

In the 1990's molecular genetic techniques allowed fine scale alteration of DNA and inquiry into the genetic processes driving the emergence of specific cell types. Early genetic screens found that cell identity could be ablated by single-gene knockouts. One excellent example is *shortroot* (*shr)* in *Arabidopsis thaliana* (Benfey et al., 1993). In *shr* mutant plants, root endodermis cells fail to form, resulting in significantly stunted root

growth and illustrating that *SHR* is indispensable for endodermis cell-type identity. Further analysis of *SHR* revealed it is a mobile transcription factor critical for cell fate differentiation (Helariutta et al., 2000). The identification of *SHR,* and and other transcription factors that defined cell identity generated questions aimed at how cell fates were encoded within the genome. These questions remain the subject of active investigation, with ongoing experiments continually offering deeper insights into the molecular events that drive plant cell-fate decisions.

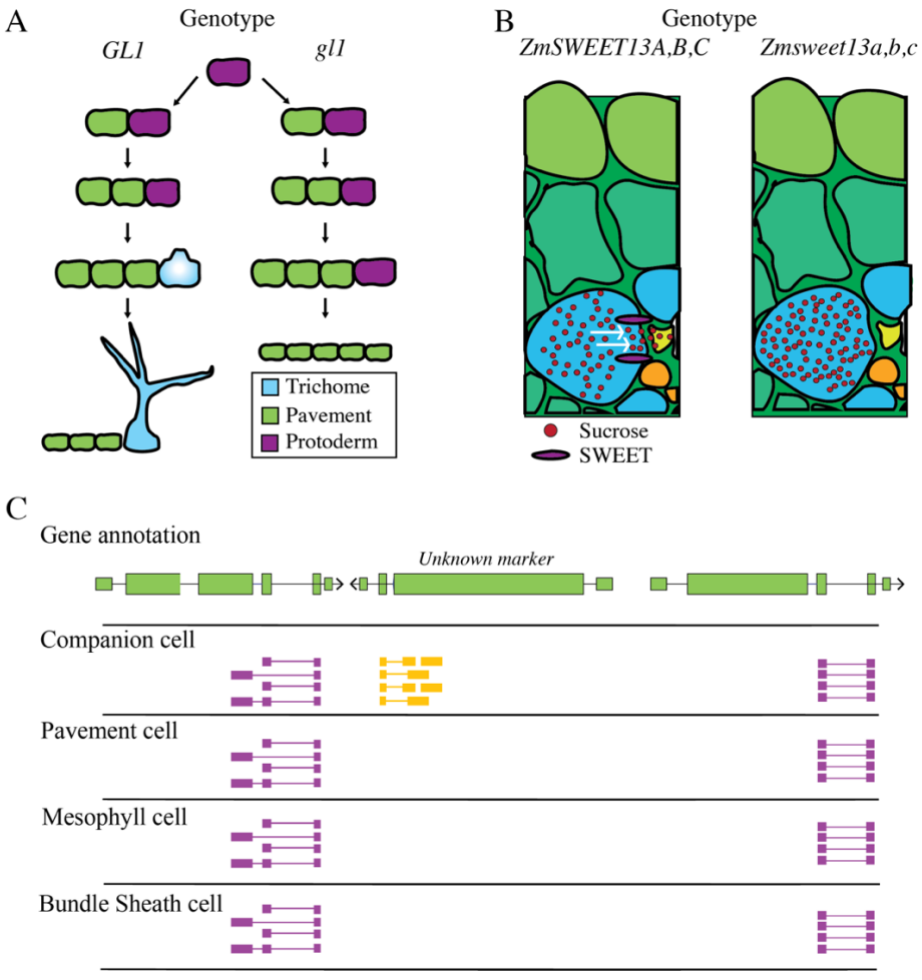

Figure 1.1: Plant cell-type markers define either unique developmental, metabolic or physiological states **A)** Model for proper function of *GLABRA1* in *A. thaliana (GL1),* which promotes trichome development (left). Knockouts of *gl1* removes the capacity for protoderm cells to differentiate into trichomes, generating additional pavement cells (right). **B)** *ZmSWEET13s* (purple) are required for transport of sucrose from bundle

sheath cells into the vasculature in *Zea mays*. *zmsweet13* knockouts raise sucrose concentrations in bundle sheath cells. **C)** Hypothetical example of a *de novo* discovered marker gene identified by single-cell RNA-seq. Expression of the *de novo* discovered marker, *Unknown,* is limited to companion cells, as opposed to pavement cells, mesophyll cells, and bundle sheath cells. Single cell RNA-seq reads are colored by their strand, with purple reads representing the positive strand and yellow reads representing the negative strands.

The Genetic Underpinnings of Plant Cell Identity:

Plant cell development and function result from a complex interplay of genes responsible for determining cell fate and maintaining cellular identity. Identification of key developmental regulators, like *SHR,* demonstrated that the development of entire cell lineages depended on the expression of a few genes. Determining how and where these essential "marker" genes of cell identity were expressed became a central question in plant genetics. Subsequently, molecular genetic approaches such as mutagenesis screens, and reporter gene assays, were developed to assess the cellular context in which these marker genes were expressed. These advancements resulted in the identification of many other genes critical in cell-type identity. For example, *GLABRA1* (*GL1*) in *A. thaliana* controls trichome fate, as *gl1* null plants generated by T-DNA insertion had no trichomes on the leaf and stem **(Figure 1.1A)** (Herman & Marks, 1989; Oppenheimer et al., 1991). Despite establishing the necessity of *GL1* for trichome formation, this finding did not elucidate its expression pattern or how *GL1* facilitated trichome development. This knowledge gap led to the creation of promoter reporter lines, in which a gene's transcriptional regulatory sequences (promoters and CREs) are fused to a reporter (e.g.,

GUS, GFP) to illuminate where and when the gene is expressed (Birnbaum et al., 2003; Brady et al., 2003; Helariutta et al., 2000; Stadler et al., 2005). In *GL1* reporters, expression was found to change throughout development; in early development, *GL1* is expressed throughout the early leaf primordia, but, as the epidermis matures, only trichomes precursors maintain high *GL1* expression (Kirik et al., 2001; Larkin et al., 1994; Oppenheimer et al., 1991). Research into genes crucial for cell development expanded reporter methods by combining cell-type reporters with protoplast isolation to isolate cell populations and conduct genome-wide identification of transcription factors associated with specific cell types (Birnbaum et al., 2005; Toufighi et al., 2005). Application of these genome-wide assays identified genes crucial for the development of particular cell types. Presently, cell-type-specific genetic inquiry in plants has the potential to be significantly enhanced through single-cell methodologies, allowing for refined discrete measurement from individual cells empowering our understanding of plant cell fate decisions.

While genes important in the development of cell types are critical to our understanding of how cell types differentiate, they do not reveal much about plant cell-type function. This has led to researchers looking for genes which are important to the function of mature cell-types. For instance, genes such as *SUGARS WILL EVENTUALLY BE EXPORTED TRANSPORTER 13* (*ZmSWEET13)*, a sucrose transporter, is expressed specifically bundle sheath cells and phloem parenchyma (Bezrutczyk et al., 2018, 2021). Knockouts of *ZmSWEET13* impair phloem loading increasing sucrose concentrations in leaves **(Figure 1.1B)**. Although not required for abaxial bundle sheath cell development, *ZmSWEET13* represents a key gene required for cell-type-specific function. Similarly,

Arabidopsis *SUCROSE-PROTON SYMPORTER 2* (*SUC2*) drives sieve element sucrose loading through companion cell-specific expression (Stadler & Sauer, 1996). *suc2* knockout plants are stunted due to impaired sucrose transport, but companion cell identity is unaffected (Gottwald et al., 2000). Genes such as *SUC2* and *ZmSWEET* further our understanding of the genetic partitioning of functions between plant cell types. This genetic division enhances our perspective of what constitutes a plant cell type, transitioning from definitions based on histology, to those based on gene expression and function of discrete genetic loci. Although finding cell-type-specific functional genes is valuable, their identification is generally done by investigating a single gene at a time, requiring significant investment of time and resources. Single-cell genomics provides an opportunity to discover additional genes important to function of specific cell types on a genome-wide scale, across all cell types sampled at a single time. This influx of information will quickly evolve our understanding of cell types from a few key loci, to combinations of genes critical for both development and function.

Single-cell genomics allows measurement of chromatin states and mRNAs in thousands of individual cells (Buenrostro, Wu, Litzenburger, et al., 2015; Cusanovich et al., 2015; Jaitin et al., 2014). Plant single-cell genomics is especially exciting given the lack of cell-type-specific genomic measurements outside of model plant systems. The information-richness and high-complexity of single-cell datasets are useful because it allows for a detailed understanding of how different cell types utilize the genome. However, single-cell technologies remove cells from the sampled tissue, erasing any knowledge about position or identity, and complicating the identification of each cell's cell type. Therefore, annotation relies on molecular marker genes to reveal cell identity

post hoc. This annotation is confounded by the gradient of transitional cell identities that underpin differentiation. For instance, in *A. thaliana* guard cell differentiation from protoderm involves five state transitions, necessitating additional markers to accurately delineate cellular states (L. Chen et al., 2020). These transitory states make having well-established developmental marker genes critical to accurate annotation of single-cell datasets. For this reason, the first plant single-cell RNA-seq (scRNA-seq) analysis was conducted on *A. thaliana* roots because root cell types have well described genes associated with specific cell types and developmental stages (Ryu et al., 2019; Shulse et al., 2019). Once single-cell datasets are accurately annotated, they can be leveraged in powerful ways. Testing for differentially expressed genes in annotated cell types allows for the identification of novel genes potentially critical in proper cell-type function. One scRNA-seq A. *thaliana* study used annotated root cell types to discover 50 genes with cell-type-specific expression patterns **(Figure 1.1C)** (Zhang et al., 2019). This "*de novo* " discovery of cell-type-specific genes provides a wealth of candidate genes to target and study, which will further reveal their importance in specific cell types of interest. Single-cell genomics will increase the speed of cell-type-specific gene identification, improving our understanding of which loci are critical for proper cell-type function and development in plants.

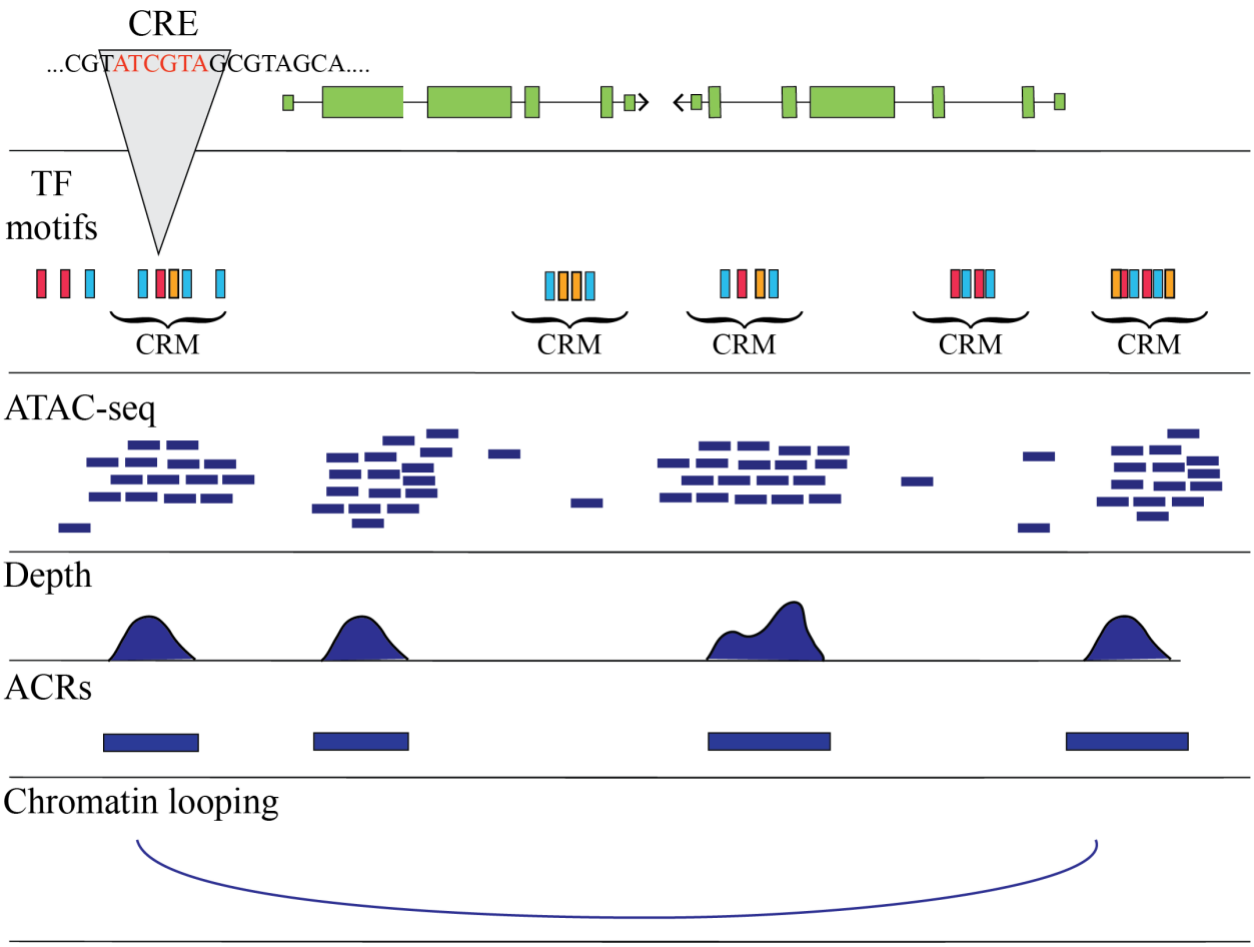

Figure 1.2: Deciphering the regulatory genome with ATAC-seq. Top: Gene annotation TF Motifs: Transcription factor binding motifs. Colors represent different motifs. Motifs occur in clusters referred to as *cis*-regulatory modules (CRMs). Gray arrow highlights a specific TF motif acting as a CRE. **ATAC-seq**: Example of aligned reads from an ATAC-seq experiment. Note that reads align heavily to CRM regions. **Depth:** Histograms of the above read depth. **ACRs:** Identified accessible chromatin regions from the read depth above. **Chromatin looping:** An example of how ACRs do not always operate on the closest gene. The line represents chromatin interaction between two ACRs.

The Regulatory Genome Specifies How Cell types are Established

Although scRNA-seq will improve our knowledge of cell-type-specific gene expression, our understanding of the processes driving these expression patterns remains poor. Pairing of single-cell technologies with assays identifying the regulatory genome stand to greatly enhance our understanding of how the genome can regulate the expression of both developmental and functionally important genes. Cell-type-specific expression is the result of different cells using the same genetic blueprint encoded in the genome in different ways. Cell fates are determined by the interpretation, enhancement, or silencing of instructions encoded in DNA which are driven by CREs (Andersson et al., 2015). CREs are non-coding sequences of DNA composed of transcription factor (TF) binding sites. TFs bind CREs within nucleosome depleted sequences, to recruit co-factors, remodel chromatin, and regulate gene transcription (Lai et al., 2019). This *cis*-regulation has implications on plant development, environmental response, and evolution (Cramer, 2019).

CREs often work in concert and are then referred to as *cis*-regulatory modules (CRMs) (**Figure 1.2**) (Schmitz et al., 2022; Shlyueva et al., 2014) . CRMs are further subdivided as "enhancers," or "silencers," based on the ability to recruit co-activators or co-repressors to genes (Gisselbrecht et al., 2020; Pang & Snyder, 2020; Shlyueva et al., 2014). Identification of CRMs genome wide is routinely performed with assays that measure accessible chromatin environments, as these are the regions that are open to TF binding. Methods such as DNase-seq, MNase-seq, as well as FAIRE-seq have been used to study CRMs genome wide (Boyle et al., 2008; Giresi et al., 2007; Johnson et al., 2006). Currently, the most widely adopted method to investigate accessible chromatin is

Assay for Transposase-Accessible Chromatin followed by sequencing (ATAC-seq) (Buenrostro, Wu, Chang, et al., 2015). In brief, ATAC-seq works by utilizing a hyperactive Tn5 transposase to directly insert sequencing adapters into accessible chromatin regions of DNAs (**Figure 1.2**). The fragments generated are then amplified, sequenced, aligned to the genome, and areas more accessible than genomic background are computationally identified (**Figure 1.2**) (Yan et al., 2020) . These peaks, named accessible chromatin regions (ACRs), are well accepted proxies for CRMs, and thus collections of CREs (Bajic et al., 2018; Lu et al., 2017).

With the widespread adoption of ACR identification, numerous discoveries have been made about the regulatory nature of DNA in plant genomes. For instance, it has recently been revealed that ACRs frequently operate on genes >50 kilobases away in plants with large genomes (Ricci et al., 2019). Additionally, variable ACR usage has been implicated in biotic and abiotic stress responses, providing more insights into how the genome tunes expression to the environment. (Han et al., 2020; Raju, 2020; Z. Zeng et al., 2019; Zhou et al., 2022). However, ACRs provide no information about whether these regions are enhancing or repressing transcription. This can be predicted by overlaying ACRs with ChIP-seq data which measures the histone modifications nearby (Lu et al., 2019; Oka et al., 2017; Ricci et al., 2019). ACRs that are active are flanked by histone acetylation, whereas those that are actively repressing a target gene are flanked by histone methylation and Polycomb silencing (Lu et al., 2019; Ricci et al., 2019). Recently, the ability to apply ATAC-seq to single cells (scATAC-seq) was developed, allowing for cell-type-specific ACR identification and measurement (Buenrostro, Wu, Litzenburger, et al., 2015; Cusanovich et al., 2015).

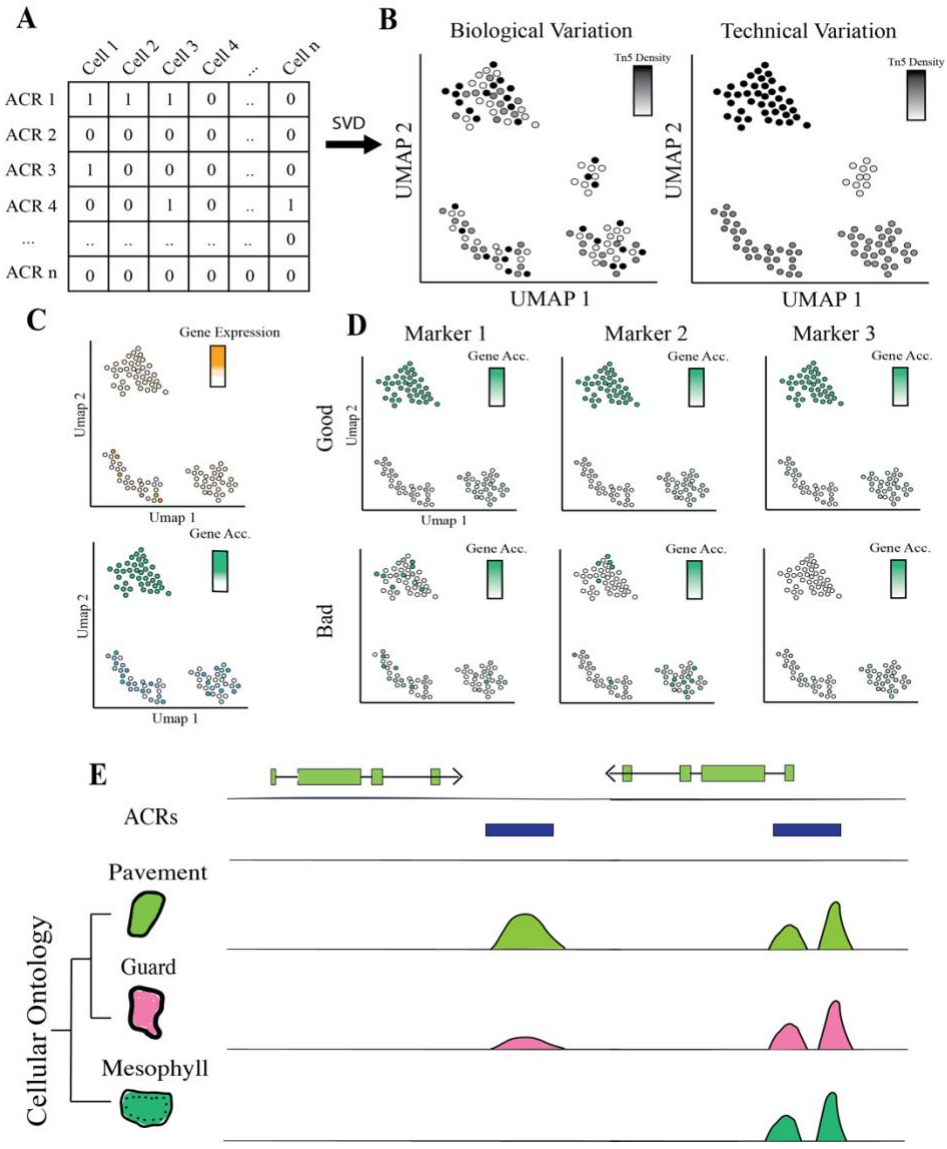

Figure 1.3: Schematic of analysis paradigms and challenges of single-cell ATAC-seq data **A)** An example binary matrix being transformed into a UMAP by means of SVD where rows are ACRs identified from a bulked dataset and columns are cells. The input matrix is binary if a cell either has a Tn5 integration event in that ACR or not **B)** Left) An example of a good UMAP embedding where Tn5 insertion density is not driving the clustering of cells. Right) An example of a poor UMAP embedding where technical artifacts due to Tn5 insertion density are driving the clustering **C)** An instance where chromatin accessibility of a gene (Gene Acc.) does not align with expectations based on gene expression (Gene Expression). Here, the upstream region of the gene depicted could harbor a silencer that recruits TFs to repress its target gene. **D)** Top) An example of where the UMAP embedding correlates well with the chromatin accessibility of multiple marker genes for a specific cell type. Bottom) Example of a poor UMAP embedding

where markers do not agree with each other. Whether this is due to poor markers or a poor selection of features to generate the embedding requires additional inquiry. **E)** Challenges associated with the assignment of single-cell ACRs. Whether the middle ACR is cell-type specific or restricted to a few key cell types is challenging to determine. This potentially reflects issues in annotation, or that related cell-types share similar chromatin environments. The assignment of cell-type specific ACRs is non-trivial and requires careful considerations by the researcher of both biological and technical challenges.

Single Cell ATAC-seq, Emerging Paradigms and Tangible Value

ScATAC-seq has revealed differential usage of ACRs in cellular identity and development in plant models (Dorrity et al., 2020; Farmer et al., 2021; Marand et al., 2021). Although the application of scATAC-seq techniques to plants stands to teach us much about *cis*-regulatory biology, implementing these techniques are non-trivial and come with a series of caveats and challenges. The ability to deconvolute cellular heterogeneity in plant tissue allows for the identification of cell-type-specific ACRs, which can be used to identify TFs and CREs important to cell function.(Marand et al., 2021). Intriguingly, scATAC-seq also offers a method to study developmental trajectories within cell types. Key genes or CREs that operate differently through development can be identified by ordering cell lineages from progenitor to mature cell type (Nelms & Walbot, 2019; Trapnell et al., 2014). This “pseudo time” method was applied to root hair development in *Oryza sativa*, as well as phloem companion cell development in *Z. mays* root. In *O. sativa,* pseudotime analysis found 13,000 ACRs and 3,000 genes important in the transition into root hair cell-type identity (Zhang et al., 2021). In *Z. mays,* it was found that as cells differentiate from quiescent center cells to phloem companion cells the fractions of ACRs which were accessible decreased significantly (Marand et al., 2021). Pseudotime analyses are just the beginning as application of scATAC-seq to plants will reveal roles of *cis*-regulation in evolution, stress responses, and adaptation. While

exciting, the analysis and annotation of these datasets is computationally challenging and requires awareness of current limitations.

The computational challenges associated with scATAC-seq data analysis are primarily due to the low number of Tn5 integration events per cell. For example, upwards of 99% of the chromatin accessibility measurements genome wide are often missing from any particular cell (Buenrostro, Wu, Litzenburger, et al., 2015). This data scarcity has significant ramifications in scATAC-seq analysis. The first step of scATAC-seq analysis is isolating high quality cells. One way of doing this is by "pseudo-bulking," which mimics bulk ATAC-seq by aggregating the reads from all nuclei, to identify peaks (ACRs) (H. Chen et al., 2019). Then broken nuclei are removed by estimating the Fraction of Reads in Peaks (FRiP) per nucleus, and removing nuclei with Tn5 integration events below a FRiP threshold (generally >.25) . Next, doublets, which are instances where two cells are mistakenly sequenced as one are removed by comparing them against an *in-silico* generated doublet set of cells (Wolock et al., 2019). Based on the single cell technologies used, the top 5-10% of cells with the highest "doublet score" are removed. The next steps annotate similar cells into cell types. Annotation starts by generating a binary matrix of Tn5 insertions in ACRs by cells, which is fed into dimensionality reduction algorithms. These algorithms, such as singular value decomposition (SVD) or principal component analysis (PCA), cluster cells into similar groups by identifying correlated features (ACR presence/absence), which reveal underlying patterns and relationships among the cells **(Figure 1.3A-B)**. The resulting principal components, or meta-features, represent high-dimensional data (ACR accessibility) in low-dimensional space.. Cell proximity in this low dimensional is a proxy for cell relatedness, either

biological or technical (**Figure 1.3B)**. Presently, either Uniform Manifold Approximation and Projection (UMAP) or t-Distributed Stochastic Neighbor Embedding (t-SNE) are used to visualize scATAC-seq data. These techniques plot cells in 2D, while trying to preserve the high dimensional space computed above (**Figure 1.3B**) (Maaten & Hinton, 2008; McInnes et al., 2020). One should not make biological conclusions about relative distance and space between cells, as recent evidence points to the inherent flaws in this approach (Chari et al., 2021). For instance when using three dimensional datasets with known spatial relationships between points, t-SNE or UMAP processing scrambles the relative distance between points, indicating that the 2D distances generated are artifactual (Chari et al., 2021). From this embedding, discrete cell clusters are assigned using community detection methods such as Louvain or Leiden algorithms (Blondel et al., 2008; Waltman & van Eck, 2013). In brief, these methods work by trying to identity clusters of cells in high-dimensional space, which maximizes the differences between groups and minimizes the differences within groups based off a given parameter set. Clusters are then analyzed with the assumption that they are representative of roughly homogenous cell types. Annotating clusters to a cell type involves approximating gene expression of cell-type markers by summing gene body and promoter chromatin accessibility (Cusanovich, Hill, et al., 2018). This approximation, while valuable, is imperfect, as chromatin accessibility does not always correlate with gene expression (**Figure 1.3D**). Based on the specific chromatin accessibility patterns of known marker genes, clusters are assigned cell types (**Figure 1.3C**). The clustering and annotation of cell types remains one of the most time consuming and difficult steps in scATAC-seq analysis. Current heuristic methods rely on user based decisions that are often difficult to

replicate (Gibson, 2022). As the field matures, more consistent annotation metrics are needed to ensure proper and timely cell type assignment.

ScATAC-seq analysis is a deeply iterative process. For instance, selecting different ACRs to include in dimensionality reduction can drastically alter cluster membership and generate different results. This requires researchers to try different selections of ACRs to find a set that reduces technical artifacts but maximizes biological interpretation. Additionally, technical artifacts can have significant effects on annotations and interpretation of results. For example, cells with a high density of Tn5 insertion events per cell can cluster together, thus the underlying embedding doesn't represent one of biological variation, but technical (**Figure 1.3B**). Technical artifacts can be even more misleading, with cells being assigned specific clusters due to the lack of data, rather than the presence of genuine differences.

Once annotations are finalized, cell-type-specific ACRs are identified. Combining cells of the same cell type via "pseudo-bulking" allows for the robust identification of ACRs for individual cell types (Cusanovich, Reddington, et al., 2018; Domcke et al., 2020). This deconvolution of tissue-level chromatin accessibility to cell-type resolved accessibility is where the power of single cell lies. While identifying ACRs from cell-type-level data is straightforward, classifying these ACRs as cell-type-specific or broadly accessible is challenging and is heavily impacted by the statistical approach chosen (**Figure 1.3E**). Making this categorization more opaque is cell types which share developmental origins often have similar chromatin accessibility patterns (Cusanovich, Hill, et al., 2018; Domcke et al., 2020). This leads to an additional class of ACRs that are cell-type-restricted or limited in their chromatin accessibility to a few cell types. Recent

plant scATAC-seq studies have found between 23-27% cell-type-restricted ACRs in given species (Marand et al., 2021; Zhang et al., 2021). However, the number and proportion of ACRs that are cell-type specific is unknown, and being established in model systems with more exhaustive sampling (X. Chen et al., 2018; Domcke et al., 2020). Whether these cell-type-specific ACRs are critical to cell-type function is uncertain and requires follow-up molecular genetics studies. Finally, scATAC-seq provides exciting opportunities to begin deciphering both *cis* and *trans* regulators of the genome. Recent studies have shown the ability to link TFs with their likely binding sites in a cell-type-specific manner, by correlating the chromatin accessibility of transcription factor gene bodies with the accessibility of their corresponding binding sites (Marand et al., 2021). This allows for the identification of cell-type-specific gene regulatory networks which have been long elusive. While the computational workflow and challenges labeled here may seem daunting, rigorous data analysis avoids many of these pitfalls. However, it should be noted that these specific computational challenges aren't the only issue. Quirks associated with evolution, genome structure, and the unique ways plant cell-type identity can be modified also need to be considered.

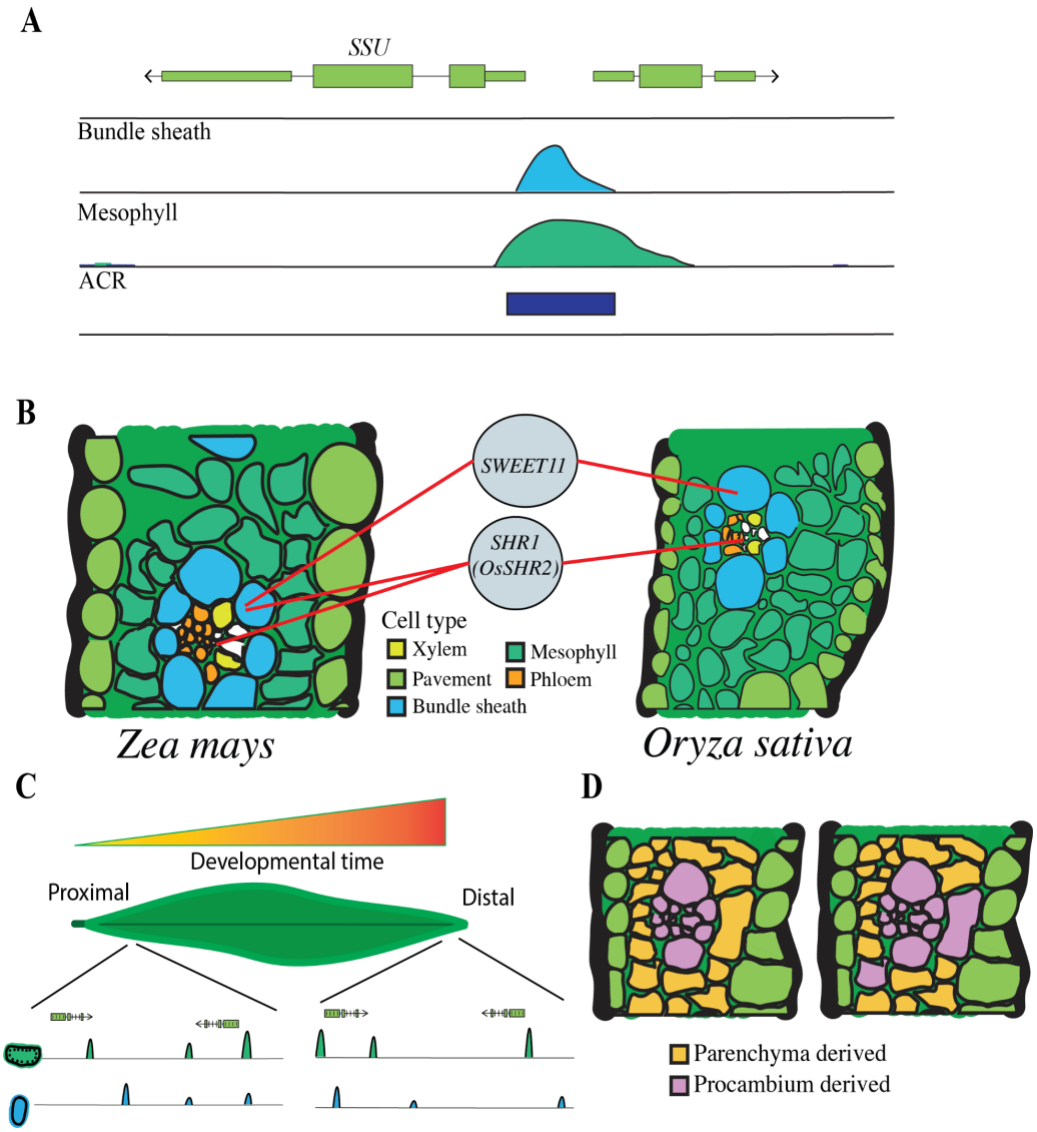

Figure 1.4: Biological challenges in single-cell ATAC-seq data take many forms and unique situations. **A)** The gene *SHORT SUBUNIT (SSU)* is known to have specific function in bundle sheath cells in *Z. mays*. However, SSU may not behave as a useful marker for annotation via scATAC-seq due to the proximity of its promoter to the start of an uncharacterized gene. **B)** Leaf cross sections of two species, *Z. mays*, and *O. sativa.* Cell types are color coded, and important marker genes are labeled in gray circles. Red lines point to the cell types these genes are active in. Although *SHR1* is expressed in bundle sheath and vascular cells in *Z. mays*, it is not found in *O. sativa* bundle sheath cells. **C)** Leaf laid out from proximal to distal ends with developmental gradient overlaid on top, with the oldest cells being at the tip and newest cells at the base. It is currently unknown whether different regions of the leaf might have different regulatory chromatin environments and subfunctionalization (bottom). **D)** Leaf cross sections where colors indicate cells originating from either parenchyma (ground tissue) or procambium. Left) Normal development Right) Abnormal development with position dependent effects altering the origin of a mesophyll cell derived from the procambium.

Biological Challenges Associated with the Analysis of scATAC-seq in Plants

While the computational challenges associated with scATAC-seq are laid out and addressable, certain unique features of plant biology complicate analysis. Variable genome sizes, rapid changes in gene function caused by molecular evolution, and the totipotent nature of plant cells all alter the interpretation of plant scATAC-seq data. However, while analytically challenging, these features offer unique opportunities to study CREs and their relationship to plant biology generally.

Significant variations in genome size can add additional hurdles to analyzing scATAC-seq data. For instance, genome size affects the use of gene proximal chromatin accessibility as a proxy for gene expression. In compact genomes, the reduced proximity between transcriptional start sites (TSSs) means ACRs often encompass two promoters, which convolutes correlating chromatin accessibility with gene expression **(Figure 1.4A)**. This is in stark contrast to larger genomes which result from the expansion of intergenic and intronic gene space often as a result of increased transposon load (S.-I. Lee & Kim, 2014; Lu et al., 2019). This size expansion moves CRMs important for gene expression further upstream of the TSS, increasing the prevalence of gene-distal ACRs (Lu et al., 2019; Oka et al., 2017; Ricci et al., 2019; Zhao et al., 2018). Linking these distal ACRs to their target genes is challenging and requires additional data from proximity-ligation based methods. Hi-C is the most commonly used, as it captures chromatin interactions ranging from 1 kb to >100kb depending on the experimental setup and sequencing depth. **(Figure 1.2)** (Eagen et al., 2015; Lieberman-Aiden et al., 2009; Mifsud et al., 2015). However, Hi-C remains restricted to bulk tissues, limiting the detection or confirmation of gene-ACR interactions in rare cell types. Predictions about

chromatin contacts can be made from scATAC-seq itself, but requires further experimentation for validation (Marand et al., 2021; Pliner et al., 2018). Genome size variation necessitates adapting analysis strategies on a per-genome basis, impeding the standardization of scATAC-seq analysis between species.

Although variation in plant genome size complicates scATAC-seq, the lack of high-quality markers for plant species remains the biggest challenge in the annotation and analysis of single-cell ATAC-seq datasets. Since currently no species has an exhaustive list of cell-type-specific genetic markers, markers are generally borrowed between species. For instance, in non-model plants cell types are annotated using gene orthologs of known markers with cell type specificity from model plant species. However, it is known that gene expression changes rapidly due to molecular evolution (Hill et al., 2020). Gene duplication followed by neo-functionalization, whole genome duplications rewiring large scale expression patterns, and rapid gene family expansion are a few of the many ways molecular evolution can reshape gene function, and expression (Birchler & Yang, 2022; Hughes et al., 2014; Panchy et al., 2016). Even in a relatively short evolutionary time span of 65-70 million years key developmental genes can shift their cell-type expression context, complicating their use in a cross species context (Hughes & Langdale, 2022). For example, *SHR* has different cell-type specificity in *Z. mays* and *O. sativa* leaves; in *Z. mays*, expression of *ZmSHR1* is limited to vasculature and bundle sheath cells, whereas the *O. sativa* ortholog, *OsSHR2,* has limited vasculature expression, and is absent from bundle sheath cells (**Figure 1.4B)** (Schuler et al., 2018)**.** Due to the unreliability of individual markers, non-model systems need to use sets of markers to annotate cell types in order to Minimize incorrect cell annotation. However, to date the

number of markers per cell type is limited, restricting the ability to apply single cell techniques to non-model plants.

Plants cells have a unique relationship between cell identity by descent, and cell position in the plant. Plant cell fates are not genetically hardwired based off precursors. For instance, although plant cell types generally develop in predictable lineages, such as the procambium giving rise to the vasculature, exceptions can occur. Instead, the location of a plant cell during development can have greater impacts on cell fate than their stem cell niche of origin (Reinhardt et al., 2003). This has been shown in *Z. mays,* where the mesophyll cells neighboring the bundle sheath lineage may be derived from procambial cells, or ground meristem cells **(Figure 1.4D)** (Esau, 1943; Langdale et al., 1989; Sharman, 1942). This position-dependent effect is well documented for vasculature and epidermal cell types in species including cotton, tobacco, and sunflower (Dolan & Poethig, 1998; Esau, 1954; Hung et al., 1998; Jegla & Sussex, 1989; Poethig & Sussex, 1985b, 1985a). This poses a challenge, as these cells which have undergone position dependent effects are likely to cluster with cells that share the same precursor, and not with cells with the same terminal identity. This makes annotation more difficult and increases the heterogeneity in identified cell types. Finally, since these events are rare, isolating and studying these populations is challenging, but could pose a valuable study system to understand the role of how variable precursors alter the chromatin environment of differentiated cells.

Finally, the gradient nature of plant growth provides additional challenges. Plant organs grow in a gradient of development, with younger cells found closer to the dividing meristem, and older cells further away. Continual organogenesis and development results

in gene expression profiles that are dependent on the section sampled within an organ **(Figure 1.4C)**. In *Z. mays* this developmental progression yields differences in expression of key carbon metabolism enzymes at different sections of the leaf (Li et al., 2010; Pick et al., 2011; Wang et al., 2013). Whether these different sections of the leaf, and the cell types found within, constitute different cell types or specific sub-functionalization is up for debate, and further complicates placing cells into discrete categories (H. Zeng, 2022). This heterogeneity has already been hinted at in some studies. A combinatorial scATAC-seq and scRNA study of *A. thaliana* roots, found unique genetic and epigenomic markers in three different clusters of endodermal cells, illustrating that discrete sub-functionalization of may happen within previously described cell types (Dorrity et al., 2020). The extent to which these clusters represent unique sub-functional cell types remains open and requires further exploration.

The Age of Single Cell Regulatory Genomics:

scATAC-seq enables the genome-wide investigation into the function and importance of plant cell-type-specific CREs. Although we can now identify cell-type-specific CREs in plant genomes, our understanding of how these regions interact with the coding genome is still quite limited. Leveraging intra- and inter-genetic diversity, along with treatment conditions, stands to greatly improve our understanding of CREs in plant biology and their role in responding to environmental stimuli, population adaptation and diversity, as well as reveal their importance over evolutionary time.

Performing scATAC-seq on a phenotypically diverse intra-specific population will clarify the influence of genetic CRE variation on phenotypes with cell-type resolution. Genetic variation in regulatory sequences can result in species adaptation to novel environments in both plant and animal systems (Cleves et al., 2014; Studer et al., 2011; van der Burg et al., 2020; Wucherpfennig et al., 2022). In plants, CRE variation in the flowering time gene CONSTANS underlies flowering time diversity in natural accessions of *A. thaliana* (Rosas et al., 2014). However, most studies addressing CRE genetic variation lack cell-type resolved data and therefore may overlook genetic variance in rare cell-type CREs that underpins local adaptation. Combining quantitative genetic approaches, like genome-wide association (GWA), with scATAC-seq, phenotypic associations and chromatin accessibility variation can be correlated, potentially identifying the CREs, and cell types, underpinning trait variation within distinct populations (Das et al., 2022). Although a nascent area of study, the combination of scATAC-seq and population diversity may reveal how CRE genetic diversity alters the regulatory epigenome to shape species adaptation.

Beyond applying scATAC-seq to single species populations, comparative genomics focused on diverse species offers the opportunity to examine plant CRE evolution at a deeper timescale. Plant genomes exhibit a high rate of structural variation and sequence turnover as compared to animal genomes, causing rapid CRE turnover (Paterson et al., 2010). Highlighting the high rate of CRE change between even closely related species, a study comparing distal CREs between sister species *Z. mays* and *S. bicolor*, found approximately one-third of CREs were shared and accessible in the same tissue, one-third were novel to each lineage, and one-third shared sequence similarity but

were not within accessible chromatin in the tissue examined (Lu et al., 2019). While CREs sequences change quickly, the gene regulatory networks they influence may be more stable. Investigating root hair cell type development in four eudicots found that although few orthologous CREs were conserved across all species, TF binding at key genes was preserved (Maher et al., 2018). Pairing comparative genomics analyses with scATAC-seq will allow investigation into the pace of CRE sequence changes in specific cell types within individual plant lineages. This approach will enhance both our understanding of plant CRE evolution and uncover conserved mechanisms underpinning plant adaptation and resilience to environmental changes.

Finally, CREs drive responses to environmental stimuli. Differential CRE usage is vital in response to disease, cold, drought, and hormonal signals (Azodi et al., 2020; Moore et al., 2022; Reynoso et al., 2019; Zou et al., 2011). One comparative genomics study examined CRE usage with a flooding treatment and identified root-specific CREs associated with flooding, which revealed shared motifs within flood-responsive CREs across four species studied, representing 123 million years of evolutionary divergence (Reynoso et al., 2019). This flooding research suggests that regulatory networks behind abiotic stress responses may be conserved for millions of years. Integrating scATAC-seq with environmental treatments will identify the CREs crucial for cell-type-specific environmental responses. Beyond discovering environmentally dynamic CREs, this approach will find the cell types with the most responsive CRE usage in different conditions, revealing which cell types drive stress adaptation. This focus on cell-type responses could have far-reaching implications for our understanding of environmental

response in plants, as previous study has traditionally been restricted to organismal response.

Plant Cell Types – Definitions in Flux:

While the age of single-cell genomics stands to alter our understanding of plant cell biology, it is important to acknowledge that the definition of a cell type is in flux. In this perspective, we define a "cell-type" as a cell with unique molecular signatures, and that alteration of this signature modifies the form or function of a given cell type. However, although valuable, this definition has limitations. For instance, what is the threshold of molecular changes needed to separate related cells into distinct cell types? How many differentially expressed genes or differentially accessible CREs are needed to constitute a novel cell type? This problem becomes especially acute when trying to delineate plant cells in transitional identities, as developing plant cells exist along a continuum of maturity with few discrete stages. While the discussions surrounding plant cell-type classifications may appear semantical, it underpins real biological questions. How we define "cell types" will have real implications for biologist moving forward (Efroni, 2018; H. Zeng, 2022).

Despite their immense development, maturity, and anatomical differences, inevitably, knowledge of plant cell types will be compared to what we know about animal cell types. IN plants there is wide variation in the number of identified cell types, with 55 being identified in *Z. mays* and 180 in *A. thaliana* (T. A. Lee et al., 2023; Marand et al., 2021). This contrasts significantly with animals, as in mouse brains alone there exists 45 types of inhibitory neurons (Hodge et al., 2019). The definition of fewer plant

cell types could be explained by more technological limitations and less intensive study than that found in animal models. Alternatively, the paucity of plant cell types may reflect real biological differences between plants and animals. Unlike mammals, plant cell divisions result in the incomplete separation of nuclei; cytokinesis ends with the deposition of a new cell wall (cell plate) that contains plasmodesmata pores that retain cytoplasmic, and endoplasmic reticulum, connections between the daughter cells (Burch-Smith & Zambryski, 2012). The interconnectedness of plant cells through plasmodesmata has large implications in plant biology and may fuel the differences between plant and animal cell types, as plant cells exist as a connected community not individuals. This interconnectedness has led some to propose a more holistic 'organism-level' view. Instead of focusing on cells or cell types as the biologically meaningful units of study, the organismal theory proposes to focus on the entire organism, as plant cells rarely work in isolation (Kaplan & Hagemann, 1991). However, this organism-level perspective conflicts with the severe phenotypic alterations caused by mutants that eliminate specific cell types as detailed above. In either case, single-cell (epi)genomics will reveal more about why plant cell types are less numerous than their mammalian counterparts. These techniques provide unprecedented cellular resolution, and if the lack of plant cell types is driven by past technical limitations, single-cell genomics will usher in an era of discovery wherein many new discrete plant cell types will be unveiled. Alternatively, if these new techniques confirm a relatively small number of more homogenous plant cell types, it may provide credence to the notion that plant cells should be studied as a physiological unit, highlighting the importance of intercellular cooperation in plant biology. Single-cell

regulatory genomics stands to enliven plant research and provides the toolset to address these basic questions about the cell-type composition of plants.

Acknowledgments

We would like to thank KD for feedback on this perspective. Research reported in this publication was supported by the National Institute of General Medical Sciences of the National Institute of Health under award number T32GM007103. Additionally, this work was supported by the National Science Foundation (IOS-1856627, IOS-2026554 and MCB-2120132) to RJS. This material is based upon work supported by the U.S. Department of Energy, Office of Science, Biological and Environmental Research Program under Award Number DE-SC0023338. Please note that all supplemental files and tables in this dissertation can be found here: https://github.com/Jome0169/Mendieta_Dissertation_github

Chapter 2

Leveraging histone modifications to improve genome annotations[1]

[1]Mendieta, John Pablo. Accepted by Genes|Genomes|Genetics. Reprinted here with the permission of the publisher 3/13/2024

Abstract:

Accurate genome annotations are essential to modern biology; however, they remain challenging to produce. Variation in gene structure and expression across species, as well as within an organism, make correctly annotating genes arduous; an issue exacerbated by pitfalls in current *in-silico* methods. These issues necessitate complementary approaches to add additional confidence and rectify potential misannotations. Integration of epigenomic data into genome annotation is one such approach. In this study, we utilized sets of histone modification data, which are precisely distributed at either gene bodies or promoters to evaluate the annotation of the *Zea mays* genome. We leveraged these data genome wide, allowing for identification of annotations discordant with empirical data. In total, 13,159 annotation discrepancies were found in *Zea mays* upon integrating data across three different tissues, which were corroborated using RNA-based approaches. Upon correction, genes were extended by an average of 2,128 base pairs, and we identified 2,529 novel genes. Application of this method to five additional plant genomes identified a series of misannotations, as well as identified novel genes, including 13,836 in *Asparagus officinalis,* 2,724 in *Setaria viridis*, 2,446 in *Sorghum bicolor*, 8,631 in *Glycine max*, and 2,585 in *Phaseolous vulgaris*. This study demonstrates that histone modification data can be leveraged to rapidly improve current genome annotations across diverse plant lineages.

Index Words: epigenomics, genome annotation, histone modification, plant genomes

Introduction:

Accurate genome annotations and assemblies are an essential resource for modern biology. Their capacity to facilitate genetic inquiry, as well as operate as the backbone for genome biology makes their production vital. However, while the creation of gapless, and near-perfect genome assemblies is becoming commonplace (Liu et al., 2020; Miga et al., 2020), genome annotation remains challenging (Salzberg, 2019). Generation of a genome annotation requires multiple lines of evidence in the form of mRNA expression data, homology-based inference, and *in-silico* prediction algorithms, which are synthesized into a single concordant annotation (Yandell & Ence, 2012). The challenges of such complex data synthesis, potentially compounded by the generation of *in-silico* artifacts at each aforementioned stage of analysis, makes accurate genome annotation precarious at best (Salzberg, 2019).

The epigenome provides an invaluable untapped resource which adds additional support to increase confidence in genome annotation. Generally, eukaryotic genomes are divided into two distinct domains, 1) euchromatin, which is gene-rich and has abundant transcriptional activity, and 2) heterochromatin, which is gene-poor, densely populated with repeats and transposable elements and mostly devoid of transcriptional activity (Hannah, 1951; McClintock, 1950). These two major domains of the epigenome are defined by their occurrence with specific covalent modifications to DNA and to the alpha globulin tail of histones, which together comprise chromatin (Luger, 1997). Histone modifications are diverse and they correlate with a wide range of biological phenomena. Some have proposed that chromatin comprises a "language" or code all its own in the genome, with different combinations and permutations of histone modifications correlating to distinct biological outputs (Rando, 2012; Strahl & Allis, 2000). Evolutionarily, histone modifications and their functions are deeply conserved, with eukaryotes using similar sets of histone modifications around transcriptionally active and inactive regions of the genome (Bernstein et al., 2005; Morris et al., 2007; Schübeler et al., 2004), suggesting their essentiality to eukaryotic genomes.

