PHD

Chromatin-level regulation of clustered genes

Bishop, Jade

Award date:
2023

Awarding institution:
University of Bath

Chromatin-level regulation of clustered genes

A thesis submitted for the degree of Doctor of Philosophy
University of Bath
Department of Life Sciences

Jade Bishop

Supervisor: Dr Hans-Wilhelm Nützmann

December 2022

Contents

Acknowledgments

My gratitude first goes to Dr Hans-Wilhelm Nützmann for supporting me throughout my entire PhD project and teaching me how to learn from my mistakes, being patient and giving me great advice. I would like to thank Dr Volkan Cevik for his technical wisdom on USER cloning and CRISPR/Cas9 genome editing. Rob Clayton has been a big support, helping me maintain plant material as well as aiding me in the growth of new plants. Many thanks go to Toby Parkes, who allowed me to ask him unlimited questions throughout the day and made me feel welcome. I could not have done any of my bioinformatics work without the help of Soheila Bayat to troubleshoot my coding, and the Mathematic Resources Centre at the University of Bath, specifically Andrew Chapman who took the time to understand my project and being a sounding board for analysis I wanted to carry out.

Next, my thanks to go students and research assistants that joined our lab and provided experimental help as well as advice. George helped visualise my genome edited plant material using the fluorescent microscope and by Gus staining, and Francesco carried out a large number of genotyping PCRs. The undergrads who joined the lab kept me on my toes, but also provided valuable data that guided my experiments, Max, Megan, Maria, Jack, Hazel and Megan.

Many thanks to the technical team, health and safety and the supply ordering team who were always really friendly and helpful, which allowed my project to run smoothly.

Thanks to my friend, Kyle, who allowed me to rant, gave me scientific support as well as a sound advice throughout my project.

I would also like to thank my family, firstly my parents, for supporting me throughout my Undergraduate degree and always pushing me to do more, and never thinking that any goal of mine was unobtainable. My pets, especially my puppy Nova, for giving me a reason to get out of the house and being there for emotional support.

Massive thanks go to my husband, Ben, for giving me unfaltering support and encouragement and keeping me fed throughout my writing up.

Finally, thanks go to the Royal Society and the University of Bath University Research Studentship Award.

Abstract

The order of genes in a eukaryotic genome has classically been considered random. However, large-scale genomics and transcriptomics studies as well as the frequent discovery of gene clusters have established that eukaryotic gene order is far from random. Indeed, there is increasing realisation that the positioning of genes influences expression patterns. For example, a characteristic feature of gene clusters, groups of neighbouring and functionally related genes, is the co-ordinate regulation of the clustered genes.

Here, using the model organism *Arabidopsis thaliana* and its specialised metabolic gene clusters we have investigated the importance of genomic integrity in gene cluster regulation and aimed to identify key regulatory factors that lead to its co-regulation.

By exploiting T-DNA mutant lines with disrupted linear and, putatively, the 3D cluster organisation, we have analysed how these disruptions change the expression profile of an exemplar gene cluster. Furthermore, we have applied genome editing technologies to re-localise a cluster gene to assess if the native genomic context is important for gene regulation. We have applied bioinformatic data analyses, carried out mutant screens and quantitative gene expression analyses to unveil novel chromatin regulatory enzymes of gene clusters. Moreover, we have investigated sequence and expression conservation of metabolic gene clusters across *A. thaliana* ecotypes.

Our results show that the genomic integrity of the entire cluster space as well as its neighbouring genomic sites is important for cluster regulation. We present preliminary evidence that changing the genomic environment of a gene cluster results in the drastic misregulation of the gene. In addition, we have identified the chromatin marks H3K4me3, H3K18ac, H3K36me3 and H3K9me2 as putative epigenetic regulators of metabolic gene clusters in *A. thaliana* and discovered novel chromatin mutant lines with a role in gene cluster regulation. Our pilot evolutionary analyses indicates that the expression of one gene within the cluster can alter the expression profile of the remaining genes. Collectively, the work presented here has provided novel insights into the coordinate regulation of metabolic gene clusters and the importance of genomic integrity and chromatin environment. It forms a foundation for future research that will design synthetic gene clusters, further advancing our understanding of transgenesis, and unveiling the evolutionary principles of gene cluster formation.

Abbreviations

ABA	Abscisic Acid
AGE	Agarose Gel Electrophoresis
ATAC	Assay for Transposase-Accessible Chromatin
ATX	Arabidopsis Trithorax
bp	Base pairs
Cas9	CRISPR-associated Protein 9
ChIP	Chromatin Immunoprecipitation
COMPASS	Complex Proteins Associated with Set1
CRISPR	Clustered Regularly Interspaced Palindromic Repeats
crRNAs	CRISPR-RNAs
C-TAD	Centromere-Proximal TAD
DARs	Differentially Accessible Regions
DMNT	(*E*)-4,8-dimethyl-1,3,7-nonatriene
DSB	Double Stranded Break
Eaf3	Esa1 associated factor
ERF	Ethylene Response factor
ETP	epipolythiodioxopiperazine
FLC	Flowering Locus C
FLD	Flowering Locus D
GA	Gibberellic Acid
GAI	Gibberellic-Acid Insensitive
GFP	Green Fluorescent Protein
HDACs	Histone Deacetylases
HDR	Homology Directed Repair
Hox	Homeobox
IDD	Intermediate Domain
IME	Intron-Mediated Enhancement
INO80	Inositol Requiring 80
JA	Jasmonate/Jasmonic Acid
*JMJ*C	Jumonji C

*JMJ*D6	Jumonji Domain Containing 6
KDM	Lysine-Specific Demethylase
LCR	Locus Control Region
LDL	LSD-like
LSD	Lysine Specific Demethylase
mRNA	messenger RNA
NASC	Nottingham Arabidopsis Stock Centre
NHEJ	Non-Homologous End Joining
nt	Nucleotides
ORF	Open Reading Frame
OSCs	Oxidosqualene Cyclases
P450s	P450 Monooxygenases
PAM	Protospacer Adjacent Motif
PcG	Polycomb Group
PCR	Polymerase Chain Reaction
PRC2	Polycomb Repressive Complex 2
qPCR	quantitative PCR
RFP	Red Fluorescent Protein
RGA	Repressor Of GAI
Rpd3	Reduced Potassium Dependency 3
rRNA	ribosomal RNA
SCR	Scarecrow
SD	Segmental Duplication
SET	SET (Su(var) 3-9,E(z),Trithorax
sgRNA	single-guide RNA
SIGnAL	SALK Institute Genome Analysis Laboratory
TADs	Topologically Associating Domains
TALENs	Transcription Activator-Like Effector Nucleases
TALEs	Transcription Activator-Like Effectors
TEs	Transposable Elements
TFIID	Transcription Factor II D
tracrRNA	Trans-Activating Crrna

T-TAD	Telomere-Proximal TAD
USER	Uracil-Specific Excision Reagent
ZFNs	Zinc finger nucleases

1 Introduction

The regulation of genes in any genome is multifaceted and relies on a complex orchestration of many layers of molecular systems that gets more complicated as organisms evolve. Over time the genome evolves by changing its structure, sequence and size through mutation, horizontal gene transfer, recombination and other processes.[1]

Here, we focus on the importance of gene order in eukaryotes. We ask why the order in which genes occur on a linear level along chromosomes can alter gene expression by analysing metabolic gene clusters in *Arabidopsis thaliana (A. thaliana)*. We also look on a larger scale at how these gene clusters are co-regulated and which enzymes are involved in their regulation.

In this section, we will first explain the organisation of the three-dimensional eukaryotic genome and the key elements for maintaining and altering its structure. Then, we will provide detailed information about gene clusters. We will describe exemplar clusters and their regulation. Finally, we will introduce the *A. thaliana* genome and focus the rest of the thesis on the thalianol gene cluster.

1.1 Chromatin organisation

A chromosome is a dynamic and complex structure consisting of DNA and tightly associated proteins. In contrast to prokaryotes, where DNA is usually arranged in a circular single chromosome within the cell, eukaryotes usually contain multiple linear chromosomes localised within a distinct cellular compartment, the nucleus. Eukaryotic chromosomal DNA is wound around histone proteins, together forming chromatin.[2] A nucleosome is a basic unit of chromatin which is made up of an octamer of core histones, made up of two copies of each histone protein: H2A, H2B, H3 and H4, with 147 base pairs of DNA wrapped around it (***Figure 1***).[3] Nucleosomes are connected to each other by short stretches of linker DNA typically associated with histone H1 proteins. The repetitive arrangement of the core histone octamers along a DNA molecule is called a nucleosomal array.[4] The dynamic structural organisation of the nucleosomal array has been linked to changing transcriptional activities of genes, and the level of chromatin compaction has been inversely correlated with the transcription rate.[5] Chromatin is a flexible structure that can be actively re-arranged by chromatin remodelling enzymes and by modification with small chemical residues. Highly condensed chromatin is typically referred to as heterochromatin while less condensed 'open' chromatin is called euchromatin.[6] The specific structural properties of chromatin at a chromosomal locus define its accessibility to proteins and therefore the efficiency of all chromosome related processes, from transcription to repair and replication.[7]

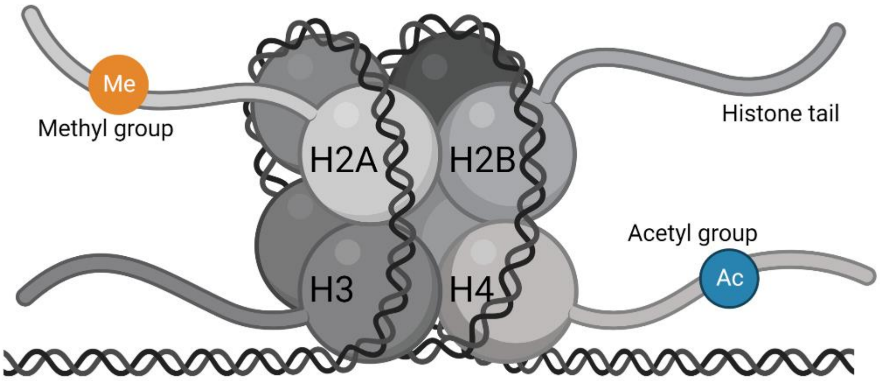


Figure 1: Nucleosome unit. *Histone octamers make up a nucleosomal unit, 147 base pairs are wrapped around 2 pairs of four histone proteins. Histone proteins have a histone tail where histone modification marks can occur, such as methylation and acetylation. (Created with BioRender.com)*

1.1.1 How chromatin can be modified

Chromatin structure can be modified locally and dynamically by addition and removal of small chemical groups. By modifying the histone variant protein or the histone tail post translationally, the chromatin state can be altered to either allow or restrict access of the transcription machinery to the DNA. A nucleosome unit contains histone tails at the N-terminal end of histones. These protrude from the histone octamer[3] and are exposed to possible modifications, such as acetylation, phosphorylation, methylation, ubiquitination and ADP-ribosylation[8]. Histone modifications affect inter-nucleosomal interactions and therefore modify chromatin structure.

1.1.1.1 Histone acetylation

The acetylation of histones, first described in 1964[9], is catalysed by a family of enzymes called histone acetyltransferases. Histone deacetylases (HDACs) are responsible for the removal of acetyl groups. The cofactor acetyl CoA is utilised to catalyse the transfer of an acetyl group to the ε-amino group of lysine side chains. Most commonly, acetylation occurs on the H3 histone protein at lysine residues 9, 14, 18 and 23.[8,10] The acetylation of lysine residues is accompanied by a neutralisation of the positive charge of lysine. This has been suggested to weaken the interaction between DNA and histones and therefore loosening chromatin structure and increasing accessibility of the transcriptional machinery. As such, histone acetylation is widely linked to active transcription.

1.1.1.2 Histone methylation

Methylation of histones occurs on lysine and arginine amino acids. The lysine residue holds up to three methyl groups whereas arginine only holds up to two methyl groups.[11] Methylation marks are typically found on lysine residues (K) 4, 9, 27 and 36. In contrast to acetylation, methylation is associated with both activation, at K4 and K36, and silencing at K9 and K27. The deposition of methyl groups is carried out by a group of enzymes called methyltransferases, and the removal is carried out by demethylases.[12]

The normal development of animals and plants requires H3K27me3 to suppress the expression of genes in specific tissues at appropriate developmental stages.[13] In *A. thaliana* H3K27me3 is present at 15 to 60% of protein-coding genes in a given tissue[14]. Polycomb group (PcG) proteins have been shown to be involved in the deposition of H3K27me3, specifically the polycomb repressive complex 2 (PRC2) which acts as a methyltransferase and is conserved in plants and animals.[15] In plants PRC2 is an important regulator of developmental transitions such as seed formation and flowering and inactivation of PRC2 function leads to embryo abortion or non-viable callus-like structures.[16] H3K27me3 associated gene repression is described as facultative because promoters remain accessible for transcription factor binding and often a paused RNA polymerase remains at the site.[17–20]

Constitutive heterochromatin is typically marked by di- and tri-methylation of histone 3 lysine 9 (H3K9me2/3) throughout eukaryotes.[21] However, in plants constitutive heterochromatin is enriched in H3K9me1/2 and euchromatin is associated with H3K9me3.[22] H3K9me3 is a mark enriched in animals over gene family clusters, this mark has been suggested to protect repetitive gene clusters from recombination.[23] In *A. thaliana* H3K9me3 marks approximately

40% of genes, but low levels are generally detected at regions enriched with transposons.[24] H3K9me2 is the most studied H3K9 methylation mark in *A. thaliana* and is linked to regulating seed dormancy and suppressing genes relates to the ABA (abscisic acid) signalling pathway.[25,26]

Methylation marks at histone 3 lysine 36 (H3K36) are often linked with active transcription, however, they are associated with both active and repressed genes.[27,28] H3K36 methylation is implicated in diverse processes such as DNA repair, recombination and alternative splicing.[29] Genome-wide analysis of H3K36me3 showed that the mark is typically enriched at the 3′ end of the gene body.[30] In *Saccharomyces,* H3K36me3 is bound by Eaf3 (Esa1 associated factor 3) histone acetylase proteins, which recruit the Rpd3 (Reduced Potassium Dependency 3) HDAC complex to nucleosomes. These HDACs remove transcription-coupled hyperacetylation and reduce transcriptional activity.[31] The enzymes that facilitate the transfer of a methyl group to the histone, histone methyltransferases, contain a catalytic SET (SET (Su(var) 3-9,E(z),Trithorax) domain.[29] In *A. thaliana,* H3K36 di- and tri- methylation is associated with active transcription of the Flowering Locus C (FLC).[32] FLC encodes a transcription factor which represses genes that initiate flowering, thus when plants are flowering the FLC locus is repressed[33]. H3K36me also has major functions in the temperature response and nitrate signalling, indicating a role for H3K36 methylation in the abiotic stress response.[34,35]

Active chromatin, or euchromatin, is frequently associated with the methylation of histone 3 lysine 4.[36] H3K4me1 colocalises with the acetylation mark H3K27ac and is typically found in enhancer regions. It is acting as a signature mark for enhancers to increase gene transcription.[37] H3K4me2 typically labels the centre of genes, usually at a higher density than H3K4me1, however its function remains unclear.[38] Highly enriched near the transcription start sites of genes is the activation mark H3K4me3.[39] Although, H3K4me3 has been extensively studied it remains unclear whether H3K4me3 triggers active transcription or is associated with transcriptional memory.[38,39]

1.1.1.3 Histone variants

Some histone genes encode distinct paralogs, known as histone variants, that differ in amino acid sequence from the original. Such variants primarily exist of H2A and H3.[40] Histone variants change the structural properties of the canonical nucleosome and function in chromatin remodelling and modulation of transcription. [41]

1.1.1.3.1 H2A variants

The entry and exit position for DNA at the nucleosome is occupied by H2A. Therefore, it is a key histone in controlling DNA access.[40] In *A. thaliana* nucleosomes usually only contain one single type of H2A variant at a time.[42]

The H2A.X histone variant is an essential component of the DNA damage repair. The C-terminal phosphorylation motif of H2A.X is differing from the core histone structure. The variant is usually deposited in response to DNA damage to recruit enzymes that repair double stranded breaks.[43] In *A. thaliana*, the partial loss of two H2A.X variants leads to weak sensitivity to DNA damage.[44,45] In some single-cell eukaryotes such as yeast, H2A.X is the primary form of H2A.[41] Due to its function, genomic sites prone to DNA damage during cell cycle often become enriched with H2A.X, whereas those not involved in cell cycle show lower levels of enrichment.[46] In mouse embryonic stem cells, H2A.X has also been shown to be deposited at rRNA (ribosomal RNA) promoter regions, repressing rDNA transcription to limit cell proliferation.[47] Mice without H2A.X are viable, however they show genomic instability and males are infertile.[48]

The best studied histone variant, H2A.Z, plays a prominent and complex role in transcriptional regulation, DNA repair and centromeric heterochromatin regulation.[49] H2A.Z only shares 60% amino acid similarity with the canonical H2A, and the major differences which directly influence nucleosome stability are in the C and N terminal domains and the L1 loop.[50,51] H2A.Z deposition promotes RNA polymerase II recruitment, stimulating ATP-dependent chromatin re-modelers and poising genes for transcription[52–54]. Commonly, H2A.Z has been found to be enriched within the nucleosomes that surround the transcriptional start site in genome-wide studies of fungi, animals, plants and protozoa.[55,56] H2A.Z has been found to be a key regulator of genes in response to environmental changes in *Saccharomyces cerevisiae* and

Schizosaccharomyces pombe.[57–59] In *A. thaliana*, impaired deposition of H2A.Z results in the misregulation of many genes involved in the innate immune response.[60] H2A.Z also aids in the correct transcription of genes by facilitating the deposition of H3K27me3 during developmental processes.[61]

The histone variant H2A.W is a plant specific histone variant that is strictly and specifically localized to constitutive heterochromatin in *A. thaliana*.[45] H2A.W has an extended C-terminal tail with a SPKK motif, which is often present in linker H1 which enables it to bind more DNA.[62] It is suggested to have evolved to facilitate the DNA damage response in highly condensed heterochromatin. In conjunction with H1 and H3K9me, H2A.W works to silence a distinct set of transposable elements, thus aiding to prevent transposition.[63]

1.1.1.3.2 H3 variants

There are two major histone H3 variants in eukaryotes: H3.1 and H3.3, which differ only by a few amino acid residues.[64]

An important variant for the development of *Drosophila,* mice and the reprogramming events during development in animals and plants is H3.3.[65–67] H3.3 was originally described as being incorporated at actively transcribed genes at transcription end sites however recent data has shown that H3.3 can accumulate at silent loci in pericentric heterochromatin and telomeres in mice.[68] The mechanisms by which H3.3 promotes gene transcription is unclear. Studies in *A. thaliana* indicate that lines deficient in H3.3 had a decrease in gene body methylation marks. H3.3 has been shown to prevent the deposition of the linker protein, H1, thus relaxing the chromatin to allow access to DNA methyltransferases to increase transcription.[69]

The histone variants H3.1 and H3.2, however, are enriched in silent areas of the genome marked by H3K27 and H3K9 methylation.[70,71] Only differing by one amino acid, H3.1 and H3.2 display different expression pattern and associate with different histone modification marks.[72] [73,74]

1.1.2 Nuclear organisation of chromosomes

Many factors contribute to the regulation of gene expression. The folding of chromatin is a central unit in the interplay of these factors in gene regulation. The plasticity of the chromosome architecture allows genes to switch between active and inactive folding states and enables the interaction of genes with enhancers, promoters, non-coding RNAs and transcription factors. Eukaryotic genomes are typically organised into compartments, domains and loops.

Typically, eukaryotic genomes are segregated into the A and B compartment. The A compartment is associated with open chromatin and active genetic elements, while the B compartment is comprised of closed chromatin and inactive genes.[75] Within the same compartment pairs of loci will interact at a higher frequency and more consistently than pairs of loci in two different compartments.[75] These compartments are enriched with distinct histone modifications that are representative of the observed transcriptional state.[76] The open and closed domains occupy different spatial compartments in the nucleus, the A compartment is often associated with nuclear bodies whereas the B compartment is preferentially associated with the nuclear envelope or nucleolus.[77,78]

Regions that interact with each other more frequently segregate into sub-megabase scale domains, known as topologically associating domains (TADs). TADs are regions of the genome with a high degree of self-interaction and are formed by active extrusion of chromatin loops by cohesin.[79] They are generally conserved between different cell types and can switch between compartments A and B in a cell-type specific manner.[80] The partitioning of the genome into TADs correlates with many of the regulatory processes of the linear genome such as co-ordinated gene expression, DNA replication and deposition of histone modification marks alongside the facilitation of enhancer-promoter interaction.[81,82] In mammals, TADs are a prominent feature that can range in size from 40 kb to 3 Mb[83], whereas in plants they are not common. Larger plant genomes, such as maize, rice and tomato, have TAD-like features which in rice have been shown to cover approximately 25% of the genome and be around 45 kb in size.[84] Chromatin architectural proteins, often the CTCF-cohesin complex, demarcate the TAD boundaries and promote their formation by loop extrusion.[81,85–87] Importantly, the loss of the CTCF-cohesin complex or of the cohesion loading factor eliminate TADs, however A/B compartments remain.[88]

Plant genomes do not contain a CTCF encoding gene, thus do not contain canonical TADs.[89] A/B compartments and chromatin-interacting domains have however been identified in many plant species.[90] In rice genomes, TAD-like domains have been detected and TAD boundaries regions with properties similar to animal TAD boundaries have been identified, however they are not a prominent feature of the *A. thaliana* genome.[84,91] A unique feature of plant genomes is that they form compartment domains that interact with each other on an inter- and intrachromosomal level.[92] These domains often contain either active or inactive genes which are defined by the sharp change in the corresponding histone modification mark, for example repressive domains contain inactive genes and histone marks such as H3K27me3 which shows a sharp depletion at the border.[92]

Chromatin loops bring stretches of genomic sequences that are separated by intervening sequences in close proximity to each other.[93] These loops often aid the regulation of genes by bringing enhancer regions in closer proximity to its target gene.[94] The locus control region (LCR), which will be explained in more detail later, is a well-known example of a chromatin loop. Gene-loops, a prominent feature of the *A. thaliana* genome,[95] are another type of chromatin loop in which the transcription termination site of a gene loops upstream to be in closer contact with the gene promoter, thus reinforcing the directionality of RNA synthesis.[96] In *A. thaliana,* the FLC gene forms a gene-loop which is disrupted in the early phase of vernalization.[97] Loop formation relies on RNAs and multivalent proteins such as the mediator protein complex. In *A. thaliana*, the mediator complex enhances chromatin looping between jasmonic acid (JA) enhancers and their promoters to trigger the jasmonic acid signalling pathway.[98]

On a smaller scale, beyond loops, dynamic nucleosome-nucleosome interactions have been described. These may organise cell-type specifically into linear or clutched arrangements. [99]

1.2 Importance of gene order

There is increasing evidence of a non-random gene order in eukaryotes and that the position of genes along chromosomes serves a function. For example, neighbouring genes often show similar expression profiles[100] and genes can share regulatory elements such as bidirectional promoters.[101] A genome wide study in eukaryotes analysed cell-cycle-dependent expression patterns, and found that a quarter of genes which are induced in the same phase of the cell cycle were directly adjacent to one another[102], thus providing evidence of a non-random gene order. Furthermore, a random gene order would give rise to random transcription levels of genes across the eukaryotic genome, however is has been found that genes neighbouring highly transcribed genes are more likely to have a higher transcription level and vice versa[100]. These neighbouring genes often share promoters and transcriptional regulatory signals, resulting in them being co-expressed[103]. Gene clusters provide good evidence and a model for non-random gene order. These clusters are often functionally related and associated with metabolism, development and immunity.[104]

1.3 Gene clusters

1.3.1 What are gene cluster?

The co-localisation of functionally related genes into operons is a defining feature of bacterial genomes. In contrast, the organisation of functionally related genes in eukaryotes has historically been considered random. However, large scale genomics and transcriptomics experiments have shown that gene order is not random. Furthermore, it was found that genes neighbouring highly transcribed genes are more likely to have a higher transcription level and vice versa[100]. Similar to prokaryotic operons, functionally related genes have also been found to be clustered within the eukaryotic genome which are often co-regulated, and it has been suggested that co-ordinated expression of genes is favoured by selection[100]. Different types of gene clusters can be identified in eukaryotic genomes, such as arrays of homologous or non-homologous genes and clusters of functionally related genes. Although they may appear similar to operons, clustered genes in eukaryotes differ in that they transcribe each gene within the cluster into monocistronic pre-messenger RNA (mRNA), with rare exceptions in some nematodes[105]. Gene clusters encode for vital functions such as immunity in animals and plants, for example the major histocompatibility complex and resistance factors, and for embryonic patterning, such as *Hox* transcription factors.[106] Secondary metabolic gene clusters are prominent in fungal and plant genomes and play important roles in the ability of these organisms to interact with their environment.[107]

An example of clusters of homologous genes is the *Hox* (homeobox) gene cluster that encodes a major class of transcription factors required during embryonic development.[108] The precise activation of the *Hox* genes allows for the specific co-ordinated expression of the anterior to posterior body axis, in bilaterian animals they are responsible for the patterning of the main body axis.[109] Mammals have four *Hox* clusters which appear to be the same across tetrapods.[110] Following transcription, the chromatin at the specific loci is condensed, showing a rapid transition in chromatin state.[111] It was shown that co-ordinate activation of the *Hox* genes is associated with H3K4me3 and transcriptional repression with H3K27me3.[112] Two distinctive TADs have been identified in the *HoxD* cluster, which is involved in vertebrate limb development, the T-TAD and C-TAD (telomere-proximal and centromere-proximal).[113] Interestingly, this cluster goes through two subsequent waves of transcription, the centre of the cluster is activated first followed by the flanking genes.[114] The change from the activation of central cluster genes to outer genes is associated with the remodelling of the cluster specific TAD structure.[113]

Metabolic gene clusters, or biosynthetic gene clusters, are widespread in the genomes of plants and fungi and encode for a significant proportion of their metabolic capacity.[115] These gene clusters provide a great insight into the evolutionary processes as these genes have been brought together by genome reorganisation, whereas *Hox* gene clusters have arisen through gene duplication.[106,110] Fungal and plant metabolic gene clusters are primarily associated with secondary metabolism. Secondary metabolic pathways are described as being non-essential but with important ecological functions, such as initiating an immune response to a pathogen or acting as an attractant to pollinators[116]. The immense diversity of secondary metabolites provides a rich source of drugs and antibiotics, as almost 25% of modern medicine is derived from products naturally occurring in plants[117,118]. Secondary metabolic gene clusters encode the different enzymatic steps of a biosynthetic pathway[115] that metabolises primary metabolites into secondary metabolites, as shown in ***Figure 2***. The enzymes encoded within secondary metabolic gene clusters that divert primary metabolites into the metabolic pathway are often called scaffolding enzymes. The enzymes that modify the metabolic scaffold (such as methyltransferases and acyltransferases[115]) are called tailoring enzymes . The overexpression or silencing of metabolic cluster genes is often detrimental to the organism due to the accumulation of toxic pathway intermediates[106,115]. Plant metabolic gene clusters consists of three or more neighbouring genes and range from approximately 35kb to several hundred kb in size[119].

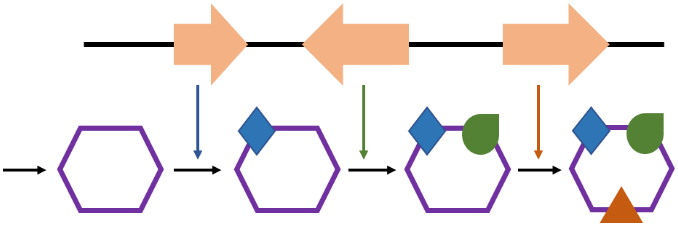


Figure 2: Secondary metabolic gene cluster. *Secondary metabolic gene clusters encode enzymes that catalyse each step of a metabolic pathway to further metabolise primary metabolites into secondary metabolites.*

Fungi produce a large array of diverse secondary metabolites, with a wide range of functions such as metabolites for interaction with other organisms or defense.[120,121] Interestingly, within the fungal genome secondary metabolic pathways are typically arranged into clusters adjacent to each other, whereas genes required for the synthesis of primary metabolites are randomly dispersed throughout the genome.[121] The most commonly known product of biosynthetic gene clusters in fungus is the β-lactam antibiotic penicillin, which consists of three neighbouring genes arranged the same order in *Penicillium chrysogenum, Aspergillus nidulans* and *Penicillium nalgiovense.*[122,123] These biosynthetic gene clusters in fungi are co-regulated, the whole cluster can be actively transcribed in response to an environmental stimuli; for example in response to bacteria[124]. A wide range of transcription factors contribute to the regulation of biosynthetic gene clusters in fungus such as PacC, an important regulator for pH.[125] Co-regulation within fungal clusters is suggested to be partly due to the chromatin structure.[126] The first example for the involvement of chromatin modifying enzymes in fungal gene cluster regulation showed that the deletion of an eraser of histone acetylation led to an increase in two secondary metabolites, fumitremorgin B and pseurotin, but a reduced amount of gliotoxin in *Aspergillus fumigatus*.[127] Studies in *Aspergillus nidulans* show that clusters often share common chromatin features but have distinctive features in terms of combination of histone modification marks, which are required for the activation and repression of transcription.[126]

1.3.2 Mechanisms of co-regulation in gene clusters

A typical feature of clustered genes is their co-ordinated transcription. Such co-ordination of transcription may arise from a variety of mechanisms as shown in ***Figure 3***. Bidirectional promoters, for example, where one promoter element drives the expression of two divergently orientated genes, enables tight co-ordination of two neighbouring genes.[128] The delineation of an entire gene cluster or its segments by unique histone modifications may enable the formation of a shared chromatin environment surrounding adjacent genes. This may establish a distinct platform for transcriptional co-ordination. Local 3D domains encompassing gene clusters may separate them from neighbouring chromatin environment and promote the formation of efficient promoter-enhancer contacts. Such DNA loops allow distant regulatory sites, for example enhancers, to come in close proximity to its target promoters. Enhancer DNA elements act independent of orientation and location and may activate target genes from 1 million bps away.[129,130] They are described as clusters of DNA sequences that can recruit various transcription factors. These transcription factors then interact with components of the Mediator complex or TFIID (Transcription factor II D) by looping out the intervening sequences to facilitate the engagement of the RNA polymerase II with the promoter region. They also recruit chromatin remodelling complexes and histone modifying enzymes which in turn will increase the accessibility of other proteins to the DNA[131,132]. The patterning of enhancers, their DNA methylation status and specific binding of transcription factors is suggested to control enhancer activity and as such target gene expression.

The facilitated co- positioning of cluster genes within the nucleus, for example at the centre, near the nucleolus or the nuclear envelope, may also play a role in the co-ordinate regulation of clusters.[133]

Super enhancers are described as being *cis*-regulatory elements with unusually strong enrichment of Mediator and polymerases.[134] A prominent example of a super-enhancer and the strong effects of it when deleted is the locus control regions (LCRs) of the human β-globin locus.[135] When this gene segment along with a 1.5kb promoter region was expressed in erythroleukemia cell lines, the genes were only expressed in a small proportion of cells and expression was significantly low.[136] Loss of β-globin gene cluster expression in patients suffering from β-thalassemia, an inherited blood disorder, was identified to be caused by the deletion of the β-globin LCR. Loss of the LCR causes chromatin compaction and prevents polymerase binding.[137,138]

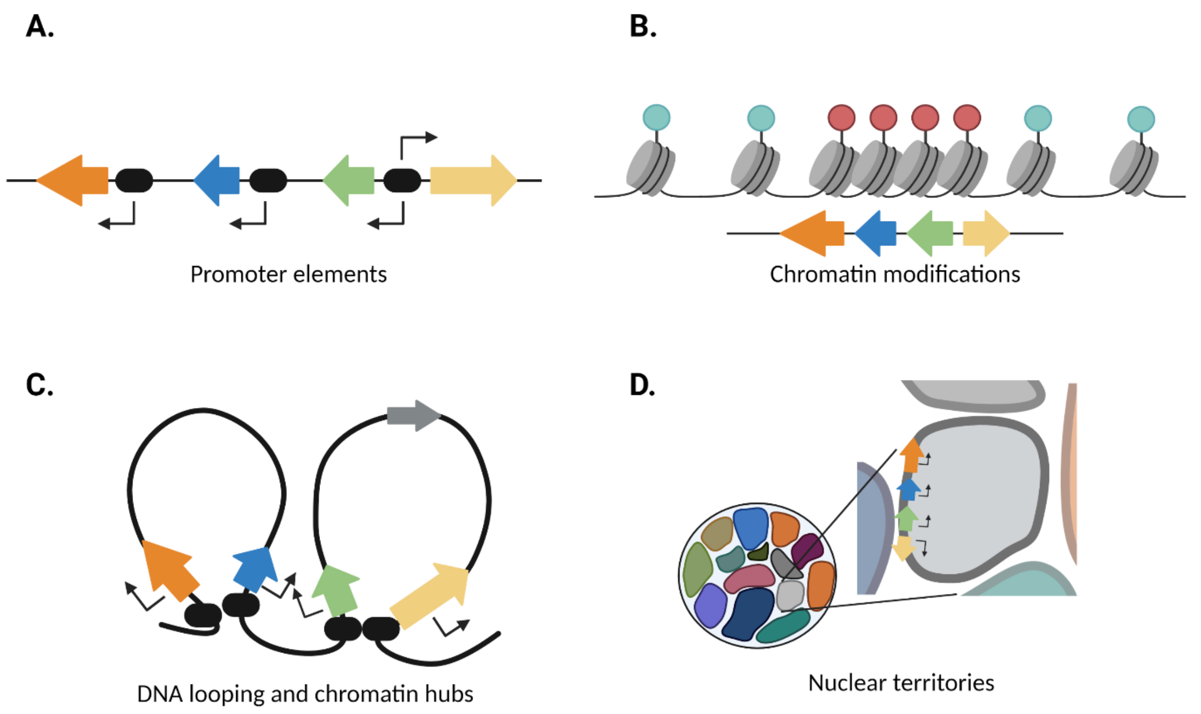


Figure 3: Depiction of the potential regulatory factors of gene clusters. *A. Conserved promoter motifs, black circles, are located upstream of each cluster gene. B. Chromatin modifications facilitate the formation of open chromatin (red circles) and closed chromatin conformation (green circles). Open chromatin facilitates transcription and closed chromatin suppresses it. C. Chromatin hubs form in the 3D chromatin structure. Intervening DNA and genes (grey arrow) are looping out, bringing cluster genes in a closer proximity to regulatory elements. Genes close (coloured arrow) to the regulatory elements are co-expressed (black circles). D. Gene clusters may locate to discrete nuclear territories with characteristic chromosomal conformations.*

1.3.3 Metabolic gene clusters in *Arabidopsis thaliana*

More than twenty secondary metabolic gene clusters have been identified in plants thus far.[139] *A. thaliana* contains four secondary metabolic gene clusters, thalianol, marneral, arabidiol/baruol and tirucalladienol, all of which are related to triterpene metabolism.[140,141]

Triterpenes are a large and structurally diverse class of natural products with over 23,000 known triterpene structures derived from natural sources.[142] The plant kingdom has the highest amount of structurally diverse triterpenes and are often synthesised in response to developmental or environmental cues.[143,144] The biological function of most triterpenes are unknown, however triterpenes have been shown to play a critical role in plant and root growth[145,146], while others confer a defence response.[147,148] In plants, the most common precursor for triterpenes is 2,3-oxidosqualene, which is converted into a variety of triterpene scaffolds by oxidosqualene cyclases (OSCs).[144] Tailoring enzymes, such as P450 monooxygenases (P450s), catalyse diverse reactions to add functional groups to the scaffold.[149] In *A. thaliana* 13 OSCs have been identified, some of which are encoded in gene clusters.[150]

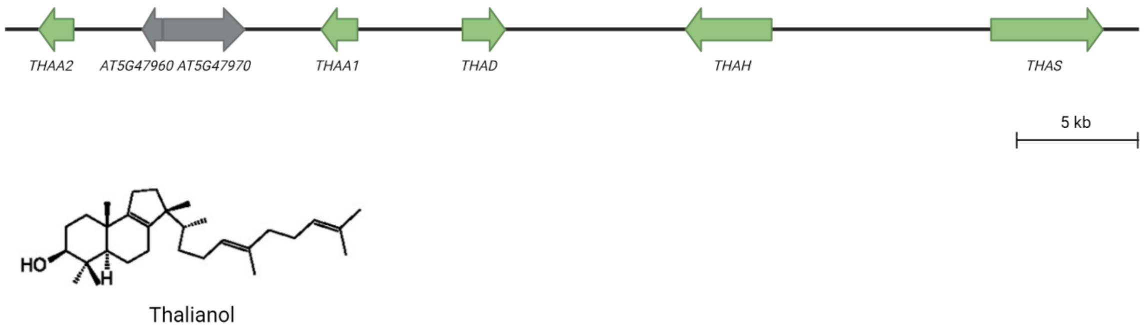


Figure 4: Thalianol gene cluster and its product, thalianol. *The organisation of the thalianol gene cluster (to scale), including the five cluster genes (green arrows) and the two intervening genes (grey arrows). The thalianol compound depicted.*

The marneral cluster consists of three clustered genes that are co-expressed, *MRN1*, *CYP71A16* and *CYP705A12*.[151] *MRN1* encodes a synthase that converts 2-3-oxidosqualene to marneral.[152] Similar to the thalianol cluster, the marneral cluster is exclusively expressed in roots.[153] Overexpression of the marneral cluster leads to a pronounced dwarf phenotype of the plant, indicating that tight co-regulation of the clusters is important for plant fitness.[151]

Five genes make up the tirucalladienol gene cluster which spans about 47 kb, which produces tirucalla-7,24-dien-3β-ol.[154] *PEN3* encodes a triterpene synthase, *AT5G36130* and *CYP716A2* encode a single cytochrome protein with an unknown function and *CYP716A1* was shown to encode an enzyme involved in the hydroxylation of tirucalla-7,24-dien-3β-ol when co-expressed with *PEN3.*[154]

The arabidiol/baruol metabolic gene cluster produces arabidiol as well as downstream toxic breakdown products that protect the *A. thaliana* roots from *Pythium irregulare*, and oomycete which causes root rot.[155] The volatile organic compound (*E*)-4,8-dimethyl-1,3,7-nonatriene (DMNT) and the non-volatile apo-arabidiol is produced in response to *Pythium* infection. This cluster is made up of eleven genes, two OSCs, seven CYPs and two acetyltransferases.[156] The OSC, *PEN1*, is involved in the degradation of arabidiol in response to jasmonate (JA) treatment to produce DMNT.[155] Whereas *BARS1* produces baruol as its main product.[141]

The thalianol cluster has an important role in ecological interactions by shaping the root microbiota to tailor it to its own needs for plant growth and health, which is reflected in its high conservation across different *Arabidopsis* accessions.[151,157] The accumulation of thalianol and its derivatives is partly regulated by the phytohormone JA.[158] The thalianol biosynthetic gene cluster consists of five adjacent cluster genes distributed over a region of 45 kb, shown in ***Figure 4***. Originally, it was reported that the thalianol cluster consists of four physically adjacent co-expressed genes[106], *THAS, THAH, THAO,* and *THAA1*. This set of genes is conserved in both *A. thaliana* and *A. lyrata*. However, recent findings have shown that a nearby acyltransferase gene, *THAA2*, as well as two non-linked genes, *THAR1* and *THAR2,* are required for completion of the metabolic pathway.[159] The four core genes of the thalianol cluster encode an OSC, two CYP450s and a BAHD family acyltransferase. All of these enzymes are exclusively active in the thalianol metabolic pathway. The four core cluster genes are highly co-expressed in the root and silenced in all other organs of *A. thaliana*.[106] The overexpression of *THAS* results in thalianol accumulation in leaves and dwarfing of the plant.[106] The BAHD acyltransferase gene *THAA2*, separated by two intervening genes from the core cluster*,* is co-expressed with the other four core cluster genes. However, the encoded enzyme is active in multiple triterpene pathways, amongst them the thalianol and arabidiol pathways.[157] A genome-wide study analysing the polymorphisms in *A. thaliana* showed that the thalianol gene cluster was one of the most conserved regions of the genome.[106,153,160]

The tight co-regulation of the thalianol gene cluster has been investigated in multiple studies, providing extensive prior data and information on its regulatory properties[106,153,155,161–163], making it an interesting candidate for gene regulation and gene order analysis. Furthermore, the opposing and defined change in thalianol gene cluster expression between organs provides an informative and clear experimental system for the investigation of transcriptional co-regulation in eukaryotic gene clusters. Moreover, working with the plant model system *A. thaliana* offers multiple advantages.

A. thaliana is a key model organism in plant science and genetics due to its short life cycle which allows for rapid cultivation of large numbers of plants in a controlled environment. It has a relatively small genome of around 135 Mb in size. 27,655 coding genes and 6,607 small or long non-coding genes are distributed across 5 chromosomes (TAIR10.1).[164] Recently, the first complete telomere to telomere *A. thaliana* genome assembly has been published. Rapid and reproducible methodologies for transformation are available, enabling efficient genome manipulation.[165] Available resources include genome-wide mutant libraries and more than 1000 sequenced ecotypes with world-wide distribution.[166] This plethora of existing knowledge and resources makes *A. thaliana* an ideal organism to work with to investigate genome organisation and its regulation.

