Decoding *cis*-regulatory control and evolution of conserved and divergent phenotypes in plants

A dissertation presented by

Danielle Ciren

to the

Cold Spring Harbor Laboratory School of Biological Sciences

in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

in

Biological Sciences

Cold Spring Harbor Laboratory

September 2023

Abstract

Eukaryotic gene expression is regulated by specific genetic and physical interactions among *cis*-regulatory elements (CREs), which ensure gene expression is coordinated to a precise level, in a particular cell type, and at a specific time during development. This precision is facilitated by multiple epigenetic modifications, constantly remodeling the chromatin to exclude or encourage transcription factor binding at CREs. CREs are located in every genomic context, including upstream, downstream, within the gene itself (in the UTRs, introns, and exons), and even at distal sites that create DNA loops to contact promoters. It is the combination of multiple CREs, in multiple genomic locations, that come together to regulate genes in space and time. Throughout this thesis work, we have studied CREs and their interactions within an evolutionary framework, using a functional genetics approach. We sought to explore two main questions about the evolution of regulatory regions. The first, how do conserved genes in distantly related organisms maintain similar functions and spatiotemporal expression patterns, often amidst drastic *cis*-regulatory sequence divergence? And the second, how does variation in CREs contribute to phenotypic divergence? Using CRISPR-Cas9, we generated 89 unique mutations in *cis*-regulatory regions upstream and downstream of the dosage-sensitive developmental genes *CLAVATA3 (CLV3)* and *SELF PRUNING 5G (SP5G),* and quantified the phenotypic effect. *CLV3* is highly conserved among *Arabidopsis* and tomato, despite extremely diverged regulatory regions. We find evidence supporting a billboard model of CRE organization, in which the particular transcription factor bindings sites regulating *CLV3* in both species are conserved, however their arrangement (spacing, order, orientation, etc.) is more flexible to change. In contrast, closely related species of wild and domesticated tomato have different flowering time responses to daylength, which has been attributed to differences in expression of *SP5G.* We found that multiple CREs are involved in the regulation of *SP5G* in the wild species, although none of our engineered *cis*-regulatory alleles mimicked the domestication phenotype on its own. Further investigation into the genetic and physical interactions among *SP5G* CREs, as well as their molecular consequences, will

enhance our understanding of potential mechanisms of phenotypic divergence. Thus, we have found evidence for robustness and higher order complexity within *cis*-regulatory regions, in the context of both conserved and diverged traits. This work has explored fundamental principles of gene regulation using a functional genetics approach rarely applied in the field. Here we present a new perspective on *cis*-regulatory mechanisms of evolution, using *in vivo* mutagenesis experiments that have provided an improved understanding of the functional, phenotypic relevance of CREs and their interactions in the regulation of genes.

Acknowledgements

"You will also find that help will always be given at ~~Hogwarts~~ Cold Spring Harbor to those who ask for it." – Harry Potter and the Chamber of Secrets

I have many people to acknowledge for their support during the course of this thesis project. First and foremost, I have to thank my thesis advisor, Zach Lippman, for his continued guidance during the last four years. He has been a committed and patient mentor, always making time to discuss experiments and concerns, and provide feedback. He led the lab through a very uncertain global pandemic, and without his support I know I would not have made it this far.

I would also like to thank everyone in the Lippman lab. I feel very fortunate to have found a lab with such intelligent and kind people. I have enjoyed getting to know them better both within the lab and outside of it, playing volleyball and building rafts! I have appreciated learning from and collaborating with so many lab members over the years, including Xingang, Sophie, Jia, Anat, Lyndsey, Amy, Jack, Hagai, Miguel, Iacopo, Gina, Brooke, Uma, Tak, Matthias, Sebastian, Jennifer, and Andrew. A special thanks to Sophie Zebell and Sebastian Soyk for their direct mentoring in the ways of genetics and molecular biology. Thank you to the greenhouse crew, Blaine and Sharron, for always diligently taking care of my plants.

I have also been fortunate to be part of the best group at Cold Spring Harbor Lab, the plant people. Thank you to all of my colleagues in Delbruck for creating a nice environment to do science. Thanks to Michelle Murrell for ensuring the smooth operation of the building and constant supply of reagents and equipment. Thanks also to the farm crew, Kyle Schlecht and Tim Mulligan, who kept my plants (and my dreams of graduating) alive at Uplands and Woodhouse.

I also have to thank Joyce VanEck, and everyone on the Boyce Thompson Institute plant transformation team, for working incredibly hard to transform all of the tomato CRISPR constructs used in this thesis project.

I'd like to thank my thesis committee for their feedback, advice, and scientific discussions over the years. Thank you to Ullas Pedmale for being a supportive academic mentor, as well as Jessica Tollkuhn for agreeing to be my committee chair, and Adam Siepel for being a committee member. Thank you to all of my teachers in first year, who gave me the unique opportunity to learn about so many different areas of biology, which I know will continue to aid me throughout my career. And thank you to everyone at the Cold Spring Harbor School of Biological Sciences – Alyson, Big Kim, Little Kim, Monn, Alex, and Zach. I always felt at ease knowing that I could approach the school with anything, and receive immediate support and advice.

I need to express my gratitude to my fellow classmates, many of whom I lived with at various points during the last five years. I don't know how I would have survived first year without them, let alone five. I feel extremely grateful to have been grouped with this particular collection of people. Thank you to Jenelys, Alexa, Amritha, Marie, Teri, Mo, Connor, Jonathan, and Asad. I am excited to see all of your success in the future! A special shoutout to Jenelys and Teri, for the emotional support and good times exploring Long Island, NYC, DC, and Japan!

I would also like to thank my past scientific mentors at Queen's University and the Sainsbury Laboratory. They introduced me to scientific research, and the wonderful world of plant biology. I am particularly grateful to Jacqueline Monaghan for mentoring me through the experience of working in my first research lab. I truly feel that the course of my scientific career would have been entirely different had I not joined her lab at Queen's.

Lastly, I have to thank my family and friends, whose love and support has been there since the beginning. Thank you to Kat and Anna, my childhood friends who have stayed in touch despite the large distances constantly separating us. And lastly, to my most constant supporters, in the good times and the bad, I extend the deepest gratitude to my parents, Allison and Ron, and my sister Alex. They have always made me feel as if I could do anything I set my mind to, even if I didn't believe it myself. For them I will always be deeply thankful.

Table of Contents

List of Abbreviations

3C	chromosome conformation capture
ABE	adenine base editor
ACR	accessible chromatin region
AG	AGAMOUS
AN	ANANTHA
AP1	APETALA1
ARF	auxin response factor
ATAC-seq	assay for transposase-accessible chromatin with sequencing
AtCLV3	*Arabidopsis thaliana* CLAVATA3
CETS	CENTRORADIALIS/TERMINAL FLOWER 1/SELF-PRUNING
CDF1	CYCLING DOF FACTOR 1
ChIA-PET	chromatin interaction analysis with paired-end tag sequencing
ChIP-seq	chromatin immunoprecipitation with sequencing
CLV1	CLAVATA1
CLV3	CLAVATA3
CNE	conserved non-coding element
CNS	conserved non-coding sequence
CO	CONSTANS
CRE	*cis*-regulatory element
CRISPR	clustered regularly interspaced short palindromic repeats
CRM	*cis*-regulatory module
CTCF	CCCTC-binding factor
CZ	central zone
DAP-seq	DNA affinity purification and sequencing
DPE	downstream promoter element
DNase-seq	DNase I hypersensitive sites sequencing
EMSA	electrophoretic mobility shift assay
eve	*even-skipped*
FA	FALSIFLORA
FIMO	find individual motif occurrences
FLC	FLOWERING LOCUS C
FT	FLOWERING LOCUS T

FTL1	FT-LIKE 1
FWA	FLOWERING WAGENINGEN
gRNA	guide RNA
GUS	β-glucuronidase
GWAS	genome-wide association study
HAT1	Histone acetyltransferase 1
HDR	homology-directed repair
Hi-C	genome-wide chromosome conformation capture
IL	introgression line
INTACT	isolation of nuclei tagged in specific cell types
LFY	LEAFY
LRR-RLK	leucine-rich repeat receptor-like kinase
MEME	multiple em for motif elicitation
MNase-seq	micrococcal nuclease digestion with deep sequencing
MOA-seq	MNase-defined cistrome-occupancy analysis with sequencing
MTE	motif ten element
MY	million years
n	sample number
ns	not significant
PAM	protospacer-adjacent motif
pegRNA	prime editing guide RNA
PEPB	phosphatidylethanolamine-binding protein
PFM	position frequency matrix
QTL	quantitative trait loci
RdDM	RNA-directed DNA methylation
SAM	shoot apical meristem
sd	standard deviation
SEM	scanning electron microscope
SFT	SINGLE FLOWER TRUSS
shh	*sonic hedgehog*
SlCLV3	*Solanum lycopersicum* CLAVATA3
SlSP5G	*Solanum lycopersicum* SELF PRUNING 5G
SlWUS	*Solanum lycopersicum* WUSCHEL
SNP	single nucleotide polymorphism

SP5G	SELF PRUNING 5G
SpSP5G	*Solanum pennellii* SELF PRUNING 5G
STARR-seq	self-transcribing active regulatory region sequencing
STM	SHOOT MERISTEMLESS
svb	*shavenbaby*
SVP	SHORT VEGETATIVE PHASE
TAD	topologically associating domain
TALEN	transcription activator-like effector nucleases
TE	transposable element
TET1	Tet methylcytosine dioxygenase 1
TF	transcription factor
TFBS	transcription factor binding site
TFL1	TERMINAL FLOWER 1
TM3	TOMATO MADS-BOX GENE 3
TMF	TERMINATING FLOWER
TSS	transcription start site
UTR	untranslated region
WOX9	WUSCHEL HOMEOBOX9
WT	wild type
WUS	WUSCHEL
Y1H	yeast one-hybrid

List of Figures and Tables

Chapter 1: Introduction

1.1 Overview

From a single cell, and an identical DNA code faithfully copied, multicellular organisms develop into complex amalgamations of hundreds of unique cell types. This presents an apparent conundrum: how, from a static string of nucleotide bases, are complex multicellular organisms formed? While the DNA sequence may be primarily unchanged from cell to cell, the chromatin (DNA and its associations with shaping proteins and RNA) is constantly undergoing dynamically orchestrated processes of remodeling. This association between DNA, proteins, and RNA is altered through precise mechanisms that facilitate the reliable development of complex multicellular forms, often without error. Furthermore, all of these factors are in a constant flux throughout evolutionary time, as the sequence and structure of the chromatin is altered in beneficial and non-beneficial ways, and selected upon to generate the diversity of species on our planet today. Thus, the key to understanding development and growth stems from an in depth study of non-coding DNA sequences, as well as their constantly evolving physical forms and associations.

During the course of this thesis research, I have endeavored to uncover key truths about this process, referred to broadly as gene regulation. Unique cell types are formed through the differential expression of genes. Gene expression (transcription) is coordinated by combinations of *cis*-regulatory elements (CREs), the accessibility of which are carefully regulated in time and space by factors controlling chromatin structure. CREs are short DNA sequences that are bound by specific transcription factors, and are often near or within the genes they regulate (although not always). Multiple CREs work together to regulate the transcription of a gene in *cis*-regulatory modules (CRMs), or they may also have separate roles in specific tissues or developmental time points to coordinate the pleotropic roles of a gene.

The main aim of this thesis research was to explore the *cis*-regulatory control of conserved and divergent phenotypes in an evolutionary context. To this end, we chose to study CREs and their

interactions involved in the regulation of two dosage sensitive developmental genes: *CLAVATA3* and *SELF PRUNING 5G. CLV3* is a highly conserved plant stem cell regulator, and we explored how *cis*-regulatory control of this gene evolved between *Arabidopsis thaliana* and tomato over a divergence time of ~125 million years (MYs). In contrast, phenotypic variation for flowering time between wild and domesticated species of tomato was derived from differential expression of *SP5G,* and we explored potential *cis*-regulatory determinants of this variation. We were curious about what we could reveal about the nature of *cis*-regulatory control – for example, how stringent are regulatory regions in their sequence, grammatical organization, interactions, and chromatin structure? Are they robust to the mutations introduced throughout evolutionary time, or are small changes often disruptive enough to have a selectable consequence on fitness? We approached these questions using functional genetics, revealing the phenotypic relevance of CREs to gene regulation via *in vivo* mutagenesis of CREs in their native context. Within this introduction I will outline what has already been learned about CREs in animal and plant model systems, and how we hope to expand upon this knowledge.

1.2 Defining the components of *cis*-regulation

The DNA elements typically involved in eukaryotic *cis*-regulatory regions include the core promoter, enhancer elements, silencer elements, and insulators. I will briefly describe the main characteristics of each. In our own research, we are interested in how core promoters, enhancers, and silencers interact genetically and physically in plants to regulate gene expression.

1.2.1 Core promoters

The core promoter of a gene typically spans ~50 bp on either side of the transcription start site (TSS), and encompasses the minimal sequence needed to direct initiation of transcription. Motifs within the core promoter are bound by general transcription factors (TFs), which recruit RNA polymerase II to form the preinitiation complex and initiate basal transcription (Atkinson and Halfon 2014). There are both general and specific motifs found within core promoters. For example, the TATA box (about ~35 bp upstream of TSSs) is shared among eukaryotes. However, some core

promoter motifs are specific to particular species or kingdoms. Some core motifs found in plant promoters include the Y patch, CCAAT box, BRE elements, and an initiator region (Inr), while Motif Ten Elements (MTEs) and Downstream Promoter Elements (DPE) have been extensively studied in human and *Drosophila* promoters (Zhong et al. 2023). Some core promoters may not have any characteristic motifs. When active, core promoters are generally accessible and non-methylated, with histone modifications such as acetylation and H3K4me3, and histone variant H2A.Z (Schmitz et al. 2021).

1.2.2 Enhancers

In addition to these core promoter elements, transcription factor binding sites (TFBSs) are crucial to transcription by recruiting gene-specific transcription factors. Enhancers are *cis*-regulatory elements that increase the rate of transcription when bound by transcription factors. They are often responsible for coordinating the transcription of a gene at the proper time and cell type during development. While core promoter elements are generally in close proximity to the TSS, enhancers can be found both proximal and distal to the TSS, as well as within the gene itself (Marand et al. 2023). In animals, active and repressed enhancers are marked by characteristic histone modifications: for example, H3K27ac and H3K4me1, and H3K27me3, respectively (Calo and Wysocka 2013). The enhancers of animals also sometimes produce transcripts, called enhancer RNAs (Schmitz et al. 2021). However, their existence does not seem to be widespread in plants.

1.2.3 Silencers

Silencers are sequences that bind TFs to decrease gene expression. Accessible regions enriched with H3K27me3 are more likely to be silencers, and recruit Polycomb repressive complex components (Marand et al. 2023). However, in animals various histone modifications have been associated with silencers, making it hard to pinpoint a single characteristic mark. It should also be noted that despite these general definitions, the situation can be more complicated, with certain elements acting as enhancers or silencers depending on the context. For example, the transcription

factor WUSCHEL can be an activator or repressor depending on its concentration in the cell, making the CREs that it binds to contextual enhancers or silencers (Perales et al. 2016).

1.2.4 Insulators

Insulators are elements that prevent promoter-enhancer interactions when bound by certain proteins. While there are examples in animals, none have been validated in plants yet. In animals, CCCTC-binding factor (CTCF) binds specific sequences to establish distinct chromatin domains (known as topologically associating domains or TADs) (Atkinson and Halfon 2014). Only CREs and genes within the same domain can interact. Plants lack a CTCF homolog, even though large genome plant species do organize their genomes into TAD-like domains (Dong et al. 2017). It is possible that another, non-homologous TF was co-opted for the role in plants, and its discovery would be the first evidence of insulator activity in plant gene regulation (Kurbidaeva and Purugganan 2021).

1.2.5 Post-transcriptional mechanisms of gene regulation

Whilst this thesis will generally focus on *cis*-regulatory elements that affect transcription rate, it should also be noted that multiple mechanisms of post-transcriptional control exist to modulate transcript or protein abundance. Many 3' UTRs in animals and plants contain motifs that influence mRNA stability, localization, translation, and polyadenylation (Mayr 2019). Alternative splicing can generate different mRNAs from the same gene, leading to various protein isoforms with different functions or localizations. Non-coding RNAs such as microRNAs and small interfering RNAs can bind mRNA molecules and prevent their translation. Finally, the rate of protein turnover is impacted by various mechanisms, such as polyubiquitination which targets proteins for degradation by the 26S proteosome (Deribe et al. 2010).

1.3 Epigenetic regulation of gene expression

Epigenetic forces shaping the genome are key to cell type specific gene regulation, and include changes to chromatin accessibility, DNA methylation, histone modifications, and 3D

chromatin conformation. All of these features can be used to predict CREs and follow the dynamic nature of gene expression during development.

1.3.1 Chromatin accessibility

There is a strong association between chromatin accessibility and the ability of a DNA sequence to be a CRE. Open regions of DNA are accessible to TFs, making them more likely to be CREs than closed DNA (both enhancers and silencers). In humans, accessible chromatin regions make up just 2-3% of the genome, but contain 94% of all ENCODE TFBSs (Thurman et al. 2012). Therefore, whole-genome assays for chromatin accessibility in specific tissues reveal potential regions of importance for gene regulation. DNase-seq, MNase-seq, and ATAC-seq have been key methods in the identification of open chromatin. DNase I hypersensitive sites sequencing (DNase-seq) pairs DNase I digestion with high-throughput sequencing to profile chromatin accessibility (Boyle et al. 2008). DNase I preferentially digests regions devoid of nucleosomes. These digested fragments can then be sequenced to reveal accessible regions of the genome. This technique is also useful for mapping TF footprints, which manifest as short regions of uncleaved DNA within larger accessible regions. In plants, studies of chromatin accessibility have helped track gene regulation over multiple developmental stages. For example, in *Arabidopsis* DNase-seq was used to follow the dynamics of chromatin accessibility at multiple stages of flower development (Pajoro et al. 2014). Chromatin accessibility, gene expression, and binding of MADS-box proteins were evaluated at four stages, revealing a correlation between changes in TF binding and changes in gene expression, as well as dynamic remodeling of chromatin accessibility over time to orchestrate binding of specific regulators at key moments in development. A similar technique, micrococcal nuclease digestion with deep sequencing (MNase-seq), can be used to reveal accessible chromatin regions (ACRs) and even nucleosome positioning at finer scale (Pajoro et al. 2018). MNase-defined cistrome occupancy analysis (MOA-seq) is used to determine TF binding footprints through the recovery of short fragments (Savadel et al. 2021). The assay for transposase-accessible chromatin with sequencing

(ATAC-seq) is an easier and lower-input method to profile chromatin accessibility (Buenrostro et al. 2013). It utilizes the Tn5 transposase, which inserts sequencing adapters into accessible regions. An ATAC-seq study of chromatin accessibility in 13 angiosperm species revealed an association between increasing genome size and the number of distal ACRs in 13 angiosperm species (Lu et al. 2019). All of these assays have been used successfully to profile chromatin accessibility in both animals and plants, however the production of cell type specific ACR maps has been difficult until recently.

Since TF footprints are known to be highly tissue specific, tissue specific ACR maps are vital to the detection of functional CREs. One approach used to isolate nuclei from specific cell types of *Arabidopsis* involved combining INTACT (isolation of nuclei tagged in specific cell types) with ATAC-seq. While INTACT is an effective technique for nuclei isolation in plants, it does require the generation of stable transgenic lines for the cell type of interest, which limits its ease of execution. Regions of accessibility were characterized in *Arabidopsis* roots through INTACT+ATAC-seq (Tannenbaum et al. 2018). Root specific ACRs were found to correlate with root specific expression patterns of certain genes, and they were enriched for motifs of TFs known to be important to root development. Furthermore, half of ACRs were located outside of the promoter (in the intergenic space), suggesting the increased importance of enhancers/silencers in orchestrating dynamic, tissue-specific developmental processes. More recently, single cell RNA-seq and ATAC-seq have facilitated the creation of cell-type specific expression and ACR maps from bulk tissue samples (Marand and Schmitz 2022). In droplet based single cell methods, a microfluidic chip promotes the formation of oil droplets with single nuclei attached to a barcoded bead. Inside this droplet they undergo the initial steps of RNA-seq or ATAC-seq (incubation with Tn5 and NGS adapters), library amplification, and barcoded primers are added. Individual cellular profiles of gene expression and chromatin accessibility can be distinguished, since each cell is uniquely barcoded. Single-cell ATAC-seq was also used to profile the accessibility of different cell types within the root tip, revealing variance in accessibility within different cell types of the same organ (Feng et al. 2022).

Therefore, studies of chromatin accessibility have revealed multiple trends in the dynamic orchestration of gene regulation. However, although ACRs can be used to predict CREs, they cannot predict the magnitude of their contribution to gene expression, or their relevance to the phenotypic expression of a trait. Furthermore, accessibility is a key feature of both enhancers and silencers. Therefore, other epigenetic and functional assays are needed to fully characterize the specific role of a CRE in gene regulation.

1.3.2 DNA methylation

DNA methylation is a key epigenetic mechanism used to control gene expression. Methylation of DNA within promoters inhibits transcription by blocking the binding of transcription factors and RNA polymerase, as well as physically condensing the chromatin. Therefore, active regulatory regions are associated with lower methylation patterns in all sequence contexts (CG, CHG, CHH, where H=A, T, or C), while most intergenic regions are highly methylated (Lloyd and Lister 2022). However, GC methylation within the gene body itself is actually associated with active gene expression (Bewick and Schmitz 2017). Bisulfate sequencing and nanopore sequencing can both map 5mC methylation. Patterns of DNA methylation are a proposed approach to identify functional non-coding regions regardless of tissue type. The maize genome contains many short unmethylated regions shared among distinct tissue types, in contrast to chromatin accessibility which is cell type specific (Crisp et al. 2020). However, a subsection of DNA methylation patterns are cell type specific, such as those within gametes, and are modified by chromatin remodeling factors such as the CLASSY family in *Arabidopsis* (Zhou et al. 2022).

1.3.3 Histone modifications

Whilst the genome can generally be categorized into heterochromatin (filled with repeats and transposons) and euchromatin (filled with transcribed genes), the reality is often more complex than that, with certain regions changing their chromatin state in response to developmental stage or cell type, as well an environmental stimuli. The DNA is wrapped tightly around histone octamers (called

nucleosomes), which can undergo chemical modifications such as methylation and acetylation in order to alter their association with DNA, and thus chromatin accessibility. For example, histone acetylation generally results in open chromatin due to reduced binding affinity between the histones and DNA. Conversely, many types of histone methylation are generally associated with a closed chromatin state, and attracting Polycomb group proteins (responsible for remodeling chromatin to promote gene silencing) (Schmitz et al. 2021).

Active enhancers are typically accessible with acetylated histones, and repressed enhancers have inaccessible chromatin and H3K27me3, while poised enhancers contain features of both (Zhang et al. 2015). Repressed, poised, and active enhancers are well known to be associated with specific histone marks in animals, however the case is less clear in plants. Active enhancers in plants have been found to be associated with the histone variant H2A.Z (Lu et al. 2019; Ricci et al. 2019). However, most unmethylated, accessible regions in plants lack most of the known histone modifications, or they are associated with many different histone modifications, none of which characterize the majority (Ricci et al. 2019). Interestingly, H3K4me1, which is found in the majority of active mammalian enhancers, is not significantly associated with ACRs in plants (Lu et al. 2019). These findings indicate that there is still much unknown about the epigenetic features that allow sequences to behave as CREs in plants.

1.3.4 3D chromatin conformation

The study of the 3D structure of chromatin has been a vital area of research illuminating important principles of gene regulation in eukaryotes. The genomes of animals and plants are folded into reproducible 3D structures that have consequences for gene expression. Studies of eukaryotic genomes have revealed the presence of conserved topologically associating domains (TADs), the boundaries of which are distinguished by particular proteins in animals (Schoenfelder and Fraser 2019). TADs are defined as genomic regions that have greater contact frequencies with one another than sequences in other domains. TADs with similar properties can be grouped into active “A”

compartments and repressive “B” compartments based on epigenetic features and gene expression. TADs are proposed to facilitate gene expression by bringing multiple promoters into closer contact with long-range regulatory elements. At a finer scale, genomes are also defined by numerous looping interactions – long range interactions, often between promoters and distal elements, that may or may not come into direct contact to control gene expression (Dong et al. 2017; Schoenfelder and Fraser 2019). Small scale promoter-downstream looping may impact gene expression by promoting/inhibiting transcription factor or RNA polymerase recruitment, recycling RNA polymerase, suppressing bidirectional transcription, terminating transcription, and promoting gene body enhancement of transcription (Liu et al. 2016).

In plants, chromatin architecture does play a role in gene regulation, although this role is less defined than in mammals. Both short and long range chromatin looping is prevalent among plant species. Notably, *Arabidopsis* does not seem to form TADs – due to its small genome size most CREs are gene proximal (Feng et al. 2014). However, chromatin looping is still highly prevalent, especially across gene bodies, and there are several documented cases (Liu et al. 2016). The *Arabidopsis* flowering gene *FLOWERING LOCUS C* (*FLC*) has been highly studied for its unique mechanism of transcriptional regulation following cold induced flowering (vernalization). A promoter-downstream loop is formed across the *FLC* gene, and is specifically disrupted following vernalization (Crevillén et al. 2013). This disruption is followed by transcription of the antisense RNA *COOLAIR* and Polycomb-mediated repression of the locus (Rosa et al. 2016). This example highlights the importance of chromatin looping for gene regulation, as well as the potential for multiple functions of downstream regulatory elements (i.e. as antisense transcription initiation sites, and promoter looping anchors). Additionally, a loop between the transcription start site and downstream region of the flowering repressor *TERMINAL FLOWER 1 (TFL1)* dissociates upon binding of a MADS-box transcription factor complex downstream, leading to reduced expression (Kaufmann et al. 2010).

In contrast to *Arabidopsis*, several crop species are known to have complex 3D structures. As plant genomes increase in size, and CREs are pushed further from the genes they regulate, these

plants likely had to develop strategies to physically connect CREs and their genes. One Hi-C study found that the crop species tomato, maize, sorghum, rice and foxtail millet have complex 3D genome arrangements characterized by compartments, domains, and loops, similar to mammals and unlike *Arabidopsis* (Dong et al. 2017). Specifically, tomato and maize demonstrate extensive chromatin looping among gene islands, suggesting the potential for numerous distal regulatory elements. One of the first examples of a distal enhancer element controlling gene expression in crops was that of the *b1* gene in maize, where a loop forms between the gene body and an enhancer 100 kb upstream to mediate high gene expression (Louwers et al. 2009). Other examples of long range enhancers have since been found in maize, such as the DICE enhancer 140 kbp upstream of the enzyme *Bx1*, and the KRN4 enhancer 60 kbp downstream of *UB3* (Zheng et al. 2015; Du et al. 2020). In rice, an enhancer 10 kb upstream of *LG1* regulates the closed panicle phenotype (Zhu et al. 2013). More research is required to reveal the extent of these long range interactions among the plant kingdom, and the mechanisms involved in their establishment.

It can often be difficult to pair distal CREs with their target genes, since CREs can skip over multiple genes to interact with a gene much farther away. In maize, 40% of distal CREs actually skip over closer genes in order to contact more distal genes (Li et al. 2019). As such, a better approach to enhancer prediction in plants with large genomes is likely chromatin accessibility assays and conservation analysis combined with some variation of chromosome conformation capture (3C) technology. Hi-C is an unbiased variation of 3C that captures all chromatin-chromatin interactions genome-wide occurring in a specific tissue type, and at a specific developmental period (Lieberman-Aiden et al. 2009). Hi-C identifies genome wide long range interactions, albeit at low resolution. Higher resolution derivatives such as capture Hi-C and ChIA-PET can provide higher resolution data by enriching for specific promoters or histone modifications associated with gene regulation (ex. H2K27ac, H3K27me3, RNApolII) (Li et al. 2019). Such an approach was recently used to reveal numerous long range CREs in the maize genome (Ricci et al. 2019). Distal ACRs in maize are

enriched for transcription factor binding sites, gene loops, intergenic quantitative trait loci (QTL), and enhancer activities, suggesting that at least a portion of these sites are involved in gene regulation.

1.4 Validating CREs and transcription factor binding

While certain epigenetic and sequence features of a DNA element can help researchers predict the presence of an enhancer or silencer, they do not guarantee its functionality in controlling gene expression. For this purpose, a number of assays have been developed to verify both CRE activity, and the particular TF(s) interacting with a CRE.

1.4.1 Validating *cis*-regulatory elements

Several assays have been developed to verify the activity of predicted CREs. Reporter assays can be useful to determine both the level and location of gene expression. A predicted CRE placed upstream of a minimal promoter may be used to drive the expression of luciferase or GFP to validate enhancer ability, or to drive the expression of B-glucuronidase (GUS) *in vivo* to determine expression domains (Weber et al. 2016). Many enhancers can be tested at once using the genome wide assay: self-transcribing active regulatory regions sequencing (STARR-seq). STARR-seq uses a reporter library, with various genomic sequences cloned downstream of a reporter gene containing a minimal promoter (Arnold et al. 2013). Transcript abundance can be measured as a proxy for enhancer activity. STARR-seq has been used on few plant species to date, including rice, maize, and tobacco (Ricci et al. 2019; Sun et al. 2019; Jores et al. 2020).

Reporter assays have contributed valuable knowledge about the function of CREs, however they are only a proxy for CRE activity, without truly validating it. Most reporter assays do not place CREs in their native environment, so false negatives and positives can occur frequently. Even reporter assays *in vivo* have traditionally inserted transgenes into random regions of the genome. Furthermore, isolating CREs from their native context ignores the complex interactions among multiple CREs in the regulation of a gene, including distal CREs that interact with promoters through physical looping of the DNA. We also do not yet fully understand how the activity of a CRE and gene expression itself

correlate with phenotypic effect, which is likely complicated by the specific tissue and time during development that the CRE is utilized. For our research, we used CRISPR-Cas9 genome editing to study CREs in their native context, thus providing a direct measure of their phenotypic effect.

1.4.2 Identifying the *trans*-factors in gene regulation

Also of interest to the study of gene regulation is the identification of the *trans* factors that bind CREs in order to bring about changes in gene expression – namely, the transcription factors. General TFs are responsible for initiating gene expression, while cell type specific TFs bind CREs to fully activate certain genes only in particular developmental contexts. While we have multiple indicators to predict which sequences may constitute CREs (accessibility, methylation, 3C assays, reporters, conservation, etc.), predicting the TFs that they bind to can be more challenging. Multiple factors dictate whether a specific TF will bind a DNA sequence. Firstly, TFs show a preference for specific DNA motifs, often 5-11 bp long. However, this is only part of the story, since TF motifs can be found numerous times throughout the genome, and the majority of them are unbound (~99%) (Hajheidari and Huang 2022). Sequence context will therefore play a huge role in whether a motif will translate to a bona fide transcription factor binding site, especially various epigenetic features of the corresponding chromatin.

The main assays to map TFBSs genome-wide are chromatin immunoprecipitation sequencing (ChIP-seq) and DNA affinity purification sequencing (DAP-seq). ChIP-seq captures stable DNA-protein interactions through sequencing of DNA fragments attached to the TF of interest, which is isolated via a TF-specific antibody (Johnson et al. 2007). Large-scale ChIP-seq experiments can be used to map transcriptional regulatory networks – for example, ChIP-seq of 104 TFs expressed in the maize leaf was used to trace these networks (Tu et al. 2020). DAP-seq eliminates the requirement for a TF-specific antibody by tagging the TF of interest and expressing it *in vitro.* The tagged TF is incubated with a genomic DNA library *in vitro,* and bound genomic fragments are captured via tag-

specific beads. DAP-seq was first used to map the binding sites of 529 TFs in *Arabidopsis thaliana* (O'Malley et al. 2016).

Alternatively, if you have a CRE of interest and you want to know which TF's bind to it, there are also a number of low throughput experimental methods. For example, a yeast one-hybrid (Y1H) or reporter assay. Y1H is a crude approach to test the binding of many TFs to a specific CRE. Of course, you need a TF library for the species of interest, and TF binding in yeast will have many differences to binding in the native context. To test candidate TFs, dual luciferase assays in *Nicotiana benthamiana* can be used to validate an interaction between a specific CRE and TF of interest (McNabb et al. 2005). Lastly, a classical assay for the detection of TF-DNA binding is electrophoretic mobility shift assay (EMSA). In this assay, TF-DNA complexes will migrate more slowly on a gel than free DNA (Hellman and Fried 2007).

Although these assays can potentially predict functional CREs, again they are prone to false positives and negatives. For example, the quality of ChIP-seq data fundamentally relies on the quality of the antibody used, while DAP-seq, Y1H, and dual luciferase assays survey TF binding completely out of context. Regardless, these experiments are still vital to understanding the mechanism of regulation of a gene, along with the CREs.

1.5 Where CREs are located

Understanding the regulatory grammar, including the relative location of CREs with respect to the genes they regulate, is crucial for identifying CREs and establishing strategies for genetic engineering and synthetic promoter design. In plants and animals, CREs are located in every possible genomic context – within regions proximally upstream and downstream of the gene, as well as within introns, exons, 5'UTRs, 3'UTRs, and at distal sites that mediate their effects through long range loops. Since chromatin accessibility is a main indicator of CREs, it is generally used to describe trends in CRE location. For example, in tomato meristem tissue, 53% of ATAC-seq peaks are located within 5 kb upstream of genes, ~6% within 3 kb downstream, ~20% are distal intergenic (>3 kb

away), and the remaining ~21% are split amongst introns, exons, and UTRs (Hendelman et al. 2021). Conserved regions have a very similar distribution.

Examples of CREs in all of these sequence contexts have been documented. While proximal regions upstream are thought to harbor the majority of CREs, the first intron of many plant genes can be quite large, and often harbors CREs with important contributions to gene expression. For example, the first intron of *FLC* and *AG* in *Arabidopsis,* and *KN1* in maize, all possess validated CREs (Greene et al. 1994; Sieburth and Meyerowitz 1997; Qüesta et al. 2016). The regulation of meristem size in tomato involves both upstream and downstream CREs. A QTL downstream of the meristem promoting TF *WUSCHEL* weakly regulates tomato size, while CRISPR-Cas9 engineering of CREs proximally upstream of the negative regulator *CLV3* causes severely fasciated fruits (Rodríguez-Leal et al. 2017). In *Arabidopsis*, a downstream CRM is known to partially regulate *CLV3* expression (Perales et al. 2016). An enhancer located within the 3'UTR of tomato *SP5G* mediates high expression of this anti-florigen in wild tomato species (Zhang et al. 2018). Predicted enhancers are highly prevalent in the introns, 3'UTRs, and downstream regions of floral regulators in *Arabidopsis*, many of which were tested for activity in GUS reporter assays (Yan et al. 2019).