In plants specifically, recent large-scale studies have corroborated histone modification function, and co-localization to specific regions of the genome (X. Li et al., 2008; Z. Lu et al., 2019; Mahrez et al., 2016; Ricci et al., 2019; Shi & Dawe, 2006). For example, transcribed genes generally possess Histone H3 Lysine 4 trimethylation (H3K4me3) and Histone H3 Lysine 9/27/56 (H3K9/27/56ac) acetylation near their transcriptional start sites and H3K4me1 and H3K36me3 throughout their gene bodies (Q. Li et al., 2015; Z. Lu et al., 2019; Oka et al., 2017; Ricci et al., 2019; Roudier et al., 2011; X. Zhang et al., 2009), whereas actively silenced genes often possess Histone H3 Lysine 27 trimethylation (H3K27me3) throughout their gene bodies and promoter-proximal regions (Bernatavichute et al., 2008; X. Li et al., 2008; X. Zhang et al., 2006, 2007; Zilberman et al., 2007). The epigenomic landscape of heterochromatin is quite distinct, as repeats and transposable elements are highly enriched for DNA methylation, H3K9me2 and small RNAs (Bernatavichute et al., 2008; C. Lu et al., 2005; X. Zhang et al., 2007; Zilberman et al., 2007). The unique patterns and distributions of histone modifications throughout the genome, especially within transcribed genes, provides a unique opportunity to improve efforts in genome annotations.

Histone modifications associated with transcription reflect various features of transcriptional units. For example, in *Arabidopsis thaliana,* H3K4 can be either mono- di- or tri-methylated by ARABIDOPSIS HOMOLOG OF TRITHORAX1 (ATX1) and ARABIDOPSIS HOMOLOG OF TRITHORAX2 *(*ATX2), (Alvarez-Venegas et al., 2003; Nislow et al., 1997; Saleh et al., 2008). These histone modifications primarily occur at genic regions of the genome, with H3K4me2 and H3K4me3 being distributed specifically around transcriptional start sites (X. Zhang et al., 2009). H3K4me3 as well as ATX1 are also found tightly linked to Pol II occupancy, as ATX1 and specific subunits of Pol II are consistently found to co-localize at promoters (Fromm & Avramova, 2014) . Binding of ATX1 and Pol II form a transcriptional initiation complex, allowing for rapid transcriptional responses (Song et al., 2015). Paired with this,

increased proportions of H3K4me3 at promoters correlate with enhanced transcriptional rates (X. Zhang et al., 2009).

A histone modification which is intimately linked to transcription elongation is Histone H3 Lysine 36 methylation. During transcription the phosphorylated carboxy terminal domain of RNA Pol II recruits the histone methyltransferase Su(var)3-9, Enhancer-of-zeste and Trithorax 2, or SET2 (homolog SET DOMAIN GROUP 8, or SDG8 in *A. thaliana*), to methylate H3K36 (Wagner & Carpenter, 2012). Much like H3K4, H3K36 can be mono- di - or tri- methylated, but only di- and tri- methylation correlate with transcription in plants (Xu et al., 2008). SET2 limits the occupancy of Pol II in yeast, indicating its essential role during transcription elongation (Kizer et al., 2005). In *A. thaliana,* mutation of *SDG8* has been implicated in a range of phenotypic phenomena from development, to timing of flowering (Bu et al., 2014; Cartagena et al., 2008; Cazzonelli et al., 2009; Jin et al., 2015). In plants, H3K36me3 co-occurs with the length of the transcribed units, demonstrating its deeply conserved function (X. Li et al., 2008; Z. Lu et al., 2019). Uniquely in plant genomes, H3K36me3 is correlated with the histone modification H3K4me1 across the length of the transcribed unit (Ricci et al., 2019; van Dijk et al., 2010; X. Zhang et al., 2009). This is in stark contrast to metazoan genomes where H3K4me1 primarily denotes intergenic enhancers (Bannister & Kouzarides, 2011; Rada-Iglesias et al., 2011).

Unlike the histone residues which are methylated, histone residues which are acetylated have a direct functional impact on transcription. Whereas methylated histones often act indirectly by recruiting protein complexes that impact the chromatin landscape, acetylated histones physically alter how DNA wraps around the nucleosome (Allfrey et al., 1964; He et al., 2003). The negatively charged acetyl groups added on the histone protein repel negatively charged DNA promoting a more permissive environment for transcription (Allfrey et al., 1964; Earley et al., 2007). In plant genomes, acetylated histones co-occur with other transcription initiation histone modifications, such as H3K4me3 around the promoter sequence of actively transcribed genes (Z. Lu et al., 2019; Roudier et al., 2011). Interestingly, in plants, histone acetylation can also indicate

accessible chromatin in proximal and distal cis-regulatory elements (Z. Lu et al., 2019; Oka et al., 2017; Ricci et al., 2019; Zhao et al., 2018).

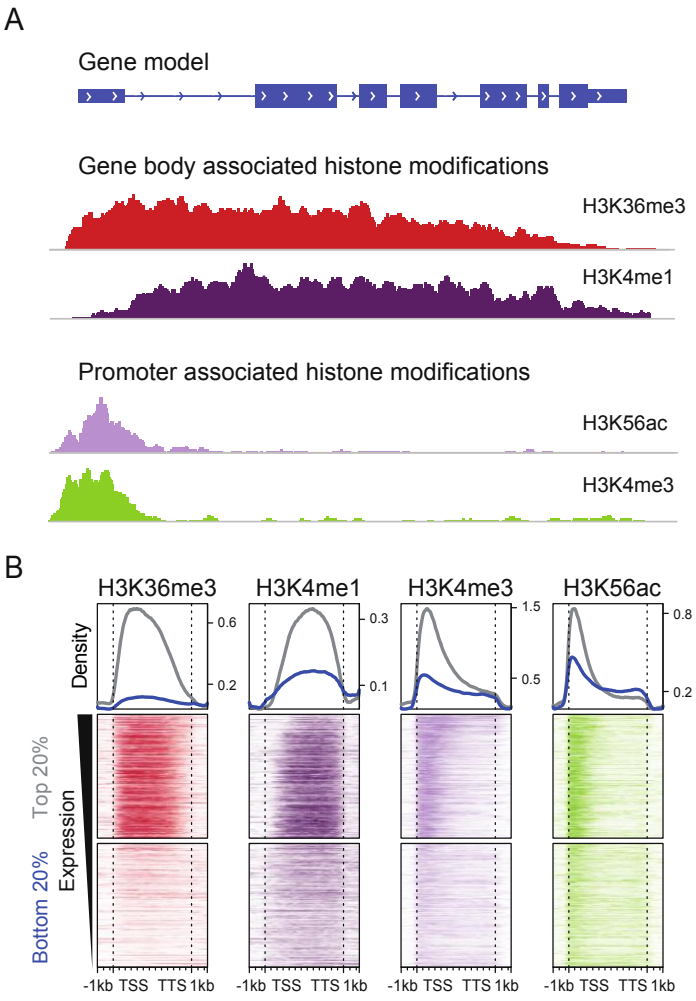

Figure 2.1: The distribution of histone modifications across expressed genes : A) An example of an annotated gene body, with corresponding histone modifications indicative of transcription. Histone modifications associated with transcription initiation (H3K56ac and H3K4me3) are known to correspond to the promoters, whereas histone modifications associated with transcription elongation (H3K36me3 and H3K4me1) occur across the length of the gene body. B) Metaplots of the top and bottom 20% of expressed genes in the genome of *Z. mays* ordered by expression.

Previous studies demonstrated that histone modification data can be leveraged on a genome-wide scale for a multitude of uses. For instance Sartor et al utilized epigenomic data identify the expressed regions of the maize genome, in what is sometimes called the "expressome" (Sartor et al., 2019). This utilization further allowed them to identify regions of the

maize genome, which are likely functional, and not constitutively repressed. However, although this utilization of epigenomic data provides valuable insights into expressed regions of the genome, it doesn't seek amend potential annotations issues present in current annotations (Sartor et al., 2019). Histone modification data has been used to annotate regions of the genome potentially harboring unannotated genes, or long non-coding RNAs (lncRNAs) (Guttman et al., 2009; Jarroux et al., 2017). A recent analysis of the *Z. mays* epigenome identified signals of actively transcribed transcriptional units outside of currently annotated gene features (Ricci et al., 2019). However, to date, few studies have leveraged epigenomics to improve the quality of genome annotations (Dozmorov, 2017; Ernst & Kellis, 2017). In this study, we show that integration of RNA-sequencing (RNA-seq) data with histone modification data significantly improves plant genome annotations.

Results:

To determine if histone modification data could be leveraged to improve genome annotations, we used previously published ChIP-seq data of histone modifications from leaf, root, and inflorescence tissue of *Zea mays* (Ricci et al., 2019), which is known to be challenging to annotate (Wang et al., 2016). These data were used to define the chromatin landscape around expressed genes. As previously reported, these data showed the expected enrichment of histone modifications around expressed genes, with H3K36me3 and H3K4me1 occurring across the gene body of actively transcribed genes as indicated by RNA-seq data, and histone modifications H3K4me3 and H3K56ac at promoters (Chunyan et al., 2010; He et al., 2003; Kizer et al., 2005) (**Fig. 2.1, Fig. 2.2A**). Hereafter, we refer to these histone modifications collectively as either "transcription initiation" (H3K56ac and H3K4me3) or "transcription elongation" (H3K4me1 and H3K36me3). These histone modifications together are representative of the chromatin environment around transcribed genes, considering many other modifications correlate with those chosen (Berr et al., 2011; B. Li et al., 2007). We evaluated the co-occurrence of regions enriched

for either histone modification comprising transcription initiation, and found that 64% of these regions co-occurred, as compared to 74% of gene body (H3K36me3 and H3K4me1) histone modifications (**Fig. 2.2B**). The percentage co-occurrence between similar histone modifications is consistent across additional sampled tissues: root and inflorescence (**Supplementary Figure. 2.1-2**).

A greater number of transcription initiation enriched regions than transcription elongation enriched regions were found in all tissues sampled (7,719 excess enriched domains in leaf, 9,327 in inflorescence, and 12,164 in root). This larger number of transcription initiation enriched domains as compared to transcription elongation modifications can be explained by multiple reasons. In total, 19,724 transcription initiation regions in leaf, 23,387 in root, and 20,941 in inflorescence overlapped genes. This discrepancy compared with the transcription elongation modifications can be in part explained by the fact that 917 genes in leaf, 1,580 in root, and 1,331 in inflorescence overlapped greater than one transcription initiation enriched regions, totaling an additional 1,981 transcription initiation domains in leaf, 3,235 in root, and 2,736 in inflorescence. Additionally, of genes that overlapped transcription initiation modifications, 2,584 in leaf, 3,441 in root, and 2,107 in inflorescence overlapped H3K27me3, a known repressive histone modification. A total of 4,822 in leaf, 6,018 in root, and 5,939 transcription initiation regions did not overlap with any annotated gene. Interestingly, of the subset of these transcription initiation

modifications; 550 in leaf, 805, in inflorescence, and 752 in root overlap H3K27me3 domains, possibly indicating silenced unannotated genes in the genome. Additionally, 1,669 transcription

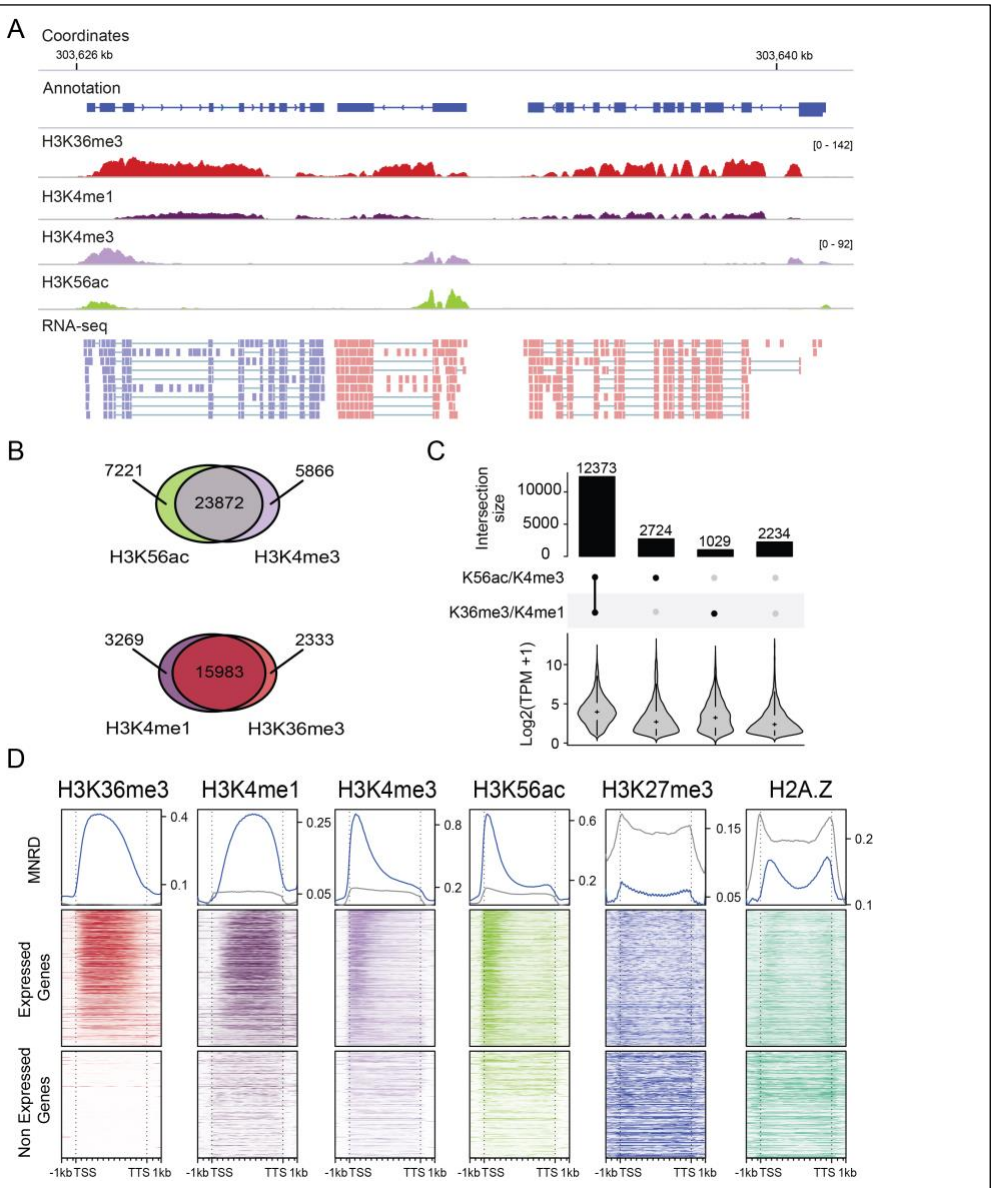

Figure 2.2: The histone modification landscape of expressed and unexpressed genes in *Z. mays*. A) Representative screenshot of expressed genes in the genome, and corresponding histone modifications. The histone modifications H3K36me3 and H3K4me1 are found across gene bodies, whereas H4K4me3 and H3K56ac are found in promoters of genes. B) Venn diagrams of enriched domains for co-occurring histone modifications that reflect either transcription initiation or transcription elongation in the maize leaf. C) Upset plot with number of genes with greater than 1 TPM, and their intersection with histone modifications associated with either transcription elongation and transcription initiation. D) Metaplot profiles of expressed and unexpressed genes.

initiation enriched loci in leaf, 2,795 in inflorescence and 1,782 in root, overlapped a region also enriched for at least one transcriptional elongation histone modification, possibly representing a

set of unannotated genes. Finally, a subclass of 2,340 inflorescence, 2,412 leaf, and 3,486 root transcription initiation enriched regions show no overlap with transcription elongation modifications (**Supplementary Figure.** 2.**3**). These regions are generally small with a mean size of 678 bp,and are on average 42,088 bp away from the nearest gene (**Supplementary Figure.** 2.**3**). We were additionally interested in seeing if these regions were conserved *in Sorghum bicolor*, a relative of *Zea mays* which shared a recent common ancestor 13 million years ago (Wang et al., 2018). By searching the Sorghum genome for initiation only regions we're able to ascertain whether these regions are unique to the maize genome, and additionally ask whether conserved regions represent potential missing genes missed in the annotation of *Zea mays*. To contrast the conservation of these unknown initiation only regions, we took an equal sample of initiation modification regions which overlapped genes in the same tissue type (2,412 in Leaf, 3,486 in Root, and 2,340 in Inflorescence). The underlying nucleotide sequences from these initiation domains were gathered, and orthologous regions in the sorghum genome were identified using Blastn (Camacho et al., 2009). In total, of the initiation only modification regions we were able to identify 236 leaf initiation only regions (9%), 411 root (11%) and 257 inflorescence regions (10%) in the Sorghum genome. This is in contrast to the control initiation regions of which we were able to identify 2,350 (90%) leaf regions, 3,191 (91%) root, and 2,118 (90%) inflorescence regions. These results indicates that a large majority of these initiation only enriched domains are unique to the genome of *Zea mays*. Finally, of the sequences which we could identify in Sorghum, we were curious to see if they overlapped genes, indicating maize specific annotations errors which have been corrected in the Sorghum genome. Of the initiation modification only regions, 178 leaf domains were found to overlap genes (75%), 287 root (69%) , and 201 inflorescence regions (78%), as compared to 2318 (98%) leaf control regions, 3125 (97%) root control regions, and 2077 (98%) inflorescence control regions. Demonstrating that of those initiation only regions which could be identified in Sorghum, a large majority belong to annotated genes, indicating further missing genes in the maize genome.

The concordance of histone modification data around expressed protein-coding genes was used to evaluate their potential to identify actively transcribed regions of the genome. Genes that had a transcript per million (TPM) value greater than 1 were labeled as "active" whereas those which had a TPM value less than 1 TPM were labeled as "inactive". To ensure that the analysis did not suffer from *in-silico* biases created by mappability issues in the maize genome, only genes that were greater than 70% mappable were used for analysis (see methods). Overall, 67% of active genes had both histone modifications indicative of transcription initiation and elongation (**Fig. 2.2C**). Genes that had both histone classes were likely to be more highly expressed as compared to the other three groups (harboring only one class of histone modification, as compared to two, or no domain enrichment), a trend observed across all three tissue types examined (Kolmogorov-Smirnov tests: $P < 2.2e-16$) (**Fig.** 2.**2C**; **Supplementary Figure.** 2.**1-2**).

To further demonstrate the relationship between histone modifications and transcribed regions of the genome, we evaluated the distribution of the histone modifications throughout gene bodies of active and inactive genes by generating metaplots (**Fig. 2.2D**). Transcribed genes generally show enrichment for histone modifications of both transcription initiation, as well as transcription elongation. Active genes also display the expected meta profiles of the sampled histone modifications, with H3K36me3 showing increased enrichment at the 5' region of gene bodies, and H3K4me1 showing increased enrichment at the 3' end. In contrast, inactive genes show no enrichment for transcriptionally related histone modifications, but do show enrichment for histone modifications (H3K27me3) and variants (H2A.Z) associated with facultative heterochromatin (Luo & Lam, 2010). These modifications are well documented to be present in genes silenced by polycomb repressive groups of proteins, generally demarcating developmental or environmental specific genes (Coleman-Derr & Zilberman, 2012). The slight enrichment in H3K4me3 around these silenced genes likely represents a set of genes bivalently modified, likely poised for rapid upregulation (Zeng et al., 2019). These results are similar for both inflorescence

and root tissues as well, and are consistent with expectations based on previous findings about the distribution of histone modifications around active and inactive genes (**Supplementary Figure** 2.**1 and 2.2**).

In the analysis of the histone modifications around expressed genes, we identified two distinct subclasses of genes which violated the expected distributions of histone modifications. One such subset of genes only co-occurred with transcription elongation histone modifications, whereas the other exclusively co-occurred with transcriptional initiation histone modifications (**Fig. 2.2C**). After manually inspecting a set of genes from each of these classes, we realized that a substantial proportion could be explained by misannotations, with the histone modification data clearly denoting the true extent of the gene model. For instance, in the transcriptional elongation only class, oftentimes the correct transcription initiation start site was clearly evident directly upstream. This led us to speculate that histone modification data can be leveraged to improve gene annotations and to identify novel genes not previously annotated in the genome.

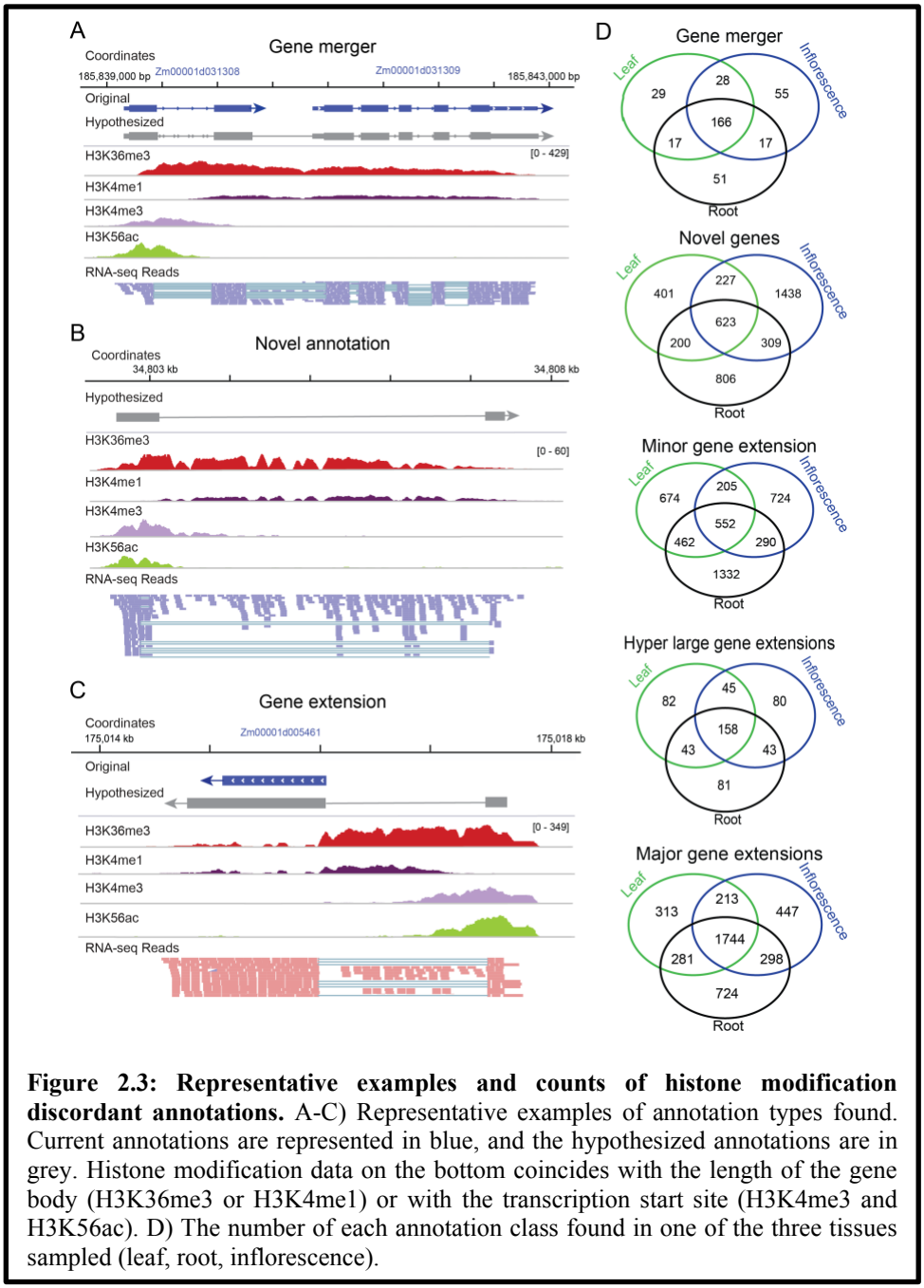

Figure 2.3: Representative examples and counts of histone modification discordant annotations. A-C) Representative examples of annotation types found. Current annotations are represented in blue, and the hypothesized annotations are in grey. Histone modification data on the bottom coincides with the length of the gene body (H3K36me3 or H3K4me1) or with the transcription start site (H3K4me3 and H3K56ac). D) The number of each annotation class found in one of the three tissues sampled (leaf, root, inflorescence).

Identification of Previously Ambiguous Annotation Classes

After manually inspecting regions of the genome where the histone modification data was discordant with the annotation, we identified three distinct classes of putative misannotations. One class labelled the "Gene merger" class featured histone modification data supporting a single transcriptional unit, but instead, multiple gene annotations existed at these loci in the reference (**Fig.** 2.**3A**). Further, alignment of RNA-seq data clearly shows reads bridging the gap between many of these putative misannotations, further supporting that these are a single transcribed unit. A second class of annotation issues found was regions of the genome that had evidence of transcription, and yet had no annotation present in the reference annotation. This class, labelled the "Novel class", likely identifies novel protein coding genes or lncRNAs (**Fig. 2.3B**). Finally, we identified an annotation class based off of missing downstream or upstream regions of the transcribed unit that we labeled the "Extension class" (**Fig.** 2.**3C**). This annotation class is defined by signals of transcription initiation appearing upstream of the annotation, or transcription elongation histone modifications extending past their current length of the full transcript. We further subdivided the "Extension class" based off the distance added to the original annotation, with minor extensions being annotations which were only extended by less than 500 bp or the length of a single exon, major extensions comprising regions falling between 500 and 2,000 bp, and hyper large extension being those greater than 2,000 bp.

Using these defined classes, we implemented a method to identify these regions genome wide (see Methods section), across three different maize tissues (inflorescence, leaf, and root). In total, we identified 4,004 potential novel annotations, with 66% (2,645 loci) being identified in only a single tissue. We found 363 potential gene merger events, with 166 (45%) of these mergers being found in all tissues sampled. Of the potential mergers, 357 (98.3%) of the predicted mergers consisted of gene pairs, with the remaining six (2.65%) representing loci where three or more genes were hypothesized to be a single transcriptional unit, in total encompassing 732 gene features. Further, 108 (29.8%) of the potential merger events have identical Gene Ontology (GO)

terms, possibly indicating a single locus which was divided into two during the annotation process. To rule out potential assembly errors being the main cause of this merger class, we intersected our merger class with a list of B73 contigs, and found all but one (99.72%) were found on a single contig, ruling out large scale genomic assembly errors as a potential cause of these merged genes. Additionally, to ensure that this approach wasn't merging tandem gene duplicates, we used BLASTP to compare the protein coding sequences of merged pairs, and looked for sequence identity (States & Gish, 1994). We found that only four (2.4%) of these merged pairs had any significant sequence identity between them and of these four pairs, only two had identical GO terms. Finally, of the three extension classes, 4,252 minor extensions, 4,064 major extensions and 543 hyper large extensions were found. For both major, and minor extensions, root comprises the highest proportion of uniquely identified extensions, comprising 17% of the major extension class and 31% of the minor extension class. Transcripts found in each annotation class were additional scanned for functional domains, we found that within the hyper large gene class 433 (80%) had a functional domain, as compared to 441 (60%) genes in the merger class, 2,627 (61%) in the minor extension class, and 2,944 (72%) in the major extension class. In total, using histone modification data we were able to identify 13,159 loci requiring further investigation. Either encompassing misannotations or potential novel loci which have gone unannotated until now. With these regions identified, we were then interested to see if we could validate these hypothesized annotations.

Validation of Hypothesized Annotations

After identifying putative misannotations, we sought to validate these hypothesized annotations by reassembling transcripts at the specified locus using more inclusive computational parameters. In parallel to assembling transcripts from RNA-seq data, we also utilized full-length transcript isoform sequencing using PacBio Iso-seq reads from multiple studies (see methods) to evaluate hypothesized annotations. Overall, 67% (335) of the hyper large class of genes were

validated by both long-read sequencing, as well as re-assembled transcripts from short reads, and 22.5% (115) were supported by one of these data types (**Fig. 2.4A**). For the major extension class, 45.7% (1,856) of regions were supported by both RNA-seq and Iso-seq reads, with 28.5% (1,157) being validated by a single data type, and 25.9% (1,051) of major extension annotations being unsupported. In the gene merger class, 47.9% (174) of the hypothesized mergers were validated with RNA-seq and Iso-seq, 13.77% (50) supported by a single data type, and 38.9% (139) had no additional support. In total, 68% (2,698) of the minor extension class were validated by at least one alternative data source. For the novel class of annotations, we re-assembled regions using RNA-seq from the corresponding tissue in which the novel region was identified. In total, 72% (3,253) of the novel loci were supported by an assembled transcript. Overall 6,385 out of 9,213 of the potential misannotations were found to be corroborated from orthogonal datasets, demonstrating the capacity of histone modification data to allow for identification of potentially misannotated regions, and hypothesis-driven annotation correction.

We next evaluated how the distribution of gene length shifted after reannotation for each

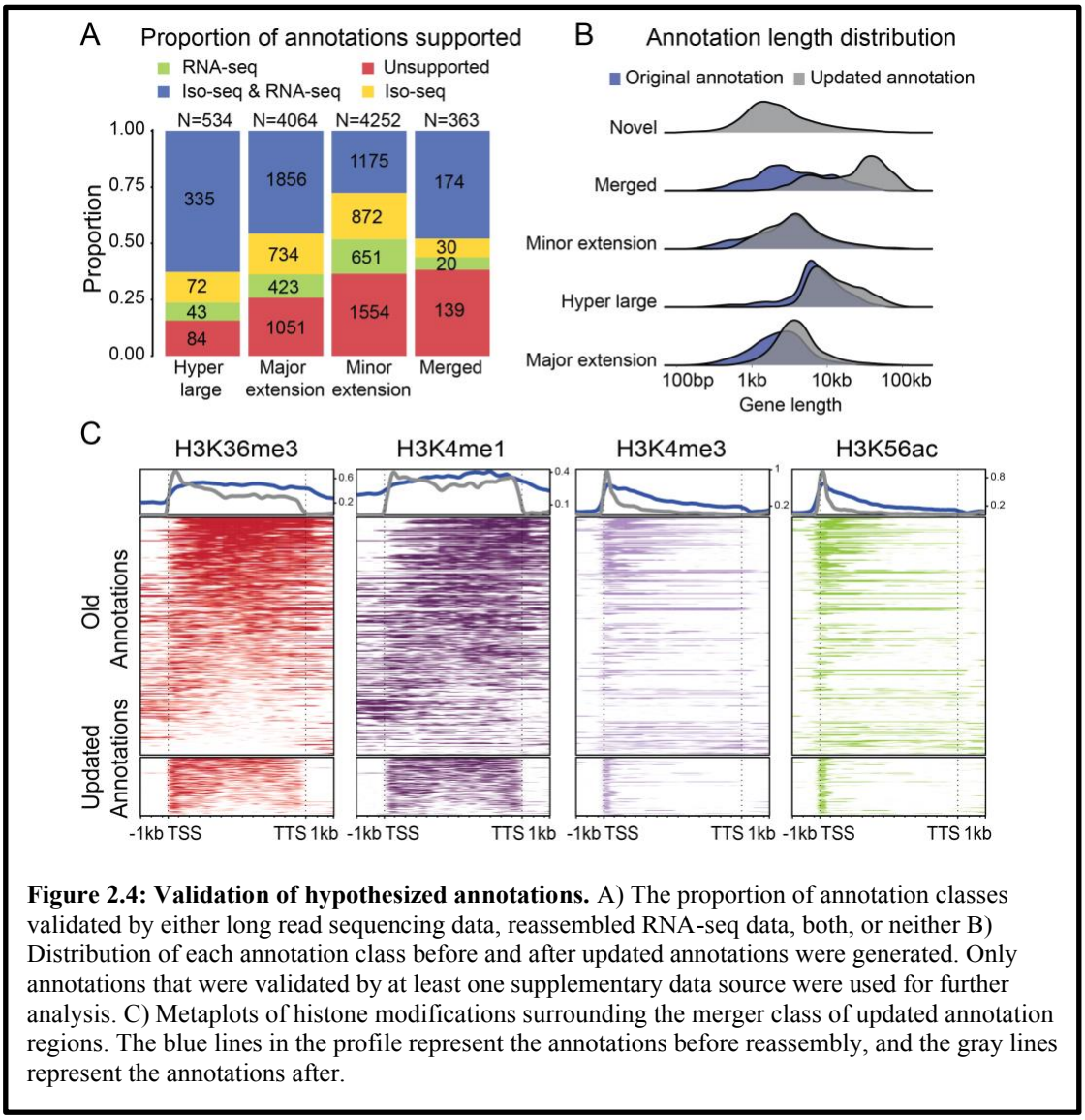

Figure 2.4: Validation of hypothesized annotations. A) The proportion of annotation classes validated by either long read sequencing data, reassembled RNA-seq data, both, or neither B) Distribution of each annotation class before and after updated annotations were generated. Only annotations that were validated by at least one supplementary data source were used for further analysis. C) Metaplots of histone modifications surrounding the merger class of updated annotation regions. The blue lines in the profile represent the annotations before reassembly, and the gray lines represent the annotations after.

class. Only loci which had at least one type of supplementary support (Iso-seq reads or RNA-seq) were used for this analysis (**Fig. 2.4B**). Overall, the distribution of the merged class was the most radically changed, as the median gene size shifted from 3,089 bp in length to 29,704 bp (**Supplementary Figure 4)**. In contrast, the median gene size for the novel class is 1,962 bp, smaller than the median known gene size for maize which is 2,568 bp (Portwood et al., 2019). The major extension gene class shifted from a median size of 2,363 bp to 3,818 bp , the minor extension class shifted from 3,165 bp to 3,631 bp, and the hyper large gene class shifted from 6,838 bp to 10,909 bp. To determine if the re-annotated regions more accurately recapitulated the

expected distribution of histone modifications around a transcribed gene, we regenerated metaplots. The updated annotation sets more accurately reflect the known landscape of histone modifications around transcribed units (**Fig. 2.4C**). This trend appears similarly in all found annotation classes (**Supplementary Figure 2.5-6**). This implementation of histone modification data allowed us to recapture previously unannotated regions in the genome of *Z. mays* while also improving existing annotations. All updated annotation coordinates are found in **Supplementary Table 2.2.** Additionally, we compared the class of merged annotations against a known list of 78 split annotations pairs available on Gramene (Tello-Ruiz et al., 2020). In total, 31 (40%) of the Gramene split annotations were concordant with the annotation mergers identified by our methods. The remaining 47 split gene pairs were either in regions where there was missing data (20/78), mappability issues (20/78), or were missed due to complex loci with multiple gene features in diverging directions (7/78). Comparisons between the merged dataset and the Gramene split gene dataset is in **Supplementary Table 2.2**. Although the Gramene split list is a set of well documented split errors in the maize genome, more recent studies using comparative annotation-based approaches have also been implemented, posing an excellent opportunity to compare and contrast the use of histone modifications.

We were interested in comparing the merged annotation group against a recent study that aimed to improve annotation of the maize genome by comparing annotations of numerous *Zea mays* cultivars (B73, PH207, and W22) against one another (Monnahan et al., 2020). By utilizing a blastp based approach for identification of potential gene merger pairs, followed by an analysis focused on variation in expression patterns across tissues, they identified split gene pairs that should be merged across the genome (Monnahan et al., 2020). In total, 109 (48%) of the merged annotation class identified in this study intersected gene merger pairs identified in the Monnahan et al study. Out of these 109 cross captured merger pairs, 34 were represented in the high confidence gene merger class identified in Monnahan et al. Additionally, 60 out of the 109 (55%) of the mergers found at the intersection of our studies fall into instances where they were

identified in Monnahan et al, but unable to be confidently classify based off of differential RNA-seq analysis. The histone modification data in our study provides clear evidence independent of RNA-seq that these 60 loci should be merged **(Supplementary Figure 2.7)**. Finally, there was a small class of 15 loci (14%) that were discordant between the two methods. With the histone modification data supporting gene merger, whereas the analysis by Monnahen et al. identifies that these loci should remain as split pairs. All intersecting annotations found between our studies are in **Supplementary Table 2.2**. Overall, the concordance between these gene merger sets demonstrates the inherent challenge associated with genome annotation while also demonstrating the advantage ChIP-seq provides as an orthogonal assay to RNA-seq based methods for gene annotation.

Knowing that Monnahan et al identified 96 high-confident gene merger pairs, we were interested in further investigating the remaining 62 pairs to ascertain why these potential misannotations were not identified by our method. Of the remaining 62 high confidence candidates identified by Monnahan et al., the histone modification data indicates that 30 of them should not be merged **(Supplementary Figure 2.8)**. The data provided by ChIP-seq provides strong evidence of distinct genes possessing their own transcription start sites and evidence of unique transcriptional elongation activities at each gene **(Supplementary Figure 2.8)**. Upon individual inspection of the remaining uncaptured 32 high confidence merger pairs identified by Monnahan et al. these candidates existed in either low mappability regions of the genome (10/32) or did not intersect a combination of histone modification enriched domains within the tissues that we sampled (22/32). The lack of being able to capture these loci is a limitation of our method; demonstrating the essentiality of utilizing many methods to improve genome annotations.

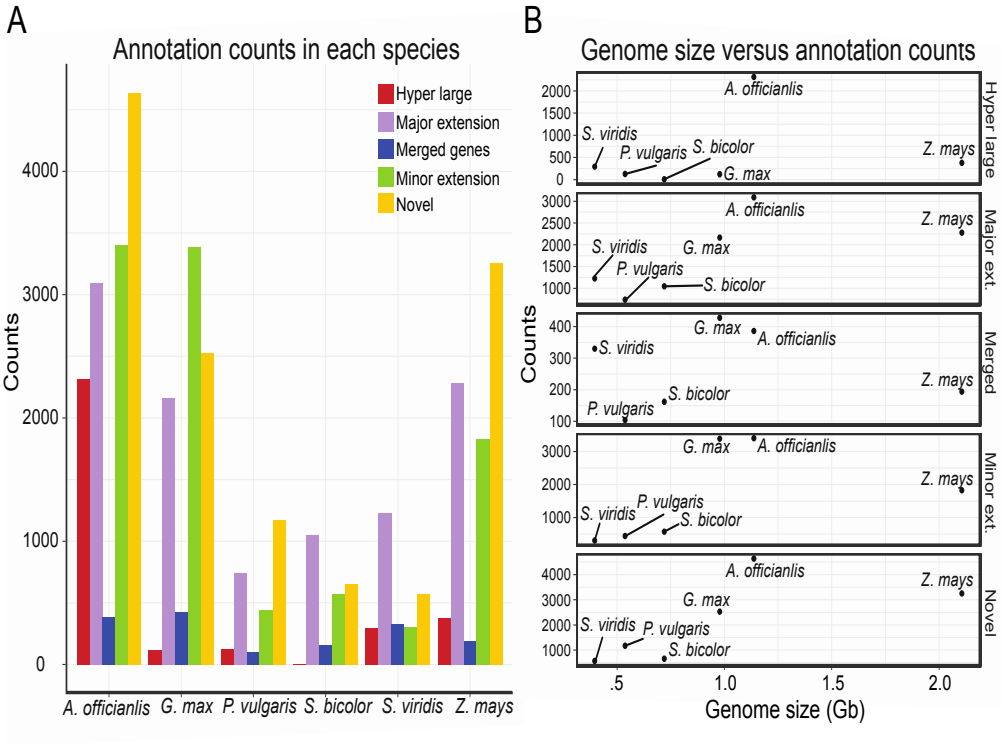

Figure 2.5: Reannotation of diverse plant genomes using epigenomic data. A) Counts of each annotation type identified in each plant genome. B) Scatterplot of genome size of each species annotated versus the number of annotation counts in each class

Reannotation of Multiple Plant Genomes

After successfully applying this method in *Z. mays*, we were interested in extending this method to other plant genomes with available high-quality histone modification data. In total, we included an additional five species, *Asparagus officinalis*, *Setaria viridis*, *Sorghum bicolor*, *Glycine max*, and *Phaseolous vulgaris* (Z. Lu et al., 2019). In total, we identified 4,640 novel annotations present in the *Asparagus officinalis*, 3,404 minor extensions, 386 potential gene mergers, 3,090 major extensions, and 2,316 hyper large extensions (**Fig. 2.5A**). The abundance of potential novel transcripts identified in *A. officinalis* was unexpectedly high, given that only a single tissue type, leaf, was used for this analysis. This stands in contrast to *Z. mays*, where across the three tissue types sampled, we found 4,004 potential novel regions. In *G. max* we found a

further 121 hyper large extensions, 2,165 major extensions, 3,388 minor extensions, 428 merged genes, 2,529 novel annotations. The annotations and relative counts of each annotation class found in each species are found in **Supplementary Table 2.2**. This analysis clearly demonstrates that histone modification data can be utilized on diverse plant genomes to quickly assay the quality of genomic annotations in a given tissue type.

Upon generating a list of hypothesized annotations, we noticed a slight trend in regards to genome size, and putative annotation errors. We noticed that the smaller genomes that we sampled, namely *P. vulgaris, S. bicolor, and S. viridis* appeared to have smaller number of potential genome annotation errors as compared to larger genomes. By correlating genome size with the counts of annotation error of each type, we found that larger genomes have more errors in the extension and novel classes of genes (**Fig 2.5B**). Although this trend appears to be true for *G. max,* and *A. officinalis,* it breaks down for the largest genome sampled here, *Z. mays.* However, the fact that *Z. mays* doesn't continue this trend may reflect the attention that this plant garners; with less annotation errors reflecting the abundance of resources, and groups working it. This large proportion of hyper large and major extension genes also appears to reflect a certain level of bias when annotating plant genomes, as we capture more issues in regards to large gene classes in plant genomes which are larger, and likely have a history of transposon expansion around gene features, causing increased intron size (**Fig. 2.5B**).

Discussion:

In the post genome assembly era, annotation represents the next great hurdle in accurate genomic resource creation. Here, we demonstrate that histone modification data offers a valuable untapped resource to precisely improve plant genome annotations. By easily assessing the transcribed space of the genome and identifying domains enriched with histone modifications that correlate with specific transcriptional events, valuable hypotheses about annotation features can be generated. These hypotheses, such as identifying potential transcript length and location of

transcription start sites, can be used in a manner complementary to RNA-based methods to provide a way to quickly fix gene models, and generate more accurate genome annotations.

This study demonstrates the power and advantages of using histone modification data to generate hypothesis about the transcribed genic space, offering valuable orthogonal assay. By utilizing histone modifications on a genome-wide scale, we identified consistent trends where annotations were discordant with the expected distribution of histone modification data and identified five distinct classes of annotation errors. We validated a set of these annotations using RNA-based methods. In total, we were able to identify, and validate 7,930 annotation errors. Of these updated transcripts, 3,253 represent novel transcripts, demonstrating the capacity of histone modification data to capture previously unannotated genes. Upon correction and reannotation, these updated annotations more accurately reflected what is known about the histone modification landscape of transcribed genes and captured previously unannotated gene space.

Additionally, this study shows that the usefulness of epigenomic data is not unique to *Z. mays***.** To demonstrate this we assayed five additional plant genomes for possible annotation errors using this method, and found varying abundances of either novel or misannotations. We correlated the counts of annotation errors with genome size, and found a slight correlation between the two, although additional studies of a great number of species will be required to know if this is significant. The abundance of potential annotation errors found across these five species demonstrates the importance of having orthogonal support for gene annotations and illustrates the challenges in making accurate *a priori* assumptions about gene features in plants.

Annotation errors are a natural part of generating genomic resources. The complexity of genic space, paired with the tissue and cell type specificity of many genes, and the assumptions required in each *in-silico* step of annotation converge to create an exceptionally challenging problem. These myriad challenges make annotation errors an inevitability, and downstream curation a necessity. Currently, sophisticated community-driven approaches exist to identify and fix annotation errors, but these large-scale efforts are limited to only well-studied species. This

bias in community size greatly inhibits the potential value, as the species with assembled genome increase in diversity.

The methodology presented here offers a protocol to appraise current annotations, and potentially fill in this downstream gap. However, while valuable, it is important to note that this method is not a panacea. ChIP-seq remains a challenging experiment and is not used as frequently as compared to RNA-seq. The lack of publicly available data, as well as the limited number of the tissue types sampled diminishes utility of this method. However, the increased accuracy added to genome annotations due to this method certainly introduce the potential of ChIP-seq becoming a standard protocol when considering genome annotation methods. Having a sequenced genome is only the first step to creating a valuable biological resource, and the challenges facing the production of accurate genome annotations remain. Epigenomic data offers one powerful orthogonal resource which, when utilized correctly, can strengthen current efforts and mitigate some, but not all, issues of genomic annotation moving forward.

Methods:

Genome Versions and Annotation: The maize genome V4 and annotation set version 4.38 of the annotation were acquired from gramene and used for all analysis (Jiao et al., 2017). The asparagus genome was taken from the asparagus genome project (http://asparagus.uga.edu/tripal/). The genomes for other genomes were retrieved from phytozome version 13 with the most recent annotations used.

ChIP-seq Data Processing Peaks: Raw reads from five different ChIP-seq libraries consisting of two replicates each were used to identify regions of enrichment for the histone modifications H3K36me3, H3K4me1, H3K56ac, and H3K4me3, as well input genomic. Reads were trimmed using trimmomatic, and aligned to the genome using bowtie2, `--very-sensitive` (Bolger et al., 2014; Langmead & Salzberg, 2012). Only uniquely mapping reads were used for downstream

analysis. Peak calling was done for histone modifications known to have broad peaks (H3K36me3, and H3K4me1) using the software epic2 with the parameters "--false-discovery-rate-cutoff .1 --keep-duplicates", as well as MACS2 to identify smaller regions of enrichment using the parameters `callpeak --keep-dup all -g 2.6e9 -q .1`. Narrow peaks (H3K56ac and H3K4me3) were called using MACS2 with the parameters " --keep-dup all --extsize 147 -g 2.6e9 -p .05' (Stovner & Sætrom, 2019; Y. Zhang et al., 2008). Peaks which were within 480 bps of each other were merged. Intersection between replicates of the same histone modification were taken. The minimum and maximum distanced regions were taken between intersecting regions, and the results merged to give a single peak which overlapped the extent of both peaks.

RNA-seq Data Processing: Raw reads were trimmed using trimmomatic with default parameters (Bolger et al., 2014). Reads were aligned to the *Z. Mays* reference genome version 4 using the STAR aligner, and the values `--outSAMstrandField intronMotif --outSAMmapqUnique 255 --alignIntronMax 50000` (Dobin et al., 2013). TPM values were calculated using TPMCalculator from NCBI .

Generation of Heatmaps and Metaplots: ChIP-seq data was handled similarly to ChIP-seq peak calling, with only uniquely mapping reads used. Libraries were normalized by read number using the 'bamCoverage' command found in deepTools version 3.3.1, and normalized using Counts Per Million mapped reads (CPM) (Ramírez et al., 2014). Matrices were generated with the compute matrix function 'scale-regions' with parameters '-bs 20 -b 1000 -a 1000 --regionBodyLength 5000'. Matrices were loaded into a custom R script and the R library EnrichedHeatMap was used to plot heatmaps (Gu et al., 2018). Genomic input reads were subtracted from ChIP-seq signal to account for genome bias, and the 95 percent quantile of each data set was selected as the upper value.

Mappability Control: In order to ensure that we were controlling for potential mappability issues in our analysis we utilized Genmap version 2.3.0 (Pockrandt et al., 2020). We generated mappability scores at single base pair resolution for unique kmer size 75 (size of our ChIP-seq reads) for the entire maize genome using the flags '-K 75'.

Annotating The Genome Using ChIP-seq: A custom pipeline was developed to annotate the genome using peak calls from ChIP-seq. Current annotations were categorized as either being expressed, or unexpressed based on alignment of stranded RNA-seq reads (Greater than 5 RNA-seq reads), as well as overlapping peak calls correlating with gene body extensions (H3K36me3, H3K4me1). Annotations were considered "good" or "unaltered" if histone modifications H3K36me3 or H3K4me1 overlapped the length of the gene body, and the annotation overlapped a peak correlating with promoter transcription initiation (H3K56ac or H3K4me3) in the first 50% of the gene body. Expressed annotations which did not contain a peak correlating with a promoter were then further explored by searching upstream of the transcription start site. These extensions were only carried out when the transcription initiation peak, and region in between the gene body, and the transcription initiation peak had similar coverages of transcription elongation modification across these regions. This class dubbed the "extension class" was further sub-categorized based on the length of extension. Minor extensions being defined as an annotation being increased by less than 500 bp, or the length of a single exon, major extensions defined as increasing the length of annotation between 500 and 2,000 bp, and hyper large extension with protein-coding genes needing to be extended upwards of 2,000 bp. Novel annotations were classified as those regions with a corresponding transcription elongation peak, as well as a corresponding transcription initiation peak that did not overlap within any known protein-coding, or non-coding gene. Finally, the merged class of annotations were those in which extension

caused overlap with another annotation. At these loci the coordinates were shifted to encompass both annotations.

We avoided utilizing this method to split annotations due to the possibility of potential "split" annotations representing separate isoforms of the same transcriptional unit. Due to the hereteogeneous nature of cell types within plant tissues, the aggregate ChIP-seq signal wouldn't provide clear evidence of variable isoforms versus two separate transcriptional units.

Tandem duplication analysis: To test for tandem duplicates in the merger class, we generated a blast protein database containing the original protein coding sequences of all genes found in this class. Tandem duplicates were defined as those which had a percent identity greater than 50%, and could align to at least 50% of the query protein sequence length. Dotplots were also generated for all pairs, and manually inspected for obvious signs of duplication. Additionally, 68 gene pairs out of the 363 were removed from this analysis, as we identified a set of annotations with multiple genes annotated which were completely overlapping, and annotated to the same transcriptional start sites. These loci likely represent different isoforms of the same gene which have been misannotated, and were thus discarded from our tandem duplicate analysis.

Assembly and validation of updated annotations in maize: To validate updated loci, reads overlapping the hypothesized annotation regions were pulled from 23 strand specific tissue types of the maize tissue atlas (Walley et al., 2016). StringTie was used to assemble transcripts in each region with parameters "--rf -f 0.01 -a 2 -m 50 -c 3.0 -f 0.0". Updated transcripts were then compared to old annotations, and categorized as correct if the updated transcript was larger than the original annotations. For further validation, Iso-seq reads were gathered from three different array express projects E-MTAB-7837, E-MTAB-7394, E-MTAB-3826, E-MTAB-5957, E-MTAB-5915, and E-MTAB-5956, aligned using STARlong "--outFilterMultimapScoreRange 1 --outFilterMismatchNmax 2000 --winAnchorMultimapNmax 200 --scoreGapNoncan -20 --

scoreGapGCAG -4 --scoreGapATAC -8 --scoreDelBase -1 --scoreDelOpen -1 --scoreInsOpen -1 --scoreInsBase -1 --seedSearchLmax 30 --seedSearchStartLmax 50 --seedPerReadNmax 100000 --seedPerWindowNmax 1000 --alignTranscriptsPerReadNmax 100000 --alignTranscriptsPerWindowNmax 10000“ (Dobin et al., 2013; Wang et al., 2016, 2018, 2020). Predicted annotation regions were compared to Iso-seq alignments, and regions that had a corresponding Iso-seq alignment which was greater than the original annotation were considered as passing.