1.3.4 Regulation of metabolic gene clusters

1.3.4.1 Key histone modification marks

It is important to study the regulation of gene clusters to improve our understanding of location-specific effects on gene expression. This is crucial for further developments in genetic manipulations, biotechnology and safe gene therapy. Advancing our understanding of gene clustering in specialised metabolism may also provide valuable new insight for the industrial production of higher-value metabolites. The model clusters in this study, the thalianol, marneral and arabidiol/baruol clusters, are exclusively expressed within root tissues and tightly repressed in the aerial parts of the plant.[140] Previously, two post-translational histone modification marks, H2A.Z and H3K27me3, have been identified as being key regulators of these clusters.[161] A genome-wide analysis[13] of H3K27me3 histone modification marks in *A. thaliana* showed a pronounced increase of the mark at all three clusters in both genic and intergenic regions.[161] Importantly, the immediate flanking genes were not enriched for H3K27me3 marks.[161] The enrichment of H3K27me3 was also not detected at non-clustered biosynthetic pathway genes.[167] Deposition of H3K27me3 at the three gene clusters is negatively correlated with gene expression.[167] *A. thaliana* mutant plants with lowered H3K27me3 levels show increased cluster expression.[163] Interestingly, strong H3K27me3 labelling has also been shown for the rice momilactone and phytocassane as well as the maize DIMBOA gene clusters. In contrast to H3K27me3 marks, the deposition of the histone variant H2A.Z at gene clusters is positively associated with transcription. H2A.Z levels are elevated in gene cluster expressing tissues compared to non-expressing tissues. H2A.Z mutant lines show strongly reduced cluster expression levels coinciding with reduced nucleosome accessibility across the cluster region[161].

1.3.4.2 Cluster 3D chromatin structure

While the importance of cluster associated chromatin marks had been identified for both fungal and plant specialised metabolic gene clusters over the past 15 years, the 3D organisation of metabolic gene clusters remained elusive. However, recent investigations have shown that distinct 3D domains are established at metabolic gene clusters in plants. These domains separate the cluster chromatin environment from the neighbouring chromatin environment. Indeed, at the thalianol gene cluster a conformational switch in domain structure can be identified in expressing and non-expressing tissues.[163] When the cluster is inactive a chromosomal loop domain is formed encompassing the cluster whereas a highly compact locus domain conformation is established during expression. The compact domain at expressing clusters consists of two layers that have variable strength, the smaller layer that has a higher intensity starts at the peripheral cluster gene, *THAA2*, and extends to the final cluster gene, *THAS*, covering about 50 kb in size whereas the larger less intense domain also includes the non-cluster upstream gene *At5g47910*. In leaves the interacting region is larger than in the roots, shifting downstream covering the cluster as well as a region downstream, however the intensity of the interaction is reduced.[140] The silencing interactive domain is about 110 kb in size and starts at *THAA2* and ends at the non-cluster genes *At5g48150* and *At5g48160* which are not co-expressed with the cluster. The silencing domain structure falls into a B compartment while the expressing domain is associated with an A compartment.[140] Similar results were also found for the marneral and arabidiol/baruol gene clusters, indicating the importance of the 3D architecture for the controlled transcriptional regulation of these biosynthetic gene clusters. The marneral cluster forms a large 300 kb interactive domain when the cluster is inactive in the leaves and a more local interactive domain with intense interaction when active in the roots. A single local interactive domain which encompasses the arabidiol/baruol cluster in both the leaves and the roots, however the interaction is stronger in the leaves compared to the roots when the cluster is inactive.[140]

When silenced, the thalianol, marneral and arabidiol/baruol clusters were also found to be engaging in interactions with the pericentromeric parts of the chromosome. In expressing tissues these interactions were lost. In accompanying FISH analysis, cluster visualisation showed that the thalianol clusters is located at the nuclear envelope in non-expressing tissues.[140]

1.3.5 Evolution of metabolic gene clusters

The evolution of metabolic gene cluster is under intense investigation and debated in the scientific community. Due to the similarities between prokaryotic operons and eukaryotic gene clusters, one proposed evolutionary path has been horizontal gene transfer from microbes. Indeed, some fungal biosynthetic gene clusters have been found to be acquired by horizontal gene transfer, such as the epipolythiodioxopiperazine (ETP) gene cluster implicated in animal and plant disease.[168] However, there is currently no scientific evidence backing a microbe to plant transfer of metabolic gene clusters[161]. Interestingly, however, recent work has suggested a plant-to-plant transfer of the DIMBOA gene cluster.[169]

The current studies suggest that metabolic gene clusters in plants have arisen by gene duplications and re-organisation of the host genome. Thereby, most gene clusters are thought to have evolved independently of each other.[106] The genes encoding the specialised metabolic enzymes often share homology with primary or other specialised metabolic genes and it is therefore thought that they have either arisen from a common ancestor, or evolved by gene duplication and acquired a new function.[170] The thalianol gene cluster is thought to have formed around the progenitor of the lineage-specific OSC clade II using phylogenetic analysis, with sequential rearrangements, duplications and gene loss. The similarities between the P450 enzymes in the cluster belonging to the same clade indicates that they may have formed via segmental duplication (SD) of an ancestral cluster region[153], however the different orientations of the genes indicates a gene-inversion would have occurred following the SD. Also, both the marneral and thalianol clusters contain an excess number of transposable elements (TEs) to gene density ratio, which makes them potential key acceptor sites for SDs. [106]

The raison d’etre of gene clustering in specialised metabolism is unknown. However, a set of evolutionary advantages have been proposed that may explain the clustering phenomenon. For example, co-location of pathway genes may increase the likelihood of co-inheritance of all genes and thus reduce the risk of loss of individual genes only.[171] Similarly, the co-localisation of genes may support their co-regulation due to shared regulatory elements, epigenetic signatures, and chromatin domains. Both co-inheritance and co-regulation may be essential in the orderly completion of the metabolic pathway and an avoidance of an accumulation of potentially toxic intermediates.[172] In *A. thaliana* for example, it has been shown that overexpression of *THAS* and *MRN1* leads to dwarfing of the respective mutant plants.[106]

1.4 Research scope and aims

The importance of gene order in the regulation of genes is exemplified by the co-regulation of neighbouring genes in gene clusters. Pioneering research on model clusters in humans, animals and yeasts has shown that these clusters are often regulated on chromatin-level and by DNA sequence elements distant from the cluster. As such, they provide evidence that structural integrity of the genome is important for gene activity and serve as reminder that genome editing may have unexpected consequences to non-target genes.

Here, we use metabolic gene clusters in plants as model systems to investigate the co-ordinate regulation of neighbouring genes and the importance of gene positioning for gene activity in plants.

We aim to shed light on the importance of sequence integrity of gene clusters in plants by disrupting the thalianol gene cluster and repositioning its genes. We will use genome editing technology, mutant libraries and quantitative transcript level analyses to investigate the relationship between cluster integrity and expression.

Furthermore, we aim to characterise the code of chromatin modifications associated with metabolic gene clusters and identify novel regulators of chromatin-level cluster regulation. Here, we will use bioinformatic data analyses, epigenetic mutant lines, chromatin immunoprecipitation and quantitative transcript level analyses for our investigations.

Finally, we aim to analyse sequence and expression conservation of metabolic gene clusters within a species. We will apply bioinformatic sequence comparison as well as transcript level quantification for these analyses.

2 Materials and Methods

2.1 Materials

2.1.1 Plant growth and conditions

2.1.1.1 Plant material used for genotyping

Plants used were in the *A. thaliana* Columbia (Col-0) background, also referred to as wild-type (WT). *A. thaliana* T-DNA insertion lines were obtained from the Nottingham Arabidopsis Stock Centre (NASC).[173] For genotyping, these plants were sown into F2 fine sand compost (Levingtons) and kept in an environment-controlled growth room (Sanyo) for up to four weeks. Fresh leaf samples were collected in 2ml microcentrifuge screw cap tubes containing four to five zirconia beads (2.0mm diameter, BioSpec) and immediately submerged into liquid nitrogen. Samples were either stored in a -80°C freezer or used for DNA isolation. Plants that tested homozygous for the insert during genotyping were left to complete their growth cycle and seeds were collected at around eight weeks. The growth rooms were kept at 16-hour day length and a temperature of 22°C and an 8-hour night length at 18°C, and at a constant humidity of 40%.

2.1.1.2 Plant material used for qPCR and ChIP

For qPCR and ChIP *A. thaliana* seeds (approximately 0.01g per plate) were sterilised and grown on MS media (***Table 1***) plates. Seeds were pipetted onto MS media plates in two rows and left for fifteen minutes or until dry. The plates were taped with 3M Micropore tape and placed in growth rooms vertically until harvest. For qPCR, seeds were grown on Sterilin 100 mm Square Petri Dishes to make one biological replicate. On day seven, root and leaf material were harvested separately using a razor blade. MiraCloth dried material, was transferred into 2 mL microcentrifuge tubes and immediately submerged in liquid nitrogen. Samples were either stored in a -80°C freezer or directly used for RNA isolation. For ChIP, plants were grown on Gosselin Square Petri Dish, on day 10 plant material was harvested and directly used in a ChIP experiment. Plants were grown vertically on plates in the growth room which kept a 16-hour day length at a temperature of 22°C and an 8-hour night length at 18°C, at a constant humidity of 40%.

Table 1: Ingredients to make 1000 mL MS media

Component	Volume per 1000 mL
Murashige & Skoog Medium	4.3g
Sucrose	10.0g
Phytagel agar	5.0g

2.1.1.3 Plants used for floral dip

A. thaliana seeds, kindly donated by Miki et al[174], that express a Cas9 driven by a DD45 promoter were sown onto F2 fine sand compost and kept in an environment-controlled growth room (Sanyo) which kept a 16-hour day length at a temperature of 22^{o}C and an 8-hour night length at 18^{o}C, at a constant humidity of 40%. The primary bolts were clipped, and the plants were used in a floral dip when the secondary bolts reached 10 to 20 cm in size and had flowered. Following floral dip, the plants completed their growth cycle and bagged when the siliques started to brown. After 10 days the plants were no longer watered and left for another 7 days at which seeds were harvested.

2.2 Methods

2.2.1 Polymerase Chain Reactions

The following protocols for polymerase chain reactions (PCR) were used:

2.2.1.1 Fragment amplification

For construction of insertion, deletion and random integration constructs a PCR protocol using the KAPA polymerase was used. KAPA PCRs allow the integration of uracil into PCR products. KAPA PCR products were used for USER reactions and fusion PCRs. Bacterial artificial chromosomal DNA was used as a template for PCR amplification of DNA fragments in a 1:100 ratio, ***Table 2,*** sequences shown in ***Supplementary 7.1***. PCR primer sequences are shown in ***Table 3***. PCR reactions were run on a thermocycler (Applied Biosystems MiniAmp Thermal Cycler) using the reagents in ***Table 4*** and programmes in ***Table 5***.

Table 2: Templates used for DNA amplification

Artificial Chromosomal DNA template number	Use for:
2020	*THAS* insert homologous arms
2813	*MRN* insert homologous arms
106	All *PP2AA3* regions
2013	*THAS* regulatory elements

Table 3: PCR primers. *Primer sequence and associated fragments, fragment length and construct are shown. The underlining indicates the USER reaction site.*

Use in construct	Fragment	Fragment size (bp)	Primers
T1	*THAS* promoter	3416	GGCTTAUCCACTCTCTCCTGGTAGATG AGGTTGCUCTAAGTTTACTTGGACAAGG
T1	*THAS* terminator	1349	ATGTCACAUCTATCTTCTTCACCGT GGTTTAUTTTCAGTGTTCAGCCGTGAACC
P1	*PP2AA3* promoter	2031	GGCTTAUCTAACAACAATAATACAACAGATTG AGCTCCUCGCCCTTGCTCACCATGACAAAGCTGACCATATATTG
P1	*PP2AA3* terminator	861	AGAACAUGATACGGCCATGCTTG GGTTTAUGGTTACGTAACTATTAAGCCAGAATC
T1	eGFP/Gus	2532	AGCAACCUTACAAAATGGTGAGCAAGGGCGAGGAGCTG ATGTGACAUGGTCATTGTTTGCCTCCCTGCTGC
P1			AGGAGCUGTTCACCGGGGTGGTGCC ATGTTCUCCACAATCATTGTTTGCCTCCCTGCTGC
T2B	*MRN* LA	900	ACATGGCUGGACAACATAGAAATCATTTATG ATGCACAUAAGATTTAATCGATGTG
T2B	*MRN* RA	950	AATGGTAUAATGATGTAGCACTTTAGG GGTTTAUAGTAACACTTAGTTTAGTTTTTACCC
T2C	*THAS* LA	913	ACATGGCUGTCTGACCCGTTGGGATGTTAATTTTC ACTCGTUAGTTACAACGGTATTGTATG
T2C	*THAS* RA	950	ACGGCAAUAACCTAAAACAATAACC GGTTTAUCCAATAACAACAGTAGAAGCC
T2A	*PP2AA3* LA	945	ACATGGCUGTTCGTCAACAAAAGCAACTTCG AGCAGGUGTGTTGTATGTAATCATTATAATAG
T2A	*PP2AA3* RA	928	ATTCGGUAAGAGAGATCTTTATTTTC GGTTTAUTGTTTGAAGTGCGTGGATAATG
T2/P2	PU6-26::TU6-26	358	GGCTTAUCATCTTCATTCTTAAGATATG AGCCATGUACCCCAGAAATTGAACGCCGAAG

Table 4: Volumes for reagents to use in one reaction for a KAPA PCR protocol

Reagent	Volume (per 1 reaction, µL)
KAPA enzyme mix	20
Forward Primer (10nm)	0.8
Reverse primer (10nm)	0.8
Template DNA	0.5
Nuclease free water	17.9
Total volume	40

Table 5. KAPA PCR cycle conditions. *The denaturation, annealing and extension cycles were repeated under indicated conditions.*

KAPA standard	Initial denaturation	Denaturation	Annealing	Extension	Final extension	Number of cycles
Temperature (ºC)	95	98	55	72	72	30
Time length	3 minutes	20 seconds	15 seconds	1 minute per 1kb	2 minutes 30 seconds	

2.2.1.2 Fusion PCR

Fusion PCRs were used to fuse fragments of DNA together. The KAPA protocol was used as described in **2.2.1.1**. 0.25 µL of each fragment in equimolar concentrations was used as template. The primers used are shown in ***Table 3***.

2.2.1.3 Colony PCR

Colony PCRs were performed to screen *E. coli* transformants for correct inserts. PCRs were carried out with the DreamTaq polymerase using the conditions shown in ***Tables 6***. All PCR reactions were run on a thermocycler (Applied Biosystems MiniAmp Thermal Cycler) using the programme indicated in ***Table 7***.

Table 6: Volumes for reagents used in one reaction for a DreamTaq PCR protocol.

Reagent	Volume (per 1 reaction, µL)
DreamTaq Buffer	2.5
Forward Primer (10nm)	0.5
Reverse primer (10nm)	0.5
Template DNA	1
DreamTaq polymerase	0.075
Nuclease free water	12.925
Total volume	25

Table 7: DreamTaq PCR cycle conditions. *The denaturation, annealing and extension cycles were repeated under indicated conditions.*

DreamTaq	Initial denaturation	Denaturation	Annealing	Extension	Final extension	Number of cycles
Temperature (°C)	94	94	55	72	72	35
Time length	2 minutes	2 minutes	30 seconds	1 minute per 1kb	1 minute 30 seconds	

2.2.1.4 Genotyping PCR

Genotyping PCRs were carried out to screen plant material for insertion of DNA constructs using the Phusion PCR protocol. PCR using Phusion polymerase using the reagent volumes in ***Table 8*** and ran on a thermocycler using conditions described in ***Table 9***.

Table 8: Volumes for reagents used in one reaction for a Phusion PCR protocol.

Reagent	Volume (per 1 reaction, µL)
Phusion Buffer	4
Forward Primer (10nm)	0.5
Reverse primer (10nm)	0.5
dNTP mix (10nm)	0.5
Template DNA	1.5
Phusion polymerase	0.15
Nuclease free water	12.85
Total volume	20

Table 9: Phusion PCR cycle conditions. *The denaturation, annealing and extension cycles were repeated under conditions as described.*

Phusion	Initial denaturation	Denaturation	Annealing	Extension	Final extension	Number of cycles
Temperature (ºC)	98	98	55	72	72	25
Time length (seconds)	30	10	30	30 per 1kb	10	

2.2.1.5 Quantitative PCR

For quantification of gene expression, quantitative PCRs (qPCR) were carried out. Synthesised cDNA was diluted to a ratio of 1:6 with nuclease free water. 1 µL of diluted cDNA was first added to the well in a 96-well plate (Applied Biosystems MicroAmp Fast Optical 96-Well Reaction Plate) followed by the master mix described in ***Table 10***, then sealed with Microseal 'B' PCR Plate Sealing Film (BioRad). To normalise the data the *PP2AA3* (*At1g13320*) gene was used in each qPCR, primers shown in ***Supplementary 7.3***. The qPCR was run in a qPCR machine (ARIA), using the cycle shown in ***Table 11***. All qPCR experiments were performed using three biological replicates, unless otherwise indicated, and three technical replicates. Data analysis was performed using a ΔΔCT method.

Table 10: Volumes for reagents used in one reaction for a MyTaq qPCR protocol.

Reagent	Volume (per 1 reaction, µL)
MyTaq HS-mix	10
Forward Primer (10nm)	0.4
Reverse primer (10nm)	0.4
EvaGreen dye	1
Nuclease free water	6.7
Total volume	18.5

Table 11: MyTaq quantitative PCR cycle conditions. *The denaturation, annealing and extension cycles were repeated under indicated conditions.*

MyTaq	Initial denaturation	Denaturation	Annealing	Melting curve			Number of Cycles
Temperature (^{o}C)	95	95	62	95	65	95	40
Time length (seconds)	120	5 minutes	30	30	30	30	

2.2.1.6 PCR clean-up

PCR products required for downstream applications were purified using the Monarch PCR & DNA Cleanup Kit (5µg, New England BioLabs) or by QIAquick Gel Extraction Kit (QIAGEN) according to the manufacturer's instructions. Recommended procedures for salt-sensitive applications and heat treatment were included. DNA concentrations of purified products were determined by using a spectrophotometer (Thermo Scientific NanoDrop One Microvolume UV-Vis Spectrophotometers) and 1.2 µL of each sample. The Monarch PCR & DNA Cleanup Kit elution buffer was used as a blank.

2.2.2 Agarose gel electrophoresis

2.2.2.1 DNA electrophoresis

Electrophoresis of DNA samples was carried out in 1% agarose (Bioline) gels for approximately 30 minutes at 100V. 1x TAE was used as running buffer. The Thermo Scientific GeneRuler 1 kb DNA Ladder or 100bp ladder (***Figure 5***) were applied to gels for DNA size estimation. Gels were stained with ethidium bromide (1µl of ethidium bromide per 25ml of gel) and visualised with an UV transilluminator.

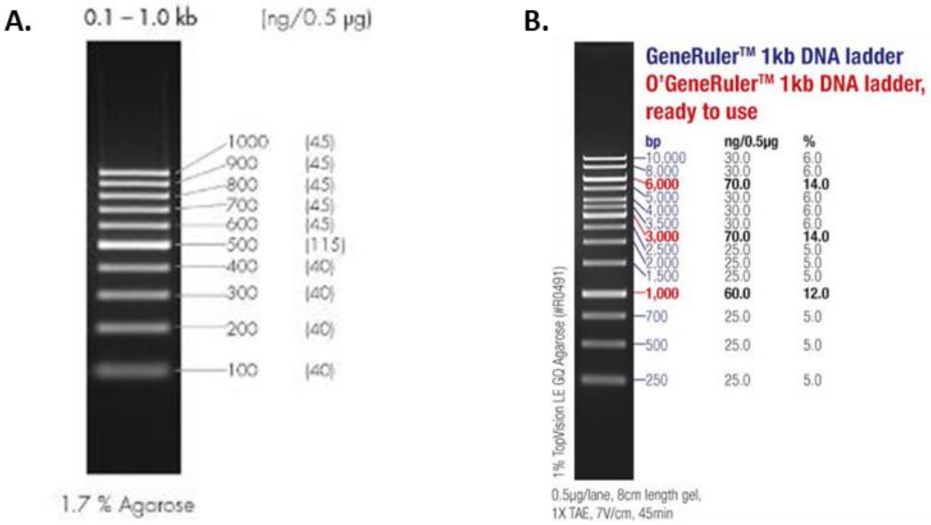


Figure 5: 100 bp and 1 kb DNA ladder. *DNA ladder used in AGE to identify size of PCR products.*

2.2.3 Isolation of plant genomic DNA

Plant material was collected as described in **2.1.1.1**. The sample was homogenised using a Precellys Tissue Homogenizer at 6000 rpm for 1 minute. 400 µL of DNA extraction buffer (volumes shown in ***Table 12***) was added to each sample and homogenised. The sample was then incubated at 65^{o}C for 15 minutes at 400 rpm, followed by a 10-minute stationary incubation period. The DNA pellet was recovered by a 5-minute centrifugation step at 13000 rpm. 300 µL of supernatant was added to 300µl of 100% isopropanol, followed by 3 minutes of centrifugation. After removal of all the supernatant the pellet was washed with 500 µL 70% ethanol and centrifuged for 10 minutes. The supernatant was removed, and the pellet was left to dry at room temperature and resuspended with 60 µL sterile milli-Q H_20. The DNA was then genotyped by PCR reaction, using the Phusion and DreamPCR protocols, and target specific primers (see ***Supplementary data 7.3***).

Table 12: Reagent and working concentration for DNA extraction buffer

Reagent	Working concentration
TRIS-HCl	0.2M
NaCl	0.25mM
EDTA	25mM
SDS	1%

2.2.4 RNA isolation

2.2.4.1 Isolate RNA

Plant material previously harvested using the method described in 2.1.1.2 was transferred to a 2 mL microcentrifuge tube containing 4-5 zirconia beads. 600 µL TRIzol was added and incubated at room temperature for ten minutes. Tissue was disrupted in a Precellys twice at 6000 rpm for one minute with a thirty second break. 120 µL chloroform was added to the homogenised solution, vortexed for fifteen seconds and incubated at room temperature for up to three minutes. Samples were centrifuges for fifteen minutes at 4°C at 15,000 x g. The aqueous phase was transferred to a new tube and one volume of 70% ethanol added and vortexed. The following steps used Invitrogen PureLink RNA Mini Kit spin tubes and buffers with an adapted protocol. The solution was transferred to a spin column and centrifuged for fifteen seconds at 10,000 x g, flow through was discarded. 700 µL wash buffer I was added and centrifuged for fifteen seconds at 10,000 x g, flow through was discarded. 500 µL wash buffer II was added and centrifuged for fifteen seconds at 10,000 x g, flow through was discarded, this step was repeated once. The spin column was centrifuged at 10,000 x g for ninety seconds and the spin cartridge was transferred to a new 1.5 mL Eppendorf. 35 µL nuclease free water was added directly to the filter of the spin column and centrifuged at 10,000 x g for thirty seconds to elute the RNA.

2.2.4.2 Removal of genomic DNA

Ambion Turbo DNAse kit was used to remove the genomic DNA. 1/10 volume of reaction buffer and 1 µL DNase was added to the sample and incubated at 37°C for thirty minutes. 1/10 volume DNase inactivation agent was added, vortexed and incubated at room temperature for ninety seconds. The samples were vortexed and left to incubate at room temperature for ninety seconds. Samples were centrifuged at 12,000 x g for two minutes at room temperature and the supernatant was collected in a new 1.5 mL Eppendorf. Concentration of RNA in the sample was quantified using Nanodrop 2000/2000c spectrophotometer by Thermo Fisher Scientific. RNA was stored at -80°C and used for cDNA preparation. The absence of DNA was confirmed using a genomic primer during qPCR.

2.2.4.3 cDNA preparation

cDNA synthesis was performed using solutions from the RevertAid First Strand cDNA Synthesis Kit.

1900 ng of RNA template was prepared with nuclease free water to a final volume of 11 µL. 1 µL of oligo (dT)18 primer was added to each sampled. Samples were incubated in a thermocycler at 65°C for five minutes. Temperature was reduced to 42°C, and cDNA mix (***Table 13***) was added and incubated for ninety minutes. cDNA was stored at -20°C or used for qPCR analysis.

Table 13: Volume and reagents for one reaction of cDNA mix.

Reagent	Volume (µL)
5x Reaction Buffer	4
10 nM dNTP mix	2
RiboLock RNase Inhibitor (20U/µL)	1
Revert Aid Reverse Transcriptase	1

2.2.5 Seed sterilisation

Seeds were sterilised prior to pipetting onto sterile MS media plates. 1mL 5% sodium hypochlorite solution (with a drop of trytone) was added to the seeds and incubated on a tube rotator at a medium-high speed for up to ten minutes. Seeds were briefly centrifuged, liquid was removed and 1 mL of sterilised ddH_2O added to resuspend seeds, this was repeated three times. All liquid was removed and 140 µL of sterile ddH_2O was added per number of plates to sow onto.

2.2.6 Cloning *E. coli*

To transform *A. thaliana* using *Agrobacterium tumefaciens*-mediated infiltrations, vectors were created using USER reactions. The vectors were then transformed into *E. coli*, colonies positively selected for containing the vector then undergo plasmid extraction for transformation into *A. tumefaciens.* The USER reaction to create the vector and processes to transform *E. coli* will be described next.

2.2.6.1 USER reaction

The USER reaction allows ligation of multiple DNA fragments into a vector by creating sticky ends which combine with other DNA fragments. To increase ligation efficiency all fragments were diluted to equimolar levels. The USER vector to fragment ratio was 5:1. The USER reaction was set up to a final volume of 10 µL as shown in ***Table 14*** and incubated at 37°C for 20 mins followed by 25°C for 20 mins. 1.2 µL of T4 DNA ligase buffer and 1 µL of T4 DNA ligase was added and thoroughly mixed and left to incubate for 1 hour or overnight (~16 hours). The ligated vector is then transformed into *E. coli*.

Table 14: Reagent volumes for the USER cloning reaction.

Reagent	Volume (µL)
DNA fragment(s)	1 (each)
Cutsmart buffer	1
USER enzyme	1
LBJJ233 USER vector	1
Nuclease free water	Up to 10 µL

2.2.6.2 *E. coli* transformations

E. coli was transformed with a ligated vector either using heat shock or electroporation:

2.2.6.2.1 *E. coli* heat shock

For *E. coli* heat shock transformations, 5 µL of USER reaction was used to transform 25 µL of chemically competent *E. coli* (NEB® 5-alpha Competent *E. coli*) cells. Both solutions were mixed by swirling with a pipette, and left on ice for 5 minutes, followed by a 42°C water bath for 30 seconds. The solution was briefly put on ice and after addition of 100 µL of SOC outgrowth (NEB) media incubated at 37°C and 200 rpm for 1 hour. Transformed *E. coli* cells were spread on ½ salt LB media (***Table 15***) plates with kanamycin (50mg/ml, Melford) and incubated overnight at 37°C.

Table 15: Reagents for 1 litre ½ salt LB media. *Addition of agar to make solid LB media, exclusion will produce liquid LB.*

LB media ingredients (per litre)	**Volume (g)**
Peptone	10
Yeast extract	5
NaCl	10
Agar (plates only)	20

2.2.6.2.2 *E. coli* electroporation

For *E. coli* transformations by electroporation, 50 µL of *E. coli* (NEB 10-beta Electrocompetent *E. coli*) was added to 5 µL of the USER reaction, swirled with a pipette tip and then added to a 0.1 cm cuvette (Gene Pulser/MicroPulser Electroporation Cuvettes, BioRad) on ice. The cuvette was placed inside the electroporator (MicroPulser Electroporator, BioRad) and pulsed with 1800v. 500 µL of SOC (NEB 10-beta/Stable Outgrowth Medium) was immediately added and the sample incubated at 37^{o}C for at least 1 hour at 180 rpm. The cells were then spread on ½ salt LB media plates containing kanamycin (50mg/ml) and incubated overnight at 37^{o}C.

2.2.6.3 Identification of positive transformants

One or more single *E. coli* colonies were picked for PCR colony screening. Colonies were either directly added to the DreamTaq PCR mix or before addition to the PCR mix dissolved in 1 µL of nuclease free water and incubated at 95^{o}C for 5 min. For screening of multiple single colonies, up to 5 colonies were dissolved in 10 µL sterile water and incubated at 95^{o}C for 10 minutes followed by a brief vortex. 1.25 µL of this solution was added to the DreamTaq PCR mix. Primer pairs used for colony screening are shown in ***Table 10***.

2.2.6.4 Plasmid extraction

Colonies that were positively identified from the colony screen were grown overnight (12-16 hours) at 37^{o}C and 180 rpm in 10 mL liquid LB media containing kanamycin (50 mg/mL). Cultures were centrifuged for ten minutes at 5000 rpm and the plasmid was extracted using the Monarch Plasmid Miniprep Kit (New England BioLabs) according to manufacturer's instructions. Elution buffer was heated to 50^{o}C. Concentrations were analysed as described in **2.2.1.6** with a spectrophotometer, using the Monarch Plasmid Miniprep Kit elution buffer as a blank.

2.2.6.5 HindIII restriction digest

To verify colony PCR results, restriction digestions were carried out on extracted plasmids. The reaction volumes are shown in ***Table 16***.

Table 16: Reagents used for one reaction of a plasmid restriction digest with HindIII.

Reagent	Amount
Plasmid DNA	1 µg
2.1 Cutsmart buffer	2 µL
HindIII enzyme	0.1 µL
Nuclease free water	Make up to 20 µL

2.2.6.6 Sequencing

Positively identified extracted plasmids were sequenced by Sanger Sequencing using primers (***Table 11***) and a Mix2Seq Eurofins sequencing kit. 50 to 100 ng/ µL of plasmid DNA is added to 2 µL of primer and the final volume is made up to 15 µL using nuclease free water.

2.2.7 *Agrobacterium tumefaciens* transformation

For transformation of plasmids into *A. tumefaciens*, 3 µl (or 1 µg) of plasmid DNA was mixed with 500 µL of chemically competent GV3101 *Agrobacterium tumefaciens* (*A. tumefaciens*) cells and submerged in liquid nitrogen for 5 minutes. Then, samples were incubated at 37°C for 5 minutes and after addition of 500 µL low salt liquid LB media incubated at 28°C for 2 hours. The solution was then spread onto LB media plates containing Kanamycin (50 mg/mL) and Rifampicin (50 mg/mL) and incubated overnight at 28°C. Colonies can be stored in 60% glycerol stock in -80°C.

2.2.8 Transformation of *A. thaliana* by floral dip

A. tumefaciens cultures were grown overnight at 28°C at 200 rpm in 10 mL liquid LB containing Kanamycin (50 mg/mL) and Rifampicin (50mg/mL. Overnight culture was centrifuged at room temperature at 5000 rpm for 15 minutes and the supernatant was removed. The cell pellet was then resuspended in 500mL 5% sucrose solution and 200 µL Silwet L-77 was added (0.04%). The solution was transferred to a 500mL beaker for transformation. *A. thaliana* plants were inverted into the *A. tumefaciens* sucrose solution and immersed for 15 seconds. Once removed from the solution the plants were blotted dry and placed in the dark overnight.

After around 16 hours the plants were removed from the dark and further incubated in long-day *A. thaliana* growth conditions as described in **2.1.1.1**. The plants were bagged as soon as siliques started to brown. After 10 more days the plants were no longer watered and left for another 7 days to dry before seed harvest.

2.2.9 Chromatin Immunoprecipitation

Chromatin immunoprecipitation (ChIP) was carried out to identify histone modifications at specific target sites across the gene clusters of interest. The buffers used throughout the experiment are listed in ***Table 17***. Protease inhibitor cocktail (cOmplete, Roche, 4693116001) was added at 1 tablet per 50ml solution.

Table 17: Buffers used for chromatin immunoprecipitation, and their components.

Buffer	Reagents
Formaldehyde buffer	1% formaldehyde in 1 x Phosphate Buffered Saline
Glycine buffer	2M Glycine in ddH2O.
Buffer A	0.4 M sucrose, 10 mM Tris–HCl, pH 8.0, 5 mM β-mercaptoethanol, protease inhibitor cocktail
Buffer B	0.25 M sucrose, 10 mM Tris–HCl, pH 8.0, 10 mM MgCl2, 1% Triton X-100, 5 mM β-mercaptoethanol, protease inhibitor cocktail
Buffer C	1.7 M sucrose, 10 mM Tris–HCl, pH 8.0, 0.15% Triton X-100, 2 mM MgCl2, 5 mM β-mercaptoethanol, protease inhibitor cocktail
Buffer D	50 mM Tris–HCl, pH 8.0, 10 mM EDTA, 1% SDS, protease inhibitor cocktail
ChIP buffer	1.1% Triton X-100, 1.2 mM EDTA, 16.7 mM Tris–HCl, pH 8.0, 167 mM NaCl, protease inhibitor cocktail
Wash buffer A	150 mM NaCl, 0.1% SDS, 1% Triton X-100, 2 mM EDTA, 20 mM Tris–HCl, pH 8.0
Wash buffer B	500 mM NaCl, 0.1% SDS, 1% Triton X-100, 2 mM EDTA, 20 mM Tris–HCl, pH 8.0
Wash buffer C	0.25 M LiCl, 1% NP-40, 1% sodium deoxycholate, 1 mM EDTA, 10 mM Tris–HCl, pH 8.0
TE buffer	10 mM Tris–HCl, pH 8.0, 1 mM EDTA.

2.2.9.1 Fixation and sonication

10-day old seedlings were harvested by separating roots and cotyledons using a razor blade and forceps. Material was dried lightly using Miracloth and a paper towel before being placed into a 50 mL falcon tube containing formaldehyde buffer. Gauze Mesh was inserted into the tube and a vacuum was applied three times for five minutes each allowing light bubbles to be established in the solution. To end the chromatin crosslinking, 0.125 M glycine buffer was added, and a vacuum was applied for another five minutes. Next, material was washed three times with ddH20 and thoroughly dried with Miracloth and paper towel. Subsequently, the material was transferred into a 1.5 mL Eppendorf tube, immediately submerged in liquid nitrogen and then either stored at -80^{o}C or directly used in the next step.

For further sample processing, material was ground to a fine powder in liquid nitrogen with precooled pestle and mortar. The powder was placed into a falcon tube containing 30 mL of buffer A, vortexed and incubated on ice for five minutes. The solution was then filtered twice through a double layer of Miracloth into a cooled 50 mL falcon tube. The samples were centrifuged at 2800 x g at 4^{o}C for 20 minutes and the supernatant was removed. The pellet, now containing chromatin, was resuspended in 1 mL of buffer B and transferred to a 1.5 mL Eppendorf tube followed by centrifugation at 12,000 x g at 4^{o}C for ten minutes. Following removal of the supernatant the pellet was resuspended in 300 µL buffer C and transferred onto 900 µL buffer C in a new 1.5 mL Eppendorf tube and centrifuged at 14,000 x g at 4^{o}C for one hour. The supernatant was removed and resuspended in 300 µL buffer D in a 1.5 mL Bioruptor Microtube (C30010016, Diagenode).

For chromatin shearing, the extracted chromatin was sonicated using the BioRuptor (Diagenode) using the conditions shown, which was optimised for both leaves and root samples, conditions shown in ***Table 18.***

Table 18: Sonication conditions used for chromatin shearing in leaf and root material.

Sample	On (seconds)	Off (seconds)	Cycles
Leaf	15	90	5
Root	10	90	5

2.2.9.2 Immunoprecipitation and DNA recovery

For immunoprecipitation, first, 15 µL of Dynabeads (Invitrogen) were washed using 1 mL wash buffer A and incubated at 4^{o}C for five minutes on a tube rotator on a low-speed setting. The reaction tube was then placed into a magnetic rack. This led to attachment of the beads to the magnet and allowed safe removal of wash buffer A. The wash procedure was repeated twice. Next, the magnetic beads were resuspended in 50 µL ChIP buffer including the antibody of choice, (target histone modification H3K4me3 Millpore 3660317, core histone H3 abcam ab176842 and mock control IGG) and incubated for 1 hour at 4°C. Afterwards, the wash step with buffer A was repeated three times. The chromatin solution was then diluted with 10x ChIP buffer and 900µL was added to the prepared magnetic beads and incubated overnight on a tube rotator on the lowest setting at 4°C. The next day, the chromatin solution was placed on the magnetic rack and supernatant removed. Then, magnetic beads were washed again twice in 1mL wash buffer A. Subsequently, beads were washed once in 1 mL wash buffer B, once in 1 mL wash buffer C and twice in 1 ml TE buffer. Before the last wash step, beads were transferred into a new clean reaction tube. To recover the chromatin, beads were resuspended in 100 µL 10% Chelex 100 resin (BioRad, 1422822) and incubated at 95°C at 1300 rpm for ten minutes. Samples were briefly centrifuged, 2 µL Proteinase K added, and incubated at 50°C for 30 minutes, and 95°C for ten minutes. Samples were then cooled on ice and 10 µL StrataClean resin (Agilent, 400714) added, vortexed and incubated for ten minutes at room temperature. Finally, the reaction tube was centrifuged at 12,000 x g at room temperature for two minutes. The chromatin containing supernatant was transferred to a clean 1.5 mL Eppendorf tube and stored at -20°C. 1.5 µL of chromatin solution were used for qPCR analysis.

2.2.10 Histone modification data analysis

Histone modification data for *A. thaliana* was downloaded from ReMap2020[175]. In total 143 ChIP-seq data sets were downloaded as .broadPeak files and analysed for the position of histone modification marks across the genome and narrowed down for analysis. R scripts were used to extract histone modification data for regions of interest, carry out a hypergeometric test and fold change using the following packages: MPFR version 4.1.0[176], dplyr version 1.1.2 and tidyverse 1.3.1[177], (see code in **Supplementary 7.4**).

100 regions were selected to be controls with a region size between 20,000 and 200,000 bp, across all five of the chromosomes (**Supplementary 7.4.3, *Table S1***). To allow the files to be easily analysed, they were converted to .csv files. The total size of the region of interest covered in the histone modification mark of interest was extracted from the file (using the

code described in **Supplementary 7.4.4**) using the co-ordinates of the gene clusters of interest in ***Table 19*** and an average of the marks was used from the co-ordinates of the 100 randomly selected regions (***Table S1***). The fold enrichment of the histone modification marks across a region of interest was then calculated in comparison to either the average of the histone modification marks across the 100 random regions or the whole genome using (x/k)/(m/N). Where x is the marks across the region of interest, k is the size of the region of interest, m is the total marks and N is the size of the region to compare (more detail in ***Supplementary 7.4.5)***. The hypergeometric (or Fisher's exact test) using the code supplied in ***Supplementary 7.4.6.***

Table 19: Co-ordinates of the gene clusters

Region of interest	Chromosome	From	To
Marneral cluster	5	19200000	19550000
Arabidiol cluster	4	8700000	8830000
Active thalianol cluster domain	5	19412077	19462516
Inactive thalianol cluster domain	5	19380000	195300000

2.2.11 Genome rearrangement data assembly

2.2.11.1 *A. thaliana* genome assemblies

Seven *A. thaliana* accession de novo assemblies (An-1, C24, Cvi, Eri, Kyo, Ler and Sha) were downloaded from the 1001 Genomes Project. The whole genome sequence for the reference genome, Col-0, was obtained from TAIR using TAIR10.1 release. FASTA and GenBank files for the biosynthetic gene cluster locations for the accessions were obtained from NCBI, Genbank assembly IDs: An-1 GCA_902460265.3, C24 GCA_000222345.1, Cvi GCA_902460275.1, Eri GCA_902460315.1, Kyo GCA_902460305.1, Ler-0 GCA_000835945.1 and Sha GCA_902460295.1.

2.2.11.2 Identification of biosynthetic gene clusters and flanking genes in multiple accessions

Biosynthetic gene cluster composition and positioning in the reference genome Col-0 was determined according to Huang.[157] Respective cluster positioning in non-reference accessions was identified by NCBI nucleotide BLAST of all cluster and neighbouring genes. Cluster regions for each accession were then downloaded from NCBI in both GenBank and FASTA file formats, files in **Supplementary 7.5**. These files were then used for downstream analyses in R and EasyFIG.

2.2.11.3 Bioinformatic analysis

2.2.11.3.1 EasyFig

EasyFig ver. 2.2.5 (Sullivan, Petty, and Beatson, 2011) with BLAST+ version 2.11.0 (Camacho et al., 2009) was used to computationally generate figures to show gene locations, areas of similarity and differences between each genome in a pairwise manner, and where inversions and rearrangements may have occurred.

GenBank format sequences were loaded into EasyFig. BLAST output files and images were generated within the program via BLASTn as a Bitmap file. The default parameters within the program were used aside from colouring: the maximum (100%) BLAST sequence identity with the same orientation in white, while the opposing orientation in green and the minimum BLAST sequence identity (65%) with the same orientation in red while the opposing orientation in red.