In addition to gene proximal CREs, it has long been understood that many distal CREs frequently regulate genes through long range chromatin interactions. In plants, the frequency of distal CREs has been associated with genome size – in general, the larger the plant genome, the more distal CREs can be found. A study of open chromatin in 13 angiosperm genomes of varying sizes confirmed this trend. For example, in the genomes of *Arabidopsis thaliana* (genome size 119 Mb) and *Spirodela polyrhiza* (143 Mb), distal ACRs (>2 kb from gene) in leaves are just 5.9% and 5.7% of total ACRs (Lu et al. 2019). In contrast, in the largest genomes assayed, maize (2124 Mb) and barley (4834 Mb), 32.8% and 45.9% of ACRs are distal. A proposed hypothesis for this trend is that transposon expansion in the regulatory regions of larger genomes pushed CREs further and further away over time (Ricci et al. 2019). There are multiple examples of distal CREs among plants, as discussed previously.

It is still unclear how important the relative position of a CRE to its gene is to its function. A few studies have attempted to investigate this in an evolutionary context. For example, between mosquitos and flour beetles (~333 MY diverged), five putative orthologous enhancers for the TF Dorsal are located in similar genomic positions, even though their sequence is not conserved (Cande et al. 2009a). In each species, an enhancer of *cactus* is conserved within an intron, a *brinker* enhancer is conserved within a neighboring gene, and a distal 5' enhancer of *twist* is positioned similarly in both species, despite sequence divergence. Similarly, in a study of *Drosophila* and sepsid *even skipped* enhancers, although sequence is not conserved, relative genomic positioning is (Hare et al. 2008). These findings suggest that evolution may favor modification of pre-existing enhancers, and enhancer position may be constrained, perhaps due to the ability to interact with the promoter only in certain positions. Furthermore, an enhancer driving similar patterns of *islet* expression in sponge, zebrafish, and mice is highly diverged in sequence, but it is located in a neighboring bystander gene in each species (Wong et al. 2020). It is possible that these enhancers are constrained partially by a need to not disrupt the function of the bystander gene. In contrast, enhancers of the *yellow* gene have different positions in various *Drosophila* species, except one enhancer whose position was conserved in an intron (Kalay and Wittkopp 2010). However, the *Drosophila* species examined do have some species specific pigment patterns. Thus, it is possible that more constrained developmental expression patterns require conserved enhancer positioning, like in the case of Dorsal enhancers, whereas genes with rapidly evolving expression patterns have less constrained enhancer architecture, in the case of *yellow.* Interestingly however, species of nymphalid butterflies that have evolved color pattern variations within their wings seem to have a conserved regulatory architecture at the *WntA* locus (a gene responsible for color patterning) (Mazo-Vargas et al. 2022). Based on tissue-specific ATAC-seq, orthologous CREs controlling expression in particular wing regions of nymphalid butterfly species are conserved in their position upstream of *WntA*, even though color patterning is highly diverse among them. Hi-C also demonstrated a conserved pattern of physical interactions among multiple CREs and the *WntA* promoter. This suggests a mode of evolutionary variation that favors

modification of pre-existing CREs, rather than gains and losses of CREs. Therefore, in circumstances of both conserved and diverged phenotypes, there is evidence of conserved CRE positioning.

During the course of this thesis work, we consider CREs in multiple genomic contexts, as well as their interactions in the control of gene regulation. We also contribute to this ongoing debate about the importance of relative genomic positioning in the evolution of *cis*-regulatory regions.

1.6 The evolution of *cis*-regulatory regions

The extent to which CREs are constrained throughout evolution remains a significant subject of debate, especially among genes with otherwise conserved coding sequence. While genes tend to have the greatest level of conservation between species, conserved non-coding sequences (CNSs) can still be detected in the genome, and generally overlap with known functional genomic features, such as accessible chromatin and TF motifs (Hendelman et al. 2021). Additionally, variation within CREs has been attributed to the generation of many divergent traits and phenotypes in the evolution of plants and animals. In this section we will discuss the evolution of CREs in both of these contexts, which we investigate further in our own research in future chapters.

1.6.1 Conserved non-coding elements and sequences

In addition to epigenetic features, non-coding sequences that have been conserved over evolutionary time are proposed predictors of CREs. Among metazoans, the deep conservation of non-coding regions is a strikingly prevalent characteristic. In animals, conserved non-coding elements (CNEs) are often quite long, and can be relatively distant from any genes. The standard definition of a CNE is a sequence over 100 bp long containing 70% sequence identity (Burgess and Freeling 2014). Metazoans also contain many examples of ultraconserved elements, which share 100% identity over 200 bp long across significant time spans of metazoan evolution. Placental mammals share 14,000 CNEs that are >100 bp long and 100% identical (Stephen et al. 2008). Some of these CNEs have confirmed roles in gene expression, such as the *sonic hedgehog* (*shh*) enhancer that controls limb development in vertebrates (Lettice et al. 2003). However, intriguingly the ability of these elements to

act as enhancers may not be guaranteed by their high conservation. For instance, only 50% of mammalian CNEs had enhancer activity in transgenic reporter assays (Visel et al. 2008). In addition, many ultraconserved elements do not have a phenotype when removed in mice, and many can maintain activity amidst multiple mutations (Ahituv et al. 2007; Snetkova et al. 2021). Thus, the role of conserved non-coding elements in gene regulation is an on-going debate.

Multiple studies suggest that specific non-coding regions of the genome are under purifying selection in plants as well. For example, a conserved non-coding sequence analysis among diverse members of the *Brassicaceae* uncovered 90,000 CNSs in this family (Haudry et al. 2013). Another study identified 155,268 CNSs within the *Solanaceae*, and found evidence of phenotypic effects when CNSs of the developmental gene *WOX9* were deleted in both tomato and groundcherry with CRISPR-Cas9 mutagenesis (Hendelman et al. 2021). On a larger evolutionary time scale, CNSs can be detected throughout eudicot flowering plants, with 35 deep CNSs shared by 10 angiosperm species spanning the entire phylum (Burgess and Freeling 2014). However, the incredibly high degree of conservation found among metazoans is rarely found among plant species. The CNSs that are detected in plants are often much shorter (15-150 bp), and are more likely to be located closer to genes (Burgess and Freeling 2014). Although the size of an individual TFBS is only 5-11 bp, CNSs can be longer if they harbor multiple TFBSs acting in CRMs (thus there may be a bias for detecting these kinds of CNSs, especially in plants). Within both animals and plants, the CNSs with highest conservation are often associated with transcription factors or other developmental genes, and are themselves enriched with known TFBSs (Burgess and Freeling 2014; Van de Velde et al. 2016).

1.6.2 Conservation of function despite *cis*-regulatory sequence divergence

Despite the discovery of many CNSs within plant families, as well as the striking prevalence of CNEs among metazoans, the non-coding sequence of conserved genes in distantly related organisms is more often highly diverged. How conserved genes maintain the proper expression patterns needed to carry out their function amidst this regulatory sequence turnover is a question of

interest. Conserved non-coding sequences are determined from the alignment of the regulatory regions of orthologous genes. CREs are often only 5-11 bp long, which is extremely difficult to detect amongst the high sequence turnover of surrounding regions. This is one reason the regulatory regions of distantly related species (such as those from different families) are often impossible to align using current methods. This feature makes it very difficult to determine the extent to which CREs and their organization are constrained over evolutionary time. There are of course some examples of highly conserved non-coding elements across vast evolutionary distances, such as multiple CNEs in animals that regulate tightly controlled developmental processes (Kvon et al. 2016). However, these occurrences of deep conservation of non-coding sequences might actually be the exception, rather than the rule. Understanding the grammar of gene regulation and its dynamic evolution is a vital element to a better understanding of this evolutionary conundrum.

Enhancer grammar is defined by TFBS type, affinity, number, spacing, orientation, order, and local DNA shape (Long et al. 2016). Exactly how flexible these factors can be and still mediate proper gene expression is still largely unknown. Several different models have been proposed to describe how TFBSs interact to form enhancer architecture. The **enhanceosome model** suggests that enhancer grammar must remain quite rigid for proper enhancer function (Arnosti and Kulkarni 2005). This model may be more likely in cases in which direct cooperativity among multiple TFs is vital for gene expression. For example, many TFs have specific binding partners that impact binding specificity and affinity. MADS-box TFs form heterodimers and heterotetramers to bind to their target sequences (Lai et al. 2020). Sometimes multiple binding motifs in close proximity are required for proper TF complex formation. ARF (Auxin Response Factor) TFs are also known to depend on specific spacing, direction, and order of motifs for proper ARF binding specificity and affinity (Freire-Rios et al. 2020). Enhancers following the enhanceosome model are also more likely to be highly conserved. Highly conserved non-coding sequences tend to regulate developmental genes more often, perhaps suggesting the importance of regulatory grammar in scenarios requiring switch-like transcriptional activation (Long et al. 2016). In contrast, the **billboard model** suggests that while

the specific TFs are important, their grammar (spacing, order, orientation, etc.) is more flexible for proper enhancer function (Arnosti and Kulkarni 2005). The TF collective model also suggests motif organization can be more flexible, since some TFs are known to be recruited to DNA through interactions with other proteins. Studies of enhancer conservation among multiple species generally supports models of flexible regulatory grammar, however most enhancers likely contain a mixture of these models, with some motifs being flexible and some more rigid (for example, those involved in direct cooperative interactions).

There are many examples in the animal kingdom that seem to lend support to the billboard model of enhancer architecture, providing an explanation for conservation of gene expression amongst regulatory sequence divergence. For example, a study of the sea urchin species *Strongylocentrotus purpuratus* and *Lytechinus variegatus* (diverged ~50 MY) was conducted, looking at eight orthologous regulatory regions with previously validated TFBSs (Cameron and Davidson 2009). For each set of orthologous regulatory regions, the sequences shared similar TFBSs, but number, position, and orientation was often different, with the exception of adjacent motifs that were often conserved in their spacing and position (likely as a result of cooperativity). In another study, the enhancer sequences of the developmental patterning gene *even-skipped* (*eve*) were found to be highly diverged between sepsids and *Drosophila* (diverged ~100 MY), yet sepsid enhancers were capable of driving a near-identical expression pattern of *eve* when transformed into *Drosophila* embryos (Hare et al. 2008). These non-coding regions do share very short conserved sequences between 20-30 bp containing transcription factor motifs, suggesting that while short TFBSs may be deeply conserved, regulatory grammar and organization could be more malleable to change. Similar conclusions were derived from a recent study of a metazoan gene conserved among sponge, zebrafish, mice, and humans over a span of nearly 700 million years (Wong et al. 2020). The enhancer sequence driving *islet* gene expression is highly diverged among all of these species. Despite this, the sponge *islet* enhancer can drive a reporter gene in both zebrafish and mouse embryos with a similar expression pattern to native zebrafish/mouse *islet*. Again, although the enhancer sequences were highly diverged,

they did share some TF motifs, although the composition and frequency of those motifs was highly variable. In plants, a similar finding was discovered for tomato and *Arabidopsis WOX9* (diverged ~125 MYs). While the pleiotropic functions of this gene are deeply conserved, non-coding sequence is not, apart from very short sequences (10-30 bp) with entirely altered grouping, order, and orientation (Hendelman et al. 2021). More studies of *cis*-regulatory sequence divergence among plant species would provide valuable insight into the universality of various models of enhancer architecture amidst evolution, and this is precisely what we will explore in chapter 2.

In many of these examples, orthologous CREs are often compared through means of reporter assays. For example, by introducing the two orthologous CREs into the same organism (to ensure an identical *trans*-factor environment), the differences in expression of the reporter gene should be *cis*-regulatory in nature. However, it is still possible that a *trans*-factor regulating the CREs has diverged between the species, such that the same reporter gene will be expressed differently in a native versus heterologous environment. Overall, reporter assays may be able to detect differences in *cis*-regulatory activity, but the actual phenotypic significance of these differences cannot be derived through these means. A change in gene expression is not proof for phenotypic divergence (Wittkopp and Kalay 2012). For example, many genes are not dosage sensitive, and are able to function perfectly well within a large range of expression levels. Furthermore, CREs that cause differences in expression domains also cannot be guaranteed to have a phenotypic effect, since cells in that domain may not be poised to react to its presence. Thus, new approaches are needed to explore this age-old question, and *in vivo* editing with CRISPR-Cas9 may be just the tool for it, since mutations in orthologous CREs of distantly related species can be compared through direct phenotypic readouts.

1.6.3 CRE variation as a mechanism of phenotypic divergence

CRE variation has long been understood to be an important mechanism of phenotypic divergence during the course of evolution. Many QTL map to non-coding regions in plants and animals (Wittkopp and Kalay 2012; Meyer and Purugganan 2013; Albert and Kruglyak 2015; Han et

al. 2018b; Ricci et al. 2019). CRE manipulation is a more subtle method of fine-tuning gene expression, and often has less drastic fitness effects than null mutations. CRE variation has been pivotal to the development of many key features in crop species during the course of domestication. For example, the establishment of apical dominance in maize was facilitated by a transposable element insertion in a CRE upstream of the *b1* gene (Louwers et al. 2009). Notably, the enlargement of tomato fruit size was facilitated by a synergistic interaction between mutations in upstream and downstream CREs of two genes controlling meristem proliferation (Rodríguez-Leal et al. 2017). A non-coding variant between *indica* and *japonica* rice is responsible for differential grain number per panicle in each variety, and a SNP in the promoter of *qSH1* led to loss of seed shattering during rice domestication (Konishi et al. 2006; Wu et al. 2021).

Similar findings hold true for other kingdoms of life. Multiple human disease states can be attributed to non-coding mutations. For example, polydactyly is caused by a mutation in an enhancer element of *shh*, which is subsequently mis-expressed in the developing limb bud (Lettice et al. 2003). Sequencing the human and chimpanzee genomes resulted in a surprising revelation: humans and chimps share ~99% of their DNA, with very high conservation of the protein-coding genes. Therefore, morphological divergence between closely related species is hypothesized to result from differences in the CREs that alter the spatiotemporal expression of genes, rather than the genes themselves. A landmark study seeking to explain differences in craniofacial morphology between humans and chimps found that many enhancers of neural crest cell genes have distinct, divergent sequence features between humans and chimps (Prescott et al. 2015). Orthologous enhancers in these cells also often have variable epigenomic features (such as presence/absence of H3K27ac) and TF motifs. Importantly, many of these species-biased enhancers could be correlated with gene expression differences between the two species. These examples serve as compelling evidence that even subtle changes in CREs can have substantial impacts on phenotypic divergence, driving key evolutionary adaptations.

1.6.4 Mechanisms of CRE variation

Understanding the various ways in which variation within CREs results in phenotypic divergence has the potential to reveal fundamental properties of regulatory regions themselves, as well as the evolutionary process. CREs can conceivably be created or destroyed through multiple different mechanisms, such as SNPs, deletions, insertions, transposable elements (TEs), promoter switching, co-option, duplications, or simply *de novo* generation from non-regulatory sequence. Examples of all are prevalent in nature.

One of the simplest mechanisms of CRE variation is the introduction of SNPs. Small substitutions, deletions, or insertions can alter TF motifs and subsequently TF binding affinity. For example, loss of trichomes in *Drosophila sechellia* compared to *Drosophila simulans* is associated with 13 single nucleotide substitutions within the 1007 bp enhancer upstream of the *shavenbaby* gene (Frankel et al. 2010; Wittkopp and Kalay 2012). A follow-up study discovered that activator loss, combined with evolution of a TFBS for a domain-specific repressor, eliminated *shavenbaby* expression in *Drosophila sechellia* (Preger-Ben Noon et al. 2016). In the pepper species *Capsicum Chinense,* its extremely hot flavor relative to other pepper species is caused by a SNP in the promoter of *MYB31*, which causes stronger binding of the TF WRKY9 to the region, increasing expression and the biosynthesis of capsaicinoids (Zhu et al. 2019). *Cis*-regulatory deletions also play a role in divergence, such as the recurrent deletion of a *Pitx1* enhancer in threespine sticklebacks, which is tied to the loss of pelvic structures in freshwater populations (Chan et al. 2010). Large insertions in regulatory regions have the potential to add new CREs from other genes, or change TF binding dynamics by altering the proximity or location of other CREs to each other (Wittkopp and Kalay 2012).

Transposable elements are known to impact gene expression in multiple examples. Generally, most TEs are highly methylated and inaccessible, however a small portion contain ACRs. For example, a transposon insertion in *GDF6* is associated with increased expression and changes in body armor size in marine and freshwater sticklebacks (Rebeiz and Tsiantis 2017). In plants, TEs have influenced gene regulation in several ways. The proliferation of TEs in large crop genomes is thought

to contribute to the increase in distal CREs in these genomes (Ricci et al. 2019). TEs are often found in the regions surrounding genes, and sometimes contribute novel regulatory functions, such as through the spreading of silent chromatin or the integration of new CREs. For example, the insertion of a *MITE* TE in an enhancer of maize *RAP2.7* led to earlier flowering, and in apple a retrotransposon in the promoter of *MYB1-1* increases expression to give apples their red color (Salvi et al. 2007; Zhang et al. 2019).

There are also examples of new regulatory sequences emerging from pre-existing ones. Major genomic rearrangements have the potential to relocate genes and their associated regulatory regions, resulting in CREs being situated in novel genomic contexts. This repositioning can reassign these CREs to regulate new genes, instead of their ancestral ones. For example, an enhancer of the *ladybird* gene in *Drosophila* and honeybee was inverted in beetles, causing it to regulate the neighboring gene, *C15*, instead (Cande et al. 2009b). More often novel expression domains are established in space or time through co-option of existing CREs, leading to a novel phenotype. For example, unique pigmentation in the wings of *Drosophila guttifera* was established by modifying a pre-existing enhancer, resulting in *wingless* expression in a new domain (Rebeiz and Tsiantis 2017). In nymphalid butterflies, variation in wing patterning involves the modification of a conserved CRE ground plan at the *WntA* locus (Mazo-Vargas et al. 2022). Different limb expression domains are driven in reporter assays by human and chimp versions of the HACSN1 non-coding regulatory element (Sumiyama and Saitou 2011). Expression in additional domains in humans is associated with 13 changes in the human HACSN1 compared to the chimp. While it was originally thought that these 13 mutations were directly responsible for the additional expression domains in humans, further investigation suggested that these mutations in HACSN1 actually just disrupted the repression of an ancestral CRE capable of driving expression in these unique domains. Following duplication, there can be subfunctionalization of enhancer functions, or repurposing of one enhancer for a novel regulatory function.

Finally, CREs can emerge *de novo* from non-regulatory DNA, such as through genetic drift. A recent study creating point mutations in developmental enhancers of *Drosophila* discovered that most

point mutations do not often result in new expression patterns in the embryo (Galupa et al. 2023). In contrast, they found that random sequences could drive expression in multiple domains across many cell types, suggesting that *de novo* enhancer formation may be a more common mechanism of phenotypic novelty. Since enhancer activity is so closely tied to chromatin accessibility, including pioneer TF motifs (Grh and Zelda for *Drosophila*) within the random sequences caused a higher proportion of these sequences to be capable of driving expression. Thus, it is possible that specific motifs in the genome can prime other sequences to evolve into *de novo* enhancers.

It is interesting to consider what *cis*-regulatory sequence changes most often lead to phenotypic variation, and what this might say about the nature of regulatory regions themselves. For example, is divergence often caused by the development of new CREs, or the modification of old ones? Is disruption of individual CREs often enough to lead to phenotypic variation, or do regulatory regions typically need to accumulate mutations within multiple CREs to have a noticeable phenotypic effect? Do mutations in TFBSs themselves more often lead to phenotypic divergence, or are alterations to genetic architecture and regulatory grammar just as common mechanisms of divergence? In chapter 3, we will explore these questions further by analyzing CREs involved in phenotypic divergence for flowering time between wild and domesticated species of tomato.

1.7 Exploring genetic interactions among CREs

Exploring how multiple CREs of a gene interact genetically improves our understanding of how regulatory regions coordinate the expression of genes at the transcript and phenotypic level, as well as the evolutionary processes that shape regulatory regions. Most studies of intragenic epistasis center around evaluating the consequences of multiple amino acid mutations (Domingo et al. 2019). However, few studies have actually looked at the genetic relationships among CREs, *in vivo*, using mutations of CREs in their native location. Additionally, studies that have measured genetic relationships between CREs have mostly been conducted using reporter assays, or using non-phenotypic readouts such as gene expression to quantify effects, rather than in *vivo* functional

dissections of CRE interactions. Studies that have been done *in vivo* mostly used animal model systems.

Interactions between CREs can generally be defined as additive, synergistic, or redundant. Additive interactions between CREs have a combined effect that is equal to the sum of each individual effect, while CREs are considered to interact synergistically when their combined effect exceeds the sum of their individual effects. CREs may also display redundancy, a form of non-additive relationship. This may reflect truly redundant roles, or mutation of one enhancer may only be revealed under particular conditions, and thus the function of the enhancer could be in conveying environmental robustness and canalization (Frankel et al. 2010). Thus far, there is no universal rule to dictate how CREs must interact. Rather, it likely varies on a case-by-case basis. Additionally, the same CREs may interact entirely differently depending on developmental stage (Kim and Wysocka 2023).

1.7.1 Reporter studies of CRE interactions

Many of the original studies looking at CRE interactions used reporter assays to quantify CRE activities and interaction effects, such as several studies of phenotypic divergence in fruit flies. These include studies of the *cis*-regulatory evolution of male abdominal pigmentation in *Drosophila melanogaster*, trichome loss in *Drosophila sechellia*, and darker body pigmentation in certain populations of *Drosophila melanogaster.*

In *Drosophila melanogaster*, males have abdominal pigmentation, which is repressed in females by Bab proteins (Williams et al. 2008). *Bab* expression in the abdomen is regulated by ABD-B and sex specific isoforms of DSX (DsxM acts as a repressor in males and an activator in females) which bind to a CRE within the first intron. Comparing the CRE in *Drosophila melanogaster* and *Drosophila willistoni* (which lacks dimorphic abdominal pigmentation), differences in the number and spacing of ABD-B TFBSs, in addition to altered polarity of one DSX site, contribute to the higher expression of *Bab* in females of *D. melanogaster.* Reporter experiments mutating the

differential ABD-B sites, combined with reversing the DSX site polarity, reduced the activity of the CRE below the level of either individual mutation, in an approximately additive manner. Additionally, although only a few ABD-B sites within the CRE differed between the species, there are 12 TFBSs for ABD-B distributed throughout the 663 bp sequence. The authors investigated interactions between these binding sites, by mutating them alone or in combination, and their effects on reporter expression were analyzed in transgenic females. Mutation of individual sites had variable effects on reporter activity, but the largest reduction in CRE activity came when more binding sites were mutated in combination. Thus, enhancer activity in this case is fully regulated by interactions among several non-redundant TFBSs.

In *Drosophila sechellia*, trichomes where lost during evolution due to multiple substitutions within TFBSs in the E6 enhancer of the *shavenbaby* gene (Frankel et al. 2011). The 13 substitutions were clustered in seven distinct regions, so the researchers mutated all seven clusters individually in *D. melanogaster* to the *D. sechellia* versions, and vice versa. In reporter assays, they found that the substitutions have non-additive effects – mutating seven clusters at once had a greater effect than the sum of the individual effects of each cluster mutation.

Finally, adaptive melanism in a highland population of *Drosophila melanogaster* is attributed to multiple substitutions within an enhancer of *ebony (Rebeiz et al. 2009).* They discovered that phenotypic divergence in pigmentation could be attributed to five substitutions. These five substitutions interacted non-additively in reporter assays – mutating five sites in combination had a lesser phenotypic effect than the sum of the individual effects of each mutation. Each individual site mutation also had a drastically different impact on reporter expression.

1.7.2 *In vivo* studies of CRE interactions

A few experiments of enhancer interactions have been done using genome editing *in vivo* in animal model systems. Many metazoan super-enhancers are known to cooperate synergistically to activate their target genes (Hnisz et al. 2017). For example, in the mammary gland, the *whey acidic*

protein gene is regulated by three individual enhancers that make up a super-enhancer (Shin et al. 2016). Individual and combined mutations were made *in vivo* within these three enhancers using CRISPR-Cas9 and TALENs in mice. These experiments demonstrated that all enhancers were required for full gene expression *in vivo*, and that they cooperate synergistically to induce gene expression 1000-fold during pregnancy. In contrast, *in vivo* homologous recombination was used to delete five constituent enhancers of the α-globin super-enhancer in mice, individually and in combination, and demonstrated that each enhancer acts independently, in an additive fashion to regulate α-globin expression as well as its hematological phenotype (Hay et al. 2016). A similar finding was discovered for the enhancers of limb development genes, again using *in vivo* genome editing to delete enhancers individually or in combination (Osterwalder et al. 2018). They found that no individual limb enhancer was essential for limb development, since deletions of individual enhancers did not impact limb morphology. However, removing pairs of enhancers did, and they suggest that this redundancy is conferred by additive effects of individual enhancers on gene expression. These examples highlight the potential for additive and redundant interactions among enhancers to convey phenotypic robustness to important developmental processes.

Recently, one study attempted to find genome-wide patterns defining enhancer interactions using transdifferentiation of human leukemia B-cells to macrophages as a model system (Choi et al. 2021). They used eRNA synthesis as a proxy for enhancer activity, and created models to associate enhancer activity and gene expression. The majority of the enhancers tested drove expression by an additive model (348 genes), while 136 cooperated synergistically. Synergistic enhancers were often associated with cell type specific TFs, suggesting that synergy may be utilized to enable switch-like expression patterns. In contrast, additivity may be useful to enable fine-tuning and robustness of gene expression in most contexts.

Lastly, studies of CRE interactions in the regulation of a tomato gene were conducted in our own lab recently. CRISPR-Cas9 deletion of specific upstream regions of the meristem development gene *CLV3* individually and in pairs revealed evidence of redundant, additive, and synergistic

interactions among CREs in this 5' region (Wang et al. 2021). In chapter 2, I will expand upon this study by exploring the nature of genetic interactions among upstream and downstream CREs in gene regulation, as well as how these interactions have changed over the course of evolution.

1.8 CRISPR-Cas9 *in vivo* mutagenesis of CREs

Currently, the literature includes very few studies that connect changes in *cis*-regulatory DNA sequences with changes in *cis*-regulatory activity and organismal phenotypes *in vivo*. Furthermore, many of these studies in plants have been performed for the purpose of bioengineering to improve agricultural traits, creating targeted mutations based on QTL mapping or GWAS. For this thesis research, we were concerned with utilizing *in vivo* mutagenesis to uncover principles of *cis*-regulatory architecture and evolution. Here I discuss a few of the experiments that have been performed to date, which use tools including CRISPR-Cas9 mediated deletion or base editing, insertion of CREs/promoter swapping using homology-directed repair (HDR), and various deactivated Cas9 systems for editing epigenomic features.

1.8.1 CRISPR mediated deletion and base editing of CREs

Several experiments within our own lab have revealed the potential of CRISPR-Cas9 to tune gene expression by deleting or perturbing CREs. Our lab demonstrated the utility of CRISPR-Cas9 for engineering quantitative variation in tomato fruit size by creating an allelic series through promoter bashing of *CLV3* (Rodríguez-Leal et al. 2017). A follow-up study investigated interactions between various CREs upstream of *CLV3* through targeted mutagenesis, and revealed robustness to large perturbations upstream of tomato *WUS* (Wang et al. 2021). In maize, kernel row number was also weakly tuned by making upstream mutations in two CLE peptides (Liu et al. 2021). The same method was employed to create variation in inflorescence branching, by targeting the 5' of tomato *WOX9* (Hendelman et al. 2021). Furthermore, deletion of specific CREs within the *WOX9* 5' was able to separate the pleiotropic roles of this gene in embryonic development and inflorescence branching. In rice, the null mutant of *SWEET11* has a sterile phenotype, however deletion of a 149 bp region in

the 5' eliminates an effector binding element (EBE), leading to improved disease resistance without severe growth defects (Li et al. 2020a). These examples highlight the potential for regulatory element engineering to bypass the negative pleiotropic consequences often tied to coding sequence mutations. An evolutionary study of butterfly wing patterns used CRISPR-Cas9 to create mosaic deletions within CREs upstream of *WntA* in multiple butterfly species, revealing the function of conserved and divergent CREs in wing patterning control (Mazo-Vargas et al. 2022).

Advancements in CRISPR editing technologies are promising for the study of *cis*-regulatory regions. Base and prime editing are more precise approaches to genetic engineering, with the potential to alter specific nucleotides and TFBSs. Base editing and prime editing have some proven ability to induce precise mutations in plants, although they have mostly been tested on coding sequences to date (Lin et al. 2020; Zafar et al. 2020). These techniques are still relatively inefficient in plants. In animals, researchers have used base editing to explore genetic therapies for Huntington's disease. Base editing of the TFBS for NF-kB in the promoter of *huntingtin* led to reduced expression both in cell culture and a mouse model of Huntington's disease, highlighting the potential of base editing in future gene therapies (Lim et al. 2022). In the future, the development of more efficient PAM-less CRISPR systems will enable even more precise *in vivo* editing of CREs to better understand their function and evolution, as well as crop improvement.

1.8.2 CRISPR homology-directed repair for precise insertions

CRISPR-Cas9 systems employing homology-directed repair are able to insert specific sequences at precise genomic locations, or swap out large regions of DNA for other sequences. Unfortunately, it is still very inefficient in plants, so there are few examples of its successful implementation in crop engineering (Li et al. 2020b). HDR was used to insert a CaMV 35S promoter upstream of *ANT1* in tomatoes, leading to enhanced anthocyanin production that turned the fruits purple (Čermák et al. 2015). HDR was also used to engineer drought stress resistance in maize. ARGOS8 is a negative regulator of ethylene responses, and is lowly expressed in many maize inbred

varieties. Researchers hypothesized that increasing *ARGOS8* expression in these lines would reduce ethylene sensitivity, thereby increasing yield under drought stress (Shi et al. 2017). To achieve this, they inserted the *GOS2* promoter into the 5'UTR of *ARGOS8*, swapping out the native promoter in order to enhance expression. In field trials these alleles did have increased grain yield under stress conditions. In the future, more efficient methods of HDR will enable limitless possibilities for studying CREs within specific contexts. For example, it could enable promoter swaps of entire *cis*-regulatory regions between conserved genes of different species, enabling the comparison of these regions with a more direct approach than reporter experiments.

1.8.3 CRISPR mediated editing of the epigenome

CRISPR also has the potential to help us understand the direct relevance of the epigenetic features of CREs, for example through changing their accessibility. For example, a deactivated version of Cas9 (dCas9) fused to a methyltransferase can be used in the targeted methylation of specific sequences, with the goal of silencing them. Conversely, the dCas9-TET1 system can be used in the targeted demethylation of specific promoter sequences for gene activation. Recently, Ghoshal et al. developed a CRISPR-dCas9 fused to a bacterial CG DNA methyltransferase, using it to methylate and silence the promoter of the flowering gene *FWA* in *Arabidopsis* (Ghoshal et al. 2021). These methyl marks can be replicated through both mitotic and meiotic divisions, representing a stable method of gene expression modulation. Another group fused RNA-directed DNA methylation (RdDM) proteins to zinc fingers targeted at the *FWA* gene, promoting methylation and silencing (Gallego-Bartolomé et al. 2019). Conversely, directed demethylation was achieved by creation of a CRISPR SunTag-TET1 system using the demethylase TET1 to demethylate the *FWA* gene and the *CACTA1* transposon, leading to activation of gene expression (Gallego-Bartolomé et al. 2018). A CRISPR-dCas9 fused to a histone acetyltransferase has also been developed to promote open chromatin and gene activation at specific loci. When HAT1-dCas9 was targeted to the promoter of *AREB1* in *Arabidopsis*, a master regulator of the drought stress response, the gene was upregulated

and the plants experienced better survival under drought stress (Roca Paixão et al. 2019). Since many CREs are enhancers, this system offers great promise for engineering gene activation, which can be more difficult to achieve through direct editing of the native DNA sequence.

Chapter 2: Deep functional conservation of a plant stem cell regulator despite extreme divergence in *cis*-regulation

2.1 Summary

Gene expression is controlled by *cis*-regulatory elements that exist in multiple genomic contexts, including upstream of genes, downstream, at distal enhancer elements and even within genes themselves. Complex genetic and physical interactions between these elements mediate proper spatiotemporal regulation of gene expression. Despite their complexity and critical role in gene regulation, non-coding regions containing regulatory elements tend to evolve at a much faster rate compared to coding sequences, even when the corresponding genes are highly conserved. How regulatory regions are able to tolerate sequence divergence, while maintaining similar gene function and expression patterns in distantly related organisms, is still a question of debate. Here we examine the evolution of regulatory regions controlling expression of *CLAVATA3* (*CLV3*), a highly conserved plant stem cell regulator in *Arabidopsis* and tomato, which diverged ~125 million years ago. We used CRISPR-Cas9 to engineer over 70 unique mutations in the upstream (5') and downstream (3') regions of *CLV3* in both species and then assessed their impact, individually and in combination, on locule number. We found that tomato was highly sensitive to sequence perturbation upstream of *CLV3*, and only weakly affected by mutations downstream, while the combined effect of 5'+3' mutations suggested an additive or mildly synergistic relationship. In contrast, the *Arabidopsis CLV3* 5' and 3' were highly buffered to large sequence perturbations, such that deletion of large regions 5' or 3' only resulted in weak phenotypic effects. 5'+3' combinatorial mutations had synergistic effects on phenotype, and these alleles spanned the spectrum of phenotypic variation. A conserved sequence of 27 bp shared between tomato and *Arabidopsis CLV3,* albeit in altered locations, suggests a conserved mechanism of regulation via a common transcription factor binding site (TFBS). Our results support a mode of *CLV3* evolution in which the spatial organization of shared *cis*-regulatory elements is altered.