Reannotation of Other Species using chromatin data: Histone modification data was downloaded from a list of seven species from previous work; gene expression omnibus number GSE128434 (Z. Lu et al., 2019). Reads were downloaded from GEO, and treated identically as outlined in the ChIP-seq section of the methods. Identical read alignment, and peak calling were performed, adjusting for relative genome size in the epic2 and MACs2 to alter peak calling stringency (Stovner & Sætrom, 2019; Y. Zhang et al., 2008, p. 2). No replicates existed for other species.

Data Access: All novel data generated for this analysis can be found under the GEO accession number GSE160944. The code used to run the above analysis can be found on GitHub in the following repository, https://github.com/Jome0169/MendietaPablo_Annotation_Paper_scripts**.** Of special interest is the script Update_annotation.py, which implements the re-annotation pipeline discussed in the method section. Updated annotations and gene models for *Z. mays* can be viewed at the Plant Epigenome JBrowse Genome Browser.

Author Contributions: JPM and RJS led the conceptualization of this project. JPM led the analysis and writing of this manuscript. AM contributed assistance in editing figures, as well as

provided valuable feedback throughout the study. XZ and WAR both contributed ChIP-seq experiments.

Acknowledgments:

Thank you to all the Schmitz lab members for your consistent feedback, and special thanks to Katie Duval for her willingness to look over multiple rounds of figure generation and editing. Research reported in this publication was supported by the National Institute of General Medical Sciences of the National Institute of Health under award number T32GM007103. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. AM was supported by the NSF Postdoctoral Fellowship in Biology (DBI-1905869). This study was funded by support from the National Science Foundation (IOS-1856627) and the UGA Office of Research to RJS. RJS is a co-founder of REquest Genomics, LLC, a company that provides epigenomic services. JPM, XZ, and WAR declare no competing interests.

Chapter 3

Investigating the *cis*-Regulatory Basis of C_3 and C_4 Photosynthesis in Grasses at Single-Cell Resolution[1]

[1] Mendieta, John Pablo. Submitted to PNAS, 1/22/2024

Abstract:

While considerable knowledge exists about the enzymes pivotal for C_4 photosynthesis, much less is known about the *cis*-regulation important for specifying their expression in distinct cell types. Here, we use single-cell-indexed ATAC-seq to identify cell-type-specific accessible chromatin regions (ACRs) associated with C_4 enzymes for five different grass species. This study spans four C_4 species, covering three distinct photosynthetic subtypes: *Zea mays* and *Sorghum bicolor* (NADP-ME), *Panicum miliaceum* (NAD-ME), *Urochloa fusca* (PEPCK), along with the C_3 outgroup *Oryza sativa*. We studied the *cis*-regulatory landscape of enzymes essential across all C_4 species and those unique to C_4 subtypes, measuring cell-type-specific biases for C_4 enzymes using chromatin accessibility data. Integrating these data with phylogenetics revealed diverse co-option of gene family members between species, showcasing the various paths of C_4 evolution. Besides promoter proximal ACRs, we found that, on average, C_4 genes have two to three distal cell-type-specific ACRs, highlighting the complexity and divergent nature of C_4 evolution. Examining the evolutionary history of these cell-type-specific ACRs revealed a spectrum of conserved and novel ACRs, even among closely related species, indicating ongoing evolution of *cis*-regulation at these C_4 loci. This study illuminates the dynamic and complex nature of CRE evolution in C_4 photosynthesis, particularly highlighting the intricate *cis*-regulatory evolution of key loci. Our findings offer a valuable resource for future investigations, potentially aiding in the optimization of C_3 crop performance under changing climatic conditions.

Introduction:

Photosynthesis is one of the most critical chemical reactions on the planet whereby CO_2 is metabolized into glucose. Plants have evolved numerous variations of photosynthesis. The most common type of photosynthesis uses the enzyme ribulose 1,5-biphosphate carboxylase oxygenase (RuBisCO) in combination with CO_2 to generate phosphoglyceric acid. This three-carbon compound is then used in a redox reaction within the Calvin Benson cycle, where glucose is made. The production of this three-carbon compound is what gives this type of photosynthesis, C_3, its name. However, although widely evolved and found in many crop plants, C_3 photosynthesis struggles to perform in hot, arid conditions. In non-ideal conditions, O2 can competitively bind the RuBisCO active site, causing the formation of a toxic intermediate, and reducing photosynthetic efficiency and plant performance (Bowes et al., 1971). Due to increasing temperature caused by anthropogenic climate change, this reduction in photosynthetic capacity for key crop plants poses a major agricultural challenge (Wheeler & von Braun, 2013). However, other types of photosynthesis have evolved in hotter conditions and offer a model to potentially alter key C_3 crop plants to be more efficient.

The C_4 photosynthetic pathway is an example of a modified style of photosynthesis that is able to perform in hot conditions. In brief, C_4 works by sequestering key photosynthetic enzymes into two different compartments in the leaf made up of different cell types. These two cell types/compartments are bundle sheath (BS) cells, which in C_4 plants generally form a concentric ring around the vasculature, and mesophyll (MS) cells, which make up large portions of the non-vascularized leaf internal cells (Hatch, 1987). In the MS, CO_2 is imported, and converted to bicarbonate (HCO3-) by the enzyme carbonic anhydrase (CA). Bicarbonate is then converted to a four-carbon molecule oxaloacetate (OAA) by the O2-insensitive phosphoenolpyruvate carboxylase (PEPC). This OAA molecule made of a four-carbon compound (where C_4 derives its name) is finally converted into a stable metabolite, malate. Malate is then transported to the BS where it undergoes a decarboxylation process, by one of three different

types of decarboxylases, NAD-dependent malic enzyme (NAD-ME), NADP-dependent malic enzyme (NADP-ME), or phosphoenolpyruvate carboxykinase (PEPCK). This decarboxylation reaction releases a CO_2 molecule that enters into the Calvin Benson cycle. The generation and processing of intermediate molecules in cellular compartments allows for concentrated levels of CO_2 to interact with RuBisCO, reducing the inefficiencies mentioned above. Current C_4 crops such as maize (*Zea mays*), sorghum (*Sorghum bicolor*), pearl millet (*Cenchrus americanus)*, and finger millet (*Setaria italica*) excel in their ability to operate in adverse conditions.

Although the evolution of C_4 photosynthesis is a complex process, there is tantalizing evidence that engineering C_3 crops to do C_4 photosynthesis might be possible. One piece of evidence that points to this is that C_4 photosynthesis has evolved independently 61 times in different lineages of plants (Sage, 2016). These results indicate that most plant lineages have the genetic material capable of evolving into C_4 photosynthesizers. The *Poaceae* lineage of grasses exemplifies this, as C_4 photosynthesis has evolved independently at least 18 times (Sage et al., 2011). Interestingly, all of these species use the same core C_4 enzymes and steps, but many use different decarboxylation enzymes as mentioned above (Gowik & Westhoff, 2011; Grass Phylogeny Working Group II, 2012; Rao & Dixon, 2016). Furthering this hypothesis is the fact that many C_4 related genes originally evolved from either C_3 photosynthetic genes or key enzymes critical in core metabolism (Kajala et al., 2012; Sheen, 1999). For instance, PEPC is a key metabolism enzyme in the glycolytic pathways of the Krebs Cycle, with some copies being important in guard cell metabolism (Chollet et al., 1996; O'Leary, 1982; Outlaw, 1990). Instead of novel gene content being the main driver of C_4 photosynthesis, it's more likely due to the correct timing and compartmentalization of key enzymes into specific cell types. This raises the question, how is gene expression of these key C_4 enzymes regulated? Moreover, as C_4 has evolved multiple times convergently, have similar regulatory networks and paradigms been co-opted to alter when and where these key genes are expressed?

Cis-regulatory elements (CREs) are key players in gene regulation, as they both fine tune expression and provide cell-type specificity (Gowik et al., 2004; Kim et al., 2022; Marand et al., 2021; Meng et al., 2021). In brief, these regions operate as binding sites for transcription factors (TFs) that modulate molecular phenotypes. Previous work has shown that CREs could be key players in the transition to C_4 photosynthesis. This was demonstrated by taking C_4 genes from *Z. mays* and transforming them into *Oryza sativa,* a C_3 species (Matsuoka et al., 1993, 1994), which revealed that CREs from *Z. mays* genes were able to drive cell-type-specific expression in MS in *O. sativa* (Matsuoka et al., 1993, 1994). Additional analyses have implicated CREs as drivers in the evolution of C_4 photosynthesis. In the genus of plants *Flaveria,* which contains both C_4 and C_3 plants, one key difference in C_4 plants was a specific CRE driving gene expression in MS cells. This 41 bp motif named *Mesophyll expression module 1* is critical for cell-type-specific expression of *PEPC* in MS cells, a critical first step in the C_4 pathway (Gowik et al., 2004, 2017). Finally, four conserved non-coding sequences were identified to be critical in MS-specific expression of *PEPC* in monocots (Gupta et al., 2020). Furthermore, a recent cross-species study examining the binding sites of GLK, a conserved TF regulating photosynthetic genes, revealed that CREs can undergo rapid changes and result in diverse gene expression patterns without the need of altering the TF itself (Tu, Ren, et al., 2022). These findings show that CREs are important genetic elements that plants use for the evolution of C_4 photosynthesis.

Although some CREs critical for cell-type-specific expression of key photosynthetic genes have been identified, they've been restricted to those nearby the transcriptional start sites. This is due, in part, to the challenge of identifying CREs genome wide, as well as limitations in the isolation of BS and MS cells which is labor intensive and challenging. However, a recent study used a multi-omic approach in *Z. mays* BS and MS cells and found CREs genome-wide that might be critical in the cell-type-specific regulation of genes (Dai et al., 2022). One example is the identification of a potential distal CRE ~40 kb upstream of *SULFATE TRANSPORTER4* (*ZmSFP4)*, a BS-specific sulfate transporter (Dai et al., 2022). These results highlight the

complexity of *cis* regulation and the importance of identifying all CREs for a gene, not just those nearby the transcriptional start site. During the evolution of C_4 photosynthesis, it's unclear whether these CREs have been pre-established during evolution and co-opted for C_4 photosynthesis or if they evolved independently numerous times. Understanding the ways in which *cis* regulation evolves to control timing and cell-type-specific expression of C_4 photosynthesis genes would greatly assist efforts in engineering C_3 plants to be more C_4 like.

To investigate the role of CREs and their potential contribution in controlling key C_4 genes, we used single-cell indexed Assay for Transposase Accessible Chromatin sequencing (sciATAC-seq) to identify cell-type-specific CREs from five grass species representing diverse C_4 subtypes, as well as an additional C_3 outgroup. We investigated the cell-type specificity of both the core C_4 enzymes, and those which are unique to each photosynthetic subtype. Further, we identify CREs of C_4 genes, and find previously unknown cell-type-specific CREs that might be critical in C_4 gene expression. We find that some of these regulatory regions appear not just conserved in a single C_4 subtype, but in all of the C_4 species we studied. Finally, we leverage these data to find transcription factor binding motifs enriched in MS and BS cell types and use these motifs to catalog these regulatory loci.

Results:

Identification and Annotation of Cell Types in Diverse Species:

To investigate CREs in BS and MS cells potentially important in C_4 photosynthesis, we generated replicated sciATAC-seq libraries for four different C_4 species, comprising three different C_4 subtypes NADP-ME (*Z. mays, S. bicolor*), NAD-ME (*Panicum miliaceum)*, and PEPCK (*Urochloa fusca),* and a C_3 outlier species (*O. sativa*) (**Figure 3.1A**). Libraries were filtered for high-quality cells by first pseudo-bulking the sciATAC-seq libraries, and identifying accessible chromatin regions (ACRs). Using these ACRs, per nuclei quality metrics were then calculated such as fraction of reads in peaks, transcriptional start site enrichment, and total

integration events per nucleus (**Methods**). Nuclei found to have a high proportion of organellar reads were also removed, with values being adjusted on a per library basis (**Methods**). Clustering of cells was done on genomic bins, and with additional cells removed that had a high correlation with *in-silico* generated doublets, and clusters were removed that were skewed towards one replicate by greater than 75% (**Methods**)**.** After filtering on per nucleus quality metrics, we identified 16,060 nuclei in *Z. mays*, 15,301 nuclei in *S. bicolor*, 7,081 nuclei in *P. miliaceum*, 19,110 nuclei in *U. fusca,* and 5,952 nuclei in *O. sativa* (**Supplemental Figure 3.1**).

Due to variation in genome size and content, cell-type annotation for each dataset was done independently using the reference genome for each species (**Figure 3.1B**). We used multiple approaches to annotate cell types. Orthologs of key marker genes from *Z. mays* and *O. sativa* were identified using a phylogenetics based approach (**Methods**). This allowed for the identification of marker genes for specific cell types in a cross species context. To gauge gene activity of these marker genes, gene body chromatin accessibility was used as a proxy for expression (**Figure 3.1D**) (Cusanovich et al., 2018; Marand et al., 2021). Cell-type annotation was done manually taking into consideration marker gene chromatin accessibility, marker enrichment in clusters, as well as ontological relationships between cell types (**Supplemental Figure 3.2-19**). Due to the lack of marker genes for many cell types in plants, as well as the challenge of annotating a broad sample of species, we reduced resolution of our annotation across our datasets to ensure accurate comparisons between variable species (**Figure 3.1B**).

Deeper exploration of the list of marker genes from *Z. mays* showed conservation of gene body chromatin accessibility in markers for certain cell types (**Supplemental Table 1-2**) . As expected, for the C_4 plants, *RIBULOSE BISPHOSPHATE CARBOXYLASE SMALL SUBUNIT1 (SSU1) and RIBULOSE BISPHOSPHATE CARBOXYLASE SMALL SUBUNIT2* (*SSU2*) were enriched in BS cells compared to MS cells (**Figure 3.1C**), a pattern that was not found in *O. sativa.* Additionally, *PEPC1* showed MS-specific chromatin accessibility in all of the C_4 species sampled (**Figure 3.1D**). Additionally, we found conservation of marker genes like

SUCROSE TRANSPORTER 1 (*SUT1*) in companion cells and sieve elements, and *GLOSSY1* (*GL1*) in epidermis cells, indicating that these historically described marker genes are likely important in this diverse set of species. This analysis provides a first examination of core-C_4 marker genes' chromatin accessibility across a diverse sample of plant species at cell-type resolution.

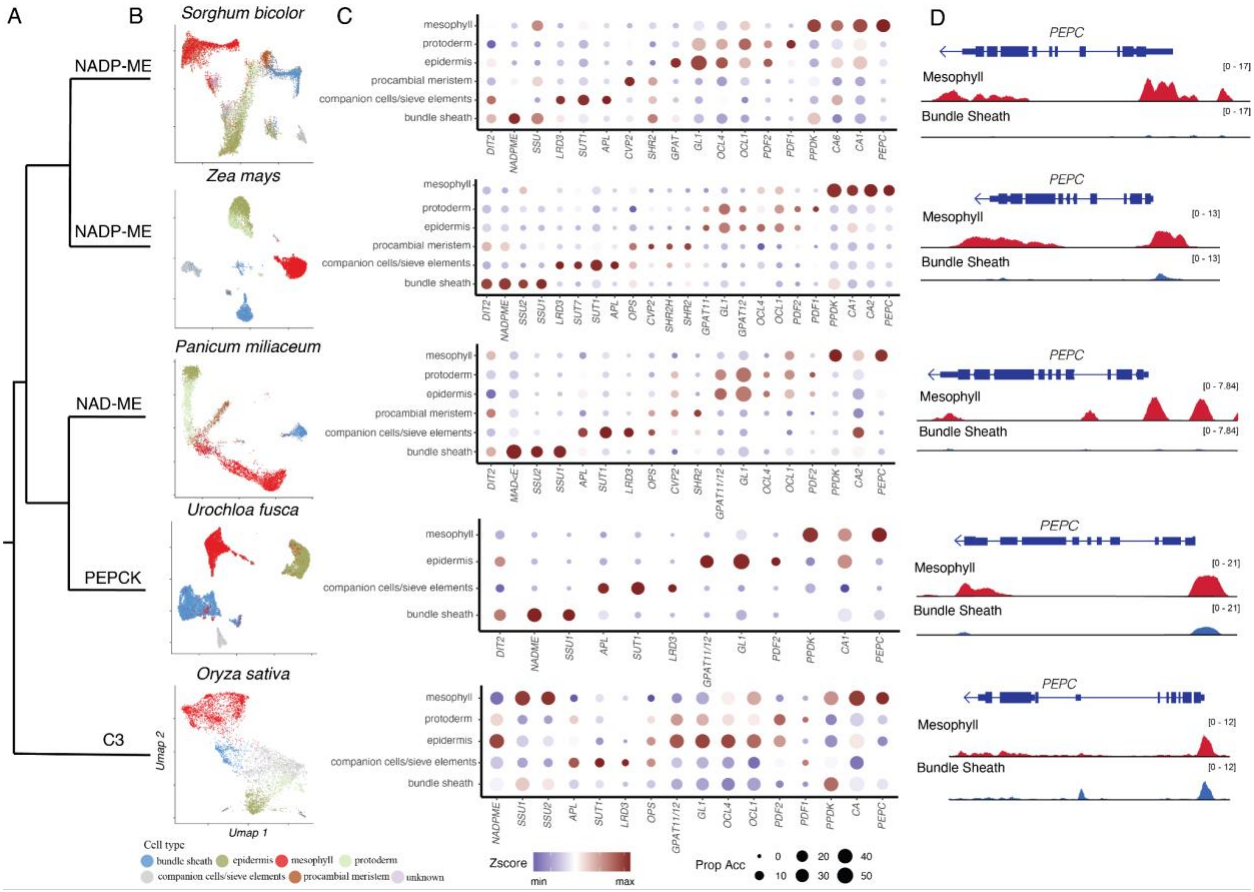

Figure 3.1: Annotation of cell types in diverse grass species at single-cell resolution **A)** A phylogeny indicating the relationship of various C_3 and C_4 photosynthesizers sampled. In this sample, two NADP-ME subtypes are represented, one NAD-ME subtype, a PEPCK subtype, as well as a C_3 species. **B)** UMAP embedding showing the annotation for each species. A cell type legend is below. **C)** Dotplots for various marker genes used to annotate each species. The y-axis represents cell types, and the x-axis is a list marker genes used to annotate different cell types. The size of each circle is proportional to the number of cells within that cell type that showed chromatin accessibility of the marker. Color is z-score transformed values across clusters of gene chromatin accessibility across the clusters. **D)** Screenshots of the *PEPC* locus for all sampled species. For each screenshot, the top track shows the protein coding, the red track is chromatin accessibility of MS cells, and the blue track is the chromatin accessibility of the BS cells.

Chromatin Accessibility of Core C_4 Enzymes Shows Similar Cell-Type Bias, but Differing Evolutionary Origins:

We measured the chromatin accessibility bias of the C_4-associated enzymes. Due to the diverse nature of the plants sampled, and the C_4 photosynthetic subtypes, we separated enzymes into core- and subtype-specific groups. This list comprised nine core C_4 enzymes, and nine variable enzymes. These enzymes were assigned to one of these two groups based on if they are found in all C_4 subtypes (core) or are specific to only one or two subtypes (variable). One example of a core enzyme is carbonic anhydrase, which is used to generate bicarbonate from CO_2, as well as for the regeneration of phosphoenolpyruvate from oxaloacetate in the BS cells by means of PEPCK (**Figure 3.2A**). The list of gene families that we considered as core or variable is found in (**Supplemental Table 3**).

To investigate the cell-type bias of these enzymes, we used chromatin accessibility of the gene (gene body as well as 500 bp upstream of the transcriptional start site) (**Figure 3.2B**). Cell-type bias was calculated as the $\log_2$ fold change of BS/MS chromatin accessibility. To identify core C_4 enzymes across these species, we used OrthoFinder, named and numbered the enzyme models based off of their relatedness to *Z. mays* copies of known core C_4 genes (Emms & Kelly, 2019). Using only cell-type-specific chromatin accessibility data, we observed expected cell-type bias with many orthologs of the maize MS-specific core C_4 genes showing MS-specific bias as compared to BS (**Figure 3.2C**). For instance, in all C_4 species, *PEPCK*, which regenerates PEP from OAA in BS cells, always showed a BS-specific bias (**Figure 3.2 A & C**). Additionally *PEPC*, which conconverts bicarbonate to OAA in MS cells, showed MS-specific bias for all species sampled, except the C_3 outgroup *O. sativa* (**Figure 3.2A & C**). These results highlight the quality of the data and the cell-type annotations for these single-cell datasets.

When analyzing these data in tandem with the phylogenetic trees, we noticed that some of the key enzymes showed different cell-type specificity based on their evolutionary origin (**Supplemental Figure 3.20-21A**). For instance, for carbonic anhydrase in *P. miliaceum,* the

orthologs that showed the largest bias between MS and BS cell types were not the copies that were the most evolutionary closely related to the *Z. mays* and *S. bicolor* cell-type-specific copies (Here *PmCA1* and *PmCA2*). Rather, a copy found in a separate clade (*PmCA3*) showed the most MS-specific bias (**Figure 3.2C**). This indicates that during the evolution of C_4, different sets of carbonic anhydrases were likely co-opted for C_4. One challenge using chromatin accessibility in this context, however, is the fact that neighboring gene models can occlude cell-type-specific signals. For instance, in the *S. bicolor* copy of *SSU1* (here *RBCS1)*, a BS-specific gene has a neighboring gene model directly upstream which shares a promoter region making measurement of the cell-type-specific bias of some loci challenging when using chromatin accessibility data (**Supplemental Figure 3.22**).

One unexpected result from this analysis was the lack of cell-type-specific bias for *MALATE PHOSPHATE ANTIPORT 1 (DIC1),* also known as *DICARBOXYLATE/TRICARBOXYLATE TRANSPORTER 1* (*DTC1*) in *Z. mays*. It has been previously reported that *DIC1* had BS-specific expression bias in *Z. mays* as well as in *P. miliaceum* (M. Taniguchi & Sugiyama, 1997; Y. Taniguchi et al., 2004; Tausta et al., 2014). However, there is not a clear signal based on the chromatin accessibility data. This could indicate that some ACRs harbor multiple CREs active in different cell types that are not obvious in chromatin accessibility data or that the cell-type-specificity observed is not due to *cis*-regulation, possibly involving post-transcriptional processes (**Figure 3.2C**). Lastly, as expected, there was very little bias in the C_3 outgroup (*O. sativa*). In total, 12/13 of the core C_4 enzymes showed cell-type-specific bias in *Z. mays*, 7/12 in *S. bicolor*, 16/21 in *P. miliaceum*, 11/13 in *U. fusca*, and finally 0/16 in *O. sativa*. These data demonstrate that chromatin-accessibility data can be leveraged to investigate the cell-type regulation of C_4 genes while also taking into consideration their evolutionary relationships in a cross species context.

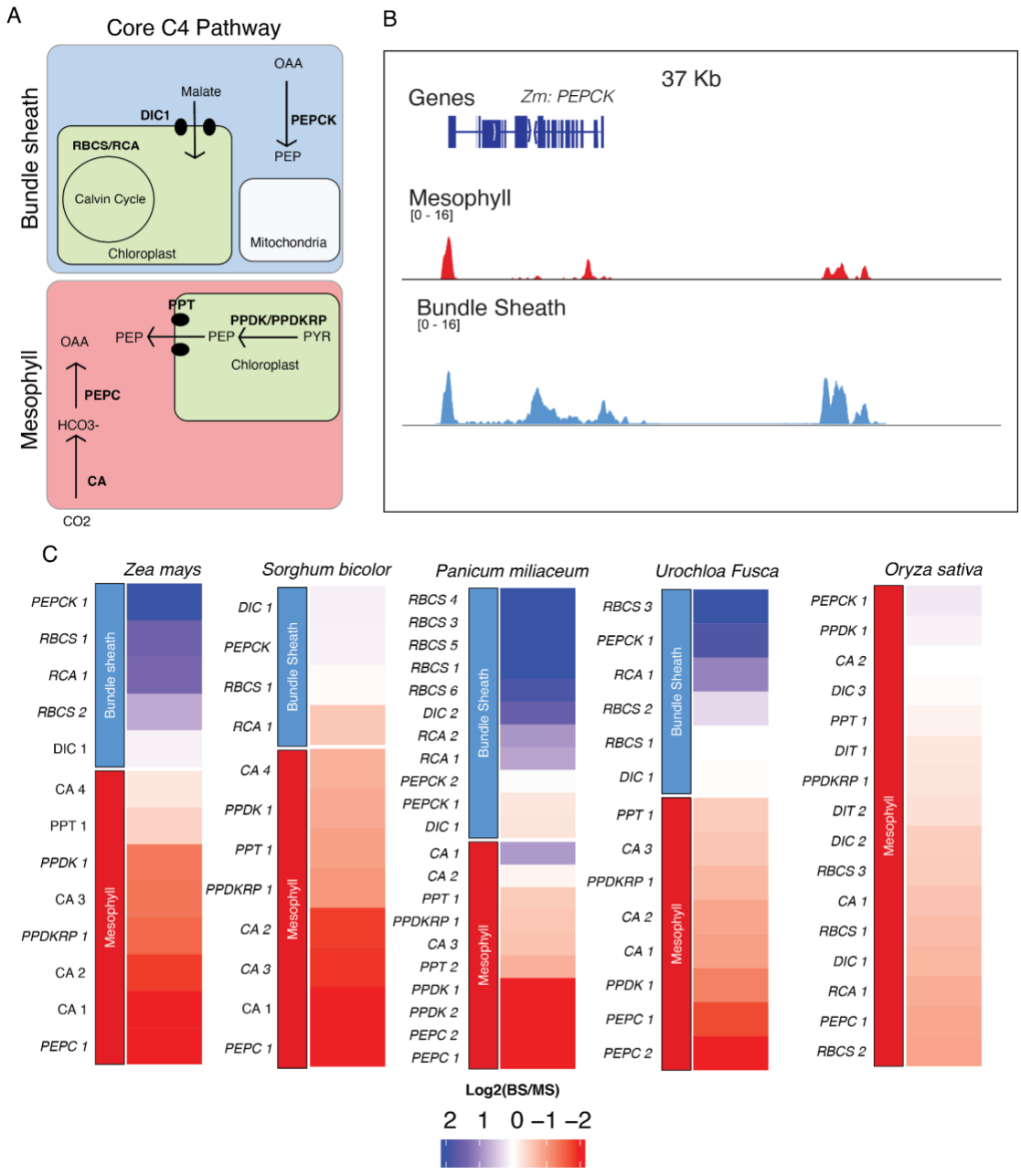

Figure 3.2: Cell-type chromatin-accessibility bias for core enzymes in C_4 and C_3 species. **A)** Schematic of the core C_4 enzymatic pathway. The red and blue squares represent MS and BS cells, respectively. Enzymes are labeled in bold, and transporters are denoted by shapes. Intermediate molecules are indicated by non-bolded text. **B)** Screenshot of *PEPCK* in *Z. mays*. Blue tracks correspond to BS chromatin accessibility and red tracks show MS chromatin accessibility. Tracks are equally scaled to facilitate comparison. **C)** Heatmaps of chromatin accessibility bias of the core C_4 enzymes. Values within each heatmap correspond to Log2(BS/MS). Blue indicates increased BS chromatin accessibility and red indicates increased

MS chromatin accessibility. Each species column and subtype was clustered independently, and genes were assigned as being MS- or BS-specific (top/bottom of heatmap) based on literature. Enzyme copies were distinguished phylogenetically.

Key C_4 Subtype Enzymes Show Potential Convergent Evolution in Cell-type-specific Bias:

We investigated the variable enzymes that give each C_4 subtype its unique properties by focusing on two species (*S. bicolor* and *Z. mays*) from the *NADP-ME* subtype (**Figure 3.3A**). As expected, chromatin accessibility bias was observed for enzymes previously reported as having cell-type-specific expression patterns, similarly to the core C_4 enzyme set (Borba et al., 2023; Dai et al., 2022). Reassuringly, one of the most biased enzymes identified was *NADP-ME*, the key enzyme of the redox step in *NADP-ME* subtypes. More specifically, of the multiple copies of *NADP-ME* that exist in *Z. mays,* we observed the expected cell-type bias for the known BS-specific copy, *ME3,* a key factor in C_4 (here *ZmNADP-ME1*) (**Figure 3.3B**). We noticed in *S. bicolor*, the BS-specific *NADP-ME* and the MS-specific *NADP-malate dehydrogenase* (*NADP-MDH*) gene copies are recent tandem duplications, each maintaining their respective cell-type specific chromatin accessibility (**Figure 3.3B** & **C**, **Supplemental Figure 3.21**). The malate transporters *DICARBOXYLIC ACID TRANSPORTER1/2 (DIT1/2)* also demonstrated their expected cell-type-specific bias with *DIT1* being MS specific and *DIT2* being BS specific in both species (**Figure 3.3B** & **C**). However, upon further inspection of the phylogenies of the *DITs* in *S. bicolor,* we noticed a pattern where the most BS-biased copy, *SbDIT2 (*Sobic.004G035500*)*, was phylogenetically more closely related to the *ZmDIT1*. These results indicate that over evolutionary time, even members of the same C_4 photosynthetic subtype, which likely share a C_4 ancestor, can use different paralogous loci to achieve cell-type-specific expression. This highlights that C_4 evolution is an ongoing process.

NAD-ME subtypes in *P. miliaceum* are interesting, as the intermediate molecule being passed between MS and BS doesn't take the form of malate, but instead aspartate, alanine, and oxaloacetate (**Figure 3.3D**). At least one copy of all of the key redox enzymes, NAD-ME and the

NAD-dependent malate dehydrogenase (NAD-MDH), show BS-biased chromatin accessibility (**Figure 3.3E** & **F**). Interestingly, of the three copies of *NAD-MDH* analyzed, only two showed bias for BS. Next, we evaluated two key enzymes associated with the generation of critical intermediate metabolites, Aspartate aminotransferase (AspAT), and Alanine aminotransferase (AlaAT). It has been reported that some AspAT have cell-type-specific expression patterns, with the MS-specific copy of the protein being transported to the cytosol and the BS-specific copy being transported to the mitochondria (**Figure 3.3E** & **F**) (Nomura et al., 2005; M. Taniguchi et al., 1992, 1995)**.** Of the four copies of AspAT we examined, two (*PmAspAT3/4*) showed significant MS-specific bias, whereas the other two copies (*PmAspAT1/2*) didn't show significant deviation towards BS (**Figure 3.3E**). This possibly indicates differing levels of regulation for the AspAT copies that did not show the expected BS bias, or missing copies of AspAT that we have not investigated. Within AlaAT, however, we identified one copy, *PmAlaAT1,* showing MS-specific bias, and *PmAlaAT6* showing BS-specific bias; something that has been previously hypothesized based on biochemical information (Son et al., 1991). Additionally, somewhat unexpectedly is that we didn't observe clear bias for sodium bile acid symporters (*BASS*) and sodium:hydrogen antiporters (*NHD*) (**Figure 3.3E**). These two proteins together form a functioning sodium bile acid symporter system, which balances the ratio of sodium and is important in the transport of pyruvate into the chloroplast of MS cells (Furumoto et al., 2011)**.** Although two copies of the *BASS* genes were MS biased, only a single copy of *NHD* was slightly MS biased. Surprisingly, we do observe slight cell-type-specific chromatin accessibility bias for malate transporter *DIT1/DIT2* in *P. miliaceum*. This is somewhat surprising, as malate is not the main 4-carbon intermediate used by NAD-ME subtypes (Rao & Dixon, 2016). This highlights the flexible nature of *P. miliaceum* in terms of its C_4 photosynthetic style, as it has been implicated that it can perform some of the metabolite shuttling as the NADP-ME subtype (Rao & Dixon, 2016; Wang et al., 2014; Zou et al., 2019). The potential flexibility of *P. miliaceum* in its style of C_4 makes it an extremely interesting species to study, especially when considering that it doesn't

share common C_4 ancestry with *Z. mays* or *S. bicolor.* This lack of evolutionary relationship makes the analysis of the more distantly related species *U. fusca* all the more valuable. These observations point to the complicated nature of some of these C_4 photosynthetic subtypes. While the obvious subtype-specific enzymes show expected chromatin-accessibility bias, others do not.

Using the *PEPCK* subtype in *U. fusca,* we evaluated cell-type bias of enzymes that operate as an intermediate between NAD-ME and NADP-ME subtypes (**Figure 3.3G).** Copies of *NAD-ME* and *PEPCK* showed significant BS bias (**Figure 3.3H** & **I**). Additionally, *NADP-MDH* was significantly biased towards MS, reflecting its critical role in the regeneration of malate from pyruvate (**Figure 3.3H**). We also observed one copy of *BASS,* which was heavily MS biased, as well as the only copy of *NHD* being highly MS biased (**Figure 3.3G**) (Washburn et al., 2021). Within the *BASS* family, based on the phylogenies, it appears one clade of *BASS* genes was co-opted to be MS specific, whereas the other clade remained somewhat BS specific. This potentially indicates that this co-opted clade may have been predisposed for C_4 photosynthesis at the common ancestor of *P. miliaceum* and *U. fusca.* Additionally, we also find one MS-biased and one BS-biased version of AlaAT (**Figure 3.3H**).

Finally, when evaluating genes in the C_3 outgroup *O. sativa,* we only observed significant chromatin accessibility bias for three of the 14 enzymes. This is expected given the overall lack of enzymatic bias seen in C_3 species (**Figure 3.3K**). Interestingly though, we did find a single instance where one copy of *AspAT* is BS specific, suggesting that this copy of *AspAT* might slowly be co-opted into being more BS-specific (**Figure 3.3K**). Even more interesting is the slight BS-specific bias of the rice *NAD-MDH*, a BS-specific enzyme in the *NAD-ME* subtypes. These results show a series of complex evolutionary relationships where many different genes can be co-opted into the C_4 pathway, and highlights the myriad ways in which C_4 evolution occurs.

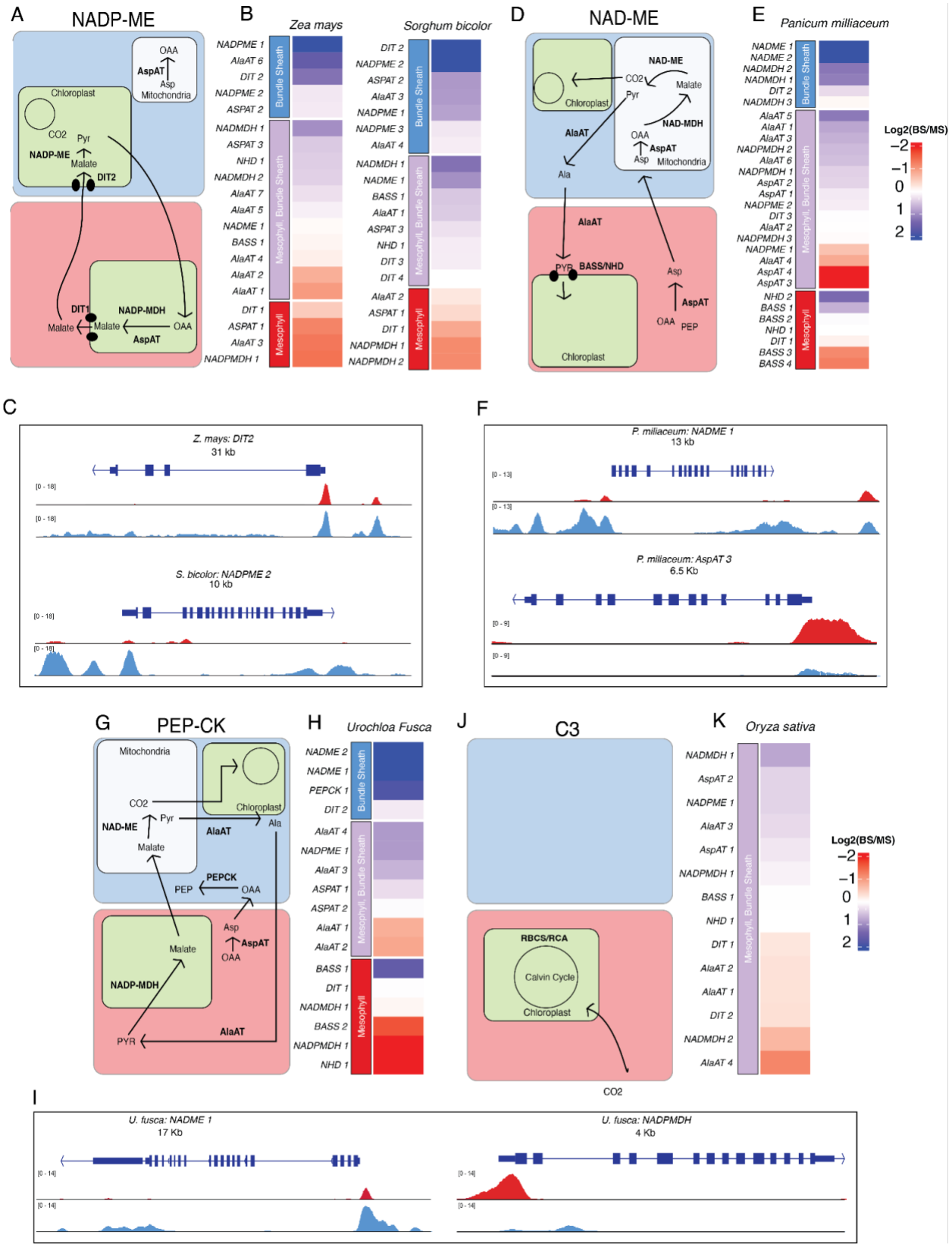
A
NADP-ME
OAA
AspAT
Asp
Mitochondria
Chloroplast
CO2
Pyr
NADP-ME
Malate
DIT2
DIT1
NADP-MDH
Malate
Malate
OAA
AspAT
B
Zea mays
NADPME 1
AlaAT 6
DIT 2
NADPME 2
ASPAT 2
NADMDH 1
ASPAT 3
NHD 1
NADMDH 2
AlaAT 7
AlaAT 5
NADME 1
BASS 1
AlaAT 4
AlaAT 2
AlaAT 1
DIT 1
ASPAT 1
AlaAT 3
NADPMDH 1
Bundle Sheath
Mesophyll, Bundle Sheath
Mesophyll
Sorghum bicolor
DIT 2
NADPME 2
ASPAT 2
AlaAT 3
NADPME 1
NADPME 3
AlaAT 4
NADMDH 1
NADME 1
BASS 1
AlaAT 1
ASPAT 3
NHD 1
DIT 3
DIT 4
AlaAT 2
ASPAT 1
DIT 1
NADPMDH 1
NADPMDH 2
D
NAD-ME
Chloroplast
CO2
Pyr
Malate
NAD-MDH
AlaAT
OAA
AspAT
Asp
Mitochondria
Ala
AlaAT
PYR
BASS/NHD
Asp
AspAT
OAA
PEP
Chloroplast
E
Panicum milliaceum
NADME 1
NADME 2
NADMDH 2
NADMDH 1
DIT 2
NADMDH 3
AlaAT 5
AlaAT 1
AlaAT 3
NADPMDH 2
AlaAT 6
NADPMDH 1
AspAT 2
AspAT 1
NADPME 2
DIT 3
AlaAT 2
NADPMDH 3
NADPME 1
AlaAT 4
AspAT 4
AspAT 3
NHD 2
BASS 1
BASS 2
NHD 1
DIT 1
BASS 3
BASS 4
Log2(BS/MS)
−2
−1
0
1
2
C
Z. mays: DIT2
31 kb
[0 - 18]
[0 - 18]
S. bicolor: NADPME 2
10 kb
[0 - 18]
[0 - 18]
F
P. miliaceum: NADME 1
13 kb
[0 - 13]
[0 - 13]
P. miliaceum: AspAT 3
6.5 Kb
[0 - 9]
[0 - 9]
G
PEP-CK
Mitochondria
Chloroplast
CO2
NAD-ME
Pyr
Malate
AlaAT
Ala
PEPCK
PEP
OAA
Asp
AspAT
OAA
Malate
NADP-MDH
PYR
AlaAT
H
Urochloa Fusca
NADME 2
NADME 1
PEPCK 1
DIT 2
AlaAT 4
NADPME 1
AlaAT 3
ASPAT 1
ASPAT 2
AlaAT 1
AlaAT 2
BASS 1
DIT 1
NADMDH 1
BASS 2
NADPMDH 1
NHD 1
Bundle Sheath
Mesophyll, Bundle Sheath
Mesophyll
J
C3
RBCS/RCA
Calvin Cycle
Chloroplast
CO2
K
Oryza sativa
NADMDH 1
AspAT 2
NADPME 1
AlaAT 3
AspAT 1
NADPMDH 1
BASS 1
NHD 1
DIT 1
AlaAT 2
AlaAT 1
DIT 2
NADMDH 2
AlaAT 4
Mesophyll, Bundle Sheath
Log2(BS/MS)
−2
−1
0
1
2
I
U. fusca: NADME 1
17 Kb
[0 - 14]
[0 - 14]
U. fusca: NADPMDH
4 Kb
[0 - 14]
[0 - 14]

Figure 3.3: Cell-type chromatin accessibility bias for variable C_4 genes associated with C_4 subtypes. **A/D/G/J)** Schematic of C_4 enzymatic pathways for various C_4 subtypes. The red and blue squares represent MS and BS cells. Enzymes are labeled in bold, and transporters are denoted by shapes. Intermediate molecules are indicated by non-bolded text. For clarity, core enzymes have been removed. **B/E/H/K)** Heatmaps of chromatin accessibility bias in C_4 subtype enzymes. Values within the heatmap correspond to Log2(BS/MS). Blue indicates increased BS-chromatin accessibility and red indicates increased MS-chromatin accessibility. Genes were labeled as being BS specific (blue) BS/MS specific (purple) or MS specific (red) based on previous literature. **C/F/I)** Screenshot of various C_4 sub-type enzymes and their chromatin accessibility profiles around the TSS. Blue tracks correspond to BS chromatin accessibility and red tracks show MS chromatin accessibility. Tracks are equally scaled to facilitate comparison.

Cell-type-specific Accessible Chromatin Regions of Both Core- and Subtype-Specific Enzymes:

Although measuring the gene body chromatin accessibility of C_4 enzymes is valuable, it doesn't inform us about the cell-type-specific *cis*-regulatory environment controlling these genes, as we only included 500 bp upstream in this initial analysis. To identify all potential CREs important for regulation of C_4 enzymes, we identified cell-type-specific ACRs using a modified entropy metric (**Methods; Supplemental Figure 3.33**)**.** In short, cell-type-specific ACRs are those which are unique to either a single cell-type or two or three cell-types in contrast to broadly accessible ACRs which are accessible in many different cell-types. For each C_4 enzyme, in both the core and the non-core set, we identified ACRs around them. We only considered ACRs to be potential regulators of a locus based on distance, with assigned ACRs needing to be less than 200 kb away from the target enzyme, and requiring that no other gene intervenes between the ACR and enzyme in question. In total, across all variable and core enzymes and taking into consideration only C_4 species, we find that on average, C_4 genes have between 2-3 cell-type-specific ACRs, with an additional 2-3 broadly-accessible ACRs (**Figure 3.4A**).

For all C_4 subtypes, the key redox enzymes all showed BS cell-type-specific ACRs, potentially identifying critical CREs for proper cell-type-specific expression. For instance, in *Z. mays, NADP-ME1* had five BS-specific ACRs, in *S. bicolor, NADP-ME2* had five BS-specific ACRs, in *P. miliaceum, NAD-ME1* had four BS-specific ACRs, and in *U. fusca, PEPCK,* had

three BS-specific ACRs (**Figure 3.4 A & C**). Additionally, of the MS-specific enzymes, we consistently observed numerous cell-type-specific ACRs around the carbonic anhydrase family. On average, there were 3.5 MS-specific ACRs for each copy of carbonic anhydrase across all of the species. This likely reflects the fact that carbonic anhydrase is critical in the initial steps of C_4, and also important in CO_2 sensing (Engineer et al., 2016). We also noticed an intriguing pattern where enzymes which were accessible in one cell type had cell-type-specific ACRs of the other cell type. For instance, around *SSU2,* a BS-specific enzyme, we found a series of MS-specific ACRs (**Figure 3.4D**). On average, we found 2.5 BS-specific ACRs around *SSU* and 1.5 MS-specific ACRs. This contrasting pattern was observed in key photosynthetic enzymes in all of the C_4 subtypes. This likely indicates that some of these ACRs contain CREs that negatively regulate *SSU* in MS, as cell-type-specific CRE usage has been implicated as being an important driver in proper compartmentalization (Bansal et al., 1992; Viret et al., 1994). The identification of ACRs around key C_4 enzymes provides a detailed map about potential *cis*-regulators of these loci, which provides the basis for future investigation into the direct function of each of these ACRs and how they might be altering transcription in multiple different ways. These results show that there are likely multiple ACRs important to cell-type specificity of these enzymes.

Traditionally, the field has focused on *cis*-regulation within a set distance from the transcriptional start site, often 1-2 kb, which is thought to generally encompass the promoter (Lu et al., 2019). However, we observed abundant distal cell-type-specific ACRs for many of these key genes (**Figure 3.4B**). For instance, the average distance of an ACR to its C_4 enzyme is 10,080 bp (*Z. mays*), 3,017 bp (*S.bicolor*), 4,260 bp (*P. miliaceum*), 2,358 bp (*U. fusca*), and 4,730 bp (*O. sativa*), indicating that the *cis*-regulatory space for these enzymes is far greater than previously appreciated, where a majority of the focus in the literature is on putative promoters. The genome of *Z. mays* emphasizes this point, as the subtype-specific enzyme *NADP-ME* has three cell-type-specific BS ACRs distal to the transcriptional start site, with the furthest being 34,336 bp away (**Figure 3.4C**). Interestingly, we found some enzyme/ACR pairs with opposite

cell-type-specificity (*i.e.* BS-specific enzyme, MS-specific ACR). Many of these ACRs were distally located. For example, in *Z. mays,* the MS-specific ACR of *RBCS* was 36,171 bp upstream (**Figure 3.4D**). When investigating ACRs around promoters, we were struck at how often cell-type-specific ACRs occurred outside of the bounds of previously analyzed promoters. For example, in *PEPC* in *P. miliaceum*, a recent analysis demonstrated that a series of conserved non-coding sequences found between species were able to drive MS expression (Gupta et al., 2020). When we looked at chromatin accessibility data of the promoter fragment which was cloned from *PEPC*, we identified many MS-specific ACRs within the cloned fragment, but an additional one upstream. This results shows the advantage of using scATAC-seq data to identify candidate CREs for certain genes, removing the guesswork of cloning fragments to investigate and providing a detailed cell-type-specific regulatory map of the locus (**Figure 3.4E**). Thus, scATAC-seq greatly improves the search space of the active CREs potentially driving cell-type-specific gene expression patterns.

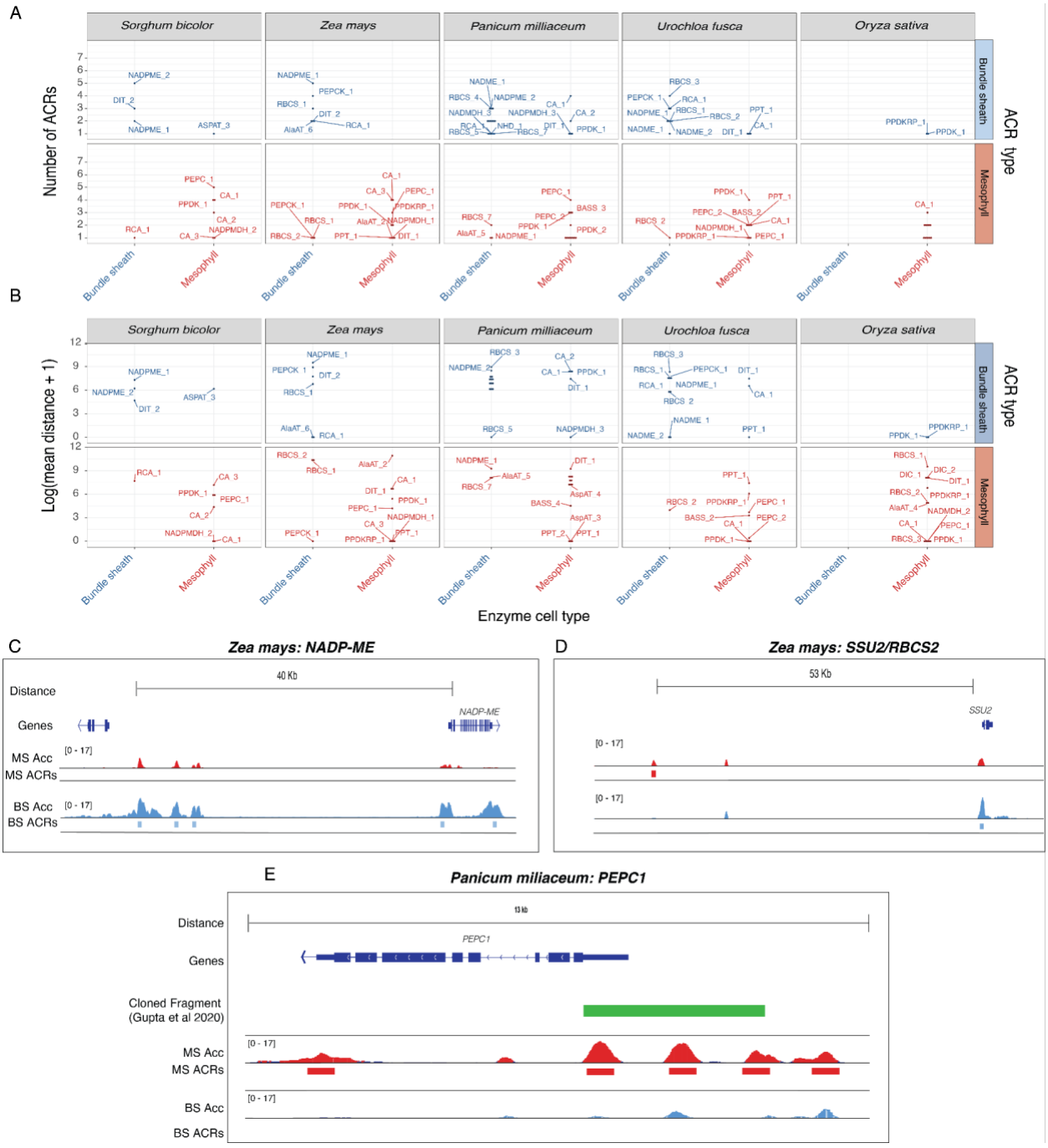

Figure 3.4: Investigating the number and distance of cell-type-specific ACRs around C_4 enzymes across subtypes. **A)** Dot plots showing the number of cell-type-specific ACRs around each enzyme. The x-axis indicates which cell type these enzymes are found in. The y-axis is counts of ACRs. The graph is further subdivided with the top panel being broad ACRs, middle panel BS-specific ACRs, and the bottom being MS-specific ACRs. Enzymes are labeled. **B)** Dotplots showing the mean distance of cell-type-specific ACRs to their closest C_4 enzyme. The x-axis indicates which cell type these enzymes are found in. The x-axis is the genomic distance to the C_4

enzyme in question. If an enzyme had multiple cell-type-specific ACRs, the distance was averaged (mean). **C)** Screenshot of *NADP-ME1* in *Z. mays*. Blue tracks correspond to BS chromatin accessibility and red tracks show MS chromatin accessibility. Tracks are equally scaled to facilitate comparison. **D)** Screenshot of *SSU2* in *Z. mays*. Blue tracks correspond to BS chromatin accessibility and red tracks show MS chromatin accessibility. Tracks are equally scaled to facilitate comparison. **E)** Screenshot of *PEPC1* in *P. miliaceum.* The green fragment represents the cloned promoter from Gupta et al 2020, which was identified by minimap2 alignment. Blue tracks correspond to BS chromatin accessibility and red tracks show MS chromatin accessibility. Tracks are equally scaled to facilitate comparisons.