2.2.11.3.2 IGV

ATAC-seq data was obtained from NCBI Gene Expression Omnibus under Series GSE116287[178], which show the accessibility of the chromatin across the genome of 10-day old *A. thaliana* seedlings. The data was visualised on IGV_2.8.0 with the Col-0 tracks alongside to view the annotations along the genome.

3 Disrupting cluster integrity

3.1 Overview

The thalianol gene cluster provides an excellent example of non-random gene order in eukaryotes. How the co-ordinated regulation of the five thalianol cluster genes is maintained remains largely unknown. Here, we hypothesise that there are key regions along the cluster, outside of the core promoter elements, that are vital for its regulation.

To identify such regulatory sites, we screened a collection of *A. thaliana* T-DNA mutant lines for their impact on cluster expression. The integration of *A. tumefaciens* derived T-DNA into the *A. thaliana* genome causes natural disruptions of genomic integrity at target sites. As such, they provide an efficient tool to investigate the role of specific genomic sites for genome activity. A previous study, using a collection of 12 mutant lines with T-DNA integrations surrounding the *AtTERT* telomerase gene identified multiple novel *cis* regulatory elements.[179] Here, we aimed to analyse a range of T-DNA lines with disruptions across the thalianol interacting active and inactive domains, within the cluster genes as well as in regions we predict could be regulatory regions using available data.

The process for selecting T-DNA lines will be outlined, alongside genotyping and the analysis of the qPCR data for the mutant lines. By looking at the changes in expression patterns of the thalianol cluster genes we can investigate the potential factors involved in the regulatory outcome of the genes. This can then be linked to the 3D chromosome conformation and its impact on the co-regulation of the thalianol cluster, thus extending the knowledge on gene expression and the eukaryotic genome.

Alongside this, we aim to create our own insertional mutations using CRISPR/Cas9 as well as deleting regions we deem key for the thalianol clusters regulation. We will outline the CRISPR/Cas9 process, selection of mutagen sites, cloning and screening of potential mutants.

3.1.1 Genome editing techniques

An organism's genome can be manipulated in various ways, such as deleting regions of DNA, inserting DNA fragments or by altering base pairs and genes[180]. Genetic engineering allows the organism's genome to be studied intracellularly within its own environment to develop novel medicinal and biotechnological applications[180]. Specific nucleases are often used to create site-specific changes in the genome, such as TALENS or ZFNs [181]. Zinc-finger nucleases (ZFNs) contain a nuclease domain of the FokI restriction enzyme and are sequence-specific endonucleases that are customised to cleave a DNA target of choice[182]. The low success rate of engineering of target site specificity is the biggest downside to using this technique[183]. An alternative is using transcription activator-like effector nucleases (TALENs). TALENs comprise of a FokI nuclease domain fused to a domain customizable for target DNA binding[184]. Proteins secreted from bacteria, derived from highly conserved repeats (transcription activator-like effectors, TALEs) alter gene transcription by binding to host DNA. An artificial DNA-binding domain alongside a Fokl nuclease can be created to infect an organism, thus causing a double stranded break.[184,185] The engineering of the nuclease protein to guide the DNA cleavage required on both techniques described is often expensive and time consuming [181].

CRISPR/Cas9 is an alternative gene editing method that relies on endonuclease activity guided by a programmable RNA, which is often a quicker and cheaper method [186], overview depicted in ***Figure 6***.

This system is based on an adaptable immune response evolved in bacteria and archaea which is mediated by RNA to protect the organism from invading viruses and plasmids[187]. The microbe integrates short fragments of the foreign nucleic acids into its own chromosome at a region of tandem repeats known as clustered regularly interspaced palindromic repeats (CRISPR) in response to an invading pathogen[188]. A protospacer, the integrated DNA, is directly adjacent to noncoding, repetitive element of RNA which allows spacing between exogenous DNA and contains a sequence associated with each protospacer called a protospacer adjacent motif (PAM)[189]. The PAM allows for these systems to recognise self from non-self (foreign genetic material), thus varying between organisms.[190] Upon secondary infection of the pathogen the CRISPR loci are transcribed into a long primary transcript, then processed into short CRISPR-RNAs (crRNAs) of around 20 nt that is complementary to previously invading pathogens[191]. This system requires fusion with an auxiliary trans-activating crRNA (tracrRNA) to facilitate the processing of the crRNA array and to form the single-guide RNA (sgRNA)[192]. A complex then forms between the sgRNA and the Cas9

(CRISPR-associated protein 9) nuclease which directs the protein to the foreign target DNA using the complementary sequence provided by the crRNA. The The target DNA must directly precede a PAM sequence which in this case is 5'-NGG[186]. The Cas9 will then cause a double stranded break 3 nt upstream from the PAM sequence leading to erroneous DNA with the aim to cause cell death for invading pathogens[193].

This system can be programmed for targeted genome editing by expressing Cas9 and a sgRNA (with a complementary sequence to the target) in an organism, the only drawback being the requirement of a directly adjacent PAM in the host's genome[194]. The target sequence will then undergo a DSB leading to a DNA repair mechanism, which can result insertions or deletions at the re-ligated site[195]. Double stranded breaks in DNA are considered the most dangerous as they can result in the loss of huge chromosomal regions[196], organisms therefore have several mechanisms to repair the DSB.

The double-stranded break is recognised by a Ku70-Ku80 heterodimer protein (Ku) which acts to recruit and activate other proteins to the site, such as the DNA-dependent protein kinase catalytic subunit (DNA-PKcs) to repair the DNA by non-homologous end joining (NHEJ), depicted in ***Figure 6B*** [197]. The DNA-PKcs triggers cell signalling to orchestrate a downstream repair process such as recruitment of scaffold proteins to bind DNA Ligase 4 to seal the DSB[198], if the DSB does not have blunt ends but rather sticky ends, an endonuclease is recruited to excise the overhangs[199]. In contrast, homology directed repair (HDR), shown in ***Figure 6C***, a copy of a homologous sequence to align the DSB ends prior to ligation is required[200]. HDR usually takes place during the S phase of the cell cycle where there is a sister chromatid to be used as the homologous template to accurately copy and restore the DSB[201]. NHEJ is typically the method sought after while attempting to cause deletions and mutations within a target genome, whereas HDR is preferred to insert new DNA fragments into the genome by providing the organism with a copy of the broken DNA which contains errors.

Although the CRISPR/Cas9 system is highly adaptable to all organisms providing a rapid method for genome editing[202], it is still challenging in more recently evolved plants such as *A. thaliana* due to the low efficiency of HDR and limited donor template delivery into the plant[181]. *A. thaliana* can be transformed much more rapidly than other plant species through the floral dip method[203], but due to low heritability rates and transformation efficiencies[204,205] CRISPR/Cas9 is challenging. The floral dip method has significant benefits over tissue culture approaches but the frequency of homozygous CRISPR/Cas9 induced mutations reported in T1 lines are very low[203]. Many protocols have been developed to increase the efficiency of homozygous T1 mutations with the most successful being a sequential transformation method using a germ-line specific promoter (DD45) to drive Cas9 expression followed by a second transformation of the sgRNA construct, which showed a knock-in efficiency of 16–55% in the T2 generation[174]. This indicates that it is important that Cas9 is expressed in germ cells or during early embryogenesis for increased HDR-directed gene editing in *A. thaliana* [203]. Vector design has also proved to be a limiting factor of genome editing in *A. thaliana,* thus improvement in design can lead to higher frequencies of large deletions, up to 70 kb[206].

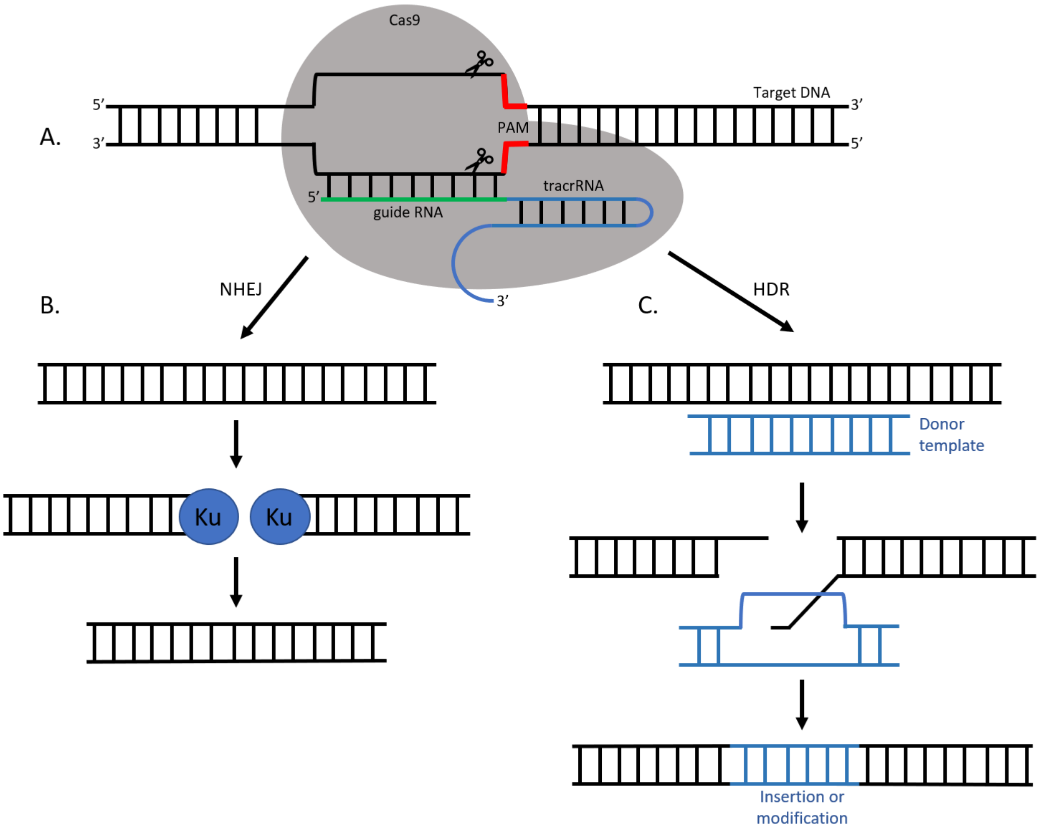


Figure 6: CRISPR/Cas9 genome editing system and the double stranded break DNA repair mechanisms. *A. Diagram representation of the Cas9 nuclease binding to the target DNA, guided by the 20nt sgRNA (green) when fused to the trans-activating CRISPR-RNA (tracrRNA, blue). Cas9 will cleave the target DNA ~3bp upstream of the protospacer adjacent motif (PAM, red) leading to a double stranded break (DSB). The DSB is repaired either by B non-homologous end joining (NHEJ) that can lead to random insertions or deletions at the cleavage site, or C homology directed repair (HDR) that utilised a copy of the DNA to repair itself. By providing a donor template with insertions or mutations these can get inserted into the genome during HDR following a DSB.*

3.1.2 USER cloning and fragment assembly

CRISPR/Cas9 requires a DNA construct to be assembled made up of multiple sequence-specific DNA fragments of various sizes. Multiple methods can be used to precisely join two or more DNA fragments together. For example, fusion PCR allows for the seamless fusion of several PCR fragments by simply amplifying each fragment with overlapping complementary sequences and subsequent fusion PCR reaction with the combined fragments as templates[207].

Fusion PCR is a highly versatile and efficient assembly technology, yet, it is prone to the incorporation of errors due to the requirement of several PCR reactions[208]. To avoid multiple PCR reactions several methods have been developed to assemble multiple DNA fragments in one cloning step such as Golden Gate cloning and Gibson Assembly. Golden Gate cloning allows for the assembly of up to nine DNA fragments without the addition of sequence homology onto the end of DNA fragments[209]. This method is based on two DNA fragments flanked by restriction sites and digestion by a Type IIS enzyme and ligation[210]. Assembled products do not contain restrictions sites and, therefore, are not subject to redigestion following ligation. Golden Gate cloning requires the addition of restriction sites onto each DNA fragment and the domestication of DNA to remove any restriction sites within each fragment[209]. Gibson Assembly allows for the assembly of DNA constructs of up to 900 kb by creating overhangs on DNA fragments by T5 exonuclease and subsequent ligation by a DNA ligase[211].

As an alternative to the classical restriction digest based methods, Uracil excision based USER (Uracil-Specific Excision Reagent) cloning has been developed[212]. USER cloning requires a single-step process and approximately 1 hour to assemble a vector containing ligated DNA fragments (depicted in ***Figure 7***). For USER cloning, DNA fragments are amplified with primers containing a deoxyuridine (U) to replace a deoxythymidine (T) approximately 7 to 9 bases from each end of the fragment[208]. All the fragments to be ligated are then put into the same reaction with the USER vector and USER enzyme mix (a mixture of UDG and DNA glycosylase-lyase endonuclease VIII) to cleave the U, leaving a 7 to 9 nt complementary sticky end to fuse each fragment. The USER vector cassette must be cleaved with two restriction enzymes, Nt.BbvCI and PacI, prior to cloning. The end fragments must be amplified with specific USER primers which are complementary to the vector, also eliminating insertion of flipped fragments[209].

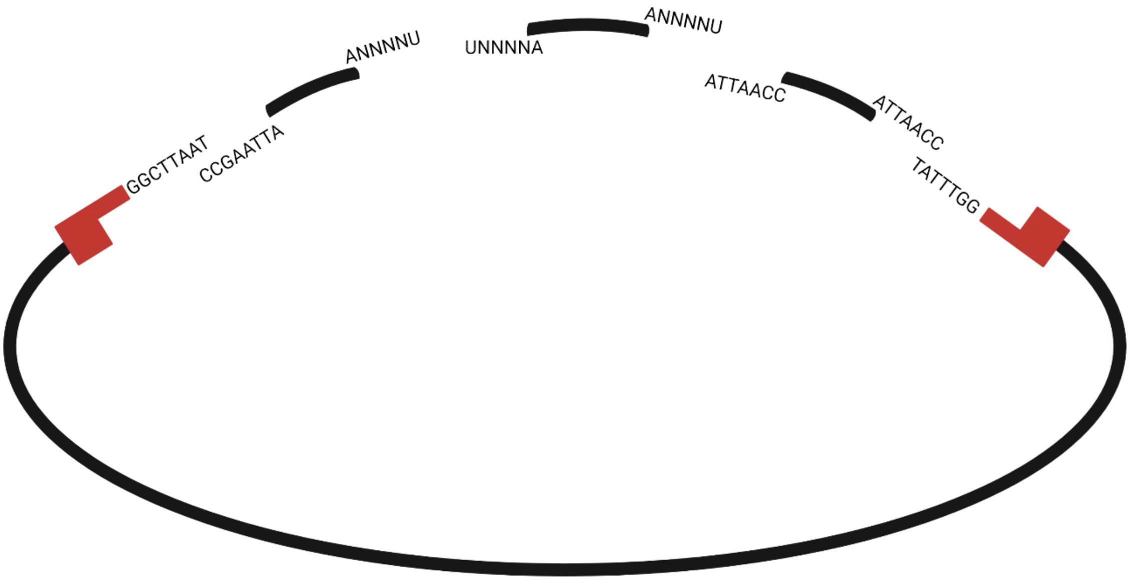


***Figure 7. Schematic of the USER cloning reaction**. The USER reaction allows ligation of many DNA fragments (grey) into a USER vector (blue). The USER vector is cleaved with Nt.bbvci/paci restriction enzymes, exposing sticky ends. Those fragments to be inserted into the vector must be amplified with complementary sticky ends to the vector, and the internal fragments amplified with primers containing uracil. The USER enzyme cleaves the uracil exposing sticky ends. The sticky ends bind complementary to its neighbouring fragment, followed by ligation to create a construct.*

3.1.3 Project objectives

We next will explain the selection of T-DNA mutagens, the data obtained and discuss how the insertion may have disrupted the 3D architecture across the genome. Then, establish a CRISPR/Cas9 and cloning protocol, and describe the outcomes of the genome editing.

3.2 Disruption of the gene order results

3.2.1 Selection of T-DNA lines

There are thousands of T-DNA lines that have an insert of a known size DNA fragment at known locations to cause disruption of that site. We used previously published 2Chi-C data[140] as well as ATAC-seq data[178] to identify regions that would be of interest if they were disrupted by a T-DNA mutant line.

3.2.1.1 Thalianol cluster interacting domains

Published Hi-C data identified two large interacting regions in the thalianol cluster that are of interest to us, shown in ***Figure 8***. In the roots where the thalianol cluster is expressed an active interacting domain is formed spanning 50 kb from *THAA2* to *THAS*, whereas in leaves an inactive domain forms spanning 110 kb from *THAA2* to the non-cluster genes *At5g48150* and *At5g48160 (PAT1* and *OBE2*, respectively).[163] This contrasting chromatin conformation change between leaves and roots is suggested to ensure organ-specific gene expression. Thus, we selected 19 T-DNA lines in the larger leaf inactive domain, which covers the thalianol cluster and a large region downstream of the cluster. The large interacting region which forms in leaves will be referred as the inactive domain, and the smaller active interacting region which forms in the roots as the active domain, see ***Figure 8C***.

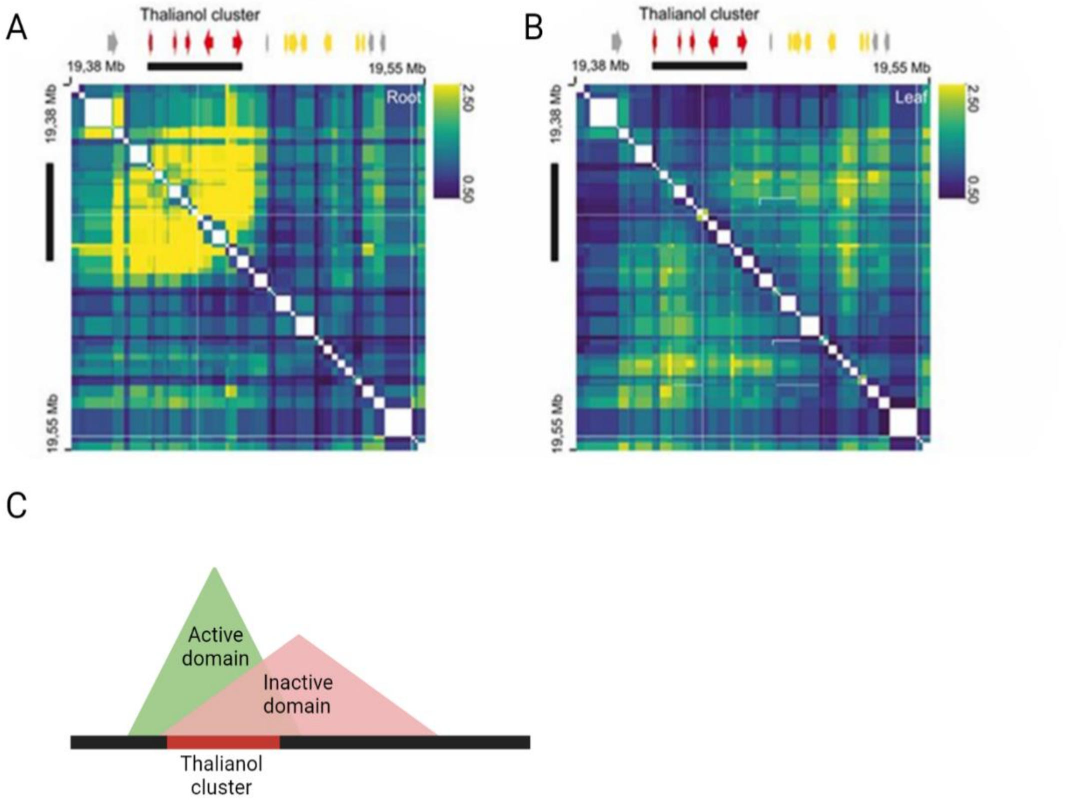
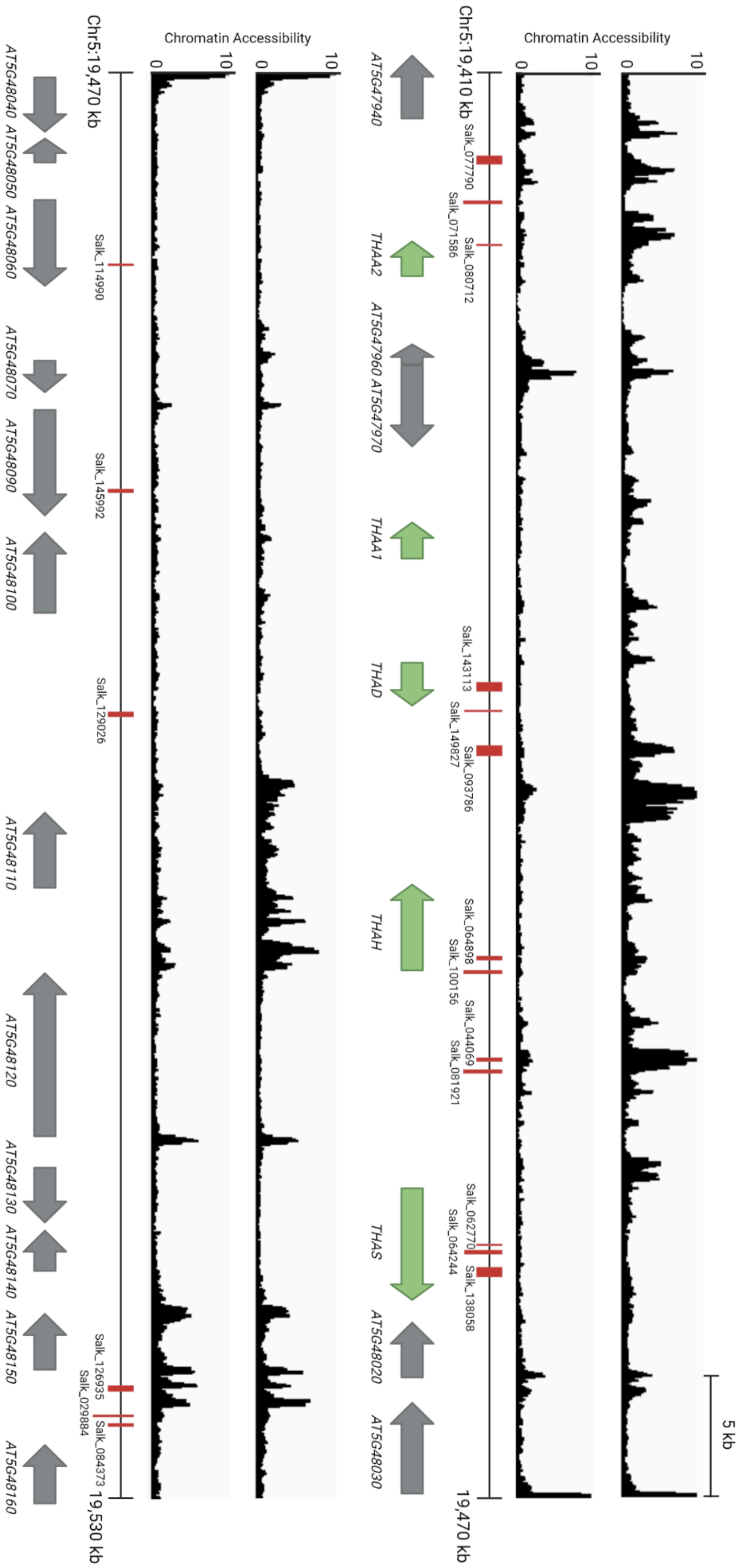


Figure 8. Interacting regions of the thalianol gene cluster. *Interaction map of the thalianol gene cluster region in roots (A) and leaves (B) from Hi-C data. Strong to weak interaction tendency is indicated by yellow to blue colouring. Model of the interacting domains formed in the thalianol gene cluster, indicated in red. The inactive domain is weaker and forms in leaves during transcriptional repression, while the active domain is much stronger and forms during transcriptional activation in roots. (Adapted from Nützmann et al, 2020)*

3.2.1.2 Identifying regions of interest in the thalianol gene cluster using ATAC-seq data

Chromatin accessibility data has been shown to be a good predictor of important regulatory sites. The Genomic areas with ATAC-seq (Assay for Transposase-Accessible Chromatin) peaks are more accessible to binding of transcription factors and other DNA binding proteins, thus indicating their regulatory role. Available ATAC-seq data for *A. thaliana* allows us to identify differentially accessible regions (DARs) in roots and shoots throughout and nearby the thalianol cluster. Using this data, we have been able to narrow down the key target sites for T-DNA library screening to analyse for the effects of disrupting the cluster architecture on co-regulation of the thalianol cluster. The lines selected, shown in ***Figure 9***, contain those that are in more accessible regions, in less accessible regions, in intragenic as well as intergenic regions to allow for a comparative analysis.

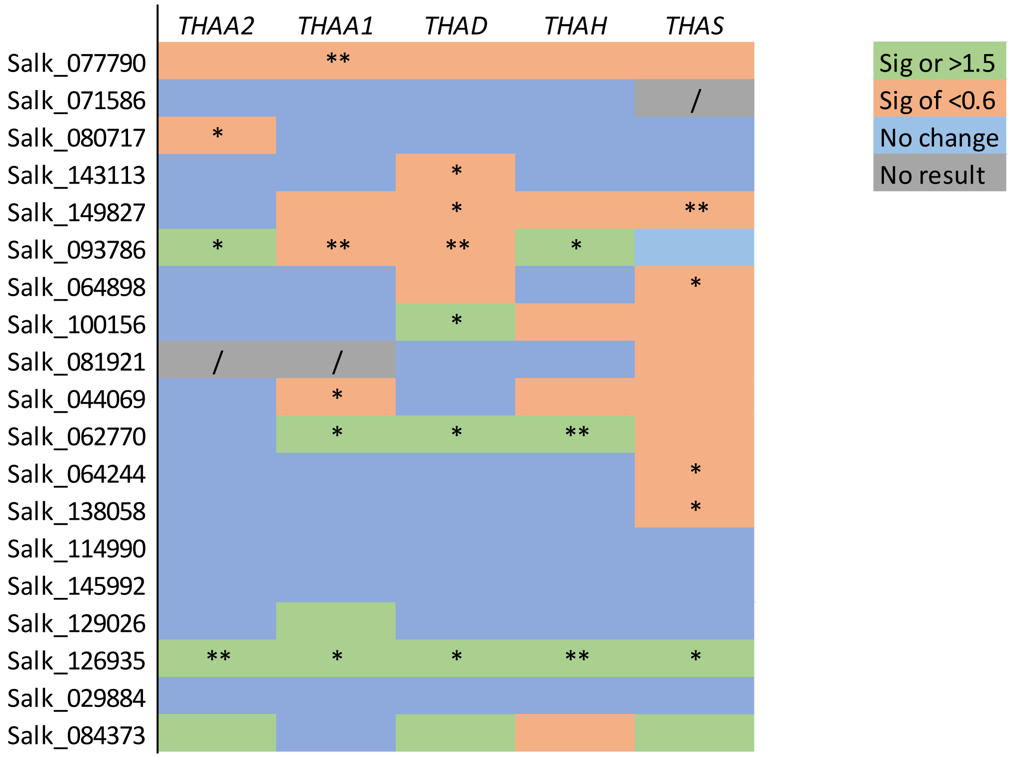

Chromatin Accessibility
0
10
0
10
Chr5:19,410 kb
19,470 kb
5 kb
Salk_077790
Salk_071586
Salk_080712
Salk_143113
Salk_149827
Salk_093786
Salk_064898
Salk_100156
Salk_044069
Salk_081921
Salk_062770
Salk_064244
Salk_138058
AT5G47940
THAA2
AT5G47960
AT5G47970
THAA1
THAD
THAH
THAS
AT5G48020
AT5G48030
Chromatin Accessibility
0
10
0
10
Chr5:19,470 kb
19,530 kb
Salk_114990
Salk_145992
Salk_129026
Salk_126935
Salk_029884
Salk_084373
AT5G48040
AT5G48050
AT5G48060
AT5G48070
AT5G48090
AT5G48100
AT5G48110
AT5G48120
AT5G48130
AT5G48140
AT5G48150
AT5G48160

3.2.2 Genotyping

To analyse the impact of disruption in the gene cluster and to identify points of co-regulation, 19 different T-DNA lines were chosen. T-DNA lines used for this experimentation were previously genotyped and selected for homozygosity using guidelines published by The SALK Institute Genome Analysis Laboratory (SIGnAL).

3.2.3 Transcript analysis of T-DNA cluster disruption lines

To assess if the T-DNA insertion lines have an impact on the expression of genes within the cluster, qPCRs were carried out on homozygous T-DNA lines. Each set of lines was sown alongside a Col-0 reference to maintain comparable growth conditions. Only root samples were used in the experiments described below. qPCR analyses were performed to test the transcript levels of *THAA2, THAA1, THAD, THAH* and *THAS* with the constitutively expressed gene, *At1g13320* (*PP2AA3)* as a reference gene. Most qPCR analysis was carried out with three biological replicates, however some had two biological replicates where indicated. A depiction showing an overview of the data collected is shown in ***Figure 10*** and ***Figure 11***.



	THAA2	*THAA1*	*THAD*	*THAH*	*THAS*
Salk_077790	🟥	🟥 **	🟥	🟥	🟥
Salk_071586	🟦	🟦	🟦	🟦	▨ /
Salk_080717	🟥 *	🟦	🟦	🟦	🟦
Salk_143113	🟦	🟦	🟥 *	🟦	🟦
Salk_149827	🟦	🟥	🟥 *	🟥	🟥 **
Salk_093786	🟩 *	🟥 **	🟥 **	🟩 *	🟦
Salk_064898	🟦	🟦	🟥	🟦	🟥 *
Salk_100156	🟦	🟦	🟩 *	🟥	🟥
Salk_081921	▨ /	▨ /	🟦	🟦	🟥
Salk_044069	🟦	🟥 *	🟦	🟥	🟥
Salk_062770	🟦	🟩 *	🟩 *	🟩 **	🟥
Salk_064244	🟦	🟦	🟦	🟦	🟥 *
Salk_138058	🟦	🟦	🟦	🟦	🟥 *
Salk_114990	🟦	🟦	🟦	🟦	🟦
Salk_145992	🟦	🟦	🟦	🟦	🟦
Salk_129026	🟦	🟩	🟦	🟦	🟦
Salk_126935	🟩 **	🟩 *	🟩 *	🟩 **	🟩 *
Salk_029884	🟦	🟦	🟦	🟦	🟦
Salk_084373	🟩	🟦	🟩	🟥	🟩

🟩	Sig or >1.5
🟥	Sig of <0.6
🟦	No change
▨	No result

Figure 10: Overview of transcript analysis data for T-DNA mutants of the thalianol cluster genes in roots. *Blue represents no significant change in transcript levels, grey represents no result taken, green represents a significant increase or an increase larger than 1.5 and red represents a significant decrease in or a decrease less than 0.6. Statistical significance (paired t-test): *P < 0.05, **P < 0.01.*

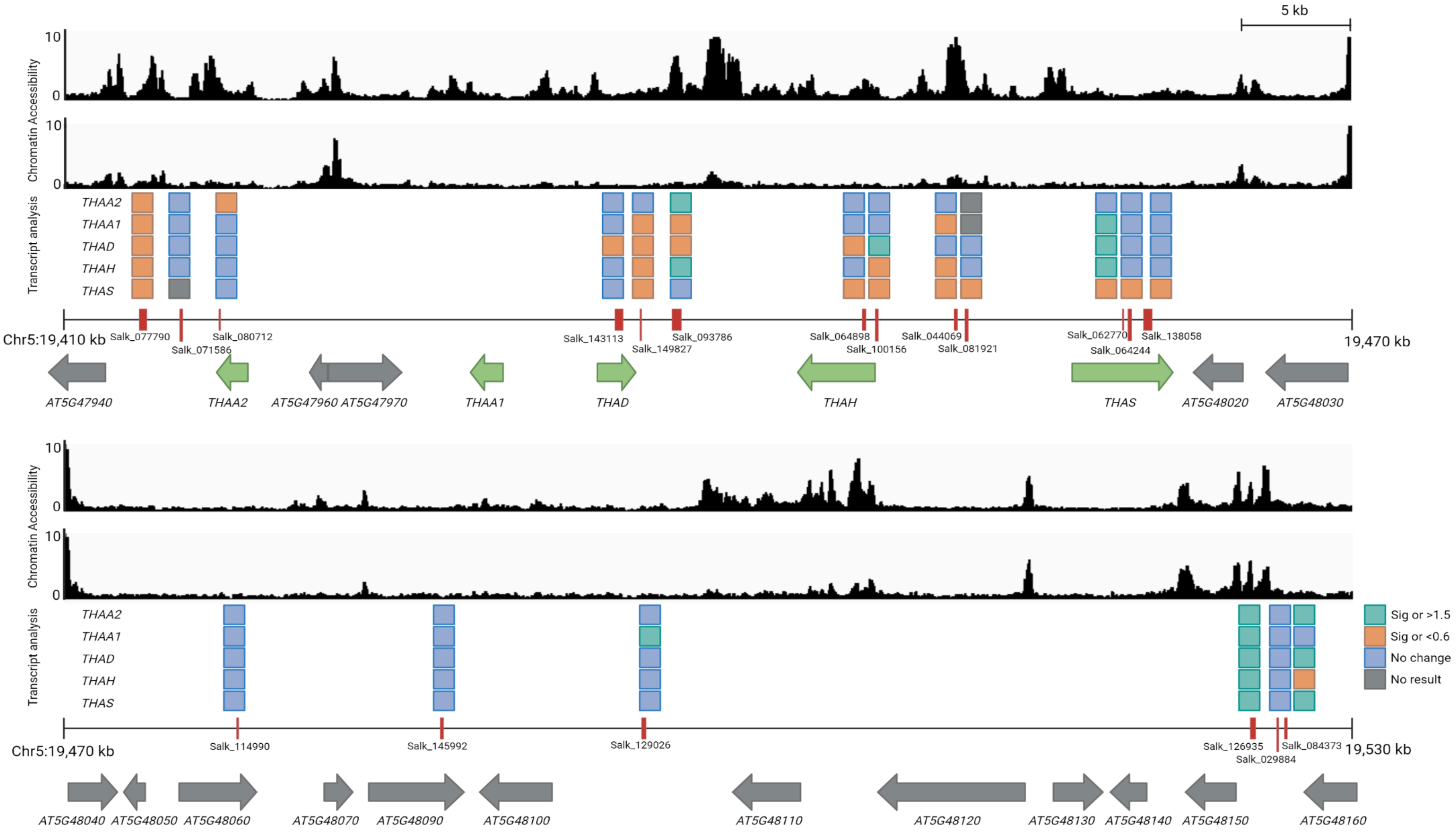


Figure 11: ATAC-seq data across the thalianol gene cluster and interacting regions overlaid by an overview of transcript analysis for T-DNA mutants *ATAC-seq data indicates regions of accessible chromatin in the roots and shots of A. thaliana. T-DNA lines (red) were selected based on their positioning throughout the thalianol cluster (cluster genes indicated by green arrows) and the extended downstream region (non-clustered genes in grey) shown to scale. Blue squares represent no significant change in transcript levels, grey squares represent no result taken, teal squares represent a significant increase or an increase larger than 1.5 and orange squares represents a significant decrease in or a decrease less than 0.6.*

3.2.3.1 Mutants downstream of *THAA2* lead to downregulation of the thalianol cluster

Two T-DNA lines, downstream of the *THAA2* gene, were analysed to assess if any intergenic regions outside of the gene cluster could contain key regulatory elements. Salk_077790 showed a trend for overall downregulation of the four core thalianol cluster genes tested alongside *THAA2, **Figure 12***. Moderate reduction in the transcript levels for *THAA1* was found, which was significant (P-value or P < 0.01) compared to the wild-type. Although the transcript levels were all similarly reduced for the other genes analysed in Salk_077790, they were not significant due to the variation between biological samples, a common limitation to this experimental procedure alongside *THAA2* having two biological replicates. When we analysed the transcript levels of Salk_071586, an insertion closer to the *THAA2* gene no expression trends or significant difference to wild type expression levels were observed.

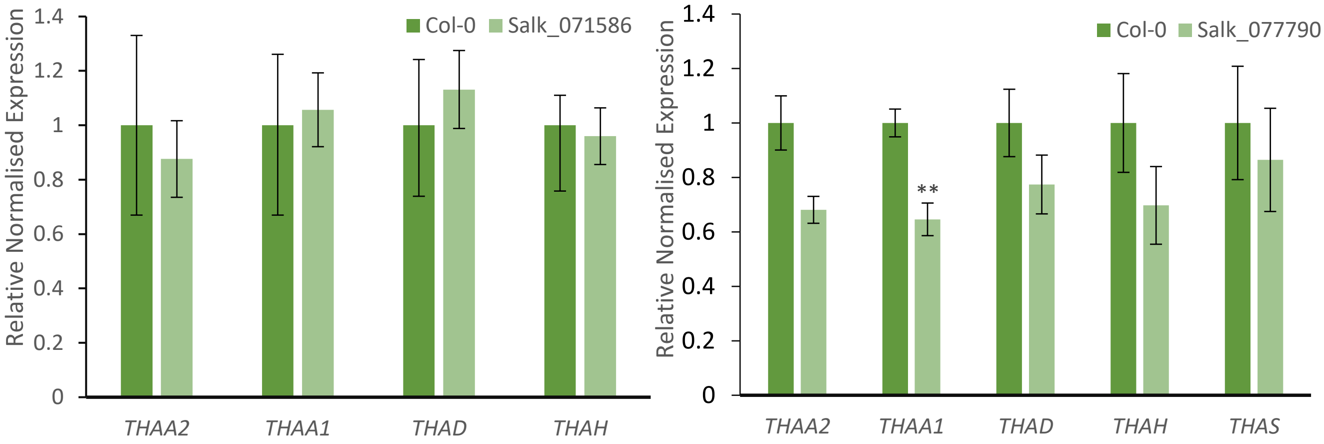


Figure 12: Transcript analysis of two T-DNA mutants, Salk_071586 and Salk_077790, of the five thalianol cluster genes. *Gene expression was measured by qPCR and compared to the wild-type (Col-0) transcript level and normalised to 1. An internal control was used, PP2AA3 (AT5G13320). Error bars indicate standard error of the mean for three biological replicates. Statistical significance (paired t-test): *P < 0.05, **P < 0.01. Two biological replicates were used for Salk_077790 primer THAA2.*

3.2.3.2 *THAA2* misregulation does not alter the four core thalianol cluster genes

Salk_080717 contains a T-DNA insert near the end of the *THAA2* gene, as this gene is debated if it is a core thalianol cluster gene, a mutation in this locus could allow a better definition of its role in the cluster. Relative transcript levels of *THAA2* are significantly and substantially decreased compared to the wild type (P < 0.05), ***Figure 13***. A minor yet non-significant trend towards reduced expression levels is observed for the four core thalianol cluster genes tested.

This indicates that the expression levels of *THAA2* does not alter the expression levels of the core thalianol cluster genes.

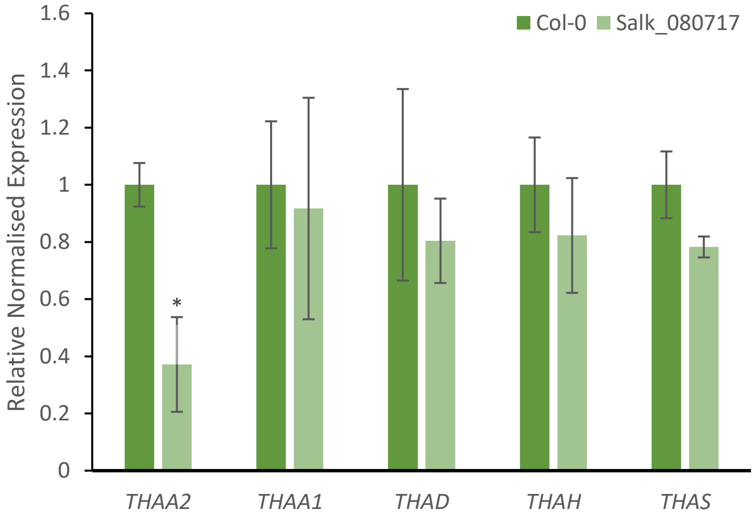


Figure 13: Transcript analysis of the T-DNA mutant Salk_080717. *Gene expression was measured by qPCR and compared to the wild-type (Col-0) transcript level and normalised to 1. An internal control was used, PP2AA3 (AT5G13320). Error bars indicate standard error of the mean for three biological replicates. Statistical significance (paired t-test): *P<0.05, **P<0.01*

3.2.3.3 Disruption in *THAD* does not lead to cluster misregulation, however T-DNA insertion directly downstream of *THAD* causes gene cluster down regulation

To assess if *THAD* contains any important regulatory elements for the co-regulation of the thalianol gene cluster a T-DNA line, Salk_143113, containing a T-DNA insertion within the ORF (open reading frame) of the *THAD* gene was selected.

As expected, transcript levels of the *THAD* gene were dramatically and significantly reduced (P <0.05) in Salk_143113 compared to the wild type, ***Figure 14***. However, no changes to gene expression were detected for the other cluster genes. Expression levels for *THAA1* and *THAH* were characterised by a high variation between biological replicates.