Predicting the effects of engineered *cis*-regulatory variation in new plants therefore depends on an understanding of the underlying spatial architecture of gene regulation.

2.2. Introduction

Changes in the sequence of DNA underlie the emergence of new species, as well as optimization of crops during domestication and breeding. However, over the course of evolution, mechanisms of selection have ensured the conservation of thousands of genes vital to survival and reproduction, even between species from highly diverged lineages. Interestingly, while many of these genes have significant conservation of protein sequence, function, and spatiotemporal patterns of expression, their *cis*-regulatory regions are often highly diverged, and are thus evolving at a faster rate relative to coding sequences (Koonin and Wolf 2010). The lack of detectable conservation among the *cis*-regulatory sequences of conserved genes is still not completely understood. A better understanding of the architecture of *cis*-regulatory regions would help to clarify how this sequence divergence is tolerated.

Cis-regulatory elements (CREs) and their interactions mediate proper spatiotemporal regulation of gene expression during development and growth. Regulatory regions are composed of transcription factor binding sites (TFBSs), whose organizational grammar in number, spacing, orientation, order, and cooperativity can be vital to produce specific expression patterns in some circumstances, while in others it may be more flexible. These two scenarios are described by the enhanceosome and billboard models of enhancer architecture. The enhanceosome model describes CREs that must remain quite rigid in their identity and organization in order to function properly (Arnosti and Kulkarni 2005). This is often the case for TFs that bind cooperatively, for example, and these CREs are more likely to be highly conserved. In contrast, the billboard model suggests that the organizational grammar of specific TFBSs is flexible for proper gene expression (Arnosti and Kulkarni 2005). Since TFBSs are often only 5-11 bp long, detecting conservation of CREs organized

in this way is more difficult. Of course, genes may also be regulated by CREs involved in a mixture of these models.

Thus, while conservation may be a useful tool to detect a portion of CREs, it is not sufficient to predict the existence of all CREs. However, identification of conserved non-coding sequences (CNSs) between closely and distantly related species has been successful at identifying CREs in a number of documented cases. For example, analyzing sequence conservation over a shorter evolutionary time-frame, such as CNSs within the *Solanaceae* family, revealed that many CNSs overlap with regions of open chromatin upstream of the tomato *WUSCHEL HOMEOBOX9* (*WOX9*) gene, and these regions were shown to have a conserved function in mediating the pleiotropic roles of WOX9 in embryonic development and inflorescence branching (Hendelman et al. 2021). There is also some evidence of functional CNSs among distantly related species. For example, many ultraconserved non-coding elements (which have 100% identity over 200 bp long) have been identified among distantly related animals, such as humans and mice (Bejerano et al. 2004). Their functional relevance is still under debate following findings that some do not have a phenotype when removed in mice, and many can maintain activity amidst multiple mutations (Ahituv et al. 2007; Snetkova et al. 2021). Despite the ongoing debate, some ultraconserved elements are clearly vital to development. For example, a core enhancer sequence of the *sonic hedgehog* gene has been highly conserved among vertebrates with limbs, while accumulating mutations in limb-less snake lineages (Kvon et al. 2016). Nonetheless, deep conservation of non-coding sequence seems to be the exception, rather than the rule. While CNSs can be discovered between family members, the regulatory regions of genes from species that are more distantly related are typically unable to be aligned due to sequence degradation. There are many examples of this phenomenon among the animal kingdom. Enhancer sequences of the developmental patterning gene *even-skipped* (*eve*) are highly diverged between sepsids and *Drosophila,* yet sepsid enhancers are capable of driving a near-identical expression pattern of *eve* when transformed into *Drosophila* embryos (Hare et al. 2008). These non-coding regions do share very short conserved sequences between 20-30 bp containing

transcription factor motifs, suggesting that while short TFBSs may be deeply conserved, regulatory grammar and organization could be more malleable to change. A similar finding was discovered for tomato and *Arabidopsis WOX9*. While the pleiotropic functions of this gene are deeply conserved, non-coding sequence is not at this level of species divergence, apart from very short sequences (10-30 bp) with entirely altered grouping, order, and orientation (Hendelman et al. 2021). Thus, studies of both plants and animals have provided support for both the enhanceosome and billboard models of enhancer architecture. The complex interactions and mechanisms involved in maintaining conserved spatiotemporal expression over evolutionary time is still an active area of research.

Complex interactions between CREs in various genomic contexts, including upstream and downstream of genes, within genes, and at distal sites >10 kb away, mediate gene regulation in space and time. Understanding how the regulatory grammar, genomic context, and genetic and physical interactions of CREs are conserved (or not) over evolutionary time is fundamental to understanding gene regulation. From our previous studies, we discovered that various regions upstream of the stem cell regulator *CLV3* interact additively, synergistically, and redundantly in their control of meristem size in tomato (Wang et al. 2021). We aimed to expand upon this study by including CREs from other genomic contexts, such as downstream of genes, as well as providing an evolutionary perspective on CRE organization and interactions. Previous studies of the evolution of *cis*-regulatory regions have provided valuable insights, however they have relied on indirect methods such as reporter assays to predict enhancers and their contribution to gene regulation (Hare et al. 2008; Cameron and Davidson 2009; Wong et al. 2020). In our study, we explore CREs, their organization, and their interactions with a functional genetics approach, using CRISPR-Cas9 to create allelic diversity in the regulatory regions of a highly conserved gene in tomato and *Arabidopsis*. Through this approach we were able to manipulate CREs in their native context, *in vivo,* and thus characterize the functional relevance of the loss of CREs in various genomic contexts directly, through phenotypic output. Our results support previous findings of highly divergent *cis*-regulatory regions between distantly related species, with

conserved gene function/expression despite altered organization of CREs and their genetic interactions.

2.3 Results

2.3.1 Conserved function of the CLV3 peptide despite regulatory sequence divergence in *Arabidopsis* and tomato

Here we introduce *CLV3* as a model gene to gain further insight into CRE evolution. CLV3 is a signaling peptide that negatively regulates meristem size in a feedback loop with the stem cell promoting transcription factor WUSCHEL (WUS) (**Fig. 2-1A**) (Somssich et al. 2016). As an integral regulator of stem cell development, it is highly conserved among land plants (Fouracre and Harrison 2022). Thus, although *Arabidopsis thaliana* (*Arabidopsis*) and *Solanum lycopersicum* (tomato) belong to different angiosperm lineages separated by ~125 MY of evolution (even greater than the time separating humans and mice), the function of CLV3 in meristem regulation is conserved (Somssich et al. 2016). Null mutations in both *Arabidopsis* and tomato *CLV3* lead to stem cell overproliferation, which results in enhanced size and number of floral organs including sepals, petals, stamens, and carpels (**Fig. 2-1B, C**). Furthermore, the functional, processed 12 amino acid peptide sequence is highly conserved, as well as several post-translational modifications including hydroxylation and arabinosylation (Ohyama et al. 2009; Xu et al. 2015). In both *Arabidopsis* and tomato, CLV3 peptide is known to bind to leucine-rich repeat receptor-like kinases (LRR-RLKs), including CLAVATA1 (CLV1) (Somssich et al. 2016). Furthermore, they share similar expression domains within the central zone (CZ) of the shoot apical meristem (SAM), although tomato *CLV3* (*SlCLV3*) seems to be absent from the L1 layer compared to *Arabidopsis CLV3* (*AtCLV3*) (**Fig. 2-1D, E**) (Fletcher et al. 1999; Xu et al. 2015).

We began our analysis of *cis*-regulatory evolution by attempting to align the regulatory regions of *Arabidopsis* and tomato *CLV3* with several of their close relatives. Through this analysis we were able to identify several regions of partially and highly conserved non-coding sequence

within each family (the *Brassicaceae* or *Solanaceae*), however the non-coding regions of *Arabidopsis* and tomato *CLV3* were not able to be aligned (**Fig 2-1D, E**). Thus, despite conservation of peptide sequence, function, and expression, the regulatory regions of *Arabidopsis* and tomato *CLV3* are remarkably diverged. Given these characteristics, we determined that *CLV3* is a good model gene to investigate CRE evolution, and specifically how genes in distantly related species are able to maintain conserved spatiotemporal expression and function, despite highly divergent regulatory regions.

Figure 2-1. Conserved function of the CLV3 peptide despite regulatory sequence divergence in *Arabidopsis* and tomato.

A. Representative diagram of a shoot apical meristem (SAM), demonstrating the conserved negative feedback loop between the signaling peptide CLV3 and the transcription factor WUS. CLV3 peptide indirectly inhibits WUS expression, while WUS promotes CLV3 expression.

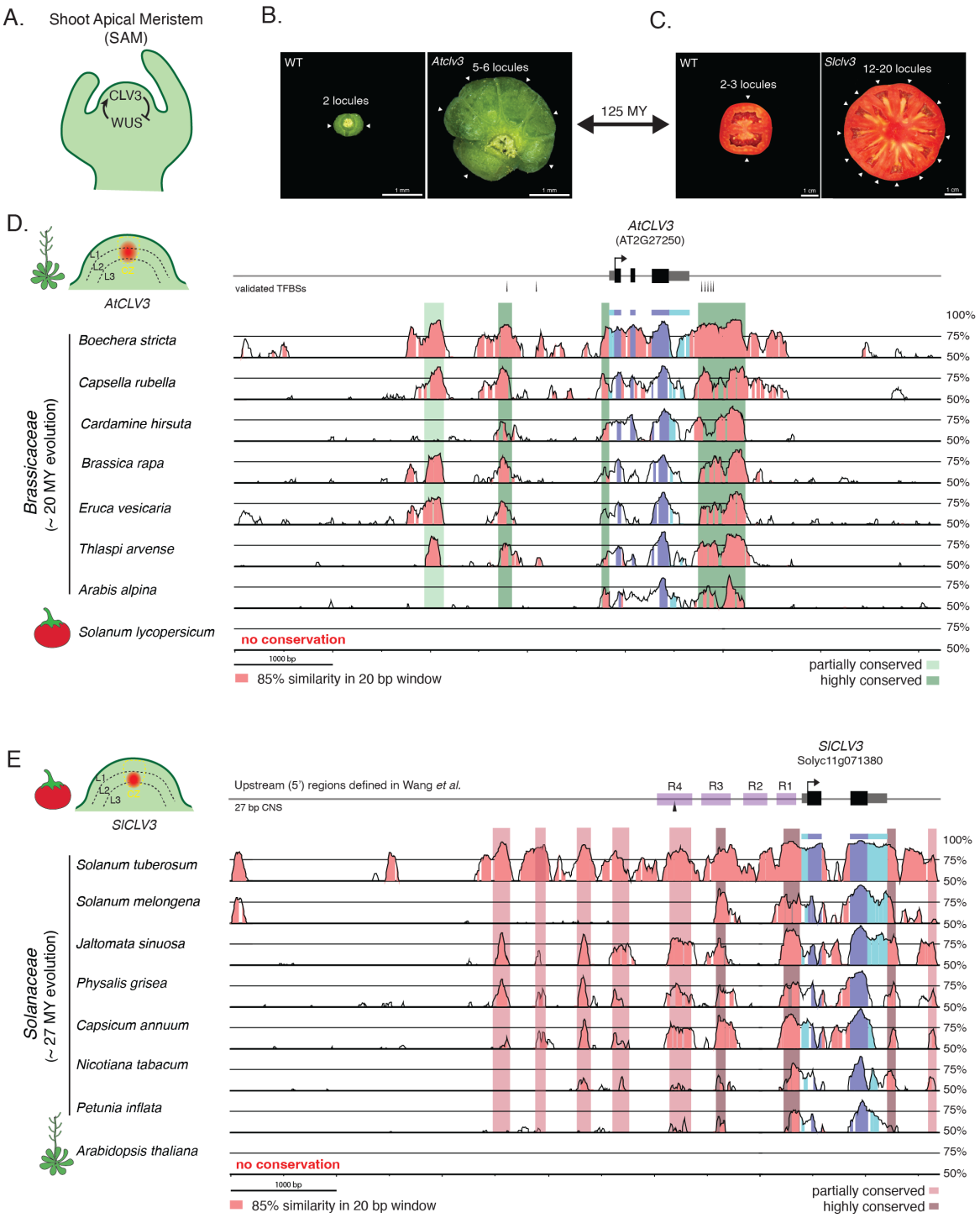
B. Top-down view of *Arabidopsis* siliques from wild type (WT) and an *Atclv3* null mutant. Individual locules are denoted by white arrows.

C. Tomato fruits from WT and a *Slclv3* null mutant, sliced in half. Individual locules are denoted by white arrows.

D. mVISTA DNA sequence alignments of *CLV3* orthologs from various *Brassicaceae* species, using the *AtCLV3* gene and its entire 5' and 3' regions as the reference sequence. Conservation is calculated as sequences with 85% similarity in 20 bp windows. Conserved UTRs are light blue, and conserved exons are dark blue. Regions of within-family conservation are represented by light and dark green bars. *SlCLV3* could not be aligned to *AtCLV3.* A representative diagram of the SAM of *AtCLV3* is shown, indicating the location of *AtCLV3* RNA expression relative to previously defined regions. L1, L2, and L3 layers are denoted by dotted black lines, the central zone is outlined in yellow, and *CLV3* transcripts are represented in red.

E. mVISTA DNA sequence alignments of *CLV3* orthologs from various *Solanaceae* species, using the *SlCLV3* gene and its entire 5' and 3' regions as the reference sequence. Conservation is calculated as sequences with 85% similarity in 20 bp windows. Conserved UTRs are light blue, and conserved exons are dark blue. Regions of within-family conservation are represented by light and dark red bars. *AtCLV3* could not be aligned to *SlCLV3.* A representative diagram of the SAM of *SlCLV3* is shown, indicating the location of *SlCLV3* RNA expression relative to previously defined regions.

Figure 2-1.


A.
Shoot Apical Meristem (SAM)
CLV3
WUS
B.
C.
WT
2 locules
Atclv3
5-6 locules
125 MY
WT
2-3 locules
Slclv3
12-20 locules
D.
L1
L2
L3
CZ
AtCLV3
Brassicaceae (~ 20 MY evolution)
Boechera stricta
Capsella rubella
Cardamine hirsuta
Brassica rapa
Eruca vesicaria
Thlaspi arvense
Arabis alpina
Solanum lycopersicum
AtCLV3
(AT2G27250)
validated TFBSs
100%
75%
50%
no conservation
1000 bp
85% similarity in 20 bp window
partially conserved
highly conserved
E
SlCLV3
Solanaceae (~ 27 MY evolution)
Solanum tuberosum
Solanum melongena
Jaltomata sinuosa
Physalis grisea
Capsicum annuum
Nicotiana tabacum
Petunia inflata
Arabidopsis thaliana
SlCLV3
Solyc11g071380
Upstream (5') regions defined in Wang et al.
27 bp CNS
R4
R3
R2
R1
100%
75%
50%
no conservation
1000 bp
85% similarity in 20 bp window
partially conserved
highly conserved

2.3.2 Mutations affecting the 5' or 3' region of *AtCLV3* have weak effects on fruit locule number

To understand how such divergent non-coding regions nevertheless support similar gene functions, we compared the organization of the *CLV3* regulatory regions in tomato and *Arabidopsis*, using a functional genetics approach. We used CRISPR-Cas9 to dissect 5' and 3' CREs and their genetic interactions in the regulation of *CLV3*. Previously, we used CRISPR-Cas9 multiplex mutagenesis to create allelic diversity in the proximal 2 kb upstream of tomato *CLV3*, and discovered several regions of importance for *SlCLV3* regulation (Rodríguez-Leal et al. 2017; Wang et al. 2021). In order to study the relative organization of CREs in a distant homolog, we selected *Arabidopsis thaliana,* a model plant in which *CLV3* is well-studied and known to have a similar phenotype to tomato *CLV3.* We took a similar, unbiased approach to explore the relative contribution of both the 5' and 3' to the regulation of *Arabidopsis CLV3.* We used two different 8-gRNA arrays to generate deletions within a 1.5 kb region upstream of the 5'UTR, as well as the entire 3.8 kb region between the 5'UTR of *AtCLV3* and the next gene upstream. Together these approaches generated 11 alleles with unique mutations in the 5' of *AtCLV3* (**Fig. 2-2A**). We counted locule number, a sensitive and easily quantifiable phenotype of *CLV3* perturbation. Five of these 5' alleles had no impact on locule number at all, while six had a very weak increase in average locule number compared to WT. The most significant increase in locule number was observed when almost the full ~3.8 kb of upstream sequence was deleted ($AtCLV3^{pro\text{-}11}$). However, large perturbations to the distal 5' region had no effect on locule number on their own ($AtCLV3^{pro\text{-}9,10}$), while certain alleles with gene proximal deletions had a slight increase in locule number ($AtCLV3^{pro\text{-}3,7}$) (**Fig. 2-2A, B**). Notably, none of these 5' alleles were able to recapitulate the phenotype of the null allele (*Atclv3*). Thus, unlike tomato, the 5' non-coding region of *Arabidopsis CLV3* does not appear to be critical for the regulation of *CLV3* function, even though it contains binding sites for both WUS and SHOOT MERISTEMLESS (STM), another *CLV3* regulator.

Since the 5' of *AtCLV3* was highly buffered to large sequence perturbations, we hypothesized that critical CREs may be present downstream instead. Several lines of evidence supported this hypothesis. Previous work showed that *AtCLV3* GUS reporter constructs require 1.2 kb of the 3' region in order to recapitulate endogenous expression (Brand et al. 2002). Furthermore, ChIP-qPCR experiments confirmed the binding of WUS to sites in both the 5' and 3' of *AtCLV3,* and binding of STM to a site within the 5' (Perales et al. 2016; Su et al. 2020). Five of the six WUS binding sites are clustered within a 116 bp region 3' of *AtCLV3,* and form a *cis*-regulatory module that relies on cooperativity and WUS concentration to control *AtCLV3* expression and domain (Perales et al. 2016). We therefore tested the effects of 3' deletions, targeting 1.6 kb of the 3' region proximal to the 3'UTR of *AtCLV3*, and isolated three alleles with large 3' deletions (**Fig. 2-3C**). Deletion of a ~600 bp sequence proximal to the end of the gene had a weak effect on locule number ($AtCLV3^{3p\text{-}1}$), while disrupting the distal 3' region had no significant effect ($AtCLV3^{3p\text{-}2}$). Deletion of a portion of the 3' region has a comparable effect to deleting the entire 5' region of *AtCLV3,* although again none of the 3' alleles recapitulate the null phenotype. Together, these results demonstrate that neither the 5' nor 3' region alone is critical to *AtCLV3* function.

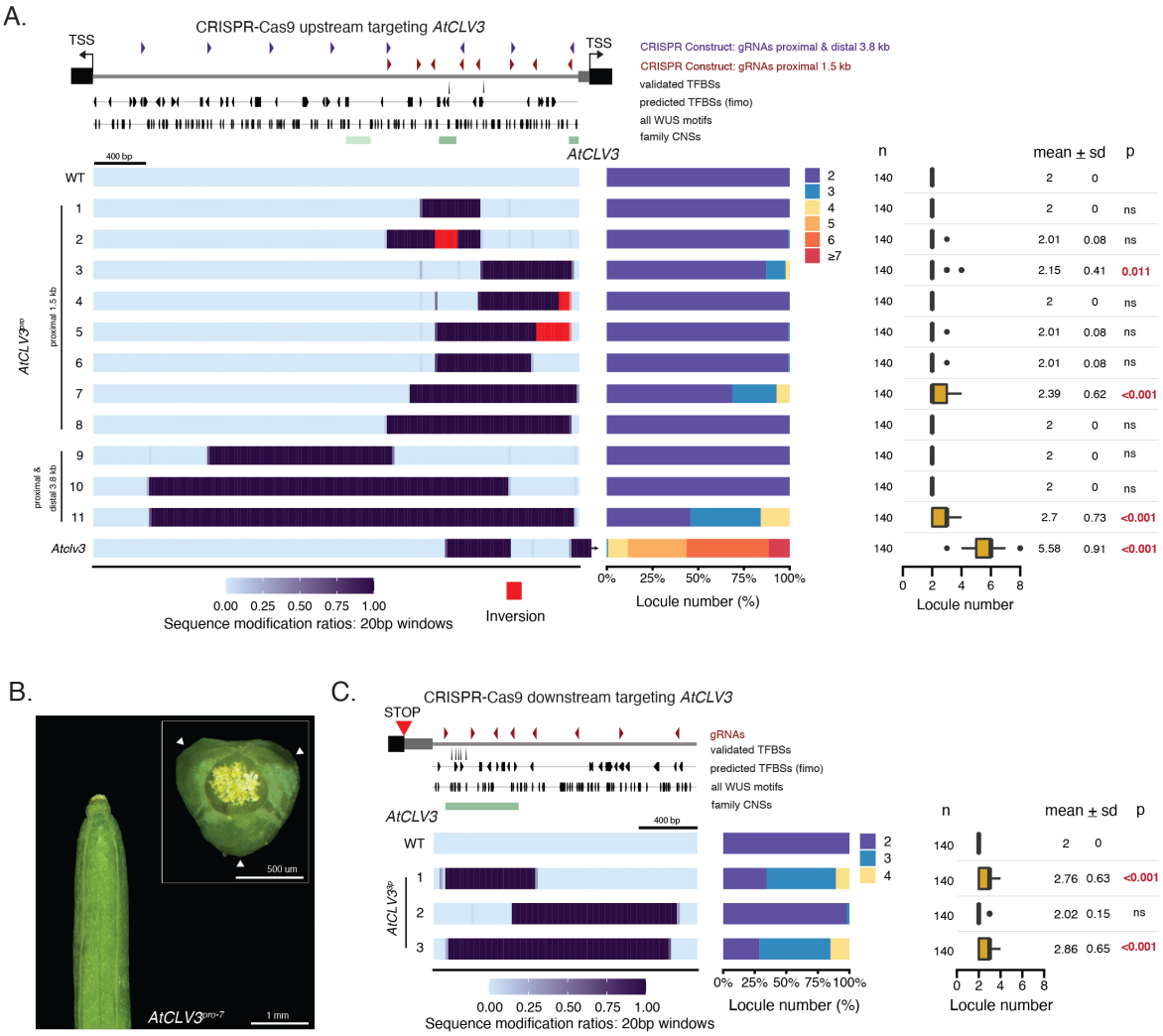
Figure 2-2. Mutations affecting the 5' or 3' region of *AtCLV3* have weak effects on fruit locule number.

A. Encoded representation of the 11 alleles generated from targeting the 5' of *AtCLV3* with a gRNA array spanning the 1.5 kb proximal to the 5'UTR (red), and a gRNA array targeting the entire 3.8 kb region between the 5'UTR of *AtCLV3* and the next gene upstream (purple). The alleles have been encoded, such that perturbations to the region are represented as the degree of sequence modification relative to WT within 20 bp windows. Inversions within the alleles are shown in red in the encoding. Validated and predicted TFBSs are indicated by black arrows. Family CNSs identified in Fig. 2-1 are represented on the sequence in green. Locule number quantifications are represented by stacked bar plots and box plots (with outliers as black points). Sample number (n) is shown to the left, and mean and standard deviation (sd) are shown to the right. A two-sided Dunnett's compare with control test was performed to compare all 5' alleles to WT, and the p-values are included to the right. "ns" means not significant.

B. A silique with three locules from the 5' allele $AtCLV3^{pro\text{-}7}$. A top-down view is shown in the inset.

C. Encoded representation of the three alleles generated from targeting the 3' region of *AtCLV3* with an 8-gRNA array, and their locule number quantifications. A two-sided Dunnett's compare with control test was performed to compare all 3' alleles to WT.

Figure 2-2.


A.
CRISPR-Cas9 upstream targeting AtCLV3
TSS
TSS
CRISPR Construct: gRNAs proximal & distal 3.8 kb
CRISPR Construct: gRNAs proximal 1.5 kb
validated TFBSs
predicted TFBSs (fimo)
all WUS motifs
family CNSs
AtCLV3
400 bp
WT
1
2
3
4
5
6
7
8
9
10
11
Atclv3
AtCLV3pro
proximal 1.5 kb
proximal & distal 3.8 kb
0.00 0.25 0.50 0.75 1.00
Sequence modification ratios: 20bp windows
Inversion
2
3
4
5
6
≥7
0% 25% 50% 75% 100%
Locule number (%)
n
mean ± sd
p
140 2 0
140 2 0 ns
140 2.01 0.08 ns
140 2.15 0.41 0.011
140 2 0 ns
140 2.01 0.08 ns
140 2.01 0.08 ns
140 2.39 0.62 <0.001
140 2 0 ns
140 2 0 ns
140 2 0 ns
140 2.7 0.73 <0.001
140 5.58 0.91 <0.001
0 2 4 6 8
Locule number
B.
500 um
AtCLV3pro-7
1 mm
C.
CRISPR-Cas9 downstream targeting AtCLV3
STOP
gRNAs
validated TFBSs
predicted TFBSs (fimo)
all WUS motifs
family CNSs
AtCLV3
400 bp
WT
1
2
3
AtCLV3gp
0.00 0.25 0.50 0.75 1.00
Sequence modification ratios: 20bp windows
2
3
4
0% 25% 50% 75% 100%
Locule number (%)
n
mean ± sd
p
140 2 0
140 2.76 0.63 <0.001
140 2.02 0.15 ns
140 2.86 0.65 <0.001
0 2 4 6 8
Locule number

2.3.3 Combined mutations in the 5' and 3' regions of *AtCLV3* have synergistic effects on fruit locule number

The absence of strong phenotypes from perturbing large sections of the 5' or 3' of *AtCLV3* alone suggested the likelihood of higher order interactions between these regions. We previously showed that individual regions within the *SlCLV3* 5' had multiple complex interactions when mutations in these regions were combined, including additivity, redundancy, and synergy (Wang et al. 2021). We hypothesized that combined mutations in important *AtCLV3* 5' and 3' regions would expand the range of potential locule number phenotypes. In order to test this hypothesis, we took two main approaches to create alleles with both 5' and 3' mutations. First, we chose two different transgene-free 5' alleles with large deletions proximal to the 5'UTR of *AtCLV3* (*AtCLV3*$^{pro\text{-}8}$ and *AtCLV3*$^{pro\text{-}7}$), and transformed them with the 8-gRNA array previously used to create 3' mutations (**Fig. 2-3A**). The second, complimentary approach was to take a 3' allele with a large deletion (*AtCLV3*$^{3p\text{-}3}$) and transform it with a 5' 8-gRNA array spanning the proximal 1.5 kb to the 5'UTR. We selected for alleles with mutations in the newly targeted regions, resulting in a series of 28 alleles with various combinations of 5' and 3' mutations (**Fig. 2-3B, C, D**). Intriguingly, this series of 28 alleles spanned the entire spectrum of variation for locule number, encompassing alleles with weak, moderate, and strong effects, as well as multiple alleles with a null-like phenotype.

These alleles made it possible to further identify several subregions that play a significant role in *CLV3* regulation (we note that the analysis was not systematic and is limited by the random nature of the allele-generating scheme). New 3' targeting in the background of the 5' allele *AtCLV3*$^{pro\text{-}8}$ suggests that deletion of the ~600 bp region between gRNA-1 and gRNA-5 downstream of *AtCLV3* is sufficient to produce a null-like phenotype in this 5' mutant background (*AtCLV3*$^{pro\text{-}8}$ + 3p^{l}) (**Fig. 2-3B**). This deleted region notably overlaps with the 3' WUS TFBSs. Partial deletions of this region, as well as mutations distal to this region, only had weak or moderate effects on locule number in the 5' mutant background. New 3' mutations in the background of the 5' allele *AtCLV3*$^{pro\text{-}7}$ (which has a

weak phenotype on its own) reiterate the strong effect on locule number from loss of the 600 bp region (*AtCLV3*$^{pro\text{-}7}$ + 3p^{r}) (**Fig. 2-3C**). Smaller deletions within this 600 bp region reveal a slight enhancement in locule number from deleting all five 3' WUS binding sites (*AtCLV3*$^{pro\text{-}7}$ + 3p^{o}). However, a greater enhancement in locule number results from deleting the sequence adjacent to the 3' WUS TFBSs (*AtCLV3*$^{pro\text{-}7}$ + 3p^{p} and *AtCLV3*$^{pro\text{-}7}$ + 3p^{q}), suggesting the existence of additional 3' CREs outside of those already characterized.

We were also able to further dissect the *AtCLV3* 5' by sequential mutagenesis in the background of the 3' allele *AtCLV3*$^{3p\text{-}3}$ (**Fig. 2-3D**). The majority of these alleles deleted the 5' WUS TFBS and had moderate or strong phenotypes, while one allele, *AtCLV3*$^{3p\text{-}3}$ + pro^{b}, deleted a large region of the 5' without compromising the 5' WUS binding site and had no effect. This suggests the increased importance of the 5' WUS TFBS in the absence of certain 3' CREs. However, it seems likely that additional CREs, as well as higher order interactions among them, may be present in the 5' region, since larger deletions do not necessarily yield stronger phenotypes than some smaller deletions (for example, *AtCLV3*$^{3p\text{-}3}$ + pro^{g} compared to *AtCLV3*$^{3p\text{-}3}$ + pro^{h}). Altogether, the data show that alleles combining deletions in both the 5' and 3' regions produce a range of *CLV3* loss-of-function phenotypes (**Fig. 2-3E**).

This enhancement of locule number in combined 5'+3' alleles compared to single 5' or 3' alleles prompted us to test the interaction effect between 5' and 3' mutations (**Fig. 2-3F**). Specifically, we asked whether the enhancement was equal to the sum of the individual effects of 5' and 3' mutations (additive), or whether it was greater than the sum of these effects (synergistic). For this analysis, we would've ideally compared alleles in the context of an identical genetic background. However, our sequential CRISPR-Cas9 targeting approach meant that the mutations/deletions in the 5'+3' combination alleles were not identical to both individual alleles. To minimize potential non-specific effects, we therefore chose those alleles in which the individual deletions were most similar to the corresponding deletion in the combined allele. Among these, we then focused on the six 5'+3' alleles that produced the strongest phenotypes (denoted in Fig. 2-3 by asterisks), and compared the

effect of the individual deletions (5’ or 3’) to their combination (5’+3’) using a linear model (**Supplementary 2-1**). In all cases, the combined effects were non-additive, revealing either redundancy or synergy between 5’ and 3’ CREs of *AtCLV3* (**Fig. 2-3F**).

Together, these results suggest that in *Arabidopsis,* 5’ regulatory elements can compensate for the loss of 3’ elements, and vice versa. These two regions therefore act either redundantly or in parallel to regulate *AtCLV3* activity.

Figure 2-3. Combined mutations in the 5’ and 3’ regions of *AtCLV3* have synergistic effects on fruit locule number.

A. Sequential CRISPR-Cas9 mutagenesis was used to make alleles with combinations of 5’ and 3’ mutations. For this approach, either a fixed 5’ allele was transformed with 3’-targeted gRNAs to induce new 3’ mutations, or a fixed 3’ allele was transformed with 5’-targeted gRNAs to induce new 5’ mutations. These transgenics were then screened for new mutations in the sequentially targeted region by PCR, and alleles with both 5’ and 3’ mutations were selected and bred to homozygosity in subsequent generations. The new alleles were phenotyped for fruit locule number, and genetic interaction tests were applied to explore the relationship between combined mutations in the 5’ and 3’.

B. Encoded representation of alleles generated from sequential CRISPR-Cas9 targeting with the *AtCLV3* 3’-gRNA array, in the background of the fixed 5’ mutant $AtCLV3^{pro\text{-}8}$. Known TFBSs are indicated by black arrows. Family CNSs identified in Fig. 2-1 are represented on the sequence in green. Locule number quantifications are represented by stacked bar plots and box plots. A grey box highlights an identified region of importance for regulation. Asterisks denote alleles that were tested for interaction effects. A two-sided Dunnett’s compare with control test was performed to compare WT and all alleles to $AtCLV3^{pro\text{-}8}$.

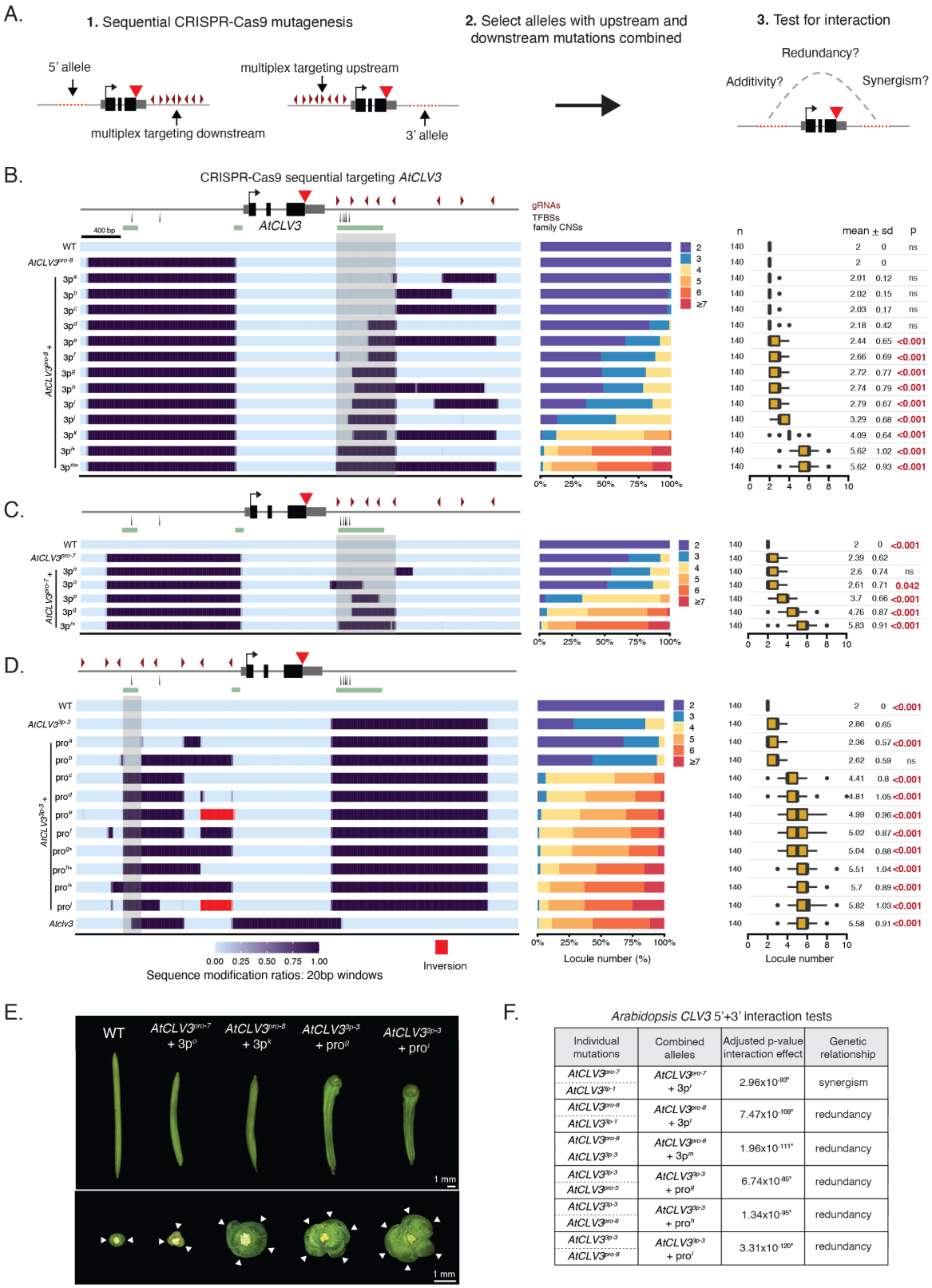
C. Encoded representation of alleles generated from sequential CRISPR-Cas9 targeting with the *AtCLV3* 3’-gRNA array, in the background of the fixed 5’ mutant $AtCLV3^{pro\text{-}7}$, and their locule number quantifications. A two-sided Dunnett’s compare with control test was performed to compare WT and all alleles to $AtCLV3^{pro\text{-}7}$.