The Evolutionary Relationships of ACRs Associated with C_4 Genes is Complex and Variable:

Next, we explored the evolutionary histories of these ACRs. Due to the fact that the C_4 subtypes come from different radiation events, (with *Z. mays* and *S. bicolor* likely sharing a C_4 ancestor and *U. fusca* and *P. miliaceum* sharing a different C_4 ancestor), we were curious to evaluate if a majority of the ACR space around these genes were either novel, or shared among these species. We implemented a pairwise sequence based approach by identifying sequence conservation of ACRs between the study species using BLAST (**Methods)**. The majority of important C_4 genes have both novel, and conserved ACRs. For example, PPDK, a MS-specific enzyme, shares ~25% of its ACRs across all species examined including the *O. sativa* C_3 outgroup (**Figure 3.5A**). Interestingly, *RUBISCO ACTIVASE (RCA)*, a critical enzyme in photosynthesis which removes inhibitory molecules from the RuBisCO active site, had novel ACRs in all of the C_4 species examined, whereas *RCA* in the C_3 species *O. sativa* shared one ACR with all of the C_4 species. This might indicate that each of the C_4 species gained regulatory sequences at *RCA* or that *O. sativa* might have lost them (**Figure 3.5A**). Focusing on NADP-ME revealed notable divergence in its associated ACRs, even among closely related species. For example, in *Z. mays*, two out of seven ACRs linked to *NADP-ME1* were unique, lacking counterparts in other species (**Figure 3.5A**). This is particularly striking given that *S. bicolor*, belonging to the same C_4 subtype, diverged from *Z. mays* only 13 million years ago (Paterson et

al., 2009). Similarly, in *S. bicolor*, the BS-specific *NADP-ME2* variant exhibited two out of five unique ACRs. This pattern underscores the rapid and distinct evolutionary trajectories of ACRs in C_4 plants. A full list of gene families, and gene models, and their relative conservation is found in **Supplemental Figure 3.24A.** Using this same approach to study all of the core class of C_4 enzymes did not reveal a generalizable pattern associated with gain or loss of ACRs around C_4 genes (**Supplemental Figure 3.24A**). Our findings not only confirm the dynamic evolution of *cis*-regulatory sequences in C_4 enzymes but also align with existing research that highlights rapid *cis*-regulatory changes among closely related species (Lu et al., 2019; Maher et al., 2018).

While investigating the ACRs around the C_4 genes is interesting, understanding how cell-type specificity is achieved across C_4 subtypes is needed for efforts to engineer C_4 photosynthesis. When looking at just the cell-type-specific ACRs around key C_4 loci, we find a similar pattern where there is a mix of both conserved and novel ACRs. For example, we discovered that some of the MS-specific ACRs associated with *PPDK* and *PEPC* are highly conserved in all of the studied species. Interestingly, the MS-specific ACRs around *PEPC* were only found in the C_4 species, and not in the C_3 outgroup, *O.sativa* (**Figure 3.5B**). This indicates that some of the CREs that allow *PEPC* expression in MS likely evolved after the split between the most recent common ancestors. We also observed that *NADP-ME* possessed numerous BS-specific ACRs that were conserved in all species, including *O. sativa* (**Figure 3.5B**). Considering the fact that proper compartmentalization of *NADP-ME* in BS cells is only critical in two of the four C_4 subtypes, this was surprising. However, in both *S. bicolor* and *Z. mays*, there were novel BS-specific ACRs associated with each key *NADP-ME*. In *Z. mays,* one out of the five BS-specific ACRs was novel to *Z. mays*, and in *S. bicolor* two out of the five were novel to *S. bicolor.* Upon inspection of all the *NADP-ME* loci in genome browsers, we were struck by the complexities and shuffling that occurred at these BS cell-type-specific ACRs (**Figure 3.5C**). These results highlight that extensive *cis*-regulatory evolution is occurring in each of these species, and in particular on a cell-type-specific level. Additionally, this may point to the fact that the novel BS-specific ACRs found

in *S. bicolor* and *Z. mays* may be more important for proper BS-specific expression than the conserved regulatory elements.

Although binary classification of ACRs was useful to decipher larger scale patterns between key enzymes, we next tested if larger segments of sequence were conserved around some C_4 genes as compared to others. We profiled the relative amount of conserved sequence at each of these ACRs, as alignment of sequence between species gives greater resolution about important ACRs. One interesting observation from this analysis was the fact that the cell-type-specific ACRs around *PEPCK* appear to be novel between *Z. mays* and *U. fusca* (**Figure 3.5D, Supplemental Figure 3.28**). This suggests that these regulatory loci emerged independently, and yet are still likely important in cell-type-specific expression of *PEPCK*. Additionally, around the *NAD-ME* loci in *P. miliaceum*, we found diverse evolutionary histories with both copies *NAD-ME1* and *NAD-ME2* having both conserved and novel BS-specific ACRs (one out of four ACRs were novel for *NAD-ME1*, and zero out of the two were conserved for *NAD-ME2*) (**Figure 3.5D**). The ACRs from *NADP-ME1* are conserved in *U. fusca*, whereas all three BS-specific ACRs are conserved in relation to *P. miliaceum.* Pointing to the fact that the ACRs have likely maintained their cell-type specificity, and are likely critical drivers in the correct expression of *NAD-ME* loci. These results highlight the dynamic evolution of cell-type-specific ACRs around key C_4 loci, and that even closely related subtypes have evolved novel ACRs potentially critical in terms of proper gene expression, as well as compartmentalization.

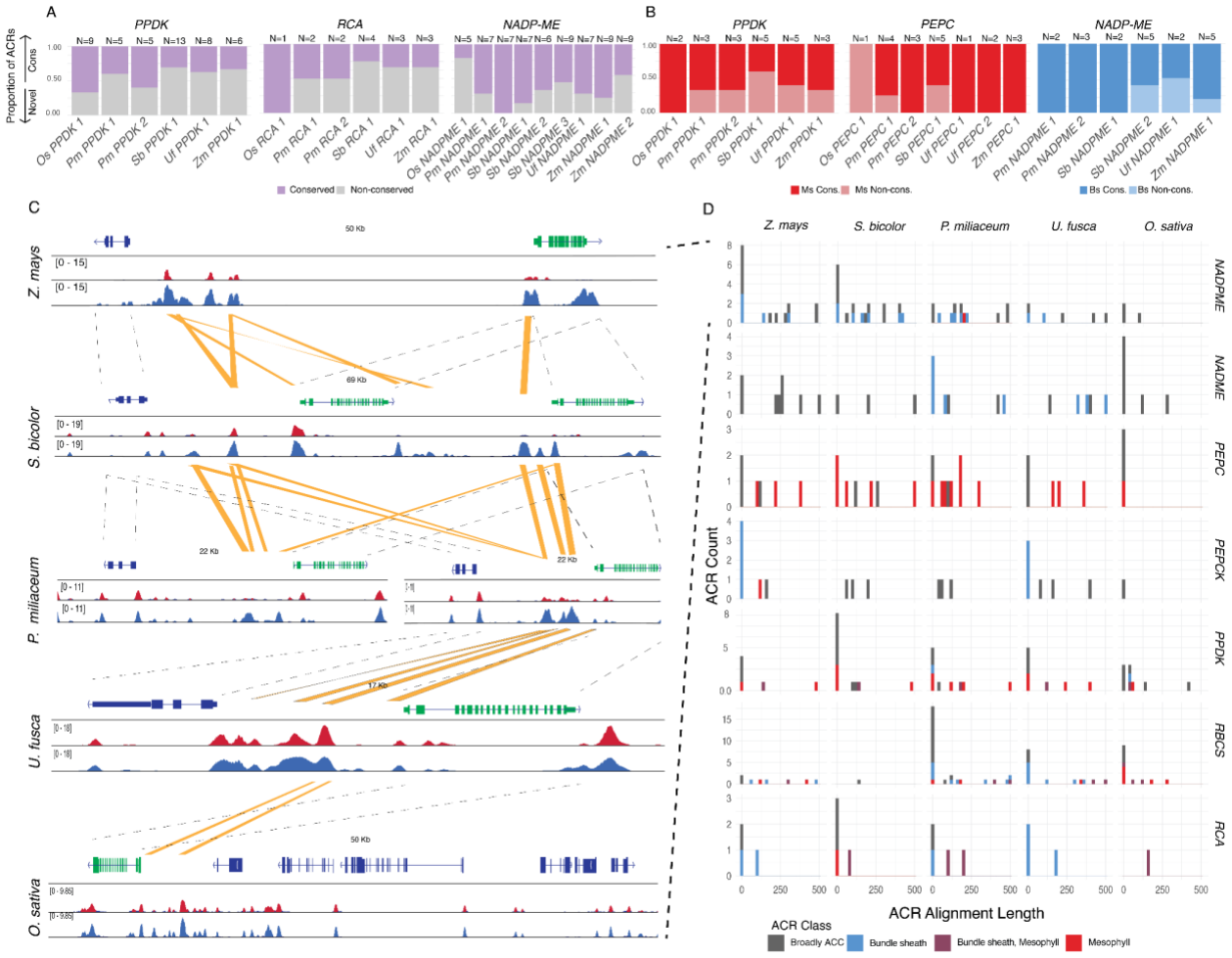

Figure 3.5: The evolutionary relationships of *cis*-regulatory regions around C4 genes is complex, being composed of both novel and conserved ACRs. **A)** The proportion of all ACRs that are conserved or novel for the following gene families *PPDK*, *RCA*, and *NADP-ME*. Purple bars represent ACRs that have any sequence aligned to them from a different species, and gray represents ACRs where sequences are not alignable. The number of ACRs in each locus is labeled at the top of each column. **B)** The proportion of cell-type-specific ACRs that are conserved and novel for the following gene families, *PPDK*, *PEPC*, and *NADP-ME*. Red bars only consider MS-specific ACRs, and blue bars only consider BS-specific ACRs. **C)** Screenshot of the conservation of BS-specific ACRs around *NADP-ME* across species. From top to bottom the species are *Z. mays, S. bicolor, P. miliaceum, U. fusca, and O. sativa. NADP-ME* is annotated in green for all species. Dashed bars between gene models represent the same gene model, and yellow bars are conserved ACRs. Browser tracks are blue for BS, and red for MS. Browser tracks are scaled within each species to allow for direct comparisons. **D)** The length of ACRs that are conserved in a cross species context. Rows represent gene families, and columns represent species. Each histogram is the number of ACRs within the loci of that gene family. The x-axis is the length of the ACR that is conserved and the y-axis is the count. ACRs are color coded according to the legend.

Identification of *de novo* TF-Binding Motifs from Cell-type-specific Chromatin Data Reveals Rapid Sequence Diversification of ACRs

Leveraging the cell-type-resolved datasets, we identified *de novo* cell-type-specific TF motifs in BS and MS ACRs (**Figure 3.6 A & B; Methods ; Supplemental Figure 3.29**). We selected the BS-specific motifs based on motif similarity within C_4 species for BS, and motif similarity seen across all species for MS. Additionally for the identification of BS specific motifs, we identified motifs which didn't appear to have a corresponding motif in *O. sativa* (**Methods**). Reassuringly, within the BS-specific motifs, we identified a DOF TF motif, which is a key driver in the switch to C_4 photosynthesis (Dai et al., 2022; Perduns et al., 2015; Yanagisawa, 2000). In total, we identified three BS-specific motifs, and three MS-specific *de novo* motifs that are shared between the species sampled (**Figure 3.6 A & B**). We surveyed the C_4 ACRs for the presence and absence of these motifs to determine if they provide the information needed for cell-type specificity. We additionally overlaid our BLAST results from the previous analysis in order to explore the relationship between these motifs and conservation (**Figure 3.6C**)**.** A substantial number of motifs were present within the non-conserved regions of the ACRs. For instance, in one MS-specific ACR associated with *ZmCA3*,12/13 MS-specific motifs were found in non-conserved regions, suggesting these regions could be critical for driving the cell-type-specificity of this locus (**Figure 3.6D**)**.**

We expanded the analysis of BS- and MS-specific motifs in conserved and non-conserved regions of ACRs across key loci in the C_4 species. On average the MS-specific motifs are more conserved than the BS-specific motifs (**Figure 3.6E-F; Supplemental Figure 3.30**). Agreeing with previous models of C_4 evolution where some motifs that are MS specific have been co-opted to operate in C_4 photosynthesis (**Figure 3.6D**) (Kajala et al., 2012). Interestingly, we noticed a pattern where around *PPDK*, many of the MS-specific motifs appeared to be in non-conserved sequences for all of our species sampled **(Figure 3.6E**). This pattern is further

highlighted in both *NADPME*, and *NADME* loci, where a majority of the BS-specific motifs occurred in non-conserved ACR regions for *NADPME*. This pattern is more nuanced in the *NADME* ACRs, as *P. miliaceum* and *U. fusca* share a significant amount of conserved sequence containing BS-specific motifs in the ACRs, suggesting that the BS-specific regulatory changes associated with these motifs are important (**Figure 3.6F**). These results highlight the capacity of genome-wide single-cell *cis*-regulatory maps to pinpoint key TF motifs important for the evolution of cell-type specificity.

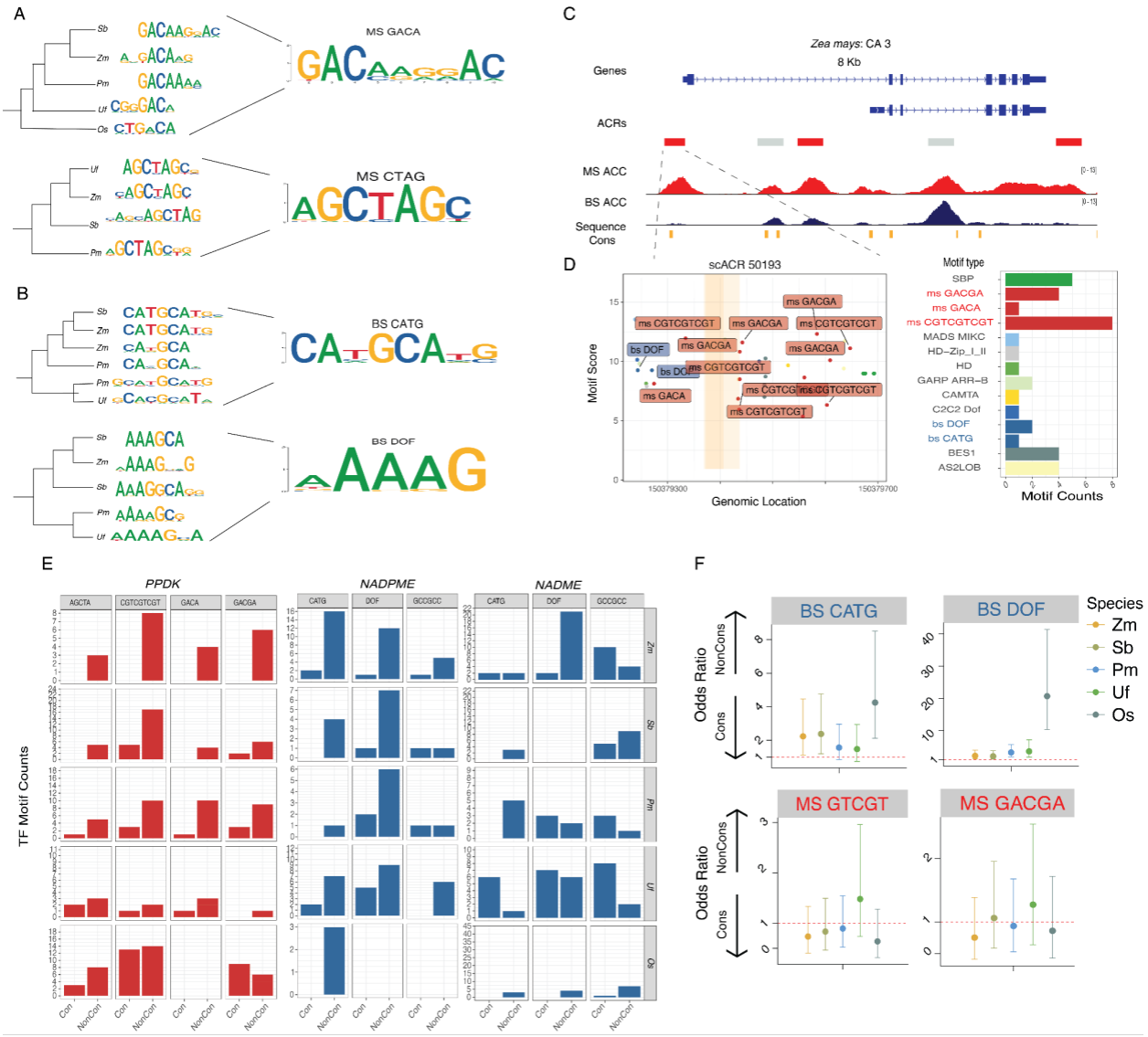

Figure 3.6: Identification of cell-type-specific TF motifs reveal a complex relationship between sequence conservation and motif prescnece. A subsample of MS- (**A**) and BS-specific (**B**) *de novo* TF motifs identified. **Left)** *De novo* motifs were clustered by the correlation of their PWMs and a correlation based tree was generated. **Right)** Representative PWMs from *de novo* discovery. **C)** Screenshot of the *ZmCA3* locus. ACRs are color coded based on their cell-type specificity. MS- and BS-chromatin accessibility tracks are equally scaled for comparison. Sequence conservation is identified by the ACR having sequence homology to other *CA* ACRs from a different species. **D)** An example of the conservation and motif landscape of one MS-specific ACR at *ZmCA3*. Left, the location of the motifs in ACRs with MS- and BS-specific motifs labeled. Orange highlighted regions correspond to the region of sequence conservation seen above. Right**,** quantification of the motifs found in the ACR. X-axis is the motif count, and the y-axis is the motif. **E)** The counts of TF motifs in conserved and non-conserved ACRs for three different genes across all five species. Y-axis is the number of ACRs of a given type, and the x-axis indicates the type of ACR. **F)** Odds ratio of four motifs when comparing their enrichment in conserved versus non-conserved regions. A higher odds ratio indicates that the motif is more often found in non-conserved regions within ACRs, whereas a lower odds ratio

means the motif is in conserved regions. The cell-type-specific motifs found in **A/B** are colored in red and blue, respectively.

The DITs in the NADP-ME Subtypes Demonstrate Dynamic CRE Evolution

Upon analyzing the malate transporters *DICARBOXYLIC ACID TRANSPORTER's (DITs)* we noticed the DITs in the NADP-ME subtypes showed an interesting pattern where the copies of *DIT1* in *Z. mays* and *S. bicolor* showed MS-specific chromatin accessibility, but the BS-specific copies of the *DITs* showed a more complex evolutionary history (**Figure 3.3B**; **Figure 3.7A**)**.** We generated a phylogeny with additional species, and found that the BS-specific copy of *ZmDIT2* is related to two additional copies of *DITs* which are not BS-specific in *S. bicolor* (Here *SbDIT3* and *SbDIT4*) (**Figure 3.7A**). *S. bicolor* has a BS-specific copy of *SbDIT2,* which shares a clade with *ZmDIT1*. These results are consistent with an earlier study that found similar patterns and expression profiles of these copies of the *DITs* in *Z. mays* and *S. bicolor* (Emms et al., 2016).

To understand how cell-type specificity changed in these DITs due to changes in *cis*-regulation, we compared the ACRs associated with the *DITs*, and mapped the TF-binding motifs found within each ACR (**Methods).** For the MS-specific *DIT1s*, we focused on a MS-specific ACR located at the 3' end of *DIT1* in *Z. mays* (**Figure 3.7B).** Upon comparing this ACR to *S. bicolor*, we were struck that the sequence found in the *Z. mays* ACR was actually split in two in *S. bicolor*, neither of which demonstrated cell-type specificity in *S. bicolor* (**Figure 3.7B ; Supplemental Figure 3.31).** A closer inspection of motifs in these ACRs showed many MS-specific motifs (**Figure 3.7B-C**). These motifs might promote MS-specific gene expression of this locus. However, many *S. bicolor* MS-specific ACRs were not found in regions with any homology to *Z. mays* (**Figure 3.7C**). These results point to the rapid change of candidate CREs (cCRES) in this locus, and likely indicate that cCREs important in cell-type-specific gene expression might not be only found in conserved regulatory regions (Yan et al., 2024). Rather, selection of MS-specific gene expression is ongoing, and may yield significantly different regulatory environments in relatively short evolutionary time scales.

Next, we examined the BS-specific *ZmDIT2* and its two orthologs *SbDIT3/4*, which are not BS specific (**Figure 3.7A, D**). The BS-specific ACR around *ZmDIT2* has many DOF TF motifs (**Figure 3.7E**). These motifs are interesting, as expression changes within the DOF TF family could be important in driving BS-specific gene expression in C_4 plants (Dai et al., 2022; Swift et al., 2023; Yanagisawa, 2000). When comparing the BS-specific ACRs around *ZmDIT2* to the more closely related copies of *SbDIT3* and *4,* we found no conservation of these DOF TF motifs, and rather a significant lack of BS-specific TF motifs (**Figure 3.7F**). Considering the fact that neither of these *DIT* copies in *S. bicolor* show BS-specific expression, this result makes sense. Potentially providing a model where the *ZmDIT2* locus either gained these cCREs allowing for this copy of *ZmDIT2* to have BS specific gene expression, or *S. bicolor* lost these BS-specific motifs, and had a gain in *SbDIT2* specificity. In either scenario, it demonstrates the rapid pace of CRE evolution, and how these regions might be altering cell-type-specific gene expression. These results are in contrast to *SbDIT2*, where the ACRs around this locus are BS specific, and contain BS-specific motifs identified in our previous analysis (**Figure 3.7F**). In total, these results highlight the rapid rate of regulatory change around key C_4 loci, and highlight the fact that there are likely key regulatory switches outside of conserved sequences. Finally, these results emphasize the fast pace in which cell-type specificity changes in plants

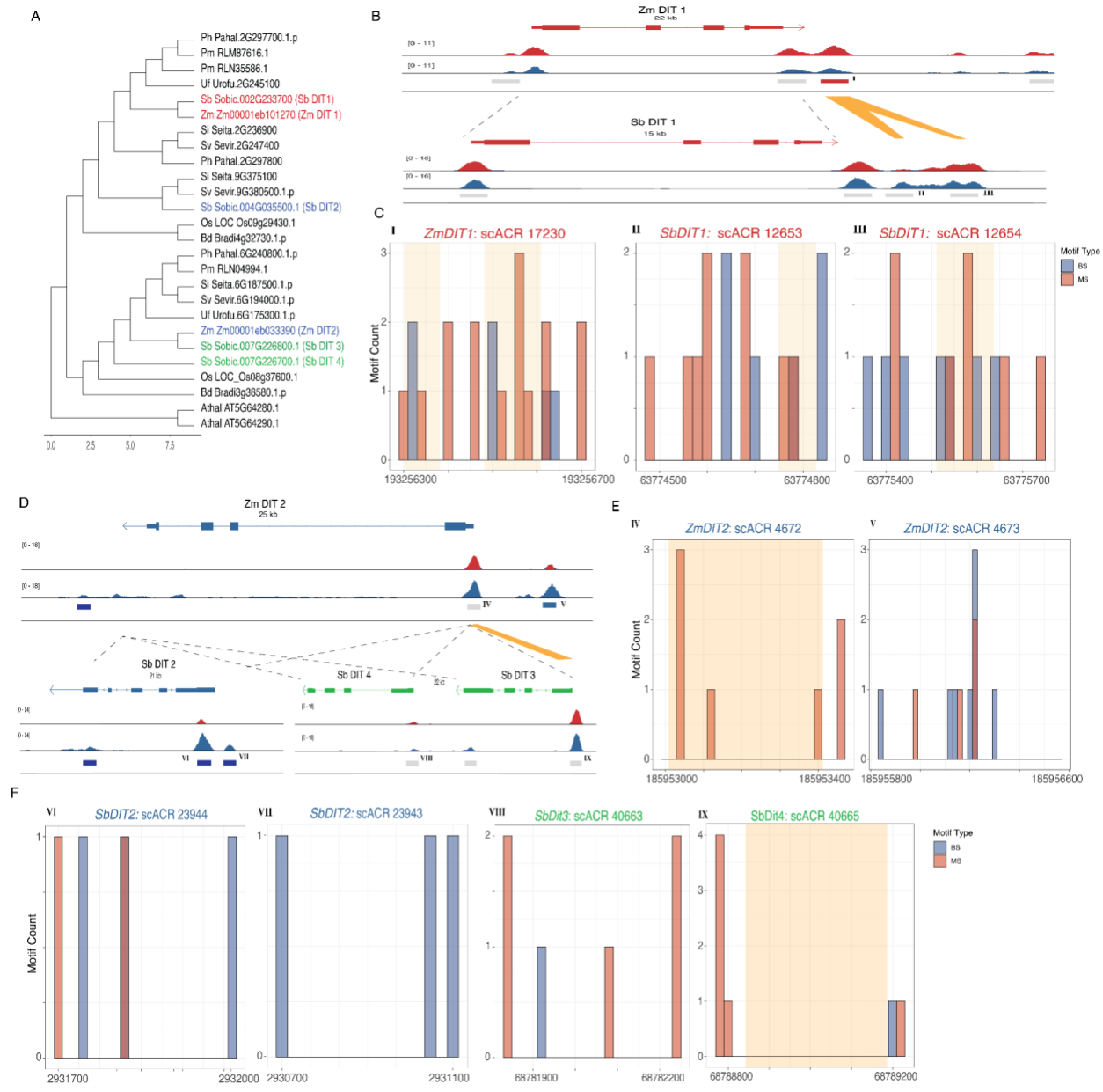

Figure 3.7: A) Phylogenetic tree showing the evolutionary relationship of the *DITs* in the monocots. *DITs* for *Z. mays* and *S. bicolor* are colored by their observed cell-type specificity, with red being MS specific, and blue being BS specific. Additional species have been added to increase resolution **B)** A screenshot of the *DIT1* between *Z. mays* (top) and *S. bicolor* (bottom). Yellow boxes indicate ACR sequences with conserved homology **C/E/F)** Motif location of BS and MS specific motifs in each ACR. The x-axis is the location within the ACR, and the y-axis is

the motif count. Yellow bars indicate regions of sequence homology. Within each histogram, the x-axis is binned into 20bp regions for ease of graphing. Roman numerals in the top corner highlight the corresponding ACR found in the screenshot. **(I-IX) top)** X-axis the genomic coordinates of the given ACR. Yellow blocks denote the sequence homology as seen above. Y-axis, the motif score as calculated by motifmatchR, higher scores indicate a more confident motif. **bottom)** The count of each motif identified in the ACR. Note that BS and MS *de-novo* identified motifs are in blue and red respectively. **D)** A screenshot of the *DIT2* loci between *Z. mays* (top) and *S. bicolor* (bottom). For the *S. bicolor* versions of the *DIT*s, *DIT2* is colored blue for its observed BS specificity and *DIT3* and *DIT4* are colored green. Yellow boxes indicate sequence homology.

Discussion:

Understanding the evolution of *cis*-regulation associated with C_4 photosynthesis has been a long standing goal in the field of plant biology. In this study, we demonstrated the utility of single-cell ATAC-seq data to investigate many aspects of the evolution of C_4 photosynthesis. By identifying cell-type-specific chromatin accessibility from four C_4 species composed of three different C_4 subtypes, as well as a single C_3 outgroup, we were able to compare and contrast key genes and their ACRs which define and distinguish C_4 photosynthesis. We have shown that by using gene-body chromatin accessibility data, we can measure cell-type-specific bias of both core, and subtype-specific C_4 enzymes. When taken into consideration with the gene family trees of many of these enzymes, we show diverse co-option of enzymes into the C_4 pathway. Additionally, we identify cell-type-specific ACRs surrounding these key C_4 enzymes. We find numerous cell-type-specific ACRs surrounding key C_4 enzymes, many of which fall outside of the core promoter region. Additionally we find that around all of the C_4 enzymes there is a mix of both conserved and novel cell-type-specific ACRs indicating that regulatory evolution of these

regions is ongoing. Finally, we use cell-type-specific ACRs to identify a series of *de-novo* binding motifs which appear to be cell-type specific, and show that these motifs surround C_4 loci, and have a mixed relationship with conservation depending on the motif. This indicates that cell-type-specific TF motifs are rapidly changing around C_4 loci.

Investigation of the CREs driving cell-type-specific expression of C_4 genes is challenging. This often requires evaluation using transgenic plants, which limits the number of CREs that can be tested. This has greatly hampered efforts at understanding how *cis*-regulation of C_4 genes evolves, whether by co-option of existing CREs or emergence of new ones. Our results show the complex nature of CRE evolution of C_4 genes, including those specific to C_4 subtypes. While we observe conservation of ACRs around many C_4 genes, we do see interesting examples where the subtype-specific enzymes have evolved novel ACRs (*NAD-ME*'s in *P. miliaceum*, and *PEPCK* in *U. fusca*). These results support that there is likely a combination of both co-opting pre-existing CREs, as well as evolving new ones to facilitate proper expression and cell-type-specification of genes. This is further exemplified by the analysis of the *DIT* family of transporters, where we show striking accumulation of cell-type-specific TF motifs in non-conserved regions of ACRs between two closely related species. This highlights that the regions of the genome promoting cell-type-specific gene expression are likely found in both conserved, and novel regions. Another recent single-cell genomic study of the evolution of CREs important for photosynthesis using a comparison between *O. sativa* and *S. bicolor* reached similar conclusions (Swift et al., 2023). They frequently found different ACRs and TF motifs in promoters of orthologous C_4 genes (Swift et al., 2023). Future efforts to assay these candidate CREs using reporter assays, transgenesis and genome editing will be required. Fortunately, these high-resolution maps of cell-type-specific ACRs of these key genes/species provide a strong foundation to build upon.

Although these studies provide a blueprint for the study of key candidate CREs associated with C_4 enzymes, profiling cell-type-specific chromatin accessibility of additional

species would be greatly beneficial. Although *O. sativa* is an invaluable outgroup for this study, additional more closely related C_3 species might make these comparisons simpler, and add additional resolution. For instance the C_3 grass species *Dichanthelium oligosanthes* is more closely related to *U. fusca* and *P. miliaceum* and has a recently completed reference genome (Studer et al., 2016). Adding more species would enable greater resolution in the comparison of cell-type-specific ACRs, as the genetic distance between the species we examined and *O. sativa* make identification of conserved and novel ACRs challenging. As an example, the ACRs associated with *NAD-ME's* in *P. miliaceum* might be co-opted instead of novel, however, based on our sampling, we cannot say.

Genome editing analysis of many of these ACRs would significantly advance which ACRs, and more specifically which CREs within the ACRs are most important for cell-type-specific expression (Meng et al., 2021). However, currently generating genome edits in monocots is challenging, time consuming and expensive. Fortunately, improvements to transgenesis are constantly improving making achieving these goals more likely in the future (Chen et al., 2022). It's also important to consider that mutational analysis of CREs is not straightforward, often requiring numerous editing events of the *cis*-regulatory landscape of each gene. Previous studies have shown that deletions of many CREs produce variable molecular and morphological phenotypes, further complicating our understanding of the *cis*-reglatory code (Ciren et al., 2023; L. Liu et al., 2021; Rodríguez-Leal et al., 2017). And finally, many species, including *P. miliaceum* and *U. fusca* have to date never been transformed. This highlights the need to continually improve transgenesis methods to help facilitate the molecular dissection of CRE. In conclusion, this study provides a comprehensive map of cell-type-specific ACRs around key C_4 genes, which reveals the dynamic evolution and diversity of *cis*-regulation of C_4 genes.

Acknowledgments:

This research was funded by awards from the National Science Foundation (IOS-2134912 and IOS-1856627) and the Office of Research to RJS and Hong Kong University Grant Council (GRF 1409420) to SZ. APM and JPM were supported by the National Institutes of Health (K99GM144742) and (T32GM142623), respectively. This research was additionally funded with support from the NSFC (32100438 & 32370247) and Shanghai Jiao Tong University 2030 Initiative (to X.T.).

Methods:

Plant Growth Conditions and Sampling:

Seedlings of all five plant species, including maize (*Zea mays* B73), sorghum (*Sorghum bicolor* BTx623), proso millet (*Panicum miliaceum* L. CGRIS 00000390), and browntop signalgrass (*Urochloa fusca* LBJWC-52), along with the C_3 plant rice (*Oryza sativa* Nipponbare), were grown under the conditions of 12:12 Light/Dark cycles at 30°C Light/22°C Dark and at 50% humidity. The sampling of the C_4 species was timed to coincide with a specific developmental stage, identified when the ligule of the third leaf became visible, marking the third leaf unfolding, yet prior to the appearance of the fourth leaf. For the C_3 species, rice, 18-day-old leaves were used to correspond with the equivalent stage of the C_4 species.

Library Preparation:

Nuclei isolation for the experiments was conducted using fresh seedlings of both the C_4 and C_3 species at their respective developmental stages. The methodology for nuclei extraction, encompassing the buffer composition and the subsequent steps, was used with procedures outlined for single-nucleus combinatorial indexing with transposed-based ATAC-seq library construction, as detailed in a prior study (Tu, Marand, et al., 2022).

Genomes:

The *Z. mays* genome version 5 was downloaded from MaizeGDB (Hufford et al., 2021; J. Liu et al., 2020). The *O. sativa* genome was downloaded from rice.uga.edu. The *S. bicolor* version v5.1 was downloaded and used from Phytozome version 13, as well as the *U. fusca* genome version 1.1 (Goodstein et al., 2012). Finally the *P. miliaceum* genome was downloaded from NCBI, bioproject number PRJNA431363 (Zou et al., 2019).

Barcode Correction Read Alignment and Mapping of Tn5 Insertions:

Read UMIs were processed using cutadapt (version 4.5) to identify UMIs (Martin, 2011). First, the index adapter sequences were trimmed from the reads. Next, the well barcodes and Tn5 barcode within the reads were identified, removed from the original sequencing read, and appended to the read header. Finally, a shell script is used to integrate all barcode information from the reads' headers and label them correspondingly in the paired-end sequencing fastq files. Reads were aligned using BWA (version 0.7.17) (Li & Durbin, 2009). Reads were filtered using samtools (version 1.16.1) for mapping quality of >10 for *Z. mays* , *S. bicolor* , *U. fusca*, and *O. sativa*. *P. miliaceum* required a greater lower threshold of 30 given its recent whole genome duplication event increasing the rate of multi-mapping reads (Danecek et al., 2021). Duplicate reads were removed using picard tools (version 2.25.0) (*Picard Tools - By Broad Institute*, n.d.). Single-base pair Tn5 integration events were mapped using the python script `makeTn5bed.py` found in the GitHub utils directory (https://github.com/Jome0169/Mendieta.C4_manuscript). Finally, for each barcode only unique Tn5 integrations sites were used for analysis. So if a nuclei had the same identical fragments multiple times, only a single event was considered.

Isolating High-Quality Cells:

Cells were filtered using Socrates (Marand et al., 2021). In short, Fraction of Reads in Peaks (FRiP) scores were calculated for each cell by pseudo bulking the libraries and identifying peaks. For each individual cell, FRiP was calculated by intersecting Tn5 integration events with peaks. Cells with a FRiP score greater than 0.2 were used. Additionally, TSS enrichment was calculated by looking at the number of Tn5 integrations around TSS. Cells that had a TSS enrichment greater than 0.15 were used. Finally, cells were compared to a random sample of low quality cells which did not pass filtering, representing the "background" of cells, and correlation was calculated between passing cells and background cells using the corr package in R. Cells which had a correlation lower than 0.3 percent as compared to background cells were used for further analysis.

UMAP embeddings were then calculated for each species utilizing genomic bins (McInnes et al., 2020). For each dataset, bins of 500 bp were calculated. To reduce the size of features to cluster on, bins had to show accessible chromatin in at least 0.005% of total cells (roughly 50~100 cells in each species). Additionally, bins that were broadly accessible across greater than 10% of cells in the given dataset were also discarded to remove regions of the genome which were constitutively accessible and wouldn't facilitate clustering. Finally, regions of the genome which were associated with either blacklist (Marand et al., 2021), or genes which were known to be related to cell cycle and circadian rhythms were removed. The final resulting matrix, which represented cell barcodes X genomic regions (here bins), were then put through the term-frequency inverse-document-frequency (TF-IDF) algorithm to identify genomic regions more descriptive of the entire dataset (Cusanovich et al., 2018). The resulting matrix was then input into Singular Value Decomposition, and clustering was then done on the remaining features with the number of principal components (PCs) equaling 50, and any PC with a correlation to read depth greater than 0.5 removed (Stewart, 1993) (Cusanovich et al., 2018). Clustering was done using the Louvain clustering algorithm in order to bin cells into similar groups based off of

the PCs calculated above, with parameters "res = 1.5, k.near = 30, m.dist = .01" in order to set K nearest neighbors to 30, minimum louvain distance to .01 in euclidean space (Blondel et al., 2008). Using the UMAP embeddings, doublets were removed using the software Scrublet as implemented in Socrates software (Wolock et al., 2019). At random, 5,000 cells were used to generate *in-silico* doublets, and cells which were scored as being likely doublets were removed. Adaptive thresholds were set on a per library basis. The doublet rate from Scrublet was compared against a mixed library where genotypes of *Z. mays* were mixed Mo17 and B73, and genotype doublets were identified. We found that Scrublet, on average, removed more cells in a conservative fashion than the birthday problem and genotype doublets identified, so we utilized the Scrublet doublet scores to be conservative. For the P. miliaceum dataset, replicates were found to integrate poorly in the UMAP embedding. Harmony (version 0.1.1) was used adjust replicate overlap with parameters "theta = 2, nclust=4, and var = "sampleID" (Korsunsky, Millard, et al., 2019). After integration, clusters which skewed greater than 75% towards one replicate were removed from downstream analysis.

Identification of Putative Orthologs:

To annotate species with less marker gene information, we identified putative orthologs or marker genes using OrthoFinder (version 2.5.4) (Emms & Kelly, 2019). For each species, the primary protein sequence of the transcript was used as input to Orthofinder. In the resulting orthofinder outputs, the script "find_markers.orthofinder.py" was used to parse the resulting phylogenies and return back putative orthologs (https://github.com/Jome0169/Mendieta.C4_manuscript). For all C_4 genes analyzed, each orthogroup was additionally annotated by hand in order to ensure accurate assignment of nearest orthologs phylogenetically.

Annotation of Cell Types:

Cell types were annotated by calculating gene chromatin accessibility for marker genes in each genome on a per cell basis. These values were then visualized on the UMAP embedding, and clusters with numerous marker genes associated with the same cell-type were used as evidence. Additionally, for each louvain cluster, enrichment of marker genes was calculated by comparing the cluster average as compared to a random shuffle of random cells. The top five most enriched markers were used in tandem with the UMAPs to ascertain cell-type identity. We also tested the statistical significance of the marker gene using Presto, a modified Wilcoxon rank-sum test in order to identify the most unique marker gene in each cluster (Korsunsky, Nathan, et al., 2019). Additionally, for specific clusters showing mixed signals from marker genes, sub-clustering was done by isolating the cluster in question, and then re-clustering these cells on a new UMAP manifold. The same steps were done to visualize marker genes, as well as test this enrichment, and statistical significance. Finally, to bolster our set of marker genes across species, we used our most confident cell-type annotation in *Z. mays* to *de novo* discover marker genes. To do so, we utilized our gene-body-accessability metrics for each annotated cell-type, and ran DESeq2 (version 1.42.0) in a replicate aware fashion using all other cells as a null (Love et al., 2014). Only statistically significant markers were kept which had a fold change greater than 1.5, and a log fold standard error of less than .6. OrthoFinder was used as mentioned above to find orthologs. To ensure that we were comparing similar cell-types, we also took an orthogonal approach where we compared the gene accessibility of the top 2000 most variable orthologs between our species. A linear model was used for each species comparison where the mean gene accessibility was taken into consideration, and the species was one-hot-encoded. Variation was calculated as the average variation between both datasets. The resulting residuals were used to generate the cell-type correlations.

Peak Identification:

To identify peaks, cells of the same annotation type were pseudo bulked in a replicate aware fashion. Within each replicate MACS2 (version 2.2.9.1) was run with parameters "--nomodel --keep-dup auto --extsize 150 --shift -75 --qvalue .05" and variable genome size flag '-g' (Y. Zhang et al., 2008). Summits for each peak identified in each replicate were extended by 250 bp in either direction. Only peaks which overlapped between replicates were used. To merge peaks from various cell types and select peak boundaries, the p-value associated with each peak in each cell type was compared by calculating the chromatin accessibility score for each peak per million, with those peaks with the highest accessibility score being selected as the representative peak. This method of identifying the most representative peaks across cell-types was inspired by previous single cell ATAC-seq papers (Cusanovich et al., 2018; Domcke et al., 2020; K. Zhang et al., 2021). Additionally, bigwigs were generated for each cell type by normalizing each dataset to the number of reads/per million scaling factor. Implementation of this algorithm is found in the script call_scACRs.py for ease of use and replication in other experiments.

Identifying Cell-type-specific ACRs:

To identify cell-type specific ACRs, a modified bootstrapping method was used which drew inspiration from the modified entropy metrics found in (K. Zhang et al., 2021). On a per ACR basis, Tn5 integrations per cell-type were summed and counts per million (CPM) normalized. These values were then converted to a probability by using the following equation (below, equation 1) where *pi* is the CPM value for the focal cell-type and *qi* is the total sum of all CPMs. From this probability statement, a modified shannon entropy metric was calculated, followed by a metric of specificity *Qpt*. For robust cell-type-specific ACR identification, the annotated cell-type was bootstrapped 5000 times, taking a sample of 250 cells from the cell population in question, and calculating both entropy and specificity scores. This was done to attempt to get a robust signal of specificity, which takes into consideration the variation in cell quality present in each

cell-type annotation. To generate the null distribution of specificity scores, individual cell annotations were scrambled to generate an equal number of null cell-type classifications. For each null value, the entropy and specificity score were calculated. Finally to calculate a p-value, a non-parametric approach was used to identify how many of the real bootstraps fell outside of the null distribution using a one tailed test. ACRs which had a p-value of <0.001 were considered to be significant. ACRs were finally classified by the number of cell types they were specific to. ACRs specific to greater than three were classified as broadly accessible, less than or equal to three as cell type restricted, and a single cell-type as cell-type specific.

1) $pi = \frac{qi}{\Sigma(qi)}$
2) $Hp = -\Sigma ptlog_2(pt)$
3) $Qpt = Hp - log_2(pt)$

Identifying Conserved ACRs Across Species

Since a majority of the C4 genes identified were not in synteny with one another, we took a gene family based approach to identify conserved and non-conserved ACRs associated with our C4 genes. In short, all ACRs within two gene models of a C4 gene are utilized for comparison. Sequences from the ACR were isolated using "bedtools getfasta" (version 2.31) (Quinlan & Hall, 2010). Then in a pairwise fashion each species had their ACRs from one C4 gene family compared to the corresponding genomic loci of the same gene family in a different species. Comparisons were made using Blastn (version 2.2.29) with the following parameters " -task blastn-short -evalue 1e-3 -max_target_seqs 4-word_size 7 -gapopen 5 -gapextend 2 -penalty -1 -reward 1 -dust no -outfmt 6" (Camacho et al., 2009). The output blast files were further filtered requiring sequence alignment to be greater than 20 nts, and have an evalue of .001. This analysis and the detailed commands ran can be found in the following snakemake file titled

"ID_syntenic_orthologous.ACRs.snake", and found in the snakemake directory in the associated github.

Identifying Cell-type-specific Motifs:

De-novo cell-type-specific motifs were identified by using XSTREME (version 5.5.3) of the MEME suite package (Bailey et al., 2015; Grant & Bailey, 2021). In brief the sequences underlying the cell-type-specific ACRs were isolated, and equally matched null set of broadly-accessible ACRs were used the comparison for genomic enrichment. These null ACRs were matched in terms of GC content, and were only allowed to be 5% different from the cell-type-specific set in question and generated using the script "gen_null_fa.py". Upon generation, motifs were analyzed using the universalmotifs package in R (version 3.18) (*Universalmotif*, n.d.). Motifs were first compared using HELL distance, and motifs which had a low correlation were discarded. In order to generate representative motifs, highly correlated motifs were merged using the function "merge_motifs" in found in the universalmotifs package. To identify the location of motifs, the R package motifmatchR were used, with a significant value cut off of .0005 (*Motifmatchr*, n.d.).

Data availability:

sciATAC-seq data for *Z. mays, S. bicolor, U. fusca,* and *P. miliceum* is found in NCBI under the following bioproject PRJNA1063172. Leaf data for *O.sativa* can be found under the following SRR bioproject PRJNA100757. All scripts used for processing and analyzing data in this manuscript can be found at the following github repository: https://github.com/Jome0169/Mendieta.C4_manuscript

Chapter 4

Evolution of plant cell-type-specific *cis*-regulatory elements[1]

[1]Mendieta, John. Submitted to Nature, 1/22/2024

Abstract:

Cis-regulatory elements (CREs) are critical in regulating gene expression, and yet our understanding of CRE evolution remains a challenge. Here, we constructed a comprehensive single-cell atlas of chromatin accessibility in *Oryza sativa*, integrating data from 104,029 nuclei representing 128 discrete cell states across nine distinct organs. We used comparative genomics to compare cell-type resolved chromatin accessibility between *O. sativa* and 57,552 nuclei from four additional grass species (*Zea mays, Sorghum bicolor, Panicum miliaceum*, and *Urochloa fusca*). Accessible chromatin regions (ACRs) had different levels of conservation depending on the degree of cell-type specificity. We found a complex relationship between ACRs with conserved noncoding sequences, cell-type specificity, conservation, and tissue-specific switching. Additionally, we found that epidermal ACRs were less conserved compared to other cell types, potentially indicating that more rapid regulatory evolution has occurred in the L1 epidermal layer of these species. Finally, we identified and characterized a conserved subset of ACRs that overlapped the repressive histone modification H3K27me3, implicating them as potentially critical silencer CREs maintained by evolution. Collectively, this comparative genomics approach highlights the dynamics of cell-type-specific CRE evolution in plants.

Main

Cis-regulatory elements (CREs) function as pivotal hubs, facilitating the binding of transcription factors (TFs) and recruitment of chromatin-modifying enzymes, thereby fine-tuning gene expression in a spatiotemporal-specific manner (Preissl et al., 2023). CREs play important roles in developmental and environmental processes, and their functional divergence frequently drives evolutionary change (Marand et al., 2023; Wittkopp & Kalay, 2012). Prior studies highlighted the dynamic nature of CREs throughout evolution and their involvement in regulating gene expression via distinct chromatin pathways (Kajala et al., 2020; Lu et al., 2019; Maher et al., 2018; Oka et al., 2017; Reynoso et al., 2019; Ricci et al., 2019). Across diverse cell types, gene expression is intricately regulated by multiple distinct CREs, each exerting control within specific cell or tissue type or particular developmental stage and environmental cue (Cusanovich et al., 2018; Domcke et al., 2020; Lu et al., 2023). Environmental sensing and adaptation, in plants, relies heavily upon epidermal cells (Javelle et al., 2011). For example, epidermal bulliform cells in grasses change their turgor pressure to roll the leaf to slow water loss under stressful conditions, with the TF, ZINC FINGER HOMEODOMAIN 1 (ZHD1), modulating leaf rolling by influencing rice (*Oryza sativa*) bulliform cell development (Kadioglu et al., 2012; Xu et al., 2014). Several studies have identified CREs functioning in a cell-type-specific manner within various plant species (Dorrity et al., 2021; Farmer et al., 2021; Feng et al., 2022; Marand et al., 2021; Nobori et al., 2023; Swift et al., 2023; Tu et al., 2022; L. Zhang et al., 2023). Despite these findings, our understanding of CREs exhibiting evolutionarily conserved or divergent cell-type-specific activities remains limited.