Similar to the results obtained for Salk_143113, in Salk_149827, harbouring a T-DNA insertion just downstream of *THAD*, *THAD* expression is dramatically repressed (P < 0.05). However, in contrast to Salk_143113, Salk_149827 also shows a dramatic reduction in transcript levels of all other cluster genes. Due to variation of transcript levels between biological replicates only the reduction in transcript levels for *THAD* and *THAS* has been determined as statistically significant, yet *THAA1* and *THAH* show a similar reduction when compared to the wild type. The reduction of *THAA2* expression is least pronounced among all 5 genes. These results indicate that the region just downstream of *THAD* could hold a key regulatory region for the co-regulation of the thalianol gene cluster.

A T-DNA insertion further downstream of Salk_149827 provides an intriguing yet different cluster expression pattern. As before, in Salk_093786 transcript levels of *THAD* are strongly reduced. Furthermore, and consistent with Salk_149827, *THAA1* levels are reduced compared to the wild type. All other cluster genes, however, show a contrasting pattern of increased transcript levels. By disrupting the architecture of the gene cluster approximately 300 bp downstream of the *THAD* gene it appears the core genes become unregulated. Both downstream genes within the cluster, *THAH* and *THAS* show a marked trend for upregulation, although only significantly in the *THAH* gene ($P < 0.05$) due to a high variation between biological samples. We find a significant dramatic decrease in transcript numbers of *THAD* ($P < 0.01$). *THAA1* also dramatically decreases ($P <0.01$). Interestingly, *THAA2* shows a marked increase in the T-DNA line which is significant ($P < 0.05$). Neither T-DNA insertion in Salk_149827 nor Salk_093786 disrupts a gene or is located in a promoter element thus suggesting a non-canonical role of the region downstream of *THAD* in cluster regulation.

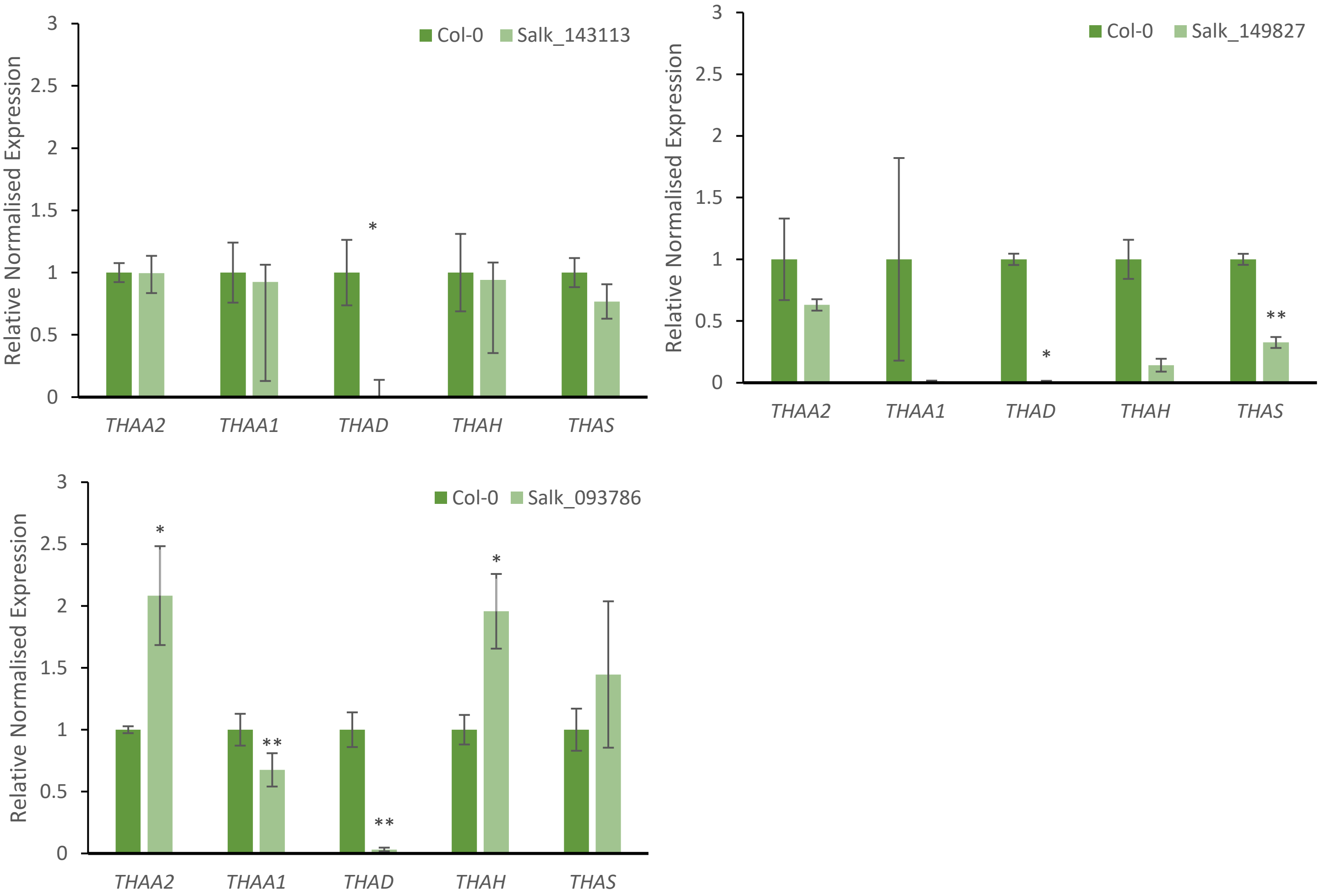


Figure 14: Transcript analysis of three T-DNA mutants, Salk_143113, Salk_149827 and Salk_093786, of the five thalianol cluster genes. *Gene expression was measured by qPCR and compared to the wild-type (Col-0) transcript level and normalised to 1. An internal control was used, PP2AA3 (AT5G13320). Error bars indicate standard error of the mean for three biological replicates. Statistical significance (paired t-test): *P<0.05, **P<0.01.*

3.2.3.4 Disruption of *THAH* leads to *THAD* misregulation

Next, we analysed the impact of disrupting the *THAH* gene using two T-DNA lines, Salk_064898 and Salk_100156. The T-DNAs in these mutant lines disrupt the gene at different positions. The thalianol cluster shows a different expression pattern in both lines, ***Figure 15***.

In particular, the *THAD* transcript levels show a pronounced upregulation in Salk_100156 which is significantly higher than the wild-type (P < 0.05), whereas there is a trend for downregulation in Salk_064898. This difference is also apparent for *THAH,* interestingly we see reduced levels in Salk_100156 but increased levels in Salk_064898. The two mutant lines show similarly reduced transcript levels for *THAS* and *THAA2*, although not significant.

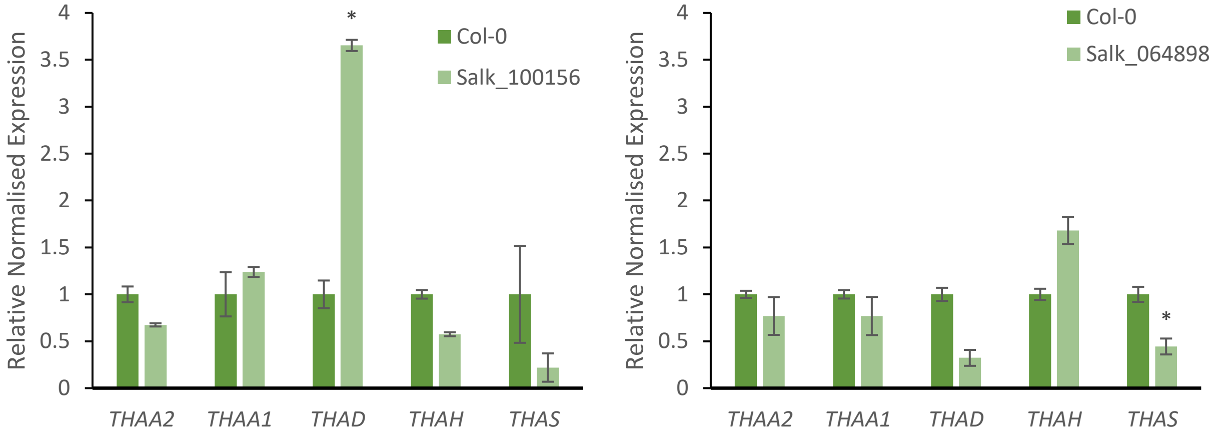


Figure 15: Transcript analysis of two T-DNA mutants inside the THAH gene, Salk_100156 and Salk_064898, of the five thalianol cluster genes. *Gene expression was measured by qPCR and compared to the wild-type (Col-0) transcript level and normalised to 1. An internal control was used, PP2AA3 (AT5G13320). Error bars indicate standard error of the mean for three biological replicates. Statistical significance (paired t-test): *P<0.05, **P<0.01.*

3.2.3.5 Mutations in the intergenic region between *THAH* and *THAS* leads to cluster disruption

Next, we analysed cluster expression levels in the mutant lines Salk_044069 and Salk_081921, which cause interruption in the intergenic region between *THAH* and *THAS* and are in a region of accessible chromatin. In Salk_044069 a T-DNA is integrated between the *THAH* and *THAS* genes. Our comparative transcript analyses, ***Figure 16***, show that both flanking genes are considerably reduced in their expression (P is equal to 0.051 and 0.053, respectively). The other three cluster genes show variable expression levels, with no change for *THAD*, a light increase for *THAA2* and a moderate, yet significant, decrease for *THAA1* compared to the wild type (P < 0.05). The misregulation of the cluster genes caused by disruption at this site indicates this could be a key regulatory region.

In Salk_081921 a T-DNA is inserted just downstream for the *THAS* gene. Here, we detected a trend for downregulation of the three cluster genes tested, *THAD, THAH and THAS,* whereby *THAS* transcript levels showed the strongest decrease.

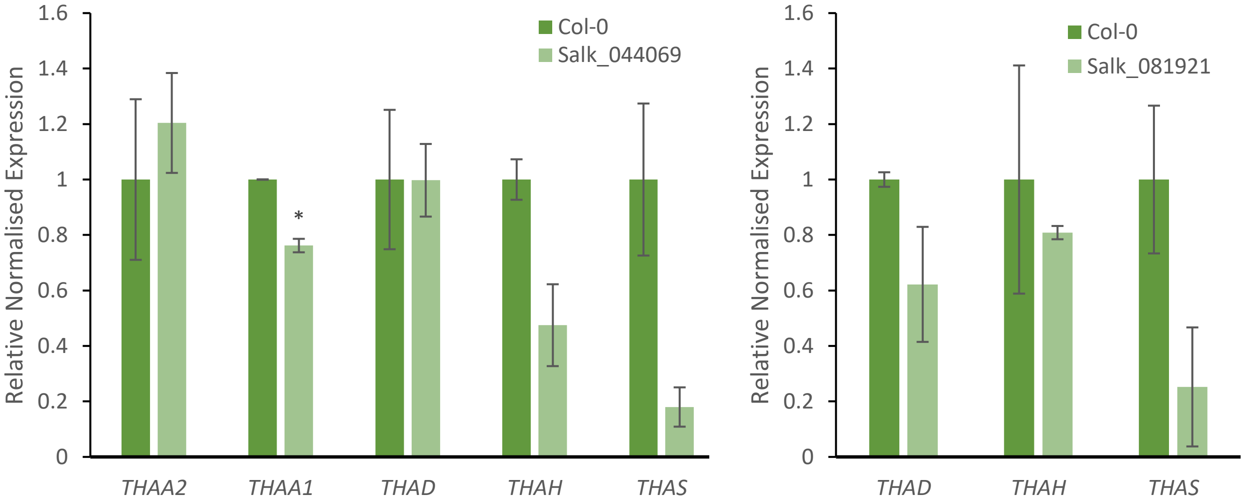


Figure 16: Transcript analysis of two T-DNA mutants, Salk_044069 and Salk_081921, of the five thalianol cluster genes and three genes for Salk_081921. *Gene expression was measured by qPCR and compared to the wild-type (Col-0) transcript level and normalised to 1. An internal control was used, PP2AA3 (AT5G13320). Error bars indicate standard error of the mean for three biological replicates. Statistical significance (paired t-test): *P<0.05.*

3.2.3.6 Mutation in eighth exon of *THAS* leads to increased expression of the cluster genes, while at a different position no change is shown

As for the *THAH* gene, we analysed multiple T-DNA lines with insertions in the *THAS* gene for their impact on cluster expression levels. The *THAS* gene encodes for the first catalytic step in the biosynthesis of thalianol and its derivatives and is therefore of central importance to the cluster. The T-DNA insertions for two of the respective lines, Salk_062770 and Salk_064244, are just approximately 270 bp away from each other and fall within the eighth and ninth exon of the gene. The T-DNA insertion of Salk_138058 falls into the twelfth exon of *THAS*.

As expected, all three lines showed considerable reduced transcript levels for the *THAS* gene, ***Figure 17***, which is significant in Salk_138058 and Salk_064244 (P < 0.05) and a P value of 0.060 for Salk_062770. In Salk_064244, the other four cluster genes remain similar to wild type expression levels. In contrast, Salk_062770 shows a significantly marked increased in transcript levels for *THAA1, THAD* and *THAH* (P < 0.05) while only *THAA2* remains stable in its expression.

Salk_138058 follows a similar, albeit much weaker, expression trend to Salk_062770. The *THAA1, THAD* and *THAH* transcripts show a slight non-significant increase and THAA2 remains stable in comparison to the wild type.

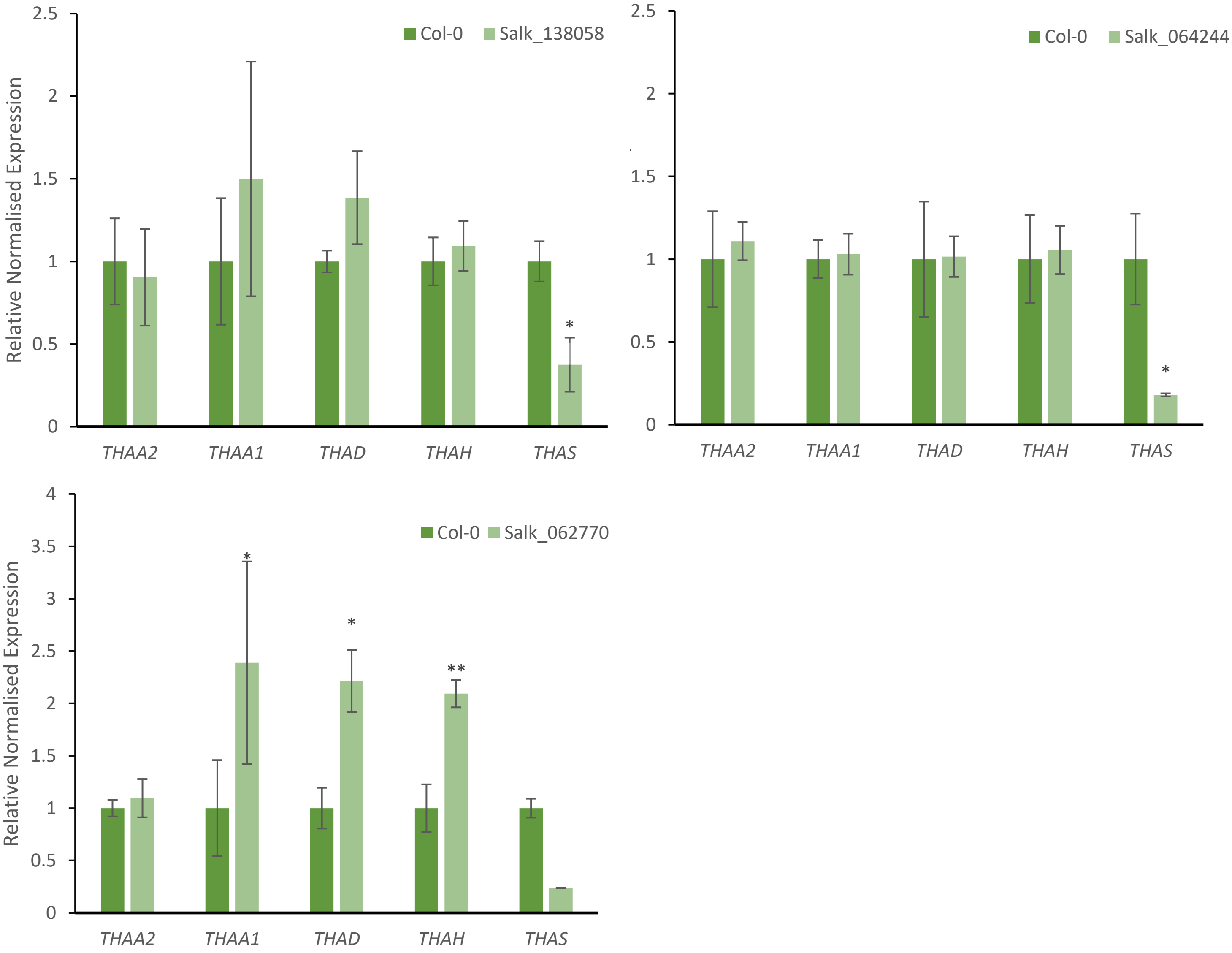


Figure 17: Transcript analysis of three T-DNA mutants in the THAS gene, Salk_138058, Salk_064244 and Salk_062770, of the five thalianol cluster genes. *Gene expression was measured by qPCR and compared to the wild-type (Col-0) transcript level and normalised to 1. An internal control was used, PP2AA3 (AT5G13320). Error bars indicate standard error of the mean for three biological replicates. Statistical significance (paired t-test): *P<0.05, **P<0.01.*

3.2.3.7 Disruption downstream of the cluster misregulates *THAA1*

Next, we analysed cluster expression levels in mutant lines with T-DNA insertions downstream of the entire cluster. This region is part of the larger 3D domain formed around the silenced gene cluster. The T-DNA insertion in Salk_129026 falls between the *At5g48100* and *At5g48110* genes about 30 kb away from the cluster. Compared to the wild type, we detected no significant changes in cluster transcript levels, ***Figure 18***. However, we noted downward trends for *THAD*, *THAH* and *THAS* and a slight upward trend for *THAA1* (P = 0.09). Our results may indicate that this downstream region could play a role in the regulation and expression of the thalianol gene cluster.

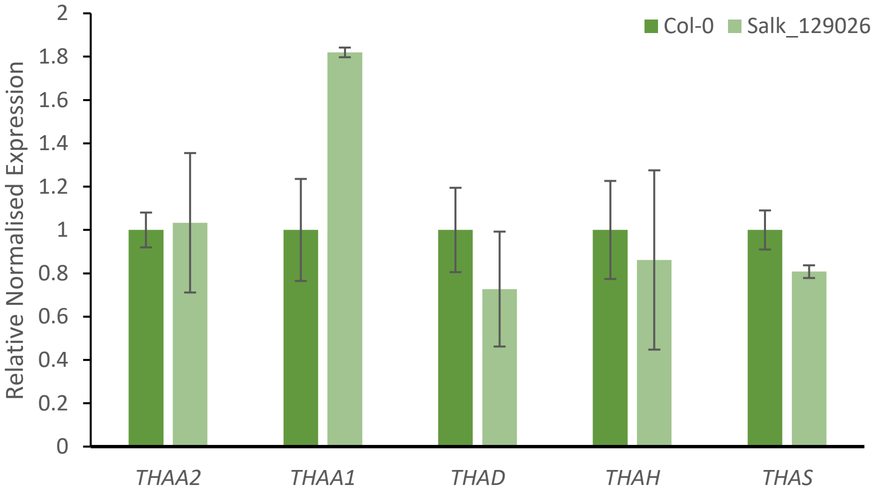


Figure 18: Transcript analysis of the T-DNA mutant, Salk_129026 of the five thalianol cluster genes. *Gene expression was measured by qPCR and compared to the wild-type (Col-0) transcript level and normalised to 1. An internal control was used, PP2AA3 (AT5G13320). Error bars indicate standard error of the mean for three biological replicates. Statistical significance (paired t-test): *P<0.05, **P<0.01.*

3.2.3.8 Insertions in *AT5G48060* and *AT5G48090* do not disrupt the thalianol gene cluster

T-DNA lines were selected to interrupt *At5G48060* and *At5g48090*, which are downstream of the thalianol cluster to assess if interrupting genes within the inactive domain will affect the transcription levels of the thalianol gene cluster. For both Salk_114990 and Salk_145992 we found no change in transcription levels, ***Figure 19***, across all five thalianol cluster genes, indicating that altering these genes does not alter the expression of the thalianol gene cluster.

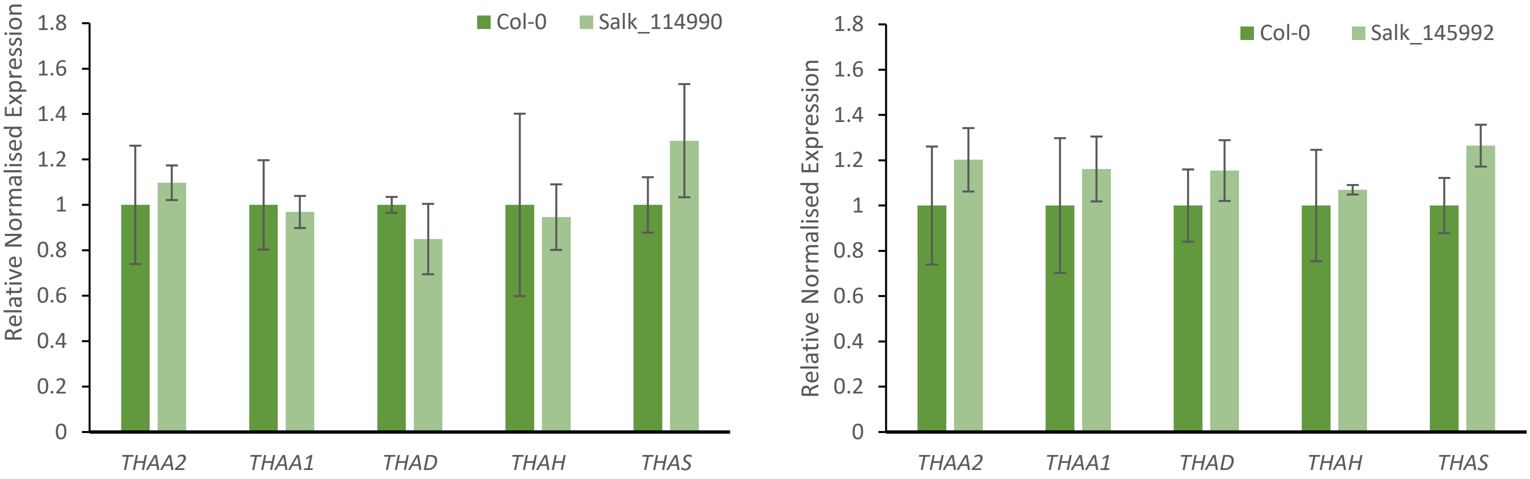


***Figure 19: Transcript analysis of two T-DNA mutants, Salk_114990 and Salk_145992, of the five thalianol cluster genes.** Gene expression was measured by qPCR and compared to the wild-type (Col-0) transcript level and normalised to 1. An internal control was used, PP2AA3 (AT5G13320). Error bars indicate standard error of the mean for three biological replicates. Statistical significance (paired t-test): *P<0.05, **P<0.01.*

3.2.3.9 Disruption downstream of the gene cluster leads to disruption of cluster gene expression

The T-DNA insertions in Salk_126935, Salk_029884 and Salk_084373 are located between the genes *PAT1* and *OBE2 (At5g48150* and *At5g48160*, respectively) and are approximately 1500 bp apart from each other. The T-DNA insertion in Salk_126935 disrupts an area with multiple ATAC-seq peaks, whereas the insertions in Salk_029884 and Salk_084373 are outside these highly accessible regions. Salk_029884 does not show significant changes in transcript levels of the thalianol gene cluster, ***Figure 20***. Salk_126935, however, shows a significant (P < 0.05) increase in all four of the thalianol cluster genes and, to a lesser but significant extent, in *THAA2.* Salk_084373, located closer to *OBE2,* shows a trend for upregulation of *THAA2, THAA1, THAD* and *THAS* compared to the wild type. For *THAH* however, we detected a considerable lowered transcript levels in this T-DNA line (P = 0.087). This data indicates that important cluster regulatory sites could be located downstream of the cluster.

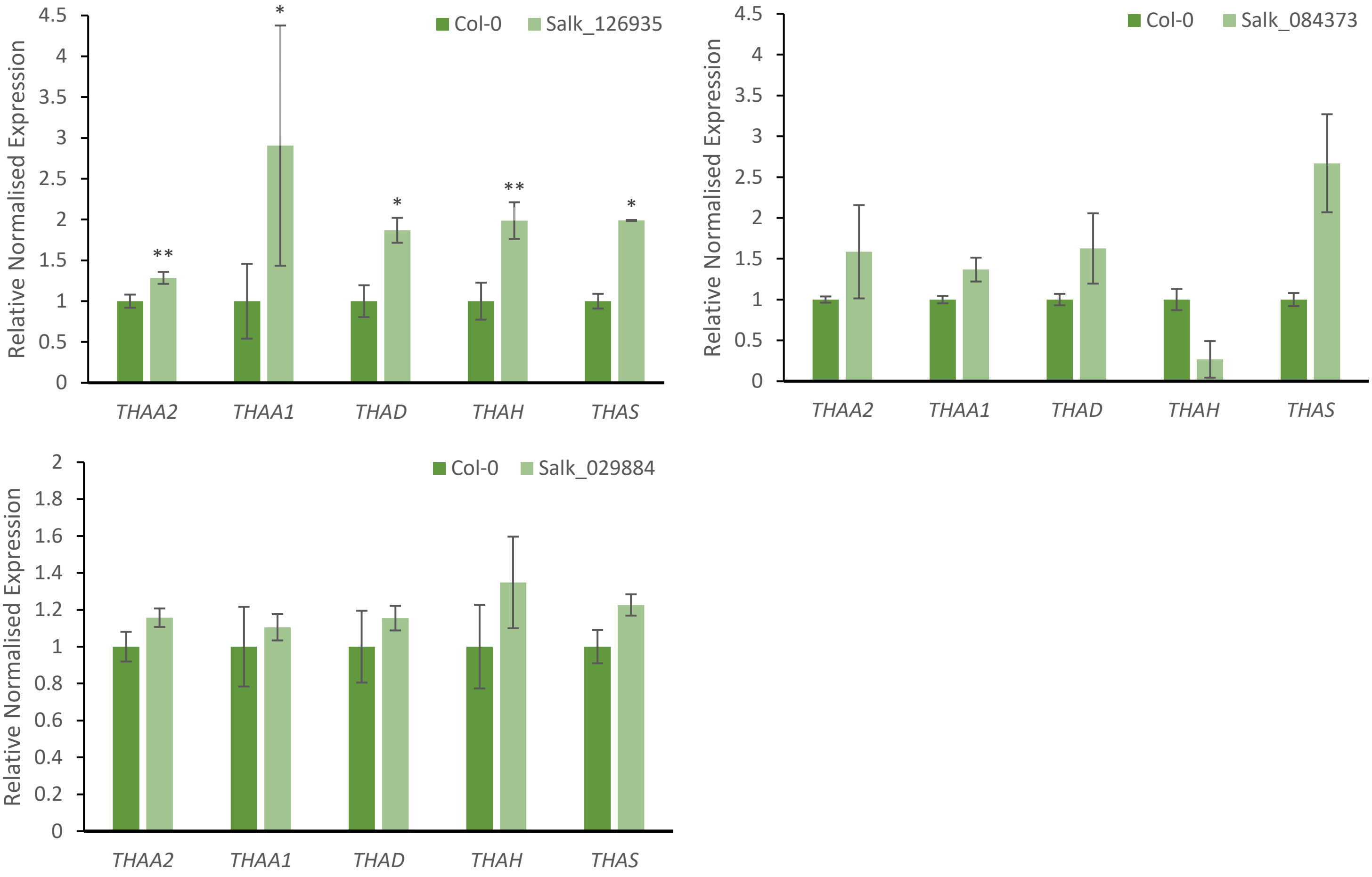


Figure 20: Transcript analysis of three intergenic T-DNA mutants, Salk_126935, Salk_084373 and Salk_029884, of the five thalianol cluster genes. *Gene expression was measured by qPCR and compared to the wild-type (Col-0) transcript level and normalised to 1. An internal control was used, PP2AA3 (AT5G13320). Error bars indicate standard error of the mean for three biological replicates. Statistical significance (paired t-test): *P<0.05, **P<0.01.*

3.2.4 Understanding gene cluster evolution

To better understand the evolution of plant metabolic gene clusters, we decided to investigate gene order and general DNA architecture of three metabolic gene clusters in multiple *A. thaliana* accessions. *A. thaliana* is a genetically very well characterised plant species and over 1000 *A. thaliana* accessions representing habitats from around the world have had their genome sequenced. Most of the accessions, however, have been studied using short-read sequencing analyses with reference-guided assemblies. These enable detection of nucleotide changes only and typically do not identify large-scale rearrangements such as inversions well.[213]

To account for potential inversion, deletion and translocation[214] across our metabolic gene clusters of interest we focused on *A. thaliana* accessions with good quality reference genome assemblies. Such large-scale rearrangements of gene clusters are likely products of abundant TE activity within clusters. We selected eight accessions for our study, An-1, C24, Cvi-0, Eri-1, Kyo, Ler, Sha and Col-0. The geographical origin of each accession is shown in ***Figure 21.***

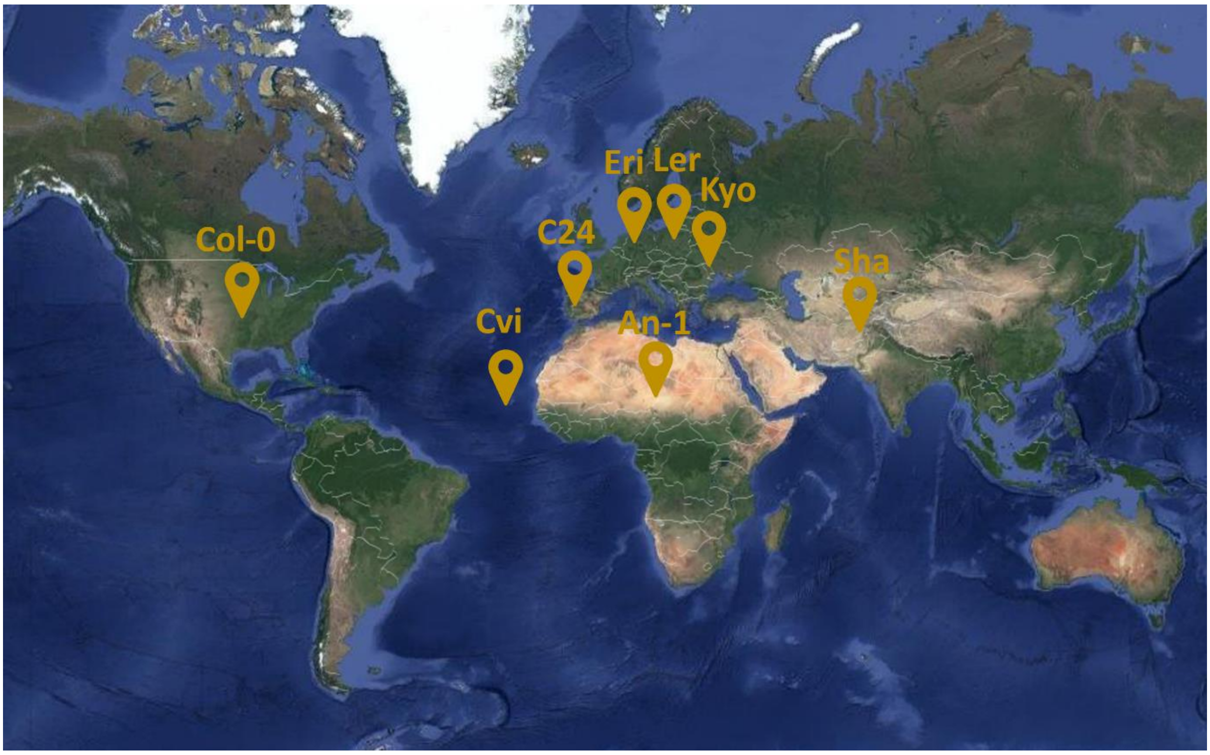


Figure 21: World map showing the geographical location of eight A. thaliana geographical locations. *Orange tab indicated the location where the accession naturally grows.*

3.2.4.1 Identifying gene cluster locations within the accessions

To analyse cluster organisation in the selected accessions, firstly, we identified the genome co-ordinates of the thalianol, marneral and arabidiol cluster genes in the respective *A. thaliana* accessions (***Table 20, 21 and 22***). By establishing the location of all cluster loci, we can determine the basic genomic differences between each accession and the wild type. We noted an almost identical chromosome positioning for all clusters across accessions. Each cluster is localised on the same chromosome at highly similar chromosome co-ordinates. In contrast to the overall location of all clusters, we identified significant changes to cluster organisation across all three clusters. For example, in the C24 accession only two of the five thalianol cluster genes were identified, *THAA2* and *THAH,* suggesting a significant reduction in cluster size. Notably, the other three thalianol cluster genes were not detected elsewhere in the genome of this accession, as determined by BLASTn analysis. For the marneral cluster, we identified an additional gene within the cluster region in the Eri accession, implicating a cluster expansion. For the arabidiol/baruol cluster, we detected both reduction and expansion of the cluster, as it is a much larger cluster, so susceptible to more variation. Only one accession, An-1, stays consistent in gene number and cluster size compared to Col-0. Both Kyo and Sha accessions show a reduced gene number of only 20 genes within the gene region, compared to 23 in the wild type, while the four accessions, C24, Cvi, Eri and Ler all gain three genes within the cluster region, as well as increasing the size by around 25,000 bp.

Table 20: Co-ordinates and features of the thalianol gene cluster across different accessions.

Thalianol gene cluster	Chromosome	Start co-ordinates	End co-ordinates	Total cluster length (bp)	Number of known genes in region
Col-0	5	19397041	19466899	69858	5
An-1	5	19181844	19252609	70765	5
C24	5	18496268	18531827	35559	2
Cvi	5	18881222	18954168	72946	5
Eri	5	19023660	19095844	72184	5
Kyo	5	19251229	19323254	72025	5
Ler	5	19002179	19073204	71025	5
Sha	5	19470109	19542297	72188	5

Table 21: Co-ordinates and features of the marneral gene cluster across different accessions.

Marneral gene cluster	Chromosome	Start co-ordinates	End co-ordinates	Total cluster length (bp)	Number of known genes in region
Col-0	5	17023502	17058245	34743	3
An-1	5	16851673	16886544	34871	3
C24	5	16145202	16179342	34140	3
Cvi	5	16583768	16617487	33719	3
Eri	5	16680205	16714643	34438	4
Kyo	5	16905939	16939982	34043	3
Ler	5	16682077	16715743	33666	3
Sha	5	17160263	17195310	35047	3

Table 22: Co-ordinates and features of the arabidiol gene cluster across different accessions.

Arabidiol gene cluster	Chromosome	Start co-ordinates	End co-ordinates	Total cluster length (bp)	Number of known genes in region
Col-0	4	8730499	8850000	119501	23
An-1	4	8999337	9118879	119542	23
C24	4	9485098	9629608	144510	26
Cvi	4	8504689	8646962	142273	26
Eri	4	8702625	8849585	146960	26
Kyo	4	8965676	9083691	118015	20
Ler	4	9173265	9320242	146977	26
Sha	4	9028894	9142034	113140	20

3.2.4.2 Basic look at the movements of genes in accessions

Then we took a closer look at the order of genes within the cluster and larger flanking regions, on a non-scale basis and not taking intergenic regions into account. We looked at a large region of the genome pinpointed by the ATAC-seq and Hi-C data as previously mentioned in section **3.2.1.1**. The thalianol cluster in Col-0 is described as having five cluster genes and two intervening genes between *THAA2* and *THAA1, **Figure 22***. In all seven of the accessions these two intervening genes have either moved upstream of the cluster genes, An-1, Eri, Kyo, Ler and Sha, or downstream, Cvi. Interestingly, in C24 only two of the cluster genes remain, *THAA2* and *THAH*.

In the marneral gene cluster, ***Figure 23***, for six of the accessions we detected the same cluster genes as for Col-0. Eri had an insertion of an intervening gene between the second and third gene of the cluster, *MRO* and *MRN1*. A BLAST was carried out on the sequence of the intervening gene and matched to *At1G40087*, a protein coding gene for a plant transposase.

The baruol/arabidiol cluster shows most organisational variation among the three clusters. In Col-0 (wild-type) 10 cluster genes, CYP702A2, CYP702A3, CYP705A1, PEN1, CYP705A2, CYP705A3, BARS1, CYP705A4, HSR201, BIA1, as well as 3 intervening genes are described. Only one of the accessions (An-1) follows the same pattern as the WT, ***Figure 24***. 4 accessions have 11 genes within the cluster with an addition of a duplicate CYP702A2, although this fragment isn’t the full length. 2 of the accessions, Kyo and Ler, have 8 of the cluster genes, missing CYP702A2 and CYP702A3 from the start of the cluster. The number of intervening genes also varies from 2 to 6 genes. The genes encode for a putative CYP, cellulose synthase-like or unknown proteins. All 10 genes downstream of the cluster are detectable in consistent positions across all accessions.