D. Encoded representation of alleles generated from sequential CRISPR-Cas9 targeting with the *AtCLV3* proximal 1.5 kb 5’-gRNA array, in the background of the fixed 3’ mutant $AtCLV3^{3p\text{-}3}$, and their locule number quantifications. A two-sided Dunnett’s compare with control test was performed to compare WT and all alleles to $AtCLV3^{3p\text{-}3}$.

E. Representative silique images from WT and several combined 5’+3’ alleles. A top-down view is shown below. White arrows denote individual locules.

F. Interaction tests performed between select individual 5’ and 3’ mutants and similar combined 5’+3’ alleles. P-values of the interaction effect were adjusted for multiple comparisons.

Figure 2-3.



Individual mutations	Combined alleles	Adjusted p-value interaction effect	Genetic relationship
AtCLV3$^{pro\text{-}7}$ / *AtCLV3*$^{3p\text{-}1}$	*AtCLV3*$^{pro\text{-}7}$ + 3p^{r}	2.96x10^{-93*}	synergism
AtCLV3$^{pro\text{-}8}$ / *AtCLV3*$^{3p\text{-}1}$	*AtCLV3*$^{pro\text{-}8}$ + 3p^{l}	7.47x10^{-109*}	redundancy
AtCLV3$^{pro\text{-}8}$ / *AtCLV3*$^{3p\text{-}3}$	*AtCLV3*$^{pro\text{-}8}$ + 3p^{m}	1.96x10^{-111*}	redundancy
AtCLV3$^{3p\text{-}3}$ / *AtCLV3*$^{pro\text{-}5}$	*AtCLV3*$^{3p\text{-}3}$ + pro^{g}	6.74x10^{-85*}	redundancy
AtCLV3$^{3p\text{-}3}$ / *AtCLV3*$^{pro\text{-}6}$	*AtCLV3*$^{3p\text{-}3}$ + pro^{h}	1.34x10^{-95*}	redundancy
AtCLV3$^{3p\text{-}3}$ / *AtCLV3*$^{pro\text{-}8}$	*AtCLV3*$^{3p\text{-}3}$ + pro^{l}	3.31x10^{-120*}	redundancy

2.3.4 *SlCLV3* 3’ deletion alleles have weak effects on locule number

The division of CREs between the 5’ and 3’ of *Arabidopsis CLV3* prompted us to question how CRE organization in the distantly related tomato *CLV3* compared. Previously, we used CRISPR-Cas9 multiplex mutagenesis to generate a series of alleles harboring deletions within the proximal 2 kb non-coding region 5’ of *SlCLV3* (Rodríguez-Leal et al. 2017; Wang et al. 2021). These alleles produced the full spectrum of phenotypic diversity in terms of fruit locule number, including null-like phenotypes, suggesting that CREs within the 5’ may be sufficient to drive *SlCLV3* expression (**Fig. 2-4A, B**). These alleles revealed a general association between phenotypic severity and deletion of distal 5’ regions (Wang et al. 2021). More specific, targeted mutations within four different 5’ regions suggested the importance of interactions among these regions to drive *SlCLV3* expression, with the most distal region (referred to as R4 for Region 4) having the greatest impact on locule number when deleted individually (Wang et al. 2021). Although mutations 5’ of *SlCLV3* were sufficient to generate a full phenotypic spectrum, the prevalence of 3’ CREs among other genes, including *AtCLV3* and *WUS,* prompted us to ask whether the 3’ region of *SlCLV3* plays any role in its regulation.

We used multiplexed CRISPR-Cas9 mutagenesis to generate allelic diversity in the 3’ region of *SlCLV3.* We isolated 12 alleles with various 3’ perturbations, and phenotyped locule number (**Fig. 2-4C**). This collection of 3’ alleles perturbed multiple predicted transcription factor binding motifs through deletions and/or insertions. Alleles with minor sequence changes or very small deletions produced either no or extremely weak phenotypes (*SlCLV3*$^{3p\text{-}1\text{-}6}$). Alleles with larger perturbations within a region proximal to the stop codon, partially overlapping a precited 3’UTR, displayed weak changes in average locule number. Alleles *SlCLV3*$^{3p\text{-}9,10,11}$ had weak loss-of-function phenotypes, while allele *SlCLV3*$^{3p\text{-}7}$ is characterized by a small deletion and 125 bp insertion that results in a gain-of-function phenotype. This 125 bp insertion integrates another copy of the 3’ region between gRNA-1 and gRNA-2, doubling the presence of any putative regulatory sequence. Allele *SlCLV3*$^{3p\text{-}12}$ has a large deletion in the distal 3’ region, however a 106 bp insertion in the proximal 3’ region precludes

any interpretation of the resulting phenotype, which is not different from WT. Taken together, this collection of 3' alleles perturbs multiple predicted TF motifs through deletions and/or insertions, including some motifs with TFs known to overlap in expression with *CLV3*. These findings reveal a relatively minor contribution of 3' elements to the regulation of CLV3 function in tomato.

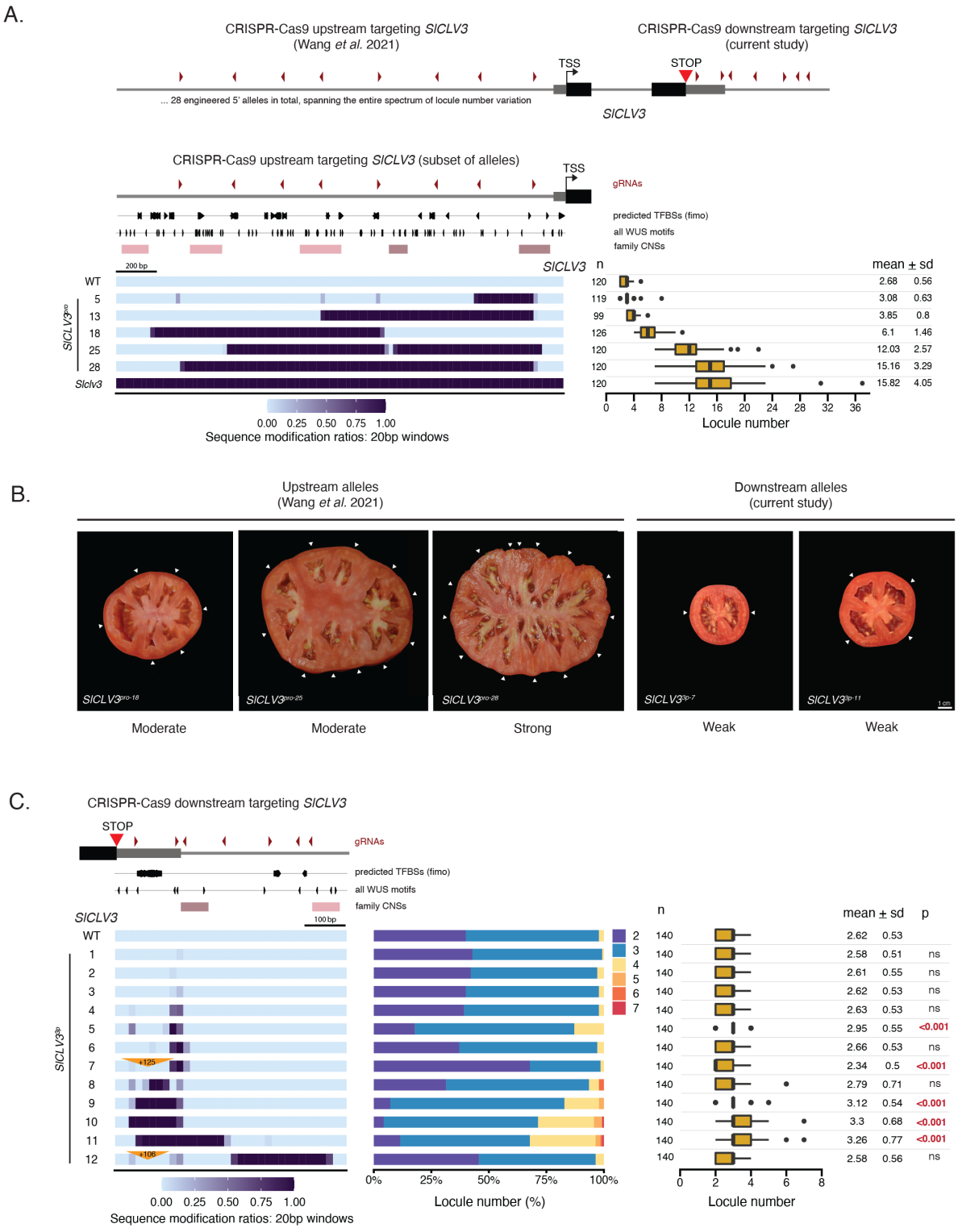
Figure 2-4. *SlCLV3* 3' deletion alleles have weak effects on locule number.

A. To facilitate comparison between tomato and *Arabidopsis*, this panel reproduces previous analysis of the 5' non-coding region of *SlCLV3*. At the top is a schematic depicting the gRNA arrays used to engineer mutations in the *SlCLV3* 5' and 3' non-coding regions using CRISPR-Cas9. gRNAs spanning the 5' of *SlCLV3* previously generated an allelic series of 28 5' mutants, with weak, moderate, and severe effects on tomato locule number. An encoded representation of a subset of these 28 alleles are shown below (Wang et al. 2021). Predicted TFBSs are indicated by black arrows. Family CNSs identified in Fig. 2-1 are represented on the sequence in red. Locule number quantifications are represented by stacked bar plots and box plots.

B. Representative images of tomatoes generated from 5' and 3' *SlCLV3* targeting.

C. Encoded representation of the 12 alleles generated from targeting the 3' region of *SlCLV3* with a 7-gRNA array, and their locule number quantifications. A two-sided Dunnett's compare with control test was performed to compare all 3' alleles to WT.

Figure 2-4.


A.
CRISPR-Cas9 upstream targeting SlCLV3
(Wang et al. 2021)
CRISPR-Cas9 downstream targeting SlCLV3
(current study)
TSS
STOP
... 28 engineered 5' alleles in total, spanning the entire spectrum of locule number variation
SlCLV3
CRISPR-Cas9 upstream targeting SlCLV3 (subset of alleles)
gRNAs
predicted TFBSs (fimo)
all WUS motifs
family CNSs
200 bp
WT
5
13
18
25
28
Slclv3
SlCLV3pro
0.00 0.25 0.50 0.75 1.00
Sequence modification ratios: 20bp windows
n
mean ± sd
120 2.68 0.56
119 3.08 0.63
99 3.85 0.8
126 6.1 1.46
120 12.03 2.57
120 15.16 3.29
120 15.82 4.05
0 4 8 12 16 20 24 28 32 36
Locule number
B.
Upstream alleles
(Wang et al. 2021)
Downstream alleles
(current study)
SlCLV3pro-18
SlCLV3pro-25
SlCLV3pro-28
SlCLV3sp-7
SlCLV3sp-11
1 cm
Moderate
Moderate
Strong
Weak
Weak
C.
CRISPR-Cas9 downstream targeting SlCLV3
STOP
gRNAs
predicted TFBSs (fimo)
all WUS motifs
family CNSs
SlCLV3
100 bp
WT
1
2
3
4
5
6
7
8
9
10
11
12
+125
+106
SlCLV3sp
0.00 0.25 0.50 0.75 1.00
Sequence modification ratios: 20bp windows
0% 25% 50% 75% 100%
Locule number (%)
2
3
4
5
6
7
n
mean ± sd
p
140 2.62 0.53
140 2.58 0.51 ns
140 2.61 0.55 ns
140 2.62 0.53 ns
140 2.63 0.53 ns
140 2.95 0.55 <0.001
140 2.66 0.53 ns
140 2.34 0.5 <0.001
140 2.79 0.71 ns
140 3.12 0.54 <0.001
140 3.3 0.68 <0.001
140 3.26 0.77 <0.001
140 2.58 0.56 ns
0 2 4 6 8
Locule number

2.3.5 Combined mutations in the 5' and 3' regions of *SlCLV3* have both additive and non-additive effects on fruit locule number

The results above hint at a divergence in the positioning of critical *CLV3* regulatory regions between *Arabidopsis* and tomato. However, despite the overall lack of non-coding sequence alignment between distantly related homologs, short 10-30 bp sequences of similarity can often still be discovered, albeit in drastically altered arrangements (Hare et al. 2008; Wong et al. 2020; Hendelman et al. 2021). This prompted us to perform a cross-species analysis for short regions of similarity between the entire 5' and 3' regions of *SlCLV3* and *AtCLV3.* This analysis identified a conserved 27 bp sequence shared between the R4 region of the *SlCLV3* 5' (a region previously characterized), and a 3' WUS binding site in the *AtCLV3* 3' region (**Fig. 2-5A**). Notably, the ATTA motif that WUS is known to weakly bind to is completely conserved within the 27 bp sequence.

Evidence of strongly synergistic interactions between *AtCLV3* 5' and 3' regions prompted us to question the nature of these interactions in tomato. Despite the finding that large 5' deletions alone are able to recapitulate a null phenotype, mutations 3' of *SlCLV3* do have a weak effect on locule number compared to WT. This suggests the possibility that 3' CREs may interact redundantly and/or additively/synergistically with specific 5' CREs, which may only be revealed by combining 3' mutations with smaller 5' mutations isolated to specific regions. From a previous dissection of *SlCLV3* 5'regions, it was found that mutations in a distal region of the *SlCLV3* 5' (R4), as well as a proximal region (R1), had a weak effect on tomato locule number when each was deleted individually (Wang et al. 2021). Mutations in the R4 and R1 regions also interacted to enhance locule number, additively and synergistically. Given the phenotypic relevance of these regions, as well as the conserved 27 bp element within the R4 region, we questioned whether R4 and/or R1 might also interact with regions 3' of *SlCLV3* to enhance locule number.

First, we performed sequential CRISPR-Cas9 targeting of the *SlCLV3* 3' region in the background of a R4 mutant (R4-5). This generated 10 alleles with combinations of mutations in the

SlCLV3 5' and 3' (**Fig. 2-5B**). Half of these alleles had small indel mutations, with little or no effect on locule number compared to the R4-5 mutant. However, four alleles clearly displayed enhanced locule number compared to R4-5 (R4 + 3p^{g}, R4 + 3p^{h}, R4 + 3p^{i}, and R4 + 3p^{j}). Interaction tests were performed to investigate the type of interaction between three of these 5'+3' alleles, R4 and the most similar single 3' mutants (**Fig. 2-5E, Supplementary 2-2**). All of these tests suggested an additive relationship between R4 mutants and 3' mutants of *SlCLV3.*

We also performed sequential mutagenesis of the R4 or R1 region in the background of the weak 3' allele *SlCLV3*$^{3p\text{-}11}$. Combined R4 and 3' mutations had enhanced locule number compared to individual mutants, however sequential mutations in the R4 region were different than the R4-4 allele, preventing a more meaningful analysis of interaction type (**Fig. 2-5C**). In contrast, combined R1 and 3' mutations had various interactions (**Fig. 2-5D, Supplementary 2-2**). The 3' deletion allele combined with partial deletions in the R1 region (allele *SlCLV3*$^{3p\text{-}11}$ + pro^{d}) did not show an enhancement in locule number compared to *SlCLV3*$^{3p\text{-}11}$. However, the 3' deletion allele combined with a full deletion of the R1 region had an enhancement in locule number that was mildly synergistic (allele *SlCLV3*$^{3p\text{-}11}$ + pro^{e}). Thus, tomato 5' and 3' regions demonstrate both additive and non-additive interactions depending on the 5' region evaluated.

Figure 2-5. Combined mutations in the 5' and 3' regions of *SlCLV3* have both additive and non-additive effects on fruit locule number.

A. Plant PAN3.0 cross species analysis was used to search for short sequences of similarity between the non-coding sequences of tomato and *Arabidopsis CLV3,* revealing a conserved 27 bp sequence which overlaps with the distal R4 region in the tomato 5' (outlined by a purple dashed box), and a known WUS TFBS in the *Arabidopsis* 3' (outlined by a blue dashed box). The DNA sequence in these two regions are shown, with the 27 bp sequence colored in red, with nucleotide mismatches highlighted in red, and the core ATTA WUS binding element in bold. All five of the previously characterized *AtCLV3* 3' WUS binding elements are also bolded and named according to their position, as defined previously (Perales et al. 2016).

B. Encoded representation of alleles generated from sequential CRISPR-Cas9 targeting with the *SlCLV3* 3'-gRNA array, in the background of the fixed R4-5 mutant. Predicted TFBSs are indicated by black arrows. Family CNSs identified in Fig. 2-1 are represented on the sequence in red. Locule number quantifications are represented by stacked bar plots and box plots. Asterisks denote alleles that were tested for interaction effects. The R4 and R1 regions previously defined are highlighted by purple boxes on the *SlCLV3* 5' non-coding sequence (Wang et al. 2021). A two-sided Dunnett's compare with control test was performed to compare WT and all alleles to R4-5.

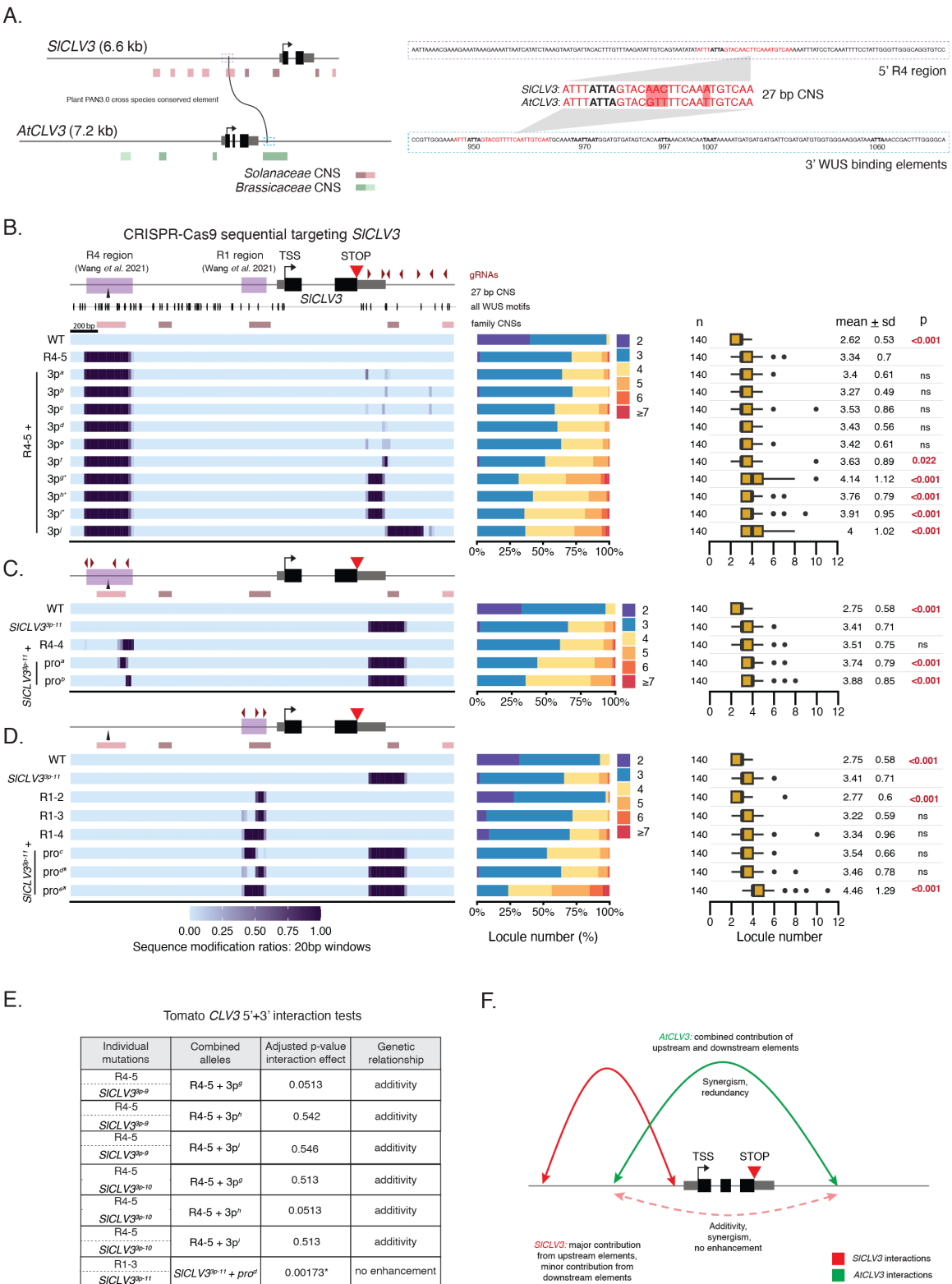
C. Encoded representation of alleles generated from sequential CRISPR-Cas9 targeting with the *SlCLV3* R4-gRNA array, in the background of the fixed $SlCLV3^{3p\text{-}11}$ allele, and their locule number quantifications. A two-sided Dunnett's compare with control test was performed to compare WT and all alleles to $SlCLV3^{3p\text{-}11}$.

D. Encoded representation of alleles generated from sequential CRISPR-Cas9 targeting with the *SlCLV3* R1-gRNA array, in the background of the fixed $SlCLV3^{3p\text{-}11}$ allele, and their locule number quantifications. A two-sided Dunnett's compare with control test was performed to compare WT and all alleles to $SlCLV3^{3p\text{-}11}$.

E. Interaction tests performed between select individual 5' and 3' mutants and similar combined 5'+3' alleles. P-values of the interaction effect were adjusted for multiple comparisons.

F. Model summarizing the relative contribution of the 5' and 3' region, as well as their interactions, to the regulation of *SlCLV3* and *AtCLV3.*

Figure 2-5.



Individual mutations	Combined alleles	Adjusted p-value interaction effect	Genetic relationship
R4-5 / *SlCLV3*$^{3p-9}$	R4-5 + 3p^{g}	0.0513	additivity
R4-5 / *SlCLV3*$^{3p-9}$	R4-5 + 3p^{h}	0.542	additivity
R4-5 / *SlCLV3*$^{3p-9}$	R4-5 + 3p^{i}	0.546	additivity
R4-5 / *SlCLV3*$^{3p-10}$	R4-5 + 3p^{g}	0.513	additivity
R4-5 / *SlCLV3*$^{3p-10}$	R4-5 + 3p^{h}	0.0513	additivity
R4-5 / *SlCLV3*$^{3p-10}$	R4-5 + 3p^{i}	0.513	additivity
R1-3 / *SlCLV3*$^{3p-11}$	*SlCLV3*$^{3p-11}$ + pro^{d}	0.00173*	no enhancement
R1-4 / *SlCLV3*$^{3p-11}$	*SlCLV3*$^{3p-11}$ + pro^{e}	0.0145*	synergism

2.4 Discussion

Despite a high degree of functional conservation, our study provides evidence for divergent regulatory strategies between two distantly related *CLV3* orthologs, with substantial alterations in regulatory sequences, their spatial arrangement, and their relative effects on *CLV3* regulation (**Fig. 2-5F).** Using CRISPR-Cas9 mutagenesis, we functionally dissected the contribution of 5' and 3' regions to the regulation of tomato and *Arabidopsis CLV3.* In tomato, our results here and in previous work show a significant contribution to *SlCLV3* regulation by elements in the 5' non-coding region, with only a minor contribution from the 3' non-coding region. Although mutagenesis of a large section of the *SlCLV3* 5' is sufficient to produce a null-like phenotype, the 3' region still plays a role in regulation, and interacts additively and synergistically with specific 5' regions to produce enhanced or non-enhanced locule number phenotypes. In contrast, we show that *AtCLV3* regulation seems to depend on a more even division of functional CREs between the 5' and 3' non-coding regions, with weak effects on locule number from individual 5' and 3' mutations that have a strongly synergistic effect in combination. At least one short 27 bp sequence containing a WUS binding site is conserved between tomato and *Arabidopsis CLV3*, albeit on opposite sides of the gene, suggesting a potential mechanism for similar spatiotemporal expression. Therefore, our results support a model of *cis*-regulatory sequence evolution in which CRE organization is malleable to change, while short sequences required for the binding of specific transcription factors are conserved. This is consistent with the billboard model of enhancer architecture, as well as the results of similar studies in animals (Hare et al. 2008; Cameron and Davidson 2009; Wong et al. 2020).

Although it is clear that the non-coding sequence and regulatory grammar of conserved genes is often not conserved between divergent species, it is still unclear how mechanisms of gene regulation are able to tolerate this extreme shuffling. Transcription factors still need to control the rate of transcription in the correct cell type, and at the correct time during development. For a dosage-sensitive gene like *CLV3,* the precise level of gene expression is likely important to maintain to some

degree. One previously proposed hypothesis is that although the binding sites of specific TFs are shuffled, their general genomic position (5', 3', intronic, exonic, within a UTR) relative to the gene may be conserved, thus preserving relative genetic and physical interactions among TFBSs and promoters. There is some evidence for this – for example, in the study of *Drosophila* and sepsid *eve* enhancers, although sequence was not conserved, relative genomic positioning was (Hare et al. 2008). Similarly, a study of five orthologous enhancers of developmental genes in mosquito and flour beetle (~333 MY diverged) found that their genomic positions are conserved (Cande et al. 2009a). For example, despite sequence divergence, an orthologous enhancer of the *cactus* gene is located within an intron in both species. In contrast, enhancers of the *yellow* gene have different genomic positions in various *Drosophila* species, however these species do have species-specific pigment patterns (Kalay and Wittkopp 2010). These findings suggested that more constrained developmental expression patterns may require conserved enhancer positioning, whereas genes with rapidly evolving expression patterns have less constrained enhancer positioning. Inconsistent with this view, our results suggest that the genomic position of CREs was shuffled between the 5' and 3' of *Arabidopsis* and tomato *CLV3* during evolution, despite the importance of *CLV3* expression to meristem development. One explanation for our finding is that it is possible for species to maintain expression patterns and evolve new CRE positioning in specific circumstances, as long as they also evolve new mechanisms of communicating with the promoter effectively (such as through physical looping of the DNA, for example).

Over the course of domestication, mutations within CREs have led to quantitative variation in agriculturally relevant traits. Notably, the enlargement of tomato fruit size was facilitated by a synergistic interaction between mutations in 5' and 3' CREs of two genes controlling meristem proliferation (van der Knaap et al. 2014). The prediction of CREs using conservation analyses could therefore serve as a tool for further crop bioengineering. However, our study underlines the limitations of this approach, especially in the detection of CREs across very large distances. Although short sequences for functional transcription factor binding sites may be highly conserved, other

features, including their relative importance to gene regulation, their positioning relative to the gene, and their interactions with other regulatory elements may diverge, as demonstrated in this study. Furthermore, while functional TFBSs may often be conserved across deep time, their small size, variable weight of importance of specific residues, and altered sequence context make them difficult to detect. Simplistic or repetitive TFBS sequences are missed by many alignment algorithms, due to the high frequency of these sequences that occur by chance. For example, a core WUS TFBS is composed of the sequence ATTA, which occurs frequently in non-coding DNA (Sloan et al. 2020). Many algorithms designed to detect CNSs discard AT repeats longer than a given size, which in our case could prevent detection of conservation within both the *SlCLV3* R4 region and *AtCLV3* 3' WUS binding site region (Hendelman et al. 2021). Experiments such as ChIP-seq may be required to overcome this challenge, and such an experiment for tomato WUS would help further clarify a mechanism for *SlCLV3* regulation. Nonetheless, sequence alignments between members of the same family may still be a useful tool for predicting CREs. For example, a region adjacent to the *AtCLV3* 3' WUS binding site is conserved among several *Brassicaceae* species, and clearly contributed to enhanced locule number when mutated (**Fig. 2-3C).** In lieu of developing new algorithms or deep learning approaches for the detection of CNSs, it is clear that we cannot solely rely on conservation to predict enhancers in diverged species.

In the future, identifying all of the TFBSs regulating *CLV3* in *Arabidopsis* and tomato would clarify the extent of CRE shuffling between these orthologs. WUS transcription factor binding and regulation has been previously demonstrated as a mechanism involved in *AtCLV3* regulation, and multiple WUS motifs can be found in *SlCLV3* 5' and 3' regions, including the 27 bp conserved element. However, there are dozens of transcription factor motifs within all of these regions, which could easily work in tandem with WUS to fully regulate *CLV3.* For example, the 3' alleles $SlCLV3^{3p\text{-}7,9,10}$ all perturb one ATTA element, yet $SlCLV3^{3p\text{-}7}$ has a weak decrease in locule number, and $SlCLV3^{3p\text{-}9}$ and $SlCLV3^{3p\text{-}10}$ both have a weak increase in locule number. This suggests that additional copies of other motifs within the $SlCLV3^{3p\text{-}7}$ insertion have the opposite effect on locule number as loss of those

motifs in the *SlCLV3*$^{3p\text{-}9}$ and *SlCLV3*$^{3p\text{-}10}$ deletion alleles. Indeed there are multiple motifs within this region, including those for MADS-box TFs that are known to regulate floral development (Wang et al. 2019). In the future, more precise experiments targeting particular TFBSs *in vivo* will provide improved motif validation, as CRISPR technologies such as base and prime editing improve in efficiency. Furthermore, post-transcriptional mechanisms may also be at play, given that the 3' non-coding region of *SlCLV3* overlaps with the 3'UTR of some transcripts, raising the possibility of effects on transcript stability and/or translation (Mayr 2019).

Our findings suggest that genetic interactions between 5' and 3' regions may be a common and important mechanism of gene regulation. Such interactions may also explain regulation of other genes, such as the tomato *WUS* (*SlWUS*) gene. Previously, large deletions we generated within the 5' of *SlWUS* did not have an effect on locule number alone (Wang et al. 2021). A locule number QTL called *lc* is known to be caused by two SNPs that disrupt a MADS-box motif downstream of *SlWUS* (Muños et al. 2011). This evidence indicates a likely division of CREs between the *SlWUS* 5' and 3' region, suggesting the hypothesis that both regions may need to be mutated in combination to affect meristem proliferation. Going forward, studies of *cis*-regulatory regions should include CREs in all genomic contexts, including 5', 3', and regions within the genes such as introns and UTRs.

While we focused primarily on genetic relationships between 5' and 3' mutations in this study, we have not dismissed the possibility that physical interactions between these regions may also be at play. Animal genomes are known to adopt specific 3D chromatin conformations that impact gene regulation (Lieberman-Aiden et al. 2009). Many animal genes are regulated by distal enhancer elements that associate with gene-proximal regions via looping (Dong et al. 2020). The genomes of many crops also form these 3D associations, although their impact on gene regulation is less clear (Dong et al. 2017). Interestingly, higher order chromatin interactions seem to be absent from the much smaller *Arabidopsis* genome, although Hi-C as well as 3C studies have revealed the large scale presence of gene loops, which in at least a number of examples have a direct influence on gene expression (Crevillén et al. 2013; Liu et al. 2013, 2016). Physical looping between 5' and 3' regions

presents a potential mechanism for the coordination of multiple CREs in gene regulation. It would be interesting to further explore looping at gene-level resolution in the future, with a technique such as 3C or capture-Hi-C, both in wild type and 5'/3' mutant genotypes. Additionally, the role of long-range enhancer elements is still missing from these functional dissections of regulatory elements in plants. We would benefit from genome-wide experiments, such as Hi-C, to further our understanding of the complexity of gene regulation in plants.

For the sake of this study, we have focused on the similarities between tomato and *Arabidopsis CLV3* function and expression, however it is interesting to consider whether regulatory sequence divergence may help to explain some slight observed differences between the two genes. For example, absence of *SlCLV3* expression from the L1 layer could be explained by altered WUS binding dynamics, due to reduced number or modified spacing of WUS TFBSs. Indeed, reporter experiments in *Arabidopsis* found that ablation of specific 3' WUS binding sites could eliminate GUS expression in the L1 layer of the SAM (Perales et al. 2016). In the future it would be interesting to drive expression of a reporter gene with *SlCLV3* 5' and 3' regions in *Arabidopsis*, or vice versa, to see if expression of the reporter overlaps with that of tomato or *Arabidopsis CLV3.* There is also a phenotypic difference between average locule number in WT tomato and *Arabidopsis*. Namely, *Arabidopsis thaliana* locule number is canalized at two, while tomato locule number varies between two and three. Differences in regulatory sequence could subtly alter either the expression level or domain of *CLV3* expression to create this variance in tomato compared to *Arabidopsis.* It would therefore be interesting to determine if *AtCLV3* driven by *SlCLV3* regulatory regions disrupts locule number canalization in *Arabidopsis.* These reporter experiments would be a nice complement to the functional, *in vivo* work of this study.