Through single-cell assay for transposase accessible chromatin sequencing (scATAC-seq), we constructed an expansive single-cell reference atlas of accessible chromatin regions (ACRs) within rice. We then leveraged these data in tandem with four additional scATAC-seq leaf datasets from diverse grasses (*Zea mays*, *Sorghum bicolor*, *Panicum miliaceum*, and *Urochloa fusca*) allowing us to compare ACRs across species and cell types (Mendieta et al., 2024). We quantified the proportion of ACRs that were conserved in these monocots and found high rates of cell-type-specific ACR turnover, particularly in epidermal cells. This indicates that the CREs associated with specific cell types are rapidly changed by evolution. Finally, we used both conserved non-coding sequences (CNS) and H3K27me3 to find a series of conserved ACRs and the candidate CREs within them that are potentially important for recruitment of Polycomb-mediated gene silencing.

Construction of an ACR atlas at Single-cell Resolution in Rice

To create a cell-type-resolved ACR rice atlas, we conducted scATAC-seq across a spectrum of nine organs in replicate (Figure 4.4.1a and b). Data quality metrics, such as correlation between biological replicates, transcription start site enrichment, fraction of reads in peaks, fragment size distribution, and organelle content, revealed excellent data quality (Supplementary Figure 4.4.1 and 2). Following strict quality control filtering, we identified 104,029 high-quality nuclei, with an average of 41,701 unique Tn5 integrations per nucleus. Based on a nine-step annotation strategy, which included RNA *in situ* and spatial-omic (slide-seq) validation of cell-type specificity, we identified a total of 128 cell states, encompassing 60 main cell types across various developmental stages from all the

organs sampled (Figure 4.4.1b; Extended Data Figure 4.1a; Supplementary Note 1; Supplementary Figure 4.3-12; Supplementary Tables 1-5).

By analyzing cell-type-aggregated chromatin accessibility profiles, we identified a total of 120,048 ACRs (Extended Data Figure 4.1b and c). Among these ACRs, 42,649 were categorized as 'cell-type-specific ACRs', exhibiting accessible chromatin in less than 17% (10/60) of the main cell types, whereas approximately 77,399 were classified as 'broad ACRs' with chromatin accessibility in more than 17% of the cell types (Figure 4.1c and d; Extended Data Figure 4.1d and e). When analyzing their proximity to genomic features in the rice genome, about half of the ACRs were gene proximal (52%; located within 2 kb of genes; Figure 4.1e). These proximal ACRs had higher but less variable chromatin accessibility than genic and distal ACRs (Extended Data Figure 4.1f). In contrast, about 19% of the ACRs overlapped genes, mostly in introns, and the remaining 29% were categorized as distal (Figure 4.1e; situated more than 2 kb away from genes). The greater chromatin accessibility variance in non-proximal ACRs, suggests these regions may be important in different cellular contexts or tissues.

We were interested in leveraging the *O. sativa* atlas to understand how ACRs change during grass evolution. The *O. sativa* ACRs were overlapped with syntenic regions defined by their relationship to four different grass species *Z. mays*, *S. bicolor*, *P. miliaceum*, and *U. fusca* that have single-cell ATAC sequencing using combinatorial indexing (sciATAC-seq) data from leaves (Mendieta et al., 2024). The analysis revealed that 34% (40,477) of the *O. sativa* ACRs were within 8,199 syntenic regions (~86 Mb in *O. sativa*) between *O. sativa* and at least one of four examined grass species, whereas the majority of ACRs (66%; 79,571) were located in non-syntenic regions. Notably, these

non-syntenic ACRs were enriched (p = 6e-248 to 0.0031; Fisher's exact test) for cell-type specificity, particularly for proximal and distal ACRs (Figure 4.1f; Extended Data Figure 4.1g). This reveals that most of the ACRs in the grass species examined occurred in non-syntenic regions.

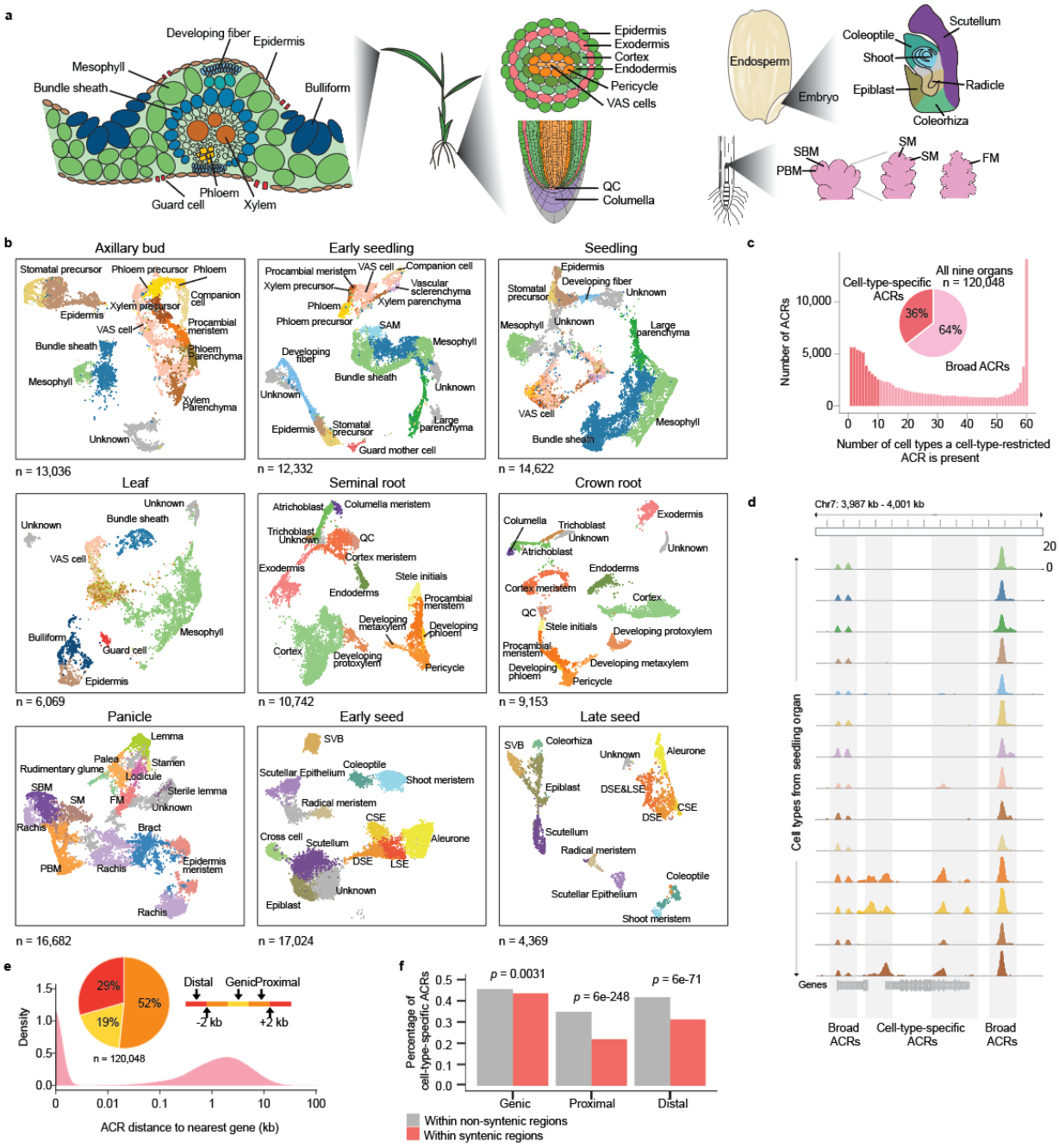

Figure 4.1. Identifying Cell types and characterizing ACRs in rice using scATAC-seq data. a, Overview of cell types in leaf, root, seed, and panicle organs. VAS cells: vasculature-related cells. QC: Quiescent Center. SBM: Secondary Branch Meristem. PBM: Primary Branch Meristem. SM: Spikelet Meristem. FM: Floral Meristem. **b,** UMAP projection of nuclei, distinguished by assigned cell-type labels in axillary bud, early seedling (7 days after sowing), seedling (14 days after sowing), leaf (V4 stage; four leaves with visible leaf collars), seminal root, crown root, panicle, early seed development (6 DAP; days after pollination), and late seed development (10 DAP). SAM: Shoot Apical Meristem. CSE: Central Starchy Endosperm. DSE: Dorsal Starchy

Endosperm. LSE: Lateral Starchy Endosperm. SVB: Scutellar Vascular Bundle. **c,** Evaluation of proportions of ACRs that are cell-type specific versus broad. **d,** A screenshot illustrates the examples of cell-type-specific and broad ACRs. **e,** Accessible chromatin regions (ACRs) show a bimodal distribution of distance to the nearest gene. The ACRs were categorized into three major groups based on their locations to the nearest gene: genic ACRs (overlapping a gene), proximal ACRs (located within 2 kb of genes), and distal ACRs (situated more than 2 kb away from genes). **f,** The enrichment of cell-type-specific ACRs in non-syntenic regions as opposed to syntenic regions. Significance testing was performed using a two-sided Fisher's exact test.

The Landscape of Cell-type-specific ACRs Across Grass Species

To determine to what degree ACR number, genomic position and cell-type specificity differs amongst grasses, we compared the composition and distribution of leaf ACRs across the five species (Figure 4.2a). Using previous cell-type annotations, we calculated the proportion of both broad and cell-type-specific ACRs across all species (Mendieta et al., 2024). An average of ~53,000 ACRs were identified across the five species, with 15-35% of the ACRs classified as cell-type specific (Figure 4.2b). Broad and cell-type-specific ACRs were equivalent in their distributions around promoters, distal, and genic regions (Figure 4.2c; Extended Data Figure 4.2a-c).

Prior hypotheses have suggested that large scale regulatory rewiring could play a key role in cell-type adaptation to environmental conditions (Kajala et al., 2020; Lv et al., 2023). To explore instances where divergent TF activity in orthologous cell types occurred, we associated gene body chromatin accessibility of TFs with their cognate TF motifs across different species and cell types. Approximately 71% of the TFs examined exhibited a positive correlation between the local chromatin accessibility of their gene body and global enrichment of their cognate binding motifs within ACRs across all cell types (Extended Data Figure 4.3a). This finding suggests a relationship between the local

TF gene chromatin accessibility and TF protein activity in the same cell. Moreover, the genomic sequences from all ACRs discovered in all species and cell types exhibited enrichment of TF motifs compared to the control set of sequences (Extended Data Figure 4.3b). Furthermore, TF motif enrichment analysis revealed known TF-cell-type specificities (Figure 4.2d). For example, the HOMEODOMAIN GLABROUS 1 (HDG1) TF is critical for epidermis and cuticle development and its motif was enriched in epidermis cells in all five species (Extended Data Figure 4.3c) (R. Wu et al., 2011). We also observed motif enrichments of WRKY, HOMEODOMAIN-LEUCINE ZIPPER (HD-ZIP), and PLANT ZINC FINGER (PLINC) in epidermal cells across all five species examined (Figure 4.2e; Extended Data Figure 4.3d; Supplementary Table 6). This result indicates that these TFs play a conserved role in the development of the grass epidermis. Phloem companion, sieve element cell, and bundle sheath cell TFs exhibited similar enrichments across the species (Figure 4.2e). However, species-specific motif patterns were also observed, with *O. sativa* being the most different. For example, the DNA-BINDING ONE ZINC FINGER (DOF) TF family exhibited higher enrichment scores in four species (*Z. mays*, *S. bicolor*, *P. miliaceum*, and *U. fusca*) that are C_4 photosynthesizing species, as opposed to *O. sativa*, which uses C_3 photosynthesis (Figure 4.2e; Extended Data Figure 4.3e and f). The DOF TF family regulates the formation and function of vascular tissues in *Arabidopsis thaliana*, and has been implicated in the switch from C_3 to C_4 (Figure 4.2e; Extended Data Figure 4.3e and f) (Dai et al., 2022; Le Hir & Bellini, 2013; Swift et al., 2023; Yanagisawa, 2000). These results suggest that distinct DOF TF motif enrichment in phloem and bundle sheath cells might fulfill different vasculature roles in C_3 and C_4 grass leaves (Gao et al., 2018; UENO et al.,

2006). Taken together, our findings demonstrate the power of scATAC-seq data in a comparative framework to explore regulatory evolution, both based on the relationship of ACRs to TF motifs, as well as the relationship between TFs and their corresponding motifs.

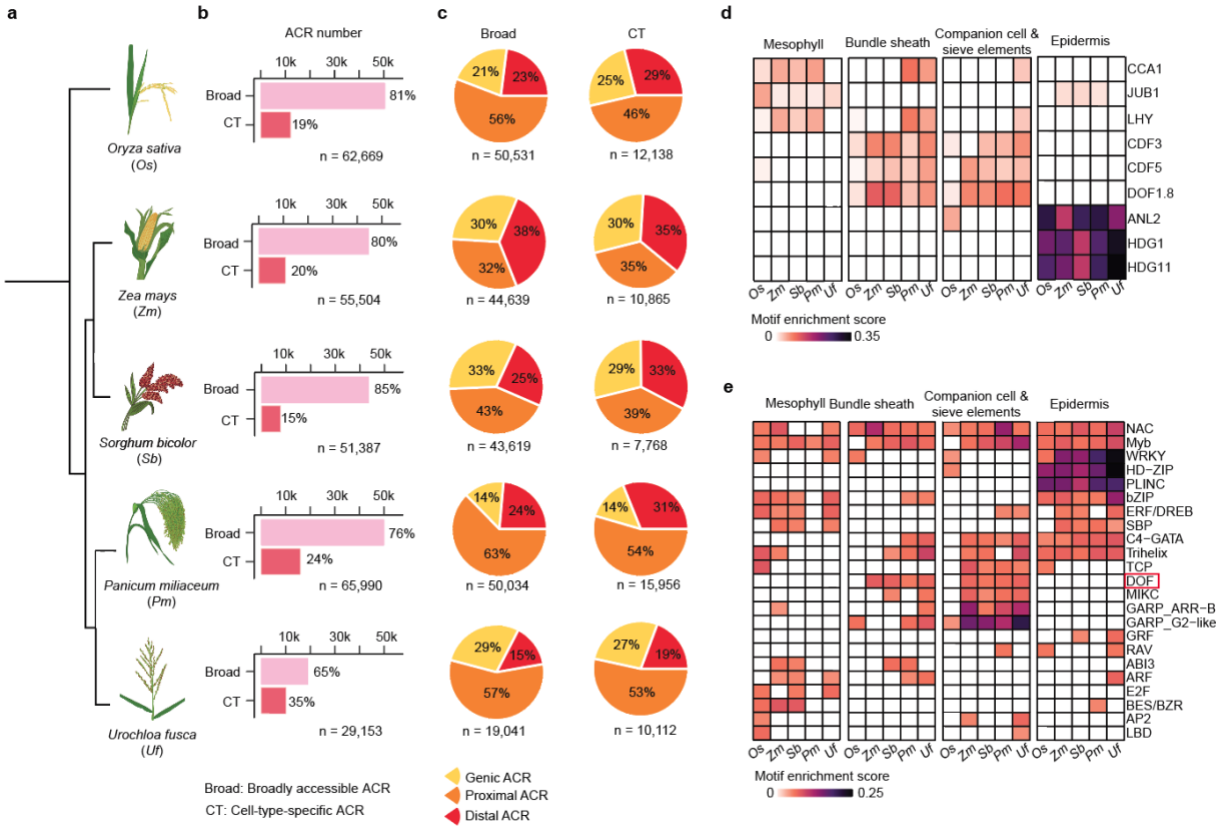

Figure 4.2. Position and motif enrichment of cell-type-specific ACRs across species. **a,** A phylogenetic tree illustrates five species under examination. **b,** The count of broad and cell-type-specific ACRs. **c,** Broad and cell-type-specific ACRs were classified into three main groups based on their proximity to the nearest gene: genic ACRs (overlapping a gene), proximal ACRs (located within 2 kb of genes), and distal ACRs (situated more than 2 kb away from genes). *O. sativa*, *P. miliaceum*, and *U. fusca* showed a higher percentage of proximal ACRs, but a lower percentage of distal ACRs compared to *Z. mays* and *S. bicolor*, likely reflecting differences in intergenic space and overall genome sizes. **d**, A heatmap illustrates nine TF motif enrichments, consistent with the known TF dynamics among cell types [CIRCADIAN CLOCK ASSOCIATED 1 (CCA1), LATE ELONGATED HYPOCOTYL (LHY) and JUNGBRUNNEN1 (JUB1): mesophyll; CYCLING DOF FACTOR 3 (CDF3) and CDF5: companion cells; DOF1.8: vascular-related cells; ANTHOCYANINLESS2 (ANL2), HOMEODOMAIN GLABROUS 1 (HDG1), and HDG11: epidermis (Amanda et al., 2016; Kubo et al., 2008; Otero &

Helariutta, 2017; Ramachandran et al., 2020; Shimadzu et al., 2023; R. Wu et al., 2011). **e,** A heatmap illustrates collapsed TF motif enrichment patterns into super motif families across various species for each cell type. The motif enrichment score cutoff was set to 0.05. The DOF TF motif family was highlighted by a red frame. To mitigate the impact of substantial variations in cell numbers across species or cell types, we standardized (down-sampled) the cell counts by randomly selecting 412 cells per cell type per species. This count represents the lowest observed cell count for a given cell type across all species (See Methods: Linear-model based motif enrichment analysis).

Species-specific Evolution of Cell-type-specific ACRs

To understand how cell-type-specific and broad ACRs changed over evolution, we examined ACRs within syntenic regions among the studied species. To compare ACRs, we devised a synteny-based BLASTN pipeline that allowed us to compare sequences directly (See Methods: Identification of syntenic regions; Figure 4.3a; Extended Data Figure 4.4a; Supplementary Table 7). Using *O. sativa* ACRs as a reference, we identified three classes of cross-species ACR conservation: 1) ACRs with matching sequences that are accessible in both species (shared ACRs), 2) ACRs with matching sequences, but are only accessible in one species (variable ACRs), and 3) ACRs where the sequence is exclusive to a single species (species-specific ACRs; Figure 4.3b). The shared ACR BLASTN hits were often small syntenic sequences, highlighting the large divergence of grass ACRs sequences. However, the majority (92-94%) of these shared BLASTN sequences encoded known TF motifs (Supplementary Figure 4.13), indicating that shared ACRs are conserved regulatory regions. In contrast, variable ACRs represent a blend of conserved and divergent regulatory elements, and species-specific ACRs likely indicate novel regulatory loci. We found that, on average, shared ACRs were enriched (p = 9e-55 to 4e-17; Fisher's exact test) for broad ACRs, whereas the variable (p = 2e-16 to 2e-04; Fisher's exact test) and species-specific (p = 2e-06 to 2e-03; Fisher's exact test) classes

were enriched for cell-type specificity (Figure 4.3c; Extended Data Figure 4.4b and c). Moreover, we observed that the genomic distribution of shared ACRs were biased towards proximal ACRs (Extended Data Figure 4.4d). This contrasts with cell-type-specific ACRs, which are overrepresented in the species-specific class, and tend to reside in distal genomic regions (Extended Data Figure 4.4e).

We further investigated whether the cell-type-specific ACRs were conserved in their cell-type specificity by evolution. Pairwise comparison between *O. sativa* and the other grasses revealed that a low number (128 to 299) of shared ACRs retained cell-type specificity in both species (Extended Data Figure 4.4f). Of these few shared cell-type specific ACRs, the majority (62%-68%), were accessible in the identical cell-type in both *O. sativa* and the corresponding species (Figure 4.3d; Extended Data Figure 4.4f). For example, the promoter ACR associated with *LATERAL ROOT DEVELOPMENT 3* (*LRD3*), a gene critical in companion cell and sieve element development , showed sequence conservation between *O. sativa* and *S. bicolor* (Figure 4.3d) (Ingram et al., 2011, p. 3). Interestingly, ACRs which were mesophyll specific in *O. sativa* changed their cell-type specificity to bundle sheath 17%-41% of the time, while bundle sheath ACRs changed to mesophyll 9%-25% of the time (Figure 4.3d; Extended Data Figure 4.4g). This result is likely due to the functional divergence associated with the shift from C_3 (*O. sativa*) to C_4 (all other species sampled) photosynthesis (Mendieta et al., 2024; Swift et al., 2023). Of all the classes of cross-species ACR conservation, species-specific ACRs were the most predominant in every cell type (Figure 4.3e). These findings suggest a dynamic and rapid evolution of cell-type-specific ACRs within the examined species. Notably, ACRs in L1 layer (epidermis and protoderm) cells exhibited the highest

proportion of species-specific ACRs (Figure 4.3e; Extended Data Figure 4.5a-c), suggesting that L1 ACRs are more evolutionarily divergent compared to the other cell types.

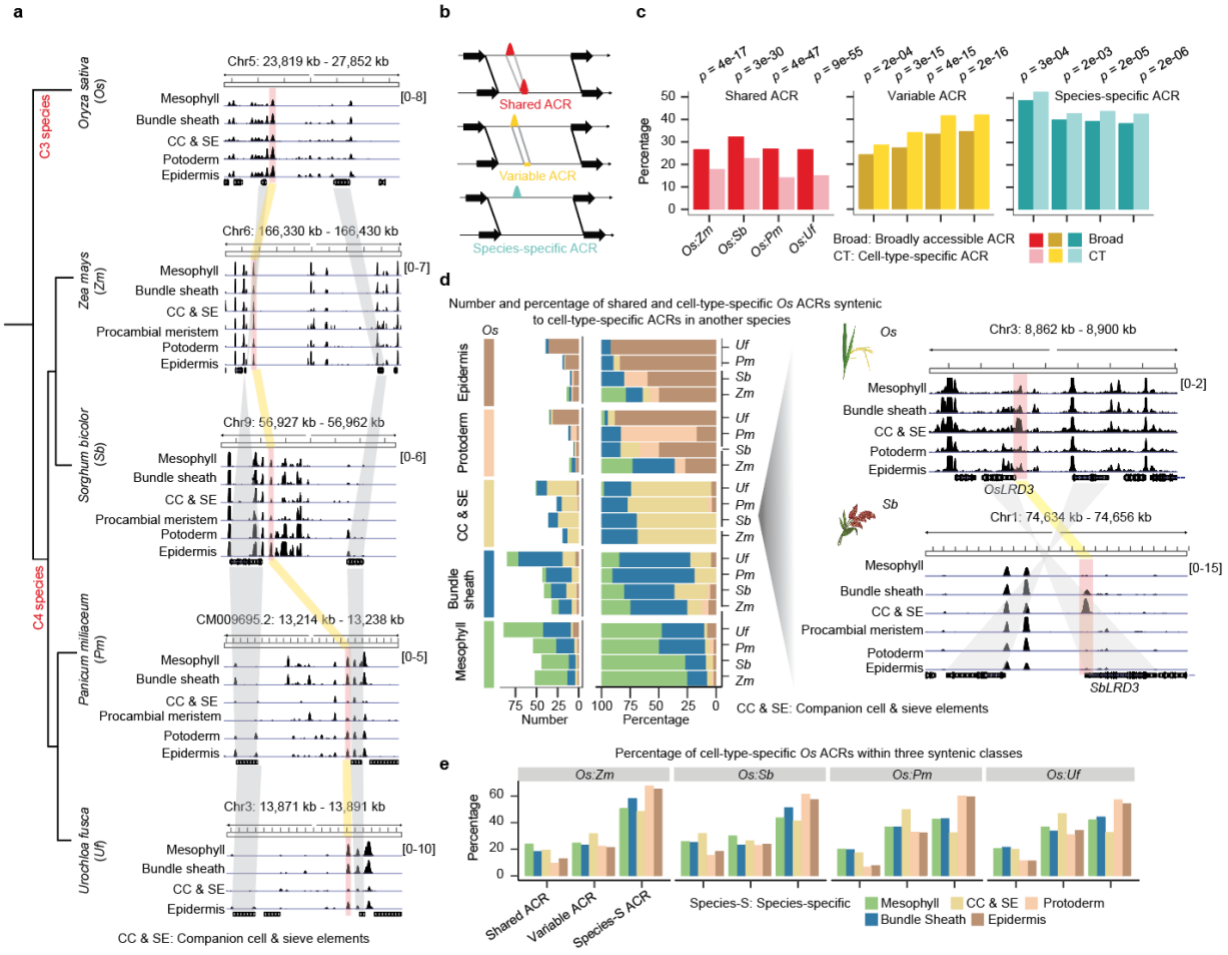

Figure 4.3. Cell-type-specific ACRs are frequently species-specific. a, A screenshot illustrating syntenic regions capturing shared ACRs across five species. The red bars denote syntenic ACRs within regions flanked by corresponding syntenic gene pairs, while the gray color highlights these syntenic gene pairs. **b,** Three classes depicting variations in ACR conservation between two species. 'Shared ACRs': ACRs with matching sequences that are accessible in both species; 'Variable ACRs': ACRs with matching sequences but are only accessible in one species; 'Species-specific ACRs': ACRs where the sequence is exclusive to a single species. **c,** The percentage of broad and cell-type-specific ACRs underlying three classes shown in panel **b**. The significance test was done by using the Fisher's exact test (alternative = 'two.sided'). **d, Left**, the number and percentage of *O. sativa* shared ACRs that retain or change cell-type specificity amongst the other four species. **Right**, a screenshot of a *O. sativa* phloem specific ACR that retains phloem specificity in *S. bicolor*. This ACR is situated at the promoter region

of *LRD3* which is specifically expressed in companion cell and phloem sieve elements (Supplementary Table 2). The gray shaded region highlights the syntenic gene pair. **e,** The percentage of cell-type-specific ACRs identified across all cell types within each species pair split into three classes shown in panel **b**. The percentage for each cell type within the three classes collectively sum to 100%.

Delving deeper into the L1 TF families, we focused on species-specific L1 ACRs. Within the species-specific syntenic ACRs, we observed a predominance of TF motifs for the HD-ZIP, SQUAMOSA PROMOTER BINDING PROTEIN (SBP), PLINC families (Extended Data Figure 4.6a). Many of these, such as HDG1, ZHD1, ATHB-20, SPL3, SPL4, and SPL5, function in epidermal cell development (Chen et al., 2010; Denyer et al., 2019; Fang et al., 2021; Horstman et al., 2015; Xu et al., 2014). The predominance of these motifs in the species-specific ACR class suggests that although L1 TF motif sequences are well conserved (Figure 4.2d and e), yet the positions of these motifs are not conserved in grass genomes. Upon comparing TF-motif enrichment in syntenic and non-syntenic ACRs, we observed the presence of these epidermal motif families in both groups (Extended Data Figure 4.6b; Supplementary Table 8), indicating their essential roles in both plant epidermal cell development and rapid gene-regulatory co-option in species-specific sequences. Notably, some TF-motif families such as WRKY in epidermal cells, were more enriched in non-syntenic ACRs (Extended Data Figure 4.6b), further supporting that this family may drive phenotypic innovation of the epidermal layer.

To look for derived species-specific ACRs associated with the altered expression of surrounding gene orthologs in epidermal cells, we performed a single-nucleus RNA sequencing (snRNA-seq) analysis in *O. sativa* (Supplementary Figure 4.7;

Supplementary Table 9). By integrating with a snRNA-seq data from *Z. mays* , we identified 87 orthologous genes, irrespective of syntenic regions, which exhibited elevated L1 *O. sativa* expression compared to *Z. mays* (Extended Data Figure 4.6c; Supplementary Table 10) (Marand et al., 2021). A gene ontology enrichment test for these 87 genes revealed eight genes that were significantly enriched in the lipid metabolic process (Extended Data Figure 4.6d), likely related to epidermal cuticle metabolism. Among the eight genes, one orthologous gene to *A. thaliana GDSL LIPASE GENE* (*LIP1*) has been shown to be specifically expressed in the epidermis (Rombolá-Caldentey et al., 2014). We further identified 102 L1 cell-type-specific ACRs from *O. sativa* that were the closest to the 87 orthologous genes and observed 11 TF motifs were enriched (q = 3e-10 to 5e-04; Binomial test) in these ACRs (Extended Data Figure 4.6c). These included TF family motifs known for their roles in epidermal cell development such as ZHD1, HDG11, ZHD5, HDG1, and WRKY25 (Hong et al., 2011; Rosado et al., 2022; R. Wu et al., 2011; L.-H. Yu et al., 2016). For example, within the *OsLIP1* intron, we identified two ZHD1 motifs within a species-specific ACR that was specifically accessible in L1 cells (Extended Data Figure 4.6e). We also flipped this comparison by identifying 166 orthologs with elevated maize epidermal expression compared to rice (Supplementary Table 10). This revealed 196 L1 cell-type-specific ACRs in *Z. mays*. Within these ACRs, the most enriched (q = 0.0129 to 0.0392; Binomial test) TF motif was MYELOBLASTOSIS 17 (MYB17; Extended Data Figure 4.6f and g). This R2R3 MYB family TF is associated with epidermal cell development, specifically for its involvement in the regulation of epidermal projections (Brockington et al., 2013). Furthermore, we hypothesized that some of these novel motifs could be related to *O.*

sativa transposable element (TE) expansion. Notably, we found long terminal repeat retrotransposon (LTR)-associated ACRs from the *Gypsy* family were enriched (p = 0.0012 to 0.0415; Fisher's exact test) in *O. sativa* epidermal cell-type-specific ACRs (Extended Data Figure 4.6h). The ZHD1 motif was enriched within these *Gypsy*-associated ACRs (p = 0.0011 to 0.0052; Binomial test) (Extended Data Figure 4.6i). By linking snRNA-seq to scATAC-seq data, we tied gene-proximal CRE changes to elevated epidermal expression of conserved orthologs 50 million years derived (Kh et al., 1989). These CRE changes drive variance in species-specific L1 layer development, potentially contributing to species differences in environmental adaptation.

CNS are Enriched in Cell-type-specific ACRs

To augment our syntenic ACR BLASTN approach, we sought to identify CNS through comparative genomics (Babarinde & Saitou, 2016; Lu et al., 2019; Song et al., 2021; Woolfe et al., 2004). Outside of UTRs, CNS are thought to encompass CREs that drive processes too critical to be lost during evolution (Nelson & Wardle, 2013). To our knowledge, no plant study has overlapped CNS with cell-type-specific chromatin accessibility measurements, leaving it unknown whether CNS retain the same cell-type accessibility in different species. 53,253 *O. sativa* and 284,916 *Z. mays* published CNS overlapped with the leaf ACRs identified in this study (Hendelman et al., 2021). Excluding CNS overlapping with untranslated regions, 30.8% and 21.3% of CNS overlapped with the leaf-derived ACRs in *O. sativa* and *Z. mays*, respectively (Figure 4.4a). Using all ACRs in the *O. sativa* atlas, this ratio increased to 65.0% (Figure 4.4a), indicating that a significant portion of these CNS likely function in specific cell types and

tissues. One common assumption is that CNS that overlap with ACRs (CNS ACRs) retain the exact same function between species. We observed 39% to 51% of total CNS ACRs within the 'shared CNS ACR' class (Extended Data Figure 4.7a and b), suggesting these ACRs have conserved functions and conserved cellular contexts between *O. sativa* and *Z. mays.* Within syntenic regions, ACRs with CNS are more cell-type specific than those without (Figure 4.4b). This enrichment was consistent for all classes of ACRs we identified, except for species-specific ACRs which were equivalently cell-type specific and broad (Figure 4.4b). The enrichment of CNS in cell-type-specific ACRs stress the importance of rare cell-type function, as the cell-type-specific CREs are preferentially retained during evolution.

However, the majority (49-61%) of all CNS ACRs differed in cell-type specificity between *O. sativa* and *Z. mays* (Extended Data Figure 4.7a and b). Specifically, we examined the CNS found in *Z. mays* ACRs that did not have a corresponding leaf ACR in *O. sativa*. Leveraging the *O. sativa* atlas, we identified these sequences had divergent cellular or tissue chromatin accessibilities. 249 (75%) of the *Z. mays* leaf variable CNS ACRs were accessible in non-leaf cell states (Figure 4.4c-e; Extended Data Figure 4.7c), highlighting the instability of the cellular context in which CNS acts. This suggests frequent co-option of CNS CREs into different tissues or cell types. Investigating the CNS ACRs that lost leaf cell-type specificity, we observed that these ACRs were accessible in many non-leaf cell types, uniformly distributed amongst the atlas cell annotations (Extended Data Figure 4.7c-g). Consistent with our findings that epidermal-specific ACRs tend to have the most species-specific ACRs in syntenic regions (Figure 4.3e; Extended Data Figure 4.5a-c), L1 cells showed a significantly lower ratio of non-

syntenic CNS ACRs compared to other cell types. This lower ratio demonstrates the frequent loss of L1 CNS, further supporting the rapid evolution of epidermal transcriptional regulation.

Interestingly, we noticed a pattern where some CNS within ACRs also overlapped domains of H3K27me3 (Figure 4.4f). H3K27me3 is a histone modification associated with facultative heterochromatin established by the POLYCOMB REPRESSIVE COMPLEX 2 (PRC2) (Cao et al., 2002; Lu et al., 2019; Schmitz et al., 2022; Xiao et al., 2017). Genes silenced by PRC2 and H3K27me3 are important regulators that are only expressed in narrow developmental stages or under specific environmental stimuli, where they often initiate important transcriptional changes (W. Ouyang et al., 2023; Xiao et al., 2017). This importance makes the identification of key CREs controlling H3K27me3 silencing especially interesting. Upon closer examination of the ACRs overlapping domains of H3K27me3 to ACRs away from H3K27me3, we observed that H3K27me3 ACRs were significantly enriched for CNS (Figure 4.4g). This enrichment supports that some of these CNS underpin conserved, and critical components, of H3K27me3 silencing.

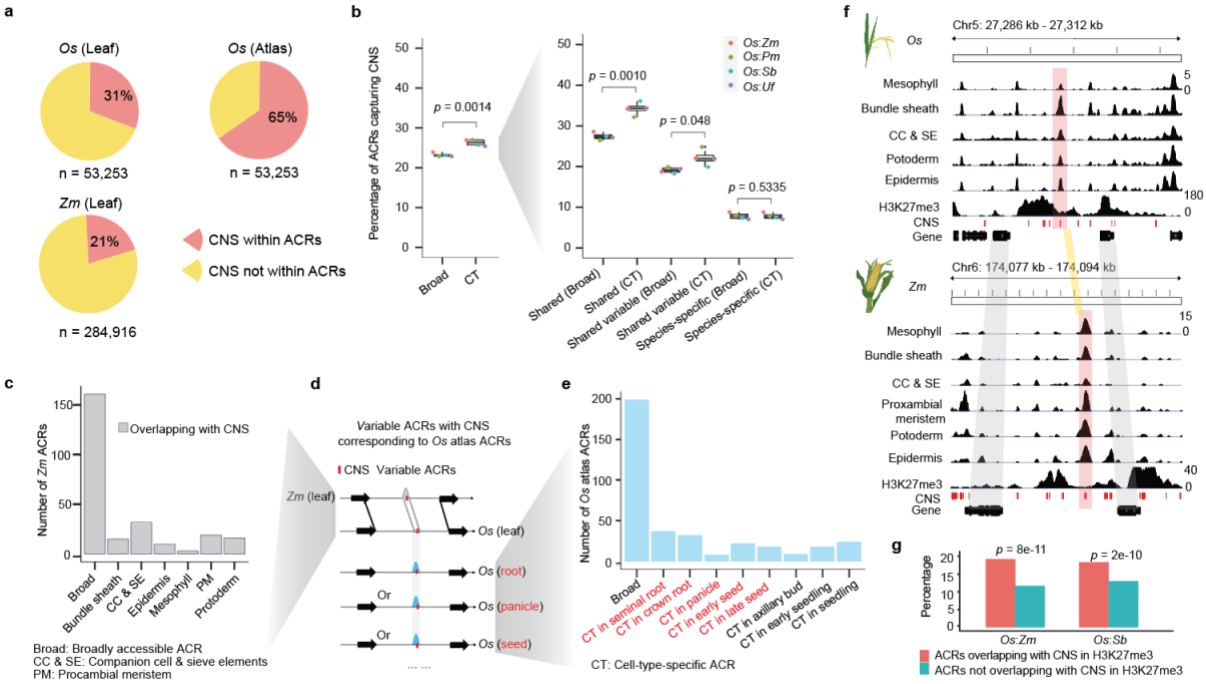

Figure 4.4. Cell-type-specific ACRs exhibit an enrichment of CNS. a, The percentage of CNS overlapping with ACRs. 'n' indicates the number of CNS. 'Atlas' means the ACRs were from the *O. sativa* atlas in Figure 4.1b. **b, Left,** the percentage of broad and cell-type-specific ACRs within syntenic and non-syntenic regions overlapping with the CNS. **Right,** this panel presented similar meaning as the left panel but focusing on three classes within syntenic regions shown in Figure 4.3b. Broad: Broadly accessible ACR; CT: Cell-type-specific ACR. Significance testing was performed using the t-test (alternative = 'two.sided'). **c,** The bar plot showcases the count of *Z. mays* variable ACRs accessible in leaf cell types. **d,** A sketch illustrating whether variable ACRs containing CNS in *Z. mays* capture ACRs derived from the *O. sativa* atlas. **e,** The bar plot represents the count of *O. sativa* atlas ACRs accessible in non-leaf cell types. **f,** An example of a syntenic block containing *O. sativa*-to-*Z. mays* conserved ACRs within a H3K27me3 region. CNS were highlighted using red color. **g,** The percentage of ACRs capturing CNS in and outside of H3K27me3 regions. The percentage for each group within H3K27me3 and not within H3K27me3 regions collectively sum to 100%. Significance testing was performed using Fisher's exact test (alternative = 'two.sided').

Candidate Silencer CREs are Enriched in Broad ACRs

To assess the stability and change of H3K27me3 related ACRs across grass lineages, we focused on comparing *O. sativa*, *Z. mays*, and *S. bicolor* using previously published H3K27me3 ChIP-seq data (Lu et al., 2019). We considered ACRs near or within H3K27me3 regions and classified them into two groups: H3K27me3-broad, representing H3K27me3 associated ACRs with chromatin accessibility in many cell types and H3K27me3-cell-type specific, those H3K27me3 associated ACRs with chromatin accessibility in few cell types (Figure 4.5a). The composition of H3K27me3-broad and H3K27me3-cell-type-specific ACRs was consistent across all species (Extended Data Figure 4.8a). H3K27me3-broad ACRs exhibited a depletion of H3K27me3 at the ACR (Figure 4.5b), consistent with the absence of nucleosomes in ACRs (Minnoye et al., 2021). In contrast, H3K27me3 depletion was not observed in H3K27me3-cell-type-specific ACRs, with most cells in the bulk H3K27me3 measurement likely containing nucleosomes with H3K27me3 (Figure 4.5b). This is consistent with the H3K27me3-cell-type-specific ACRs potentially acting after the removal of facultative heterochromatin in a specific cell type(s). However, the chromatin accessibility of the H3K27me3-broad ACRs appears to be concurrent with H3K27me3, suggesting these ACRs may regulate H3K27me3 maintenance and removal across the majority of cellular contexts.

To assess the transcriptional state of genes near the H3K27me3-broad ACRs, we evaluated snRNA-seq/single-cell RNA sequencing (scRNA-seq) from *O. sativa* seedling (Supplementary Figure 4.7) and root, and *Z. mays* seedling (T.-Q. Zhang et al., 2021). The results revealed significantly lower expression (p = 3e-34 to 4e-06; Wilcoxon test) for H3K27me3-broad ACRs associated genes across many cell types (Figure 4.5c;

Extended Data Figure 4.8b and c; Supplementary Table 11). Moreover, 58 bulk RNA-seq libraries from *O. sativa* organs, demonstrated that the expression of genes near H3K27me3-broad ACRs were significantly lower (p = 2e-07 to 0.0265; Wilcoxon test) than genes near H3K27me3-absent broad ACRs (Extended Data Figure 4.8d) (Y. Yu et al., 2022). To dissect the roles of H3K27me3-broad ACRs, we identified 2,164 H3K27me3-broad ACRs and measured neighboring gene expression in *O. sativa* cells (Supplementary Table 12). About 926 (~42.8%) of the H3K27me3-broad ACRs were associated with 838 genes that exhibited no expression across any sampled cell type (Extended Data Figure 4.8e; Supplementary Table 12), consistent with these H3K27me3 proximal genes only being expressed under specific conditions. The 1,108 expressed genes associated with H3K27me3-broad ACRs were enriched (p < 2e-16; Fisher's exact test) for cell-type specificity compared to genes without H3K27me3 (Extended Data Figure 4.8f). In summary, single-cell expression analysis revealed that the genes linked to H3K27me3-broad ACRs exhibited the hallmarks of facultative gene silencing.

We hypothesized that the H3K27me3-broad ACRs would be enriched for PRC2 silencer elements, as their consistent chromatin accessibility provides an avenue to recruit PRC2 to maintain H3K27me3 throughout development. Supporting the presence of silencer CREs with these ACRs, a known silencer CRE ~5.3 kb upstream of *FRIZZY PANICLE* was within a H3K27me3-broad ACR (Extended Data Figure 4.8g) (Bai et al., 2017). To exploit the known Polycomb *A. thaliana* targets, we used scATAC-seq and H3K27me3 data from *A. thaliana* roots and annotated H3K27me3-broad ACRs (Bai et al., 2017; Marand et al., 2021; Wang et al., 2023). The *A. thaliana* H3K27me3-broad ACRs significantly (p < 2e-16; Binomial test) captured 53 of the 170 known Polycomb

responsive elements compared to a control class of ACRs, supporting their putative silencer function (Extended Data Figure 4.8h) (Xiao et al., 2017). Furthermore, we implemented a *de novo* motif analysis on the 170 *A. thaliana* elements and identified all reported Polycomb response element (PRE) motifs (CTCC, CCG, G-box, GA-repeat, AC-rich, and Telobox) (Figure 4.5d) (Xiao et al., 2017). Using these motifs and our chromatin accessibility data, we predicted putative binding sites in *O. sativa*, *Z. mays*, and *S. bicolor*, and observed that five motifs were significantly (p = 2e-178 to 1e-05; Binomial test) enriched in the H3K27me3-broad ACRs compared to a genomic control (Figure 4.5d). About 88.0% to 92.7% of H3K27me3-broad ACRs contained at least one PRE motif and 0.1% to 0.2% encompassing all the six of the major PRE classes (Extended Data Figure 4.8i). EMF2b is an essential component of the PRC2 complex (Cao et al., 2002; Tan et al., 2022; Tonosaki & Kinoshita, 2015). About 63.7% (4,053/6,364) of the H3K27me3-broad ACRs significantly (p < 2e-16; Binomial test) overlapped peaks from ChIP-seq targeting EMF2b compared to a control class of ACRs (Figure 4.5e). EMF2b occupancy was enriched at H3K27me3-broad ACRs (Figure 4.5f), supporting H3K27me3-broad ACRs acting as PRC2 recruiting silencers. Beyond PRE motifs, we observed significant enrichment (p = 1e-16 to 0.0376; Hypergeometric test) of four TF family motifs in H3K27me3-broad ACRs: APETALA2-like (AP2), basic Helix-Loop-Helix (bHLH), Basic Leucine Zipper (bZIP), and C2H2 zinc-finger (ZnF) (Extended Data Figure 4.8j; Supplementary Table 13). AP2 and C2H2 are known to recruit PRC2 (Xiao et al., 2017). In general, the motif discovery within the H3K27me3-broad ACRs supports them harboring H3K27me3 silencer CREs.

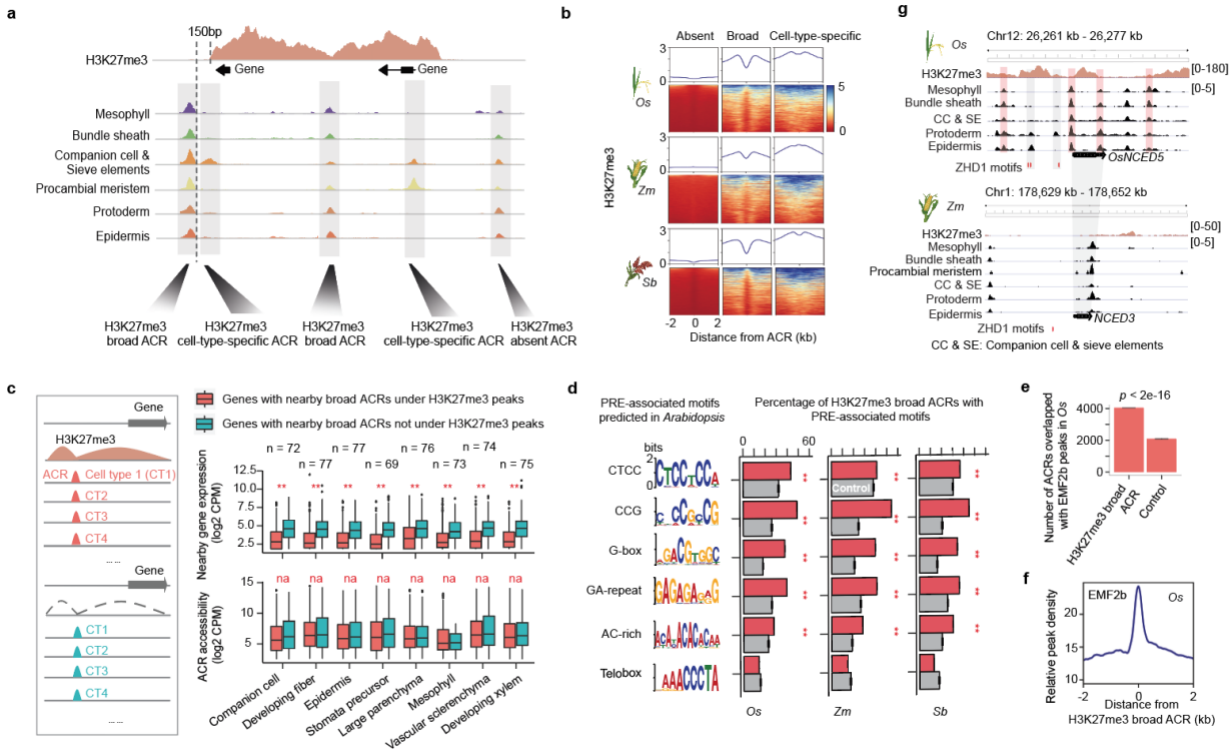

Figure 4.5. Discovery of candidate silencer CREs across species. a, A sketch graph illustrates the classification of ACRs based on their proximity to H3K27me3 peaks. We classified the ACRs into two groups: H3K27me3-associated ACRs (found within or surrounding H3K27me3 peaks) and H3K27me3-absent ACRs. The H3K27me3-associated ACR were further divided into broad ACRs, characterized by chromatin accessibility in at least five cell types, and cell-type-specific ACRs, accessible in less than three out of six examined cell types across all the species. **b,** Alignment of H3K27me3 chromatin attributes at summits of distinct ACR groups. **c,** A comparative analysis of expression levels and chromatin accessibility of genes surrounding broad ACRs under and outside of H3K27me3 peaks. ** indicate p value < 0.01, which was performed by the Wilcoxon test (alternative = 'two.sided'). The broad ACRs where the H3K27me3 region overlapped >50% of the gene body were positioned within 500 to 5,000 bp upstream of the transcriptional start site of their nearest gene. **d,** Percentage of H3K27me3-broad ACRs in *O. sativa*, *Z. mays*, and *S. bicolor* capturing six known motifs enriched in PREs in *A. thaliana*. ** indicate p value < 0.01, which was performed by the Binomial test (alternative = 'two.sided'; See Methods: Construction of control sets for enrichment tests). **e**, The number of H3K27me3-broad-ACR capturing EMF2b ChIP-seq peak. The significance test was done by using the Binomial test (See Methods: Construction of control sets for enrichment tests; alternative = 'two.sided'). **f**, A metaplot of EMF2b ChIP-seq profile at summits of H3K27me3-broad ACRs identified from *O. sativa* leaf. **g,** a screenshot of *OsNCED5* accessibility in *O. sativa* and *NCED3* accessibility in *Z. mays* L1 cells which contains four H3K27me3-broad ACRs and two *O. sativa* epidermal specific and species-specific ACRs with three ZHD1 motif sites.

Mirroring our previous syntenic ACR analysis, we split the *O. sativa* H3K27me3-broad ACRs into three classes, shared, variable and species specific, to investigate the relationship between these candidate silencers and species divergence (Extended Data Figure 4.9a-c; Supplementary Table 14). Between 54% to 61% of the H3K27me3-broad ACRs were present in the species-specific class, and the H3K27me3-broad ACRs were more likely to be species-specific compared to broad ACRs without H3K27me3 (Extended Data Figure 4.9d). Since these H3K27me3-broad ACRs exhibit hallmarks of PRC2 recruitment, we suspect that altered silencer CREs use context to drive species-specific developmental and environmental responses.

Since we identified L1 cells as being enriched in species-specific ACRs, we sought to examine the changes in H3K27me3 regulation within this tissue. We examined our previously identified 87 *O. sativa*-to-*Z. mays* orthologs to see if these genes contained H3K27me3. We observed 18 of these 87 genes were close to H3K27me3-broad ACRs (Supplementary Table 15). For example, we identified four H3K27me3-broad ACRs, and three ZHD1 motifs within two species-specific ACRs surrounding *OsNCED5* that were specifically accessible in L1 cells (Figure 4.5g). OsNCED5 TF is known to regulate tolerance to water stress and regulate leaf senescence in *O. sativa* (Huang et al., 2019). These results highlight that H3K27me3 mediated silencing may play a critical role in divergent regulation in L1 layer development.

Discussion

Our comparison of *O. sativa* with four other grasses revealed patterns in the evolutionary dynamics of *O. sativa* ACRs within syntenic and non-syntenic regions and

discovered that the grass L1 layer exhibits elevated rates of transcriptional regulatory divergence compared to other tissue/cell types. *O. sativa* had an increased prevalence of the ZHD1 TF motif, partly driven by past LTR TE expansion, that contributed to elevated ortholog L1 expression compared to *Z. mays* (Figure 4.6a). This was driven by a contrasting dichotomy; the epidermal TF motifs were the most cell-type specific of all those studied, yet their cognate ACRs exhibited the strongest target divergence amongst the measured species. This duality highlights tandem conservation of core epidermal CRE and motif targets, and the rapid co-option of novel regulatory regions into these existing regulatory frameworks. This rapid regulatory evolution might relate to the dynamic environmental pressures the epidermis has evolved to withstand (Javelle et al., 2011). Although to a lesser extent than the epidermis, this interesting contrast, where the cell-type restricted TF motifs are conserved and the cell-type-specific chromatin accessibility of cognate CREs are not, extends to other cell types. This supports a larger pattern of novel CRE target evolution that co-opts established cell-type-specific TF networks.