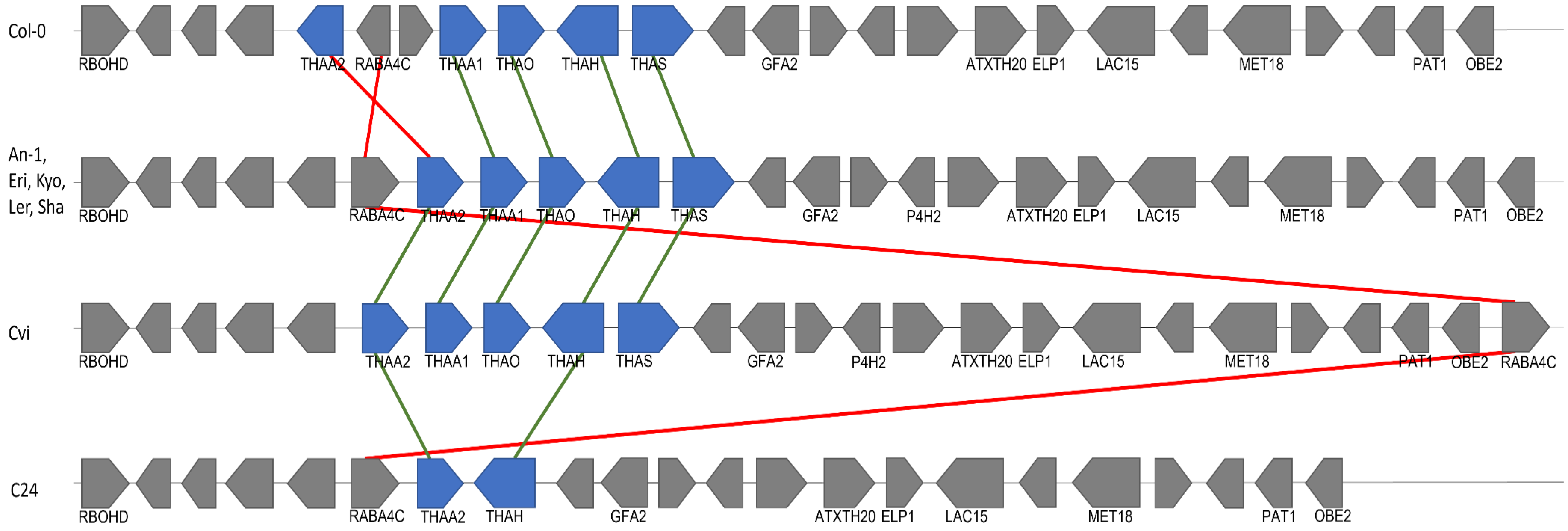


Figure 22: A depicted overview of the thalianol gene cluster in different accession. *The thalianol cluster genes, blue arrows, are typically arranged into four core genes directly neighbouring one another, and a fifth cluster gene separated by two interrupting genes. The four core cluster genes show no movement (gren line) in 7 of the 8 accessions, aside from in C24, there only two of the cluster genes are identified. The non-cluster genes (grey), also show little movement apart from the fifth cluster gene and the intervening genes that show an inversion or movement elsewhere (red line). Not to scale.*

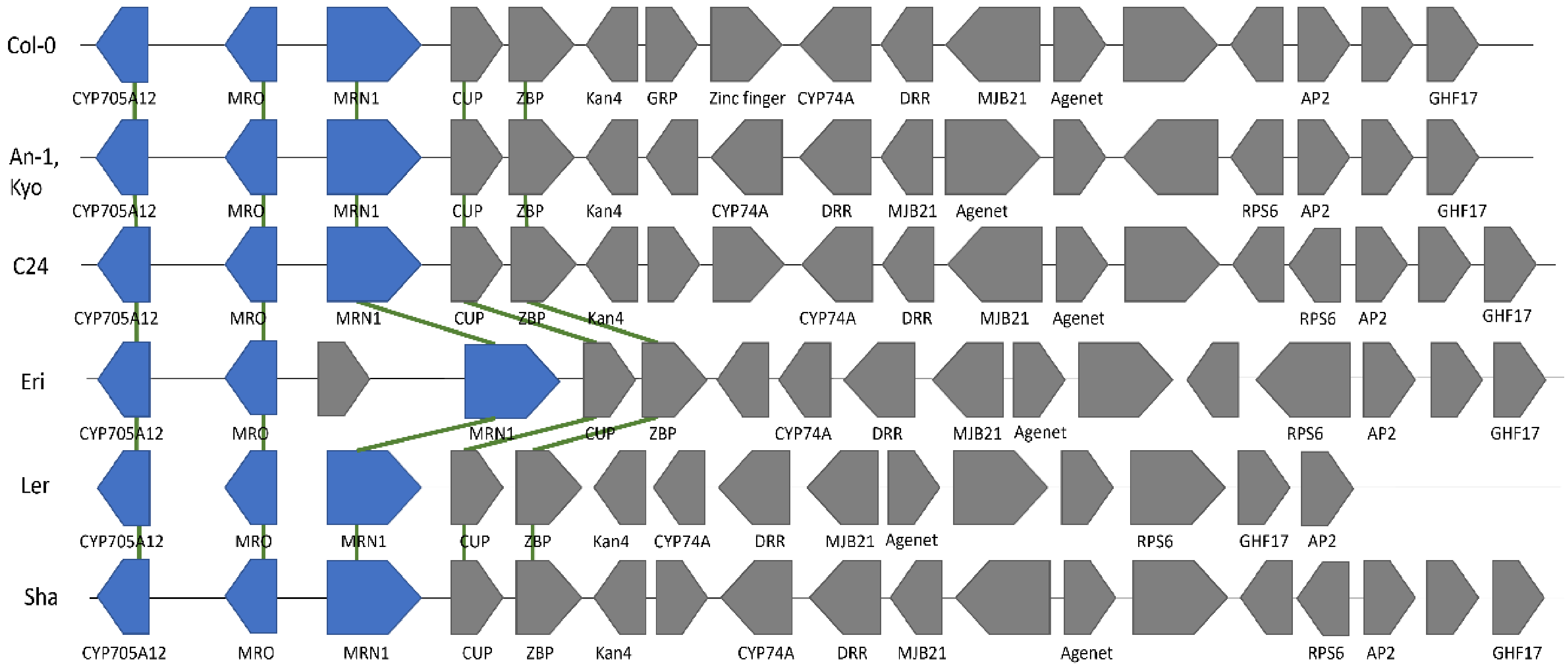


Figure 23: Depicted overview of the marneral gene cluster in different accession. *The three marneral cluster genes (blue) are highly conserved in the order, all aside from one accession, Eri, keep the same gene order. Eri contains an intervening gene between two of the cluster genes MRO and MRN1. The non-cluster genes (grey) shwo some minor changes between accessions.Not to scale.*

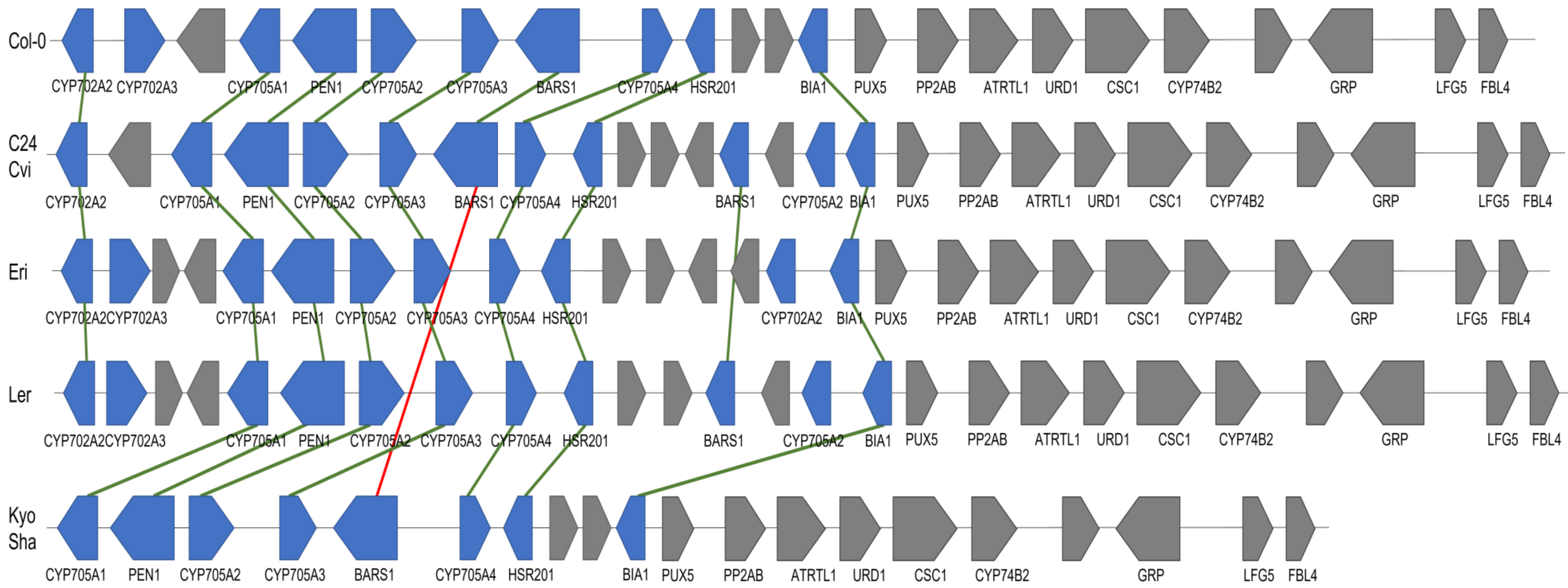

Col-0
CYP702A2 CYP702A3 CYP705A1 PEN1 CYP705A2 CYP705A3 BARS1 CYP705A4 HSR201 BIA1 PUX5 PP2AB ATRTL1 URD1 CSC1 CYP74B2 GRP LFG5 FBL4
C24
Cvi
CYP702A2 CYP705A1 PEN1 CYP705A2 CYP705A3 BARS1 CYP705A4 HSR201 BARS1 CYP705A2 BIA1 PUX5 PP2AB ATRTL1 URD1 CSC1 CYP74B2 GRP LFG5 FBL4
Eri
CYP702A2CYP702A3 CYP705A1 PEN1 CYP705A2 CYP705A3 CYP705A4 HSR201 CYP702A2 BIA1 PUX5 PP2AB ATRTL1 URD1 CSC1 CYP74B2 GRP LFG5 FBL4
Ler
CYP702A2CYP702A3 CYP705A1 PEN1 CYP705A2 CYP705A3 CYP705A4 HSR201 BARS1 CYP705A2 BIA1 PUX5 PP2AB ATRTL1 URD1 CSC1 CYP74B2 GRP LFG5 FBL4
Kyo
Sha
CYP705A1 PEN1 CYP705A2 CYP705A3 BARS1 CYP705A4 HSR201 BIA1 PUX5 PP2AB ATRTL1 URD1 CSC1 CYP74B2 GRP LFG5 FBL4

Figure 24: Depicted overview of the Arabidiol/baruol gene cluster in different accession. *The ten cluster genes (blue) show large variations between all the accessions (red lines) as well as the cluster gene number changing between each accession. The non-cluster genes (grey) show variation in the cluster intervening regions and in the downstream region.*

3.2.4.3 A more in depth look at the variations of the cluster between accessions

The above-described studies allowed us to look at gene diversity between accessions. Using different analyses, we have been able to look at the clusters in greater depth including variability in the intergenic space. For example, EasyFig cluster visualisation allows us to identify changes in the gene organisation and structure and establish the conservation of the clusters between accessions. We will focus on the marneral and thalianol clusters for the next section.

Previously, we identified that the C24 accession had three of the thalianol cluster genes missing, this more detailed look at the cluster region determines that there are large deletions throughout the cluster. The two remaining cluster genes, *THAA2* and *THAA1*, do not have most of the non-coding regions. Using ATAC-seq data, the intergenic non-coding regions that have remained were in highly conserved regions which displayed differential chromatin accessibility between roots and shoots upstream of each gene of the cluster. In the remaining six accessions we identify that the two intervening genes that were now outside of the cluster in the accessions had undergone an inversion including the first core cluster gene, *THAA2*, and the two intervening genes, ***Figure 25,*** green triangles. There are also regions of low homology around *CYP708A2 (THAD),* which shows a large amount of rearrangement between Col-0 and An-1.

For the marneral cluster, our EasyFig visualisation confirms that there are no large insertions in any of the accessions, ***Figure 26***. The region between *CYP71A16* and *MRN1*, where we have identified an additional gene in the Eri accession, is highly variable between all accession sequences. Notably, the gene insertion in Eri appears to have some homology to regions in Cvi and Kyo. Indicating that this gene has potentially been lost in the other accessions rather than been inserted into the Eri accession. The cluster shows low sequence homology between *CYP714A16* and *MRN*, BLAST identity of 65% with inversions.

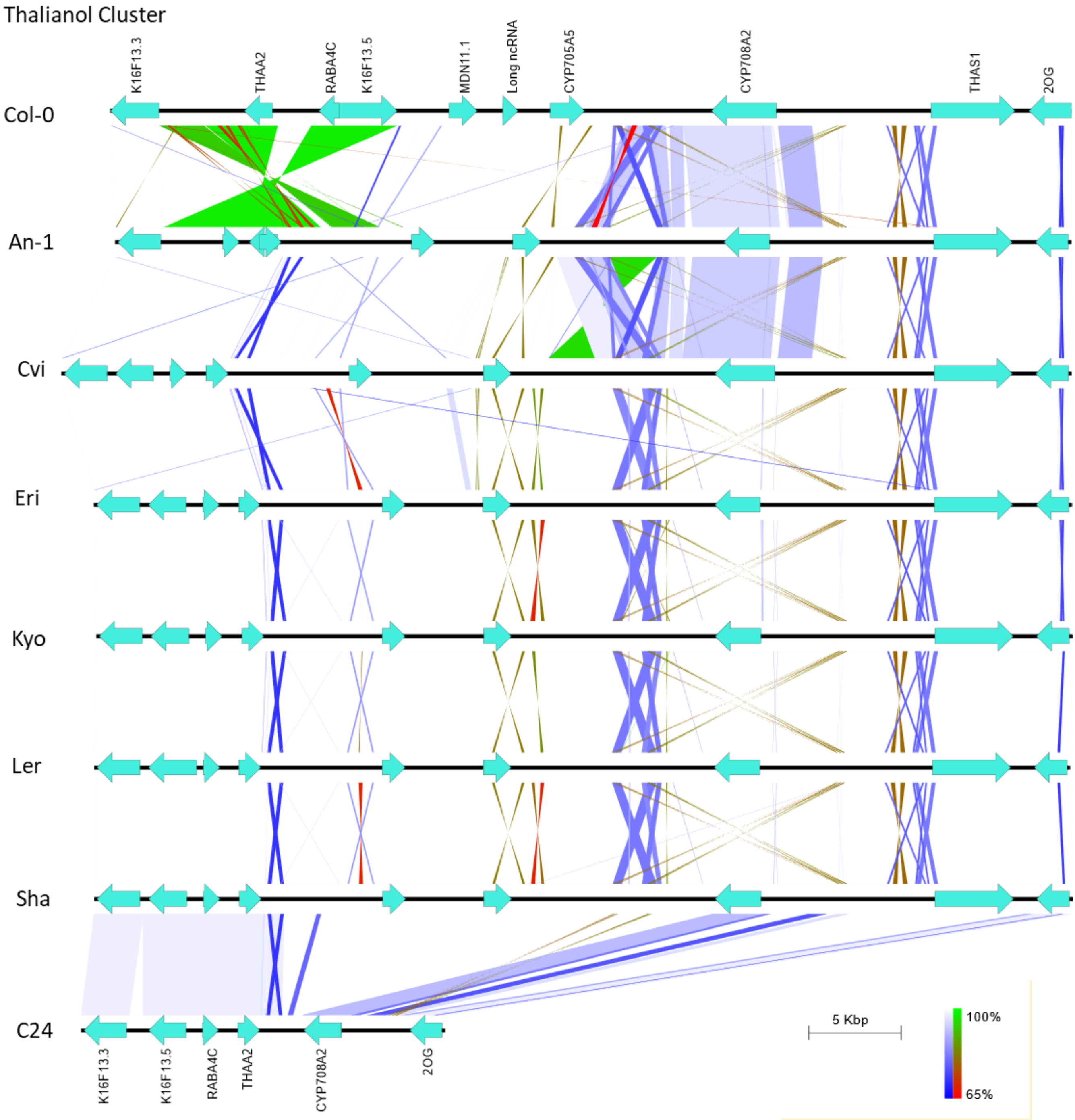


Figure 26: Pairwise BLASTn comparison of the thalianol cluster in 8 A. thaliana accessions created by EasyFig. *Comparisons are drawn pairwise in relation to the previous accession. The five thalianol gene clusters, the intervening and the flanking genes are shown by teal arrows. White regions indicate regions in the genome with a 100% BLAST sequence identity and in the same orientation, while green blocks are 100% sequence identity, however in the reverse orientation. 65% sequence homology is indicated by red and blue blocks in which red shows the reverse orientation. The gradients between the colours show a sequence homology between 65% and 100%. Crossed lines connect homologous and inverted homologous regions, homology indicated by the colour.*

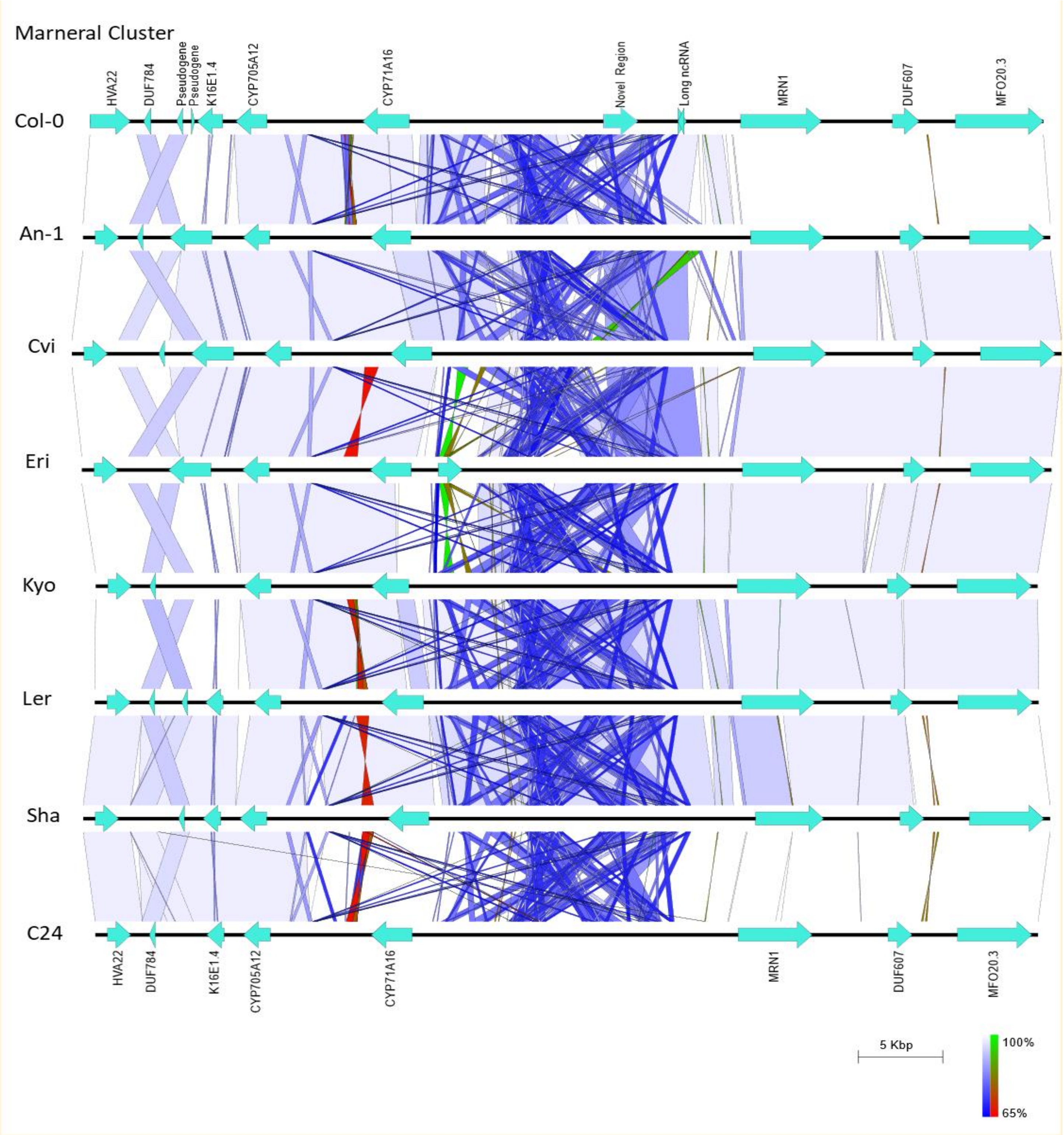


Figure 27: Pairwise BLASTn comparison of the marneral cluster in 8 A. thaliana accessions created by EasyFig. *Comparisons are drawn pairwise in relation to the previous accession. The three marneral gene clusters, the intervening and the flanking genes are shown by teal arrows. White regions indicate regions in the genome with a 100% BLAST sequence identity and in the same orientation, while green blocks are 100% sequence identity, however in the reverse orientation. 65% sequence homology is indicated by red and blue blocks in which red shows the reverse orientation. The gradients between the colours show a sequence homology between 65% and 100%. Crossed lines connect homologous and inverted homologous regions, homology indicated by the colour.*

3.2.4.4 Transcript level changes in accessions

We previously demonstrated, in sections **3.2.4.1**, that the thalianol gene cluster isn't as conserved across different *A. thaliana* accessions as previously described[153], therefore we analysed the thalianol cluster transcript levels in four of the accessions to determine if this has an impact on the co-regulation, as shown in ***Figure 27***.

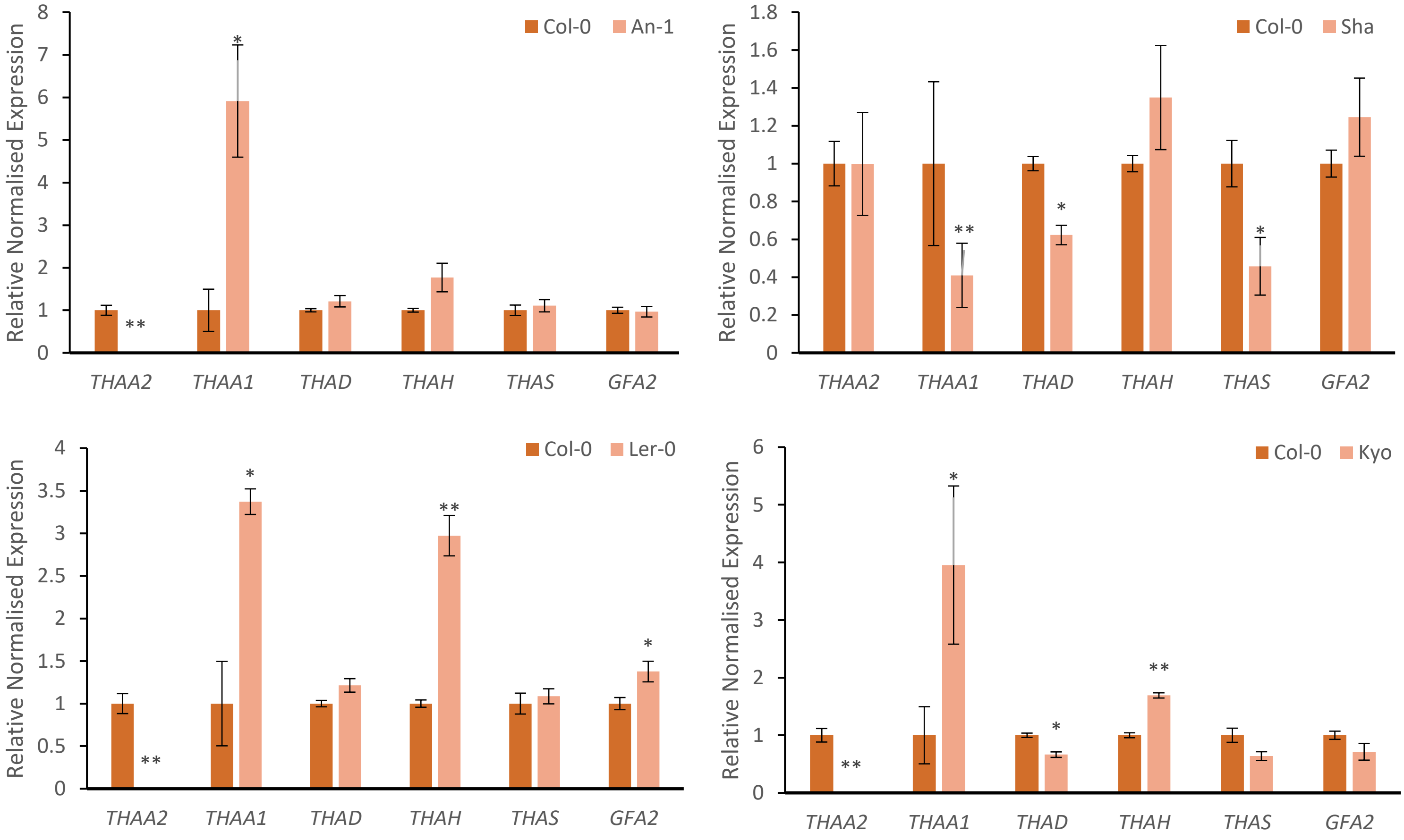


Figure 28: Transcript analysis of four A. thaliana accessions, An-1, Sha, Ler-0 and Kyo, of the five thalianol cluster genes and the flanking gene. *Gene expression was measured by qPCR and compared to the wild-type (Col-0) transcript level and normalised to 1. An internal control was used, PP2AA3 (AT5G13320). Error bars indicate standard error of the mean for three biological replicates. Statistical significance (paired t-test): *P<0.05, **P<0.01.*

All four analysed accessions show in inversion of *THAA2* and the two intervening genes, bringing *THAA2* directly adjacent to the core cluster gene *THAA1*. Despite this, the four accessions do not display the same gene expression levels when transcript number was analysed. In three of the accessions, An-1, Ler and Kyo, we detected a dramatic and significant (P < 0.05) downregulation of *THAA2*. In contrast, *THAA1* transcript levels show a pronounced increased in these three accessions. In two of the accessions, Ler and Kyo, *THAH* transcript levels significantly increased (P < 0.01). The transcript numbers for *THAH* in An-1 were also increased at a lower level, however not significantly (P = 0.061) due to variation between biological samples. Interestingly, in An-1 and Ler the *THAD* transcript levels did not show any change but in Kyo a significant decrease in transcript number was detected (P < 0.05). *THAS* transcript levels remained stable across these three accessions. In Ler, but not in the other two, *GFA2* transcript levels significantly increased (P < 0.05).

In contrast to Eri, Kyo and Ler, the Sha accession displayed a different gene expression pattern for the thalianol cluster. Here, no change in transcript levels for *THAA2* was detected and the *THAA1* gene was significantly decreased (P < 0.01) in its expression level. *THAD* and *THAS* transcript levels were also significantly reduced (P < 0.05). Transcript levels for the *THAH* and *GFA2* genes showed a trend for minor upregulation.

Three of the accessions showed similar expression profiles with some differences, while Sha showed more of a disruption of the core thalianol gene cluster genes.

3.2.1 Insertional mutagenesis using CRISPR/Cas9

Our work described so far has indicated that disrupting the thalianol gene cluster architecture in specific loci affect the co-reregulation of the cluster. Therefore, we decided to analyse how relocation of the regulatory regions of a gene within the thalianol gene cluster affects its transcriptional pattern. To do so, we attempted to integrate the *THAS* promoter and terminator regions together with two reporter genes into (a) another position within the thalianol cluster, (b) the marneral cluster and (c) in front of the constitutively expressed and non-clustered *PP2AA3* gene. As a control, we aimed to integrate the *PP2AA3* promoter and terminator regions into the same positions with the omission of the marneral cluster region. To enhance the integration of the regulatory regions into specific sites, it is required to flank the regulatory DNA fragments with approximately 1000 bp of DNA homologous to the insertion site. Thus, if the insertion does not take place at the correct loci, the insertion can still be used to assess how the architecture of the homologous arms affect the transcription of the reporter genes.

3.2.1.1 Construct selection

In order to insert the regulatory regions (promoter and terminator) of the *THAS* and *PP2AA3* gene into different loci using CRISPR/Cas9, multiple complex DNA constructs were assembled using design principles reported by Miki et and Castel et al[174,195]. Two reporter genes, eGFP and GUS, will be inserted in these constructs flanked by the respective promoters and terminators for analysis of gene expression, shown in ***Figure 28A/B***. These fragments will be referred to as the reporter fragment T1 (*THAS*) or P1 (*PP2AA3*). Homologous arms for the targeted genomic region of insertion will be integrated into the construct to increase the chances of promoter-reporter construct insertion at the correct loci, ***Figure 28C***. sgRNA will be designed to ensure the highest chance of success, based on findings by Castel et al[195], and the transcription will be regulated using the ubiquitous promoter and terminator *U6*-26 (PU6-26 and TU6-26).

3.2.1.2 Selecting sgRNA

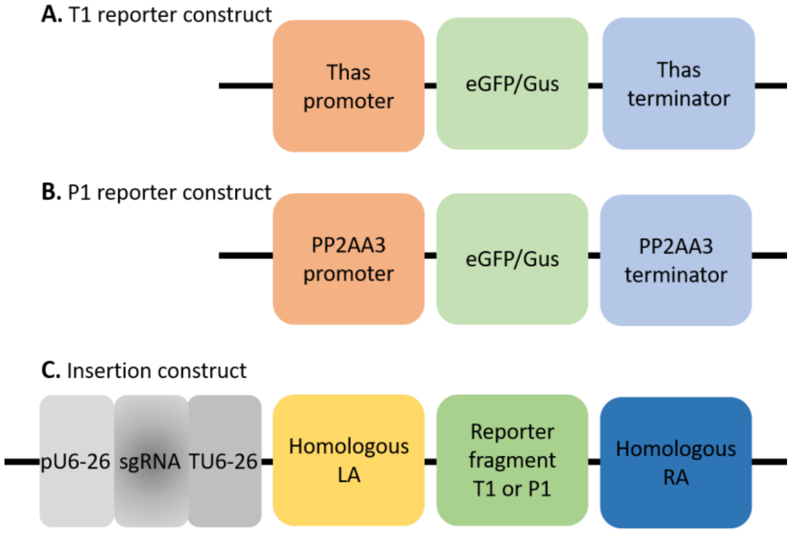


Figure 29: Final design of the constructs to be inserted into A. thaliana using CRISPR/Cas9 genome editing. *A and B depict the reporter fragment that will be inserted into the A. thaliana genome using the CRISPR/Cas9 construct C. The promoter and terminator region for THAS and PP2AA3 flank the reporter genes, GFP and Gus to analyse the potential changes in gene expression. C. Contains the sgRNA (to guide the Cas9 endonuclease to the target cut site) flanked by its regulatory regions. Homologous arms flank the reporter construct, T1 or P1, to aid the insertion.*

sgRNAs guide Cas9 nucleases to the target DNA in the nucleus. The target site must be directly preceded by a protospacer adjacent motif or PAM (-NGG) sequence element to activate the Cas9 protein. Importantly, the PAM sequence is not included in the construct sequence. Off-target binding, in which the sgRNA will bind and cleave genomic DNA with similar sequence homology, was minimised by selecting guides with the lowest number of off-target sites. On-target score and GC content of all sgRNAs are shown in ***Table 23.***

Table 23: sgRNAs selected for targeted genome editing for insertions and deletions.

Target site	Position of cleavage	Guide RNA **PAM**	Off target sites	On target	% GC
Chr3 deletion	Chr3:4251768	GCTGGAGCATCATCTCCGGG**TGG**	1	0.8678	65
Chr3 deletion	Chr3:4290494	CCCCATCTATAACCGAGCCG**AGG**	0	0.8263	60
THAS deletion	Chr5:19488206	TCTATACCTTTGGAGCCGAA**GGG**	7	0.8924	45
THAS deletion	Chr5:19521379	AAGTTGGGGAGATATCCCCG**AGG**	3	0.5915	55
THAS insertion	Chr5:19438321	TTGTAACTAACGAGTAACGA**CGG**	4	0.7475	35
PP2AA3 insertion	Chr1:4536129	TCTCTTACCGAATAAAAGCG**TGG**	6	0.9318	40
MRN insertion	Chr5:17028589	TGCATACGACCCCAATCGAA**TGG**	2	0.5040	50

3.2.1.3 Construct assembly

To assemble complete constructs containing sgRNA, homologous arms, regulatory regions and reporter genes multiple cloning steps using different cloning techniques were applied. Our workflow is shown in ***Figure 29***. In the following, we will outline the results of each experimental step in the assembly.

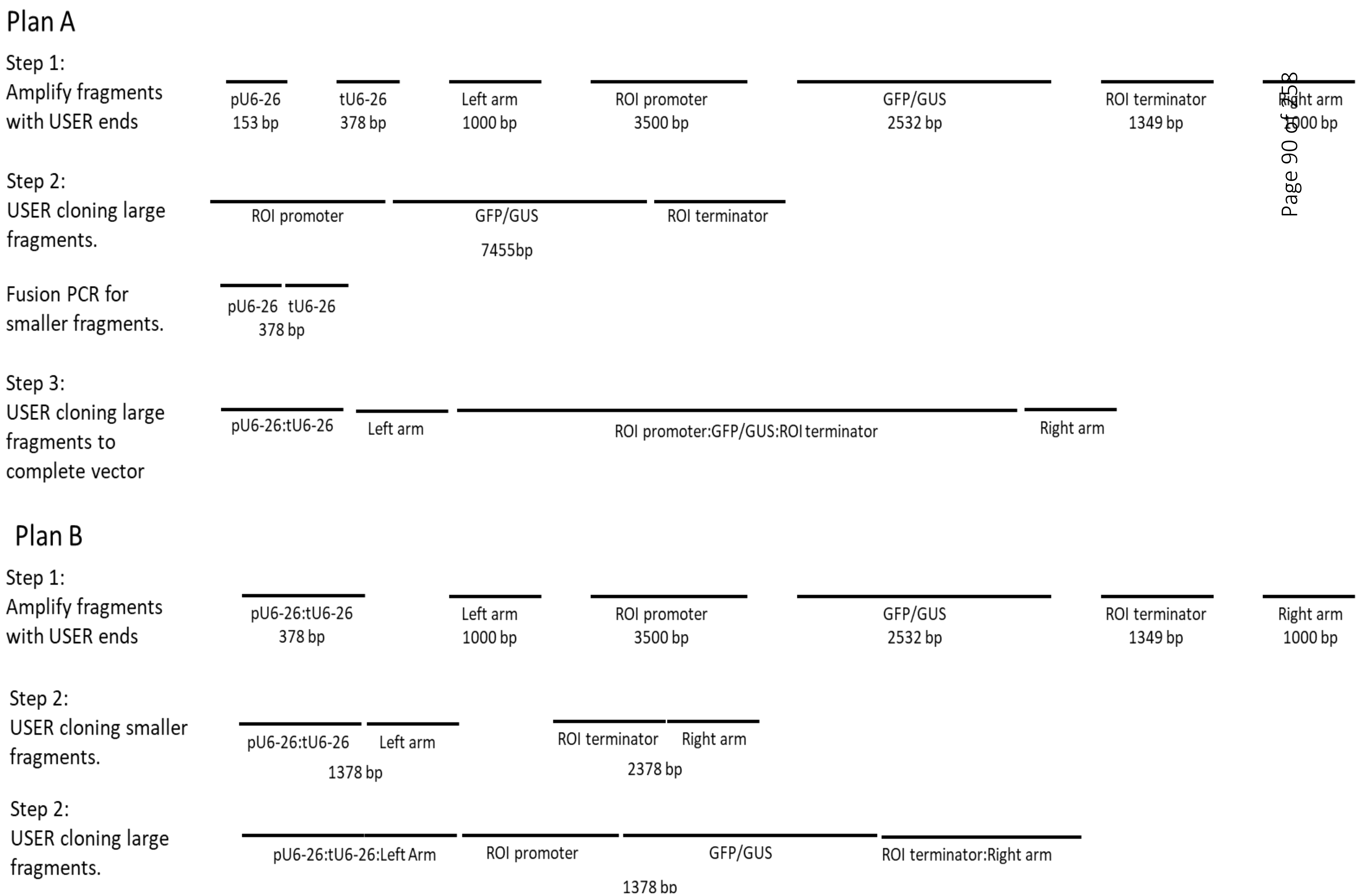


Figure 30: Outline of two plans for creating the insertion constructs. *Plan A shows the initial plan we used, step 1 is to amplify all DNA frgaments with their USER ends, step 2 is creating the reporter construct by ligating DNA fragments with USER cloning and step 3 is a single USER cloning step to ligate all the fragments together. Plan B is the altered plan to create the same construct. Step 1 reamplifies fragemtns with new USER cloning overhangs, Step 2 requires two USER cloning steps to ligate together similar sized DNA fragments, and finally Step 3 ligates DNA frgaments of a similar size to create the final insertion construct.*

)

3.2.1.3.1 Amplification of fragments for cloning

Using the KAPA PCR protocol, each individual DNA fragment was amplified by KAPA PCR using primers with USER reaction specific overhangs. All individual DNA fragment were successfully amplified (***Figure 30***).

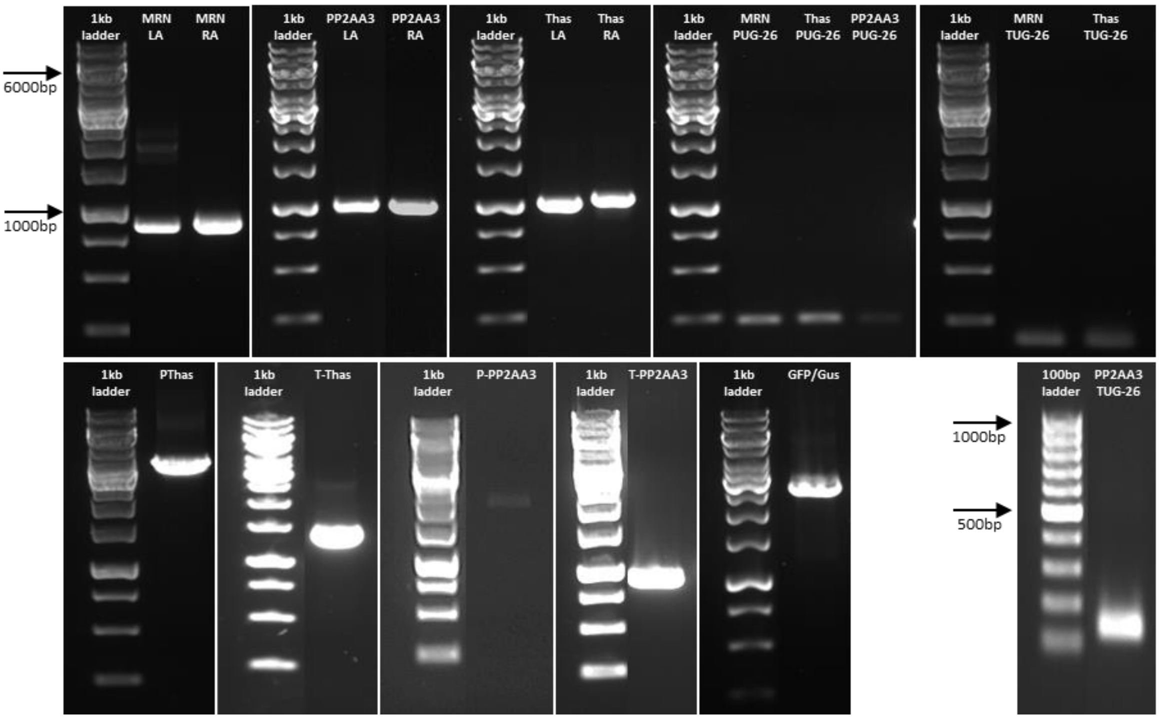


Figure 31: Agarose gel following amplification of DNA fragments for USER cloning. *Each gel shows a fragment of the correct size (brightest band) and the corresponding DNA ladder for confirmation (example of band sizes shown).*

Next, we carried out Fusion PCR to combine the *U6-26* promoter (PU6-26), sgRNA and U6-26 terminator (TU6-26) using the KAPA protocol. Three U6-26 promoter and U6-26 terminators were fused using the KAPA fusion protocol with a 100% success rate, shown in ***Figure 31.***

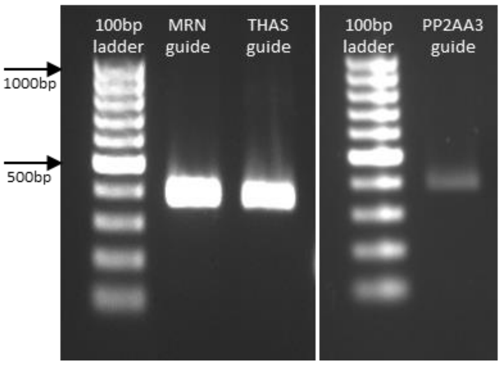


Figure 32: Agarose gel following fusion PCR to ligate to fragments together to create three sgRNAs. *Each gel shows the correct fragment size, and the corresponding 100 bp ladder for confirmation (example of band size shown).*

3.2.1.3.2 Creation of reporter construct by USER cloning

To fuse gene promoter, GFP/Gus and gene terminator together, USER cloning was carried out using a standard 1-hour ligation protocol and transformation into *E. coli*. Transformed colonies were screened for successful construct ligation by single colony DreamTaq PCR.

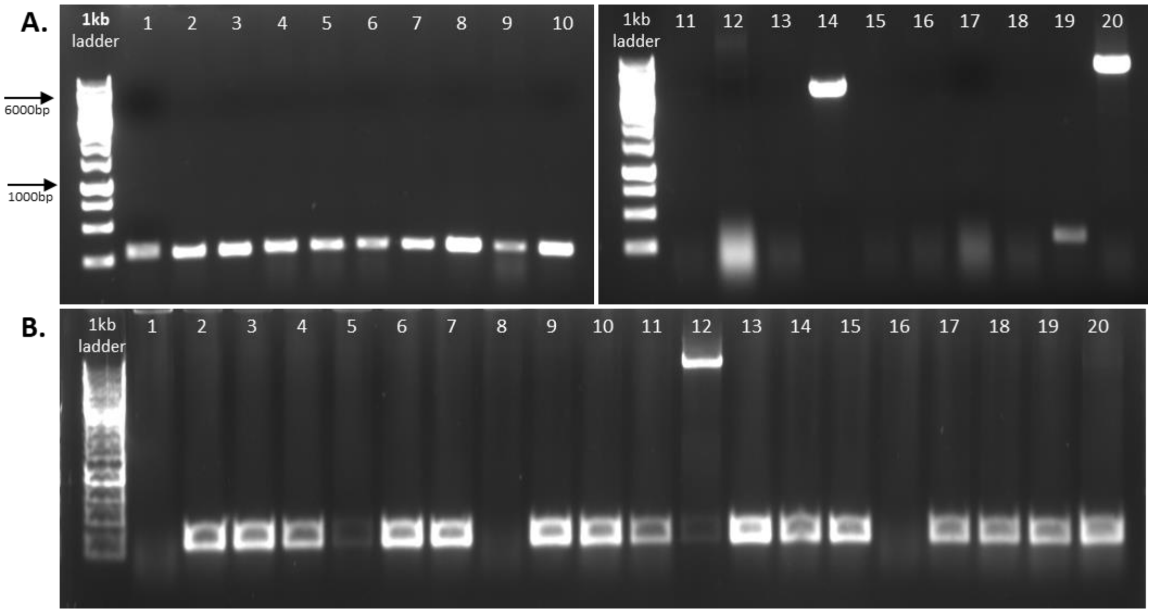
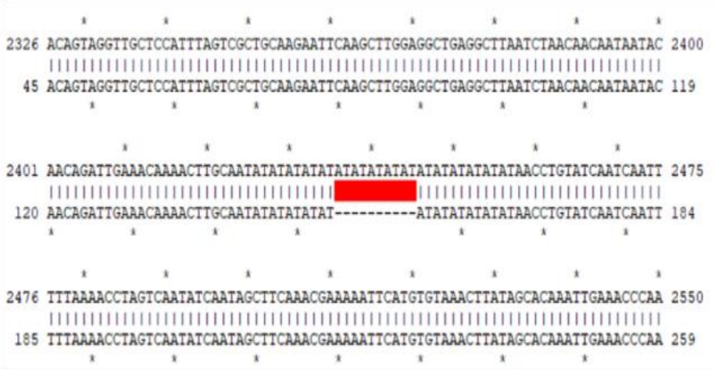


Figure 33: Colony screening PCR following USER cloning. *Single colony screening of P1 (A) and T1 (B) to confirm the insertion of the correct fragment size was carried out. One colony, 20, was positive for P1 and one colony was positive for T1, 12. The 1 kb ladder was used for size determination, example of band sizes shown.*

We obtained a positive PCR signal for 1 colony of T1 (colony 12) and 1 colony of P1 (colony 20) (***Figure 32***). Verification of inserts by DNA sequencing showed 100% sequence homology to the template DNA for T1-12 and P1-20. Of note, an AT- rich region of the *PP2AA3* promotor region in P1 showed fewer AT nucleotides compared to the reference genome. This is likely due to miss-annotation in the reference genome or problems during sequencing (***Figure 33***).



```
         *         *         *         *         *         *         *         *
2326 ACAGTAGGTTGCTCCATTTAGTCGCTGCAAGAATTCAAGCTTGGAGGCTGAGGCTTAATCTAACAACAATAATAC 2400
     |||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
  45 ACAGTAGGTTGCTCCATTTAGTCGCTGCAAGAATTCAAGCTTGGAGGCTGAGGCTTAATCTAACAACAATAATAC 119
          *         *         *         *         *         *         *

              *         *         *         *         *         *         *
2401 AACAGATTGAAACAAAACTTGCAATATATATATATATATATATATATATATATATATAACCTGTATCAATCAATT 2475
     |||||||||||||||||||||||||||||||||||██████████||||||||||||||||||||||||||||||
 120 AACAGATTGAAACAAAACTTGCAATATATATATAT----------ATATATATATATAACCTGTATCAATCAATT 184
     *         *         *         *                   *         *         *

         *         *         *         *         *         *         *         *
2476 TTTAAAACCTAGTCAATATCAATAGCTTCAAACGAAAAATTCATGTGTAAACTTATAGCACAAATTGAAACCCAA 2550
     |||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
 185 TTTAAAACCTAGTCAATATCAATAGCTTCAAACGAAAAATTCATGTGTAAACTTATAGCACAAATTGAAACCCAA 259
          *         *         *         *         *         *         *
```

Figure 34: Sequencing alignment of the template (top row) and plasmid DNA (bottom row) showing the missing AT repeat region*. The P1 construct shows mismatches in the PP2AA3 promoter region shown by the red block.*

3.2.1.3.3 Creating the whole construct by a one-step USER reaction

After successful ligation of the T1 and P1 reporter constructs, we next attempted to ligate this segment together with the other DNA fragments previously amplified. The USER cloning protocol was carried out with a 1-hour incubation period followed by transformation into *E. coli* by electroporation to create T2A, T2B, T2C, P2A and P2B (***Table 24*** for naming constructs). Colony PCR screening showed that the assembly of all fragments was unsuccessful.