In conclusion, our results demonstrate the capacity for substantial shuffling of CREs during the course of evolution, even between upstream and downstream regions. The genomes of plants have been subject to many rearrangements and intervening mutations, most notably through the expansive proliferation of transposable elements during evolution and domestication (Ricci et al. 2019). It is

possible that these intervening sequences necessitate a certain degree of CRE flexibility. From our research, it is clear that regulatory regions are riddled with complexity, including interactions among multiple CREs that can be additive, synergistic, or redundant in nature. Somehow, amidst substantial rearrangements, this complexity is maintained at its core, through evolutionary mechanisms of selection, to ensure that specific expression patterns vital to survival are maintained. This is fundamentally different to how gene sequences evolve, since they are constrained by specific mutations that alter protein structure (and thus function) (Koonin and Wolf 2010). We are only beginning to understand the constraints on CRE evolution to reproduce vital expression patterns. This study offers a new perspective on this question for all eukaryotes, through *in vivo* engineering of *cis*-regulatory alleles. In the future, more of these deep functional dissections of regulatory regions in an evolutionary context will begin to uncover potential rules for the maintenance of gene expression patterns.

2.5 Methods

2.5.1 Plant material, growth conditions and phenotyping

Seeds of *Solanum lycopersicum* cv. M82 from our own stocks were used as the background for WT and CRISPR-Cas9 tomato mutagenesis experiments. During initial allele isolation, tomato plants were sown and grown in 96-well flats for ~4 weeks before being transplanted to pots, and grown in greenhouse conditions. The greenhouse operates under long days (16h light, 8h dark) with natural and artificial light (from high pressure sodium bulbs ~250 umol/m^2), at a temperature between 26-28°C (day) and 18-20°C (night), with relative humidity 40-60%. For phenotyping, tomato plants were sown and grown in 96-well flats before being transplanted to Uplands field at Cold Spring Harbor Laboratory. Plants in the field were grown under drip irrigation and standard fertilizer regimes. For each unique genotype, locule number was quantified for 140 fruits, taken from 7-12 individual plants. Seeds of *Arabidopsis thaliana* (ecotype Col-0) from our own stocks were used as the background for WT and CRISPR-Cas9 *Arabidopsis* mutagenesis experiments. Arabidopsis plants

were germinated on ½ MS plates and transplanted to 32-well flats for growth. During initial allele isolation, plants were grown in growth chambers under long days (16h light, 8h dark) at 22°C and light intensity ~100 umol/m^2. For phenotyping, *Arabidopsis* plants were grown on ½ MS plates in a growth chamber for 1 week (continuous light, 22°C, ~100 umol/m^2) before being transplanted to 32-well flats and grown in greenhouse conditions. The greenhouse for *Arabidopsis* growth operates under long days (16h light, 8h dark) with natural and artificial light, at a temperature between 20-25°C. For each unique genotype, locule number was quantified using the stereo microscope for 140 siliques, taken from 7-10 individual plants.

2.5.2 CRISPR-Cas9 mutagenesis, plant transformation, and selection of mutant alleles

Generation of transgenic tomato with CRISPR-Cas9 mutagenesis was performed as previously described (Brooks et al. 2014). Briefly, gRNAs were designed with Geneious Prime (https://www.geneious.com). The Golden Gate assembly method was used to clone gRNAs into a binary vector with Cas9 and kanamycin selection (Rodríguez-Leal et al. 2017; Werner et al. 2012). Binary vectors were introduced into tomato plants through *Agrobacterium tumefaciens* mediated transformation in tissue culture (Van Eck et al. 2019a). Transgenic plants were screened for mutations using PCR primers surrounding the gRNA target sites. PCR products were screened for obvious shifts in size by gel electrophoresis, and mutations were characterized by Sanger sequencing. First or second generation transgenics (T0 or T1) were backcrossed to WT to eliminate the Cas9 transgene and purge the genome of potential off-target mutations. F2 or F3 plants from these crosses that were homozygous for the CRISPR-induced mutation were used for phenotypic analysis. Generation of binary vectors for *Arabidopsis* CRISPR-Cas9 mutagenesis also utilized the Golden Gate assembly method. *Arabidopsis* constructs used an intronized Cas9 previously demonstrated to increase editing efficiency (Grützner et al. 2021). The intron-Cas9 (L0 pAGM47523) was cloned with RPS5a promoter (L0 pICH41295) and NOS terminator sequence (L0 pICH41421) into the L1 plasmid

pICH47822. This was assembled into the L2 vector pAGM4723 with NPTII for kanamycin resistance (pICSL70004 in L1 pICH47732), pFAST-R selection cassette (pICSL70008 in L1 pICH47742), and the gRNAs (each with U6 promoter and gRNA scaffold). *Arabidopsis* plants were transformed with binary vectors using *Agrobacterium tumefaciens* floral dip (Zhang et al. 2006). Transgenic seed was selected by fluorescence, germinated on ½ MS plates, and transferred to soil, after which plants were subjected to a heat cycling regime that fluctuated between 37°C for 30 h and 22°C for 42 h over the course of 10 days. This protocol was previously described to increase Cas9 editing efficiency in *Arabidopsis* (LeBlanc et al. 2018). Following heat treatment, flower DNA was genotyped for mutations in the target region, and individuals with evidence of editing were selected to be grown in the next generation for screening of plants that were Cas9 negative with stabilized mutations. T3 or T4 plants homozygous for the CRISPR-induced mutation were used for phenotypic analysis. All gRNA and primer sequences are listed in Supplementary Table 1-3.

2.5.3 *Cis*-regulatory sequence conservation analyses, TFBS prediction, and Plant PAN3.0 cross species analysis

Within-family conservation analysis was performed to predict conserved non-coding sequences within the 5' and 3' of *CLV3* in *Arabidopsis* and tomato that were shared among several *Brassicaceae* and *Solanaceae* species, respectively. The closest *CLV3* ortholog from each species was determined based on the ortholog with the greatest similarity to *Arabidopsis* or tomato *CLV3* within the 5' and 3' regions. 40 kb of sequence upstream and downstream of the *CLV3* ortholog was extracted, and aligned to *Arabidopsis* or tomato *CLV3* using mVISTA Shuffle-LAGAN (http://genome.lbl.gov/vista/mvista/submit.shtml) (Frazer et al. 2004). Conservation was calculated in 20 bp windows, with an 85% similarity threshold. TFBSs were predicted by scanning the *Arabidopsis* and tomato *CLV3* 5' and 3' regions for motifs using FIMO in the MEME suite (http://meme-suite.org/doc/fimo.html) (Grant et al. 2011). Position frequency matrices for known plant transcription factors were obtained from the JASPAR CORE PFMs of plants collection 2022 (Castro-

Mondragon et al. 2022). A p-value cutoff of 0.00001 was used to predict TFBSs. To search for short, conserved non-coding sequences shared between *Arabidopsis* and tomato *CLV3,* the Plant Promoter Analysis Navigator (PlantPAN) 3.0 cross species analysis function was used (http://PlantPAN.itps.ncku.edu.tw) (Chow et al. 2019). The *Arabidopsis* and tomato *CLV3* gene with 5' and 3' regions were used as input.

2.5.4 Statistical methods

Pairwise comparisons between various alleles were performed using two-sided Dunnett's compare with control tests. A p-value cutoff of <0.05 was used. For testing the genetic interaction between 5' and 3' mutations, a linear model was used. Each four-way comparison (between WT, single 5' allele, single 3' allele, and the combined 5'+3' allele) was modelled with a linear model in R with interaction effect included. A p-value of <0.05 was used as a cutoff for a significant interaction effect. P-values were adjusted for multiple comparisons using the Benjamini-Hochberg method in R.

Chapter 3: *Cis*-regulatory elements controlling flowering time divergence in wild and domesticated tomato

3.1 Summary

Phenotypic divergence during the course of evolution is thought to frequently derive from variation within *cis*-regulatory elements, rather than coding sequence variation which is more likely to have negative pleiotropic effects. Divergence in flowering time response to daylength between wild and domesticated tomato species was recently attributed to expression differences in the dosage-sensitive anti-florigen gene *SELF PRUNING 5G (SP5G).* Wild species of tomato flower earlier under short days, while domesticated species are daylength neutral. The specific CREs underlying this trait variation are unknown, however a 52 bp enhancer within the 3'UTR of daylength sensitive species, and absent in daylength neutral species, was suggested as a candidate. Previous studies of phenotypic variation have relied on indirect methods of validating the effects of CRE variation, such as reporter assays and association analyses. Therefore, we decided to take a functional genetics approach to study the *cis*-regulatory determinants of phenotypic divergence directly (through their phenotypic outputs), using *SP5G* control of flowering time as our model gene/trait. We used CRISPR-Cas9 to engineer 18 unique mutations upstream (5') and downstream (3') of *SP5G* in the daylength sensitive introgression line IL5-4, which carries the *Solanum pennellii* (wild) variant of the *SP5G* locus, in the background of *Solanum lycopersicum* cv. M82 (a domesticated species). We found that deleting most of the predicted 52 bp enhancer element in the 3'UTR did not dramatically alter flowering time under long days, while large mutations within the non-coding region upstream generated weak and moderate flowering time phenotypes. Targeting conserved ATAC-seq peaks also had variable effects, including no effect, weak effects, and one allele with a strong effect on flowering time. Taken together, no single *cis*-regulatory allele that we generated matched the flowering time response of domesticated tomato, and multiple CREs are clearly involved in *SP5G* regulation. Thus, it is possible that multiple mutations within CREs were required to generate phenotypic divergence in flowering time, and in the

future we will explore this hypothesis with further experiments aimed at exploring genetic and physical interactions among CREs in the regulation of *SP5G.* Findings from many of our studies of *cis*-regulatory regions *in vivo* suggest that regulatory regions are often robust, containing multiple CREs interacting in various ways to mediate proper expression patterns. Thus, multi-step mutations during the course of evolution may often be required to elicit substantial phenotypic divergence for selection to act upon.

3.2 Introduction

The process of flowering - transitioning from the vegetative to the reproductive phase - is regulated by a combination of genes and environmental inputs in flowering plants. Plants are able to sense the daylength in a particular geographical region through the use of photoreceptors (phytochromes and cryptochromes), and integrate this information in order to trigger changes in expression of the flowering hormone florigen (Osnato et al. 2022). Some plants are day-neutral, flowering at the same stage despite the relative number of light/dark hours. Others, such as *Arabidopsis thaliana*, are long-day plants, flowering earliest when the number of daylight hours exceed a certain amount. Alternatively, wild species of tomato are short-day plants, flowering earlier with fewer hours of daylight. Variations in light sensitivity are necessary to facilitate the fine-tuning of flowering time to specific environments, thereby triggering reproduction during ideal periods.

In plants, the CETS (CENTRORADIALIS/TERMINAL FLOWER 1/SELF-PRUNING) gene family encodes phosphatidylethanolamine binding proteins (PEBPs), and is responsible for controlling important developmental transitions, including flowering (McGarry and Ayre 2012). It is the balance between florigens and anti-florigens belonging to this family that fine-tunes the initiation of flowering. In *Arabidopsis*, the PEBP gene *FLOWERING LOCUS T (FT)* is the major downstream component responsible for inducing flowering (i.e. florigen), and is regulated by photoperiod. The zinc-finger transcription factor CONSTANS (CO) is stabilized in long days in *Arabidopsis,* and induces *FT* expression (Suárez-López et al. 2001). Upon expression, FT protein translocates from the

leaves to the meristem through phloem, where it forms a complex with the transcription factor FD and scaffolding protein 14-3-3 (Wigge et al. 2005). This complex is thought to bind to promote the expression of several genes, including floral meristem identity genes such as *APETALA1 (AP1)* and *LEAFY (LFY)* (*FALSIFLORA* in tomato). This general module is conserved in some short-day plants, such as rice, except the CO homolog is stabilized in short days rather than long days (Ishikawa et al. 2011). In addition to transcriptional activators such as CO in *Arabidopsis*, flowering is antagonistically regulated by a variety of repressors, such as FLC (controlling the flowering response to temperature), SHORT VEGETATIVE PHASE (SVP), and CYCLING DOF FACTOR 1 (CDF1) (Osnato et al. 2022). The tomato genome contains 13 PEBP genes, 6 of which are FT-like: SFT, SP6A, SP5G, SP5G1, SP5G2, and SP5G3 (Cao et al. 2016). SINGLE FLOWER TRUSS (SFT) is the main floral inducer in tomato (i.e. tomato florigen), while SP5G, SP5G2, and SP5G3 are known inhibitors of flowering (Molinero-Rosales et al. 2004). Interestingly, CO-like proteins in tomato do not effect flowering time when overexpressed, suggesting that day-length sensitivity in tomato is controlled by different genes than those in *Arabidopsis* and rice (Ben-Naim et al. 2006).

Domesticated varieties of tomato are day-neutral, a trait selected for when tomato began to be cultivated at Northern latitudes. The day-length neutral flowering response of domesticated tomato was recently mapped to two genes, one controlling the flowering response in short days and one in long days. The PEBP gene *FTL1* (an FT-paralog) regulates daylength-neutrality in short days (Song et al. 2020). A coding sequence mutation in *FTL1* generated a truncated protein in domesticated tomato, leading to flowering two leaves later than wild species under short days. The other gene, *SP5G*, is an anti-florigen PEBP, and contributes to daylength-neutrality by reducing the long-day response in domesticated tomato (Soyk et al. 2017). Daylength sensitivity in wild tomato species is based on differences in *SP5G* expression, which is comparatively elevated in wild species under long days (most strikingly in the cotyledons). Notably, this elevated gene expression may be partially mediated by a looping interaction between the promoter and a 52 bp sequence in the 3'UTR (Zhang et al. 2018). Domesticated day-neutral tomatoes have lost this 52 bp enhancer and have a weakened

looping interaction, a proposed explanation for flowering earlier than wild species under long days. Indeed, these kinds of looping interactions seem to be widespread in some plants, such as numerous gene loops in *Arabidopsis,* including the FLC gene loop that regulates flowering in response to cold (Crevillén et al. 2013). Thus, domesticated tomato flowers similarly in short days and long days (i.e. is day-neutral) due to loss of *FTL1* function and *SP5G* expression. Mechanistically, FTL1 is proposed to activate expression of the florigen *SFT*, while SP5G represses it (Soyk et al. 2017; Song et al. 2020).

The mechanisms by which phenotypic variability is acquired is of great interest to evolutionary biologists. While variation is often derived from coding sequence mutations affecting protein function, such as the case with *FTL1*, more often QTL are mapped to intergenic regions (Wittkopp and Kalay 2012; Meyer and Purugganan 2013; Albert and Kruglyak 2015; Han et al. 2018b; Ricci et al. 2019). Throughout evolution, phenotypic variation has often been derived from *cis*-regulatory mutations, rather than coding sequence mutations that can have severe fitness consequences. The *SP5G* gene in domesticated and wild tomato species presents a model system to study this question in plants. It seems likely that *cis*-regulatory mutations are responsible for altered *SP5G* expression in domesticated tomatoes. Researchers comparing both the coding sequence and *cis*-regulatory DNA surrounding *SP5G* of daylength-neutral and daylength-sensitive tomato plants were unable to associate a single amino acid substitution with daylength sensitivity, and only the 52 bp sequence in the 3'UTR was found to consistently differ among the plants evaluated (Soyk et al. 2017; Zhang et al. 2018). However, additional *cis*-regulatory mutations could also have contributed to loss of daylength-sensitivity. For example, we could hypothesize that the acquisition of a transposable element could lead to reduced *SP5G* expression, as well as the loss of an enhancer, or gain of a silencer. Indeed, altered *SP5G* expression could conceivably come down to a single SNP in an important TFBS. It could also be attributed to a combination of *cis*-regulatory mutations that work additively or synergistically to alter *SP5G* expression. Although the 52 bp 3'UTR enhancer does

exhibit enhancer activity in reporter assays, it was not tested *in vivo,* and thus the extent of its importance to *SP5G* expression has not been verified.

To date, much of our understanding of phenotypic variation during the course of evolution has been derived from comparing naturally occurring alleles of the ancestral and derived trait. These studies often rely on indirect methods of validation, such as reporter or genomic assays that place CREs outside of their native context, or else provide information about chromatin state or TF binding that cannot in itself guarantee the functional, phenotypic relevance of *cis*-regulatory variation. Additionally, when comparing a particular trait that has diverged between different species, the comparison can potentially be confounded by many other variants between the species, especially cryptic variants that exist in one species but not the other. Thus, a more straightforward approach to the identification of CREs and their true impact on a trait of interest is a functional genetics approach – creating allelic diversity in the *cis*-regulatory region of a gene of interest *in vivo*, thus manipulating CREs in their native context, and providing a clear readout of their contribution to the trait through the phenotypic output of dosage dependent genes. We used CRISPR-Cas9 to target potential CREs controlling the expression of *SP5G* in an introgression line of the wild species *Solanum pennellii*, and observed the effect of various mutations on quantitative changes in flowering time. This approach also allowed us to explore potential CREs or combinations of CREs responsible for flowering time divergence during tomato domestication, which were not feasible to investigate through other approaches.

3.3. Results

3.3.1 Deletion of a predicted enhancer element in the 3'UTR of *SP5G* leads to slightly earlier flowering under long days

We are using the gene *SP5G* as a model to better understand *cis*-regulatory mechanisms of phenotypic divergence in evolution, as well as features of CREs and their genetic and physical interactions in the coordination of gene regulation. For our study, we compared the domesticated day-

neutral tomato species *Solanum lycopersicum cv.* M82, and the short-day wild species *Solanum pennellii.* However, since these two species have many different traits, and our goal was to focus on flowering time regulation by *SP5G* alone, we used an introgression line (IL) derived from a cross between *Solanum pennellii* and the domesticated cultivar M82 (Eshed and Zamir 1995). We performed all CRISPR experiments in IL5-4, an IL with a small chromosomal segment from *Solanum pennellii* in the background of M82. This introgression line carries the *SP5G* locus from *S. pennellii,* and thus provides a foundation to specifically modify the day-length sensitive allele of *SP5G,* and thus flowering time, in a standardized isogenic background. Under long days, IL5-4 flowers after about 14 leaves, while M82 flowers after eight (**Fig. 3-1A, B**). They both flower after about eight leaves under short days (Soyk et al. 2017).

Firstly, we performed a mVISTA alignment of M82 *SP5G (SlSP5G)* to *Solanum pennellii SP5G (SpSP5G*) (**Fig. 3-1C**). There are many regions of conservation, as expected for closely related species. Interestingly, all of the ATAC-seq peaks identified from M82 leaf and meristem tissue are highly conserved in *SpSP5G.* However, the 52 bp sequence in the 3'UTR is clearly not conserved, and large regions upstream, in the first intron, and proximally downstream have diverged in sequence between the two species. Although there is a large transposable element insertion in the first intron of *SpSP5G* relative to *SlSP5G,* it is not shared by other daylength-sensitive species (Soyk et al. 2017). Additionally, smaller SNPs and indels exist throughout larger blocks of conserved sequence. Alignment of *SpSP5G* to additional orthologs from the *Solanaceae,* spanning ~27 million years of evolution, revealed relatively deep conservation of the M82 ATAC-seq peaks and some proximal upstream regions, in contrast to the absence of conservation of either the 5'UTR or 3'UTR. Thus, there are many regions of *cis*-regulatory conservation and divergence, both useful predictors of potential CREs to functionally test with CRISPR-Cas9 *in vivo* editing.

Due to its prior characterization as an enhancer of *SP5G,* we aimed to determine its impact on flowering time by using a 3-guide CRISPR-Cas9 construct to delete the 52 bp sequence in the 3'UTR of *SpSP5G,* which is absent in M82. We were able to isolate one allele from this targeting, which

deleted an 87 bp sequence in the 3'UTR, encompassing 43 bp of the predicted enhancer (*SpSP5G*$^{enhancer\text{-}1}$) (**Fig. 3-1D**). Although the entire 52 bp was not deleted, a predicted CDF5 motif was included in the deletion, which is a TF known to regulate flowering time through various florigens, including *SP5G*. *SlCDF3* overexpression is known to delay flowering in long and short days in domesticated tomato by activating *SlSP5G* in long days, and activating *SlSP5G2* and *SlSP5G3* in short days (Xu et al. 2021). Plants with this enhancer mutation were phenotyped for flowering time under long days, from segregating populations to ensure an internal wild type control. *SpSP5G*$^{enhancer\text{-}1}$ homozygous mutants had a slightly earlier flowering time than their internal control (WT), a difference of one leaf (**Fig. 3-1D, E**). This CRISPR-Cas9 targeting also fortuitously generated a coding sequence mutation, allele *Spsp5g-1*, which is extremely early flowering under long days, and matches the loss of function phenotype of past *SP5G* null mutants generated in our lab in M82 (Soyk et al. 2017). From this experiment, we discovered that deleting most of the 52 bp enhancer did not even come close to recapitulating the flowering time phenotype of M82 (~eight leaves), suggesting that other CREs play a role in the regulation of *SP5G* expression. Of course, in the future it will be important to confirm this finding with a full deletion of the 52 bp.

Figure 3-1. Deletion of a predicted enhancer element in the 3'UTR of *SP5G* leads to slightly earlier flowering under long days.

A. Representative IL5-4 and M82 plants grown under long days (~16h light). White arrows denote individual leaves before the first inflorescence, and the number of leaves is indicated. "L" means leaves.

B. Flowering time of IL5-4 and M82 grown under long days. Sample number (n) is shown to the left, and mean and standard deviation (sd) are shown on the right. A two-tailed, two-sample t-test was performed to compare means, with the p-value displayed.

C. The *SlSP5G* gene sequence and surrounding non-coding DNA (from *Solanum lycopersicum* cv. M82), as well as several other *SP5G* orthologs from the *Solanaceae,* were aligned to the *SpSP5G* gene and its upstream and downstream region extending to the next closest genes, using mVISTA. Conserved regions were defined as regions with 70% similarity in a 100 bp window. Conserved UTRs are light blue, and conserved exons are dark blue. The proposed 52 bp enhancer described in Zhang et al. is highlighted in yellow. Predicted TFBSs are shown in black. The position of conserved ATAC-seq peaks discovered in M82 leaf (green) and meristem (pink) tissue are denoted on the *S. pennellii* sequence. Purple regions highlight the degree of conservation of each ATAC-seq peak among the various *SP5G* orthologs. The position of gRNAs used to mutagenize the 5' or 3' of *SpSP5G* are depicted on the sequence as grey arrows.

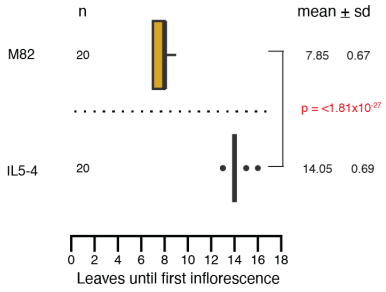
D. Encoded representation of the alleles generated by targeting the proposed 52 bp 3'UTR enhancer of *SpSP5G* with a 3-guide CRISPR-Cas9 array in IL5-4. The alleles have been encoded, such that perturbations to the region are represented as the degree of sequence modification relative to WT within 20 bp windows. Flowering time quantifications are represented by box plots (with outliers as black points). Homozygous wild type (wt) and homozygous mutant (mutant) plants were phenotyped from segregating populations. Predicted TFBSs within the 52 bp region are denoted by black arrows. gRNAs are denoted by red arrows. Sample number (n) is shown to the left, and mean and standard deviation (sd) are shown to the right. Two-tailed, two-sample t-tests were performed between wt and mutant plants from the same segregating population.

E. Representative plant of $SpSP5G^{enhancer\text{-}1}$, grown under long days.

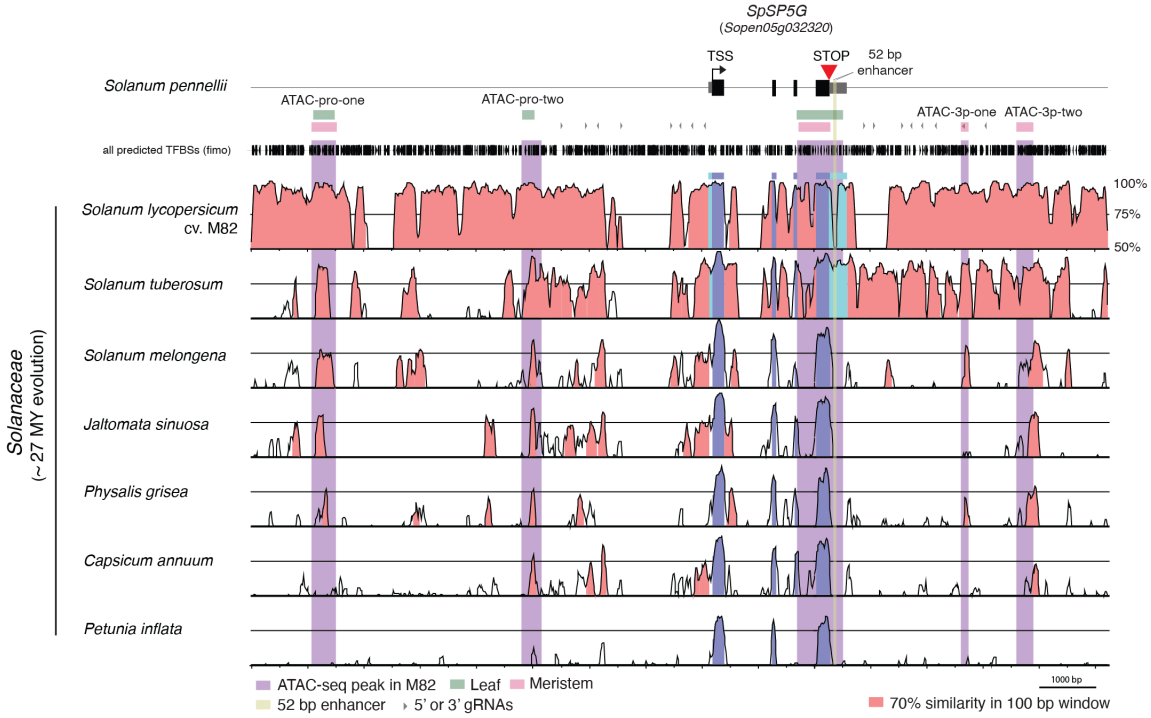
Figure 3-1.

A.

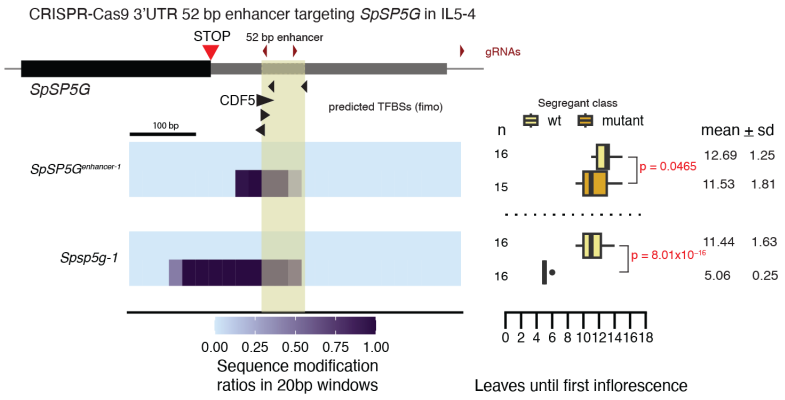
B.



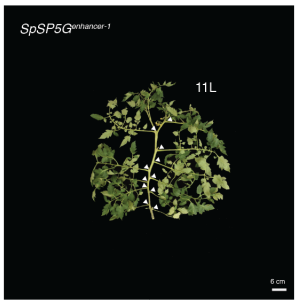
C.



D.



E.



3.3.2 Mutations upstream of *SpSP5G* generate variation in flowering time

It is not immediately clear which *cis*-regulatory regions are responsible for divergence in *SP5G* expression from comparing M82 and *Solanum pennellii SP5G* sequences. There are many regions of variation, and although deletion of most of the predicted 52 bp enhancer element did cause a slightly earlier flowering time under long days, it did not cause a M82-like flowering time. Even within regions of conservation, there are many SNPs in predicted TFBSs that could lead to altered TF binding and gene expression. Thus, we decided to take an unbiased approach to identify CREs regulating *SpSP5G* expression, using CRISPR-Cas9 8-guide arrays to target non-coding DNA upstream and downstream. Previously, our lab used CRISPR-Cas9 multiplex mutagenesis to create allelic diversity upstream of tomato *CLV3*, which translated to quantitative variation in fruit size (Rodríguez-Leal et al. 2017). We applied a similar approach to create allelic diversity 5' and 3' of *SpSP5G*.

We targeted 2.6 kb of 5' non-coding sequence proximal to the 5'UTR of *SpSP5G* with an 8-guide CRISPR-Cas9 array (**Fig. 3-2A**). In addition to a large insertion (~700 bp) relative to M82, this region has multiple areas of conservation with other *Solanaceae* family members (**Fig. 3-1C**). Multiple predicted TFBSs can be found throughout the region, including motifs associated with flowering regulation. CRISPR-Cas9 mutagenesis generated six unique alleles with various mutations upstream of *SpSP5G,* including one coding sequence allele (*Spsp5g-2*) which deletes the first 866 bp of the gene (and matches the phenotype of our other coding sequence alleles). All alleles were phenotyped for flowering time from segregating populations, under long days. Two alleles, $SpSP5G^{pro\text{-}1}$ and $SpSP5G^{pro\text{-}2}$, had a slightly earlier flowering time than WT (two or one leaves, respectively). Three alleles ($SpSP5G^{pro\text{-}3}$, $SpSP5G^{pro\text{-}4}$, $SpSP5G^{pro\text{-}5}$) had a moderately earlier flowering time than WT, flowering ~three leaves earlier under long days (**Fig. 3-2A, B**). Thus, we were able to create variation in flowering time by generating allelic diversity in the upstream region

alone. These alleles suggest that the presence of multiple CREs within the 5' region contribute to the regulation of *SP5G,* since larger deletions were correlated with earlier flowering time.

Next, we targeted 2.5 kb of the region proximally downstream of the 3'UTR of *SpSP5G* with an 8-guide CRISPR-Cas9 array (**Fig. 3-2C**). This region is largely conserved in M82 (except for one region adjacent to the 3'UTR) and includes one conserved ATAC-seq peak (**Fig. 3-1C**). CRISPR-Cas9 mutagenesis only generated two alleles, neither of which significantly impacted flowering time under long days, in segregating populations. It is possible that the 3' region deleted in $SpSP5G^{3p\text{-}2}$ does not contain CREs, or it may contain CREs that interact redundantly with CREs outside of the region. Generation of additional 3' alleles in the future would further clarify the role of the 3' region to *SP5G* regulation, especially the region of divergence adjacent to the 3'UTR.

Notably, none of these alleles mimicked either the domestication phenotype or the null phenotype, indicating the potential for the existence of many CREs tuning flowering time variation. Indeed, our previous studies of *cis*-regulatory regions support the existence of interactions among many CREs in gene regulation, and greater perturbations to regulatory regions often being necessary to elicit strong phenotypes (Wang et al. 2021).

Figure 3-2. Mutations upstream of *SpSP5G* generate variation in flowering time.

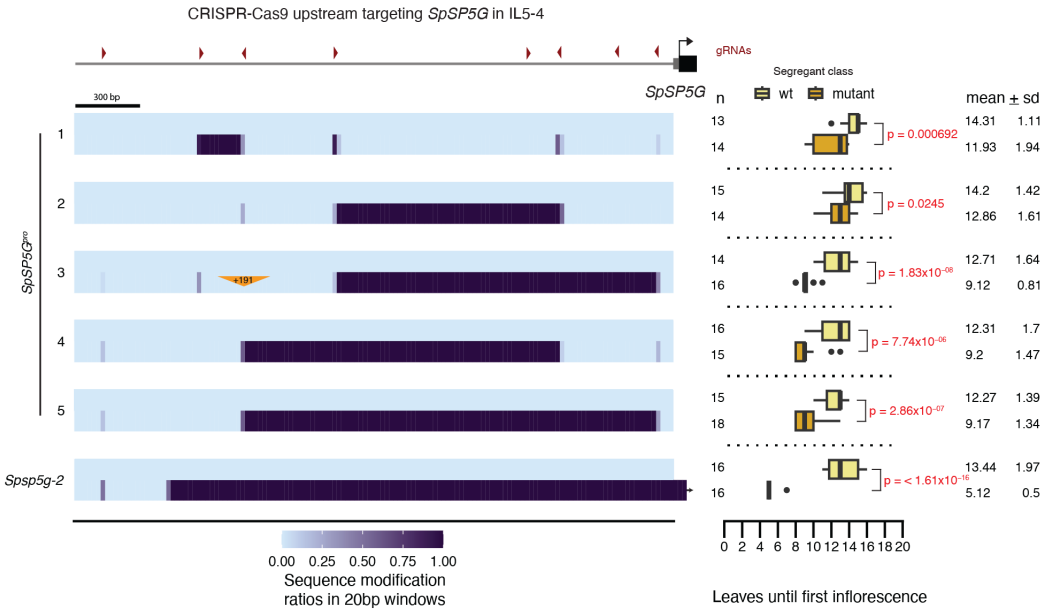
A. Encoded representation of the alleles generated by targeting the region upstream of *SpSP5G* with an 8-guide array spanning the 2.6 kb proximal to the 5'UTR (red arrows). Flowering time quantifications, from segregating populations, are represented by box plots (with outliers as black points). Two-tailed, two-sample t-tests were performed between wt and mutant plants from the same segregating population.

B. Representative *SpSP5G* 5' alleles grown under long days.

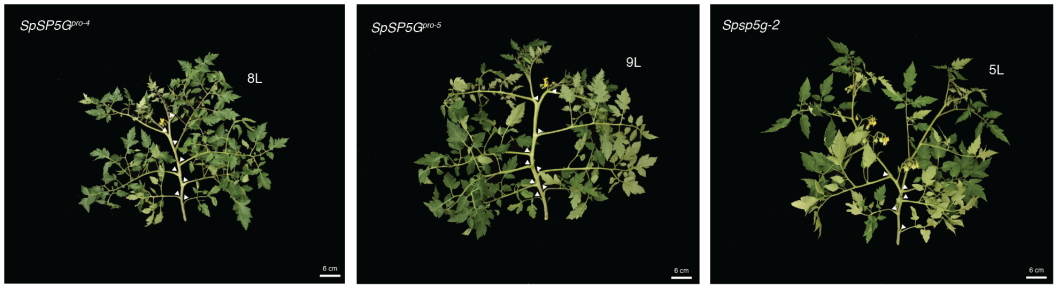
C. Encoded representation of the alleles generated by targeting the region downstream of *SpSP5G* with an 8-guide array spanning the 2.5 kb proximal to the 3'UTR (red arrows). Flowering time quantifications, from segregating populations, are represented by box plots (with outliers as black points). Two-tailed, two-sample t-tests were performed between wt and mutant plants from the same segregating population.