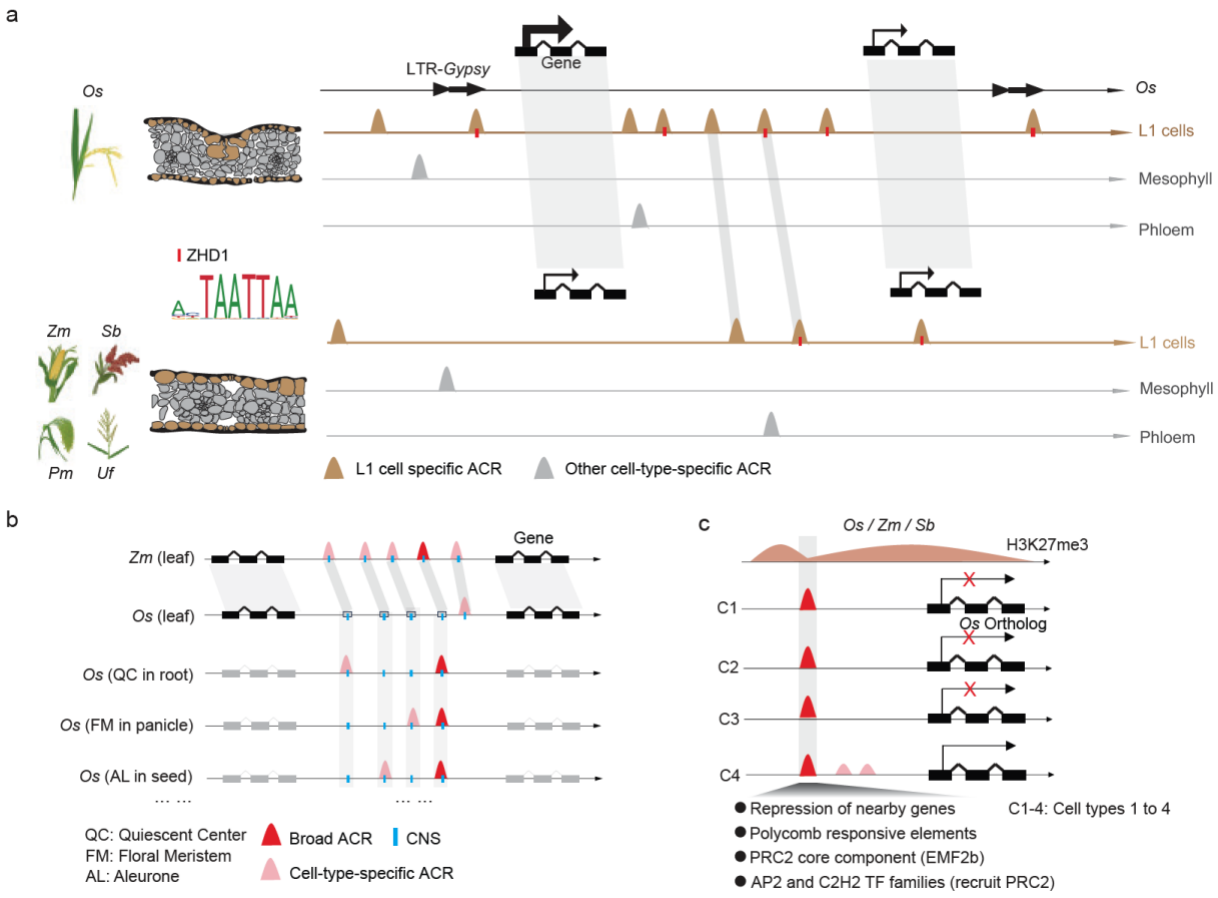

Figure 4.6. Evolution of cell-type-specific ACRs and CREs. **a,** The analysis of leaf cell types across these species revealed an enrichment of cell-type-specific ACRs in species-specific regions. Notably, these species-specific ACRs were enriched within L1 cells compared to all others examined. Additionally, it was observed that epidermal cell-specific ACRs significantly overlapped with LTR-*Gypsy* TEs, which were enriched for the ZHD1 motif known to regulate leaf curling. The L1 ACRs within the ZHD1 motif likely contribute to species-specific elevated expression of genes involved in L1 cell development. **b,** We found an enrichment of CNS in cell-type-specific ACRs. Although some CNS ACRs retained the same cell-type specificity between *O. sativa* and *Z. mays,* these CNS ACRs often switched tissue or cell-type accessibility between grass species. **c,** Despite being within facultative heterochromatin, H3K27me3-broad ACRs are accessible in many cell types, providing a physical entry point for PRC2 to bind. Several lines of evidence support that the H3K27me3-broad ACRs contain silencer CREs. Specifically, these ACRs are linked to transcriptionally silent genes, are enriched for PRE motifs, enriched for TF family motifs (AP2 and C2H2) reported to recruit PRC2, and enriched for PRC2 subunit (EMF2b) ChIP-seq peaks.

Highlighting the rapid rate of regulatory evolution, ACRs, and the CREs within them, underpin phenotypic variation within plant and mammalian species (Andrews et al., 2023; Cusanovich et al., 2018; Engelhorn et al., 2023; Marand et al., 2021). Despite the link between CREs and phenotypic variation, how these transcriptional regulatory circuits have changed during species divergence is challenging to address. This is partly due to rapid CRE changes occluding pairwise comparison; even closely related plant species share but a fraction of their ACR/CRE complements (Zhao & Schranz, 2019). We used a comparative single-cell epigenomics approach to characterize the evolution of cell-type-specific ACRs and CREs in grasses. We demonstrate that grass cell-type-specific ACRs have changed tremendously over 50 million years, with relatively few (0.2-0.4%) cell-type-specific ACRs remaining conserved across the examined plant species. This contrasts with mammals, where ~17% (66,781/384,412) of cell-type-specific ACRs remain conserved over 87 million years (Zemke et al., 2023). The huge difference in ACR conservation highlights the speed at which plant CRE evolution takes place compared to mammals. The repeated whole genome duplications in plant lineages (Clark & Donoghue, 2018), and the functional redundancy they provide, may be the fuel for rapid CRE divergence driving plants' adaptation to diverse environments (Jump et al., 2009).

Integration of the *O. sativa* atlas with CNS revealed ~65% were accessible in at least one cell type. We expect that most CNS not captured by an ACR in our study are likely accessible in an unsampled cell type or developmental condition. This stresses the need for expanded accessible chromatin atlases using more tissues, segments of development, and environment. Our results also demonstrated that CNS ACRs switch cell-type

specificity between species (Figure 4.6b). This shows that although CNS can be conserved function, this function can sometimes be altered in different cellular contexts across lineages (Ciren et al., 2023). However, it remains possible that the main CNS function conserved between *O. sativa* and *Z. mays* occurs outside of leaves; i.e. these switching CNS may underpin conserved functions in conserved non-leaf contexts. Nonetheless, the switching of CNS leaf cell-type accessibility highlights the importance of merging chromatin accessibility data with CNS datasets, as the assumption of conserved CNS sequence equaling conserved CRE function is often invalid.

Much focus has been placed on enhancer CREs within ACRs, yet silencers are equally important, as they repress gene expression until the proper developmental or environmental cues. Our prior research uncovered that some ACRs flanking H3K27me3 are linked to the suppression of nearby genes[6,9]; however, the question remains whether these ACRs function as silencers. Our *O. sativa* cell-type ACR atlas allowed the identification of ACRs within H3K27me3 regions that were accessible in most cell types (Lu et al., 2019, 2019). Several lines of evidence support these H3K27me3-broad ACRs as silencers of linked, transcriptionally repressed genes. Specifically, these putative silencers were enriched for CNS, PREs and related TF motifs, and PRC2 *in vivo* occupancy (Figure 4.6c) (Tan et al., 2022). Future gene editing of these putative silencers will reveal more about their role in PRC2 recruitment and gene silencing. Similar H3K27me3-broad ACR putative silencers were also present in the epigenomic landscape of *O. sativa*, *Z. mays*, and *S. bicolor*, suggesting this is a conserved feature of grass genomes. H3K27me3 silencing is deeply conserved in eukaryotes, and we hypothesize

that other single-cell comparative genomic investigations will find this pattern of broadly accessible silencers in all species with H3K27me3 (Wiles & Selker, 2017).

Our rice atlas of cell-type-specific ACRs and this cross-species analysis provides a useful resource to enhance our understanding of regulatory evolution more broadly. This resource, and these observations, will fuel research into identifying key CREs controlling specific genes by demarcating high-confident targets for genome editing.

Methods

Preparation of plant materials

Early seedlings, specifically seedling tissues above ground, were collected 7 and 14 days after sowing. Flag leaf tissue was harvested at the V4 stage, characterized by collar formation on leaf 4 of the main stem. Axillary buds were obtained from rice plants grown in the greenhouse at approximately the V8 stage. Rice seminal and crown root tips (bottom 2 cm) were gathered at the same stage as seedling tissues, 14 days after sowing. Panicle tissue was acquired from rice plants grown in the greenhouse. Inflorescence primordia (2-5 mm) were extracted from shoots harvested at the R1 growth stage, where panicle branches had formed. Early seeds were harvested at approximately six days after pollination (DAP), and late seeds at approximately ten DAP. All tissues were collected between 8 and 9 am, and all fresh materials were promptly utilized for library construction starting at 10 am.

Single-cell ATAC-seq library preparation

Nuclei isolation and purification were performed as described previously (X. Zhang et al., 2024). In brief, the tissue was finely chopped on ice for approximately 2 minutes using 600 µL of pre-chilled Nuclei Isolation Buffer (NIB, 10 mM MES-KOH at pH 5.4, 10 mM NaCl, 250 mM sucrose, 0.1 mM spermine, 0.5 mM spermidine, 1 mM DTT, 1% BSA, and 0.5% TritonX-100). After chopping, the entire mixture was passed through a 40-µm cell strainer and then subjected to centrifugation at 500 rcf for 5 minutes at 4°C. The supernatant was carefully decanted, and the pellet was reconstituted in 500 µL of NIB wash buffer, which consisted of 10 mM MES-KOH at pH 5.4, 10 mM NaCl, 250 mM sucrose, 0.1 mM spermine, 0.5 mM spermidine, 1 mM DTT, and 1% BSA. The sample was filtered again, this time through a 10-µm cell strainer, and then gently layered onto the surface of 1 mL of a 35% Percoll buffer, prepared by mixing 35% Percoll with 65% NIB wash buffer, in a 1.5-mL centrifuge tube. The nuclei were subjected to centrifugation at 500 rcf for 10 minutes at 4°C. Following centrifugation, the supernatant was carefully removed, and the pellets were resuspended in 10 µL of diluted nuclei buffer (DNB, 10X Genomics Cat# 2000207). About 5 µL of nuclei were diluted 10 times and stained with DAPI (Sigma Cat. D9542) and then the nuclei quality and density were evaluated with a hemocytometer under an epifluorescence microscope. The original nuclei were diluted with a DNB buffer to a final concentration of 3,200 nuclei per uL. Finally, 5 uL of nuclei (16,000 nuclei in total) were used as input for scATAC-seq library preparation. scATAC-seq libraries were prepared using the Chromium scATAC v1.1 (Next GEM) kit from 10X Genomics (Cat# 1000175), following the manufacturer's instructions. (10xGenomics,

CG000209_Chromium_NextGEM_SingleCell_ATAC_ReagentKits_v1.1_UserGuide_RevE). Libraries were sequenced with Illumina NovaSeq 6000 in dual-index mode with eight and 16 cycles for i7 and i5 index, respectively.

Single-nuclei RNA-seq library preparation and data analysis

The protocol for nuclei isolation and purification was adapted from the previously described scATAC-seq method. In summary, to minimize RNA degradation and leakage, the tissue was finely chopped on ice for approximately 1 minute using 600 µL of pre-chilled Nuclei Isolation Buffer containing 0.4U/µL RNase inhibitor (Roche, Protector RNase Inhibitor, Cat. RNAINH-RO) and a comparatively low detergent concentration of 0.1% NP-40. Following chopping, the entire mixture was passed through a 40-µm cell strainer and then subjected to centrifugation at 500 rcf for 5 minutes at 4°C. The supernatant was carefully decanted, and the pellet was reconstituted in 500 µL of NIB wash buffer, comprising 10 mM MES-KOH at pH 5.4, 10 mM NaCl, 250 mM sucrose, 0.5% BSA, and 0.2U/µL RNase inhibitor. The sample was filtered again, this time through a 10-µm cell strainer, and gently layered onto the surface of 1 mL of a 35% Percoll buffer. The Percoll buffer was prepared by mixing 35% Percoll with 65% NIB wash buffer in a 1.5-mL centrifuge tube. The nuclei were then subjected to centrifugation at 500 rcf for 10 minutes at 4°C. After centrifugation, the supernatant was carefully removed, and the pellets were resuspended in 50 µL of NIB wash buffer. Approximately 5 µL of nuclei were diluted tenfold and stained with DAPI (Sigma Cat. D9542). Subsequently, the nuclei's quality and density were evaluated with a hemocytometer under a microscope. The original nuclei were further diluted with a DNB buffer to

achieve a final concentration of 1,000 nuclei per μL. Ultimately, a total of 16,000 nuclei were used as input for snRNA-seq library preparation. For snRNA-seq library preparation, we employed the Chromium Next GEM Single Cell 3'GEM Kit v3.1 from 10X Genomics (Cat# PN-1000123), following the manufacturer's instructions (10xGenomics, CG000315_ChromiumNextGEMSingleCell3-_GeneExpression_v3.1_DualIndex_RevB). The libraries were subsequently sequenced using the Illumina NovaSeq 6000 in dual-index mode with 10 cycles for the i7 and i5 indices, respectively.

The raw BCL files obtained after sequencing were demultiplexed and converted into FASTQ format using the default settings of the 10X Genomics tool cellranger mkfastq (v7.0.0) (Zheng et al., 2017). The raw reads were processed with cellranger count (v7.0.0) using the Japonica rice reference genome[57] (v7.0) (Zheng et al., 2017). Genes were kept if they were expressed in more than three cells and each cell having a gene expression level of at least 1,000 but no more than 10,000 expressed genes. Cells with over 5% mitochondria or chloroplast counts were filtered out. The expression matrix was normalized to mitigate batch effects based on global-scaling normalization and multicanonical correlation analysis in Seurat (v4.0) (Satija et al., 2015). The Scrublet tool was employed to predict doublet cells in this dataset. SCTransform in Seurat was used to normalize the data and identify variable genes (Wolock et al., 2019).The nearest neighbors were computed using FindNeighbors using 30 PCA dimensions. The clusters were identified using FindClusters with a resolution of 1. The cell types were annotated based on the marker gene list (Supplementary Table 2). To identify genes exhibiting

higher expression in a particular cell type than in the others, we utilized the 'cpm' function from edgeR (v3.38.1), for normalizing the expression matrix (Robinson et al., 2010). Genes within a specific cell type that displayed more than 1.5-fold change in log2 Counts Per Million (CPM) values compared to the average log2 CPM across all cell types were determined as specifically expressed genes in that particular cell type.

Slide-seq library preparation and data analysis

Root tissues from rice seedlings 14 days after sowing were used for the Slide-seq V2 spatial transcriptomics. The tissues were embedded in the Optimal Cutting Temperature (OCT) compound, snap-frozen in a cold 2-methylbutane bath, and cryosectioned into 10 um thick slices. The spatial transcriptome library was constructed following a published method (Rodriques et al., 2019; Stickels et al., 2021). In brief, the tissue slices were placed on the Slide-seq V2 puck and underwent the RNA hybridization and reverse transcription process. After tissue clearing and spatial bead collections, the cDNA was synthesized and amplified for a total of 14 cycles. The library was constructed using Nextera XT Library Prep Kit (Illumina, USA) following the manufacturer's instructions.

The reads alignment and quantification were conducted following the Slide-seq pipeline (https://github.com/MacoskoLab/slideseq-tools). The data processing is similar to the procedures applied in the analysis of snRNA-seq but setting a resolution of 0.7 for the FindClusters function in Seurat (v4.0) (Satija et al., 2015). The cell types were annotated based on the histology of cross-sectioned roots. The marker genes of each cluster were identified using the Wilcoxon Rank Sum test in FindAllMarkers.

RNA *in situ* Hybridization

The rice samples were put into the vacuum tissue processor (HistoCore PEARL, Leica) to fix, dehydrate, clear, and embed, and were subsequently embedded in paraffin (Paraplast Plus, Leica). The samples were sliced into 8 μm sections with a microtome (Leica RM2265). The cDNAs of the genes were amplified with their specific primer pairs in situ_F/in situ_R and subcloned into the pGEM-T vector (Supplementary Table 16). The pGEM-gene vectors were used as the template to generate sense and antisense RNA probes. Digoxigenin-labelled RNA probes were prepared using a DIG Northern Starter Kit (Roche) according to the manufacturer's instructions. Slides were observed under bright fields through a microscope (ZEISS) and photographed with an Axiocam 512 color charge-coupled device (CCD) camera.

Raw reads processing of scATAC-seq

Data processing was executed independently for each tissue and/or replicate. Initially, raw BCL files were demultiplexed and converted into FASTQ format, utilizing the default settings of the 10X Genomics tool cellranger-atac make-fastq (v1.2.0). Employing the Japonica rice reference genome (v7.0), the raw reads underwent processing using cellranger-atac count (v1.2.0) (S. Ouyang et al., 2007; Satpathy et al., 2019). These steps encompassed adaptor/quality trimming, mapping, and barcode attachment/correction. Subsequently to the initial processing, reads that were uniquely mapped with mapping quality > 10 and correctly paired were subjected to further refinement through SAMtools view (v1.7; -f 3 -q 10; Li et al. 2009). To mitigate the impact of polymerase chain

reaction (PCR) duplicates, the Picard tool MarkDuplicates (v2.16.0) was applied on a per-nucleus basis (*Picard Tools - By Broad Institute*, n.d.). To elevate data quality, a blacklist of regions was devised to exclude potentially spurious reads. The methodology involved the exclusion of regions displaying bias in Tn5 integration from Tn5-treated genomic DNA. Specifically, regions characterized by 1-kilobase windows with coverage exceeding four times the genome-wide median were eliminated. We further leveraged ChIP-seq input data[6], to filter out collapsed sequences in the reference using the same criteria. This blacklist also incorporated sequences of low-complexity and homopolymeric sequences through RepeatMasker (v4.1.2) (Tarailo-Graovac & Chen, 2009). Moreover, nuclear sequences exhibiting homology surpassing 80% to mitochondrial and chloroplast genomes[87] (BLAST+; v2.11.0) were also included within the blacklist (Camacho et al., 2009). Furthermore, BAM alignments were converted into BED format, wherein the coordinates of reads mapping to positive and negative strands were subjected to a shift by +4 and -5, respectively. The unique Tn5 integration sites per barcode were finally retained for subsequent analyses.

Identifying high-quality nuclei

To ensure the acquisition of high-quality nuclei, we harnessed the capabilities of the Socrates package for streamlined processing (Marand et al., 2021). To gauge the fraction of reads within peaks, we employed MACS2 (v2.2.7.1) with specific parameters (genomesize = 3e8, shift= -75, extsize = 150, fdr = 0.05) on the bulk Tn5 integration sites (Y. Zhang et al., 2008). Subsequently, we quantified the number of integration sites per barcode using the callACRs function. Next, we estimated the proximity of Tn5

integration sites to genes, focusing on a 2 kb window surrounding the TSS. This estimation was achieved through the buildMetaData function, which culminated in the creation of a meta file. For further refinement of cell selection, we harnessed the findCells function, implementing several criteria: 1) A minimum read depth of 1,000 Tn5 integration sites was required. 2) The total number of cells was capped at 16,000. 3) The proportion of reads mapping to TSS sites was above 0.2, accompanied by a z-score threshold of 3. 4) Barcode FRiP scores were required to surpass 0.1, alongside a z-score threshold of 2. 5) We filtered out barcodes exhibiting a proportion of reads mapping to mitochondrial and chloroplast genomes that exceeded two standard deviations from the library mean. 6) We finally used the detectDoublets function to estimate doublet likelihood by creating synthetic doublets and conducting enrichment analysis. These multiple steps ensured the meticulous identification and selection of individual cells, facilitating a robust foundation for subsequent analyses.

Nuclei clustering

For the nuclei clustering, we leveraged all functions from the Socrates package (Marand et al., 2021). We binned the entire genome into consecutive windows, each spanning 500 bp. We then tabulated the count of windows featuring Tn5 insertions per cell. Barcodes falling below one standard deviation from the mean feature counts (with a z-score less than 1) were excluded. Moreover, barcodes with fewer than 1,000 features were eliminated. We pruned windows that exhibited accessibility in less than 0.5% or more than 99.5% of all nuclei. To standardize the cleaned matrix, we applied a term frequency-inverse document frequency normalization function. The dimensionality of the

normalized matrix underwent reduction through the utilization of non-negative matrix factorization, facilitated by the R package RcppML (v0.3.7) (DeBruine et al., 2021). We retained 50 column vectors from an uncentered matrix. Subsequently, we selected the top 30,000 windows that displayed the highest residual variance across all cells. This selection was based on fitting a model where the proportion of cells with accessibility served as an independent variable and variance as the dependent variable. To further reduce the dimensionality of the nuclei embedding, we employed the UMAP technique using umap-learn (k = 50, min_dist = 0.1, metrix = 'euclidean') in R (v0.2.8.0) (McInnes et al., 2020). Furthermore, we clustered nuclei using the callClusters function within the Socrates framework (Marand et al., 2021). Louvain clustering was applied, with a setting of k=50 nearest neighbors at a resolution of 0.7. This process underwent 100 iterations with 100 random starts. Clusters with an aggregated read depth of less than 1 million and 50 cells were subsequently eliminated. To filter outlying cells in the UMAP embedding, we estimated the mean distance for each nucleus using its 50 nearest neighbors. Nuclei that exceeded 3 standard deviations from the mean distance were deemed outliers and removed from consideration.

Estimation of gene accessibility scores

To estimate the gene accessibility, we employed a strategy wherein the Tn5 insertion was counted across both the gene body region as well as a 500 bp extension upstream. Subsequently, we employed the SCTransform algorithm from the Seurat package (v4.0) to normalize the count matrix that was then transformed into a normalized accessibility score, with all positive values scaled to 1 (Satija et al., 2015).

Cell type validation

Upon completing the initial annotation process, which was based on a curated list of marker genes, we further expanded our marker repertoire by incorporating markers collected from published bulk RNA-seq data encompassing a diverse array of cell types, which were acquired *via* laser capture dissection. In brief, we collected the markers from several studies, including three distinct cell types within rice leaves[91] and ten cell types across rice seed organs[92,93] (Hua et al., 2021; Itoh et al., 2016; T.-Y. Wu et al., 2020). From these sources, we selected the top 100 variably expressed markers for each cell type and employed them to compute cell identity enrichment scores. We undertook a comprehensive assessment of markers linked to the different cell types. Subsequently, we randomly drew 100 markers from this pool and repeated this procedure 1,000 times to construct a null distribution based on marker chromatin accessibility scores. For each target cell type marker, we compared their accessibility scores to this null distribution. This facilitated the derivation of an enrichment score per cell, delineating the marker's significance for each representative cell type. We next employed a MAGIC algorithm to refine these enrichment scores (van Dijk et al., 2018). These scores were then mapped onto a UMAP plot, enhancing the cell identity annotation.

Furthermore, we undertook validation of our single-cell chromatin accessibility atlas through integration with published scRNA-seq from rice root tissue. This validation was achieved through two distinct approaches (Supplementary Figure 4.8). In the first approach, we leveraged the marked enrichment technique, adapting the above mentioned

methodology with the incorporation of the top 20 markers derived from marker identification using the 'FindMarkers' function in Seurat[78] (v4.0) (Satija et al., 2015). Following the acquisition of a smoothed score for each cell type, individual cells were annotated to specific cell types based on the largest enrichment score within that cell type. A threshold was further set, requiring the maximum score to exceed 0.5 for confident labeling; otherwise, the cell was labeled as 'Unknown'. The second approach entailed employing a k-nearest neighbor (knn) strategy. This strategy commenced with the normalization of scRNA-seq datasets, mirroring the process applied to scATAC-seq datasets. The top 3,000 most variable genes within the scATAC-seq dataset were then identified using the Seurat function 'FindVariableFeatures', subsequently filtering to include only genes common to both datasets. By treating the scRNA-seq cells as a reference, a dimension reduction process was conducted to generate a loading matrix, which was then utilized to project the scATAC-seq cells onto the scRNA-seq cell embedding. The integration of these two datasets was achieved through the Harmony algorithm (v0.1.0) (Korsunsky et al., 2019). Within the dual embeddings, the 20 nearest neighbors of each scATAC-seq cell in the scRNA-seq dataset were computed. The most frequent label among these RNA neighbors (> 10 cells) was subsequently assigned as the label for each scATAC-seq cell or designated as NA if no label meeting this threshold was identified.

Cell cycle prediction

The prediction of cell cycle stages per nucleus was executed similarly to annotating cell identities based on the aforementioned enriched scores. In brief, we collected a set of 55

cell-cycle marker genes from a previous study (Pettkó-Szandtner et al., 2015). For every cell-cycle stage, the cumulative gene accessibility score for each nucleus was computed. These resultant scores were subsequently normalized using the mean and standard deviation derived from 1,000 permutations of the 55 random cell-cycle stage genes, with exclusion of the focal stage. Z-scores corresponding to each cell-cycle stage were transformed into probabilities using the 'pnorm' function in R. Furthermore, the cell-cycle stage displaying the highest probability was designated as the most probable cell stage.

ACR identification

Upon segregating the comprehensive single-base resolution Tn5 insertion sites BED dataset into distinct subsets aligned with annotated cell types, we executed the MACS2 tool (v2.2.7.1) for precise peak identification per cell type (Y. Zhang et al., 2008). Notably we employed non-default parameters, specifically: --extsize 150, --shift -75, --nomodel -keep-dup all. To mitigate potential false positives, a permutation strategy was applied, generating an equal number of peaks based on regions that were mappable and non-exonic. This approach encompassed the assessment of Tn5 insertion sites and density within both the original and permuted peak groups. By scrutinizing the permutation outcomes, we devised empirically derived false discovery rate (FDR) thresholds specific to each cell type. This entailed determining the minimum Tn5 density score within the permutation cohort where the FDR remained < 0.05. To further eliminate peaks that exhibited significant overlap with nucleosomes, we applied the NucleoATAC tool (v0.2.1) to identify potential nucleosome placements (Schep et al., 2015). Peaks that

featured over 50% alignment with predicted nucleosomes were systematically removed. The average fragment size of reads overlapping with the peaks were calculated and the peaks with the average fragment size > 150 bp were filtered out. Ultimately, the pool of peaks for each cell type was amalgamated and fine-tuned, yielding 500 bp windows that were centered on the summit of ACR coverage.

Identification of cell-type-specific ACRs in rice atlas

To identify cell-type-specific ACRs in the rice atlas, we implemented a series of cutoffs to determine whether the peak is accessible for a specific cell type. For each cell type, we first normalized the read coverage depth obtained from the MACS2 tool divided by total count of reads, and ensured that the maximum of normalized coverage within the peak exceeded a predefined threshold set at 2. Additionally, we calculated Tn5 integration sites per peak, filtering out peaks with fewer than 20 integration sites. Subsequently, we constructed a peak by cell type matrix with Tn5 integration site counts. This matrix underwent normalization using the 'cpm' function wrapped in edgeR (v3.38.1) and 'normalize.quantiles' function wrapped within preprocessCore (v1.57.1) in the R programming environment (*preprocessCore*, n.d.; Robinson et al., 2010). To further refine our selection, a threshold of 2 was set for the counts per million value per peak per cell type. Peaks that satisfied these distinct cutoff criteria were deemed accessible in the designated cell types. We identified these peaks as displaying chromatin accessibility in anywhere from one to ten of the primary cell types among the total pool of 60 main cell types.

Identification of cell-type-specific ACRs in leaf tissue across five species

To account for the lower cell number found in these data sets, cell-type-specific ACRs were identified using the same method as found in our prior study (Mendieta et al., 2024). In brief we utilized a modified entropy metric combined with a bootstrapping approach (Domcke et al., 2020). For each cell-type and species a sample of 250 cells were taken 5,000 times with replacement and a specificity metric was calculated. This specificity metric was then compared against a series of 5,000 null distributions consisting of a random shuffle of 250 cells which were of mixed cell populations. Then, a nonparametric test was used to the median real bootstrap specificity score versus the null, and ACRs were labeled as cell-type-specific if they had a *p* value less than 0.0001.

Correlation between chromatin accessibility of TF genes and motif deviation

We sourced rice and *A. thaliana* TFs from PlantTFDB[99] (v4.0) database (Jin et al., 2017). To identify rice orthologs of *A. thaliana* TFs, we employed BLAST[87] (BLAST+; v2.11.0) by utilizing protein fasta alignments with an e-value threshold of 1e-5 used for significance (Camacho et al., 2009). Alignments were restricted to fasta sequences categorized as TFs from either species. To further refine the putative orthologs, we applied filters based on functional similarity to *A. thaliana* TFs. Alignments with less than 15% identity were excluded, along with rice TFs associated with distinct families. From the remaining candidates, we selected the orthologs demonstrating the highest Pearson correlation coefficient concerning the motif deviation scores. Motif deviation scores of specific TF motifs within nuclei were computed *via* chromVAR[100] (v1.18.0) (Buenrostro et al., 2015).

Linear-model based motif enrichment analysis

We employed the FIMO tool from the MEME suite (v5.1.1) with a significance threshold of p value < 10^{-5} to predict motif locations (Bailey et al., 2015). The motif frequency matrix used was sourced from the JASPAR plants motif database (v9) (Castro-Mondragon et al., 2022). Subsequently, we constructed a binarized peak-by-motif matrix and a motif-by-cell count matrix. This involved multiplying the peak-by-cell matrix with the peak-by-motif matrix. To address potential overrepresentation and computational efficiency, down-sampling was implemented. Specifically, we standardized the cell count by randomly selecting 412 cells per cell type per species. This count represents the lowest observed cell count for a given cell type across all species. For each cell type annotation, total motif counts were predicted through negative binomial regression. This involved two input variables: an indicator column for the annotation, serving as the primary variable of interest, and a covariate representing the logarithm of the total number of nonzero entries in the input peak matrix for each cell. The regression provided coefficients for the annotation indicator column and an intercept. These coefficients facilitated the estimation of fold changes in motif counts for the annotation of interest in relation to cells from all other annotations. This iterative process was conducted for all motifs across all cell types. The obtained p values were adjusted using the Benjamini-Hochberg procedure to account for multiple comparisons. Finally, enriched motifs were identified by applying a dual filter criterion: corrected p values < 0.01, fold-change of the top enriched TF motif in cell type-specific peaks for all cell types should be over 1, and beta (motif enrichment score) > 0.05 or beta > 0.

Binomial test-based motif enrichment analysis

To assess the enrichment of motifs in a target set of ACRs, we performed analysis for each specific motif. We randomly selected an equivalent number of ACRs as found in the target set, repeating this process 100 times. Notably, the randomly selected ACR set did not overlap with the actual target set of ACRs. Following this, we computed the average ratio of ACRs capturing the motif within the null distribution.

Subsequently, we executed an exact Binomial test, wherein we set this ratio as the hypothesized probability of success (Wagner-Menghin, 2014). The number of ACRs overlapping the motif in the target set was considered the number of successes, while the total number of ACRs in the target set represented the number of trials. The alternative hypothesis was specified as 'two.sided'. This meticulous approach allowed us to robustly evaluate and identify significant motif enrichments within the target set of ACRs.

Construction of control sets for enrichment tests

To perform comparative analysis of expression levels and chromatin accessibility of genes surrounding broad ACRs under and outside of H3K27me3 peaks, we sampled the same number of ACRs per cell type regarding the broad ACRs not under H3K27me3 peaks. This step is to make sure that their nearby gene chromatin accessibility exhibited similar values compared to the broad ACRs under the H3K27me3 peaks.

To check if the H3K27me3-broad-ACRs could significantly capture the known PREs and capture the EMF2b ChIP-seq peaks, we generated control sets by randomly selecting not-H3K27me3-broad-ACR instances 100 times, yielding a mean number value for the control sets. The Binomial test *p* value was calculated by comparing the mean ratio to the observed number of H3K27me3-broad ACRs overlapped with the PREs.

To test if H3K27me3-broad ACRs in *O. sativa*, *Z. mays*, and *S. bicolor* significantly capture six known motifs, we generated control sets by simulating sequences with the same length as ACRs 100 times, yielding a mean proportion for the control sets. The binomial test p value was calculated by comparing the mean ratio to the observed overlapping ratio of H3K27me3-broad ACRs capturing the motifs.

De novo motif analysis

To identify position weight matrix of six known motifs within 170 *A. thaliana* PREs[55], we employed the streme function with default settings from the MEME suite[101] (v5.1.1) (Bailey et al., 2015; Xiao et al., 2017). The control sequences were built up to match each PRE sequence by excluding exons, PREs, and unmappable regions, and they possess a similar GC content (< 5% average difference) and same sequence length compared to the positive set.

Identification of syntenic regions

Identification of syntenic gene blocks was done using the GENESPACE (v1.4) (Lovell et al., 2022). In brief, to establish orthologous relationships between ACR sequences, ACRs

in the *O. sativa* genome were extended to incorporate the two closest gene models for a 'query block' since GENESPACE only draws relationships between protein coding sequences. Then the GENESPACE function 'query_hits' was used with the argument 'synOnly = TRUE' to retrieve syntenic blocks. The resulting syntenic hits were further filtered to allow only a one-to-one relationship between *O.sativa* and the corresponding species. The corresponding syntenic blocks were then named and numbered, and both the genes and genomic coordinates were recorded.

To further identify corresponding ACRs within these blocks we set up a BLASTN pipeline (v2.13.0) (Camacho et al., 2009). For each comparison of species, using *O. sativa* as the reference the underlying nucleotide sequences of the syntenic regions were extracted using Seqkit, and used as the blast reference database (v2.5.1) (Shen et al., 2016). The sequences underlying the ACRs within the same syntenic region in a different species were then used as the query. The blast was done using the following parameters to allow for alignment of shorter sequences '-task blastn-short -evalue 1e-3 -max_target_seqs 4-word_size 7 -gapopen 5 -gapextend 2 -penalty -1 -reward 1 -outfmt 6'. This procedure was run for each syntenic region separately for all species comparisons. The resulting BLASTN files were combined, and then filtered using a custom script. Alignments were only considered valid if the e-value passed a stringent threshold of 1e-3, and the alignment was greater than 20 nucleotides with the majority of the shared ACRs (92% to 94%) containing the alignment regions including TF motif binding sites (Supplementary Figure 4.13). The resulting filtered BLAST files, and the

BED files generated from these BLAST files allowed us to draw our relationships between ACRs in the corresponding syntenic space.

Estimation of conservation scores

Conservation scores were predicted using PhyloP (v1.0), where values are scaled between 0 to 1, with one being highly conserved and 0 being non-conserved (Pollard et al., 2010). Phylogenies to train PhyloP were generated using PhyloFit (v1.0), and neutral and conserved sequences were identified using the whole genome aligner progressive cactus (Hubisz et al., 2011).

ChIP-seq analysis

The clean reads of EMF2b were downloaded from a previous study (Tan et al., 2022). The reads were mapped to the rice reference genome (v7.0) using bowtie2 (v2.5.2) with the following parameters: '--very-sensitive --end-to-end'. Reads with MAPQ > 5 were used for the subsequent analysis (Langmead & Salzberg, 2012). Aligned reads were sorted and duplicated reads were removed using SAMtools(v1.7). Peak calling was performed using epic2 with the following parameters: '-fdr 0.01 --bin-size 150 --gaps-allowed 1' (Danecek et al., 2021; Stovner & Sætrom, 2019)sa. The peak 'BED' and 'BIGWIG' files of H3K27me3 ChIP-seq data for leaf, root, and panicle rice organs were downloaded from RiceENCODE[111] (http://glab.hzau.edu.cn/RiceENCODE/).

GO enrichment test

The GO enrichment tests were performed based on the AgriGO[43] (v2) by setting the Chi-square statistical test and multi-test adjustment method is Hochberg (FDR) (Tian et al., 2017).

Additional resources

Cell-type resolved data can be viewed through our public Plant Epigenome JBrowse Genome Browser (http://epigenome.genetics.uga.edu/PlantEpigenome/index.html) (Hofmeister & Schmitz, 2018).

Data Availability

scATAC-seq data encompassing 18 libraries from nine organs were accessible in NCBI (PRJNA1007577/GSE252040; https://dataview.ncbi.nlm.nih.gov/object/PRJNA1007577?reviewer=kgarq48dii11vomg44kgr1jq66; PRJNA1052039; https://dataview.ncbi.nlm.nih.gov/object/PRJNA1052039?reviewer=flhu9sl84o5m999r1ph8tlmmbg)

Acknowledgements

This research was funded by the National Science Foundation (IOS-2134912) to SRW and RJS and the UGA Office of Research to RJS. APM and JPM were supported by the National Institutes of Health (K99GM144742) and (T32GM142623), respectively.

Contributions

R.J.S., H.Y., J.P.M., A.P.M., M.A.A.M., D.W., S.Z., and S.R.W. designed and conceived experiments and managed the project. X.Z., Y.L., Z.L., X.T., S.Z., Y.W., and H.Y. participated in material collection and sample processing. H.Y., J.P.M., and T.R. performed the bioinformatics analyses. H.Y. and J.P.M. wrote the manuscript. Y.L. Y.W. and Z.L. contributed to marker validation. R.J.S., J.P.M., A.P.M., M.A.A.M., D.W., S.Z., and S.R.W. edited the manuscript.

Competing interests

The authors declare no competing interests.

Discussion

The work presented in this thesis demonstrates the ways in which comparative genomics, in tandem with plant epigenomic data, can both further genomic resource development and provide insights into regulatory sequence evolution in plant genomes.

In the first chapter we demonstrate how chromatin modification data can ameliorate plant genome annotation issues. Genome annotation is a challenging endeavor requiring multiple lines of evidence in the form of gene expression data, additional genome annotations from closely related species, as well as *ab initio* gene prediction tools (Cantarel et al., 2007; Mudge & Harrow, 2016; Salzberg, 2019). However, each of these methods come with *a priori* assumptions. For instance, when aligning RNA-seq data a parameter such as max intron size is set, establishing an assumption about gene length genome wide which likely does hold true in all instances (Arnold et al., 2013). This provides a challenging paradigm where assumptions must be made about the genome, but genes that violate these parameters still need to be caught.

We used chromatin modification data in the form of Chromatin ImmunoPrecipitation with sequencing (ChIP-seq) to provide additional data which can aid in more accurate gene model annotation. Chromatin modifications associated with transcriptional start sites (H3K4me3, and H3K56ac), and an additional set of chromatin modifications known to co-occur across the gene body (H3K36me3, H3K4me1), were used to accurately assay the genome for well defined genes. Using these histone modifications we found a plethora of poorly annotated gene models in the *Zea mays* genome, including a series of truncated genes which needed to be expanded by up to 10,000 kb. Additionally, we found sets of novel transcripts which have previously been

undescribed. All new annotations were validated with multiple methods, both by reassembling RNA-seq reads allowing for greater intron size, as well as by utilizing more recent isoform sequencing (ISO-seq) datasets in *Z. mays* (Stelpflug et al., 2016). This work demonstrates the power of using chromatin modification data to assay the genome accurately for transcription, enabling genome annotations to be done with fewer *a priori* hypotheses attached.

Additionally, we extended these methods to other plant genomes and found many similar annotation issues across *Plantae* more generally, highlighting the challenges in generating high quality resources for the plant genomics community. The successful implementation of this method to additional plant genomes demonstrated its robustness and applicability. This chapter demonstrates the power of incorporating epigenomic data to facilitate more accurate genome annotations going forward. High quality genome annotations are critical in modern biology. They facilitate all levels of genetic inquiry, so generating accurate annotations is of the utmost importance.

The second chapter of this dissertation investigated the *cis*-regualtory basis of C_4 photosynthesis at single-cell resolution. Within this chapter we annotated single-cell Assay for Transposase Accessible Chromatin (ATAC-seq) datasets from five diverse grass species which encapsulate three different types of C_4 photosynthesis (NADP-ME, NAD-ME, and PEPCK), as well as a C_3 outgroup. With these datasets, we explored the cell-type-specific accessibility bias of key C_4 enzymes. We analyzed enzymes which make the C_4 subtypes unique, as well as those enzymes which are common to C_4 metabolism more generally. We then proceeded to investigate the regulatory sequences surrounding these loci, generating high resolution maps of cell-type-specific accessible

chromatin regions (ACRs). Finally, using comparative genomics approaches, we compared these regulatory regions with each other in an attempt to understand the evolutionary basis of cell-type-specific regulatory regions potentially critical in expression of key C_4 genes.

One striking result from this analysis was the high number of cell-type-specific ACRs which had no sequence homology to neighboring species, likely indicating that cell-type-specific ACRs around key C_4 loci evolve rapidly. However, there do appear to be some interesting exceptions to this, as we identified bundle sheath specific ACRs directly upstream of NADP-ME conserved across all five species. This conservation might suggest that certain enzymes are more likely to be integrated into a C4-like photosynthetic process due to their regulatory contexts. However, it's important to note that this does not ensure these enzymes or genes will consistently be incorporated into the C_4 pathway.

After our comprehensive analysis of C_4 photosynthesis loci, we focused on one specific gene family in order to learn the nuanced ways in which cell-type-specificity can change across evolution. To this end we focused on the *DIT* gene family, known for its role in malate transport and crucial for C_4 photosynthesis in *Z. mays* and *S. bicolor*. Notably, both species exhibit cell-type-specific expression of DIT genes: each has a *DIT* gene variant expressed uniquely in either the bundle sheath or mesophyll cells. Our phylogenetic examination of the *DIT* gene family revealed that the bundle sheath-specific *DIT* in *Z. mays* is a recent duplication, situated in a different branch of the *DIT* gene family tree compared to its counterpart in *S. bicolor*, where the mesophyll and bundle sheath-specific *DIT* genes diverge into separate clades. This pattern suggests a significant

shift in cell-type specificity for the *ZmDIT2* locus which occurred during the recent divergence of *Z. mays* and *S. bicolor*, around 13 million years ago (MYA), illustrating the rapid evolution of cell-type specificity between closely related species. Further analysis aimed to connect this change in specificity to variations in cell-type-specific transcription factor (TF) binding motif presence within the regulatory sequences of the *DIT* gene family. Remarkably, we discovered a prevalence of cell-type-specific TF motifs within novel, non-conserved regulatory regions. This was particularly evident in the bundle sheath-specific version of *ZmDIT2* in *Z. mays*, where nearly all bundle sheath-specific TF motifs were located in a new regulatory area just upstream of the transcription start site. This suggests that the cell-type specificity of the *ZmDIT2* locus may have been altered through the adoption of cell-type-specific TF motifs. These insights underscore the dynamic nature of regulatory sequence evolution at the level of individual loci. Additionally, they demonstrate how, even among closely related species, changes to regulatory sequences can occur swiftly, with significant implications for gene expression and function.

Together, these results point to a somewhat non-satisfying resolution: that the evolution into a C_4 type photosynthesizer is not based solely off of co-option of existing regulatory sequences or evolving novel sequences. Rather, these analyses point to a mixture of both models, with ACRs evolving new sequences, and the neo-functionalization of existing sequences.

In addition to providing novel discoveries into the potential regulation of C_4 loci, and generating a valuable resource for the C_4 community overall, our observations in this study also open up new experimental avenues. First off, the importance of these cell-

type-specific ACRs needs to be validated using mutagenesis. Deletions of these ACRs and subsequent measurement of the expression of the most closely related C_4 gene would provide invaluable evidence that these ACRs are 1) critical for the proper expression of these enzymes and 2) likely play an important role in evolution C_4 photosynthesis regulation broadly. However, to date, editing of regulatory elements is a tedious and time consuming process (Liu et al., 2021; Rodríguez-Leal et al., 2017). Additionally, when considering the recalcitrant nature of monocots to transformation and the fact that two of the species sampled (*U. fusca* and *P. miliaceum*) have never been transformed, this stands as a daunting task (Lowe et al., 2016). An additional method which could be valuable in testing the function of putative C_4 regulatory regions would be a *β-glucuronidase* (*GUS*) reporter assay (Jefferson et al., 1987; McCabe et al., 1988). In brief, these assays work by fusing a regulatory region upstream of a minimal promoter element and the galactosidase gene, then transforming these constructs into a plant (for C_4 research this is usually the C_3 crop *Oryza sativa)*. These transgenic plants are then histologically stained to identify the spatiotemporal locations of *GUS* expression, providing evidence that the candidate regulatory region drives gene expression in a specific cellular type (bundle sheath or mesophyll). Previous work has shown this to be an effective method to identify C_4 regulatory elements important in driving cell-type-specific gene expression (Gowik et al., 2004). However, while a valuable orthogonal approach to knockouts, these assays come with their own suite of challenges. Namely, the labor associated with plant transformation is significant and it takes months for transformed plants to develop to the point where analysis is possible. Additionally, many regulatory sequences operate in tandem with each other (Avsec et al., 2019; Hendelman et al., 2021). The *GUS* assay, by isolating

specific regulatory regions, may disrupt the native genomic context of these loci, altering their natural regulatory dynamics. This limitation of GUS reporters in replicating the true *in vivo* regulation highlights the need for multiple avenues of functional genetic experimentation to thoroughly investigate the importance of regulatory sequences.

Performing similar analyses with additional species would greatly enhance our understanding of the rate at which regulatory evolution occurs at C_4 loci. The sampled species comprise a split of over 60 million years from the C_3 outgroup *O. sativa* (Kh et al., 1989). When considering the speed at which regulatory evolution occurs in plant genomes, using distantly related species greatly reduces our resolution in identifying conserved regulatory sequences (Lu et al., 2019; Maher et al., 2018). Adding additional C_3 species, such as the intermediate *Dichanthelium oligosanthes*, to our analysis would enable a deeper understanding about the changes which are required to transition from C_3 to C_4 (Studer et al., 2016). For instance, the decreased evolutionary distance of *D. oligosanthes* to *U. fusca* and *P. miliaceum* could aid in identifying the emergence of critical regulatory loci.

In the final chapter we investigate the rate of cell-type-specific ACR change genome-wide using the single-cell-ACRs identified in the previous study in tandem with a recent *O. sativa* ACR atlas. In brief, using syntenic gene pairs identified between *O. sativa* and all other sampled species, we compare the conservation of ACRs over evolutionary time (57 MYA).

Strikingly, we found that many TF families are enriched in the same cell types across species, indicating a deep evolutionary conservation of regulatory networks, and suggesting that these regulatory networks are likely crucial for proper cell-type-specific

functioning. To date this cell-type specific analysis is the first of its kind in plants and identifies highly conserved and important regulatory networks. We go on to classify and compare the genomic distribution of broadly accessible versus cell-type-specific ACRs and do not observe a large bias in genomic location or distance to the closest gene. We further explored the relative conservation of cell-type-specific ACRs and found that epidermal cells have far fewer conserved cell-type-specific ACRs as compared to all other cell-types in all sampled species. This intriguing finding suggests that significant amounts of turnover associated with the regulation of epidermal cells has occurred over evolutionary time. This is likely due to the wide range of biological pressures occurring to the L1 layer of tissues in plants. For instance, the epidermis has to combat both abiotic stresses in the form of radiation and weather, as well as biotic stresses like pathogens and herbivores (Glover, 2000).

When overlapping our conserved set of ACRs with a previously published set of conserved non-coding sequences (CNS) we were surprised by the fact that cell-type-specific ACRs were enriched for conserved non-coding sequences as compared to broad ACRs (Hendelman et al., 2021). This likely indicates that cell-type-specific regulatory sequences may be critical in the proper development and function of specific cell-types. Additionally, this observation may be indicative that it is harder to evolve cell-type-specific regulatory loci than it is broadly accessible regulatory loci.

We extended this analysis to investigate instances where a CNS found in a *Z. mays* leaf ACR was present in *O. sativa*, but was not in an accessible region. We then leveraged the *O. sativa* atlas to see whether CNSs in *O. sativa* demonstrate altered tissue accessibility patterns. Interestingly, we did find limited instances where these

accessibility patterns had shifted, for instance where a CNS is in an ACR that in *Z. mays* is specifically accessible in leaf but in *O. sativa* is broadly accessible in the cells in root. This accessibility shift likely indicates that the *O. sativa* version of these ACRs lost tissue specificity rather than the *Z. mays* ACR gaining novel specificity. However, this should be further investigated utilizing a previously published atlas of maize accessibility utilizing scATAC-seq (Marand et al., 2021).

While this analysis identified some interesting patterns about the retention and loss of ACRs, further work needs to be done to more deeply understand the evolution of regulatory sequences in plants. In brief, like the C_4 work above, adding additional species to this analysis which provide a more continuous sample of divergence times would further aid in our understanding of the ways in which ACRs change over evolutionary time. Due to the rapid rate of synteny breakdown in plants, sampling multiple species in the same genus may be more valuable than species which are highly divergent (Zhao & Schranz, 2019). Interestingly, recent papers focused on just the genus *Oryza* demonstrated rapid and widespread changes in both leaf anatomy and physiological traits (Chatterjee et al., 2016). This manuscript highlights that while the genetic composition of plant genomes are rapidly re-arranging, their gross anatomy is as well. The variation in both genomic content as well as larger anatomical features provides an exciting opportunity where the sheer diversity of plant anatomy could be utilized in tandem with genomics to see how much changes in gross anatomy seem to be correlated with changes in regulatory sequences.

While the analysis conducted here provides exciting findings about the importance of CNSs in plant genomes, they come with a few caveats. Firstly, the CNS

dataset we utilized lacked two of the species used in this analysis, limiting the comparisons we were able to make (Hendelman et al., 2021). This lack of species is almost certainly occluding additional patterns about loss and retention of CNSs over evolutionary time. Additionally, in the generation of the CNS dataset, more species which were closely related towards the *Z. mays* lineage were used, potentially biasing CNS identification.