Table 24: Description of construct shortened names. *A table showing the shortened names given to each construct and their use.*

Construct shortened name	Description of construct use
T1	Reporter construct containing *THAS* promoter and terminator region and the eGFP/Gus.
P1	Reporter construct containing *PP2AA3* promoter and terminator region and the eGFP/Gus.
T2A	Insertion of T1 into the *PP2AA3* loci.
T2B	Insertion of T1 into the marneral loci.
T2C	Insertion of T1 into the thalianol loci.
P2A	Insertion of P1 into the thalianol loci.
P2B	Insertion of P1 into the *PP2AA3* loci.
D*THAS*	Deletion of the region downstream from the thalianol cluster.
DChr3	Deletion of the region in chromosome 3.

Although there were bands on the agarose gel, the band size indicated a closed vector with no insert (***Figure 34***). The restriction digest carried out also showed DNA fragments with size corresponding to the vector only.

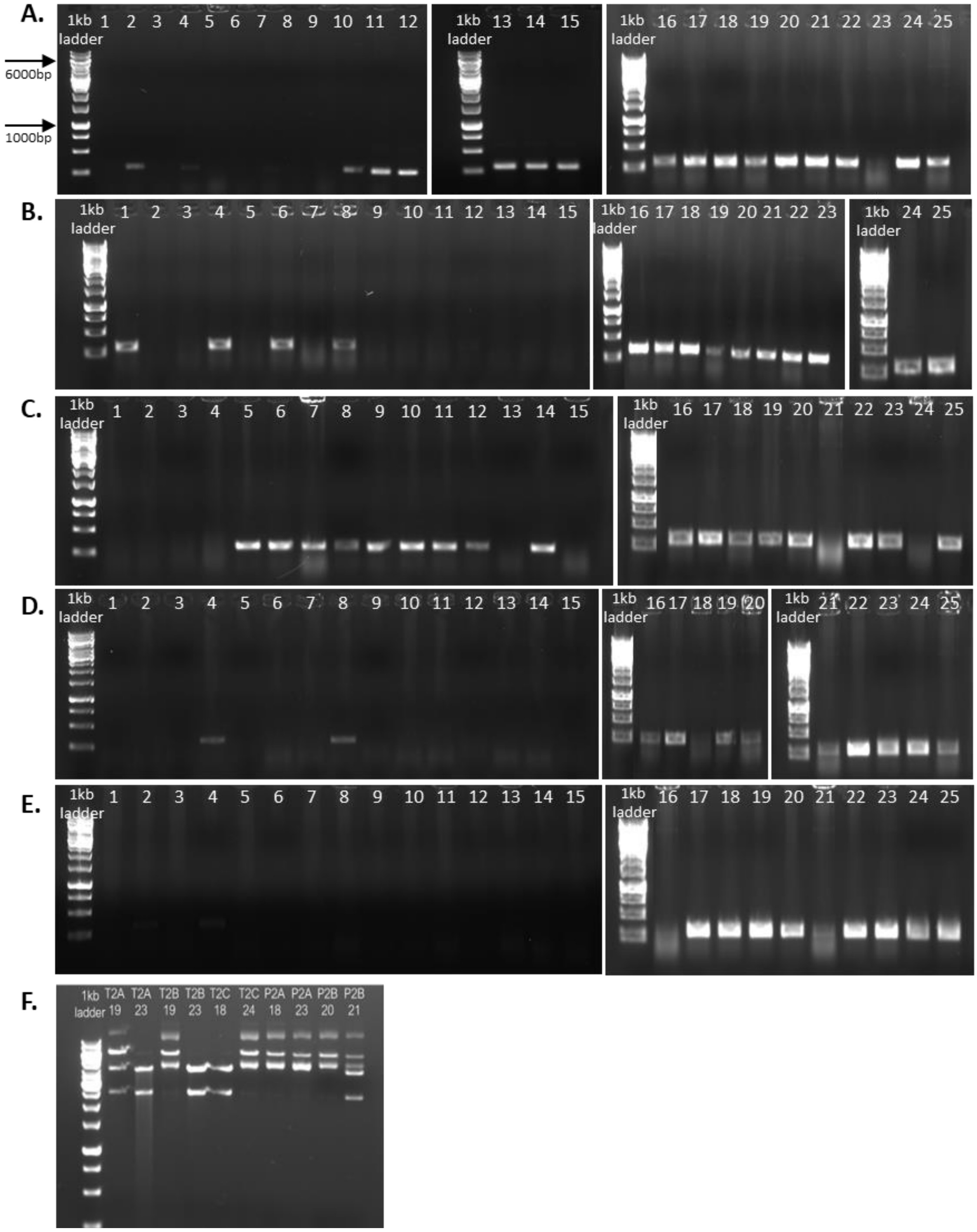


Figure 35: Agarose gel electrophoresis of colony PCR following USER cloning, and HindIII digest of the plasmid. *The T2A (A), T2B (B), T2C (C), P2A (D) and P2B (E) constructs were attempted to be inserted into a plasmid by USER cloning. No DNA bands of expected size were detected by single colony PCR. A HindIII digest was carried out to assess the plasmids in (F). DNA band sizes for a closed empty vector were detected. Expected size for A-E is a fragment size of 8000 bp. Expected digest band sizes (F): 7834, 4652 and 4305 bp.*

Therefore, we decided to adjust our experimental conditions. All DNA fragments were re-amplified and this time gel purified. The ligation period was increased to 16-hours. However, as exemplified in ***Figure 35***, no successful assembly of target constructs was achieved. After 10 transformation attempts and screening of on average 60 colonies per construct no positive plasmids were identified. A selection of the colony screening and HindIII digests to check plasmids is shown in ***Figure 35C***.

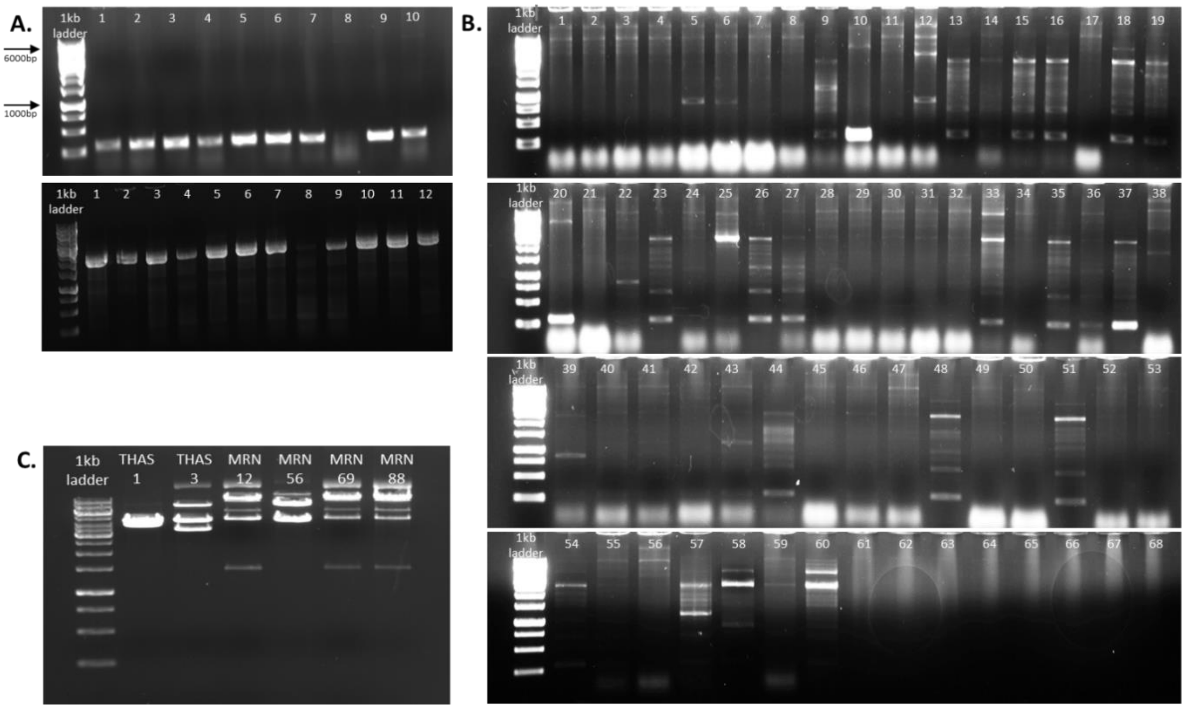


Figure 36: Agarose gel electrophoresis of colony PCR screen and HindIII plasmid digest. *Insertions were attempted by USER cloning to create constructs T2B (A) and T2C (B), however for A no bands were found and for B multiple bands were found in the colony PCR indicating an unsuccessful insertion. A HindIII digest was carried out on the plasmids (C) which indicated a closed vector with no insertions. Expected size for A-B is a fragment size of 8000 bp. Expected digest band sizes (C): 7834, 4652 and 4305 bp.*

3.2.1.3.4 Creating the constructs using smaller USER reactions

In principle, the USER cloning technique should allow for the fusion of multiple DNA fragments into a vector. However, as our attempts to generate our large DNA cassettes by using a single USER cloning step had been unsuccessful, we decided to break the assembly down into shorter ligation steps, using DNA fragments of similar sizes (***Figure 29***). We focused on inserting the *THAS* regulatory elements into the three locations, constructs T2A, T2B and T2C, from here on out.

The first step required ligating smaller fragments into the USER vector. Two USER reactions were carried out for each construct to ligate the sgRNA and promoter fragment to the respective left homologous arm and the T-*THAS* to the respective right homologous arm, ***Figure 36***. For each reaction we screened 20 colonies using DreamTaq single colony PCR. For one ligation reaction 18 colonies were positive and for another ligation 19 colonies were positive. For four ligation reactions, all 20 colonies showed a positive result for complete fragment ligation. Three positive colonies were picked at random for each fragment. The respective plasmids were extracted and sent for sequencing. At least 2 out of three plasmids were confirmed to contain the expected sequence for each fragment. These plasmids were used for amplification for the next USER cloning step.

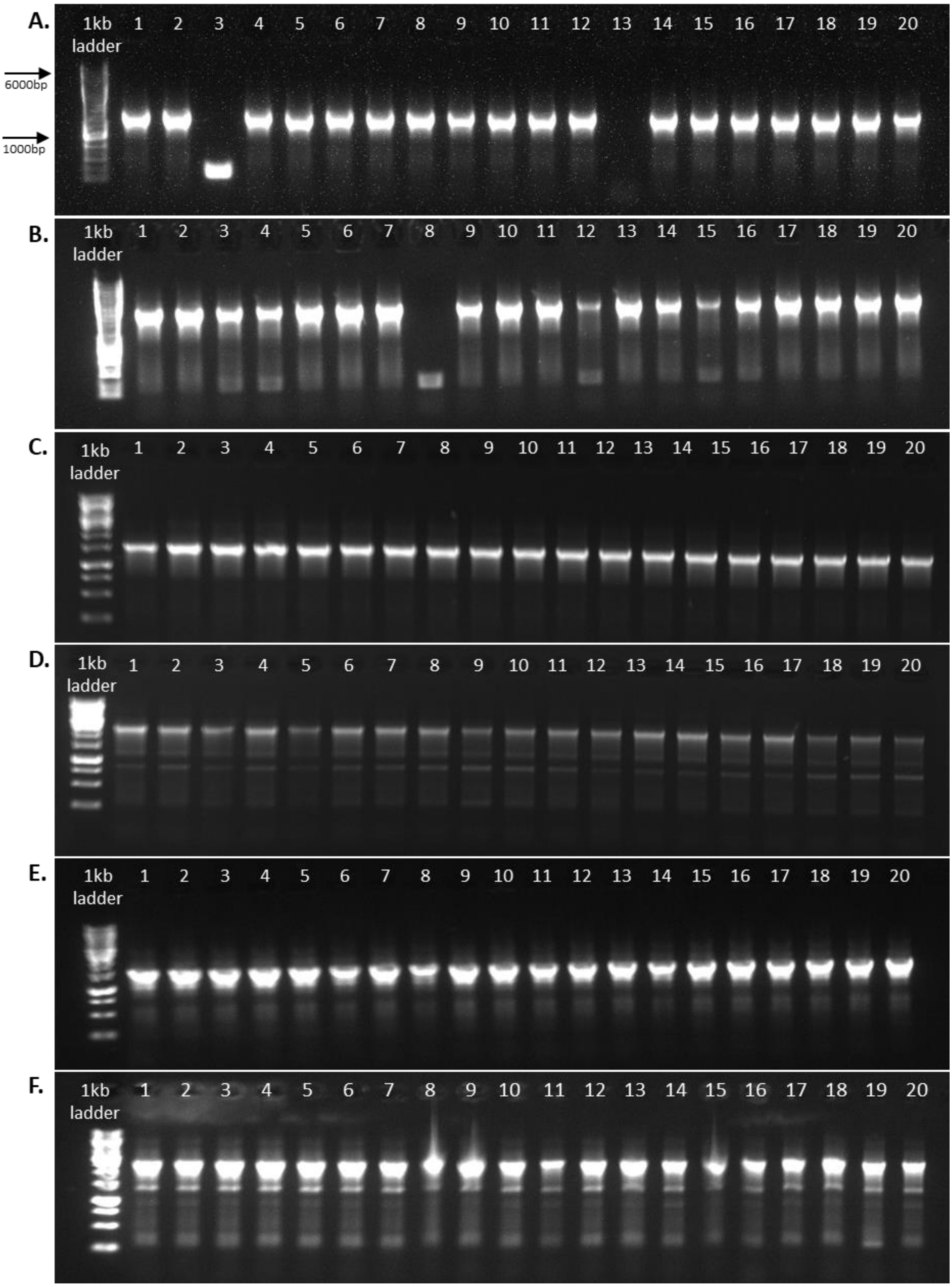


Figure 37: Colony PCR on an agarose gel following shorter ligation USER cloning reactions. *Shorter USER reactions were carried out to ligate the sgRNA to MRN LA (A), T-THAS to MRN RA (B), sgRNA to THAS LA (C), T-THAS to THAS RA (D), sgRNA to PP2AA3 LA (E) and T-THAS to PP2AA3 RA (F). All reactions show successful results, indicated by the brighted band of the correct size on the agarose gel.The 1kb DNA ladder was used for band sizing. Expected band size 1500-2000 bp.*

The second round of USER cloning used the previously ligated fragments (sgRNA ligated to the left homologous arm and T-*THAS* ligated to the right homologous arm) as a template to create three of the full constructs.

For T2A assembly, one USER reaction was carried out. 16 out of 47 analysed colonies showed AGE banding with multiple bands of which one band aligned with the expected size of vector and insert, ***Figure 37***. To get a colony with a single band of the correct size, 4 colonies were spread onto new plates and the colony PCR was repeated, ***Figure 37B.*** Two positive colonies with a single band of the expected vector plus insert size were selected for plasmid extraction and sequencing. One plasmid (T2A-25) showed 100% homology to the targeted gene cassette and put forward to the next experimental steps.

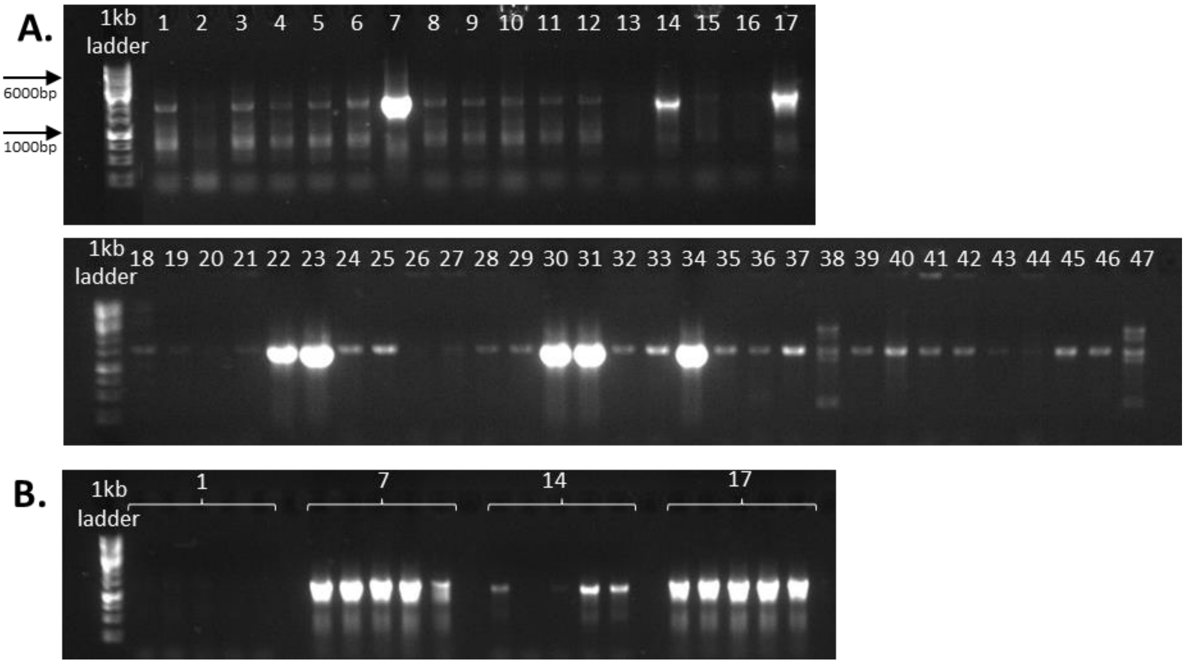


Figure 38: Agarose gel PCR of the T2A insertion colony PCR following USER cloning. *Initial single colony PCR to confirm T2A had successfully inserted into the vector by USER cloning (A) showed positive bands of the correct size (confirmed by the 1kb ladder). However, multiple bands were obtained by most colonies (A) Colony PCR on sub-cultured colonies show single bands of the expected size (B). Expected size 1500 bp.*

In the first round of USER cloning for T2B assembly, 49 colonies were screened. The AGE banding suggested a closed vector for 20 colonies, ***Figure 38***. For 6 colonies a two-band pattern was detected by AGE. As one of these bands showed the expected vector plus insert size, we deduced that the analysed colonies contained both vector only and vector plus insert plasmids. To obtain colonies with distinct plasmids, colonies were spread onto new plates and colony PCRs repeated. Three colonies were selected to be spread onto new plates and have the colony PCR repeated. Indeed, for one colony we detected vector containing insert PCR products only. However, subsequent DNA sequencing showed that the vector had multiple sequence misalignments. Therefore, we decided to repeat the USER cloning. This time, 50 colonies were screened and 69% of colonies show positive vector plus insert PCR results. Four colonies were randomly selected for plasmid extraction and sequencing. Of these, three showed sequence misalignments and one, colony 17 (T2B-17), showed 100% sequence homology to the target sequence. This plasmid was used for further experiments.

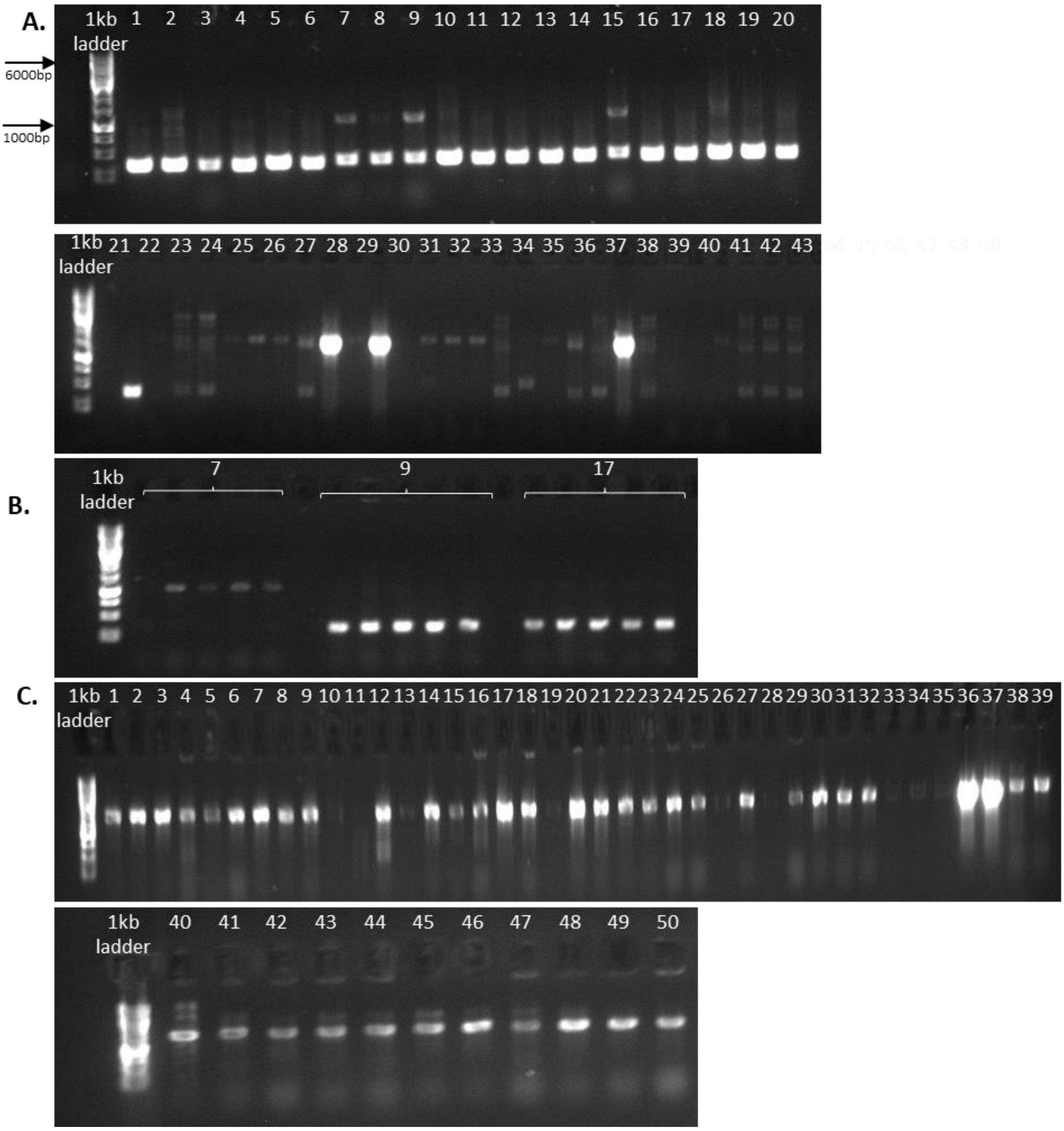


Figure 39: Agarose gel for single colony PCR. *The first colony PCR (A) to confirm insertion T2B into the vector by USER cloning showed multiple bands. 3 out of 20 repeat colony PCRs on sub-culture colonies show DNA fragment of expected size (B). Results of colony PCRs of second USER cloning attempt to vector insertion of T2B (C). Multiple colonies with expected banding were identified. Sizes of bands confirmed by the 1kb ladder, expected size 1300 bp.*

T2C required a single USER reaction, and 38 colonies were screened, ***Figure 39***. 38 colonies were screened and 97% of colonies show positive vector plus insert PCR results. Four colonies were randomly selected for plasmid extraction and sequencing, two showed sequencing misalignments and two showed 100% sequence homology, colonies 17 and 22 (T2C- 17 and T2C-22). These two plasmids were used in further experiments.

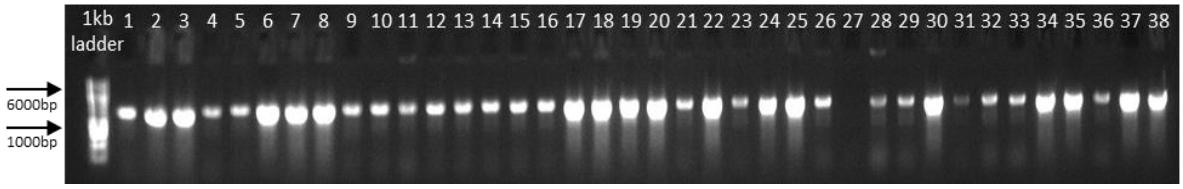


Figure 40: Agarose gel for single colony PCR for T2C. *Insertion of T2C into the USER vector was successful, indicated by a postitive band on the gel at the correct size. 1kb ladder used to confirm band sizes, expected size 1500 bp.*

3.2.1.4 CRISPR/Cas9 plant transformation and selection

Plasmids T2A-25, T2B-17 and T2C-17 were transformed into *A. tumefaciens*. All three plasmids contain antibiotic resistance genes against rifampicin and gentamycin. As such, positively transformed *A. tumefaciens* strains were identified by selection against rifampicin and gentamycin.

Next, transformed *A. tumefaciens* strains were used for standard *Agrobacterium*-mediated transformation of *A. thaliana*. Approximately 40 plants were transformed for each construct. After harvest of seeds, transformed seeds were selected by their red seed coat, ***Figure 40***, due to the vector inserted containing red fluorescent protein (RFP) which is expressed in the seed coat.[215]

Less than 1% of seeds collected showed a red seed coat. 10 seeds were planted from *A. thaliana* plants transformed with T2A containing *A. tumefaciens*, 25 seeds for T2B and 5 seeds for T2C. 5, 17 and 2 seeds germinated for each respective construct.

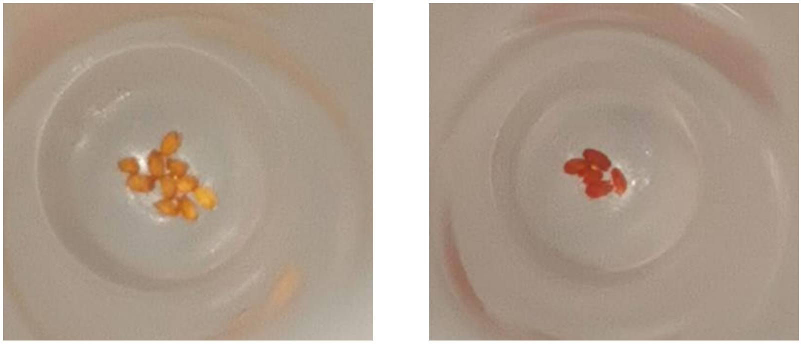


Figure 41: Wild-type seed and example of a seed expressing red fluorescent protein. *Left, brown coloured wild-type seeds and right, red coloured transformed seed.*

After the establishment of healthy rosettes, leave samples were taken from all plants that had germinated. To confirm the integration of our gene cassettes, we performed the following three PCR reactions. The first PCR identified if GFP was inserted, out of 24 plants tested half showed a positive result for containing the GFP. The second confirmed that the *THAS* promotor was inserted ligated to the GFP, 10 out of the 24 showed a positive result. The third PCR extended from the left homologous arm to the right homologous arm, 7 of which tested positive. This reaction allowed for the reporter and regulatory region to be amplified, meaning that either the insert is on the correct position or that the homologous arms have also been inserted into the genome, ***Figure 41*** and ***Table 25***. Plants with positive results for two or three PCRs were used for subsequent experiments. Although we are yet to confirm the integration occurred in the correct position, for the purposes of this section we will assume they are based on the use of the CRISPR/Cas9 integration methodology, and the findings presented.

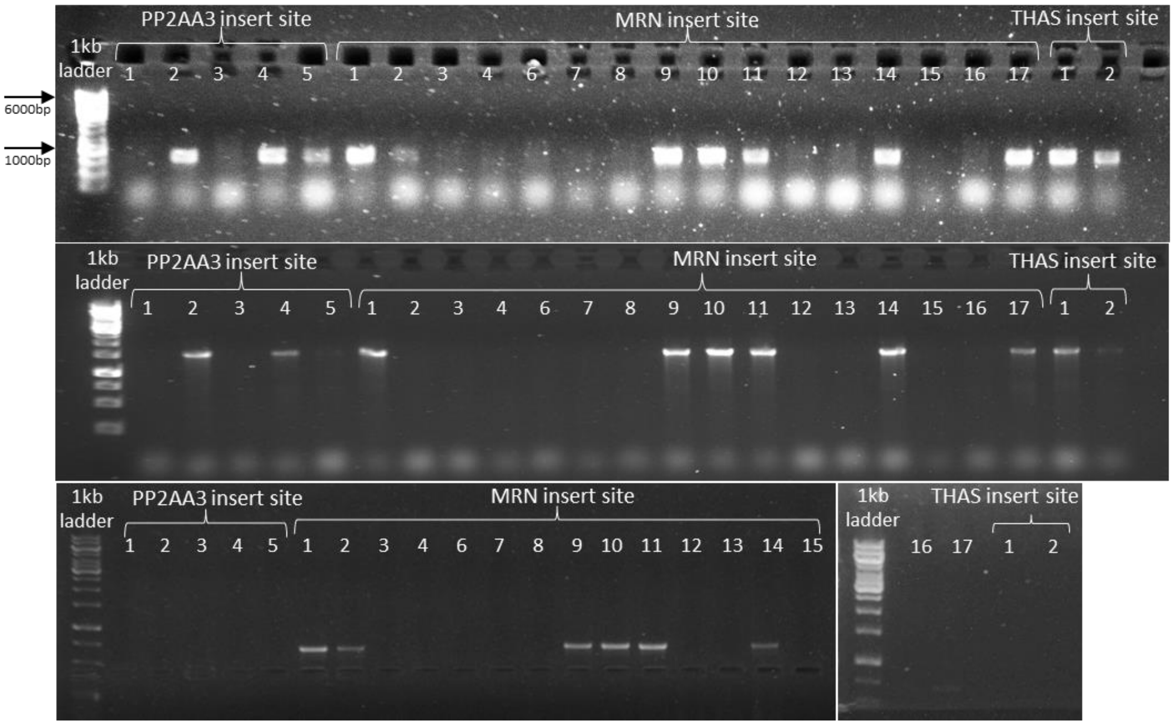


Figure 42: Agarose gel of plant genomic DNA PCR to identify gene cassette insertion. *Three PCRs were carried out on transformed plants (those with red seed coating) to confirm gene cassette integration. A bright band indicated a positive result. Sizes are confirmed with the 1kb DNA ladder, and an example of band size is shown. Expected band sizes: 1000 bp (top), 1500 bp (middle) and 700 bp (bottom).*

Table 25: Overview of PCRs carried out to identify the successful insertions . *A tick indicates that a positive band of the correct size was detected by PCR reaction.*

Insert	T2A					T2B																	T2C	
Plant number	1	2	3	4	5	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	1	2
GFP primers		✓		✓	✓	✓								✓	✓	✓			✓			✓	✓	✓
Promotor to GFP		✓		✓		✓								✓	✓	✓			✓			✓	✓	✓
LA to RA						✓	✓							✓	✓	✓			✓			✓		

3.2.1.5 Cluster regulatory regions display differing cell expression profiles

To investigate whether the different flanking sequences in each construct would influence the reporter expression pattern, we analysed GUS staining and GFP expression in young seedlings.

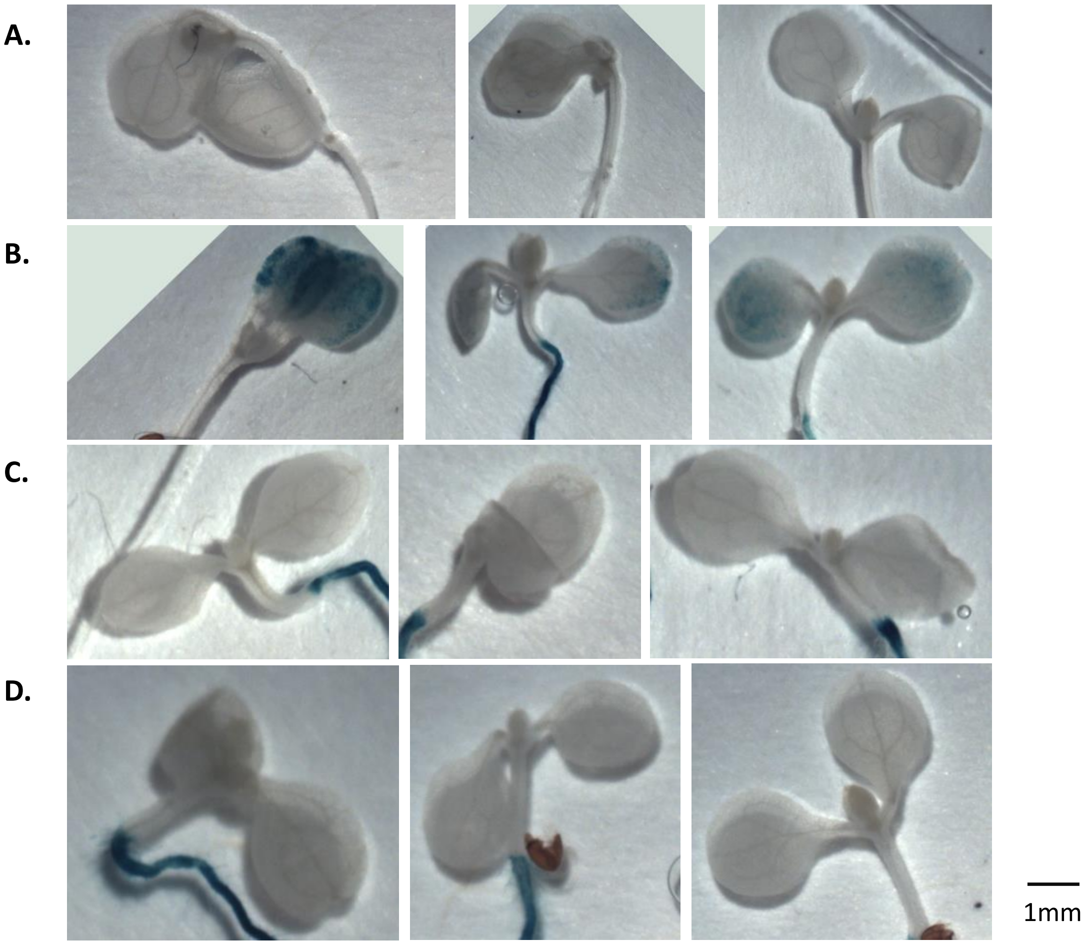


Figure 43: Gus staining of A. thaliana leaves. *Staining of A. thaliana plants to show the expression of Gus. In the wild-type (A) no Gus staining is detected. Plants transformed with the THAS reporter construct into the PP2AA3* site (B) *show expression of Gus in the leaves but not the hypocotyl. Plants transformed with the THAS reporter into the THAS (C) and MRN (D) site show no Gus expression in the leaves or hypocotyl.*

Gus staining, showed that plants containing promoter-reporter gene cassettes inserted into the *PP2AA3* sites expressed Gus in the primary and lateral root and cotyledons and not in the hypocotyl, ***Figure 42 and Figure 43***. In contrast, plants containing promoter-reporter gene cassettes into both the *MRN* and *THAS* sites showed an expression profile constrained to the root with strong signal enrichment at the root tip ***Figure 42 and Figure 43***. Notably, an expression gap between the root tip and remaining root were detected in plants with both constructs, ***Figure 43***.

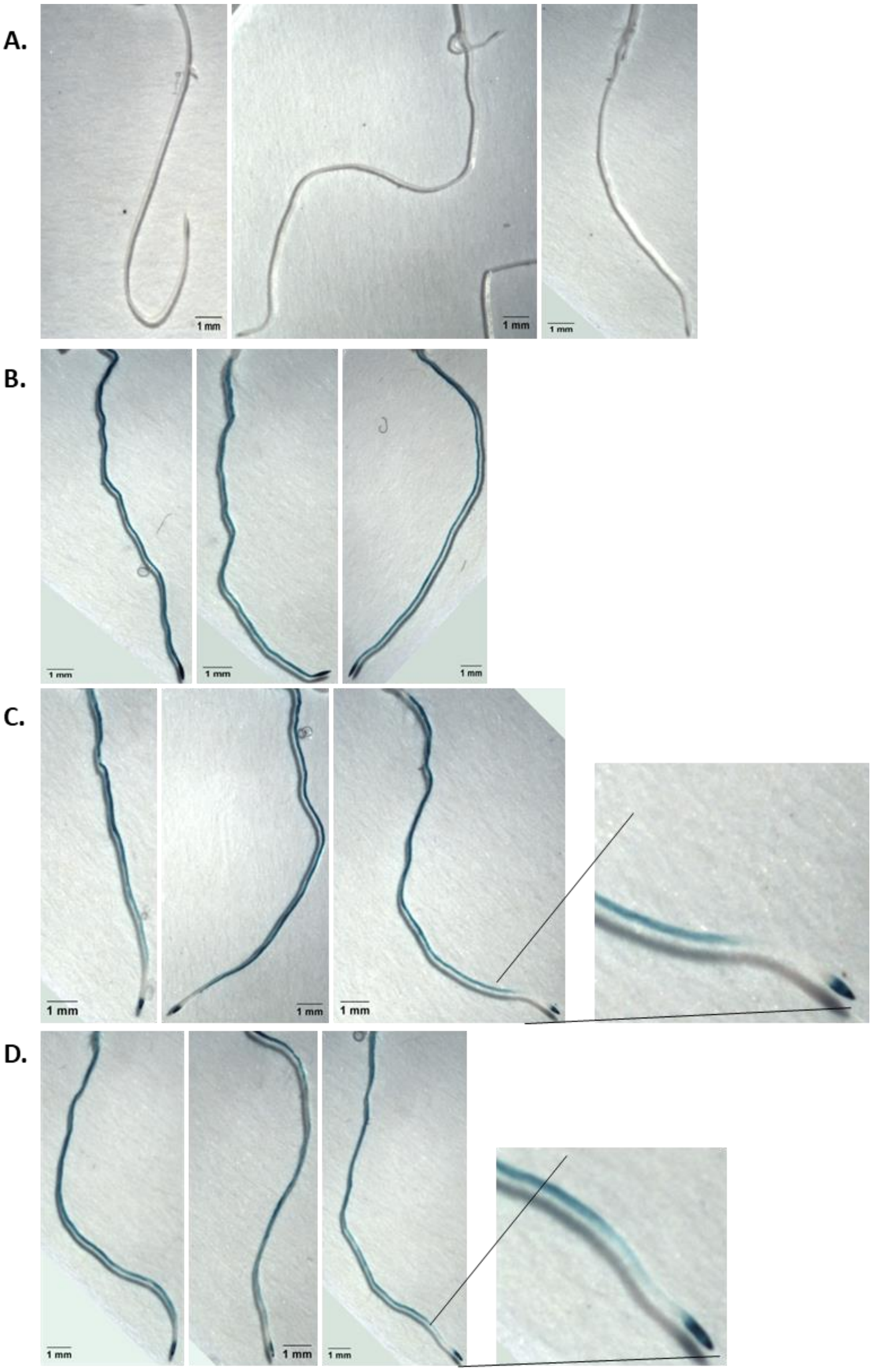


Figure 44: Gus staining of A. thaliana roots. *The wild-type no Gus staining was detected (A). Plants transformed with the THAS reporter construct into the PP2AA3 site Gus expression throughout the root and an enhanced Gus signal in the root tip. Plants transformed with both the THAS reporter construct into the THAS (C) and MRN (D) site showed a signal throughout the root with an intense signal in the root tip, however a short segment without signal was detected between the tip and the rest of the root (zoomed image).*

GFP analysis allows for a more in-depth imaging of cell specific expression, ***Figure 44***. The integration into the *PP2AA3* site expressed GFP throughout all the cells in the root, with greater expression in the root cap, with lower expression in the columella root cap. Throughout the root it appears that expression is stronger in the epidermis and cortex than in the stele. In the *MRN* and *THAS* insertion site transformed plants, at the root tip expression is contained to the lateral and the columella root cap extending to the epidermis. The rest of the root has uniform expression throughout.