Figure 3-2.

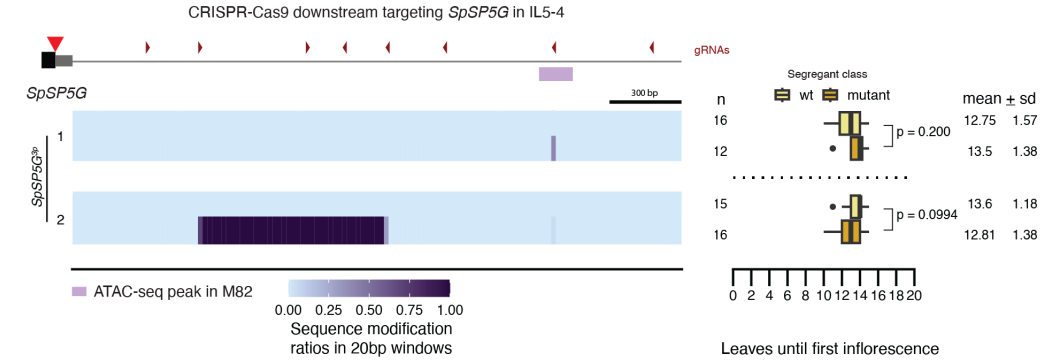
A.



B.



C.



3.3.3 Mutations in ATAC-seq peaks conserved between M82 and *S. pennellii* have various impacts on flowering time

Outside of these regions proximally upstream and downstream of *SpSP5G,* there were additional sequences we predicted might have regulatory function. We chose to specifically target four ATAC-seq peaks (identified in M82) with CRISPR-Cas9 guide arrays individually - two distally upstream and two distally downstream of *SpSP5G.* These regions were of interest because most of them were conserved in sequence in multiple *Solanaceae* species evaluated, suggesting they could be functionally relevant to *SP5G* expression (**Fig. 3-1C**). Additionally, we reasoned that although largely conserved between *S. pennellii* and M82, small indels or SNPs in these regions could also have led to flowering time divergence through disruption of a TFBS, for example.

Firstly, we separately targeted two ATAC-seq peaks downstream of *SpSP5G* with 4-guide arrays, and quantified flowering time from segregating populations (**Fig. 3-3A**). Four unique alleles were generated by targeting the more proximal ATAC-seq peak (called ATAC-3p-one). Collectively, three of these alleles ($SpSP5G^{ATAC\text{-}3p\text{-}one\text{-}1}$, $SpSP5G^{ATAC\text{-}3p\text{-}one\text{-}2}$, and $SpSP5G^{ATAC\text{-}3p\text{-}one\text{-}3}$) delete the majority of the ATAC-seq peak, although none significantly affect flowering time. One allele, $SpSP5G^{ATAC\text{-}3p\text{-}one\text{-}4}$, deletes a region 1289 bp long which encompasses ATAC-3p-one, as well as additional DNA surrounding it. This large deletion results in plants that flower significantly early under long days, after about six leaves (**Fig. 3-3B**). Interestingly, loss of function alleles (*Spsp5g*-1 and *Spsp5g*-2) only flower one leaf earlier than $SpSP5G^{ATAC\text{-}3p\text{-}one\text{-}4}$ plants (after ~five leaves compared to ~six leaves), suggesting that this *cis*-regulatory deletion may drastically reduce *SP5G* expression (or alter timing or location of expression) without completely abolishing it. Only one allele was generated from targeting the more distal ATAC-seq peak (called ATAC-3p-two). The deleted region of this allele, $SpSP5G^{ATAC\text{-}3p\text{-}two\text{-}1}$, overlaps with a small portion of the ATAC-seq peak, as well as a region widely conserved among the *Solanaceae SP5G* orthologs (**Fig. 3-3A, Fig. 3-1C**). This allele had a slightly earlier flowering time than WT (~one leaf earlier). Taken together, deletions within these

downstream ATAC-seq peaks reveal an important role of the 3' region in *SpSP5G* regulation, which contains multiple CREs.

Secondly, we separately targeted two ATAC-seq peaks upstream of *SpSP5G* with 4-guide arrays as well, and quantified flowering time within the first transgenic generation (T1) (**Fig. 3-3C**). Thus, flowering time for each allele was compared to an external wild type control (IL5-4). Thus far, we have obtained three unique alleles from targeting the more distal upstream ATAC-seq peak (ATAC-pro-one). All three alleles ($SpSP5G^{ATAC\text{-}pro\text{-}one\text{-}1}$, $SpSP5G^{ATAC\text{-}pro\text{-}one\text{-}2}$, and $SpSP5G^{ATAC\text{-}pro\text{-}one\text{-}3}$) flowered about one to two leaves earlier than the external WT control. However, in the future these alleles will need to be phenotyped in segregating populations to be more confident of their true impact on flowering time and *SP5G* regulation. Similarly, mutants generated from targeting the more proximal upstream ATAC-seq peak (ATAC-pro-two) will be isolated and phenotyped in the future.

From these experiments, smaller deletions in ATAC-seq peaks either had a weak impact or no impact on flowering time, whilst a large deletion downstream of *SpSP5G* had a very strong effect on flowering time. Whether these regions were altered during domestication to create phenotypic variation in flowering time is still unclear however, since particular alleles were both earlier and later flowering than M82.

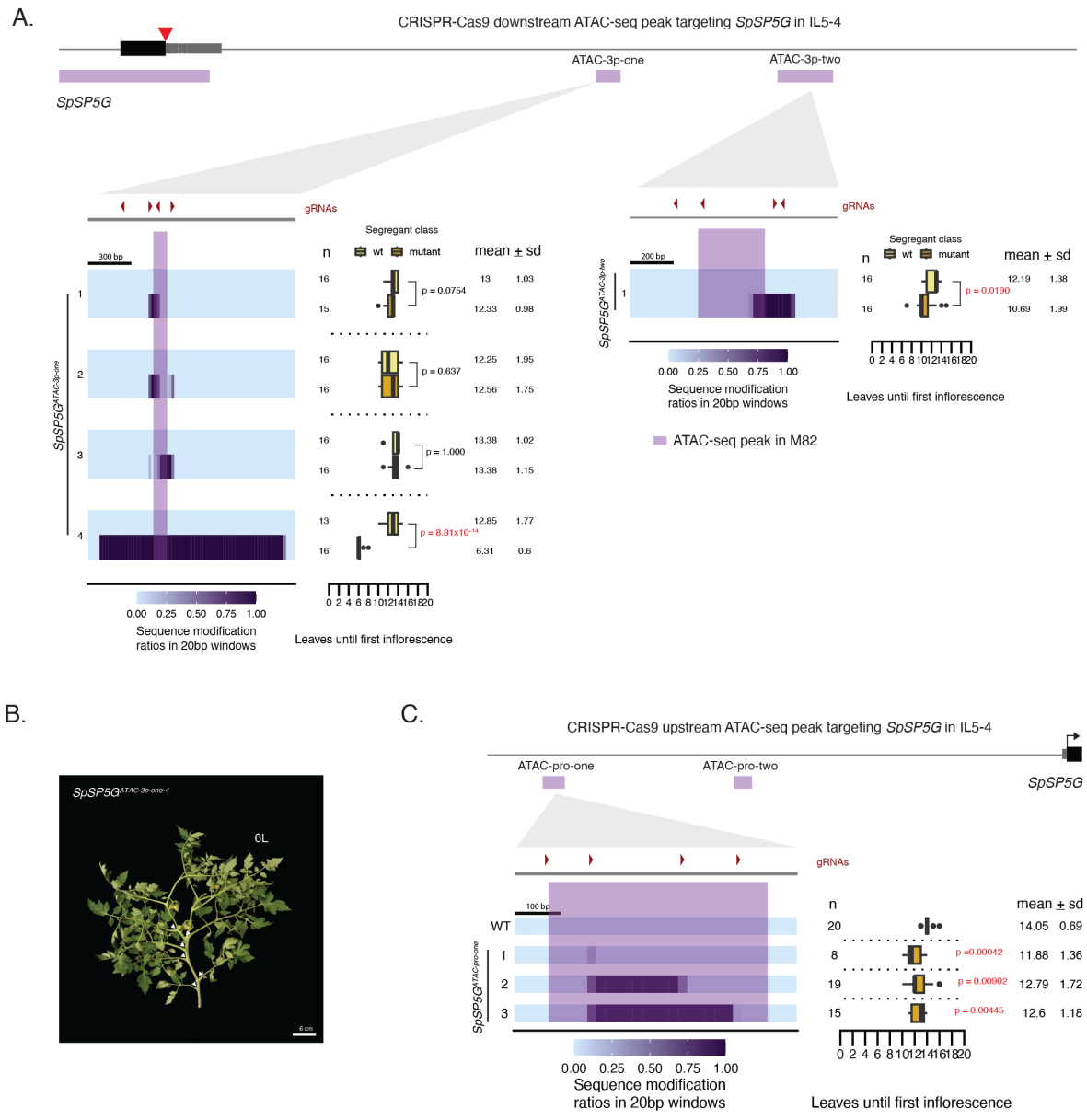
Figure 3-3. Mutations in ATAC-seq peaks conserved between M82 and *S. pennellii* have various impacts on flowering time.

A. Encoded representation of the alleles generated by targeting two different ATAC-seq peaks (identified in M82) downstream of *SpSP5G* with 4-guide constructs. Flowering time quantifications, from segregating populations, are represented by box plots (with outliers as black points). The ATAC-seq peaks from M82 are shown on the *S. pennellii* sequence, highlighted in purple. Two-tailed, two-sample t-tests were performed between wt and mutant plants from the same segregating population.

B. Representative plant of $SpSP5G^{ATAC\text{-}3p\text{-}one\text{-}4}$, grown under long days.

C. Encoded representation of the alleles generated by targeting one ATAC-seq peak (identified in M82) upstream of *SpSP5G* with a 4-guide construct. Flowering time quantifications are represented by box plots (with outliers as black points). The ATAC-seq peaks from M82 are shown on the *S. pennellii* sequence, highlighted in purple. A two-sided Dunnett's compare with control test was performed to compare all alleles to WT.

Figure 3-3.


A.
CRISPR-Cas9 downstream ATAC-seq peak targeting SpSP5G in IL5-4
ATAC-3p-one
ATAC-3p-two
SpSP5G
gRNAs
Segregant class
n
wt
mutant
mean ± sd
300 bp
200 bp
p = 0.0754
p = 0.637
p = 1.000
p = 8.81x10-14
p = 0.0190
0.00 0.25 0.50 0.75 1.00
Sequence modification ratios in 20bp windows
0 2 4 6 8 10 12 14 16 18 20
Leaves until first inflorescence
ATAC-seq peak in M82
B.
SpSP5GATAC-3p-one-4
6L
6 cm
C.
CRISPR-Cas9 upstream ATAC-seq peak targeting SpSP5G in IL5-4
ATAC-pro-one
ATAC-pro-two
SpSP5G
100 bp
WT
p =0.00042
p = 0.00902
p = 0.00445
Sequence modification ratios in 20bp windows
Leaves until first inflorescence

3.4 Discussion

This study provides evidence for the regulation of *SpSP5G* by multiple CREs. CRISPR-Cas9 mutagenesis was used to functionally dissect the *cis*-regulatory regions upstream and downstream of *SP5G* in an introgression line of the daylength-sensitive wild species *Solanum pennellii.* Deletion of 43/52 bp of a previously characterized enhancer element in the 3'UTR of *SpSP5G* was not able to recapitulate the daylength-insensitive flowering time phenotype of the domesticated tomato M82. With many potential regions of divergence and conservation shared between *S. pennellii* and M82 *SP5G,* we decided to take an unbiased approach to CRE discovery by mutating the regions upstream and downstream of *SpSP5G* with multiplex CRISPR gRNA arrays. We found that all five alleles with mutations upstream of *SpSP5G* had weak or moderate effects on flowering time. Targeting regions downstream of *SpSP5G,* including two ATAC-seq peaks conserved with M82, generated alleles which had no effect, a weak effect, or a strong effect on earlier flowering. Interestingly, no individual allele recapitulated the flowering time of M82. Although we did not definitively identify the causative *cis*-regulatory mutation(s) responsible for earlier flowering time among domesticated tomato cultivars, we did identify several potential regions of importance to the regulation of *SP5G*. Furthermore, this is an ongoing study, and several experiments are in progress to provide further insights into the principles of gene regulation and CREs.

Firstly, further genome editing methods will be applied to generate a full deletion of the proposed 52 bp enhancer sequence in the 3'UTR of *SpSP5G.* Although we deleted the only predicted CDF5 motif, there is another motif within the 9 bp not deleted (although it is not obviously tied to flowering or light perception), and still other CREs could exist that work with CDF5 to regulate *SP5G.* We were initially limited in our guide design by the PAMs available within the region, as well as repetitive regions. To overcome this, a second targeting of the 52 bp enhancer is underway utilizing a Cas9 variant that recognizes the PAM -NG, expanding our guide choices to increase the likelihood of deleting the full region. This experiment will allow us to more definitively rule out its function as

the sole region responsible for flowering time divergence during domestication. However, despite this incomplete deletion experiment, our preliminary dissection of the 52 bp enhancer does still provide substantial evidence that other CREs may have been manipulated, alone or in combination with the 52 bp enhancer, to generate daylength neutrality in M82.

New phenotypic variation for a trait can emerge through both singular and multi-step mutations within CREs. For example, large deletions can occur in one singular incidence, completely eliminating enhancer activity and instantly creating phenotypic divergence. This was the case for multiple populations of freshwater sticklebacks, which all gained deletions in a small enhancer of *Pitx1* that specifically drives expression in the pelvis, resulting in loss of the pelvic spine in these populations (Chan et al. 2010). Furthermore, one SNP within an enhancer 10 kb upstream of *LG1* is associated with altered gene expression, and led to a compact panicle architecture in cultivated rice during domestication (Zhu et al. 2013). However, many examples in the literature suggest that functional divergence of CREs emerges from smaller, multi-step mutations, which interact additively or synergically to fully alter enhancer activity. Multiple examples from studies of phenotypic divergence in *Drosophila* species highlight this (Williams et al. 2008; Rebeiz et al. 2009; Frankel et al. 2011). For example, loss of trichomes in *Drosophila sechellia* compared to *Drosophila melanogaster* was caused by multiple mutations within the E6 enhancer of the *shavenbaby (svb)* gene (Preger-Ben Noon et al. 2016). A large cluster of TFBSs for the activator Arrowhead act redundantly to make the E6 enhancer robustly activate *svb* expression in *Drosophila melanogaster.* Loss of four Arrowhead TFBSs, combined with the gain of a TFBS for the repressor Abrupt, was able to overcome the remaining Arrowhead TFBSs to eliminate enhancer function. However, neither mutation on its own is enough to overcome the robustness of the enhancer in driving *svb* expression. Therefore, whilst some traits during evolution were derived from individual mutations, there are also many examples of multi-step mutations that are necessary to drive phenotypic diversity. In the case of flowering time divergence during tomato domestication, many plausible scenarios can be imagined, from simplistic individual SNPs or indels that eliminated the majority of *SP5G* expression, to more

complex, multi-step mutations that were required for full expression loss, which interact additively or synergistically. There is also the possibility that enhancer loss was combined with silencer gain in M82.

Taken together, our functional genetic dissection of the *cis*-regulatory regions of *SpSP5G* suggests that there are many CREs upstream and downstream that contribute to its regulation. This suggests the hypothesis that perhaps many CREs are working together additively, synergistically, or redundantly to regulate *SP5G* expression, and thus multiple CREs must be mutated at once to severely impact *SP5G* and flowering time. Previous findings from our lab, as well as the previous thesis chapter, support this idea of gene expression being controlled by interactions among many CREs, not just a few (Wang et al. 2021). Several findings from our dissection of *SpSP5G* also support this reasoning. Firstly, many predicted binding sites for TFs involved in flowering and light perception can be found spread throughout the entire non-coding sequence surrounding *SP5G,* including CDF5, FLC, AP3, RVE8, bZIP16, PIF5, COG1, SPL8, SOC1, and SVP. Of course, just because these motifs exist does not guarantee that they are functional. However, we also found a general trend, in which larger mutations were associated with earlier flowering than smaller mutations. For example, smaller upstream mutations had a weak effect on flowering time, while larger mutations generally had a moderate effect (**Fig. 3-2A**). The only *cis*-mutation with a severe flowering time phenotype was caused by a large deletion of 1289 bp downstream, again supporting the idea of interaction among many CREs in the regulation of *SP5G.* Small deletions in the 3'UTR and ATAC-seq peaks either had no effect or a very weak effect on flowering time. Furthermore, no single *cis*-regulatory allele that we generated recapitulated the null phenotype (~five leaves). Thus we can hypothesize that reduced *SP5G* expression during domestication was possibly the result of multiple CRE losses/gains. In the future, we will evaluate this hypothesis by creating mutations in multiple *cis*-regulatory regions of *SP5G* separately and in combination. Combinations of 52 bp enhancer and 5' mutations are of particular interest, since neither on their own was as early flowering as the domestication allele under long days. However, we hypothesize that the additive effect of their

loss may replicate flowering time of the domesticated allele. To test this, we could transform the *SP5G* 5' guide construct into the background of the 3'UTR enhancer deletion allele. It would also be interesting to genetically dissect interactions among CREs within the upstream region alone, as well as within the 1289 bp mutated region downstream, to determine the nature of potential higher order interactions that may be occurring among multiple CREs dispersed throughout these regions. We previously demonstrated that additive, synergistic, and redundant genetic relationships between CREs in the 5' were all important to the regulation of a stem cell regulator in tomato (Wang et al. 2021). It would be illuminating to expand upon this work with another gene, especially incorporating more non-5' CREs, to discover if any general principles govern CRE interactions using a functional, *in vivo* genetics approach.

In addition to further exploration of genetic interactions in the regulation of *SpSP5G,* we would also like to explore physical mechanisms of CRE interactions in *SP5G* regulation. The 52 bp enhancer element was proposed to increase *SP5G* expression by creating a physical loop with the TSS (Zhang et al. 2018). Thus, we could evaluate how the strength of this gene loop has changed (or not) in our enhancer deletion allele compared to WT, using 3C. It would also be interesting to explore whether physical interactions are necessary for *SP5G* to utilize downstream CREs (such as those discovered within ATAC-3p-one and ATAC-3p-two), or distal upstream CREs. We are only beginning to understand how 3D chromatin interactions impact gene regulation in plants. Genome-wide assays of chromatin structure are often low resolution, and thus a high resolution exploration of looping interactions at the *SP5G* locus, using 3C or 4C for example, would provide valuable insights into short and long range looping interactions in plant gene regulation.

Lastly, a number of experiments could help us better understand the molecular consequences of our newly generated alleles. We will perform RNA-seq at different time points for all 18 alleles generated in this study, to determine the impact of each mutation on the diurnal expression of *SP5G.* Since *SP5G* expression is known to be highest 4 hours after dawn in wild type, mutations that affect either the transcript abundance or timing of gene expression could be responsible for the flowering

time phenotypes we observed. Since variance in *SP5G* expression level, rather than timing or location of expression, is altered between wild and domesticated species, mutations that impact anything other than expression level were likely not introduced to create flowering time divergence during evolution. Since another general role of 3'UTRs is to regulate mRNA stability, it will also be important to conduct RNA stability assays in the future, to ensure that the 43 bp deletion in the 3'UTR is actually disrupting a CRE, rather than destabilizing the transcript. These experiments will provide a better understanding of the relationship between gene expression and phenotypic effect, which may or may not always be correlated in expected ways.

By studying *SpSP5G* in an isogenic background, we have begun to dissect the *cis*-regulatory elements controlling its regulation using CRISPR-Cas9 genome editing. We have discovered several *cis*-regulatory regions of importance to *SpSP5G* regulation by quantifying its dosage-dependent trait, flowering time. By understanding more about the regulation of this gene, we have explored several hypotheses about how phenotypic divergence can mechanistically emerge during the course of evolution, using an *in vivo* approach rarely used. Through this study as well as previous ones, we are constantly finding that regulatory regions are robust to perturbation, often requiring multiple fortuitous mutation events to create drastic phenotypes. Thus, developing phenotypic variation in many cases is expected to be a slow process, not instantaneous. Multi-step mutations during the course of evolution may often be required to elicit substantial phenotypic divergence for selection to act upon. In the future, we hope to gain further insights into the multiple CREs regulating this gene, their genetic and physical interactions, and molecular consequences in the control of flowering time variation.

3.5 Methods

3.5.1 Plant material, growth conditions and phenotyping

Seeds of the introgression line IL5-4 were obtained from the Charles M. Rick Tomato Genetics Resource Center (TGRC) at the University of California, Davis. IL5-4 was used as the

background for WT and CRISPR-Cas9 tomato mutagenesis experiments. During initial allele isolation, tomato plants were sown and grown in 96-well flats for ~four weeks before being transplanted to pots, and grown in greenhouse conditions. The greenhouse operates under long days (16h light, 8h dark) with natural and artificial light (from high pressure sodium bulbs ~250 umol/m^2), at a temperature between 26-28°C (day) and 18-20°C (night), with relative humidity 40-60%. For phenotyping, tomato plants were sown and grown in 96-well flats before being transplanted to pots in the greenhouse, and grown under long days. For each F2 population, between 13-16 WT and 13-16 homozygous mutant plants were phenotyped for flowering time, defined as the number of leaves before emergence of the first inflorescence. Twenty plants of IL5-4 and M82 were also phenotyped for flowering time under long days.

3.5.2 CRISPR-Cas9 mutagenesis, plant transformation, and selection of mutant alleles

Generation of transgenic tomato with CRISPR-Cas9 mutagenesis was performed as previously described (Brooks et al. 2014). Briefly, gRNAs were designed with Geneious Prime (https://www.geneious.com). The Golden Gate assembly method was used to clone gRNAs into a binary vector with Cas9 and kanamycin selection (Werner et al. 2012; Rodríguez-Leal et al. 2017). Binary vectors were introduced into tomato plants through *Agrobacterium tumefaciens* mediated transformation in tissue culture (Van Eck et al. 2019b). Transgenic plants were screened for mutations using PCR primers surrounding the gRNA target sites. PCR products were screened for obvious shifts in size by gel electrophoresis, and mutations were characterized by Sanger sequencing. First or second generation transgenics (T0 or T1) were backcrossed to WT to eliminate the Cas9 transgene and purge the genome of potential off-target mutations. F2 populations from these crosses were used for phenotypic analysis. All gRNA and primer sequences are listed in Supplementary Table 3 and 4.

3.5.3 *Cis*-regulatory sequence conservation analysis, identification of ATAC-seq peaks, and TFBS prediction

Within-family conservation analysis was performed to predict conserved non-coding sequences within the 5' and 3' of *SpSP5G* in tomato that were shared among several *Solanaceae* species. The closest *SP5G* ortholog from each species was determined based on the ortholog with the greatest similarity to *SlSP5G* within the 5' and 3' regions. 40 kb of sequence upstream and downstream of each *SP5G* ortholog was extracted, and aligned to *Solanum pennellii SP5G* using mVISTA Shuffle-LAGAN (http://genome.lbl.gov/vista/mvista/submit.shtml) (Frazer et al. 2004). Conservation was calculated in 100 bp windows, with a 70% similarity threshold. ATAC-seq peaks from M82 meristem and leaf tissue were obtained from assays previously published (Hendelman et al. 2021). The sequence of the ATAC-seq peak from M82 was aligned to the *SpSP5G* locus using Geneious Prime to find the orthologous region in *S. pennellii* (https://www.geneious.com). TFBSs were predicted by scanning the *SpSP5G* 5' and 3' regions for motifs using FIMO in the MEME suite (http://meme-suite.org/doc/fimo.html) (Grant et al. 2011). Position frequency matrices for known plant transcription factors were obtained from the JASPAR CORE PFMs of plants collection 2022 (Castro-Mondragon et al. 2022). A p-value cutoff of 0.00001 was used to predict TFBSs.

3.5.4 Statistical methods

A two-tailed, two-sample t-test was used to compare flowering time between WT and mutant alleles in each segregating population. Pairwise comparisons between ATAC-pro-one alleles and WT were performed using Dunnett's compare with control test. P-value cutoff of <0.05 was used.

Chapter 4: Conclusions and perspectives

4.1 Main conclusions and significance

Throughout this thesis work, we have studied *cis*-regulatory elements within an evolutionary framework, using a functional genetics approach. We have also considered some key features of CREs, including their genetic and physical interactions, as well as the role that conservation of non-coding sequences play in the control of gene regulation. Firstly, we have explored a long-held question about how select genes are able to maintain conserved function and expression throughout evolution despite extreme *cis*-regulatory sequence divergence. Secondly, we have explored potential mechanisms of *cis*-regulatory variance involved in the phenotypic divergence of species throughout evolution. These questions have been studied before, mostly in animal model systems, and with molecular assays that only provide a proxy for phenotypic effect. While there are many features predictive of CREs, such as chromatin accessibility, methylation, histone modifications, and reporter assays, we are still unsure how reliable these techniques are at discovering bona fide CREs that impact gene regulation in a meaningful way. Reporter assays commonly used to verify the activity of CREs place them in a non-native context. We have explored the functional relevance of CREs in these two contexts, using CRISPR-Cas9 *in vivo* mutations of *cis*-regulatory regions in two model plants.

In chapter 2, we explored the evolution of CREs and their organization in the expression of a highly conserved gene regulating meristem size, *CLV3*, in *Arabidopsis* and tomato. Despite diverging approximately 125 million years ago, the functional *CLV3* peptide is still highly conserved, as is the location and timing of expression in the meristem (Somssich et al. 2016). However, the DNA upstream and downstream of *CLV3* in *Arabidopsis* and tomato is highly diverged, making identification of CNSs between the two species difficult. Therefore, we sought to discover how CREs regulating *CLV3* evolved using CRISPR-Cas9 mutagenesis of 5' and 3' *cis*-regulatory regions. Previously, mutagenesis of the region ~2 kb upstream of tomato *CLV3* was sufficient to produce the

full spectrum of quantitative variation for locule number, including a null-like phenotype (Rodríguez-Leal et al. 2017). We found that mutations downstream of tomato *CLV3* had subtle phenotypic effects on locule number, and additive and mildly synergistic interactions with upstream mutations. In contrast, mutagenesis of either the upstream or downstream region of *Arabidopsis CLV3* had a weak effect on locule number. Combined mutations in the upstream and downstream regions had a synergistic effect on locule number, spanning a range of phenotypes including null-like. While CNSs could not be detected between the species, they were discovered within each family. We found that within-family CNSs were partially predictive of a mutation having a phenotypic effect, with CNSs 5' and 3' of *Arabidopsis* and tomato *CLV3* having an effect on locule number when mutated. Additionally, a 27 bp element within a *Solanaceae* CNS upstream of tomato *CLV3* was also found in a *Brassicaceae* CNS downstream of *Arabidopsis CLV3.* This 27 bp contained an intact WUS binding site, which was previously demonstrated to bind WUS in *Arabidopsis* using ChIP-seq (Perales et al. 2016). Therefore, our results suggest that particular TFBSs may be conserved, while their organization is more malleable to change. Since TFBSs are often only 5-11 bp long, this offers a potential explanation for the difficulty in detecting these conserved sites over very long evolutionary distances.

In chapter 3, we explored the potential role of CREs in phenotypic divergence for flowering time between wild and domesticated tomato species. Variation for flowering time during long days was previously mapped to the anti-florigen *SP5G,* and difference in expression level of *SP5G* was proposed as an explanation for flowering time differences (Soyk et al. 2017). Furthermore, attempts to associate a particular mutation with daylength insensitive tomato species, in the coding sequence or surrounding *cis*-regulatory DNA, have been difficult, only identifying one consistent region of divergence in the 3'UTR (Zhang et al. 2018). When we deleted the majority of this region using CRISPR-Cas9 in the daylength sensitive introgression line IL5-4, it did not generate the daylength insensitive flowering time phenotype characteristic of domesticated species. Therefore, we hypothesized that other CRE mutations, on their own or in addition to loss of this 3'UTR element,

may have been responsible for the molecular and phenotypic shift in *SP5G* expression and flowering time during domestication. To test the possibility of this, we targeted the upstream and downstream *cis*-regulatory regions of *SP5G* in IL5-4 with multi-guide CRISPR-Cas9 constructs, generating several alleles with various deletions. These deletions led to both weak and moderate effects on flowering time, although no allele matched the early flowering time of the domesticated tomato M82. This suggested that the interaction of multiple CREs may control *SP5G* expression, and thus multiple CREs may have been altered during domestication to create daylength neutrality. In addition to what we could learn from targeting regions of divergence, we were also interested in what we could learn about the regulation of *SP5G* from CNSs and open chromatin. We targeted several conserved ATAC-seq peaks with CRISPR-Cas9, two upstream and two downstream of *SP5G.* Deletions within these regions had no effect, weak effects, or strong effects on flowering time, although again no phenotype matched M82. Taken together, our results suggest that mutations within the *cis*-regulatory elements of *SP5G* are likely sufficient to generate daylength insensitive tomatoes, however we will need to further explore interactions among several CREs to confirm this suspicion, as well as the molecular consequences of particular alleles.

From our functional dissections of the *cis*-regulation of these two genes, we have gained some insights into the relationship between regulatory regions and evolutionary processes. Mutations that occur within the coding sequence of genes often have immediate, extreme consequences for the expression of a particular trait. Due to their often severe outcome for protein function, these kinds of mutations are immediately available to be acted upon by mechanisms of natural or artificial selection. In contrast, our studies of *CLV3* and *SP5G* suggest that regulatory regions are extremely robust to perturbations, in the context of both conserved and diverged traits. This seems to stem from the extensive complexity that characterizes these regions – namely, multiple CREs and their higher order interactions in the control of gene regulation. Extreme sequence changes and rearrangements of CREs are tolerated by highly conserved genes (such as *CLV3*), and there seems to be a requirement for multiple fortuitous mutations in many CREs to enable extreme phenotypic shifts (such as in *SP5G*).

This seems to be in line with many other studies of *cis*-regulatory/phenotypic divergence (Williams et al. 2008; Frankel et al. 2011; Wittkopp and Kalay 2012). Thus, *cis*-regulatory evolutionary mechanisms of divergence might be expected to advance more gradually, as regulatory regions slowly accumulate the mutations necessary to enable a selectable level of phenotypic change. However, in the future more functional studies are needed to determine if this is a universal, generalizable trend among eukaryotic genes and traits.

In addition to providing a much needed functional perspective to the study of CRE evolution, we have also helped to inform core principles that can aide in genetic engineering of *cis*-regulatory elements for crop improvement, as well as synthetic promoter design. CRISPR has provided breeders with a tool to more rapidly and precisely engineer plants with beneficial traits for agriculture. Targeting the *cis*-regulatory elements of particular genes is often a more successful method of trait engineering, since pleiotropic effects of coding sequence mutations can be avoided. A better understanding of CREs, their conservation, and their interactions will enable breeders to better predict the phenotypic effect of their directed mutations on the trait of interest, and gene editing will provide both a faster and more efficient means of producing crops with beneficial traits.

4.2 Future directions

4.2.1 Conserved non-coding sequences in gene regulation

During this thesis project, we have explored the role of conserved non-coding sequences in gene regulation. Conserved non-coding elements are a frequent feature of animal genomes, however they have been more difficult to detect in plant genomes across large evolutionary distances, such as those separating different phyla, class, and order. Since functional TFBSs are often small, it can be hard to detect their conservation if the higher order organization of multiple TFBSs is not of vital importance to maintain proper gene regulation. We have found evidence for this in the *cis*-regulatory DNA surrounding *CLV3*, in which multiple WUS binding sites are likely conserved, however their detection is complicated by the fact that the core motif is 4 bp long and low complexity, a sequence

which frequently occurs by chance. The lab previously developed a new algorithm to detect CNSs in the *Solanaceae* and *Brassicaceae,* as well as potential CNSs between these families (Hendelman et al. 2021). In order to overcome the issue of extreme sequence divergence between different families, consensus CNSs from each family are extracted and used in the comparison, reducing the search space. Currently, this analysis is being expanded to include many new plant families, across the entire kingdom, and extract CNSs at six different levels of conservation: family, dicots, flowering plants, seed plants, land plants, and green algae. When finished, this database of plant CNSs will serve as a helpful starting point for identifying potential CREs of importance. In the future, a more in-depth functional dissection of CNSs across the plant kingdom will provide a more wholistic view of the relevance of these sequences to gene regulation and evolution. As CRISPR-Cas9 efficiency and methods of transgenesis improve, larger-scale mutagenesis experiments, with more genes from plant species spanning various evolutionary distances, will be important to carry out. For example, we could select 10-20 conserved genes, and explore the phenotypic relevance of CNSs at all different levels (from within family all the way to sequences conserved with green algae), using the numerous model plant species available to the field.

However, it should still be noted that alignment algorithms often filter out low complexity regions, which can harbor important TFBSs, so we cannot solely rely on bioinformatics approaches to detect CREs – sometimes functional experiments such as CRISPR-Cas9 mutagenesis are still the best option to dissect genetic complexity. Additionally, the CNSs that can be detected with this new approach across large evolutionary distances are still very short, and there is the possibility that a portion of these are due to chance. This was the case with *Arabidopsis* and tomato *CLV3* - no reliable CNSs were detected between the two families, likely due to WUS binding sites being low complexity. Therefore, functional mutagenesis of regulatory regions is a good approach to identify these types of CREs.