The resolution of the cell-type annotation in this study also requires discussion. A significant limitation of this study is that we were only able to achieve a coarse resolution of cell-type annotation across the non-rice species due to lack of genetic markers. It has been well established that even in short evolutionary time scales, marker genes used to distinguish key cell-types in one species can have altered expression patterns in another (Hughes & Langdale, 2022; Lucas et al., 2013). Therefore some marker genes, like the handful of markers used to identify abaxial and adaxial orientation of cell-types, are unreliable indicators in an unverified cross species context (Emery et al., 2003; Jiajia et al., 2020). While resolution of the annotations provided in these studies is still highly valuable, these caveats may mean that certain nuances are lost in our observations. Finally, conducting functional molecular genetics tests on conserved ACRs across deep evolutionary periods could significantly underscore their importance. However, challenges persist in creating genetic deletions within this group of monocots. Moreover, it is important to recognize that manipulating cell-type-specific ACRs may necessitate detailed phenotyping of specific plant cell types. For example, to assess the impact of edits on a companion cell specific ACR, it would be essential to precisely measure companion cell function across all studied species (Hunt et al., 2023; Ivashikina et al.,

2003). This would require the generation of additional single-cell libraries for mutants or involve complex histological measurements, both of which are costly and technically demanding (Ivashikina et al., 2003; Shahan et al., 2020). Despite these challenges, functional assays offer invaluable insights and could add to our understanding of these key regulatory loci.

In total this dissertation illustrates the numerous ways plant epigenomic data can be used. Both to facilitate resource curation, as well as to guide discovery. Techniques like single-cell ATAC-seq are revolutionizing our understanding of plant gene regulation by uncovering regulatory regions that were previously obscured in bulk tissue assays. To deepen our comprehension of these regions, the integration of additional data sources is crucial. This work underscores the significance of comparative genomics in shedding light on these regulatory loci. By pinpointing sequences conserved across species over extensive evolutionary periods, we've identified loci that are likely essential for specific plant cell types, opening up new paths for research. Furthermore, this dissertation provides resources that future computational biologists could use to apply machine learning techniques, aiming to understand how plant cell types interpret regulatory loci across different species. This approach, inspired by studies in mammalian systems, could lead to the creation of novel cell-type-specific regulatory sequences for use in synthetic biology, showcasing the potential of cross-species analyses of regulatory regions. Understanding the evolution of gene regulation in plants is challenging. Due to the lack of resources, as well as unique evolutionary features associated with plant genomes, novel analysis methods need to be tried and adapted. This dissertation starts addressing

this need, making one small step for plant-genomicist but one giant leap for plant-genomicistkind.

References

Alberts, B., Johnson, A., Lewis, J., Raff, M., Roberts, K., & Walter, P. (2002). The Plant Cell Wall. *Molecular Biology of the Cell. 4th Edition*. https://www.ncbi.nlm.nih.gov/books/NBK26928/

Amanda, D., Doblin, M. S., Galletti, R., Bacic, A., Ingram, G. C., & Johnson, K. L. (2016). DEFECTIVE KERNEL1 (DEK1) Regulates Cell Walls in the Leaf Epidermis. *Plant Physiology*, *172*(4), 2204–2218. https://doi.org/10.1104/pp.16.01401

Anatomie des plantes. (n.d.). Retrieved December 19, 2022, from https://bibdigital.rjb.csic.es/records/item/13576-redirection

Andersson, R., Sandelin, A., & Danko, C. G. (2015). A unified architecture of transcriptional regulatory elements. *Trends in Genetics*, *31*(8), Article 8. https://doi.org/10.1016/j.tig.2015.05.007

Azodi, C. B., Lloyd, J. P., & Shiu, S.-H. (2020). The cis-regulatory codes of response to combined heat and drought stress in Arabidopsis thaliana. *NAR Genomics and Bioinformatics*, *2*(3), lqaa049. https://doi.org/10.1093/nargab/lqaa049

Babarinde, I. A., & Saitou, N. (2016). Genomic Locations of Conserved Noncoding Sequences and Their Proximal Protein-Coding Genes in Mammalian Expression Dynamics. *Molecular Biology and Evolution*, *33*(7), 1807–1817. https://doi.org/10.1093/molbev/msw058

Bai, X., Huang, Y., Hu, Y., Liu, H., Zhang, B., Smaczniak, C., Hu, G., Han, Z., & Xing, Y. (2017). Duplication of an upstream silencer of FZP increases grain yield in rice. *Nature Plants*, *3*(11), Article 11. https://doi.org/10.1038/s41477-017-0042-4

Bailey, T. L., Johnson, J., Grant, C. E., & Noble, W. S. (2015). The MEME Suite. *Nucleic Acids Research*, *43*(W1), Article W1. https://doi.org/10.1093/nar/gkv416

Bajic, M., Maher, K. A., & Deal, R. B. (2018). Identification of Open Chromatin Regions in Plant Genomes Using ATAC-Seq. In M. Bemer & C. Baroux (Eds.), *Plant Chromatin Dynamics* (Vol. 1675, pp. 183–201). Springer New York. https://doi.org/10.1007/978-1-4939-7318-7_12

Bansal, K. C., Viret, J. F., Haley, J., Khan, B. M., Schantz, R., & Bogorad, L. (1992). Transient expression from cab-m1 and rbcS-m3 promoter sequences is different in mesophyll and bundle sheath cells in maize leaves. *Proceedings of the National Academy of Sciences*, *89*(8), 3654–3658. https://doi.org/10.1073/pnas.89.8.3654

Becraft, P. W., & Freeling, M. (1991). Sectors of liguleless-1 tissue interrupt an inductive signal during maize leaf development. *The Plant Cell*, *3*(8), 801–807. https://doi.org/10.1105/tpc.3.8.801

Benfey, P. N., Linstead, P. J., Roberts, K., Schiefelbein, J. W., Hauser, M.-T., & Aeschbacher, R. A. (1993). Root development in *Arabidopsis*: Four mutants with dramatically altered root morphogenesis. *Development*, *119*(1), 57–70. https://doi.org/10.1242/dev.119.1.57

Bezrutczyk, M., Hartwig, T., Horschman, M., Char, S. N., Yang, J., Yang, B., Frommer, W. B., & Sosso, D. (2018). Impaired phloem loading in zmsweet13a,b,c sucrose transporter triple knock-out mutants in Zea mays. *The New Phytologist*, *218*(2), 594–603. https://doi.org/10.1111/nph.15021

Bezrutczyk, M., Zöllner, N. R., Kruse, C. P. S., Hartwig, T., Lautwein, T., Köhrer, K., Frommer, W. B., & Kim, J.-Y. (2021). Evidence for phloem loading via the abaxial bundle sheath cells in maize leaves. *The Plant Cell*, *33*(3), Article 3. https://doi.org/10.1093/plcell/koaa055

Birchler, J. A., & Yang, H. (2022). The multiple fates of gene duplications: Deletion, hypofunctionalization, subfunctionalization, neofunctionalization, dosage balance constraints, and neutral variation. *The Plant Cell*, *34*(7), 2466–2474. https://doi.org/10.1093/plcell/koac076

Birnbaum, K., Jung, J. W., Wang, J. Y., Lambert, G. M., Hirst, J. A., Galbraith, D. W., & Benfey, P. N. (2005). Cell type–specific expression profiling in plants via cell sorting of protoplasts from fluorescent reporter lines. *Nature Methods*, *2*(8), 615–619. https://doi.org/10.1038/nmeth0805-615

Birnbaum, K., Shasha, D. E., Wang, J. Y., Jung, J. W., Lambert, G. M., Galbraith, D. W., & Benfey, P. N. (2003). A Gene Expression Map of the Arabidopsis Root. *Science*, *302*(5652), 1956–1960. https://doi.org/10.1126/science.1090022

Blondel, V. D., Guillaume, J.-L., Lambiotte, R., & Lefebvre, E. (2008a). Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, *2008*(10), P10008. https://doi.org/10.1088/1742-5468/2008/10/P10008

Blondel, V. D., Guillaume, J.-L., Lambiotte, R., & Lefebvre, E. (2008b). Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, *2008*(10), P10008. https://doi.org/10.1088/1742-5468/2008/10/P10008

Borba, A. R., Reyna-Llorens, I., Dickinson, P. J., Steed, G., Gouveia, P., Górska, A. M., Gomes, C., Kromdijk, J., Webb, A. A. R., Saibo, N. J. M., & Hibberd, J. M. (2023). Compartmentation of photosynthesis gene expression in C4 maize depends on time of day. *Plant Physiology*, kiad447. https://doi.org/10.1093/plphys/kiad447

Bowes, G., Ogren, W. L., & Hageman, R. H. (1971). Phosphoglycolate production catalyzed by ribulose diphosphate carboxylase. *Biochemical and Biophysical Research Communications*, *45*(3), 716–722. https://doi.org/10.1016/0006-291x(71)90475-x

Boyer, L. A., Plath, K., Zeitlinger, J., Brambrink, T., Medeiros, L. A., Lee, T. I., Levine, S. S., Wernig, M., Tajonar, A., Ray, M. K., Bell, G. W., Otte, A. P., Vidal, M., Gifford, D. K., Young, R. A., & Jaenisch, R. (2006). Polycomb complexes repress developmental regulators in murine embryonic stem cells. *Nature*, *441*(7091), Article 7091. https://doi.org/10.1038/nature04733

Boyle, A. P., Davis, S., Shulha, H. P., Meltzer, P., Margulies, E. H., Weng, Z., Furey, T. S., & Crawford, G. E. (2008). High-Resolution Mapping and Characterization of Open Chromatin across the Genome. *Cell*, *132*(2), 311–322. https://doi.org/10.1016/j.cell.2007.12.014

Brady, S. M., Sarkar, S. F., Bonetta, D., & McCourt, P. (2003). The ABSCISIC ACID INSENSITIVE 3 (ABI3) gene is modulated by farnesylation and is involved in auxin signaling and lateral root development in Arabidopsis. *The Plant Journal*, *34*(1), 67–75. https://doi.org/10.1046/j.1365-313X.2003.01707.x

Brockington, S. F., Alvarez-Fernandez, R., Landis, J. B., Alcorn, K., Walker, R. H., Thomas, M. M., Hileman, L. C., & Glover, B. J. (2013). Evolutionary Analysis of the MIXTA Gene Family Highlights Potential Targets for the Study of Cellular Differentiation. *Molecular Biology and Evolution*, *30*(3), 526–540. https://doi.org/10.1093/molbev/mss260

Buenrostro, J. D., Wu, B., Chang, H. Y., & Greenleaf, W. J. (2015). ATAC-seq: A Method for Assaying Chromatin Accessibility Genome-Wide: ATAC-seq for Assaying Chromatin

Accessibility. In F. M. Ausubel, R. Brent, R. E. Kingston, D. D. Moore, J. G. Seidman, J. A. Smith, & K. Struhl (Eds.), *Current Protocols in Molecular Biology* (p. 21.29.1-21.29.9). John Wiley & Sons, Inc. https://doi.org/10.1002/0471142727.mb2129s109

Buenrostro, J. D., Wu, B., Litzenburger, U. M., Ruff, D., Gonzales, M. L., Snyder, M. P., Chang, H. Y., & Greenleaf, W. J. (2015). Single-cell chromatin accessibility reveals principles of regulatory variation. *Nature*, *523*(7561), Article 7561. https://doi.org/10.1038/nature14590

Burch-Smith, T. M., & Zambryski, P. C. (2012). Plasmodesmata Paradigm Shift: Regulation from Without Versus Within. *Annual Review of Plant Biology*, *63*(1), 239–260. https://doi.org/10.1146/annurev-arplant-042811-105453

Burg, K. R. L., Lewis, J. J., Brack, B. J., Fandino, R. A., Mazo-Vargas, A., & Reed, R. D. (2020). Genomic architecture of a genetically assimilated seasonal color pattern. *Science*, *370*(6517), 721–725. https://doi.org/10.1126/science.aaz3017

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., & Madden, T. L. (2009). BLAST+: Architecture and applications. *BMC Bioinformatics*, *10*, 421. https://doi.org/10.1186/1471-2105-10-421

Candela, H., Johnston, R., Gerhold, A., Foster, T., & Hake, S. (2008). The milkweed pod1 Gene Encodes a KANADI Protein That Is Required for Abaxial/Adaxial Patterning in Maize Leaves. *The Plant Cell*, *20*(8), 2073–2087. https://doi.org/10.1105/tpc.108.059709

Cao, R., Wang, L., Wang, H., Xia, L., Erdjument-Bromage, H., Tempst, P., Jones, R. S., & Zhang, Y. (2002). Role of histone H3 lysine 27 methylation in Polycomb-group silencing. *Science (New York, N.Y.)*, *298*(5595), 1039–1043. https://doi.org/10.1126/science.1076997

Chari, T., Banerjee, J., & Pachter, L. (2021). *The Specious Art of Single-Cell Genomics* [Preprint]. Genomics. https://doi.org/10.1101/2021.08.25.457696

Chen, H., Lareau, C., Andreani, T., Vinyard, M. E., Garcia, S. P., Clement, K., Andrade-Navarro, M. A., Buenrostro, J. D., & Pinello, L. (2019). Assessment of computational methods for the analysis of single-cell ATAC-seq data. *Genome Biology*, *20*(1), Article 1. https://doi.org/10.1186/s13059-019-1854-5

Chen, L., Wu, Z., & Hou, S. (2020). SPEECHLESS Speaks Loudly in Stomatal Development. *Frontiers in Plant Science*, *11*. https://www.frontiersin.org/articles/10.3389/fpls.2020.00114

Chen, M., Long, X., Chen, M., Hao, F., Kang, J., Wang, N., Wang, Y., Wang, M., Gao, Y., Zhou, M., Duo, L., Zhe, X., He, J., Ren, B., Zhang, Y., Liu, B., Li, J., Zhang, Q., Yan, L., … Gao, F. (2022). Integration of single-cell transcriptome and chromatin accessibility of early gonads development among goats, pigs, macaques, and humans. *Cell Reports*, *41*(5), 111587. https://doi.org/10.1016/j.celrep.2022.111587

Chen, X., Miragaia, R. J., Natarajan, K. N., & Teichmann, S. A. (2018). A rapid and robust method for single cell chromatin accessibility profiling. *Nature Communications*, *9*(1), Article 1. https://doi.org/10.1038/s41467-018-07771-0

Chen, X., Zhang, Z., Liu, D., Zhang, K., Li, A., & Mao, L. (2010). SQUAMOSA promoter-binding protein-like transcription factors: Star players for plant growth and development. *Journal of Integrative Plant Biology*, *52*(11), 946–951. https://doi.org/10.1111/j.1744-7909.2010.00987.x

Chen, Z., Debernardi, J. M., Dubcovsky, J., & Gallavotti, A. (2022). *The combination of morphogenic regulators BABY BOOM and GRF-GIF improves maize transformation efficiency* [Preprint]. Plant Biology. https://doi.org/10.1101/2022.09.02.506370

Chollet, R., Vidal, J., & O'Leary, M. H. (1996). PHOSPHO *ENOL* PYRUVATE CARBOXYLASE: A Ubiquitous, Highly Regulated Enzyme in Plants. *Annual Review of Plant Physiology and Plant Molecular Biology*, *47*(1), 273–298. https://doi.org/10.1146/annurev.arplant.47.1.273

Ciren, D., Zebell, S., & Lippman, Z. B. (2023a). Extreme restructuring of cis -regulatory regions controlling a deeply conserved plant stem cell regulator. *bioRxiv: The Preprint Server for Biology*, 2023.12.20.572550. https://doi.org/10.1101/2023.12.20.572550

Ciren, D., Zebell, S., & Lippman, Z. B. (2023b). *Extreme restructuring of cis-regulatory regions controlling a deeply conserved plant stem cell regulator* (p. 2023.12.20.572550). bioRxiv. https://doi.org/10.1101/2023.12.20.572550

Clark, J. W., & Donoghue, P. C. J. (2018). Whole-Genome Duplication and Plant Macroevolution. *Trends in Plant Science*, *23*(10), 933–945. https://doi.org/10.1016/j.tplants.2018.07.006

Cleves, P. A., Ellis, N. A., Jimenez, M. T., Nunez, S. M., Schluter, D., Kingsley, D. M., & Miller, C. T. (2014). Evolved tooth gain in sticklebacks is associated with a *cis* -regulatory allele of *Bmp6*. *Proceedings of the National Academy of Sciences*, *111*(38), 13912–13917. https://doi.org/10.1073/pnas.1407567111

Cramer, P. (2019). Organization and regulation of gene transcription. *Nature*, *573*(7772), Article 7772. https://doi.org/10.1038/s41586-019-1517-4

Cusanovich, D. A., Daza, R., Adey, A., Pliner, H. A., Christiansen, L., Gunderson, K. L., Steemers, F. J., Trapnell, C., & Shendure, J. (2015). Multiplex single cell profiling of chromatin accessibility by combinatorial cellular indexing. *Science (New York, N.Y.)*, *348*(6237), Article 6237. https://doi.org/10.1126/science.aab1601

Cusanovich, D. A., Hill, A. J., Aghamirzaie, D., Daza, R. M., Pliner, H. A., Berletch, J. B., Filippova, G. N., Huang, X., Christiansen, L., DeWitt, W. S., Lee, C., Regalado, S. G., Read, D. F., Steemers, F. J., Disteche, C. M., Trapnell, C., & Shendure, J. (2018a). A Single-Cell Atlas of In Vivo Mammalian Chromatin Accessibility. *Cell*, *174*(5), Article 5. https://doi.org/10.1016/j.cell.2018.06.052

Cusanovich, D. A., Hill, A. J., Aghamirzaie, D., Daza, R. M., Pliner, H. A., Berletch, J. B., Filippova, G. N., Huang, X., Christiansen, L., DeWitt, W. S., Lee, C., Regalado, S. G., Read, D. F., Steemers, F. J., Disteche, C. M., Trapnell, C., & Shendure, J. (2018b). A Single-Cell Atlas of In Vivo Mammalian Chromatin Accessibility. *Cell*, *174*(5), Article 5. https://doi.org/10.1016/j.cell.2018.06.052

Cusanovich, D. A., Reddington, J. P., Garfield, D. A., Daza, R., Aghamirzaie, D., Marco-Ferreres, R., Pliner, H., Christiansen, L., Qiu, X., Steemers, F. J., Trapnell, C., Shendure, J., & Furlong, E. E. M. (2018). The cis-regulatory dynamics of embryonic development at single cell resolution. *Nature*, *555*(7697), Article 7697. https://doi.org/10.1038/nature25981

Dai, X., Tu, X., Du, B., Dong, P., Sun, S., Wang, X., Sun, J., Li, G., Lu, T., Zhong, S., & Li, P. (2022). Chromatin and regulatory differentiation between bundle sheath and mesophyll cells in maize. *The Plant Journal*, *n/a*(n/a), Article n/a. https://doi.org/10.1111/tpj.15586

Danecek, P., Bonfield, J. K., Liddle, J., Marshall, J., Ohan, V., Pollard, M. O., Whitwham, A., Keane, T., McCarthy, S. A., Davies, R. M., & Li, H. (2021). Twelve years of SAMtools and BCFtools. *GigaScience*, *10*(2), giab008. https://doi.org/10.1093/gigascience/giab008

Das, A. C., Foroutan, A., Qian, B., Hosseini Naghavi, N., Shabani, K., & Shooshtari, P. (2022). Single-Cell Chromatin Accessibility Data Combined with GWAS Improves Detection of Relevant Cell Types in 59 Complex Phenotypes. *International Journal of Molecular Sciences*, *23*(19), 11456. https://doi.org/10.3390/ijms231911456

de Velde, J. V., Heyndrickx, K. S., & Vandepoele, K. (2014). Inference of Transcriptional Networks in Arabidopsis through Conserved Noncoding Sequence Analysis. *The Plant Cell*, *26*(7), 2729–2745. https://doi.org/10.1105/tpc.114.127001

DeBruine, Z. J., Melcher, K., & Triche, T. J. (2021). *Fast and robust non-negative matrix factorization for single-cell experiments* (p. 2021.09.01.458620). bioRxiv. https://doi.org/10.1101/2021.09.01.458620

Denyer, T., Ma, X., Klesen, S., Scacchi, E., Nieselt, K., & Timmermans, M. C. P. (2019). Spatiotemporal Developmental Trajectories in the Arabidopsis Root Revealed Using High-Throughput Single-Cell RNA Sequencing. *Developmental Cell*, *48*(6), 840-852.e5. https://doi.org/10.1016/j.devcel.2019.02.022

Dixon, L. E., & Boden, S. A. (2021a). A modified intron of VRT2 drives glume and grain elongation in wheat. *Molecular Plant*, *14*(9), 1421–1423. https://doi.org/10.1016/j.molp.2021.08.016

Dixon, L. E., & Boden, S. A. (2021b). A modified intron of VRT2 drives glume and grain elongation in wheat. *Molecular Plant*, *14*(9), 1421–1423. https://doi.org/10.1016/j.molp.2021.08.016

Dolan, L., & Poethig, R. S. (1998). Clonal Analysis of Leaf Development in Cotton. *American Journal of Botany*, *85*(3), 315–321. https://doi.org/10.2307/2446322

Domcke, S., Hill, A. J., Daza, R. M., Cao, J., O'Day, D. R., Pliner, H. A., Aldinger, K. A., Pokholok, D., Zhang, F., Milbank, J. H., Zager, M. A., Glass, I. A., Steemers, F. J., Doherty, D., Trapnell, C., Cusanovich, D. A., & Shendure, J. (2020a). A human cell atlas of fetal chromatin accessibility. *Science*, *370*(6518), Article 6518. https://doi.org/10.1126/science.aba7612

Domcke, S., Hill, A. J., Daza, R. M., Cao, J., O'Day, D. R., Pliner, H. A., Aldinger, K. A., Pokholok, D., Zhang, F., Milbank, J. H., Zager, M. A., Glass, I. A., Steemers, F. J., Doherty, D., Trapnell, C., Cusanovich, D. A., & Shendure, J. (2020b). A human cell atlas of fetal chromatin accessibility. *Science*, *370*(6518), Article 6518. https://doi.org/10.1126/science.aba7612

Domcke, S., Hill, A. J., Daza, R. M., Cao, J., O'Day, D. R., Pliner, H. A., Aldinger, K. A., Pokholok, D., Zhang, F., Milbank, J. H., Zager, M. A., Glass, I. A., Steemers, F. J., Doherty, D., Trapnell, C., Cusanovich, D. A., & Shendure, J. (2020c). A human cell atlas of fetal chromatin accessibility. *Science*, *370*(6518), Article 6518. https://doi.org/10.1126/science.aba7612

Dorrity, M. W., Alexandre, C., Hamm, M., Vigil, A.-L., Fields, S., Queitsch, C., & Cuperus, J. (2020). *The regulatory landscape of Arabidopsis thaliana roots at single-cell resolution* [Preprint]. Genomics. https://doi.org/10.1101/2020.07.17.204792

Eagen, K. P., Hartl, T. A., & Kornberg, R. D. (2015). Stable Chromosome Condensation Revealed by Chromosome Conformation Capture. *Cell*, *163*(4), Article 4. https://doi.org/10.1016/j.cell.2015.10.026

Efroni, I. (2018). A Conceptual Framework for Cell Identity Transitions in Plants. *Plant and Cell Physiology*, *59*(4), Article 4. https://doi.org/10.1093/pcp/pcx172

Emery, J. F., Floyd, S. K., Alvarez, J., Eshed, Y., Hawker, N. P., Izhaki, A., Baum, S. F., & Bowman, J. L. (2003). Radial patterning of Arabidopsis shoots by class III HD-ZIP and KANADI genes. *Current Biology: CB*, *13*(20), 1768–1774. https://doi.org/10.1016/j.cub.2003.09.035

Emms, D. M., Covshoff, S., Hibberd, J. M., & Kelly, S. (2016). Independent and Parallel Evolution of New Genes by Gene Duplication in Two Origins of C4 Photosynthesis Provides New Insight into the Mechanism of Phloem Loading in C4 Species. *Molecular Biology and Evolution*, *33*(7), 1796–1806. https://doi.org/10.1093/molbev/msw057

Emms, D. M., & Kelly, S. (2019). OrthoFinder: Phylogenetic orthology inference for comparative genomics. *Genome Biology*, *20*(1), 238. https://doi.org/10.1186/s13059-019-1832-y

Engelhorn, J., Snodgrass, S. J., Kok, A., Seetharam, A. S., Schneider, M., Kiwit, T., Singh, A., Banf, M., Khaipho-Burch, M., Runcie, D. E., Camargo, V. S., Torres-Rodriguez, J. V., Sun, G., Stam, M., Fiorani, F., Schnable, J. C., Bass, H. W., Hufford, M. B., Stich, B., … Hartwig, T. (2023). *Phenotypic variation in maize can be largely explained by genetic variation at transcription factor binding sites* (p. 2023.08.08.551183). bioRxiv. https://doi.org/10.1101/2023.08.08.551183

Engineer, C., Hashimoto-Sugimoto, M., Negi, J., Israelsson-Nordstrom, M., Azoulay-Shemer, T., Rappel, W.-J., Iba, K., & Schroeder, J. (2016). CO2 sensing and CO2 regulation of stomatal conductance: Advances and open questions. *Trends in Plant Science*, *21*(1), 16–30. https://doi.org/10.1016/j.tplants.2015.08.014

Esau, K. (1939). Development and Structure of the Phloem Tissue. *Botanical Review*, *5*(7), 373–432.

Esau, K. (1943). Ontogeny of the vascular bundle in Zea Mays. *Hilgardia*, *15*(3), 325–368.

Esau, K. (1954). Primary Vascular Differentiation in Plants. *Biological Reviews*, *29*(1), 46–86. https://doi.org/10.1111/j.1469-185X.1954.tb01397.x

Evert, R. F., Esau, K., & Esau, K. (2006). *Esau's Plant anatomy: Meristems, cells, and tissues of the plant body: their structure, function, and development* (3rd ed). Wiley-Interscience.

Fang, J., Guo, T., Xie, Z., Chun, Y., Zhao, J., Peng, L., Zafar, S. A., Yuan, S., Xiao, L., & Li, X. (2021). The URL1–ROC5–TPL2 transcriptional repressor complex represses the ACL1 gene to modulate leaf rolling in rice. *Plant Physiology*, *185*(4), 1722–1744. https://doi.org/10.1093/plphys/kiaa121

Farmer, A., Thibivilliers, S., Ryu, K. H., Schiefelbein, J., & Libault, M. (2021). Single-nucleus RNA and ATAC sequencing reveals the impact of chromatin accessibility on gene expression in Arabidopsis roots at the single-cell level. *Molecular Plant*, *14*(3), 372–383. https://doi.org/10.1016/j.molp.2021.01.001

Feng, D., Liang, Z., Wang, Y., Yao, J., Yuan, Z., Hu, G., Qu, R., Xie, S., Li, D., Yang, L., Zhao, X., Ma, Y., Lohmann, J. U., & Gu, X. (2022). Chromatin accessibility illuminates single-cell regulatory dynamics of rice root tips. *BMC Biology*, *20*(1), 274. https://doi.org/10.1186/s12915-022-01473-2

Furumoto, T., Yamaguchi, T., Ohshima-Ichie, Y., Nakamura, M., Tsuchida-Iwata, Y., Shimamura, M., Ohnishi, J., Hata, S., Gowik, U., Westhoff, P., Bräutigam, A., Weber, A. P. M., & Izui, K. (2011). A plastidial sodium-dependent pyruvate transporter. *Nature*, *476*(7361), Article 7361. https://doi.org/10.1038/nature10250

Gao, Z., Shen, W., & Chen, G. (2018). Uncovering C4-like photosynthesis in C3 vascular cells. *Journal of Experimental Botany*, *69*(15), 3531–3540. https://doi.org/10.1093/jxb/ery155

Gibson, G. (2022). Perspectives on rigor and reproducibility in single cell genomics. *PLoS Genetics*, *18*(5), e1010210. https://doi.org/10.1371/journal.pgen.1010210

Giresi, P. G., Kim, J., McDaniell, R. M., Iyer, V. R., & Lieb, J. D. (2007). FAIRE (Formaldehyde-Assisted Isolation of Regulatory Elements) isolates active regulatory elements from human chromatin. *Genome Research*, *17*(6), 877–885. https://doi.org/10.1101/gr.5533506

Gisselbrecht, S. S., Palagi, A., Kurland, J. V., Rogers, J. M., Ozadam, H., Zhan, Y., Dekker, J., & Bulyk, M. L. (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

Glover, B. J. (2000). Differentiation in plant epidermal cells. *Journal of Experimental Botany*, *51*(344), 497–505. https://doi.org/10.1093/jexbot/51.344.497

Goel, M., Campoy, J. A., Krause, K., Baus, L. C., Sahu, A., Sun, H., Walkemeier, B., Marek, M., Beaudry, R., Ruiz, D., Huettel, B., & Schneeberger, K. (2024). *The majority of somatic mutations in fruit trees are layer-specific* (p. 2024.01.04.573414). bioRxiv. https://doi.org/10.1101/2024.01.04.573414

Goodstein, D. M., Shu, S., Howson, R., Neupane, R., Hayes, R. D., Fazo, J., Mitros, T., Dirks, W., Hellsten, U., Putnam, N., & Rokhsar, D. S. (2012). Phytozome: A comparative platform for green plant genomics. *Nucleic Acids Research*, *40*(Database issue), D1178–D1186. https://doi.org/10.1093/nar/gkr944

Gottwald, J. R., Krysan, P. J., Young, J. C., Evert, R. F., & Sussman, M. R. (2000). Genetic evidence for the in planta role of phloem-specific plasma membrane sucrose transporters. *Proceedings of the National Academy of Sciences*, *97*(25), 13979–13984. https://doi.org/10.1073/pnas.250473797

Gowik, U., Burscheidt, J., Akyildiz, M., Schlue, U., Koczor, M., Streubel, M., & Westhoff, P. (2004). Cis-Regulatory Elements for Mesophyll-Specific Gene Expression in the C4 Plant Flaveria trinervia, the Promoter of the C4 Phosphoenolpyruvate Carboxylase Gene[W]. *The Plant Cell*, *16*(5), 1077–1090. https://doi.org/10.1105/tpc.019729

Gowik, U., Schulze, S., Saladié, M., Rolland, V., Tanz, S. K., Westhoff, P., & Ludwig, M. (2017). A MEM1-like motif directs mesophyll cell-specific expression of the gene encoding the C4 carbonic anhydrase in Flaveria. *Journal of Experimental Botany*, *68*(2), 311. https://doi.org/10.1093/jxb/erw475

Gowik, U., & Westhoff, P. (2011). The Path from C3 to C4 Photosynthesis. *Plant Physiology*, *155*(1), 56–63. https://doi.org/10.1104/pp.110.165308

Grant, C. E., & Bailey, T. L. (2021). *XSTREME: Comprehensive motif analysis of biological sequence datasets* (p. 2021.09.02.458722). bioRxiv. https://doi.org/10.1101/2021.09.02.458722

Grass Phylogeny Working Group II. (2012). New grass phylogeny resolves deep evolutionary relationships and discovers C4 origins. *The New Phytologist*, *193*(2), 304–312. https://doi.org/10.1111/j.1469-8137.2011.03972.x

Gupta, S. D., Levey, M., Schulze, S., Karki, S., Emmerling, J., Streubel, M., Gowik, U., Paul Quick, W., & Westhoff, P. (2020). The C4Ppc promoters of many C4 grass species share a

common regulatory mechanism for gene expression in the mesophyll cell. *The Plant Journal*, *101*(1), 204–216. https://doi.org/10.1111/tpj.14532

Hake, S., & Freeling, M. (1986). Analysis of genetic mosaics shows that the extra epidermal cell divisions in Knotted mutant maize plants are induced by adjacent mesophyll cells. *Nature*, *320*(6063), Article 6063. https://doi.org/10.1038/320621a0

Han, J., Wang, P., Wang, Q., Lin, Q., Chen, Z., Yu, G., Miao, C., Dao, Y., Wu, R., Schnable, J., Tang, H., & Wang, K. (2020). Genome-wide Characterization of DNase I-hypersensitive Sites and Cold Response Regulatory Landscapes in Grasses. *The Plant Cell*. https://doi.org/10.1105/tpc.19.00716

Hatch, M. D. (1987). C4 photosynthesis: A unique elend of modified biochemistry, anatomy and ultrastructure. *Biochimica et Biophysica Acta (BBA) - Reviews on Bioenergetics*, *895*(2), 81–106. https://doi.org/10.1016/S0304-4173(87)80009-5

Hay, A., & Tsiantis, M. (2006). The genetic basis for differences in leaf form between Arabidopsis thaliana and its wild relative Cardamine hirsuta. *Nature Genetics*, *38*(8), 942–947. https://doi.org/10.1038/ng1835

He, J.-X., Gendron, J. M., Sun, Y., Gampala, S. S. L., Gendron, N., Sun, C. Q., & Wang, Z.-Y. (2005). BZR1 is a transcriptional repressor with dual roles in brassinosteroid homeostasis and growth responses. *Science (New York, N.Y.)*, *307*(5715), 1634–1638. https://doi.org/10.1126/science.1107580

Helariutta, Y., Fukaki, H., Wysocka-Diller, J., Nakajima, K., Jung, J., Sena, G., Hauser, M.-T., & Benfey, P. N. (2000). The SHORT-ROOT Gene Controls Radial Patterning of the Arabidopsis Root through Radial Signaling. *Cell*, *101*(5), Article 5. https://doi.org/10.1016/S0092-8674(00)80865-X

Hendelman, A., Zebell, S., Rodriguez-Leal, D., Dukler, N., Robitaille, G., Wu, X., Kostyun, J., Tal, L., Wang, P., Bartlett, M. E., Eshed, Y., Efroni, I., & Lippman, Z. B. (2021). Conserved pleiotropy of an ancient plant homeobox gene uncovered by cis-regulatory dissection. *Cell*, *0*(0), Article 0. https://doi.org/10.1016/j.cell.2021.02.001

Herman, P. L., & Marks, M. D. (1989). Trichome Development in Arabidopsis thaliana. II. Isolation and Complementation of the GLABROUS1 Gene. *The Plant Cell*, *1*(11), 1051–1055. https://doi.org/10.1105/tpc.1.11.1051

Hill, M. S., Vande Zande, P., & Wittkopp, P. J. (2020). Molecular and evolutionary processes generating variation in gene expression. *Nature Reviews Genetics*, 1–13. https://doi.org/10.1038/s41576-020-00304-w

Hodge, R. D., Bakken, T. E., Miller, J. A., Smith, K. A., Barkan, E. R., Graybuck, L. T., Close, J. L., Long, B., Johansen, N., Penn, O., Yao, Z., Eggermont, J., Höllt, T., Levi, B. P., Shehata, S. I., Aevermann, B., Beller, A., Bertagnolli, D., Brouner, K., … Lein, E. S. (2019). Conserved cell types with divergent features in human versus mouse cortex. *Nature*, *573*(7772), Article 7772. https://doi.org/10.1038/s41586-019-1506-7

Hofmeister, B. T., & Schmitz, R. J. (2018). Enhanced JBrowse plugins for epigenomics data visualization. *BMC Bioinformatics*, *19*(1), 159. https://doi.org/10.1186/s12859-018-2160-z

Hong, S.-Y., Kim, O.-K., Kim, S.-G., Yang, M.-S., & Park, C.-M. (2011). Nuclear import and DNA binding of the ZHD5 transcription factor is modulated by a competitive peptide inhibitor in Arabidopsis. *The Journal of Biological Chemistry*, *286*(2), 1659–1668. https://doi.org/10.1074/jbc.M110.167692

Hooke, R. (1665). *Micrographia: Or, Some physiological descriptions of minute bodies made by magnifying glasses. With observations and inquiries thereupon*. Science History Institute Digital Collections. https://digital.sciencehistory.org/works/9g54xj51s

Horstman, A., Fukuoka, H., Muino, J. M., Nitsch, L., Guo, C., Passarinho, P., Sanchez-Perez, G., Immink, R., Angenent, G., & Boutilier, K. (2015). AIL and HDG proteins act antagonistically to control cell proliferation. *Development (Cambridge, England)*, *142*(3), 454–464. https://doi.org/10.1242/dev.117168

Huang, T., Harrar, Y., Lin, C., Reinhart, B., Newell, N. R., Talavera-Rauh, F., Hokin, S. A., Barton, M. K., & Kerstetter, R. A. (2014). Arabidopsis KANADI1 Acts as a Transcriptional Repressor by Interacting with a Specific cis-Element and Regulates Auxin Biosynthesis, Transport, and Signaling in Opposition to HD-ZIPIII Factors[W]. *The Plant Cell*, *26*(1), 246–262. https://doi.org/10.1105/tpc.113.111526

Huang, Y., Jiao, Y., Xie, N., Guo, Y., Zhang, F., Xiang, Z., Wang, R., Wang, F., Gao, Q., Tian, L., Li, D., Chen, L., & Liang, M. (2019). OsNCED5, a 9-cis-epoxycarotenoid dioxygenase gene, regulates salt and water stress tolerance and leaf senescence in rice. *Plant Science: An International Journal of Experimental Plant Biology*, *287*, 110188. https://doi.org/10.1016/j.plantsci.2019.110188

Hubisz, M. J., Pollard, K. S., & Siepel, A. (2011). PHAST and RPHAST: Phylogenetic analysis with space/time models. *Briefings in Bioinformatics*, *12*(1), 41–51. https://doi.org/10.1093/bib/bbq072

Hufford, M. B., Seetharam, A. S., Woodhouse, M. R., Chougule, K. M., Ou, S., Liu, J., Ricci, W. A., Guo, T., Olson, A., Qiu, Y., Della Coletta, R., Tittes, S., Hudson, A. I., Marand, A. P., Wei, S., Lu, Z., Wang, B., Tello-Ruiz, M. K., Piri, R. D., … Dawe, R. K. (2021). De

novo assembly, annotation, and comparative analysis of 26 diverse maize genomes. *Science*, *373*(6555), 655–662. https://doi.org/10.1126/science.abg5289

Hughes, T. E., & Langdale, J. A. (2022). SCARECROW is deployed in distinct contexts during rice and maize leaf development. *Development*, *149*(7), dev200410. https://doi.org/10.1242/dev.200410

Hughes, T. E., Langdale, J. A., & Kelly, S. (2014). The impact of widespread regulatory neofunctionalization on homeolog gene evolution following whole-genome duplication in maize. *Genome Research*, *24*(8), 1348–1355. https://doi.org/10.1101/gr.172684.114

Hung, C.-Y., Lin, Y., Zhang, M., Pollock, S., David Marks, M., & Schiefelbein, J. (1998). A Common Position-Dependent Mechanism Controls Cell-Type Patterning and GLABRA2 Regulation in the Root and Hypocotyl Epidermis of Arabidopsis. *Plant Physiology*, *117*(1), 73–84.

Hunt, H., Brueggen, N., Galle, A., Vanderauwera, S., Frohberg, C., Fernie, A. R., Sonnewald, U., & Sweetlove, L. J. (2023). Analysis of companion cell and phloem metabolism using a transcriptome-guided model of Arabidopsis metabolism. *Plant Physiology*, *192*(2), 1359–1377. https://doi.org/10.1093/plphys/kiad154

Imperatorskai͡a akademīi͡a nauk (Russia), (Russia), I., & (Russia), I. (1868). *Mémoires de l'Académie impériale des sciences de St.-Pétersbourg* (Vol. 11, pp. 1–918). L'Académie. https://www.biodiversitylibrary.org/item/175756

Ingram, P., Dettmer, J., Helariutta, Y., & Malamy, J. E. (2011). Arabidopsis Lateral Root Development 3 is essential for early phloem development and function, and hence for normal root system development. *The Plant Journal: For Cell and Molecular Biology*, *68*(3), 455–467. https://doi.org/10.1111/j.1365-313X.2011.04700.x

Itoh, J.-I., Sato, Y., Sato, Y., Hibara, K.-I., Shimizu-Sato, S., Kobayashi, H., Takehisa, H., Sanguinet, K. A., Namiki, N., & Nagamura, Y. (2016). Genome-wide analysis of spatiotemporal gene expression patterns during early embryogenesis in rice. *Development (Cambridge, England)*, *143*(7), 1217–1227. https://doi.org/10.1242/dev.123661

Jaitin, D. A., Kenigsberg, E., Keren-Shaul, H., Elefant, N., Paul, F., Zaretsky, I., Mildner, A., Cohen, N., Jung, S., Tanay, A., & Amit, I. (2014). Massively Parallel Single-Cell RNA-Seq for Marker-Free Decomposition of Tissues into Cell Types. *Science*, *343*(6172), 776–779. https://doi.org/10.1126/science.1247651

Javelle, M., Vernoud, V., Rogowsky, P. M., & Ingram, G. C. (2011). Epidermis: The formation and functions of a fundamental plant tissue. *New Phytologist*, *189*(1), 17–39. https://doi.org/10.1111/j.1469-8137.2010.03514.x

Jefferson, R. A., Kavanagh, T. A., & Bevan, M. W. (1987). GUS fusions: Beta-glucuronidase as a sensitive and versatile gene fusion marker in higher plants. *The EMBO Journal*, *6*(13), 3901–3907. https://doi.org/10.1002/j.1460-2075.1987.tb02730.x

Jegla, D. E., & Sussex, I. M. (1989). Cell lineage patterns in the shoot meristem of the sunflower embryo in the dry seed. *Developmental Biology*, *131*(1), 215–225. https://doi.org/10.1016/S0012-1606(89)80053-3

Jiang, J., Ma, S., Ye, N., Jiang, M., Cao, J., & Zhang, J. (2017). WRKY transcription factors in plant responses to stresses. *Journal of Integrative Plant Biology*, *59*(2), 86–101. https://doi.org/10.1111/jipb.12513

Jin, J., Tian, F., Yang, D.-C., Meng, Y.-Q., Kong, L., Luo, J., & Gao, G. (2017). PlantTFDB 4.0: Toward a central hub for transcription factors and regulatory interactions in plants. *Nucleic Acids Research*, *45*(D1), D1040–D1045. https://doi.org/10.1093/nar/gkw982

Johnson, S. M., Tan, F. J., McCullough, H. L., Riordan, D. P., & Fire, A. Z. (2006). Flexibility and constraint in the nucleosome core landscape of Caenorhabditis elegans chromatin. *Genome Research*, *16*(12), 1505–1516. https://doi.org/10.1101/gr.5560806

Johnston, R., Candela, H., Hake, S., & Foster, T. (2010). The maize milkweed pod1 mutant reveals a mechanism to modify organ morphology. *Genesis*, *48*(7), 416–423. https://doi.org/10.1002/dvg.20622

Jump, A. S., Marchant, R., & Peñuelas, J. (2009). Environmental change and the option value of genetic diversity. *Trends in Plant Science*, *14*(1), 51–58. https://doi.org/10.1016/j.tplants.2008.10.002

Kadioglu, A., Terzi, R., Saruhan, N., & Saglam, A. (2012). Current advances in the investigation of leaf rolling caused by biotic and abiotic stress factors. *Plant Science*, *182*, 42–48. https://doi.org/10.1016/j.plantsci.2011.01.013

Kajala, K., Brown, N. J., Williams, B. P., Borrill, P., Taylor, L. E., & Hibberd, J. M. (2012). Multiple Arabidopsis genes primed for recruitment into C4 photosynthesis. *The Plant Journal*, *69*(1), 47–56. https://doi.org/10.1111/j.1365-313X.2011.04769.x

Kajala, K., Gouran, M., Shaar-Moshe, L., Mason, G. A., Rodriguez-Medina, J., Kawa, D., Pauluzzi, G., Reynoso, M., Canto-Pastor, A., Manzano, C., Lau, V., Artur, M. A. S., West, D. A., Gray, S. B., Borowsky, A. T., Moore, B. P., Yao, A. I., Morimoto, K. W., Bajic, M., … Brady, S. M. (2021). Innovation, conservation, and repurposing of gene function in root cell type development. *Cell*, *184*(12), 3333-3348.e19. https://doi.org/10.1016/j.cell.2021.04.024

Kaplan, D. R., & Hagemann, W. (1991). The Relationship of Cell and Organism in Vascular Plants. *BioScience*, *41*(10), 693–703. https://doi.org/10.2307/1311764

Kh, W., M, G., Yw, Y., Pm, S., & Wh, L. (1989). Date of the monocot-dicot divergence estimated from chloroplast DNA sequence data. *Proceedings of the National Academy of Sciences of the United States of America*, *86*(16). https://doi.org/10.1073/pnas.86.16.6201

Kim, D.-H., & Sung, S. (2014). Polycomb-Mediated Gene Silencing in Arabidopsis thaliana. *Molecules and Cells*, *37*(12), 841–850. https://doi.org/10.14348/molcells.2014.0249

Kim, E.-D., Dorrity, M. W., Fitzgerald, B. A., Seo, H., Sepuru, K. M., Queitsch, C., Mitsuda, N., Han, S.-K., & Torii, K. U. (2022). Dynamic chromatin accessibility deploys heterotypic cis/trans-acting factors driving stomatal cell-fate commitment. *Nature Plants*, *8*(12), Article 12. https://doi.org/10.1038/s41477-022-01304-w

Kirik, V., Schnittger, A., Radchuk, V., Adler, K., Hülskamp, M., & Bäumlein, H. (2001). Ectopic Expression of the Arabidopsis AtMYB23 Gene Induces Differentiation of Trichome Cells. *Developmental Biology*, *235*(2), 366–377. https://doi.org/10.1006/dbio.2001.0287

Korsunsky, I., Millard, N., Fan, J., Slowikowski, K., Zhang, F., Wei, K., Baglaenko, Y., Brenner, M., Loh, P., & Raychaudhuri, S. (2019). Fast, sensitive and accurate integration of single-cell data with Harmony. *Nature Methods*, *16*(12), 1289–1296. https://doi.org/10.1038/s41592-019-0619-0

Korsunsky, I., Nathan, A., Millard, N., & Raychaudhuri, S. (2019). *Presto scales Wilcoxon and auROC analyses to millions of observations* [Preprint]. Bioinformatics. https://doi.org/10.1101/653253

Kubo, H., Kishi, M., & Goto, K. (2008). Expression analysis of *ANTHOCYANINLESS2* gene in Arabidopsis. *Plant Science*, *175*(6), 853–857. https://doi.org/10.1016/j.plantsci.2008.08.006

Kubo, H., Peeters, A. J., Aarts, M. G., Pereira, A., & Koornneef, M. (1999). ANTHOCYANINLESS2, a homeobox gene affecting anthocyanin distribution and root development in Arabidopsis. *The Plant Cell*, *11*(7), 1217–1226.