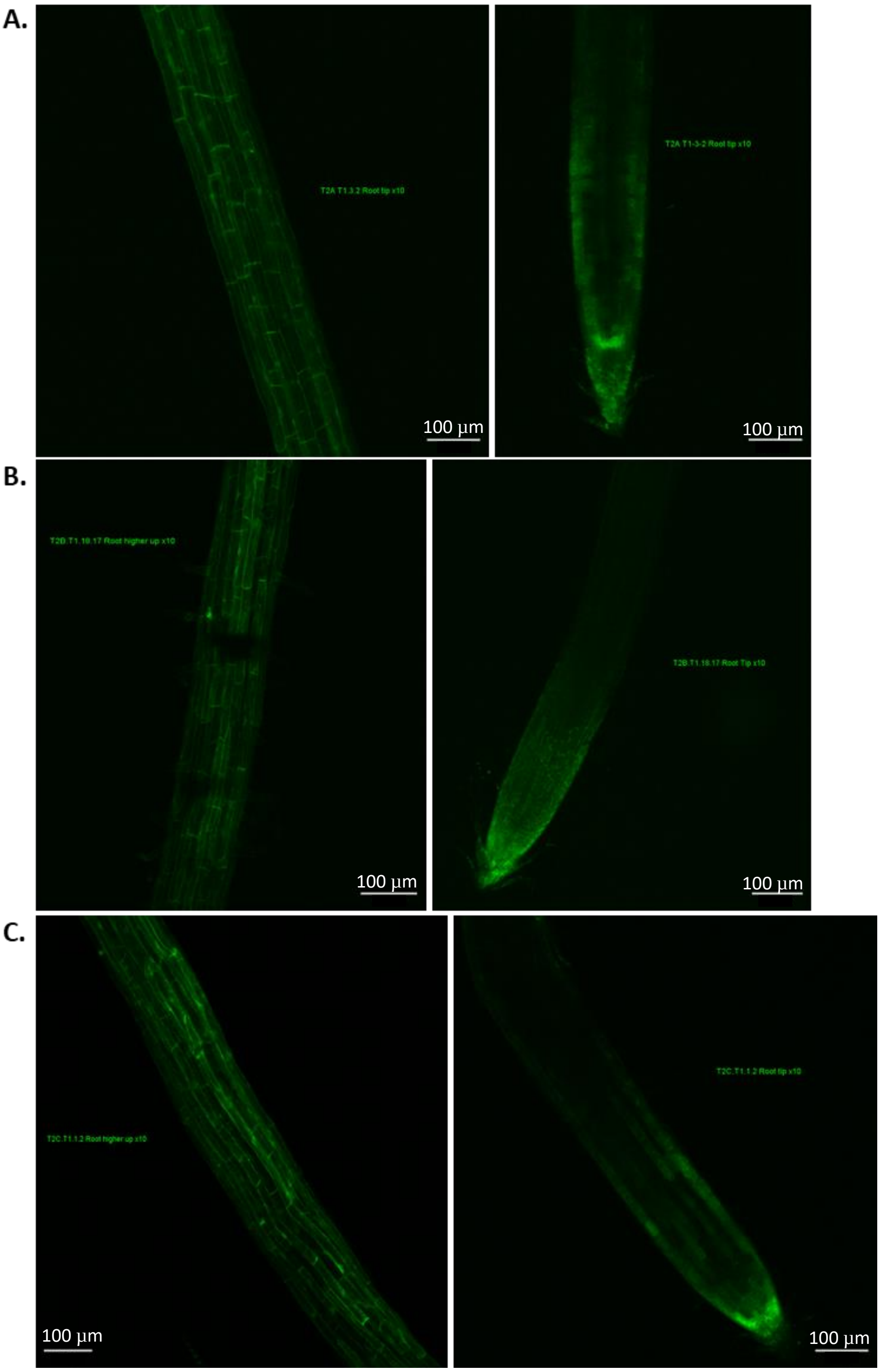


Figure 45: GFP analysis of the mutant lines. *GFP is detected throughout the root cells in all three mutant plant types. A. shows the reporter construct inserted into the* PP2AA3 *site, whereas B is into THAS and C into MRN site.*

3.2.2 CRISPR deletions

3.2.2.1 Guide selection

We have described that the thalianol gene cluster interacts with a region downstream of the cluster when the cluster is inactive[167]. Here, we aim to identify if the downstream region is imperative to the regulation of the thalianol gene cluster by attempting to delete a 33 kb region using CRISPR/Cas9 technology. Alongside this, to confirm the principles of the technologies used, a second deletion in an important region for the 3D genome organisation in chromosome 3 will also be a target for deletion. Two sgRNAs were selected for each region to delete using the conditions and variables as previously described for insertions.

3.2.2.2 Assembly of deletion cassettes

For each deletion cassette two sub-constructs were assembled, each containing, two sgRNAs flanked by the U6-26 promoter and terminator, ***Figure 45***. Individual DNA fragments were amplified with primers designed to support a fusion PCR. sgRNA ligated to the pU6-26 and TU6-26 were amplified as previously described using fusion PCR, AGE shown in ***Figure 46*** of successful amplification and ligation.

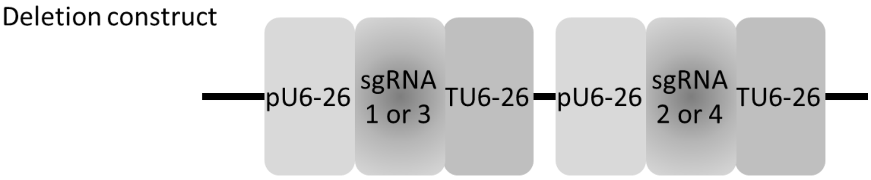


Figure 46: Final design of the constructs to be inserted into A. thaliana to cause a deletion using CRISPR/Cas9 genome editing. *Two sgRNA will be inserted into A. thaliana flanked by the U6-26 promoter and terminator with the aim to create two double stranded breaks and cause a large deletion in the genome.*

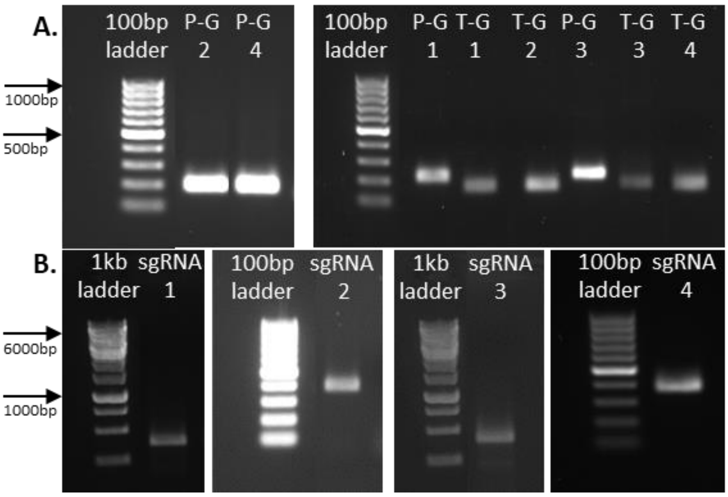


Figure 47: Agarose gel of amplification of fragments for use in fusion PCR and the fusion PCR. *The amplification of individual DNA fragments (A) shows each PCR was successful, and size was correct by comparing to the 100 bp ladder. Products of successful fusion PCR (B).*

The products of the fusion PCRs were used as a template to amplify fragments for use in the following USER reactions. The two sgRNAs required to delete a region in the genome was ligated using a USER rection with a 1-hour 16^{o}C incubation period and transformation into *E. coli* by heat shock. No colonies grew for DChr3. For D*THAS* twenty colonies were screened by a single colony PCR, ***Figure 47***. Two colonies with positive AGE banding were identified (D*THAS*-7 and D*THAS*-10). Plasmids of both colonies were extracted and sequenced. Both D*THAS*-7 and D*THAS*-10 showed 98% sequence homology to the target sequence. The mismatches were identified in the vital sgRNA sequence shown in ***Figure 48***.

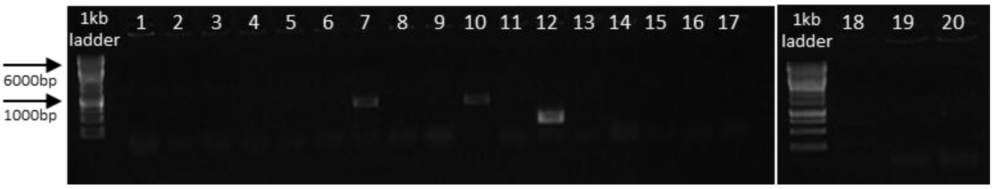
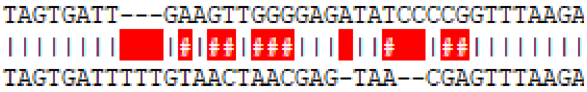


Figure 48: Agarose gel for colony PCR for DThas deletion construct. *Agarose gels for the CRISPR/Cas9 deletion constructs targeting the thalianol, expected fragment size 1000bp (DThas).*



```
TAGTGATT---GAAGTTGGGGAGATATCCCCGGTTTAAGA
||||||||   |#|##|###||| ||#  |##||||||||
TAGTGATTTTTGTAACTAACGAG-TAA--CGAGTTTAAGA
```

Figure 49: Sequence alignment for DThas-10 sgRNA region. *The alignment showing the mismatches of the thalianol deletion construct colony 10, at the vital sgRNA region.*

Fragments were re-amplified with new primers and successfully fused as previously described, followed by a repeat of the USER reaction. Three colonies were positively selected for DChr3 and two for D*THAS*, ***Figure 48***. Plasmids of these colonies were extracted and digested with HindIII. Four plasmids showed the expected digestion pattern (***Figure 49C***). Plasmids for two DChr3 colonies, 3 and 11, and one D*THAS* colony, 2, were selected for sequencing. DChr3-11 showed 99% sequence homology to the reference template (base pair mismatches in the U6-26 promoter), and DChr3-3 and D*THAS*-2 showed 100% sequence homology to the template. DChr3-3 and D*THAS*-2 were used to transform *A. thaliana* using the floral dip method. This shows that a combination of USER reaction and fusion PCR can be used as a quick method to generate smaller constructs, that do not require long ligation incubation periods.

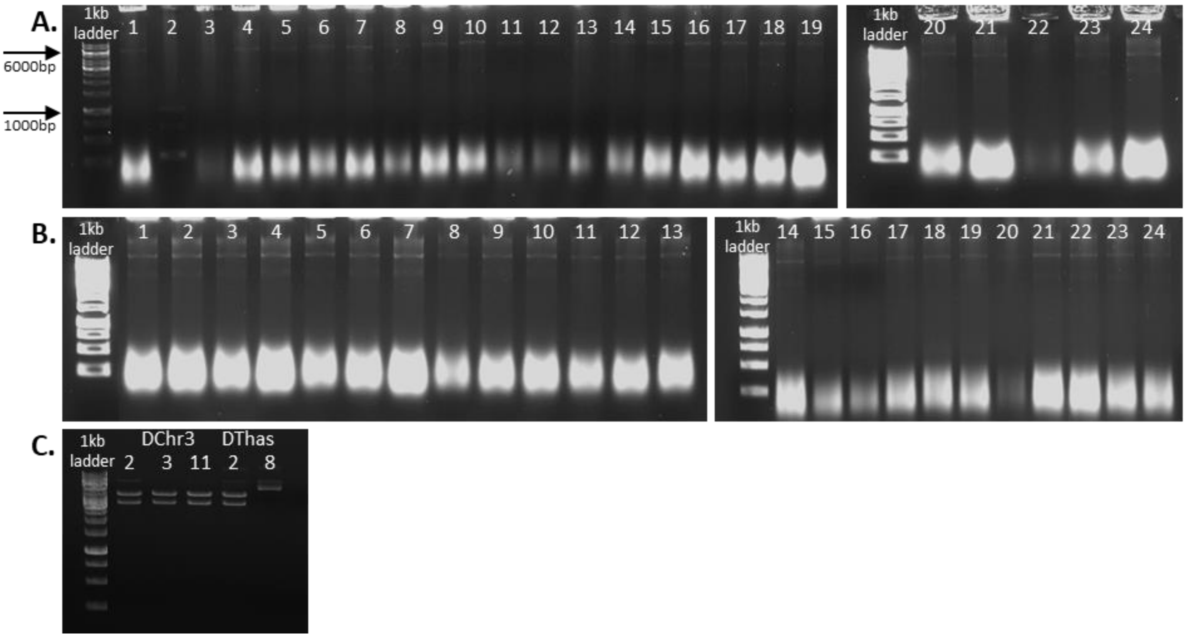


Figure 50: Agarose gel of colony PCR and restriction digest for deletion constructs. *Single colony PCRs were used to confirm insertion of the fragments by USER cloning, both DThas (A) and DChr3 (B) showed bands of the correct size, confirmed by the 1kb DNA ladder. Plasmids also underwent HindIII digestion for confirmation (C) of size. Expected size for A and B is 950 bp. Expected digest pattern (C) 4652 and 3372 bp.*

3.2.2.3 Analysis of plants with deletions

As described in section **3.2.1.4, *Figure 40*** seeds from transformed *A. thaliana* plants were collected and screened for a red seed coating, indicative of successful transformation. Subsequently, 15 red seeds were sown for D*THAS* and 8 for DChr3, of which 11 germinated for D*THAS*, and 8 for DChr3. Plant DNA was extracted from rosette leaves, and screening PCRs were performed to identify potential deletions, ***Figure 50***. First, short PCRs across the Cas9 restriction sites of D*THAS* plants were carried out. However, all PCR products showed wild type AGE banding, indicating that no deletion occurred.

Secondly, PCRs with primer targeting both border sites of the deletion were carried out. These would only result in a product if the deletions were generated. A selection of reactions were selected to provide an example of the results we obtained. For D*THAS* multiple products of various sizes were obtained out of 22 plants tested, and for DChr3 3 out of 8 samples were negative for a deletion-specific signal. All eight DChr3 and four D*THAS* PCR products were sent for sequencing. The sequencing results showed wild type sequence for all samples and as such, no deletions were successfully obtained. Notably, upon careful examination of the sequencing results, we detected one additional base pair in DChr3 plant number 3 inserted at the Cas9 target site, ***Figure 50***. This indicates that the Cas9 enzyme was successfully guided to at least one target site, implicating the successful application of the system in principle. These reactions were repeated on T1 and T2 plants. Between 100 and 400 plants were analysed for each deletion and plant generation. However, no deletions were detected.

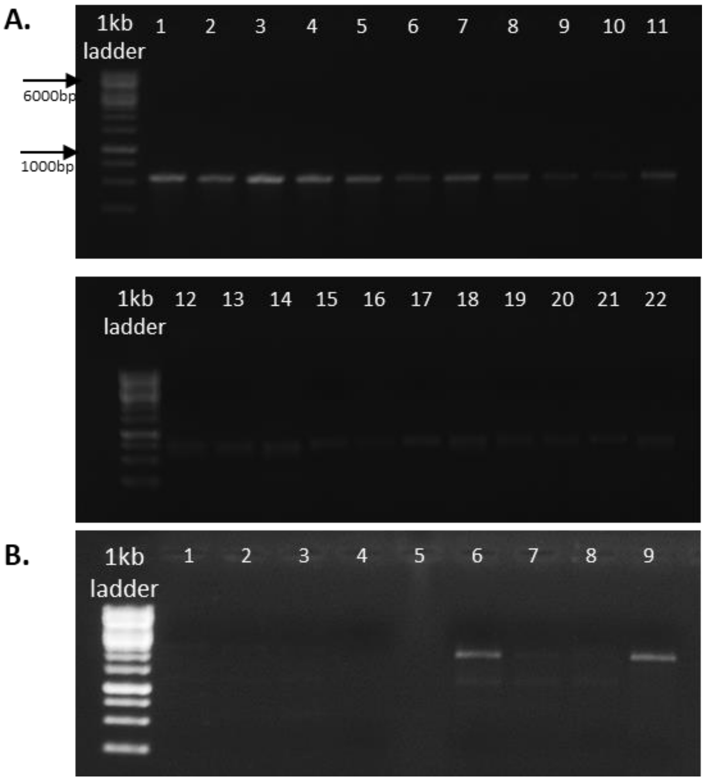


Figure 51: Agarose gel electrophoresis, genotyping transformed A. thaliana. *A. Shows the bands indicating no deletion at the thalianol gene cluster. B. Indicates bands showing no deletion at the Chr3 region.*

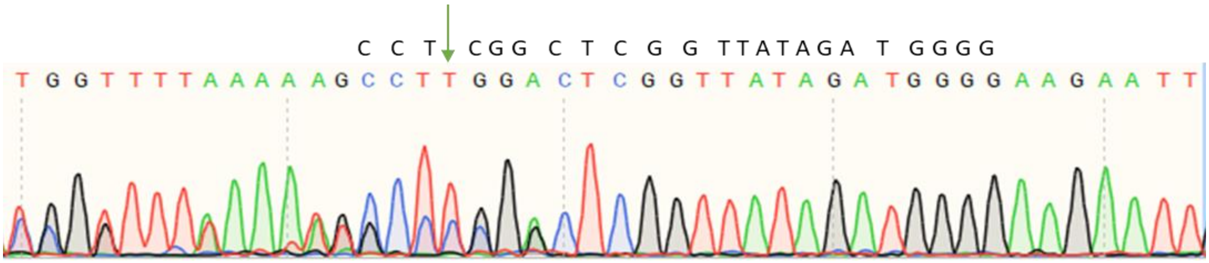


Figure 52: Sequence alignment showing additional base pair in DChr3 plant 3. *The alignment shows the insertion of 'T' (green arrow) when compared to the template sequence (top row).*

3.3 Cluster disruption discussion

3.3.1 Analysis T-DNA lines

We analysed the gene expression for 19 T-DNA lines which cause insertional mutagenesis throughout the thalianol gene cluster, extending into the silencing inactive domain. Our aim was to identify key regulatory regions that aid in the co-regulation of the thalianol gene cluster at a DNA-level. Here, the outcome of the gene expression analysis will be discussed.

3.3.1.1 Insertional mutagenesis of region upstream of *THAA2* could contain a regulatory element

Previous data had identified an active interacting 3D domain at the thalianol cluster formed in root cells. This domain encompasses the peripheral cluster gene *THAA2,* spanning 50 kb in size.[167] We identified two T-DNA mutant lines upstream of *THAA2* to investigate the boundaries of the interacting domain and its impact on cluster co-regulated gene expression, location shown in ***Figure 52***. A reduction in expression of all five thalianol cluster genes was detected in the SALK_077790 insertion line. The T-DNA insertion is located approximately 4 kB upstream of the *THAA2* gene in this line. Changes in all the genes were moderate, but we found a statistically significant reduction in the *THAA1* gene. Interestingly, when compared to cluster expression in SALK_071586, containing an insertion approximately 1,500 bp upstream of *THAA2*, no trend in gene expression changes was identified. By independent analysis of SALK_077790 we may hypothesize that the insertion has caused a 3D-conformation change and that this region is important for the folding of the 3D chromatin. However, as SALK_071586 led to no change in expression in a similar location it may be deduced that 3D domain formation is influenced by only very specific sites or other regulatory mechanism cause the change in expression. Interestingly, our re-analysed ATAC-seq data (***Figure 53)*** shows that SALK_077790 creates an insertion in an area of increased chromatin accessibility. Areas of increased chromatin accessibility indicate an important region for gene regulation[216]. The disruption of a histone modification binding site is considered unlikely as we do not find any potentially activating histone modification marks binding to this region when using our ReMap data.

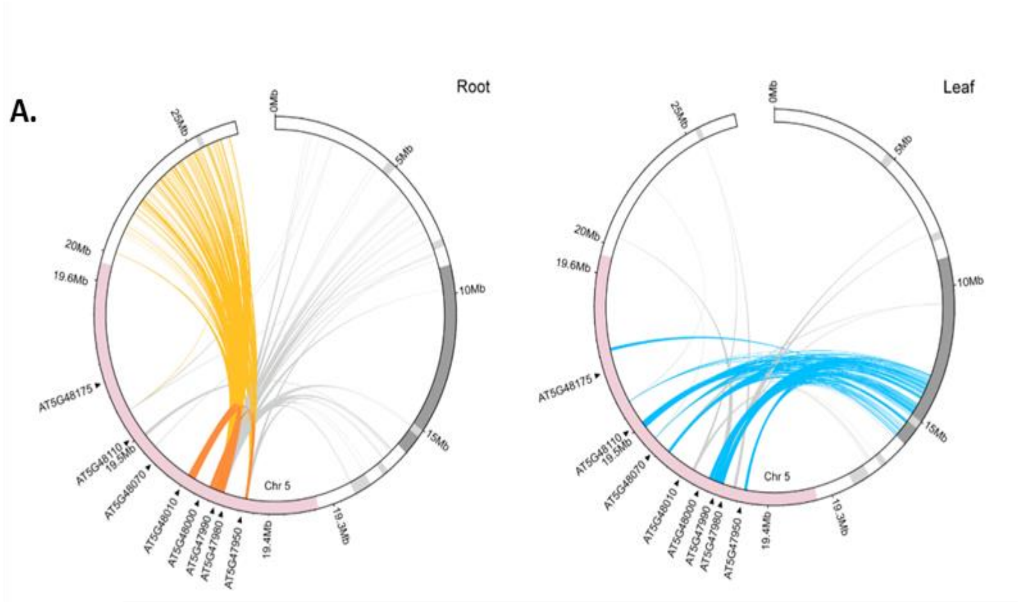


Figure 53: Circos plots showing the top 100 significant intrachromosomal interactions in the thalianol cluster in leaves and roots, identified using Hi-C. *Each connection shows a significant interaction between two regions, yellow indicates an interaction towards the long arm of chromosome 5, blue represents interactions towards the chromocenter, while orange is intra-cluster interactions, grey shows any other interaction up to the top 100 significant. (Adapted from Nützmann et al, 2020)*

Hi-C analyses of the thalianol cluster[163], ***Figure 53,*** indicates that the region directly upstream of *THAAS* significantly interacts with the region between *THAD* and *THAH* in the roots, whereas in the leaves there is interactions towards the chromocenter. Chromocenters are usually found within the nuclear periphery and are associated with heterochromatin.[217,218] Interestingly, only Salk_071586 falls into this region of intrachromosomal interaction, whereas Salk_077790 does not. This aligns with our finding that the T-DNA insertion in Salk_077790 falls into a highly accessible region. This suggests that disrupting the chromocenter-interacting region here, does not influence the regulation of the thalianol gene cluster. However, we predict that the region disrupted in Salk_077790 is a key binding domain for regulatory factors of the thalianol cluster.

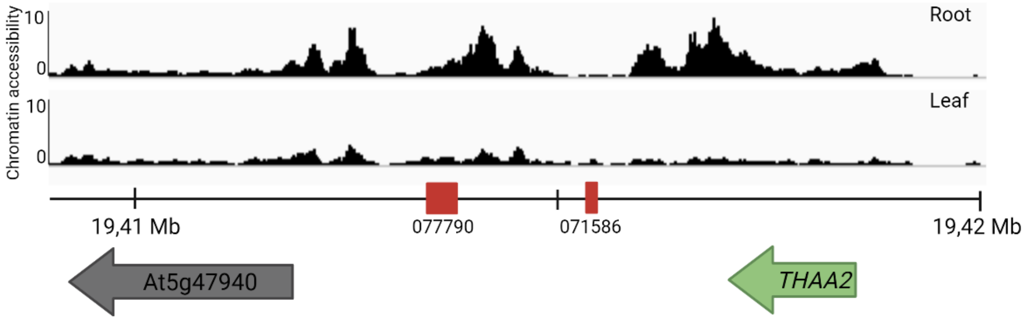


Figure 54: ATAC-seq data showing chromatin accessibility and the location of T-DNA line mutagen in the thalianol cluster. *Thalianol cluster gene shown by green arrow, non-cluster gene in grey. Red box indicated the location of the T-DNA insertion.*

3.3.1.2 Disruption in *THAS* leads to significant cluster change in specific positions only

We show how three different mutations in the *THAS* gene lead to different outcomes for the expression of the thalianol cluster. All three T-DNA mutants led to dramatic reduction in *THAS,* however Salk_064244 showed no expression change in the other four thalianol cluster genes. Both, Salk_138058 and Salk_062770 showed increased expression of *THAA1, THAD* and *THAH* with the latter mutant showing more dramatic upregulation. This leads us to believe that the expression profile of *THAS* does not affect the expression profile of the genes within the same cluster, whether that could be due to feedback mechanisms. All the mutants are in a region of increased accessibility according to the ATAC-seq data. The three mutant lines SALK_062770, SALK_064244 and SALK_135058 are each in a different exon, 8, 9 and 12, respectively.

Another factor for SALK_062770 causing a more significant change could be due to the insertion of multiple T-DNA insertions. A study showed that the average amount of insertions per T-DNA mutant lines was 1.8686, whereas previously it was only described as 1.587.[219,220] Having a larger insertion mutation than expected could lead to some mutants resulting in stronger results due to more disruption of the genome architecture and the potential binding region. This region is also more enriched with H3K4me3 and H3K18ac according to our ReMap data. These marks are both implicated as being gene activators[39,221], and thus a mutation at this site would lead to downregulation of the gene, as seen, and potential indirect affects causing dysregulation of the thalianol gene cluster. Regions less enriched with these marks may lead to a chromatin more accessible for polymerases, thus leading to upstream expression dysregulation.

We also know from previous Hi-C data that *THAS* is the gene on the border of the intense interacting region in roots[167]. Thus, an insertional mutation in the bordering region could lead to a 3D conformational change, leading to increased expression of the cluster genes, excluding the interrupted gene *THAS*. Based on our results, we predict that there is a regulatory region around exon 8 to 9 of the *THAS* gene

3.3.1.3 Reduced *THAA2* doesn't lead to sig change in gene cluster when in intron

An insertion in the intronic region of *THAA2* does not lead to any significant changes in the core thalianol cluster genes. This suggests that while the five genes are co-expressed, expression of individual genes is not a prerequisite for expression of the other genes. Due to the significant reduction in *THAA2* transcripts, it can be certain that the mutant is causing disruption of *THAA2*. We see in Section **3.2.4.2**, that multiple accessions with *THAA2* inversions show misregulation of the core cluster genes. This may indicate that disruption of the *THAA2* gene itself is not important for the regulation of the thalianol gene cluster, however the architecture and orientation of the gene is of importance.

3.3.1.4 Downstream *THAD* disruption leads to major thalianol gene cluster disruption

The complete dysregulation of the peripheral and core thalianol cluster genes was found in two mutant lines 130 bp and 300 bp downstream of *THAD.* In the mutant closer to *THAD* all the cluster genes were downregulated, while in the other mutant we found almost no expression of *THAD*, lower *THAA1* expression and in increase in the expression of *THAA2*, *THAH* and *THAS.* Hi-C data shows that the region that has been mutated is in an interacting region for the roots, therefore could lead to a conformation change in the 3D structure of the chromatin, thus disrupting the co-regulation of the cluster genes. Using ATAC-seq data it can also be confirmed that this is a region of higher accessibility in the roots. Regions that regulate gene expression are usually 5' of the gene[222], so it is interesting that a downstream region would cause disruption to the whole gene cluster. However, some studies have shown regulatory regions downstream of the gene, such as COP1 and HY5 that regulate light inducible gene expression.[223]

Recently, a super enhancer was discovered 1000 bp upstream of the SALK_093786 insertion[224]. Deletions in the super enhancers most accessible region to DNase I found downregulation of the four core cluster genes in one mutant line, but downregulation of the three upstream genes in another. Salk_149827 shows a similar expression profile to the mutant line #192 in the super enhancer. It can be confirmed that these mutant lines are not directly mutating *THAD* as the expression profile of *THAD* does not cause cluster dysregulation. We predict that the location isn't specifically causing the difference between the two mutant lines due to them only being around 200 bp apart, as well as Salk_149827 being slightly further from the super enhancer and causing a similar expression profile.

As previously mentioned, some T-DNA lines often contain multiple insertions which in this case could lead to different 3D chromatin confirmational changes, predictions depicted in ***Figure 54***. When the insertions are at a similar locus and showing different expression patterns, thus leading us to assume that Salk_149827 contains a higher number of insertions than Salk_093786. The location at which both insertional mutations occurs contains intra-cluster gene interactions between *THAA2* and *THAS*[163] allowing us to speculate the conformation of the mutants alongside our gene expression data. In Salk_093786 we predict that the insertion has led to a larger loop forming at the *THAD* locus. This may disrupt the chromatin loop for *THAA1*, creating a larger distance between both promoters and the super enhancer. *THAD* and its promoter will become looped out due to the insertion, thus further away from the super enhancer leading to almost no *THAD* transcripts detected. *THAA1* will lose some contact with the enhancer leading to reduced expression. This leads to *THAH, THAS* and *THAA2* being in closer proximity to the super enhancer, thus leading to increased expression. We speculate that In Salk_149827, the insertion (assumed to be multiple T-DNAs) has disrupted the looping of all the thalianol cluster genes resulting in loss of promoter-enhancer contacts *THAA1, THAD, THAH* and *THAS* and reduced transcription. *THAA2* continues to loosely interact with the super enhancer and shows only minor expression changes.

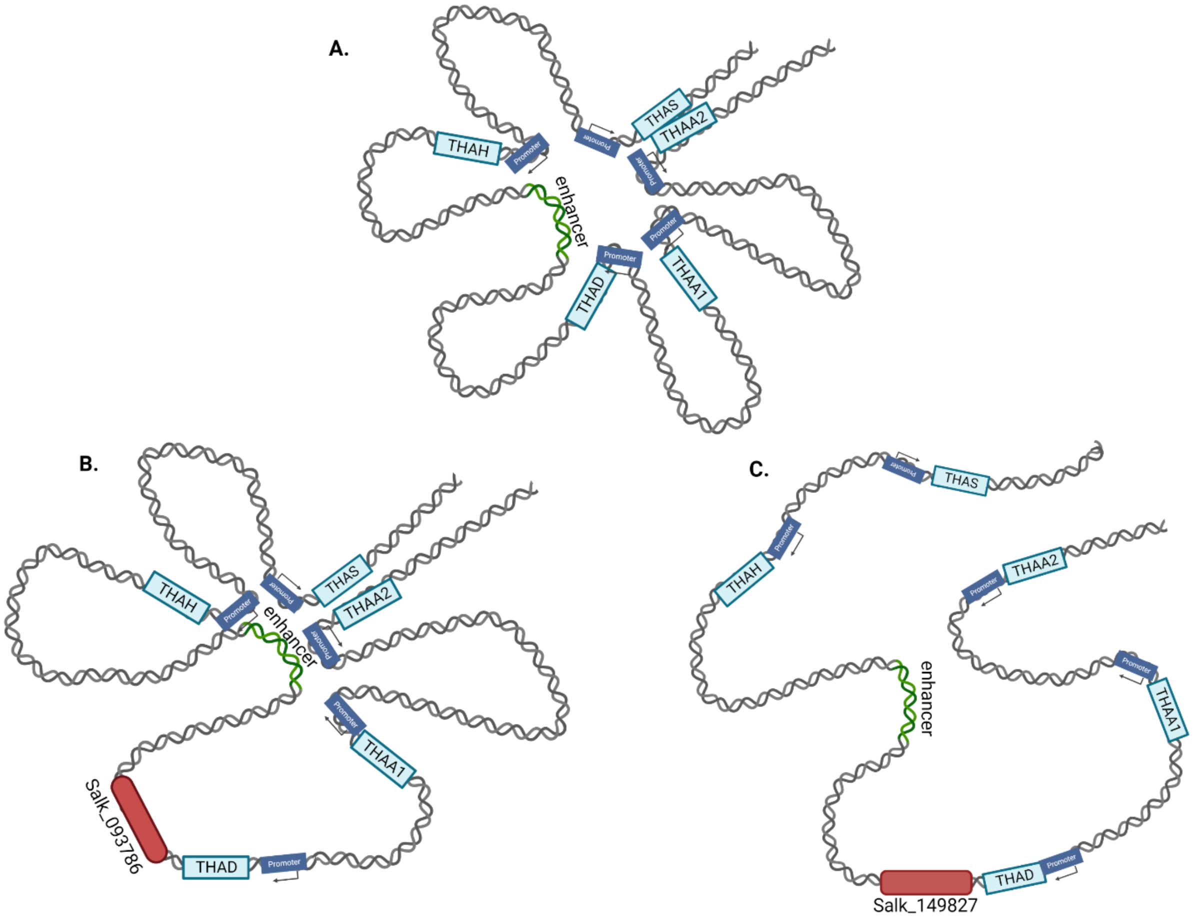


Figure 55:Depiction of the 3D chromatin architecture of the thalianol cluster, and the predicted structure based on transcript analysis of two T-DNA mutants. *A. Shows the canonical 3D architecture of the thalianol gene cluster, the genes at the region and their promotors indicated, in green is the proposed super-enhancer. B. Is the predicted structure when there is a Salk_093786 insertion, based on reduced THAD expression. C shows the reduction of all the genes in the Salk_149827 insertion. T-DNA is indicated by the red box. (Created with BioRender.com)*

3.3.1.5 *THAH* disruption

Interestingly, an insert within the *THAH* gene close to the transcription start site led to increased expression of *THAH* appearing to facilitate transcription whereas a mutation upstream in the promoter region leads to downregulation. ATAC-seq data shows that both insertions are in a region of high chromatin accessibility, however the region inside the gene is more accessible. Salk_064244 is an insertion in the intron of *THAS,* meaning transcription of the gene can still occur. Although the mechanism is largely unknown, introns within 1 kb of the transcriptional start site have been found to have a positive effect on gene expression[225], called intron-mediated enhancement (IME). IMEs have also been found to have a role in gene looping to enhance transcription in budding yeast.[226] The increased in accessibility shown in the intronic region in the ATAC-seq data has the possibility to be a IME. We suggest that the insertion in the intronic region has interrupted this intra genic looping caused by an IME leading to an increase in enhancement and increasing gene expression of *THAH.* Overexpression studies of *THAS* showed a reduced level of intermediates[157], potentially explaining why there was a trend for reduction of expression in the other thalianol cluster genes when *THAH* is being overexpressed.

While SALK_100156 is potentially in the promoter region for *THAH* leading to a reduction in its expression. Previous *THAH* knockout studies have shown accumulation of 3β,15-thaliandiol T3 and 16-keto-thalianol[227], which are produced from *THAD*, also meaning a possible overexpression of *THAD*. The insertion could have led to a conformational change in the 3D chromatin structure in which *THAD* is interacting more strongly with promoter regions thus leading to overexpression.

3.3.1.6 Intergenic disruption of *THAH* and *THAS*

Two mutant insertion lines approximately 500 bp apart show a similar expression profile of the thalianol gene cluster, with more pronounced changes in Salk_0444069 than Salk_081921. Both show a reduced expression in *THAH* and *THAS* possibly due to disruption of a bidirectional promoter, which will act as a promoter for both *THAH* and *THAS.* A more pronounced change in Salk_044069 could be due to an increase in T-DNA insertions, however this mutant is also in an area of more accessible chromatin in the roots, therefore is more likely to be a binding site for transcription machinery.[216] Hi-C data[163] shows that an intra-genic cluster interaction also occurs at this site, linking this region to *THAA2, THAA1, THAS* and the recently discovered super enhancer[224]. By causing a mutation in the area linking the promoter of *THAH* and *THAS* to the super enhancer we can expect a conformation change, as seen in ***Figure 55***. Looping is disrupted, and the promoter region is a greater distance from the super enhancer. Alongside this, Salk_044069 insertion is disrupting a H2A.Z and H3K18ac enriched region, as they have previously been implicated as gene activators, mutations in this site would lead to reduced expression of the directly adjacent genes as seen[60,221]. Both marks contribute to the loosening of the chromatin by neutralising the lysine's positive charge and by recruiting RNA polymerases[10,52], thus reduction of these marks across genes within the thalianol cluster could lead to a more compact chromatin structure leading to the reduction in gene expression.

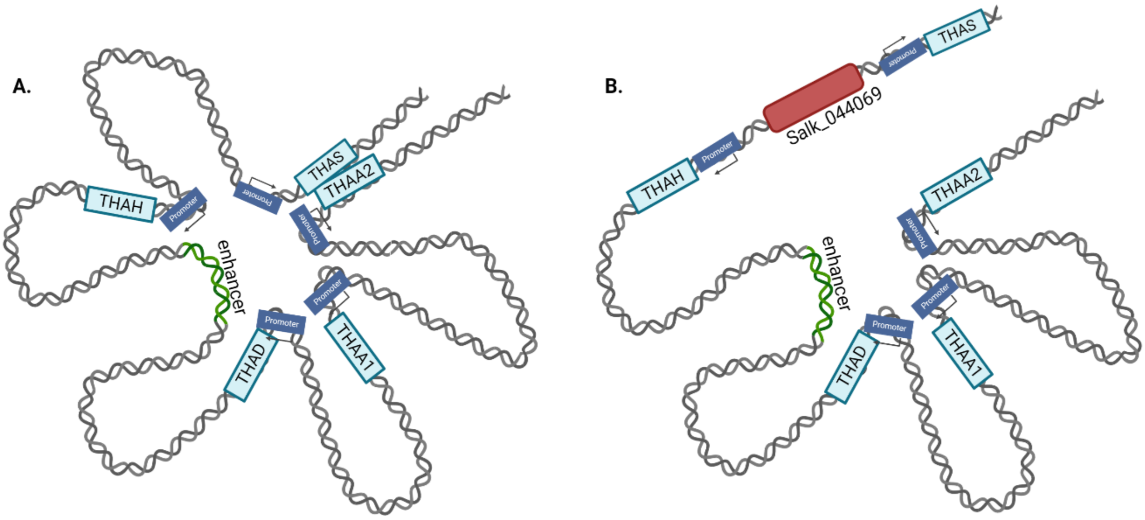


Figure 56: Depiction of the 3D chromatin architecture of the thalianol cluster, and the predicted structure based on transcript analysis of a T-DNA mutant. *A. Shows the canonical 3D architecture of the thalianol gene cluster, the genes at the region and their promotors indicated, in green is the proposed super-enhancer. B. Is the predicted structure when there is a Salk_044069 insertion, based on reduced THAH and THAS expression. T-DNA is indicated by the red box. (Created with BioRender.com)*

Both mutants also interrupt a prominent SNP-dense region of the thalianol cluster that occurs in the intergenic region between *THAH* and *THAS* (Li, unpublished results). Interestingly, Salk_081921 also interrupts a transposable element. Transposable elements play a huge role in shaping the genomic architecture and regulating phenotypic variation, as well as regulating the expression levels of genes through either *cis*- or *trans*- regulatory elements located within transposable elements sequences.[228] In *Aspergillus nidulans* the penicillin cluster has been shown to be positively regulated by transposable elements, and deletion in this region lead to decreased penicillin production.[229] Transposable elements are often associated with contributing to the formation of gene clusters due to their role in deletions, translocation and inversion, of which segmental duplication is thought to have led to the formation of the thalianol cluster.[153] Although, we don't know the function of this specific transposable element, AT5TE70000, it is possible that this element can regulate the thalianol gene cluster due to the misregulation of the genes caused by a mutation at this site.

3.3.1.7 Downstream intergenic region affects thalianol cluster gene expression

When the cluster is inactive, in leaves, there is a large silencing domain which extends from *THAA2* to the non-clustered and non-coexpressed genes *PAT1* and *OBE2* (*At5g48150* and *At5g48160)*[163], conformations shown in ***Figure 56***. We suspect a regulatory region may occur between these two genes, thus stabilising the 3D conformation in leaves to silence the thalianol gene cluster. To better define this regulatory region three T-DNA mutant lines were analysed, Salk_126935, Salk_084373 and Salk_029884. Salk_029884 showed no change in gene expression across the thalianol cluster so will not be discussed. It is determined that the promoter region could be an important region for the silencing of the root cluster. The T-DNA line approximately 550 bp upstream of *PAT1* causes significant upregulation of all five thalianol cluster genes whereas the mutant line 2500 bp upstream shows a trend for upregulation, with *THAH* being the only downregulated, however non-significantly. This could be due to minor disruption of the 3D structure or interruption of the potential regulatory region 1200 bp away. When this region is interacting with the thalianol cluster in leaves we can assume that it is acting as a repressor, however it is possible that when this region is not in contact with the cluster in roots, it is binding to an activator thus aiding the activation of the thalianol cluster. Although this region is 60 kb downstream from the cluster, chromosomal looping will allow regions a large distance away to interact.