4.2.2 Genetic interactions between CREs

We have built on previous work by lab members to define genetic interactions among CREs *in vivo.* Previously, our lab explored genetic interactions among both conserved and non-conserved 5' regions within 2.1 kb upstream of tomato *CLV3* (Wang et al. 2021). While confined mutations in individual regions (150-300 bp long) had no effect or a weak effect on locule number, pairwise combinations of mutations in these regions had enhanced effects on locule number, including additive and synergistic relationships. Furthermore, large deletions spanning the majority of this 2.1 kb region have strong effects on locule number, one allele mimicking the null mutant. Taken together, this paper demonstrated clear evidence of higher order genetic interactions among multiple CREs within the *CLV3* 5', using *in vivo* genetic perturbations and phenotypic readouts to quantify the effects. We have expanded upon this work by considering CREs in other genomic contexts, namely the 3'UTR and the region proximally downstream. Alone, mutations within these regions can have weak effects on locule number. By including these regions in interaction tests, we were able to show that they also have complex interactions with CREs upstream, including additive (in tomato *CLV3* tests) and synergistic (in *Arabidopsis CLV3* tests). We have also considered these genetic interactions between *CLV3* CREs in an evolutionary context, showing how these relationships are malleable to change over evolutionary time. In the future, it would be interesting to include CREs from other genomic regions in these interaction tests, such as CREs in introns and distal enhancer elements. The first intron of many plant genes contains CREs, and it would be useful to functionally validate their contribution to gene regulation *in vivo,* individually and in tandem with other CREs upstream and downstream (Greene et al. 1994; Sieburth and Meyerowitz 1997; Qüesta et al. 2016).

Another major finding of this paper on intragenic epistasis in tomato found that large perturbations within the conserved region of the 5' of tomato *WUS (SlWUS)* did not affect locule number or meristem termination (Wang et al. 2021). There could be several explanations for this finding, but an intriguing hypothesis (based on our findings in *CLV3*) is that CREs downstream of *SlWUS* are able to compensate for the loss of CREs upstream. The *lc* QTL, responsible for a slight increase in locule number during domestication, is caused by SNPs in a MADS-box motif located

downstream of *SlWUS,* and supports the existence of at least one CRE downstream of this gene (Somssich et al. 2016). In the future, *cis*-regulatory regions upstream *and* downstream of this gene should be further dissected. For example, it would be interesting to make 5' mutations in the background of the *lc* CRISPR mutant (generated in Rodriguez-Leal et al., 2017). An enhancer upstream may cancel out the effect of the downstream silencer. Generating allelic diversity downstream of *SlWUS,* in the background of large 5' deletion alleles, may finally lead to decreased locule number, or complete meristem termination.

The majority of CREs discovered in these experiments were enhancer elements, presumably increasing gene expression, given that increased locule number is associated with decreased *SlCLV3* expression. This may indicate something about the nature of gene regulation – perhaps enhancers and activators are simply more common than silencers and repressors. Another possibility is that our model system, *CLV3* locule number, may not be ideal to phenotypically identify silencer elements. We did not observe any uni-locular fruits in tomato or *Arabidopsis,* suggesting locule number may be constrained by a lower limit of 2 locules. Hence, the system may not be as phenotypically sensitive to small expression increases in *CLV3* as it is to small decreases. One way to overcome this particular limitation would be to introduce our mutations in a different genetic background. For example, null alleles of *Arabidopsis CLV1,* one of the receptors for *CLV3,* have a moderate effect on locule number (~4 locules on average). We introduced some of our *AtCLV3* 5' alleles with no effect on locule number into a *Atclv1* null allele background, to see if we could reveal hidden enhancer or silencer effects in this sensitized background (unpublished). We did not observe any enhancement or reduction in locule number for these alleles. However, another mutation we generated, this time in a region 12 kb upstream, which was predicted to physically interact with the *AtCLV3* promoter in one Hi-C study on whole seedlings, did not impact locule number (Liu et al. 2016). However, when introduced into the *Atclv1* background, it slightly suppressed the *Atclv1* phenotype, suggesting that it possibly functions as a silencer of *AtCLV3* (not published). The same limitation may exist in our tomato *SP5G* experiments, if *SP5G* expression is already saturated in IL5-4. Further CRISPR-Cas9 mutagenesis

experiments in M82 would be a nice complement to our work, since it is possible that formation of a new silencer element in M82 during domestication was partially responsible for decreased *SP5G* expression relative to *Solanum pennellii.* Previous attempts to mutagenize the proximal 5' of *SP5G* in M82 produced multiple large deletion alleles, none of which had any impact on flowering time (unpublished). However, a more comprehensive dissection of regions distally upstream, as well as downstream, has still not been done.

Lastly, our analysis of genetic interactions among CREs *in vivo* would benefit from more precise methods of CRISPR mutagenesis in the future. Many of our mutations are quite crude, deleting large regions of DNA at once. Since TFBSs are small, it is difficult to pinpoint the precise CREs we are perturbing, which may actually be multiple at once. Advancements in CRISPR technology will make this more feasible in the future. For example, homology-directed repair (HDR) can be used with CRISPR to repair double strand breaks by using a donor template with the desired modification, flanked by DNA homologous to the ends of the cleaved DNA (Chen et al. 2022). In our study, the combinations of 5' and 3' mutations we generated were similar, but often not precisely identical to the corresponding individual mutations we used in interaction tests. HDR could overcome this issue, introducing the precise mutations desired individually and in combination, making perfect conditions for interaction testing *in vivo.* HDR is a highly precise method of *in vivo* genome modification, however it is currently extremely inefficient in plants, occurring at low frequency. A slightly more efficient approach than HDR is base editing. A Cas9 nickase is fused to a cytosine or adenine deaminase, and directed to a specific DNA sequence using a gRNA, where nucleotides within a small region are susceptible to C to T or A to G conversions, respectively (Gaudelli et al. 2017). Although not as precise as HDR, it is still a useful approach for editing TFBSs *in vivo*.

During my thesis work, I attempted to mutate the six WUS binding sites upstream and downstream of *AtCLV3* using two more precise technologies. The first, a CRISPR-Cas9-SpRY fused to an adenine base editor (ABE8e), which is PAM-less, and converts A residues to G within a ~8 bp window. For the second, I used a CRISPR-Cas9-NG, which expanded the PAM choices such that I

was able to choose guides that would induce double strand breaks directly within the WUS motifs. Unfortunately, neither of these techniques generated any substantial edits with high efficiency. A third option, which does not include double strand breaks or HDR, is prime editing. In prime editing, a Cas9 nickase fused to a reverse transcriptase is guided to a specific site by a prime editing guide RNA (pegRNA). The pegRNA both directs the Cas9 to the specific site, and contains the desired edit to introduce at the site using reverse transcription (Anzalone et al. 2019). This technique is less constrained by PAM availability, can allow for every possible type of base-to-base change, and can precisely specify which bases are altered (unlike base editors). Prime editing has been applied in plants with some success, but is still early in its development (Li et al. 2023). However, as all of these technologies continue to be studied, they will likely improve in efficiency and replace the CRISPR-Cas9 double strand break non-homologous end joining repair method of mutagenesis in the future.

4.2.3 Physical interactions between CREs

In addition to future studies of genetic interactions among CREs, the field of plant gene regulation would also benefit from more in depth studies of the role of physical interactions among CREs, both long and short range. From assays of plant genome conformation, it is clear that large-genome plants, such as tomato and maize, have numerous long-range looping interactions, a few of which have been associated with roles in gene regulation in maize (Dong et al. 2017). And even in *Arabidopsis,* which lacks higher order genome structure, smaller gene looping interactions have been identified (Liu et al. 2016). However, the functional relevance of these looping interactions has rarely been investigated *in vivo,* nor their genetic and physical interactions with other CREs. For example, what is the relationship between non-coding regions involved in either end of a looping interaction? Do looping interactions interact additively or synergistically with other CREs outside of the loop? Are paralogous genes regulated by CREs shared through looping interactions? During this thesis research, we have provided functional evidence of CREs downstream of *SlCLV3, AtCLV3,* and *SpSP5G*. It is interesting to consider how genes utilize these CREs to affect RNA polymerase recruitment to the

TSS – one hypothesis is that they physically loop with regions upstream of the gene. Few studies have done high resolution dissections of DNA looping at one specific locus in plants. *AtCLV3* and *SpSP5G* would be good genes to test this hypothesis, since there is already evidence of looping interactions as well as downstream CREs in both cases.

There are technical challenges that have made the discovery of looping interactions difficult. Firstly, long-range enhancer elements are often difficult to assign to a specific gene, since they can often skip over many closer genes to regulate a gene farther away. Therefore, some form of chromatin conformation capture experiment must be used to predict their existence. These experiments are technically challenging, and often low resolution, preventing the discovery of weaker and shorter-range enhancer interactions. Many long-range enhancer-promoter interactions may also be transient, occurring for only a brief moment, in a particular cell type. Newly emerging single-cell Hi-C techniques may help overcome this challenge (Nagano et al. 2015). These experiments are also improving in resolution, for example with the introduction of more frequently cutting enzymes. Furthermore, locus specific 4C assays could be used in place of genome-wide assays to increase resolution at a specific gene locus (Han et al. 2018a).

Since our lab is interested in many genes expressed in the shoot apical meristem, another difficulty in loop discovery is obtaining enough meristem tissue to do some chromatin conformation capture assays, which require at least a few grams. In order to overcome this obstacle, one option is using an *ANANTHA (AN)* mutant, which overproliferates meristematic tissue. We performed Hi-C with tissue from these plants, using the Arima HiC+ kit (not published). While the libraries and sequencing data passed quality control checks, the samples were not sufficiently complex enough to generate Hi-C contact maps at greater than 5 kb resolution. This precluded the ability to detect smaller loops, such as those occurring across individual genes. We did however identify 9573 loops at 5 kb resolution, several involving meristem genes, including *SFT, TERMINATING FLOWER (TMF),* and looping between several paralogs of *TOMATO MADS-BOX GENE 3 (TM3).* Further sequencing to increase resolution, and a non-meristem tissue control, would improve our ability to predict and

locate potential physical looping interactions involved in regulating meristem genes. Once we are better able to detect these physical looping interactions among non-coding regions, we can target them with CRISPR-Cas9 mutagenesis in order to validate their contribution to gene regulation, and explore their interactions with other CREs.

4.2.4 Molecular consequences of CRE mutagenesis *in vivo*

Finally, our studies of CREs in an evolutionary context would benefit from further exploration of the molecular consequences of *in vivo* CRE mutagenesis. This would be a nice complement to the functional relevance derived from phenotyping experiments. For example, in the future we could explore the relationship between gene expression and phenotypic strength, using our series of *CLV3* and *SP5G cis*-regulatory alleles. Perhaps they do not always associate in expected ways, due to feedback mechanisms, or altered gene expression in space. Quantifying these molecular effects is challenging for meristem-expressed genes such as *CLV3.* Shifts in expression may be too subtle to pick up by crude methods such as qPCR or RNA-seq, especially considering that it is technically challenging to isolate the dozens of small meristems (at the exact same developmental stage) needed to produce enough cDNA. A more sensitive technique would need to be applied, such as droplet digital PCR, which sensitively detects absolute quantifications of lowly abundant transcripts (Taylor et al. 2017). In-situs could also be performed on meristems to identify *cis*-regulatory alleles that alter the location of expression. Expression analysis of *SP5G* should be simpler, with a qPCR experiment analyzing diurnal *SP5G* expression in the cotyledons every 4 hours over the course of a day. Collectively, these experiments would help provide a better understanding of the relationship between CREs, gene dosage (in space and time) and phenotypic output.

Another key factor to a better understanding of the consequences of our *cis*-regulatory mutations is understanding how they impact TFBSs, and TF binding dynamics. Identifying the particular TFBSs perturbed by our mutations will allow us to make more specific, directed mutations in the future, for example using base or prime editing. Currently, predicting functional TFBSs is not

trivial, and often requires some form of molecular assay. While general binding motifs are known for many TFs, not all TFs have been evaluated in every species, and every cell type. Additionally, TF motifs occur frequently just by chance, and their presence alone does not guarantee their function. Therefore, TF footprinting assays within regions of open chromatin, as well as ChIP-seq, are still useful predictors of which TF motifs are bona fide binding sites. Additionally, as algorithms for the detection of CNSs become more precise, they can help narrow down likely TFBSs that have a conserved function. For example, ChIP-seq on tomato meristem tissue would help us identify functional WUS binding sites in the regulation of *SlCLV3,* complementing existing knowledge of these sites in *Arabidopsis,* and providing a clearer picture of how these particular TFBSs were shuffled during evolution. Once we have a better prediction of these TFBSs, we can more precisely edit them using base editing, and evaluate interactions among individual TFBSs *in vivo.* For example, we could precisely genetically dissect interactions such as cooperativity between adjacent TFBSs.

Upon discovering that mutations within the 3'UTR of *SlCLV3* had a weak effect on locule number, I set out to identify likely TFBSs using dual luciferase assays. There were several MADS-box TF motifs within the mutated region associated with locule number change, and I tested the ability of several to bind to the ~90 bp 3' sequence in tobacco, including TAG1 (not published). TAG1 was a particularly interesting candidate, since its *Arabidopsis* homolog AGAMOUS is known to bind downstream of *AtWUS* (Liu et al. 2011). From these assays, relative luciferase activity was not different between wild type and mutant versions of the 3' regulatory region for any of the TFs tested. Although these assays did not provide evidence for TF binding to these 3' motifs, it is possible that this assay was not sensitive enough to detect binding if it is very weak. Additionally, this assay was performed in a heterologous system, in which particular binding partners for these MADS-box TFs may not be present. In the future, ChIP-seq or DAP-seq with these TFs, combined with precision TFBS editing, would likely be a more reliable approach to investigating these questions.

Once a TF candidate is established, there are also genetic experiments that can be performed to explore the interaction of that TF with our *cis*-regulatory alleles. Previous studies in the lab have

explored the intergenic interactions among *cis*-regulatory alleles and other genes, specifically compensators. Rodriguez-Leal et al. discovered an active mechanism of compensation between tomato *CLV3* and its paralog *CLE9* (i.e. *CLE9* is expressed in *clv3* null alleles, but not in WT) (Rodriguez-Leal et al. 2019). A follow-up study explored the intergenic relationship between a series of *CLV3* 5' alleles with locule number variation, and a null allele of *CLE9* (not published). Similarly, we could explore the genetic relationship between our *cis*-regulatory alleles and their TF null mutants. *CLV3* is not suitable for this, since null *WUS* mutations have meristem termination. However, several promising motifs for TFs involved in sensing daylength and flowering are present throughout the *SpSP5G* regulatory sequence, such as CDF5. While deletion of CDF5 and its homologs may cause earlier flowering in wild type plants, we can hypothesize that flowering time phenotypes will not be enhanced by this mutation in *cis*-regulatory alleles engineered to abolish all of its binding sites in *SP5G.* These kinds of experiments are already underway in the lab, using other model genes, and promise to provide a new perspective on the interaction between TFs and their binding sites that goes beyond studies of direct physical interaction.

4.3 Final thoughts

During the course of my thesis work, I have strived to provide a functional perspective of CREs, their organization, and their interactions in the control of expression conservation and divergence during evolution. While this work was done using two specific developmental genes, *CLV3* and *SP5G,* the principles gained likely apply to numerous genes, in both plants and animals, although further experiments of this nature will be needed to validate that in the future. With the rapid development of genome editing technologies, it is a truly exciting time in history to be a plant geneticist. I am confident that we will continue to build on our knowledge of gene regulation, thus unlocking the mysteries of development and evolution that define what it means to be alive on this earth.

References

Ahituv N, Zhu Y, Visel A, Holt A, Afzal V, Pennacchio LA, Rubin EM. 2007. Deletion of Ultraconserved Elements Yields Viable Mice. *PLOS Biology* **5**: e234.

Albert FW, Kruglyak L. 2015. The role of regulatory variation in complex traits and disease. *Nat Rev Genet* **16**: 197–212.

Anzalone AV, Randolph PB, Davis JR, Sousa AA, Koblan LW, Levy JM, Chen PJ, Wilson C, Newby GA, Raguram A, et al. 2019. Search-and-replace genome editing without double-strand breaks or donor DNA. *Nature* **576**: 149–157.

Arnold CD, Gerlach D, Stelzer C, Boryń ŁM, Rath M, Stark A. 2013. Genome-Wide Quantitative Enhancer Activity Maps Identified by STARR-seq. *Science* **339**: 1074–1077.

Arnosti DN, Kulkarni MM. 2005. Transcriptional enhancers: Intelligent enhanceosomes or flexible billboards? *J Cell Biochem* **94**: 890–898.

Atkinson TJ, Halfon MS. 2014. REGULATION OF GENE EXPRESSION IN THE GENOMIC CONTEXT. *Computational and Structural Biotechnology Journal* **9**: e201401001.

Bejerano G, Pheasant M, Makunin I, Stephen S, Kent WJ, Mattick JS, Haussler D. 2004. Ultraconserved Elements in the Human Genome. *Science* **304**: 1321–1325.

Ben-Naim O, Eshed R, Parnis A, Teper-Bamnolker P, Shalit A, Coupland G, Samach A, Lifschitz E. 2006. The CCAAT binding factor can mediate interactions between CONSTANS-like proteins and DNA. *The Plant Journal* **46**: 462–476.

Bewick AJ, Schmitz RJ. 2017. Gene body DNA methylation in plants. *Current Opinion in Plant Biology* **36**: 103–110.

Boyle AP, Davis S, Shulha HP, Meltzer P, Margulies EH, Weng Z, Furey TS, Crawford GE. 2008. High-Resolution Mapping and Characterization of Open Chromatin across the Genome. *Cell* **132**: 311–322.

Brand U, Grünewald M, Hobe M, Simon R. 2002. Regulation of CLV3 Expression by Two Homeobox Genes in Arabidopsis. *Plant Physiology* **129**: 565–575.

Brooks C, Nekrasov V, Lippman ZB, Van Eck J. 2014. Efficient Gene Editing in Tomato in the First Generation Using the Clustered Regularly Interspaced Short Palindromic Repeats/CRISPR-Associated9 System. *Plant Physiology* **166**: 1292–1297.

Buenrostro JD, Giresi PG, Zaba LC, Chang HY, Greenleaf WJ. 2013. Transposition of native chromatin for multimodal regulatory analysis and personal epigenomics. *Nat Methods* **10**: 1213–1218.

Burgess D, Freeling M. 2014. The Most Deeply Conserved Noncoding Sequences in Plants Serve Similar Functions to Those in Vertebrates Despite Large Differences in Evolutionary Rates[W]. *Plant Cell* **26**: 946–961.

Calo E, Wysocka J. 2013. Modification of enhancer chromatin: what, how and why? *Mol Cell* **49**: 10.1016/j.molcel.2013.01.038.

Cameron RA, Davidson EH. 2009. Flexibility of transcription factor target site position in conserved cis-regulatory modules. *Developmental Biology* **336**: 122–135.

Cande J, Goltsev Y, Levine MS. 2009a. Conservation of enhancer location in divergent insects. *Proceedings of the National Academy of Sciences* **106**: 14414–14419.

Cande JD, Chopra VS, Levine M. 2009b. Evolving enhancer-promoter interactions within the tinman complex of the flour beetle, Tribolium castaneum. *Development* **136**: 3153–3160.

Cao K, Cui L, Zhou X, Ye L, Zou Z, Deng S. 2016. Four Tomato FLOWERING LOCUS T-Like Proteins Act Antagonistically to Regulate Floral Initiation. *Frontiers in Plant Science* **6**. https://www.frontiersin.org/articles/10.3389/fpls.2015.01213 (Accessed July 25, 2023).

Castro-Mondragon JA, Riudavets-Puig R, Rauluseviciute I, Berhanu Lemma R, Turchi L, Blanc-Mathieu R, Lucas J, Boddie P, Khan A, Manosalva Pérez N, et al. 2022. JASPAR 2022: the 9th release of the open-access database of transcription factor binding profiles. *Nucleic Acids Research* **50**: D165–D173.

Čermák T, Baltes NJ, Čegan R, Zhang Y, Voytas DF. 2015. High-frequency, precise modification of the tomato genome. *Genome Biology* **16**: 232.

Chan YF, Marks ME, Jones FC, Villarreal G, Shapiro MD, Brady SD, Southwick AM, Absher DM, Grimwood J, Schmutz J, et al. 2010. Adaptive Evolution of Pelvic Reduction in Sticklebacks by Recurrent Deletion of a Pitx1 Enhancer. *Science* **327**: 302–305.

Chen J, Li S, He Y, Li J, Xia L. 2022. An update on precision genome editing by homology-directed repair in plants. *Plant Physiology* **188**: 1780–1794.

Choi J, Lysakovskaia K, Stik G, Demel C, Söding J, Tian TV, Graf T, Cramer P. 2021. Evidence for additive and synergistic action of mammalian enhancers during cell fate determination eds. C.P. Ponting, K. Struhl, and H. Singh. *eLife* **10**: e65381.

Chow C-N, Lee T-Y, Hung Y-C, Li G-Z, Tseng K-C, Liu Y-H, Kuo P-L, Zheng H-Q, Chang W-C. 2019. PlantPAN3.0: a new and updated resource for reconstructing transcriptional regulatory networks from ChIP-seq experiments in plants. *Nucleic Acids Res* **47**: D1155–D1163.

Crevillén P, Sonmez C, Wu Z, Dean C. 2013. A gene loop containing the floral repressor FLC is disrupted in the early phase of vernalization. *EMBO J* **32**: 140–148.

Crisp PA, Marand AP, Noshay JM, Zhou P, Lu Z, Schmitz RJ, Springer NM. 2020. Stable unmethylated DNA demarcates expressed genes and their cis-regulatory space in plant genomes. *Proceedings of the National Academy of Sciences* **117**: 23991–24000.

Deribe YL, Pawson T, Dikic I. 2010. Post-translational modifications in signal integration. *Nat Struct Mol Biol* **17**: 666–672.

Domingo J, Baeza-Centurion P, Lehner B. 2019. The Causes and Consequences of Genetic Interactions (Epistasis). *Annual Review of Genomics and Human Genetics* **20**: 433–460.

Dong P, Tu X, Chu P-Y, Lü P, Zhu N, Grierson D, Du B, Li P, Zhong S. 2017. 3D Chromatin Architecture of Large Plant Genomes Determined by Local A/B Compartments. *Molecular Plant* **10**: 1497–1509.

Dong P, Tu X, Liang Z, Kang B-H, Zhong S. 2020. Plant and animal chromatin three-dimensional organization: similar structures but different functions. *Journal of Experimental Botany* **71**: 5119–5128.

Du Y, Liu L, Peng Y, Li M, Li Y, Liu D, Li X, Zhang Z. 2020. UNBRANCHED3 Expression and Inflorescence Development is Mediated by UNBRANCHED2 and the Distal Enhancer, KRN4, in Maize. *PLoS Genet* **16**: e1008764.

Eshed Y, Zamir D. 1995. An introgression line population of Lycopersicon pennellii in the cultivated tomato enables the identification and fine mapping of yield-associated QTL. *Genetics* **141**: 1147–1162.

Feng D, Liang Z, Wang Y, Yao J, Yuan Z, Hu G, Qu R, Xie S, Li D, Yang L, et al. 2022. Chromatin accessibility illuminates single-cell regulatory dynamics of rice root tips. *BMC Biology* **20**: 274.

Feng S, Cokus SJ, Schubert V, Zhai J, Pellegrini M, Jacobsen SE. 2014. Genome-wide Hi-C Analyses in Wild-Type and Mutants Reveal High-Resolution Chromatin Interactions in Arabidopsis. *Molecular Cell* **55**: 694–707.

Fletcher JC, Brand U, Running MP, Simon R, Meyerowitz EM. 1999. Signaling of Cell Fate Decisions by CLAVATA3 in Arabidopsis Shoot Meristems. *Science* **283**: 1911–1914.

Fouracre JP, Harrison CJ. 2022. How was apical growth regulated in the ancestral land plant? Insights from the development of non-seed plants. *Plant Physiology* **190**: 100–112.

Frankel N, Davis GK, Vargas D, Wang S, Payre F, Stern DL. 2010. Phenotypic robustness conferred by apparently redundant transcriptional enhancers. *Nature* **466**: 490–493.

Frankel N, Erezyilmaz DF, McGregor AP, Wang S, Payre F, Stern DL. 2011. Morphological evolution caused by many subtle-effect substitutions in regulatory DNA. *Nature* **474**: 598–603.

Frazer KA, Pachter L, Poliakov A, Rubin EM, Dubchak I. 2004. VISTA: computational tools for comparative genomics. *Nucleic Acids Research* **32**: W273–W279.

Freire-Rios A, Tanaka K, Crespo I, van der Wijk E, Sizentsova Y, Levitsky V, Lindhoud S, Fontana M, Hohlbein J, Boer DR, et al. 2020. Architecture of DNA elements mediating ARF transcription factor binding and auxin-responsive gene expression in Arabidopsis. *Proceedings of the National Academy of Sciences* **117**: 24557–24566.

Gallego-Bartolomé J, Gardiner J, Liu W, Papikian A, Ghoshal B, Kuo HY, Zhao JM-C, Segal DJ, Jacobsen SE. 2018. Targeted DNA demethylation of the Arabidopsis genome using the human TET1 catalytic domain. *Proceedings of the National Academy of Sciences* **115**: E2125–E2134.

Gallego-Bartolomé J, Liu W, Kuo PH, Feng S, Ghoshal B, Gardiner J, Zhao JM-C, Park SY, Chory J, Jacobsen SE. 2019. Co-targeting RNA Polymerases IV and V Promotes Efficient De Novo DNA Methylation in Arabidopsis. *Cell* **176**: 1068-1082.e19.

Galupa R, Alvarez-Canales G, Borst NO, Fuqua T, Gandara L, Misunou N, Richter K, Alves MRP, Karumbi E, Perkins ML, et al. 2023. Enhancer architecture and chromatin accessibility constrain phenotypic space during Drosophila development. *Developmental Cell* **58**: 51-62.e4.

Gaudelli NM, Komor AC, Rees HA, Packer MS, Badran AH, Bryson DI, Liu DR. 2017. Programmable base editing of A•T to G•C in genomic DNA without DNA cleavage. *Nature* **551**: 464–471.

Ghoshal B, Picard CL, Vong B, Feng S, Jacobsen SE. 2021. CRISPR-based targeting of DNA methylation in Arabidopsis thaliana by a bacterial CG-specific DNA methyltransferase. *Proceedings of the National Academy of Sciences* **118**: e2125016118.

Grant CE, Bailey TL, Noble WS. 2011. FIMO: scanning for occurrences of a given motif. *Bioinformatics* **27**: 1017–1018.

Greene B, Walko R, Hake S. 1994. Mutator Insertions in an Intron of the Maize Knotted1 Gene Result in Dominant Suppressible Mutations. *Genetics* **138**: 1275–1285.

Grützner R, Martin P, Horn C, Mortensen S, Cram EJ, Lee-Parsons CWT, Stuttmann J, Marillonnet S. 2021. High-efficiency genome editing in plants mediated by a Cas9 gene containing multiple introns. *Plant Communications* **2**: 100135.

Hajheidari M, Huang SC. 2022. Elucidating the biology of transcription factor–DNA interaction for accurate identification of cis-regulatory elements. *Current Opinion in Plant Biology* **68**: 102232.

Han J, Zhang Z, Wang K. 2018a. 3C and 3C-based techniques: the powerful tools for spatial genome organization deciphering. *Molecular Cytogenetics* **11**: 21.

Han K, Lee H, Ro N, Hur O, Lee J, Kwon J, Kang B. 2018b. QTL mapping and GWAS reveal candidate genes controlling capsaicinoid content in Capsicum. *Plant Biotechnol J* **16**: 1546–1558.

Hare EE, Peterson BK, Iyer VN, Meier R, Eisen MB. 2008. Sepsid even-skipped Enhancers Are Functionally Conserved in Drosophila Despite Lack of Sequence Conservation. *PLOS Genetics* **4**: e1000106.

Haudry A, Platts AE, Vello E, Hoen DR, Leclercq M, Williamson RJ, Forczek E, Joly-Lopez Z, Steffen JG, Hazzouri KM, et al. 2013. An atlas of over 90,000 conserved noncoding sequences provides insight into crucifer regulatory regions. *Nat Genet* **45**: 891–898.

Hay D, Hughes JR, Babbs C, Davies JOJ, Graham BJ, Hanssen LLP, Kassouf MT, Oudelaar AM, Sharpe JA, Suciu MC, et al. 2016. Genetic dissection of the α-globin super-enhancer in vivo. *Nat Genet* **48**: 895–903.

Hellman LM, Fried MG. 2007. Electrophoretic mobility shift assay (EMSA) for detecting protein–nucleic acid interactions. *Nat Protoc* **2**: 1849–1861.

Hendelman A, Zebell S, Rodriguez-Leal D, Dukler N, Robitaille G, Wu X, Kostyun J, Tal L, Wang P, Bartlett ME, et al. 2021. Conserved pleiotropy of an ancient plant homeobox gene uncovered by cis-regulatory dissection. *Cell* **184**: 1724-1739.e16.

Hnisz D, Shrinivas K, Young RA, Chakraborty AK, Sharp PA. 2017. A Phase Separation Model for Transcriptional Control. *Cell* **169**: 13–23.

Ishikawa R, Aoki M, Kurotani K, Yokoi S, Shinomura T, Takano M, Shimamoto K. 2011. Phytochrome B regulates Heading date 1 (Hd1)-mediated expression of rice florigen Hd3a and critical day length in rice. *Mol Genet Genomics* **285**: 461–470.

Johnson DS, Mortazavi A, Myers RM, Wold B. 2007. Genome-wide mapping of in vivo protein-DNA interactions. *Science* **316**: 1497–1502.

Jores T, Tonnies J, Dorrity MW, Cuperus JT, Fields S, Queitsch C. 2020. Identification of Plant Enhancers and Their Constituent Elements by STARR-seq in Tobacco Leaves[OPEN]. *Plant Cell* **32**: 2120–2131.

Kalay G, Wittkopp PJ. 2010. Nomadic Enhancers: Tissue-Specific cis-Regulatory Elements of yellow Have Divergent Genomic Positions among Drosophila Species. *PLOS Genetics* **6**: e1001222.

Kaufmann K, Wellmer F, Muiño JM, Ferrier T, Wuest SE, Kumar V, Serrano-Mislata A, Madueño F, Krajewski P, Meyerowitz EM, et al. 2010. Orchestration of floral initiation by APETALA1. *Science* **328**: 85–89.

Kim S, Wysocka J. 2023. Deciphering the multi-scale, quantitative cis-regulatory code. *Molecular Cell* **83**: 373–392.

Konishi S, Izawa T, Lin SY, Ebana K, Fukuta Y, Sasaki T, Yano M. 2006. An SNP Caused Loss of Seed Shattering During Rice Domestication. *Science* **312**: 1392–1396.

Koonin EV, Wolf YI. 2010. Constraints and plasticity in genome and molecular-phenome evolution. *Nat Rev Genet* **11**: 487–498.

Kurbidaeva A, Purugganan M. 2021. Insulators in Plants: Progress and Open Questions. *Genes (Basel)* **12**: 1422.

Kvon EZ, Kamneva OK, Melo US, Barozzi I, Osterwalder M, Mannion BJ, Tissières V, Pickle CS, Plajzer-Frick I, Lee EA, et al. 2016. Progressive Loss of Function in a Limb Enhancer during Snake Evolution. *Cell* **167**: 633-642.e11.

Lai X, Stigliani A, Lucas J, Hugouvieux V, Parcy F, Zubieta C. 2020. Genome-wide binding of SEPALLATA3 and AGAMOUS complexes determined by sequential DNA-affinity purification sequencing. *Nucleic Acids Research* **48**: 9637–9648.

LeBlanc C, Zhang F, Mendez J, Lozano Y, Chatpar K, Irish VF, Jacob Y. 2018. Increased efficiency of targeted mutagenesis by CRISPR/Cas9 in plants using heat stress. *The Plant Journal* **93**: 377–386.

Lettice LA, Heaney SJH, Purdie LA, Li L, de Beer P, Oostra BA, Goode D, Elgar G, Hill RE, de Graaff E. 2003. A long-range Shh enhancer regulates expression in the developing limb and fin and is associated with preaxial polydactyly. *Human Molecular Genetics* **12**: 1725–1735.

Li C, Li W, Zhou Z, Chen H, Xie C, Lin Y. 2020a. A new rice breeding method: CRISPR/Cas9 system editing of the Xa13 promoter to cultivate transgene-free bacterial blight-resistant rice. *Plant Biotechnology Journal* **18**: 313–315.

Li E, Liu H, Huang L, Zhang X, Dong X, Song W, Zhao H, Lai J. 2019. Long-range interactions between proximal and distal regulatory regions in maize. *Nat Commun* **10**: 2633.

Li J, Zhang C, He Y, Li S, Yan L, Li Y, Zhu Z, Xia L. 2023. Plant base editing and prime editing: The current status and future perspectives. *Journal of Integrative Plant Biology* **65**: 444–467.

Li Q, Sapkota M, van der Knaap E. 2020b. Perspectives of CRISPR/Cas-mediated cis-engineering in horticulture: unlocking the neglected potential for crop improvement. *Hortic Res* **7**: 1–11.

Lieberman-Aiden E, van Berkum NL, Williams L, Imakaev M, Ragoczy T, Telling A, Amit I, Lajoie BR, Sabo PJ, Dorschner MO, et al. 2009. Comprehensive mapping of long range interactions reveals folding principles of the human genome. *Science* **326**: 289–293.

Lim CKW, McCallister TX, Saporito-Magriña C, McPheron GD, Krishnan R, Zeballos C MA, Powell JE, Clark LV, Perez-Pinera P, Gaj T. 2022. CRISPR base editing of cis-regulatory elements enables the perturbation of neurodegeneration-linked genes. *Molecular Therapy* **30**: 3619–3631.

Lin Q, Zong Y, Xue C, Wang S, Jin S, Zhu Z, Wang Y, Anzalone AV, Raguram A, Doman JL, et al. 2020. Prime genome editing in rice and wheat. *Nat Biotechnol* **38**: 582–585.

Liu C, Teo ZWN, Bi Y, Song S, Xi W, Yang X, Yin Z, Yu H. 2013. A Conserved Genetic Pathway Determines Inflorescence Architecture in Arabidopsis and Rice. *Developmental Cell* **24**: 612–622.