Lai, X., Stigliani, A., Vachon, G., Carles, C., Smaczniak, C., Zubieta, C., Kaufmann, K., & Parcy, F. (2019). Building Transcription Factor Binding Site Models to Understand Gene Regulation in Plants. *Molecular Plant*, *12*(6), Article 6. https://doi.org/10.1016/j.molp.2018.10.010

Langdale, J. A., Lane, B., Freeling, M., & Nelson, T. (1989). Cell lineage analysis of maize bundle sheath and mesophyll cells. *Developmental Biology*, *133*(1), Article 1. https://doi.org/10.1016/0012-1606(89)90304-7

Langmead, B., & Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nature Methods*, *9*(4), Article 4. https://doi.org/10.1038/nmeth.1923

Larkin, J. C., Oppenheimer, D. G., Lloyd, A. M., Paparozzi, E. T., & Marks, M. D. (1994). Roles of the GLABROUS1 and TRANSPARENT TESTA GLABRA Genes in Arabidopsis Trichome Development. *The Plant Cell*, *6*(8), 1065–1076. https://doi.org/10.1105/tpc.6.8.1065

Le Hir, R., & Bellini, C. (2013). The Plant-Specific Dof Transcription Factors Family: New Players Involved in Vascular System Development and Functioning in Arabidopsis. *Frontiers in Plant Science*, *4*. https://www.frontiersin.org/journals/plant-science/articles/10.3389/fpls.2013.00164

Lee, S.-I., & Kim, N.-S. (2014). Transposable Elements and Genome Size Variations in Plants. *Genomics & Informatics*, *12*(3), 87–97. https://doi.org/10.5808/GI.2014.12.3.87

Lee, T. A., Nobori, T., Illouz-Eliaz, N., Xu, J., Jow, B., Nery, J. R., & Ecker, J. R. (2023). *A Single-Nucleus Atlas of Seed-to-Seed Development in Arabidopsis* [Preprint]. Plant Biology. https://doi.org/10.1101/2023.03.23.533992

Leroux, O. (2012). Collenchyma: A versatile mechanical tissue with dynamic cell walls. *Annals of Botany*, *110*(6), 1083–1098. https://doi.org/10.1093/aob/mcs186

Li, H., & Durbin, R. (2009). Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics*, *25*(14), 1754–1760. https://doi.org/10.1093/bioinformatics/btp324

Li, K., Debernardi, J. M., Li, C., Lin, H., Zhang, C., Jernstedt, J., Korff, M. von, Zhong, J., & Dubcovsky, J. (2021). Interactions between SQUAMOSA and SHORT VEGETATIVE PHASE MADS-box proteins regulate meristem transitions during wheat spike development. *The Plant Cell*, *33*(12), 3621–3644. https://doi.org/10.1093/plcell/koab243

Li, P., Ponnala, L., Gandotra, N., Wang, L., Si, Y., Tausta, S. L., Kebrom, T. H., Provart, N., Patel, R., Myers, C. R., Reidel, E. J., Turgeon, R., Liu, P., Sun, Q., Nelson, T., & Brutnell, T. P. (2010). The developmental dynamics of the maize leaf transcriptome. *Nature Genetics*, *42*(12), Article 12. https://doi.org/10.1038/ng.703

Li, Y. E., Preissl, S., Miller, M., Johnson, N. D., Wang, Z., Jiao, H., Zhu, C., Wang, Z., Xie, Y., Poirion, O., Kern, C., Pinto-Duarte, A., Tian, W., Siletti, K., Emerson, N., Osteen, J., Lucero, J., Lin, L., Yang, Q., … Ren, B. (2023). A comparative atlas of single-cell chromatin accessibility in the human brain. *Science (New York, N.Y.)*, *382*(6667), eadf7044. https://doi.org/10.1126/science.adf7044

Li, Y., Yang, Z., Zhang, Y., Guo, J., Liu, L., Wang, C., Wang, B., & Han, G. (2022). The roles of HD-ZIP proteins in plant abiotic stress tolerance. *Frontiers in Plant Science*, *13*, 1027071. https://doi.org/10.3389/fpls.2022.1027071

Lieberman-Aiden, E., Berkum, N. L., Williams, L., Imakaev, M., Ragoczy, T., Telling, A., Amit, I., Lajoie, B. R., Sabo, P. J., Dorschner, M. O., Sandstrom, R., Bernstein, B., Bender, M. A., Groudine, M., Gnirke, A., Stamatoyannopoulos, J., Mirny, L. A., Lander, E. S., & Dekker, J. (2009). Comprehensive Mapping of Long-Range Interactions Reveals Folding Principles of the Human Genome. *Science*, *326*(5950), Article 5950. https://doi.org/10.1126/science.1181369

Liu, J., Chen, Z., Wang, Z., Zhang, Z., Xie, X., Wang, Z., Chai, L., Song, L., Cheng, X., Feng, M., Wang, X., Liu, Y., Hu, Z., Xing, J., Su, Z., Peng, H., Xin, M., Yao, Y., Guo, W., … Ni, Z. (2021). Ectopic expression of VRT-A2 underlies the origin of Triticum polonicum and Triticum petropavlovskyi with long outer glumes and grains. *Molecular Plant*, *14*(9), 1472–1488. https://doi.org/10.1016/j.molp.2021.05.021

Liu, J., Seetharam, A. S., Chougule, K., Ou, S., Swentowsky, K. W., Gent, J. I., Llaca, V., Woodhouse, M. R., Manchanda, N., Presting, G. G., Kudrna, D. A., Alabady, M., Hirsch, C. N., Fengler, K. A., Ware, D., Michael, T. P., Hufford, M. B., & Dawe, R. K. (2020). Gapless assembly of maize chromosomes using long-read technologies. *Genome Biology*, *21*(1), Article 1. https://doi.org/10.1186/s13059-020-02029-9

Liu, L., Gallagher, J., Arevalo, E. D., Chen, R., Skopelitis, T., Wu, Q., Bartlett, M., & Jackson, D. (2021). Enhancing grain-yield-related traits by CRISPR–Cas9 promoter editing of maize CLE genes. *Nature Plants*, *7*(3), Article 3. https://doi.org/10.1038/s41477-021-00858-5

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

Lovell, J. T., Sreedasyam, A., Schranz, M. E., Wilson, M., Carlson, J. W., Harkess, A., Emms, D., Goodstein, D. M., & Schmutz, J. (2022). GENESPACE tracks regions of interest and gene copy number variation across multiple genomes. *eLife*, *11*, e78526. https://doi.org/10.7554/eLife.78526

Lu, Z., Hofmeister, B. T., Vollmers, C., DuBois, R. M., & Schmitz, R. J. (2017). Combining ATAC-seq with nuclei sorting for discovery of cis-regulatory regions in plant genomes. *Nucleic Acids Research*, *45*(6), Article 6. https://doi.org/10.1093/nar/gkw1179

Lu, Z., Marand, A. P., Ricci, W. A., Ethridge, C. L., Zhang, X., & Schmitz, R. J. (2019a). The prevalence, evolution and chromatin signatures of plant regulatory elements. *Nature Plants*, *5*(12), Article 12. https://doi.org/10.1038/s41477-019-0548-z

Lu, Z., Marand, A. P., Ricci, W. A., Ethridge, C. L., Zhang, X., & Schmitz, R. J. (2019b). The prevalence, evolution and chromatin signatures of plant regulatory elements. *Nature Plants*, *5*(12), Article 12. https://doi.org/10.1038/s41477-019-0548-z

Lu, Z., Zhang, M., Lee, J., Sziraki, A., Anderson, S., Zhang, Z., Xu, Z., Jiang, W., Ge, S., Nelson, P. T., Zhou, W., & Cao, J. (2023). Tracking cell-type-specific temporal dynamics in human and mouse brains. *Cell*, *186*(20), 4345-4364.e24. https://doi.org/10.1016/j.cell.2023.08.042

Lv, Z., Zhao, W., Kong, S., Li, L., & Lin, S. (2023). Overview of molecular mechanisms of plant leaf development: A systematic review. *Frontiers in Plant Science*, *14*, 1293424. https://doi.org/10.3389/fpls.2023.1293424

Maaten, L., & Hinton, G. (2008). Visualizing Data using t-SNE. *Journal of Machine Learning Research*, *9*(86), 2579–2605.

Maher, K. A., Bajic, M., Kajala, K., Reynoso, M., Pauluzzi, G., West, D. A., Zumstein, K., Woodhouse, M., Bubb, K., Dorrity, M. W., Queitsch, C., Bailey-Serres, J., Sinha, N., Brady, S. M., & Deal, R. B. (2018a). 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), Article 1. https://doi.org/10.1105/tpc.17.00581

Maher, K. A., Bajic, M., Kajala, K., Reynoso, M., Pauluzzi, G., West, D. A., Zumstein, K., Woodhouse, M., Bubb, K., Dorrity, M. W., Queitsch, C., Bailey-Serres, J., Sinha, N., Brady, S. M., & Deal, R. B. (2018b). 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), Article 1. https://doi.org/10.1105/tpc.17.00581

Maher, K. A., Bajic, M., Kajala, K., Reynoso, M., Pauluzzi, G., West, D. A., Zumstein, K., Woodhouse, M., Bubb, K., Dorrity, M. W., Queitsch, C., Bailey-Serres, J., Sinha, N., Brady, S. M., & Deal, R. B. (2018c). 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, A. P., Chen, Z., Gallavotti, A., & Schmitz, R. J. (2021a). A cis-regulatory atlas in maize at single-cell resolution. *Cell*, *184*(11), Article 11. https://doi.org/10.1016/j.cell.2021.04.014

Marand, A. P., Chen, Z., Gallavotti, A., & Schmitz, R. J. (2021b). A cis-regulatory atlas in maize at single-cell resolution. *Cell*, *184*(11), Article 11. https://doi.org/10.1016/j.cell.2021.04.014

Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.Journal*, *17*(1), Article 1. https://doi.org/10.14806/ej.17.1.200

Matsuoka, M., Kyozuka, J., Shimamoto, K., & Kano-Murakami, Y. (1994). The promoters of two carboxylases in a C4 plant (maize) direct cell-specific, light-regulated expression in a C3 plant (rice). *The Plant Journal*, *6*(3), 311–319. https://doi.org/10.1046/j.1365-313X.1994.06030311.x

Matsuoka, M., Tada, Y., Fujimura, T., & Kano-Murakami, Y. (1993). Tissue-specific light-regulated expression directed by the promoter of a C4 gene, maize pyruvate,orthophosphate dikinase, in a C3 plant, rice. *Proceedings of the National Academy of Sciences*, *90*(20), 9586–9590. https://doi.org/10.1073/pnas.90.20.9586

McCabe, D. E., Swain, W. F., Martinell, B. J., & Christou, P. (1988a). Stable Transformation of Soybean (Glycine Max) by Particle Acceleration. *Bio/Technology*, *6*(8), 923–926. https://doi.org/10.1038/nbt0888-923

McCabe, D. E., Swain, W. F., Martinell, B. J., & Christou, P. (1988b). Stable Transformation of Soybean (Glycine Max) by Particle Acceleration. *Nature Biotechnology*, *6*(8), 923–926. https://doi.org/10.1038/nbt0888-923

McInnes, L., Healy, J., & Melville, J. (2020a). *UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction* (arXiv:1802.03426). arXiv. http://arxiv.org/abs/1802.03426

McInnes, L., Healy, J., & Melville, J. (2020b). *UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction* (arXiv:1802.03426). arXiv. http://arxiv.org/abs/1802.03426

McKenna, K. Z., Wagner, G. P., & Cooper, K. L. (2021). Chapter One—A developmental perspective of homology and evolutionary novelty. In S. F. Gilbert (Ed.), *Current Topics in Developmental Biology* (Vol. 141, pp. 1–38). Academic Press. https://doi.org/10.1016/bs.ctdb.2020.12.001

Meng, F., Zhao, H., Zhu, B., Zhang, T., Yang, M., Li, Y., Han, Y., & Jiang, J. (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

Mifsud, B., Tavares-Cadete, F., Young, A. N., Sugar, R., Schoenfelder, S., Ferreira, L., Wingett, S. W., Andrews, S., Grey, W., Ewels, P. A., Herman, B., Happe, S., Higgs, A., LeProust, E., Follows, G. A., Fraser, P., Luscombe, N. M., & Osborne, C. S. (2015). Mapping long-range promoter contacts in human cells with high-resolution capture Hi-C. *Nature Genetics*, *47*(6), Article 6. https://doi.org/10.1038/ng.3286

Minnoye, L., Marinov, G. K., Krausgruber, T., Pan, L., Marand, A. P., Secchia, S., Greenleaf, W. J., Furlong, E. E. M., Zhao, K., Schmitz, R. J., Bock, C., & Aerts, S. (2021). Chromatin accessibility profiling methods. *Nature Reviews Methods Primers*, *1*(1), Article 1. https://doi.org/10.1038/s43586-020-00008-9

Moore, B. M., Lee, Y. S., Wang, P., Azodi, C., Grotewold, E., & Shiu, S.-H. (2022). Modeling temporal and hormonal regulation of plant transcriptional response to wounding. *The Plant Cell*, *34*(2), 867–888. https://doi.org/10.1093/plcell/koab287

Motifmatchr. (n.d.). Bioconductor. Retrieved February 1, 2024, from http://bioconductor.org/packages/motifmatchr/

Nelms, B., & Walbot, V. (2019). Defining the developmental program leading to meiosis in maize. *Science*, *364*(6435), 52–56. https://doi.org/10.1126/science.aav6428

Nelson, A. C., & Wardle, F. C. (2013). Conserved non-coding elements and cis regulation: Actions speak louder than words. *Development (Cambridge, England)*, *140*(7), 1385–1395. https://doi.org/10.1242/dev.084459

Ngan, C. Y., Wong, C. H., Tjong, H., Wang, W., Goldfeder, R. L., Choi, C., He, H., Gong, L., Lin, J., Urban, B., Chow, J., Li, M., Lim, J., Philip, V., Murray, S. A., Wang, H., & Wei, C.-L. (2020). Chromatin interaction analyses elucidate the roles of PRC2-bound silencers in mouse development. *Nature Genetics*, *52*(3), 264–272. https://doi.org/10.1038/s41588-020-0581-x

Nobori, T., Monell, A., Lee, T. A., Zhou, J., Nery, J., & Ecker, J. R. (2023). *Time-resolved single-cell and spatial gene regulatory atlas of plants under pathogen attack* (p. 2023.04.10.536170). bioRxiv. https://doi.org/10.1101/2023.04.10.536170

Nomura, M., Higuchi, T., Katayama, K., Taniguchi, M., Miyao-Tokutomi, M., Matsuoka, M., & Tajima, S. (2005). The Promoter for C4-type Mitochondrial Aspartate Aminotransferase Does not Direct Bundle Sheath-specific Expression in Transgenic Rice Plants. *Plant and Cell Physiology*, *46*(5), 743–753. https://doi.org/10.1093/pcp/pci077

Oka, R., Zicola, J., Weber, B., Anderson, S. N., Hodgman, C., Gent, J. I., Wesselink, J.-J., Springer, N. M., Hoefsloot, H. C. J., Turck, F., & Stam, M. (2017). Genome-wide mapping of transcriptional enhancer candidates using DNA and chromatin features in maize. *Genome Biology*, *18*(1), Article 1. https://doi.org/10.1186/s13059-017-1273-4

O'Leary, M. H. (1982). Phosphoenolpyruvate Carboxylase: An Enzymologist's View. *Annual Review of Plant Physiology*, *33*(1), 297–315. https://doi.org/10.1146/annurev.pp.33.060182.001501

Oppenheimer, D. G., Herman, P. L., Sivakumaran, S., Esch, J., & Marks, M. D. (1991). A myb gene required for leaf trichome differentiation in Arabidopsis is expressed in stipules. *Cell*, *67*(3), 483–493. https://doi.org/10.1016/0092-8674(91)90523-2

Otero, S., & Helariutta, Y. (2017). Companion cells: A diamond in the rough. *Journal of Experimental Botany*, *68*(1), Article 1. https://doi.org/10.1093/jxb/erw392

Outlaw, Jr. (1990). Kinetic Properties of Guard-Cell Phosphoenolpyruvate Carboxylase. *Biochemie Und Physiologie Der Pflanzen*, *186*(5), 317–325. https://doi.org/10.1016/S0015-3796(11)80224-6

Ouyang, S., Zhu, W., Hamilton, J., Lin, H., Campbell, M., Childs, K., Thibaud-Nissen, F., Malek, R. L., Lee, Y., Zheng, L., Orvis, J., Haas, B., Wortman, J., & Buell, C. R. (2007). The TIGR Rice Genome Annotation Resource: Improvements and new features. *Nucleic Acids Research*, *35*(Database issue), D883-887. https://doi.org/10.1093/nar/gkl976

Ouyang, W., Zhang, X., Guo, M., Wang, J., Wang, X., Gao, R., Ma, M., Xiang, X., Luan, S., Xing, F., Cao, Z., Yan, J., Li, G., & Li, X. (2023). Haplotype mapping of H3K27me3-associated chromatin interactions defines topological regulation of gene silencing in rice. *Cell Reports*, *42*(4), 112350. https://doi.org/10.1016/j.celrep.2023.112350

Panchy, N., Lehti-Shiu, M., & Shiu, S.-H. (2016). Evolution of Gene Duplication in Plants. *Plant Physiology*, *171*(4), 2294–2316. https://doi.org/10.1104/pp.16.00523

Pang, B., & Snyder, M. P. (2020). Systematic identification of silencers in human cells. *Nature Genetics*, *52*(3), Article 3. https://doi.org/10.1038/s41588-020-0578-5

Paterson, A. H., Bowers, J. E., Bruggmann, R., Dubchak, I., Grimwood, J., Gundlach, H., Haberer, G., Hellsten, U., Mitros, T., Poliakov, A., Schmutz, J., Spannagl, M., Tang, H., Wang, X., Wicker, T., Bharti, A. K., Chapman, J., Feltus, F. A., Gowik, U., … Rokhsar, D. S. (2009). The Sorghum bicolor genome and the diversification of grasses. *Nature*, *457*(7229), 551–556. https://doi.org/10.1038/nature07723

Paterson, A. H., Freeling, M., Tang, H., & Wang, X. (2010). Insights from the Comparison of Plant Genome Sequences. *Annual Review of Plant Biology*, *61*(1), 349–372. https://doi.org/10.1146/annurev-arplant-042809-112235

Perduns, R., Horst-Niessen, I., & Peterhansel, C. (2015). Photosynthetic Genes and Genes Associated with the C4 Trait in Maize Are Characterized by a Unique Class of Highly Regulated Histone Acetylation Peaks on Upstream Promoters. *Plant Physiology*, *168*(4), 1378–1388. https://doi.org/10.1104/pp.15.00934

Pettkó-Szandtner, A., Cserháti, M., Barrôco, R. M., Hariharan, S., Dudits, D., & Beemster, G. T. S. (2015). Core cell cycle regulatory genes in rice and their expression profiles across the growth zone of the leaf. *Journal of Plant Research*, *128*(6), 953–974. https://doi.org/10.1007/s10265-015-0754-3

Piazza, P., Bailey, C. D., Cartolano, M., Krieger, J., Cao, J., Ossowski, S., Schneeberger, K., He, F., de Meaux, J., Hall, N., MacLeod, N., Filatov, D., Hay, A., & Tsiantis, M. (2010). *Arabidopsis thaliana* Leaf Form Evolved via Loss of KNOX Expression in Leaves in Association with a Selective Sweep. *Current Biology*, *20*(24), 2223–2228. https://doi.org/10.1016/j.cub.2010.11.037

Picard Tools—By Broad Institute. (n.d.). Retrieved January 4, 2024, from http://broadinstitute.github.io/picard/

Pick, T. R., Bräutigam, A., Schlüter, U., Denton, A. K., Colmsee, C., Scholz, U., Fahnenstich, H., Pieruschka, R., Rascher, U., Sonnewald, U., & Weber, A. P. M. (2011). Systems Analysis of a Maize Leaf Developmental Gradient Redefines the Current C4 Model and Provides Candidates for Regulation. *The Plant Cell*, *23*(12), Article 12. https://doi.org/10.1105/tpc.111.090324

Pliner, H. A., Packer, J. S., McFaline-Figueroa, J. L., Cusanovich, D. A., Daza, R. M., Aghamirzaie, D., Srivatsan, S., Qiu, X., Jackson, D., Minkina, A., Adey, A. C., Steemers, F. J., Shendure, J., & Trapnell, C. (2018). Cicero Predicts cis-Regulatory DNA Interactions from Single-Cell Chromatin Accessibility Data. *Molecular Cell*, *71*(5), Article 5. https://doi.org/10.1016/j.molcel.2018.06.044

Poethig, R. S., & Sussex, I. M. (1985a). The cellular parameters of leaf development in tobacco: A clonal analysis. *Planta*, *165*(2), 170–184.

Poethig, R. S., & Sussex, I. M. (1985b). The developmental morphology and growth dynamics of the tobacco leaf. *Planta*, *165*(2), 158–169. https://doi.org/10.1007/BF00395038

Preissl, S., Gaulton, K. J., & Ren, B. (2023). Characterizing cis-regulatory elements using single-cell epigenomics. *Nature Reviews. Genetics*, *24*(1), 21–43. https://doi.org/10.1038/s41576-022-00509-1

preprocessCore. (n.d.). Bioconductor. Retrieved March 4, 2024, from http://bioconductor.org/packages/preprocessCore/

Quinlan, A. R., & Hall, I. M. (2010a). BEDTools: A flexible suite of utilities for comparing genomic features. *Bioinformatics*, *26*(6), 841–842. https://doi.org/10.1093/bioinformatics/btq033

Quinlan, A. R., & Hall, I. M. (2010b). BEDTools: A flexible suite of utilities for comparing genomic features. *Bioinformatics*, *26*(6), 841–842. https://doi.org/10.1093/bioinformatics/btq033

Raju, S. K. K. (2020). Comparative profiling examines roles of DNA regulatory sequences and accessible chromatin during cold stress response in grasses. *The Plant Cell*. https://doi.org/10.1105/tpc.20.00471

Rao, X., & Dixon, R. A. (2016). The Differences between NAD-ME and NADP-ME Subtypes of C4 Photosynthesis: More than Decarboxylating Enzymes. *Frontiers in Plant Science*, *7*. https://www.frontiersin.org/articles/10.3389/fpls.2016.01525

Reinhardt, D., Frenz, M., Mandel, T., & Kuhlemeier, C. (2003). Microsurgical and laser ablation analysis of interactions between the zones and layers of the tomato shoot apical meristem. *Development*, *130*(17), 4073–4083. https://doi.org/10.1242/dev.00596

Reynoso, M. A., Kajala, K., Bajic, M., West, D. A., Pauluzzi, G., Yao, A. I., Hatch, K., Zumstein, K., Woodhouse, M., Rodriguez-Medina, J., Sinha, N., Brady, S. M., Deal, R. B., & Bailey-Serres, J. (2019). Evolutionary flexibility in flooding response circuitry in angiosperms. *Science*, *365*(6459), Article 6459. https://doi.org/10.1126/science.aax8862

Ricci, W. A., Lu, Z., Ji, L., Marand, A. P., Ethridge, C. L., Murphy, N. G., Noshay, J. M., Galli, M., Mejía-Guerra, M. K., Colomé-Tatché, M., Johannes, F., Rowley, M. J., Corces, V. G., Zhai, J., Scanlon, M. J., Buckler, E. S., Gallavotti, A., Springer, N. M., Schmitz, R. J., & Zhang, X. (2019). Widespread long-range cis -regulatory elements in the maize genome. *Nature Plants*, *5*(12), Article 12. https://doi.org/10.1038/s41477-019-0547-0

Richardson, A. E., & Hake, S. (2022). The power of classic maize mutants: Driving forward our fundamental understanding of plants. *The Plant Cell*, *34*(7), 2505–2517. https://doi.org/10.1093/plcell/koac081

Robinson, M. D., McCarthy, D. J., & Smyth, G. K. (2010). edgeR: A Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, *26*(1), 139–140. https://doi.org/10.1093/bioinformatics/btp616

Rodríguez-Leal, D., Lemmon, Z. H., Man, J., Bartlett, M. E., & Lippman, Z. B. (2017). Engineering Quantitative Trait Variation for Crop Improvement by Genome Editing. *Cell*, *171*(2), Article 2. https://doi.org/10.1016/j.cell.2017.08.030

Rodriques, S. G., Stickels, R. R., Goeva, A., Martin, C. A., Murray, E., Vanderburg, C. R., Welch, J., Chen, L. M., Chen, F., & Macosko, E. Z. (2019). Slide-seq: A scalable technology for measuring genome-wide expression at high spatial resolution. *Science*, *363*(6434), 1463–1467. https://doi.org/10.1126/science.aaw1219

Rombolá-Caldentey, B., Rueda-Romero, P., Iglesias-Fernández, R., Carbonero, P., & Oñate-Sánchez, L. (2014). Arabidopsis DELLA and two HD-ZIP transcription factors regulate GA signaling in the epidermis through the L1 box cis-element. *The Plant Cell*, *26*(7), 2905–2919. https://doi.org/10.1105/tpc.114.127647

Rosado, D., Ackermann, A., Spassibojko, O., Rossi, M., & Pedmale, U. V. (2022). WRKY transcription factors and ethylene signaling modify root growth during the shade-avoidance response. *Plant Physiology*, *188*(2), 1294–1311. https://doi.org/10.1093/plphys/kiab493

Rosas, U., Mei, Y., Xie, Q., Banta, J. A., Zhou, R. W., Seufferheld, G., Gerard, S., Chou, L., Bhambhra, N., Parks, J. D., Flowers, J. M., McClung, C. R., Hanzawa, Y., & Purugganan,

M. D. (2014). Variation in Arabidopsis flowering time associated with cis-regulatory variation in CONSTANS. *Nature Communications*, *5*(1), Article 1. https://doi.org/10.1038/ncomms4651

Ryu, K. H., Huang, L., Kang, H. M., & Schiefelbein, J. (2019). Single-Cell RNA Sequencing Resolves Molecular Relationships Among Individual Plant Cells. *Plant Physiology*, *179*(4), 1444–1456. https://doi.org/10.1104/pp.18.01482

Sage, R. F. (2016). A portrait of the C 4 photosynthetic family on the 50th anniversary of its discovery: Species number, evolutionary lineages, and Hall of Fame. *Journal of Experimental Botany*, *67*(14), Article 14. https://doi.org/10.1093/jxb/erw156

Sage, R. F., Christin, P.-A., & Edwards, E. J. (2011). The C4 plant lineages of planet Earth. *Journal of Experimental Botany*, *62*(9), 3155–3169. https://doi.org/10.1093/jxb/err048

Satija, R., Farrell, J. A., Gennert, D., Schier, A. F., & Regev, A. (2015). Spatial reconstruction of single-cell gene expression data. *Nature Biotechnology*, *33*(5), 495–502. https://doi.org/10.1038/nbt.3192

Sayers, E. W., Bolton, E. E., Brister, J. R., Canese, K., Chan, J., Comeau, D. C., Connor, R., Funk, K., Kelly, C., Kim, S., Madej, T., Marchler-Bauer, A., Lanczycki, C., Lathrop, S., Lu, Z., Thibaud-Nissen, F., Murphy, T., Phan, L., Skripchenko, Y., … Sherry, S. T. (2022). Database resources of the national center for biotechnology information. *Nucleic Acids Research*, *50*(D1), D20–D26. https://doi.org/10.1093/nar/gkab1112

Schep, A. N., Buenrostro, J. D., Denny, S. K., Schwartz, K., Sherlock, G., & Greenleaf, W. J. (2015). Structured nucleosome fingerprints enable high-resolution mapping of chromatin architecture within regulatory regions. *Genome Research*, *25*(11), 1757. https://doi.org/10.1101/gr.192294.115

Schmitz, R. J., Grotewold, E., & Stam, M. (2022). Cis-regulatory sequences in plants: Their importance, discovery, and future challenges. *The Plant Cell*, *34*(2), 718–741. https://doi.org/10.1093/plcell/koab281

Schuler, M. L., Sedelnikova, O. V., Walker, B. J., Westhoff, P., & Langdale, J. A. (2018). SHORTROOT-Mediated Increase in Stomatal Density Has No Impact on Photosynthetic Efficiency. *Plant Physiology*, *176*(1), 757–772. https://doi.org/10.1104/pp.17.01005

Sharman, B. C. (1942). Developmental Anatomy of the Shoot of Zea mays L. *Annals of Botany*, *6*(2), 245–282. https://doi.org/10.1093/oxfordjournals.aob.a088407

Sheen, J. (1999). C4 Gene Expression. *Annual Review of Plant Physiology and Plant Molecular Biology*, *50*(1), 187–217. https://doi.org/10.1146/annurev.arplant.50.1.187

Shen, W., Le, S., Li, Y., & Hu, F. (2016). SeqKit: A Cross-Platform and Ultrafast Toolkit for FASTA/Q File Manipulation. *PLOS ONE*, *11*(10), e0163962. https://doi.org/10.1371/journal.pone.0163962

Shimadzu, S., Furuya, T., & Kondo, Y. (2023). Molecular Mechanisms Underlying the Establishment and Maintenance of Vascular Stem Cells in Arabidopsis thaliana. *Plant and Cell Physiology*, *64*(3), 274–283. https://doi.org/10.1093/pcp/pcac161

Shlyueva, D., Stampfel, G., & Stark, A. (2014). Transcriptional enhancers: From properties to genome-wide predictions. *Nature Reviews Genetics*, *15*(4), Article 4. https://doi.org/10.1038/nrg3682

Shulse, C. N., Cole, B. J., Ciobanu, D., Lin, J., Yoshinaga, Y., Gouran, M., Turco, G. M., Zhu, Y., O'Malley, R. C., Brady, S. M., & Dickel, D. E. (2019). High-Throughput Single-Cell Transcriptome Profiling of Plant Cell Types. *Cell Reports*, *27*(7), 2241-2247.e4. https://doi.org/10.1016/j.celrep.2019.04.054

Sinha, N., & Hake, S. (1990). Mutant characters of knotted maize leaves are determined in the innermost tissue layers. *Developmental Biology*, *141*(1), 203–210. https://doi.org/10.1016/0012-1606(90)90115-y

Son, D., Jo, J., & Sugiyama, T. (1991). Purification and characterization of alanine aminotransferase from Panicum miliaceum leaves. *Archives of Biochemistry and Biophysics*, *289*(1), 262–266. https://doi.org/10.1016/0003-9861(91)90470-4

Song, B., Buckler, E. S., Wang, H., Wu, Y., Rees, E., Kellogg, E. A., Gates, D. J., Khaipho-Burch, M., Bradbury, P. J., Ross-Ibarra, J., Hufford, M. B., & Romay, M. C. (2021). Conserved noncoding sequences provide insights into regulatory sequence and loss of gene expression in maize. *Genome Research*, *31*(7), 1245–1257. https://doi.org/10.1101/gr.266528.120

Stadler, R., & Sauer, N. (1996). The Arabidopsis thaliana AtSUC2 Gene is Specifically Expressed in Companion Cells. *Botanica Acta*, *109*(4), 299–306. https://doi.org/10.1111/j.1438-8677.1996.tb00577.x

Stadler, R., Wright, K. M., Lauterbach, C., Amon, G., Gahrtz, M., Feuerstein, A., Oparka, K. J., & Sauer, N. (2005). Expression of GFP-fusions in Arabidopsis companion cells reveals non-specific protein trafficking into sieve elements and identifies a novel post-phloem domain in roots. *The Plant Journal*, *41*(2), 319–331. https://doi.org/10.1111/j.1365-313X.2004.02298.x

Stelpflug, S. C., Sekhon, R. S., Vaillancourt, B., Hirsch, C. N., Buell, C. R., de Leon, N., & Kaeppler, S. M. (2016). An Expanded Maize Gene Expression Atlas based on RNA Sequencing and its Use to Explore Root Development. *The Plant Genome*, *9*(1). https://doi.org/10.3835/plantgenome2015.04.0025

Stewart, G. W. (1993). On the Early History of the Singular Value Decomposition. *SIAM Review*, *35*(4), 551–566.

Stickels, R. R., Murray, E., Kumar, P., Li, J., Marshall, J. L., Di Bella, D. J., Arlotta, P., Macosko, E. Z., & Chen, F. (2021). Highly sensitive spatial transcriptomics at near-cellular resolution with Slide-seqV2. *Nature Biotechnology*, *39*(3), Article 3. https://doi.org/10.1038/s41587-020-0739-1

Stovner, E. B., & Sætrom, P. (2019). Epic2 efficiently finds diffuse domains in ChIP-seq data. *Bioinformatics (Oxford, England)*, *35*(21), 4392–4393. https://doi.org/10.1093/bioinformatics/btz232

Strasburger, E. (1888). *Histologische Beiträge* (pp. 1–268). G. Fischer. https://doi.org/10.5962/bhl.title.24451

Studer, A. J., Schnable, J. C., Weissmann, S., Kolbe, A. R., McKain, M. R., Shao, Y., Cousins, A. B., Kellogg, E. A., & Brutnell, T. P. (2016). The draft genome of the C3 panicoid grass species Dichanthelium oligosanthes. *Genome Biology*, *17*(1), 223. https://doi.org/10.1186/s13059-016-1080-3

Studer, A., Zhao, Q., Ross-Ibarra, J., & Doebley, J. (2011). Identification of a functional transposon insertion in the maize domestication gene *tb1*. *Nature Genetics*, *43*(11), Article 11. https://doi.org/10.1038/ng.942

Swift, J., Luginbuehl, L. H., Schreier, T., Donald, R. M., Lee, T., Nery, J., Ecker, J. R., & Hibberd, J. M. (2023). *Single nuclei sequencing reveals C4 photosynthesis is based on rewiring of ancestral cell identity networks* (p. 2023.10.26.562893). bioRxiv. https://doi.org/10.1101/2023.10.26.562893

Tan, F.-Q., Wang, W., Li, J., Lu, Y., Zhu, B., Hu, F., Li, Q., Zhao, Y., & Zhou, D.-X. (2022). A coiled-coil protein associates Polycomb Repressive Complex 2 with KNOX/BELL transcription factors to maintain silencing of cell differentiation-promoting genes in the shoot apex. *The Plant Cell*, *34*(8), 2969–2988. https://doi.org/10.1093/plcell/koac133

Taniguchi, M., Kobe, A., Kato, M., & Sugiyama, T. (1995). Aspartate Aminotransferase Isozymes in Panicum miliaceum L, an NAD-Malic Enzyme-Type C4 Plant: Comparison of Enzymatic-Properties, Primary Structures, and Expression Patterns. *Archives of Biochemistry and Biophysics*, *318*(2), 295–306. https://doi.org/10.1006/abbi.1995.1233

Taniguchi, M., Sawaki, H., Sasakawa, H., Hase, T., & Sugiyama, T. (1992). Cloning and sequence analysis of cDNA encoding aspartate aminotransferase isozymes from Panicum miliaceum L., a C4 plant. *European Journal of Biochemistry*, *204*(2), 611–620. https://doi.org/10.1111/j.1432-1033.1992.tb16674.x

Taniguchi, M., & Sugiyama, T. (1997). The Expression of 2-Oxoglutarate/Malate Translocator in the Bundle-Sheath Mitochondria of Panicum miliaceum, a NAD-Malic Enzyme-Type C4 Plant, Is Regulated by Light and Development. *Plant Physiology*, *114*(1), 285–293. https://doi.org/10.1104/pp.114.1.285

Taniguchi, Y., Nagasaki, J., Kawasaki, M., Miyake, H., Sugiyama, T., & Taniguchi, M. (2004). Differentiation of Dicarboxylate Transporters in Mesophyll and Bundle Sheath Chloroplasts of Maize. *Plant and Cell Physiology*, *45*(2), 187–200. https://doi.org/10.1093/pcp/pch022

Tarailo-Graovac, M., & Chen, N. (2009). Using RepeatMasker to identify repetitive elements in genomic sequences. *Current Protocols in Bioinformatics*, *Chapter 4*, 4.10.1-4.10.14. https://doi.org/10.1002/0471250953.bi0410s25

Tausta, S. L., Li, P., Si, Y., Gandotra, N., Liu, P., Sun, Q., Brutnell, T. P., & Nelson, T. (2014). Developmental dynamics of Kranz cell transcriptional specificity in maize leaf reveals early onset of C4-related processes. *Journal of Experimental Botany*, *65*(13), Article 13. https://doi.org/10.1093/jxb/eru152

Tian, T., Liu, Y., Yan, H., You, Q., Yi, X., Du, Z., Xu, W., & Su, Z. (2017). agriGO v2.0: A GO analysis toolkit for the agricultural community, 2017 update. *Nucleic Acids Research*, *45*(W1), W122–W129. https://doi.org/10.1093/nar/gkx382

Ton, M.-L. N., Keitley, D., Theeuwes, B., Guibentif, C., Ahnfelt-Rønne, J., Andreassen, T. K., Calero-Nieto, F. J., Imaz-Rosshandler, I., Pijuan-Sala, B., Nichols, J., Benito-Gutiérrez, È., Marioni, J. C., & Göttgens, B. (2023). An atlas of rabbit development as a model for single-cell comparative genomics. *Nature Cell Biology*, *25*(7), 1061–1072. https://doi.org/10.1038/s41556-023-01174-0

Tonosaki, K., & Kinoshita, T. (2015). Possible roles for polycomb repressive complex 2 in cereal endosperm. *Frontiers in Plant Science*, *6*, 144. https://doi.org/10.3389/fpls.2015.00144

Toufighi, K., Brady, S. M., Austin, R., Ly, E., & Provart, N. J. (2005). The Botany Array Resource: E-Northerns, Expression Angling, and promoter analyses. *The Plant Journal*, *43*(1), 153–163. https://doi.org/10.1111/j.1365-313X.2005.02437.x

Trapnell, C., Cacchiarelli, D., Grimsby, J., Pokharel, P., Li, S., Morse, M., Lennon, N. J., Livak, K. J., Mikkelsen, T. S., & Rinn, J. L. (2014). The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nature Biotechnology*, *32*(4), Article 4. https://doi.org/10.1038/nbt.2859

Tu, X., Marand, A. P., Schmitz, R. J., & Zhong, S. (2022). A combinatorial indexing strategy for low-cost epigenomic profiling of plant single cells. *Plant Communications*, *3*(4), 100308. https://doi.org/10.1016/j.xplc.2022.100308

Tu, X., Ren, S., Shen, W., Li, J., Li, Y., Li, C., Li, Y., Zong, Z., Xie, W., Grierson, D., Fei, Z., Giovannoni, J., Li, P., & Zhong, S. (2022). Limited conservation in cross-species comparison of GLK transcription factor binding suggested wide-spread cistrome divergence. *Nature Communications*, *13*(1), Article 1. https://doi.org/10.1038/s41467-022-35438-4

Tyler, S. R., Lozano-Ojalvo, D., Guccione, E., & Schadt, E. E. (2024). Anti-correlated feature selection prevents false discovery of subpopulations in scRNAseq. *Nature Communications*, *15*(1), 699. https://doi.org/10.1038/s41467-023-43406-9

UENO, O., KAWANO, Y., WAKAYAMA, M., & TAKEDA, T. (2006). Leaf Vascular Systems in C3 and C4 Grasses: A Two-dimensional Analysis. *Annals of Botany*, *97*(4), 611–621. https://doi.org/10.1093/aob/mcl010

Universalmotif. (n.d.). Bioconductor. Retrieved January 23, 2024, from http://bioconductor.org/packages/universalmotif/

Viret, J. F., Mabrouk, Y., & Bogorad, L. (1994). Transcriptional photoregulation of cell-type-preferred expression of maize rbcS-m3: 3' and 5' sequences are involved. *Proceedings of the National Academy of Sciences*, *91*(18), 8577–8581. https://doi.org/10.1073/pnas.91.18.8577

Wagner-Menghin, M. M. (2014). Binomial Test. In *Wiley StatsRef: Statistics Reference Online*. John Wiley & Sons, Ltd. https://doi.org/10.1002/9781118445112.stat06340

Waltman, L., & Eck, N. J. (2013). A smart local moving algorithm for large-scale modularity-based community detection. *The European Physical Journal B*, *86*(11), 471. https://doi.org/10.1140/epjb/e2013-40829-0

Wang, P., Kelly, S., Fouracre, J. P., & Langdale, J. A. (2013). Genome-wide transcript analysis of early maize leaf development reveals gene cohorts associated with the differentiation of C4 Kranz anatomy. *The Plant Journal*, *75*(4), Article 4. https://doi.org/10.1111/tpj.12229

Wang, Y., Bräutigam, A., Weber, A. P. M., & Zhu, X.-G. (2014). Three distinct biochemical subtypes of C4 photosynthesis? A modelling analysis. *Journal of Experimental Botany*, *65*(13), 3567–3578. https://doi.org/10.1093/jxb/eru058

Wang, Y., Strauss, S., Liu, S., Pieper, B., Lymbouridou, R., Runions, A., & Tsiantis, M. (2022). The cellular basis for synergy between *RCO* and *KNOX1* homeobox genes in leaf shape diversity. *Current Biology*, *32*(17), 3773-3784.e5. https://doi.org/10.1016/j.cub.2022.08.020

Wang, Z., Liu, M., Lai, F., Fu, Q., Xie, L., Fang, Y., Zhou, Q., & Li, G. (2023). AraENCODE: A comprehensive epigenomic database of Arabidopsis thaliana. *Molecular Plant*, *16*(7), 1113–1116. https://doi.org/10.1016/j.molp.2023.06.005

Washburn, J. D., Strable, J., Dickinson, P., Kothapalli, S. S., Brose, J. M., Covshoff, S., Conant, G. C., Hibberd, J. M., & Pires, J. C. (2021). Distinct C4 sub-types and C3 bundle sheath isolation in the Paniceae grasses. *Plant Direct*, *5*(12), e373. https://doi.org/10.1002/pld3.373

Wheeler, T., & Braun, J. (2013). Climate Change Impacts on Global Food Security. *Science*, *341*(6145), Article 6145. https://doi.org/10.1126/science.1239402

Wilhelm, K. (1880). *Beiträge zur Kenntniss des Siebröhrenapparates dicotyler Pflanzen*. Verlag von Wilhelm Engelmann.

Wittkopp, P. J., & Kalay, G. (2012). Cis-regulatory elements: Molecular mechanisms and evolutionary processes underlying divergence. *Nature Reviews Genetics*, *13*(1), Article 1. https://doi.org/10.1038/nrg3095

Wolfe, K. H., Gouy, M., Yang, Y. W., Sharp, P. M., & Li, W. H. (1989). Date of the monocot-dicot divergence estimated from chloroplast DNA sequence data. *Proceedings of the National Academy of Sciences*, *86*(16), 6201–6205. https://doi.org/10.1073/pnas.86.16.6201

Wolock, S. L., Lopez, R., & Klein, A. M. (2019a). Scrublet: Computational Identification of Cell Doublets in Single-Cell Transcriptomic Data. *Cell Systems*, *8*(4), 281-291.e9. https://doi.org/10.1016/j.cels.2018.11.005

Wolock, S. L., Lopez, R., & Klein, A. M. (2019b). Scrublet: Computational Identification of Cell Doublets in Single-Cell Transcriptomic Data. *Cell Systems*, *8*(4), 281-291.e9. https://doi.org/10.1016/j.cels.2018.11.005

Woolfe, A., Goodson, M., Goode, D. K., Snell, P., McEwen, G. K., Vavouri, T., Smith, S. F., North, P., Callaway, H., Kelly, K., Walter, K., Abnizova, I., Gilks, W., Edwards, Y. J. K., Cooke, J. E., & Elgar, G. (2004). Highly Conserved Non-Coding Sequences Are Associated with Vertebrate Development. *PLOS Biology*, *3*(1), e7. https://doi.org/10.1371/journal.pbio.0030007

Wu, R., Li, S., He, S., Waßmann, F., Yu, C., Qin, G., Schreiber, L., Qu, L.-J., & Gu, H. (2011). CFL1, a WW Domain Protein, Regulates Cuticle Development by Modulating the

Function of HDG1, a Class IV Homeodomain Transcription Factor, in Rice and Arabidopsis. *The Plant Cell*, *23*(9), 3392–3411. https://doi.org/10.1105/tpc.111.088625

Wu, T.-Y., Goh, H., Azodi, C. B., Krishnamoorthi, S., Liu, M.-J., & Urano, D. (2021). Evolutionarily conserved hierarchical gene regulatory networks for plant salt stress response. *Nature Plants*, *7*(6), 787–799. https://doi.org/10.1038/s41477-021-00929-7

Wu, T.-Y., Müller, M., Gruissem, W., & Bhullar, N. K. (2020). Genome Wide Analysis of the Transcriptional Profiles in Different Regions of the Developing Rice Grains. *Rice*, *13*(1), 62. https://doi.org/10.1186/s12284-020-00421-4

Wucherpfennig, J. I., Howes, T. R., Au, J. N., Au, E. H., Roberts Kingman, G. A., Brady, S. D., Herbert, A. L., Reimchen, T. E., Bell, M. A., Lowe, C. B., Dalziel, A. C., & Kingsley, D. M. (2022). Evolution of stickleback spines through independent cis-regulatory changes at HOXDB. *Nature Ecology & Evolution*, *6*(10), 1537–1552. https://doi.org/10.1038/s41559-022-01855-3

Xie, L., Liu, M., Zhao, L., Cao, K., Wang, P., Xu, W., Sung, W.-K., Li, X., & Li, G. (2021). RiceENCODE: A comprehensive epigenomic database as a rice Encyclopedia of DNA Elements. *Molecular Plant*, *14*(10), 1604–1606. https://doi.org/10.1016/j.molp.2021.08.018

Xu, Y., Wang, Y., Long, Q., Huang, J., Wang, Y., Zhou, K., Zheng, M., Sun, J., Chen, H., Chen, S., Jiang, L., Wang, C., & Wan, J. (2014). Overexpression of OsZHD1, a zinc finger homeodomain class homeobox transcription factor, induces abaxially curled and drooping leaf in rice. *Planta*, *239*(4), 803–816. https://doi.org/10.1007/s00425-013-2009-7

Yan, F., Powell, D. R., Curtis, D. J., & Wong, N. C. (2020). From reads to insight: A hitchhiker's guide to ATAC-seq data analysis. *Genome Biology*, *21*(1), Article 1. https://doi.org/10.1186/s13059-020-1929-3

Yan, H., Mendieta, J. P., Zhang, X., Marand, A. P., Liang, Y., Luo, Z., Roulé, T., Wagner, D., Tu, X., Wang, Y., Zhong, S., Wessler, S. R., & Schmitz, R. J. (2024). *Evolution of cell-type-specific accessible chromatin regions and the cis-regulatory elements that drive lineage-specific innovation* (p. 2024.01.08.574753). bioRxiv. https://doi.org/10.1101/2024.01.08.574753

Yanagisawa, S. (2000). Dof1 and Dof2 transcription factors are associated with expression of multiple genes involved in carbon metabolism in maize. *The Plant Journal*, *21*(3), 281–288. https://doi.org/10.1046/j.1365-313x.2000.00685.x

Yocca, A. E., Lu, Z., Schmitz, R. J., Freeling, M., & Edger, P. P. (2021). Evolution of Conserved Noncoding Sequences in Arabidopsis thaliana. *Molecular Biology and Evolution*, *38*(7), Article 7. https://doi.org/10.1093/molbev/msab042

Yu, L.-H., Wu, S.-J., Peng, Y.-S., Liu, R.-N., Chen, X., Zhao, P., Xu, P., Zhu, J.-B., Jiao, G.-L., Pei, Y., & Xiang, C.-B. (2016). Arabidopsis EDT1/HDG11 improves drought and salt tolerance in cotton and poplar and increases cotton yield in the field. *Plant Biotechnology Journal*, *14*(1), 72–84. https://doi.org/10.1111/pbi.12358

Yu, Y., Zhang, H., Long, Y., Shu, Y., & Zhai, J. (2022). Plant Public RNA-seq Database: A comprehensive online database for expression analysis of 45 000 plant public RNA-Seq libraries. *Plant Biotechnology Journal*, *20*(5), 806–808. https://doi.org/10.1111/pbi.13798

Yuan, W., Luo, X., Li, Z., Yang, W., Wang, Y., Liu, R., Du, J., & He, Y. (2016). A cis cold memory element and a trans epigenome reader mediate Polycomb silencing of FLC by

vernalization in Arabidopsis. *Nature Genetics*, *48*(12), Article 12. https://doi.org/10.1038/ng.3712

Zemke, N. R., Armand, E. J., Wang, W., Lee, S., Zhou, J., Li, Y. E., Liu, H., Tian, W., Nery, J. R., Castanon, R. G., Bartlett, A., Osteen, J. K., Li, D., Zhuo, X., Xu, V., Chang, L., Dong, K., Indralingam, H. S., Rink, J. A., … Ren, B. (2023). Conserved and divergent gene regulatory programs of the mammalian neocortex. *Nature*, *624*(7991), Article 7991. https://doi.org/10.1038/s41586-023-06819-6

Zeng, H. (2022). What is a cell type and how to define it? *Cell*, *185*(15), Article 15. https://doi.org/10.1016/j.cell.2022.06.031

Zeng, Z., Zhang, W., Marand, A. P., Zhu, B., Buell, C. R., & Jiang, J. (2019). Cold stress induces enhanced chromatin accessibility and bivalent histone modifications H3K4me3 and H3K27me3 of active genes in potato. *Genome Biology*, *20*(1), Article 1. https://doi.org/10.1186/s13059-019-1731-2

Zhang, K., Hocker, J. D., Miller, M., Hou, X., Chiou, J., Poirion, O. B., Qiu, Y., Li, Y. E., Gaulton, K. J., Wang, A., Preissl, S., & Ren, B. (2021a). A single-cell atlas of chromatin accessibility in the human genome. *Cell*, *184*(24), Article 24. https://doi.org/10.1016/j.cell.2021.10.024

Zhang, K., Hocker, J. D., Miller, M., Hou, X., Chiou, J., Poirion, O. B., Qiu, Y., Li, Y. E., Gaulton, K. J., Wang, A., Preissl, S., & Ren, B. (2021b). A single-cell atlas of chromatin accessibility in the human genome. *Cell*, *184*(24), Article 24. https://doi.org/10.1016/j.cell.2021.10.024

Zhang, L., He, C., Lai, Y., Wang, Y., Kang, L., Liu, A., Lan, C., Su, H., Gao, Y., Li, Z., Yang, F., Li, Q., Mao, H., Chen, D., Chen, W., Kaufmann, K., & Yan, W. (2023). Asymmetric

gene expression and cell-type-specific regulatory networks in the root of bread wheat revealed by single-cell multiomics analysis. *Genome Biology*, *24*(1), 65. https://doi.org/10.1186/s13059-023-02908-x

Zhang, T.-Q., Chen, Y., Liu, Y., Lin, W.-H., & Wang, J.-W. (2021). Single-cell transcriptome atlas and chromatin accessibility landscape reveal differentiation trajectories in the rice root. *Nature Communications*, *12*(1), 2053. https://doi.org/10.1038/s41467-021-22352-4

Zhang, T.-Q., Xu, Z.-G., Shang, G.-D., & Wang, J.-W. (2019). A Single-Cell RNA Sequencing Profiles the Developmental Landscape of Arabidopsis Root. *Molecular Plant*, *12*(5), 648–660. https://doi.org/10.1016/j.molp.2019.04.004

Zhang, X., Marand, A. P., Yan, H., & Schmitz, R. J. (2024). Massive-scale single-cell chromatin accessibility sequencing using combinatorial fluidic indexing. *bioRxiv: The Preprint Server for Biology*, 2023.09.17.558155. https://doi.org/10.1101/2023.09.17.558155

Zhang, Y., Liu, T., Meyer, C. A., Eeckhoute, J., Johnson, D. S., Bernstein, B. E., Nusbaum, C., Myers, R. M., Brown, M., Li, W., & Liu, X. S. (2008). Model-based analysis of ChIP-Seq (MACS). *Genome Biology*, *9*(9), Article 9. https://doi.org/10.1186/gb-2008-9-9-r137

Zhao, H., Zhang, W., Chen, L., Wang, L., Marand, A. P., Wu, Y., & Jiang, J. (2018). Proliferation of Regulatory DNA Elements Derived from Transposable Elements in the Maize Genome. *Plant Physiology*, *176*(4), Article 4. https://doi.org/10.1104/pp.17.01467

Zheng, G. X. Y., Terry, J. M., Belgrader, P., Ryvkin, P., Bent, Z. W., Wilson, R., Ziraldo, S. B., Wheeler, T. D., McDermott, G. P., Zhu, J., Gregory, M. T., Shuga, J., Montesclaros, L., Underwood, J. G., Masquelier, D. A., Nishimura, S. Y., Schnall-Levin, M., Wyatt, P. W., Hindson, C. M., … Bielas, J. H. (2017). Massively parallel digital transcriptional

profiling of single cells. *Nature Communications*, *8*(1), Article 1. https://doi.org/10.1038/ncomms14049

Zhou, P., Enders, T. A., Myers, Z. A., Magnusson, E., Crisp, P. A., Noshay, J. M., Gomez-Cano, F., Liang, Z., Grotewold, E., Greenham, K., & Springer, N. M. (2022). Prediction of conserved and variable heat and cold stress response in maize using cis-regulatory information. *The Plant Cell*, *34*(1), 514–534. https://doi.org/10.1093/plcell/koab267

Zou, C., Li, L., Miki, D., Li, D., Tang, Q., Xiao, L., Rajput, S., Deng, P., Peng, L., Jia, W., Huang, R., Zhang, M., Sun, Y., Hu, J., Fu, X., Schnable, P. S., Chang, Y., Li, F., Zhang, H., … Zhang, H. (2019). The genome of broomcorn millet. *Nature Communications*, *10*(1), 436. https://doi.org/10.1038/s41467-019-08409-5

Zou, C., Sun, K., Mackaluso, J. D., Seddon, A. E., Jin, R., Thomashow, M. F., & Shiu, S.-H. (2011). Cis-regulatory code of stress-responsive transcription in Arabidopsis thaliana. *Proceedings of the National Academy of Sciences of the United States of America*, *108*(36), 14992–14997. https://doi.org/10.1073/pnas.1103202108