Liu C, Wang C, Wang G, Becker C, Zaidem M, Weigel D. 2016. Genome-wide analysis of chromatin packing in Arabidopsis thaliana at single-gene resolution. *Genome Res* **26**: 1057–1068.

Liu L, Gallagher J, Arevalo ED, 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. *Nat Plants* **7**: 287–294.

Liu X, Kim YJ, Müller R, Yumul RE, Liu C, Pan Y, Cao X, Goodrich J, Chen X. 2011. AGAMOUS Terminates Floral Stem Cell Maintenance in Arabidopsis by Directly Repressing WUSCHEL through Recruitment of Polycomb Group Proteins[W]. *Plant Cell* **23**: 3654–3670.

Lloyd JPB, Lister R. 2022. Epigenome plasticity in plants. *Nat Rev Genet* **23**: 55–68.

Long HK, Prescott SL, Wysocka J. 2016. Ever-Changing Landscapes: Transcriptional Enhancers in Development and Evolution. *Cell* **167**: 1170–1187.

Louwers M, Bader R, Haring M, van Driel R, de Laat W, Stam M. 2009. Tissue- and expression level-specific chromatin looping at maize b1 epialleles. *Plant Cell* **21**: 832–842.

Lu Z, Marand AP, Ricci WA, Ethridge CL, Zhang X, Schmitz RJ. 2019. The prevalence, evolution and chromatin signatures of plant regulatory elements. *Nat Plants* **5**: 1250–1259.

Marand AP, Eveland AL, Kaufmann K, Springer NM. 2023. cis-Regulatory Elements in Plant Development, Adaptation, and Evolution. *Annual Review of Plant Biology* **74**: 111–137.

Marand AP, Schmitz RJ. 2022. Single-cell analysis of cis-regulatory elements. *Current Opinion in Plant Biology* **65**: 102094.

Mayr C. 2019. What Are 3′ UTRs Doing? *Cold Spring Harb Perspect Biol* **11**: a034728.

Mazo-Vargas A, Langmüller AM, Wilder A, van der Burg KRL, Lewis JJ, Messer PW, Zhang L, Martin A, Reed RD. 2022. Deep cis-regulatory homology of the butterfly wing pattern ground plan. *Science* **378**: 304–308.

McGarry RC, Ayre BG. 2012. Manipulating plant architecture with members of the CETS gene family. *Plant Science* **188–189**: 71–81.

McNabb DS, Reed R, Marciniak RA. 2005. Dual Luciferase Assay System for Rapid Assessment of Gene Expression in Saccharomyces cerevisiae. *Eukaryot Cell* **4**: 1539–1549.

Meyer RS, Purugganan MD. 2013. Evolution of crop species: genetics of domestication and diversification. *Nat Rev Genet* **14**: 840–852.

Molinero-Rosales N, Latorre A, Jamilena M, Lozano R. 2004. SINGLE FLOWER TRUSS regulates the transition and maintenance of flowering in tomato. *Planta* **218**: 427–434.

Muños S, Ranc N, Botton E, Bérard A, Rolland S, Duffé P, Carretero Y, Le Paslier M-C, Delalande C, Bouzayen M, et al. 2011. Increase in Tomato Locule Number Is Controlled by Two Single-Nucleotide Polymorphisms Located Near WUSCHEL. *Plant Physiology* **156**: 2244–2254.

Nagano T, Lubling Y, Yaffe E, Wingett SW, Dean W, Tanay A, Fraser P. 2015. Single-cell Hi-C for genome-wide detection of chromatin interactions that occur simultaneously in a single cell. *Nat Protoc* **10**: 1986–2003.

Ohyama K, Shinohara H, Ogawa-Ohnishi M, Matsubayashi Y. 2009. A glycopeptide regulating stem cell fate in Arabidopsis thaliana. *Nat Chem Biol* **5**: 578–580.

O'Malley RC, Huang SC, Song L, Lewsey MG, Bartlett A, Nery JR, Galli M, Gallavotti A, Ecker JR. 2016. Cistrome and Epicistrome Features Shape the Regulatory DNA Landscape. *Cell* **165**: 1280–1292.

Osnato M, Cota I, Nebhnani P, Cereijo U, Pelaz S. 2022. Photoperiod Control of Plant Growth: Flowering Time Genes Beyond Flowering. *Frontiers in Plant Science* **12**. https://www.frontiersin.org/articles/10.3389/fpls.2021.805635 (Accessed July 25, 2023).

Osterwalder M, Barozzi I, Tissières V, Fukuda-Yuzawa Y, Mannion BJ, Afzal SY, Lee EA, Zhu Y, Plajzer-Frick I, Pickle CS, et al. 2018. Enhancer redundancy provides phenotypic robustness in mammalian development. *Nature* **554**: 239–243.

Pajoro A, Madrigal P, Muiño JM, Matus JT, Jin J, Mecchia MA, Debernardi JM, Palatnik JF, Balazadeh S, Arif M, et al. 2014. Dynamics of chromatin accessibility and gene regulation by MADS-domain transcription factors in flower development. *Genome Biology* **15**: R41.

Pajoro A, Muiño JM, Angenent GC, Kaufmann K. 2018. Profiling Nucleosome Occupancy by MNase-seq: Experimental Protocol and Computational Analysis. In *Plant Chromatin Dynamics: Methods and Protocols* (eds. M. Bemer and C. Baroux), *Methods in Molecular Biology*, pp. 167–181, Springer, New York, NY https://doi.org/10.1007/978-1-4939-7318-7_11 (Accessed July 12, 2023).

Perales M, Rodriguez K, Snipes S, Yadav RK, Diaz-Mendoza M, Reddy GV. 2016. Threshold-dependent transcriptional discrimination underlies stem cell homeostasis. *Proceedings of the National Academy of Sciences* **113**: E6298–E6306.

Preger-Ben Noon E, Davis FP, Stern DL. 2016. Evolved Repression Overcomes Enhancer Robustness. *Developmental Cell* **39**: 572–584.

Prescott SL, Srinivasan R, Marchetto MC, Grishina I, Narvaiza I, Selleri L, Gage FH, Swigut T, Wysocka J. 2015. Enhancer Divergence and cis-Regulatory Evolution in the Human and Chimp Neural Crest. *Cell* **163**: 68–83.

Qüesta JI, Song J, Geraldo N, An H, Dean C. 2016. Arabidopsis transcriptional repressor VAL1 triggers Polycomb silencing at FLC during vernalization. *Science* **353**: 485–488.

Rebeiz M, Pool JE, Kassner VA, Aquadro CF, Carroll SB. 2009. Stepwise Modification of a Modular Enhancer Underlies Adaptation in a Drosophila Population. *Science* **326**: 1663–1667.

Rebeiz M, Tsiantis M. 2017. Enhancer evolution and the origins of morphological novelty. *Current Opinion in Genetics & Development* **45**: 115–123.

Ricci WA, Lu Z, Ji L, Marand AP, Ethridge CL, Murphy NG, Noshay JM, Galli M, Mejía-Guerra MK, Colomé-Tatché M, et al. 2019. Widespread Long-range Cis-Regulatory Elements in the Maize Genome. *Nat Plants* **5**: 1237–1249.

Roca Paixão JF, Gillet F-X, Ribeiro TP, Bournaud C, Lourenço-Tessutti IT, Noriega DD, Melo BP de, de Almeida-Engler J, Grossi-de-Sa MF. 2019. Improved drought stress tolerance in Arabidopsis by CRISPR/dCas9 fusion with a Histone AcetylTransferase. *Sci Rep* **9**: 8080.

Rodríguez-Leal D, Lemmon ZH, Man J, Bartlett ME, Lippman ZB. 2017. Engineering Quantitative Trait Variation for Crop Improvement by Genome Editing. *Cell* **171**: 470-480.e8.

Rodriguez-Leal D, Xu C, Kwon C-T, Soyars C, Arevalo ED, Man J, Lei L, Lemmon ZH, Jones DS, Van Eck J, et al. 2019. Evolution of buffering in a genetic circuit controlling plant stem cell proliferation. *Nat Genet* **51**: 786–792.

Rosa S, Duncan S, Dean C. 2016. Mutually exclusive sense–antisense transcription at FLC facilitates environmentally induced gene repression. *Nat Commun* **7**: 13031.

Salvi S, Sponza G, Morgante M, Tomes D, Niu X, Fengler KA, Meeley R, Ananiev EV, Svitashev S, Bruggemann E, et al. 2007. Conserved noncoding genomic sequences associated with a

flowering-time quantitative trait locus in maize. *Proceedings of the National Academy of Sciences* **104**: 11376–11381.

Savadel SD, Hartwig T, Turpin ZM, Vera DL, Lung P-Y, Sui X, Blank M, Frommer WB, Dennis JH, Zhang J, et al. 2021. The native cistrome and sequence motif families of the maize ear. *PLoS Genet* **17**: e1009689.

Schmitz RJ, Grotewold E, Stam M. 2021. Cis-regulatory sequences in plants: Their importance, discovery, and future challenges. *Plant Cell* **34**: 718–741.

Schoenfelder S, Fraser P. 2019. Long-range enhancer-promoter contacts in gene expression control. *Nat Rev Genet* **20**: 437–455.

Shi J, Gao H, Wang H, Lafitte HR, Archibald RL, Yang M, Hakimi SM, Mo H, Habben JE. 2017. ARGOS8 variants generated by CRISPR-Cas9 improve maize grain yield under field drought stress conditions. *Plant Biotechnology Journal* **15**: 207–216.

Shin HY, Willi M, Yoo KH, Zeng X, Wang C, Metser G, Hennighausen L. 2016. Hierarchy within the mammary STAT5-driven Wap super-enhancer. *Nat Genet* **48**: 904–911.

Sieburth LE, Meyerowitz EM. 1997. Molecular dissection of the AGAMOUS control region shows that cis elements for spatial regulation are located intragenically. *Plant Cell* **9**: 355–365.

Sloan J, Hakenjos JP, Gebert M, Ermakova O, Gumiero A, Stier G, Wild K, Sinning I, Lohmann JU. 2020. Structural basis for the complex DNA binding behavior of the plant stem cell regulator WUSCHEL. *Nat Commun* **11**: 2223.

Snetkova V, Ypsilanti AR, Akiyama JA, Mannion BJ, Plajzer-Frick I, Novak CS, Harrington AN, Pham QT, Kato M, Zhu Y, et al. 2021. Ultraconserved enhancer function does not require perfect sequence conservation. *Nat Genet* **53**: 521–528.

Somssich M, Je BI, Simon R, Jackson D. 2016. CLAVATA-WUSCHEL signaling in the shoot meristem. *Development* **143**: 3238–3248.

Song J, Zhang S, Wang X, Sun S, Liu Z, Wang K, Wan H, Zhou G, Li R, Yu H, et al. 2020. Variations in Both FTL1 and SP5G, Two Tomato FT Paralogs, Control Day-Neutral Flowering. *Molecular Plant* **13**: 939–942.

Soyk S, Müller NA, Park SJ, Schmalenbach I, Jiang K, Hayama R, Zhang L, Van Eck J, Jiménez-Gómez JM, Lippman ZB. 2017. Variation in the flowering gene SELF PRUNING 5G promotes day-neutrality and early yield in tomato. *Nat Genet* **49**: 162–168.

Stephen S, Pheasant M, Makunin IV, Mattick JS. 2008. Large-Scale Appearance of Ultraconserved Elements in Tetrapod Genomes and Slowdown of the Molecular Clock. *Molecular Biology and Evolution* **25**: 402–408.

Su YH, Zhou C, Li YJ, Yu Y, Tang LP, Zhang WJ, Yao WJ, Huang R, Laux T, Zhang XS. 2020. Integration of pluripotency pathways regulates stem cell maintenance in the Arabidopsis shoot meristem. *Proceedings of the National Academy of Sciences* **117**: 22561–22571.

Suárez-López P, Wheatley K, Robson F, Onouchi H, Valverde F, Coupland G. 2001. CONSTANS mediates between the circadian clock and the control of flowering in Arabidopsis. *Nature* **410**: 1116–1120.

Sumiyama K, Saitou N. 2011. Loss-of-function mutation in a repressor module of human-specifically activated enhancer HACNS1. *Mol Biol Evol* **28**: 3005–3007.

Sun J, He N, Niu L, Huang Y, Shen W, Zhang Y, Li L, Hou C. 2019. Global Quantitative Mapping of Enhancers in Rice by STARR-seq. *Genomics Proteomics Bioinformatics* **17**: 140–153.

Tannenbaum M, Sarusi-Portuguez A, Krispil R, Schwartz M, Loza O, Benichou JIC, Mosquna A, Hakim O. 2018. Regulatory chromatin landscape in Arabidopsis thaliana roots uncovered by coupling INTACT and ATAC-seq. *Plant Methods* **14**: 113.

Taylor SC, Laperriere G, Germain H. 2017. Droplet Digital PCR versus qPCR for gene expression analysis with low abundant targets: from variable nonsense to publication quality data. *Sci Rep* **7**: 2409.

Thurman RE, Rynes E, Humbert R, Vierstra J, Maurano MT, Haugen E, Sheffield NC, Stergachis AB, Wang H, Vernot B, et al. 2012. The accessible chromatin landscape of the human genome. *Nature* **489**: 75–82.

Tu X, Mejía-Guerra MK, Valdes Franco JA, Tzeng D, Chu P-Y, Shen W, Wei Y, Dai X, Li P, Buckler ES, et al. 2020. Reconstructing the maize leaf regulatory network using ChIP-seq data of 104 transcription factors. *Nat Commun* **11**: 5089.

Van de Velde J, Van Bel M, Vaneechoutte D, Vandepoele K. 2016. A Collection of Conserved Noncoding Sequences to Study Gene Regulation in Flowering Plants1[OPEN]. *Plant Physiol* **171**: 2586–2598.

van der Knaap E, Chakrabarti M, Chu YH, Clevenger JP, Illa-Berenguer E, Huang Z, Keyhaninejad N, Mu Q, Sun L, Wang Y, et al. 2014. What lies beyond the eye: the molecular mechanisms regulating tomato fruit weight and shape. *Frontiers in Plant Science* **5**. https://www.frontiersin.org/articles/10.3389/fpls.2014.00227 (Accessed August 23, 2023).

Van Eck J, Keen P, Tjahjadi M. 2019a. Agrobacterium tumefaciens-Mediated Transformation of Tomato. *Methods Mol Biol* **1864**: 225–234.

Van Eck J, Keen P, Tjahjadi M. 2019b. Agrobacterium tumefaciens-Mediated Transformation of Tomato. In *Transgenic Plants: Methods and Protocols* (eds. S. Kumar, P. Barone, and M. Smith), *Methods in Molecular Biology*, pp. 225–234, Springer, New York, NY https://doi.org/10.1007/978-1-4939-8778-8_16 (Accessed August 23, 2023).

Visel A, Prabhakar S, Akiyama JA, Shoukry M, Lewis KD, Holt A, Plajzer-Frick I, Afzal V, Rubin EM, Pennacchio LA. 2008. Ultraconservation identifies a small subset of extremely constrained developmental enhancers. *Nat Genet* **40**: 158–160.

Wang X, Aguirre L, Rodríguez-Leal D, Hendelman A, Benoit M, Lippman ZB. 2021. Dissecting cis-regulatory control of quantitative trait variation in a plant stem cell circuit. *Nat Plants* **7**: 419–427.

Wang Y, Zhang J, Hu Z, Guo X, Tian S, Chen G. 2019. Genome-Wide Analysis of the MADS-Box Transcription Factor Family in Solanum lycopersicum. *Int J Mol Sci* **20**: 2961.

Weber B, Zicola J, Oka R, Stam M. 2016. Plant Enhancers: A Call for Discovery. *Trends in Plant Science* **21**: 974–987.

Werner S, Engler C, Weber E, Gruetzner R, Marillonnet S. 2012. Fast track assembly of multigene constructs using Golden Gate cloning and the MoClo system. *Bioengineered* **3**: 38–43.

Wigge PA, Kim MC, Jaeger KE, Busch W, Schmid M, Lohmann JU, Weigel D. 2005. Integration of Spatial and Temporal Information During Floral Induction in Arabidopsis. *Science* **309**: 1056–1059.

Williams TM, Selegue JE, Werner T, Gompel N, Kopp A, Carroll SB. 2008. The Regulation and Evolution of a Genetic Switch Controlling Sexually Dimorphic Traits in Drosophila. *Cell* **134**: 610–623.

Wittkopp PJ, Kalay G. 2012. Cis-regulatory elements: molecular mechanisms and evolutionary processes underlying divergence. *Nat Rev Genet* **13**: 59–69.

Wong ES, Zheng D, Tan SZ, Bower NI, Garside V, Vanwalleghem G, Gaiti F, Scott E, Hogan BM, Kikuchi K, et al. 2020. Deep conservation of the enhancer regulatory code in animals. *Science* **370**: eaax8137.

Wu X, Liang Y, Gao H, Wang J, Zhao Y, Hua L, Yuan Y, Wang A, Zhang X, Liu J, et al. 2021. Enhancing rice grain production by manipulating the naturally evolved cis-regulatory element-containing inverted repeat sequence of OsREM20. *Molecular Plant* **14**: 997–1011.

Xu C, Liberatore KL, MacAlister CA, Huang Z, Chu Y-H, Jiang K, Brooks C, Ogawa-Ohnishi M, Xiong G, Pauly M, et al. 2015. A cascade of arabinosyltransferases controls shoot meristem size in tomato. *Nat Genet* **47**: 784–792.

Xu D, Li X, Wu X, Meng L, Zou Z, Bao E, Bian Z, Cao K. 2021. Tomato SlCDF3 Delays Flowering Time by Regulating Different FT-Like Genes Under Long-Day and Short-Day Conditions. *Frontiers in Plant Science* **12**. https://www.frontiersin.org/articles/10.3389/fpls.2021.650068 (Accessed July 24, 2023).

Yan W, Chen D, Schumacher J, Durantini D, Engelhorn J, Chen M, Carles CC, Kaufmann K. 2019. Dynamic control of enhancer activity drives stage-specific gene expression during flower morphogenesis. *Nat Commun* **10**: 1705.

Zafar K, Sedeek KEM, Rao GS, Khan MZ, Amin I, Kamel R, Mukhtar Z, Zafar M, Mansoor S, Mahfouz MM. 2020. Genome Editing Technologies for Rice Improvement: Progress, Prospects, and Safety Concerns. *Frontiers in Genome Editing* **2**. https://www.frontiersin.org/articles/10.3389/fgeed.2020.00005 (Accessed September 3, 2023).

Zhang L, Hu J, Han X, Li J, Gao Y, Richards CM, Zhang C, Tian Y, Liu G, Gul H, et al. 2019. A high-quality apple genome assembly reveals the association of a retrotransposon and red fruit colour. *Nat Commun* **10**: 1494.

Zhang S, Jiao Z, Liu L, Wang K, Zhong D, Li S, Zhao T, Xu X, Cui X. 2018. Enhancer-Promoter Interaction of SELF PRUNING 5G Shapes Photoperiod Adaptation1[OPEN]. *Plant Physiol* **178**: 1631–1642.

Zhang T, Cooper S, Brockdorff N. 2015. The interplay of histone modifications – writers that read. *EMBO Rep* **16**: 1467–1481.

Zhang X, Henriques R, Lin S-S, Niu Q-W, Chua N-H. 2006. Agrobacterium-mediated transformation of Arabidopsis thaliana using the floral dip method. *Nat Protoc* **1**: 641–646.

Zheng L, McMullen MD, Bauer E, Schön C-C, Gierl A, Frey M. 2015. Prolonged expression of the BX1 signature enzyme is associated with a recombination hotspot in the benzoxazinoid gene cluster in Zea mays. *J Exp Bot* **66**: 3917–3930.

Zhong V, Archibald BN, Brophy JAN. 2023. Transcriptional and post-transcriptional controls for tuning gene expression in plants. *Current Opinion in Plant Biology* **71**: 102315.

Zhou M, Coruh C, Xu G, Martins LM, Bourbousse C, Lambolez A, Law JA. 2022. The CLASSY family controls tissue-specific DNA methylation patterns in Arabidopsis. *Nat Commun* **13**: 244.

Zhu Z, Sun B, Cai W, Zhou X, Mao Y, Chen C, Wei J, Cao B, Chen C, Chen G, et al. 2019. Natural variations in the MYB transcription factor MYB31 determine the evolution of extremely pungent peppers. *New Phytologist* **223**: 922–938.

Zhu Z, Tan L, Fu Y, Liu F, Cai H, Xie D, Wu F, Wu J, Matsumoto T, Sun C. 2013. Genetic control of inflorescence architecture during rice domestication. *Nat Commun* **4**: 2200.

Supplementary Tables and Figures

Supplementary Table 1. gRNAs used in *Arabidopsis* CRISPR Chapter 2.

Name of gRNA	Sequence 5'-3' (including PAM)
***AtCLV3* upstream – proximal 1.5 kb CRISPR construct**	
gRNA1	**ATTTATAGCGTAAGCCTACAAGG**
gRNA2	**AAAGTTGTATAAAACGGCAGGGG**
gRNA3	**TGATATATTAGAGTATGTGCCGG**
gRNA4	**AATAGCATCTAAATATGAGAAGG**
gRNA5	**AATATGGATGATACCTTAATCGG**
gRNA6	**TCTGACACGTGCCCATCCGATGG**
gRNA7	**AAAAAGTAGTGGCACCTTATTGG**
gRNA8	**GATGCAGATCTTTAGCAGTATGG**
***AtCLV3* upstream – proximal and distal 3.8 kb CRISPR construct**	
gRNA1	**TTTGGTAATGAAATGAGAAGGGG**
gRNA2	**TGATATATTAGAGTATGTGCCGG**
gRNA3	**AATATGGATGATACCTTAATCGG**
gRNA4	**GATGCAGATCTTTAGCAGTATGG**
gRNA5	**GTGCAGCTCTCAACTCAAGTAGG**
gRNA6	**TTAGATGTGCATGTACATGTGGG**
gRNA7	**AAGTTGATCTATGGTGAGGGTGG**
gRNA8	**CCATTCATAGCTTATTAAGGCGG**
***AtCLV3* downstream**	
gRNA1	**TCTCCAAAGCAATGTACCGTTGG**
gRNA2	**ACCGACTTTGGGGCAGTGACAGG**
gRNA3	**TAAGGATAATAATTAGCTCTAGG**
gRNA4	**GTTATTTGAGGTGGGAAAAGTGG**
gRNA5	**AAGTCTTGGGATGACATTGGAGG**
gRNA6	**TATTGGTTAGTATAGGTGAATGG**
gRNA7	**TTAGTTTACGTCGACTAATTAGG**
gRNA8	**AGGTAGGTATATTACCCAAACGG**

Supplementary Table 2. gRNAs used in tomato CRISPR Chapter 2.

Name of gRNA	Sequence 5'-3' (including PAM)
***SlCLV3* upstream (Wang et al. 2021)**	
gRNA1	**GATATACAACAATGGCTGCATGG**
gRNA2	**GACCTTATCCCCTGCCTTTATGG**
gRNA3	**GAAACACCAAATTATGTTGTAGG**
gRNA4	**GAGATCCATAGTACAGTACTTGG**
gRNA5	**GCAGTAACAAGACAGAGTGACGG**
gRNA6	**GTCCAACAATATATGTTTATCGG**
gRNA7	**GACACCACTCGATTTAAATTTGG**
gRNA8	**GCAATGCAAGTAGCTGCAAAAGG**
***SlCLV3* downstream**	
gRNA1	**TTTAGTAAAGGGTAGTATATGG**
gRNA2	**GCTAGCCAAGTTGGAATATTAGG**
gRNA3	**TCAAAGCTATATACATATCAGGG**
gRNA4	**CTCTTCTCAAAAACGTTCGTTGG**
gRNA5	**GATTGTTAACGAATCAGTTGAGG**
gRNA6	**AACTACAAAGGACTTGCAATAGG**
gRNA7	**TACATAACATACACGTTATAAGG**
R4	
gRNA1	**GCAGTAACAAGACAGAGTGACGG**
gRNA2	**GTCCAACAATATATGTTTATCGG**
gRNA3	**ATATGTTATCAATAAAAGATCGG**
gRNA4	**GGACACCTGCCCAACCCAATAGG**
R1	
gRNA1	**GATATACAACAATGGCTGCATGG**
gRNA2	**GAAAATAGTTAAGAGGCTTTGG**
gRNA3	**GTATTGCCTCAGCATGTAGAGG**

Supplementary Table 3. Genotyping/sequencing primers used in Chapter 2.

Name of primer	**Sequence 5'-3'**
AtCLV3-pro-proximal-F	**TCTGATCTAATAAATTGTTGGCC**
AtCLV3-pro-proximal-R	**GTAGCAGAAAACTCTTCGAATC**
AtCLV3-pro-cds-F	**GCTTGCTCCATCATATGTTTG**
AtCLV3-pro-cds-R	**CTGACACTGCCTGTCACTG**
AtCLV3-pro-full-F	**CCGGAACCGAACATAGCAAA**
AtCLV3-pro-full-R	**GTAGCAGAAAACTCTTCGAATC**
AtCLV3-3p-F	**GCTGAAGTGAATGTAAGATACG**
AtCLV3-3p-R	**TGGCGAAGCGGATCATGTAA**
SlCLV3-pro-F	**AGAGCCTTCCAATAGCTGGC**
SlCLV3-pro-R	**CTGTTTAGGAGTTTCACAGGAGC**
SlCLV3-3p-F	**CACAATGGTGCTAGTCCTAAG**
SlCLV3-3p-R	**GTGTCTGGATATGTTGAAGATG**
SlCLV3-R4-F	**GAGCTAAGATCGAAAAACCGATC**
SlCLV3-R4-R	**GTAGGATCTGGAGAAAGTTGATG**
SlCLV3-R1-F	**CATAAAGGCAGGGGATAAGGTCTC**
SlCLV3-R1-R	**CTGTTTAGGAGTTTCACAGGAGC**

Supplementary Table 4. gRNAs used in tomato CRISPR Chapter 3.

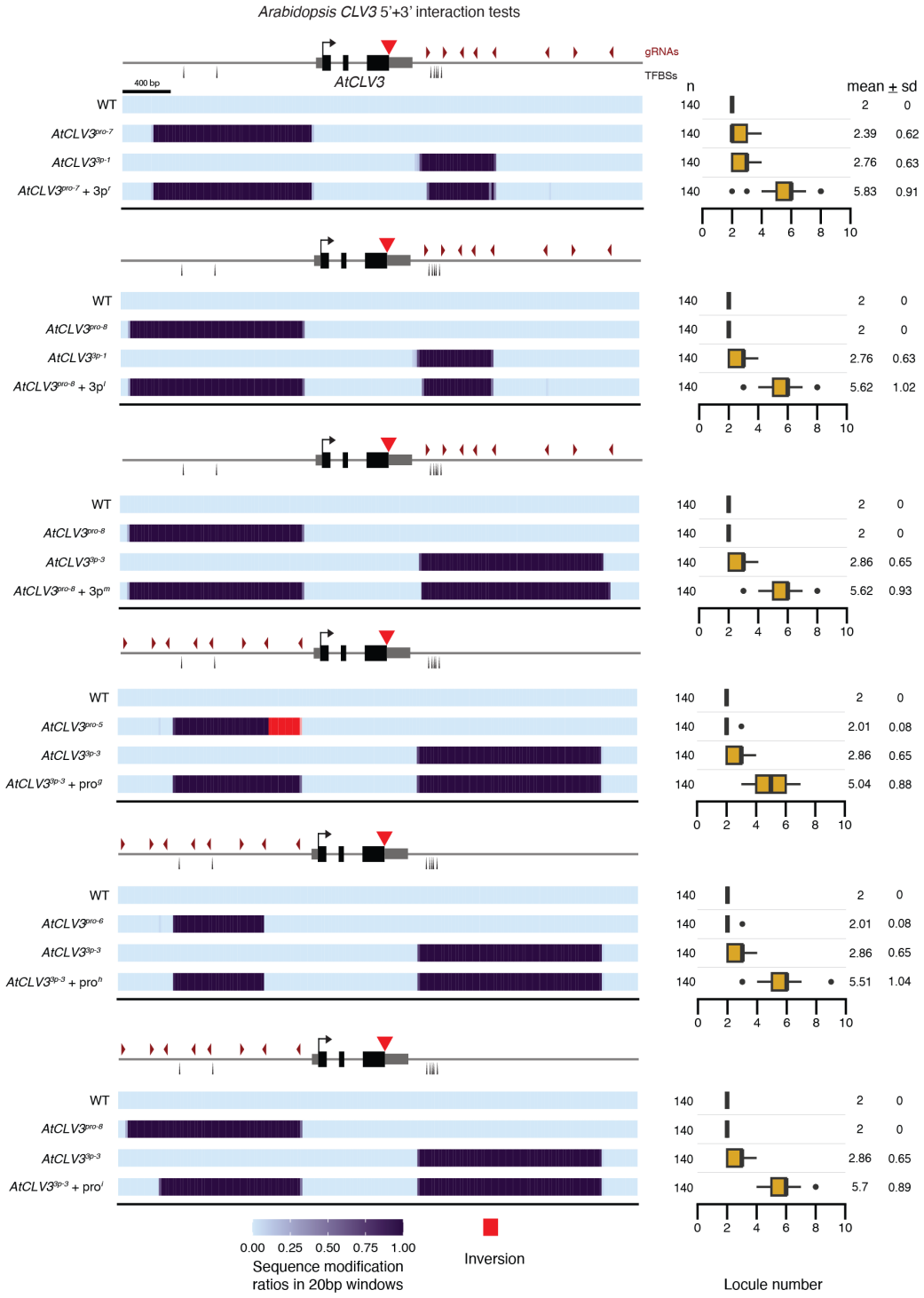
Name of gRNA	Sequence 5'-3' (including PAM)
***SpSP5G* 3'UTR 52 bp enhancer**	
gRNA1	**ATTGACACAGAGTTCGAGAACGG**
gRNA2	**TATCGATATTATATATGAGTTGG**
gRNA3	**ATCTATTTGATAAAAGAGTTTGG**
***SpSP5G* upstream**	
gRNA1	**TATGTTCTTGAGGACACATTTGG**
gRNA2	**TCAAACTGTGGGGTTCGAAGAGG**
gRNA3	**TTAATTGTTGAGCTGTAAAGTGG**
gRNA4	**TTCGAGAAGCGTTATAATTATGG**
gRNA5	**TGTCATGGCTCATGTATCCATGG**
gRNA6	**GAAGGAATGCACCTATAGAGAGG**
gRNA7	**CTCCTAGATCTCCTATCATAAGG**
gRNA8	**AGACAATCCTTACCATAAGGTGG**
***SpSP5G* downstream**	
gRNA1	**AACGGTTATACAAATTGTTATGG**
gRNA2	**TGATTGAGTTTAACTAGCGTTGG**
gRNA3	**ATAATCGTCACATTAATCTATGG**
gRNA4	**TCTCTAACTTAGCGAGGTATTGG**
gRNA5	**TGATGGTGTTGGATTCGATACGG**
gRNA6	**AATTCGATATGAACCATTTATGG**
gRNA7	**TAGGATTGAGTTGGCAGATTTGG**
gRNA8	**TTATTAAAGTTCATGCTCGATGG**
***SpSP5G* ATAC-3p-one**	
gRNA1	**TTTATGTGATGTTATTGAATTGG**
gRNA2	**ACAAACAAGTGGAATATAATTGG**
gRNA3	**TAGGATTGAGTTGGCAGATTTGG**
gRNA4	**CTATGGTACATTAGAATAATAGG**
***SpSP5G* ATAC-3p-two**	
gRNA1	**AAGGAATAATTCGAAATGTTTGG**
gRNA2	**TAAAACATTAAGGATTGTAGTGG**
gRNA3	**TAATATATAACTTGTCCCCATGG**
gRNA4	**ACACGTATTTGCTGTATCCATGG**
***SpSP5G* ATAC-pro-one**	
gRNA1	**AAAACGACGACAATTAGTTCTGG**

gRNA2	**AATGTTGATATCGTTATACTTGG**
gRNA3	**ACATGCTAAATCTCTTATTGTGG**
gRNA4	**AGAAATAATAAATGATTCTATGG**
***SpSP5G* ATAC-pro-two**	
gRNA1	**TTGGGTGGTATACGAGGAGCTGG**
gRNA2	**TTACTTCAAGTGTGGGGACATGG**
gRNA3	**TTAATTAAGATGTACATTTGAGG**
gRNA4	**AGAAACGCACACAAAGAAATCGG**

Supplementary Table 5. Genotyping/sequencing primers used in Chapter 3.

Name of primer	**Sequence 5'-3'**
SpSP5G-pro-F1	**GAACTTTGATCACTATGTGGAG**
SpSP5G-pro-R1	**TATGAGTAGACAAGAGCTAGCT**
SpSP5G-pro-R2	**CGGTGATTAAGTCTGAATGCC**
SpSP5G-3p-F	**ATACGAGTCTACATGTAAAAGTG**
SpSP5G-3p-R	**TGACAAGAATTGTGACGGGG**
SpSP5G-enhancer-F	**GGACATAATCGATTCTCGTCAA**
SpSP5G-enhancer-R	**CGGTGATTAAGTCTGAATGCC**
SpSP5G-ATAC-3p-one-F1	**CTGTATTGAATCAAACAACGTCA**
SpSP5G-ATAC-3p-one-R1	**AGTTATGAAGAGTTGCGGTTTG**
SpSP5G-ATAC-3p-one-F2	**GAGGAAACATGTCAACTAATAGC**
SpSP5G-ATAC-3p-one-R2	**GATCTGATTGGACAGATCCTTC**
SpSP5G-ATAC-3p-two-F	**CTAGCCAATGATTCTTTCTATCA**
SpSP5G-ATAC-3p-two-R	**GATCTGATTGGACAGATCCTTC**
SpSP5G-ATAC-pro-one-F	**GAGTTTGGCTGAAATCTCGATAG**
SpSP5G-ATAC-pro-one-R	**TCCGAAATTGTTTGTCATGTTGC**
SpSP5G-ATAC-pro-two-F	**CTACTCGATAGGAAGTCGAC**
SpSP5G-ATAC-pro-two-R	**GGTAAGGATTGTCTTGACGAC**

Supplementary 2-1. *Arabidopsis CLV3* alleles chosen for interactions tests.



Supplementary 2-2. Tomato *CLV3* alleles chosen for interactions tests.