At5g48150 (*PAT1* or phytochrome A signal transduction 1) is part of the plant specific GRAS regulatory protein family and controls basic plant developmental processes such as hypocotyl elongation.[230] The GRAS protein family is named after its first three members, gibberellic-acid insensitive (GAI), repressor of GAI (RGA) and scarecrow (SCR)[231], and 33 have been identified in *A. thaliana.*[232] GRAS proteins have been found to have a diverse regulatory role as transcription factors[233] in root and shoot development, gibberellic acid (GA) signalling and phytochrome A signal transduction.[234] A subfamily of GRAS proteins, DELLA, represses gene expression by binding to PIF3 and PIF4 (phytochrome interacting factors) and forming an inactive complex, thus stopping them from binding to their target promotors. However, in the presence of GA, DELLA proteins are degraded allowing PIF3 and PIF4 to regulate gene expression and promote growth.[235,236] *PAT1* has been described as specifically acting on the phyA-signalling pathway[230] due to the inactivation of *PAT1* leading to a similar phenotype as *phyA* mutant line. The studies also showed that *PAT1* is not involved in the GA signalling pathway.[230]

PAT1 has been identified in grapevine (*Vitis)* to regulate JA biosynthesis in the cold stress response.[237] JA has been identified to play a key role in abiotic stress in plants, such as in cold[238], but also in biotic stress, such as the regulation of cellular defence.[239] Overexpression of the wild Amur grape (*Vitis amurensis*) *PAT1* (known as *VaPAT1)* promotes JA biosynthesis to enhance cold stress resistance. Due to GRAS proteins often lacking DNA binding domains they often interact with intermediate domain (IDD) proteins to bind to DNA and regulate transcription of target genes.[240]

JA is a stressed-induced phytohormone and plays a crucial role in plant immunity and defence[241] has been shown to induce the activation of all five of the thalianol cluster genes.[242] A member of the Ethylene Response factor (ERF) *ERF115* promotes root tissue regeneration controlled by MYC2-mediated JA signalling.[243] A heterodimeric transcription factor complex has been shown to form with *ERF115* and *PAT1* (ERF115-PAT1) triggering cell division.[244]

This provides evidence that in roots, when the thalianol cluster is active, a molecular network including *PAT1* could be aiding the co-regulation of the clustered genes, thus when the promoter for this region is mutated dysregulation of the cluster occurs leading to overexpression. Future studies would be interesting to analyse knockouts of *PAT1*, alongside JA studies on the effect of the thalianol gene cluster.

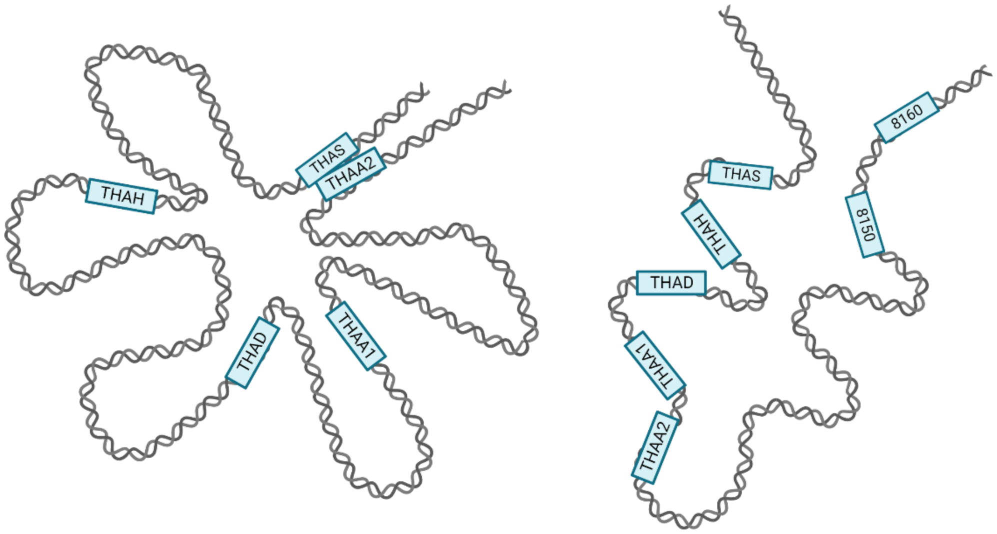


Figure 57: Depiction of the 3D chromatin architecture of the thalianol cluster in roots (left) and leaves (right). *The cluster is active in roots where the thalianol cluster is highly interacting within itself to coregulate the cluster. In leaves where the cluster is inactive, there is interaction with the downstream region to form a hairpin-like structure. (Created with BioRender.com)*

3.3.2 Accessions with similar architecture show differing expression profile

To analyse whether the order of the thalianol cluster is important for the co-regulation of the genes, we investigated the expression levels of the cluster in four different accessions compared to the wild-type, Col-0. All four of these accessions have an inversion, leading to *THAA2* being directly adjacent to *THAA1.* Interestingly, we find that when *THAA2* is not expressed, *THAA1* is being dramatically overexpressed in three of the accessions. There is also a similar trend for expression across the three other thalianol cluster genes, the most notable being an increased in expression of *THAH*. However, when *THAA2* is displaying the same level of expression in Col-0 and the Sha accession, there was a reduction in *THAA1* expression alongside a reduction in expression in *THAD* and *THAS*.

A study looking into the orientation of 21 accessions showed that only three had the same orientation as Col-0, and in 17 *THAS, THAH, THAO, THAA1*, and *THAA2* genes were contiguous[159]. These studies suggest that there was a chromosomal inversion to compact the thalianol cluster. Although *THAA2* is often described as not being a core gene for the cluster its importance is clear and Hi-C data shows that there is a chromosome loop between *THAA2* and *THAA1*[163], which appears to be an imperative loop for the regulation of the thalianol gene cluster.

For future studies, it would be interesting to look at the gene expression levels of the thalianol cluster as well as the other biosynthetic gene clusters to analyse how the gene order impacts expression. Other accessions should also be analysed, alongside obtaining more long-read sequencing data. Previously, the thalianol cluster has been described as being highly conserved, however we can conclude only the four core cluster genes are conserved. The change in expression across the cluster in different ecotypes could allow for the production of different thalianol derivatives. This may result in the establishment of unique root microbiomes [157] in different *A. thaliana* accessions enabling survival in in distinct ecological niches.

We also speculate that the orientation and position of the genes are important for the co-regulation, as when *THAA2* is downregulated due to a T-DNA insertion (SALK-080717, **Section 3.2.3.2. *Figure 12***) there is not a change in expression for the four core cluster genes. However, when there was a chromosomal inversion, indicating that the orientation and distance from the thalianol cluster had changed, the expression profiles change dramatically.

As the changes in the expression profile of the clusters will also lead to accumulation of potentially toxic intermediates[106,115], the phenotypes of these ecotypes and T-DNA mutants should also be assessed. It is predicted that the genes within a cluster are ordered in a way to allow for co-regulation of the genes within the cluster[171], thus reducing the likelihood of these toxic intermediates causing harm to the growth of the plant. However, our studies show that the expression profiles do vary between ecotypes. Although not seen in our analysis, overexpression of *THAS* leads to dwarfing of the plant.[106] This leads us to believe that the overexpression of *THAA1* as seen in An-1, Ler-0 and Kyo could lead to phenotypical changes such as a size difference when compared to Col-0. It would also be of interest to get a metabolic profile of the ecotypes as well as the T-DNA mutants to see how these changes affect the plant's ability to produce thalianol.

3.3.3 CRISPR/Cas9

The aim of this project was to create a CRISPR/Cas9 protocol alongside understanding the most efficient methods to create large constructs to generate deletions and insertions in *A. thaliana.* We utilised different methods to ligate DNA fragments together in search for the most effective way, and we conclude that a combination often works best. An extended USER cloning protocol using a lower number of DNA fragments, which had been gel purified, with a similar base pair size alongside long ligation times, electroporation and long colony growth periods was shown to be most effective when generating constructs with multiple fragments. Here, I will discuss the decision making and choice of protocol for construct design, improvements in the CRISPR protocol and the results we obtained.

3.3.3.1 Construct design

Due to the low efficiency of HDR in *A. thaliana,* CRISPR/Cas9 gene editing is often indicated as not being as effective, thus we attempted to design a construct with all the factors accounted for so that the genome can be edited successfully. One of the two most important features of CRISPR/Cas9 gene editing is the sgRNA, which guides the Cas9 endonuclease to the target site to create a double-stranded break in the DNA.[180] To have maximum efficiency, an on-target score is also considered based on research which analysed sgRNA editing efficiency by a high-throughput analysis in mammalian cells[245,246], therefore this was also evaluated to choose a guide with a high on-target score. Although ambiguous, it has also been suggested that a higher GC content within the guide leads to more efficient binding of the guide to the target sequence[195,247] Thus we selected sgRNA based on the on-target score and a higher GC content where possible. The stability of the sgRNA is also a limiting factor in the CRISPR system and optimizing the sgRNA structure improves the interaction with the Cas9 protein[248]. The whole sgRNA complex that binds to the Cas9 endonuclease is made up of the crRNA and tracrRNA, in which the crRNA corresponds to the target sequence and the tracrRNA is recognised by the Cas9.[180] A redesign of the sgRNA included an additional 4 to 10 bp fragment in the helix in the attempt to mimic the interaction between crRNA and tracrRNA, which was found to make the sgRNA complex more efficient *in vivo.*[191] An sgRNA containing an A-T inversion to remove a TTTT potential termination signal and an extended Cas9-binding hairpin structure (Extension-Flip sgRNA) was compared to the 'classic' sgRNA, ***Figure 57,*** and the efficiency of the extension-flip sgRNA was found to be higher.[195,248] Therefore, we used the extension-flip sgRNA in our constructs. Increasing the efficiency of transcription of the sgRNA was also a quick way to increase efficiency. Thus, we aimed to use a promoter and terminator that allowed the sgRNA to be transcribed in all cells by RNA polymerase III.[195] Three genes, *U6-1, U6-26 and U6-29,* are transcribed by RNA polymerase III in *A. thaliana*. Using the 205 bp region upstream of the transcription start site of each gene, the promoter region, GUS expression levels were analysed driven by each promoter, and the promoter region for *U6-26* was found to have the strongest activity. [249] Using the *U6-26* promoter region (P*U6-26*) and varied lengths of regions downstream from the *U6-26* gene, it was found that a region 67 bp in length was the most efficient.[195] Thus, we used the 205 bp *U6-26* promoter region and 67 bp *U6-26* terminator region for regulating the transcription of the sgRNA in our constructs.

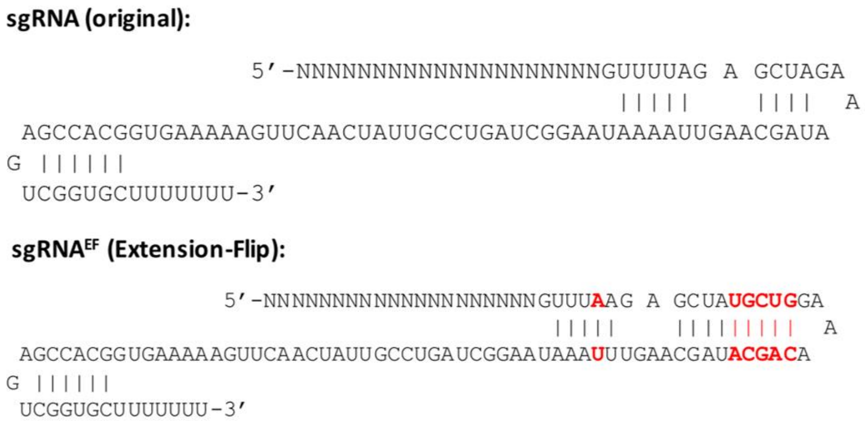

sgRNA (original):
5′-NNNNNNNNNNNNNNNNNNNNGUUUUAG A GCUAGA
||||| |||| A
AGCCACGGUGAAAAAGUUCAACUAUUGCCUGAUCGGAAUAAAAUUGAACGAUA
G ||||||
UCGGUGCUUUUUUU-3′
sgRNA^EF (Extension-Flip):
5′-NNNNNNNNNNNNNNNNNNNNGUUUAAG A GCUAUGCUGGA
||||| ||||||||| A
AGCCACGGUGAAAAAGUUCAACUAUUGCCUGAUCGGAAUAAAUUUGAACGAUACGACA
G ||||||
UCGGUGCUUUUUUU-3′

Figure 58: Comparison of the original sgRNA and the Extension-Flip sgRNA. *Diagram showing the differences between the original sgRNA compared to the Extension-Flip sgRNA. The Extension-Flip sgRNA has an A-T inversion and an extended hair-pin structure. (Adapted from Castel, 2019).*

3.3.3.2 A multi-step USER cloning protocol is more efficient to create large constructs

Successful ligation of six DNA fragments into the vector using a combination of ligation methods, showed that USER cloning can effectively be used to create large constructs, however the method does have limitations. Fusion PCR was shown to be very effective to ligate smaller DNA fragments together, however it did produce some mismatches when the products were sequenced. Due to the speed of these reactions, repetition did not delay experimentation by long periods of time. USER cloning has been described as a simple, very efficient method to seamlessly clone multiple PCR products into a vector[208]. However, little is available to troubleshoot the protocol besides efficiently designing the overhangs which complement the fragment to ligate to. To increase the efficiency of the reaction, DNA fragments were diluted to equimolar levels to account for the size difference between DNA fragments. We also identified that an extended ligation time, recommended is 1-hour[208], of 16-hours is more efficient to complete the ligation of larger DNA fragments. The USER reaction, however, has been shown to be less efficient with an increased number of DNA fragments to ligate, especially those of varying size. After multiple attempts with cloning all 6 fragments in one USER reaction step, we deduced that either the number of fragments to ligate was too high or the different sizes of the fragments was affecting the efficiency. Following the alteration of our methodology to ligate a smaller number of fragments together, thus creating extra USER reactions steps and fragments of similar size (in base pairs), we identified a high efficiency of successful assembly of large constructs. We obtained a 90% efficiency rate using a multi-step USER cloning process, the only downside being that for some constructs we found multiple banding indicating the ligation of some but not all fragments into the vector. This, however, was easily resolved by re-spreading bacterial colonies onto new plates. Thus, we recommend that when ligating DNA fragments to create a large construct that small fragments are ligated using fusion PCRs and checked using sequencing alongside a multi-step USER cloning protocol with similar DNA fragment sizes.

3.3.3.3 Effect of inserting regulatory region into different sites

Our initial aim for the insertional CRISPR/Cas9 technology was to insert regulatory regions from the *THAS* gene in the thalianol gene cluster and the constitutively expressed gene *PP2AA3*. It has long been assumed that the promoter regions are responsible for holding the key regulatory information required for the expression of a gene[250], however we previously identified in the thalianol gene cluster the possibility of regulators outside of these regions such as the super enhancer[224]. Although we are yet to confirm the location of the insertion, we show that inserting a promoter and terminator region of a gene with a well-defined expression profile into a region with alternate expression will change the expression profile of the regulatory regions. By inserting the promoter and terminator of the *THAS* gene that is exclusively expressed in the roots[153], into the *PP2AA3* region which is constitutively expressed[251] we show the regulatory regions take on the expression profile of the architectural environment, thus also becoming constitutively expressed in leaves and roots. We show that this isn't just an effect of the regulatory elements being inserted into another location as when this is also inserted into the *MRN* cluster, there is only expression in the roots. We predict that the overall architectural environment of a genomic region holds more information than previously expected to aid in the regulation of the genes at that site. Although we did not genotype these plants for homozygosity, we the pattern we observed in the mutant plants is conserved across all plants in the respective mutant type. To further expand on this study, we first aim to confirm the location of the inserts, as well as also inserting the regulatory regions for the *PP2AA3* gene into these locations. It would also be interesting to repeat with other promoter elements of the thalianol cluster and other genes from metabolic gene clusters.

3.3.3.4 Deleting regions of DNA using CRISPR/Cas9

Our aim was to create large deletions (around 30 kb) in the thalianol and in a region in chromosome 3 to determine the architecture's importance in regulating gene clusters. However, we did not obtain any positive deletions. The system has been proved to work as we identified mismatches during sequencing at the sgRNA target sites. Although large-scale deletions of up to 120 kb have been reported in *A. thaliana* these are rare and a large number of plants need to be screened.[252] Studies have shown that small deletions in *A. thaliana* of around 100 bp occur in much higher frequencies while large deletions between 5 kb and 120 kb only occur in less than 1% of plants screened.[252] The study also showed that a strategy incorporating multiple sgRNA for targeting a single loci also increased deletion efficiency in most cases. The complex chromatin architecture of these target regions may also lead to reduced efficiency in the capability of Cas9 acting as an endonuclease and sgRNA binding at these sites. To improve the efficiency of this experimentation in future to obtain large deletions, more sgRNA should be inserted in one construct, and a larger number of plants should be screened. Multiple smaller deletions may also be carried out, rather than attempted to create a 30 kb deletion in one step, a study showed a 12 kb deletion had a 51% success rate.[253] Our studies largely focused on T1 plants, however the efficiency of obtaining deletions was found to be higher in the screening of T2 plants.[254]

4 Histone modification marks

4.1 Overview

In section 3, we explored how sequence integrity of the thalianol cluster may contribute to the co-regulation of clustered genes. Here, we aim to analyse and predict key chromatin regulators of the thalianol cluster that may directly or indirectly shape co-regulation and three-dimensional chromatin structure. The patterning of histone modification marks across chromatin has been shown to influence chromatin architecture by either relaxing or compacting chromatin.[255] For example, the acetylation of histone lysine residues, typically linked to actively transcribed genes, is suggested to lower the positive charge of the histone tail and therefore reducing the interaction between histones and negatively charged DNA.[255] As described in the background section, methylation of histone marks has a more complex regulatory output with lysine residues being mono-, di- or tri-, methylated[256] For example, H3K4me3 labelling is highly correlated with transcriptionally active genes while H3K27me3 marks are predominantly found at repressed genes.[257]

The deposition and removal of histone marks are carried out by transferases and deacetylation/demethylation enzymes, respectively.[258] Lysine specific demethylase 1 (*LSD1*) was the first histone demethylase described[259].The Jumonji C (*JMJ*C)-domain containing demethylase family is the largest gene family associated with histone demethylation.[260]

4.1.1 Histone demethylases

4.1.1.1 LDL1 and LDL2

In animals only two *LSD* encoding genes are found, *LSD1* and *LSD2*. In contrast, the *A. thaliana* genome contains four homologs of *LSD1* called *LSD1-Like* (LDL) 1-3 and *FLD* (flowering locus D), each encoding for flavin-containing amine oxidases and SWIRM domains.[259] The SWIRM domain is found in a high number of proteins which regulate chromatin, however its function is not yet known[261]. *LSD* histone demethylases of *A. thaliana* have been implicated in several developmental and stress defence processes, such as *LDL1* in root elongation and lateral root initiation.[262,263] The *LDL1* in *A. thaliana* specifically demethylates H3K4me2 and H3K4me1 and can interact with SET-domain methyltransferases to form co-repressor complexes.[264,265] They have been described to be involved in flowering time control, seed dormancy, auxin homeostasis, the circadian clock, the immune response, and root development. *LDL1* and *LDL2* have been shown to work redundantly in repressing seed dormancy by reducing H3K4 methylation levels.[266] Loss of *LDL1* and *LDL2* leads to an upregulation of target genes[267]. *LDL3* specifically demethylates H3K4me2, however this change hasn't been shown to affect gene expression, however it does facilitate later activation of genes that act to form shoot progenitors.[268] The fourth homolog, *FLD*, has been shown to function as a deacetylase and demethylase to control flowering time[269].

4.1.1.2 Jumonji C (*JMJ*C)-domain containing demethylases

JMJ proteins can demethylate mono-, di- and tri- methylated lysine residues.[270] 21 *JMJ* histone demethylases, which are categorised into five groups, have been identified in *A. thaliana* thus far.[271] The five groups are *KDM5* (lysine-specific demethylase), KDM4, KDM3, *JMJD6* (Jumonji-domain containing 6) and the *JMJ*C domain-only group. The categorisation is based on *JMJ*C domain sequences and architecture. Proteins within a group tend to show similar target specificity for histone modifications.[271,272] *JMJ* type demethylases have been shown to be involved in gene repression and developmental regulation.[273]

4.1.1.2.1 JMJ14

JMJ14 is a histone demethylase suggested to be involved in the coordination of H3K4me3 levels at target loci. It have been shown to be an important regulator of local and systemic plant immune response, flowering time as well as transcriptional and posttranscriptional gene silencing.[274–276277] It has been speculated that at plant defence genes *JMJ14* is involved in both methylation and demethylation of H3K4, therefore, playing a role in gene activation and gene silencing.[277] *JMJ14* mutant leads to reactivated transgenes which are usually silenced by post-translational gene silencing, resulting in a reduced level of H3K9K14Ac levels at target loci.[278] Endogenous loci typically regulated by *JMJ14* show an increased H3K4me2/3 levels in the mutant, however transgene loci show unchanged H3K4me2 and a decrease in H3K4me3, indicating the availability for other H3K4 demethylases to act on these transgenes in the absence of *JMJ14*. [278]

4.1.1.2.2 JMJ15

The three *JMJ* demethylases, *JMJ15, JMJ15* and *JMJ18* have all been shown to have H3K4me2/3 demethylase activity and regulate diverse aspects of chromatin function and development.[275,279,280] It has been shown that an increased expression of *JMJ15* leads to downregulation of H3K4me2/3-marked stress-related genes.[281] In *JMJ15* overexpression lines, 23 significantly up-regulated and 164 significantly downregulated genes compared to the wild type were detected, indicating a repressive role of *JMJ15* in gene regulation.[281] This is in accordance with H3K4 demethylation activity of *JMJ15*.

4.1.1.2.3 JMJ16

JMJ16 is a H3K4me1/2/3 demethylase which is associated with leaf senescence and the maintenance of a reactive oxygen species homeostasis in *A. thaliana*[282–284] Overexpression of *JMJ*16 results in a global decrease in di- and tri- methylated H3K4 but an increase in H3K4me1 levels. The occupancy of *JMJ16* is taken up by the demethylation of H3K4me2/3, thus H3K4me1 is not demethylated, leading to its enhanced levels[285]. A genome-wide RNA-seq analyses of Col-0 and a *JMJ16* mutant line revealed 3355 genes upregulated and 2594 downregulated in the mutant line, out of 15988 and 15842 genes mapped in the wild-type and *JMJ16* mutant respectively.[285] Alongside this, it was shown that in the *JMJ16* mutant H3K4me3 levels were increased for 4539 genes when compared to the wild-type. 370 of these genes were also linked to transcriptional upregulation, suggestive of a direct role of *JMJ16* in regulating these loci.[285]

4.1.1.2.4 JMJ17

Like *JMJ16*, *JMJ17* is a H3K4me1/2/3 specific demethylase[286], as in the *JMJ17* mutant there was a substantial reduction in H3K4me1/2/3 levels but no changes in the levels of H3K27me3, H3K36me3 or H3K9me3.[287] *A. thaliana JMJ17* is known to play a crucial role in the response to dehydration stress and abscisic acid (ABA) signalling. *JMJ17* is selectively expressed in the veins of cotyledons, hypocotyls and roots at the seedling stage, and in in leaves, primary and secondary roots in later plant growth stages.[287] It has been shown that *JMJ17* is recruited to specific target loci by the *WRKY40* (WRKY DNA-Binding protein 40) transcription factor. [286,288]

4.1.1.2.5 JMJ27

Recently, *JMJ27* has been characterised as a H3K9 demethylase that modulates both plant defence and flowering time [289] *JMJ27* has been shown to reduce H3K9me1/2/3 methylation, with the rate of demethylation higher for H3K9me1 and H3K9me3 than H3K9me2.[289] Importantly, *JMJ27* does not show demethylation activity against H3K4, H3K27, H3K36, H3R2 and H4R3. The early flowering phenotype of *JMJ27* mutants were shown to be due to enhanced levels of H3K9me2 levels at the promoters of the key flowering genes *FLC* and *CO* (Constans) promoter. *JMJ27* has also been found to regulate several WRKY transcription factors, such as *WRKY25*.[289]

4.1.1.2.6 *JMJ*30 and *JMJ*32

JMJ30 and *JMJ32* are suggested to be involved in the demethylation of H3K27. Both enzymes are involved in the regulation of the developmental switch between vegetative and reproductive growth and expression of the flowering gene *FLC*. *JMJ30* and *JMJ32* are ubiquitously expressed in *A. thaliana*[282]. Overexpression of *JMJ30* and *JMJ32* leads to reduced H3K27me3 levels.[290] Transcriptome profiling revealed that *JMJ30/JMJ32* show an activating function in gene expression in accordance with their role in the demethylation of the repressive chromatin mark H3K27me3.[290–292]

4.1.1.3 Histone methyltransferases

All histone methyltransferases use a similar catalytic mechanism. The human genome encodes for five methyltransferases. The signature motifs of methyltransferases are conserved between plants and animals.[293] The conserved SET (Su(var) 3-9,E(z),Trithorax) domain mediates the transfer of the methyl group onto lysine or arginine residues of the histone tail.[294] *SDG* (SET domain group) proteins are a large family of methyltransferases in plants. This family is divided into the subfamilies E(z), ASH1, TRX and SU(VAR)3-9, in which *A. thaliana* contains 37 proteins.[295] The E(z) subfamily encodes for three methyltransferases in *A. thaliana* - Curly Leaf (*CLF*), Medea (*MEA*) and Swinger (*SWN*). These form the Polycomb Repressive Complex 2 (*PRC2*) protein complex that is trimethylating H3K27 and is associated with gene repression.[296] The Trithorax group (TrxG) subfamily has several functional categories, such as chromatin remodelling, histone methyltransfer and demethylation.[297] These enzymes can catalyse di- or trimethylation of H3K4me or H3K36, and are considered to have PcG antagonistic effects on gene expression.[298] The Ash subfamily also functions through TrxG and acts on H3K36, H3K27 and H3K4 methylation.[299–301] The Suv subfamily contains the largest set of genes with more complicated structures than the other subfamilies.[296] This family of proteins is primarily located in heterochromatin and can carry out mono- and demethylation of H3K9, H3K27 and H4K20.[302]

4.1.1.3.1 ATX

There are five conserved Trithorax-type histone H3K4 methyltransferase in *A. thaliana*, *ATX* 1-5 (Arabidopsis trithorax 1-5), which can further be divided into two subgroups: *ATX1*/2 and *ATX3*/4/5.[303] The promotion of H3K4 trimethylation and transcriptional activation of *FLC* to suppress flowering involves *ATX1* and *ATX2*.[304] *ATX 4 and 5* are closely related to each other, whereas *ATX3* is distantly related to *ATX4*, and they function redundantly in vegetative and reproductive development.[303,305] *ATX1*/2 are of different origin than *ATX3*/4/5, and *ATX1*/2 is likely to have originated from chromosomal segmental duplication.[303] Alongside this, *ATX1*, *ATX4* and *ATX5* have been implicated in the signalling response to abscisic acid and in dehydration tolerance.[306] *ATX4*/5 has been found to associate with a class of chromatin remodellers, *INO80*[307]. *INO80* (Inositol requiring 80) has been implicated in transcriptional regulation, DNA replication and repair in *A. thaliana*[307]. COMPASS (Complex proteins associated with Set1) are known to catalyse H3K4 methylation in *A. thaliana*[308]. The histone H3K9me2 demethylase *JMJ24* has been found to be responsible for acting as a bridge protein to assemble COMPASS on INO80 with the accessory subunits *ATX4*/5.[307] Unlike *ATX4* and *ATX5* single mutants which present a wild type morphology, *ATX4*/5 double mutant show retarded growth when compared to the Col-0 wildtype. A triple mutant of *ATX4*/5 and *JMJ24* shows a further pronounced growth retardation a significantly shortened flowering time. Genome-wide ChIPseq analyses showed that about 12 % of H3K4me3 marked genes lose this chromatin modification in the triple *ATX4*/5 and *JMJ24* mutant compared to the wild type.

Although *ATX3, ATX4* and *ATX5* have all been implicated in H3K4me3 methylation, recent data also suggest that these proteins are involved in dimethylation of H3K4 in *A. thaliana*.[305] *ATX3/4/5* triple mutants present morphological abnormalities past 2 weeks of growth are significantly smaller than the wild-type, are produce small leaf rosettes.[305] Levels of both H3K4me2 and H3K4me3 marks are consistently reduced in the triple mutant compared to the wild-type In contrast, H3K4me1, H3K9me2, H3K27me3 and H3K36me3 levels are not changed in *ATX3*/4/5. *ATX3*, *ATX4* and *ATX5* act redundantly and operate in a dosage-dependent manner. *ATX5* exhibits a stronger effect than *ATX4* and *ATX3* due to it being the only single mutant to show a decrease in H3K4me2/3, although this could be due to *ATX5* being more highly expressed[305].

4.1.1.3.2 SDG2

SDG2 is a histone methyltransferase. It plays a crucial role in both sporophyte and gametophyte development.[309] Loss of function *SDG2* mutants display fully sterile plants that are smaller than the wild-type. *SDG2* mutant plants show drastically reduced levels of H3K4me3, slightly reduced level of H3K4me2, enhanced levels of H3K4me1, yet unchanged levels of H3K36me1, H3K36me3, and H3K27me3.

4.1.1.3.3 SDG4

SDG4 is associated with the methylation of H3K4 and H3K36me3. The *SDG4* mutant line shows a significant reduction in H3K4me2/3 and H3K36me3.[310] It has been shown to be crucial for pollen germination and growth of the pollen tube[310,311]. It is suggested to function in conjunction with other proteins to methylate its target histone residues [310,312]

4.1.1.3.4 SDG25

SDG25 is a methyltransferase suggested to be involved in the di-and trimethylation of H3K4 and H3K36. *SDG25* has been shown to be regulator of the flowering genes *FLC*, *LFC* and *FT*. As such, it plays in important role in the control of flowering time in *A. thaliana*. [313]

4.1.2 Project aims

A wide range of enzymes modify and interact with chromatin to regulate gene expression. Here, we aim to identify the key chromatin modifications and chromatin modifying enzymes that are involved in the co-regulation of metabolic gene clusters in *A. thaliana*. To achieve this goal, we apply a combination of bioinformatics and molecular genetics approaches.

4.2 Identification of key histone modification marks

To identify the histone modification marks that are associated with metabolic gene clusters in *A. thaliana*, we analysed 143 publicly available histone modification datasets. These datasets enable us to monitor the distribution of 23 individual chromatin marks across the exemplar *A. thaliana* metabolic gene clusters thalianol, marneral and arabidiol/baruol. These three clusters consist of independent and different sets of metabolic genes covering 35 – 120 kb of DNA on chromosomes 4 and 5. To map the chromatin datasets to our target clusters, we used the ReMap database that provides consistently pre-analysed data tracks for all modification marks (https://remap2022.univ-amu.fr/about_atha_page). However, it should be noted that the majority of experiments were carried out under different conditions. As control regions for the analyses of the datasets, we used 100 random regions across the genome that were selected by random number generation for gene locus, chromosome, and region size (between 20,000 and 200,000 bp). The average enrichment of histone modification for the 100 random regions was used as a baseline control for statistical analysis. A hypergeometric statistical test was applied to score the probability of histone modification enrichment within our regions of interest (gene cluster) compared to the randomly expected modification enrichment (random region or whole genome). The null hypothesis is that there is no significant difference between the number of histone modification marks in the target region of interest compared to randomly selected regions. The null hypothesis can be rejected if the P-value (P) is equal to zero and accepted if the P-value is equal to one.

First, we performed the outlined analyses independently for the thalianol cluster associated active domain and the thalianol cluster associated inactive domain, using the co-ordinates of the domains based on previously published Hi-C data[140] (see Section **3.2.1.1**, ***Figure 8*** and ***Table 19***). Then, we analysed chromatin modification enrichment across all three gene clusters. For each analysis, we determined the fold enrichment of marks at our target regions compared to the whole genome.

As a result of our analyses, overview shown in ***Figure 58***, we were able to identify a core set of chromatin marks that is enriched in all three cluster regions analysed. Amongst these marks, we identified H3K18ac and H3K4me3 – histone modification marks that have not been implicated as regulators of metabolic gene clusters before. Our data also confirms that H3K27me3 is a repressive cluster mark. [163]

		Random region	Arabidiol cluster	Marneral cluster	Inactive thalianol domain	Active thalianol domain
Aerial	H2A-Z 13 days	1	1.5	1.0	1.5	1.8
	H2A-Z 3 weeks	1	1.4	1.1	1.3	1.7
Leaf	H2A.Z	1	1.6	1.0	1.4	1.7
	H2A.Z 3 weeks	1	1.4	1.1	1.2	1.6
	H2A 3 weeks	1	0.2	0.8	1.5	0.3
	H2A.X 3 weeks	1	0.3	1.0	1.3	0.0
	H3-3 4 weeks	1	0.3	0.8	1.8	0.1
	H3ac 4 weeks	1	0.7	0.6	1.4	0.3
	H3K4me3	1	0.6	0.8	1.6	0.6
	H3K27me3	1	0.6	0.5	1.3	0.2
	H3K27me3 20 days	1	3.4	2.3	1.3	3.3
	H3K36me3	1	0.6	0.5	1.2	0.7
	H3K36me3 3 weeks	1	0.4	0.8	1.4	0.2
	H3K36ac 35 days	1	0.5	0.6	1.6	0.5
Seedling	H2A.Z 6 days	1	1.3	1.0	1.1	0.9
	H2A.Z 10 days	1	1.4	1.1	1.3	1.1
	H2AK121ub 7 days	1	1.2	0.9	1.3	1.5
	H3-1 10 days	1	1.1	0.9	1.1	0.9
	H3-3 10 days	1	0.5	0.9	1.4	0.5
	H3 3 weeks	1	0.9	1.0	1.3	0.6
	H3K9ac 10 days	1	1.9	0.9	1.6	0.5
	H3K9me2 10 days	1	2.8	1.3	0.8	1.1
	H3K14ac 10 days	1	0.9	0.6	1.5	1.1
	H3K18ac 12 days	1	8.5	4.5	1.9	3.7
	H3K23ac 3 days	1	1.5	2.7	1.4	0.0
	H3K27ac 12 days	1	1.3	0.7	1.4	0.3
	H3K27me1 20 days	1	0.2	0.7	0.2	0.2
	H3K27me3 7 days	1	4.4	3.1	1.4	4.1
	H3K27me3 10 days	1	4.4	3.1	1.3	4.0
	H3K27me3 12 days	1	5.1	3.5	1.5	4.6
	H3K27me3 14 days	1	5.6	3.5	1.6	4.8
	H3K36me3 10 days	1	0.3	0.8	0.4	0.5
	H4ac 6 days	1	0.5	0.6	1.4	0.3
	H4K5ac 12 days	1	0.6	0.5	1.4	0.5
	H4K8ac 12 days	1	0.6	0.5	1.4	0.4
	H4K12ac 12 days	1	0.7	0.5	1.4	0.4
Root	H3ac 4 weeks	1	0.7	0.6	1.4	0.3
	H3ac 7 days	1	0.8	0.6	1.4	0.4
	H3K4me3 5d	1	3.1	2.4	1.2	2.3
	H3K4me3 14 days	1	2.6	0.2	1.0	0.0
	H3K27me3 5 days	1	0.4	0.8	1.3	0.2
	H3K27me3 5 days	1	0.9	0.6	1.5	1.1
Whole plant	H3-3 2 weeks	1	0.3	0.7	1.6	0.3
	H3K27me3 30 hours	1	3.5	2.5	2.1	6.8

> 1.5
< 0.6
No change

Figure 59: Overview of histone modification mark enrichment in different A. thaliana tissue. *Blue represents so change in the histone modification mark in comparison to the randomly selected regions, green represents an enrichment greater than 1.5 and orange represents a decrease less than 0.6.*

4.2.1 Comparing the thalianol interacting region to the thalianol cluster

Using the number of marks across a region of interest of a specific size, the fold enrichment of that region was calculated compared to the marks across the whole genome. The values were then normalised to that of the random regions, thus allowing to see the enrichment of marks across the region of interest.

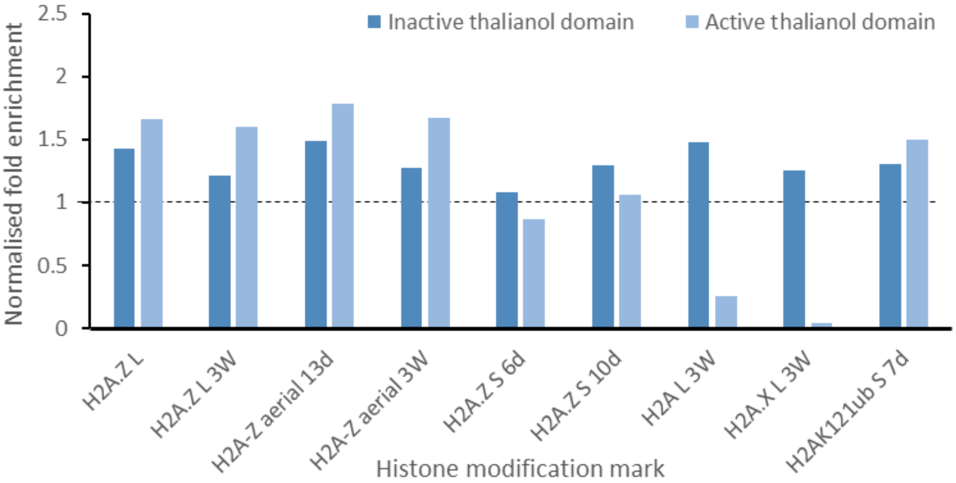


Figure 60: Enrichment of H2A variants across the core thalianol cluster and downstream interacting region. *Using data obtained from ReMap we analysed the enrichment of H2A variants across the five thalianol cluster genes (active thalianol domain) and the larger downstream region (inactive thalianol domain) that interacts with the cluster when it is repressed. Fold enrichment was determined against the histone variant enrichment across the 100 randomly selected regions for comparison, indicated by the dotted line. Plant material used for analyses is indicated by leaf (L), seedling (S) or aerial part and the time at which the sample was taken in days (d) or weeks (W).*

First, H2A histone marks and its variants were analysed, shown in ***Figure 59***. H2A.Z has been shown to be a major regulator of transcription and control of chromatin architecture. Furthermore, it has been shown that it can exert both active and repressive functions on gene transcription[314]. In the leaves of both the active and inactive thalianol domains there is enrichment of H2A.Z when compared to the random region. The level of H2A is enriched in the inactive thalianol interacting region but is reduced in the active thalianol cluster in leaves. Interestingly, we also see a reduction in H2A.X at the active thalianol cluster in leaves.

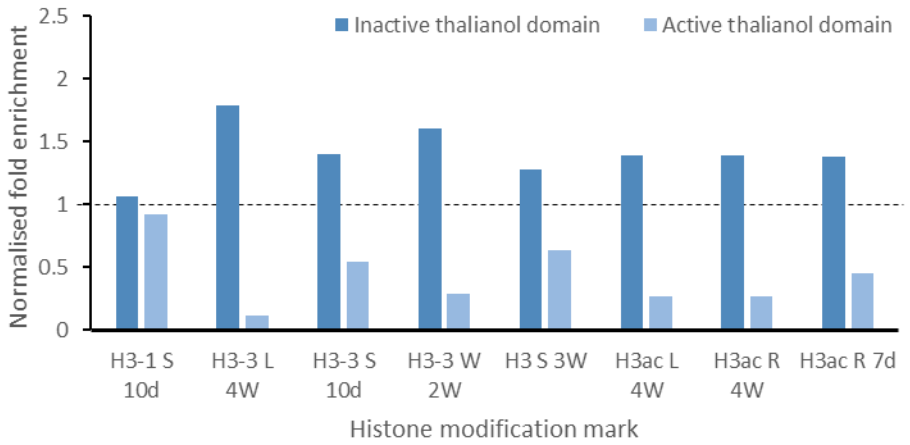


Figure 61: Enrichment of H3 variants and H3ac across the core thalianol cluster and downstream interacting region. *Using data obtained from ReMap we analysed the number H3 variants and H3ac across the five thalianol cluster genes (active thalianol domain) and the larger downstream region (inactive thalianol domain) which interacts with the cluster when it is repressed. Fold enrichment was determined against the histone variant enrichment across the 100 randomly selected regions for comparison, indicated by the dotted line. Plant material used for analyses is indicated by leaf (L), seedling (S), whole plant (W) or root (R) and the time at which the sample was taken in days (d) or weeks (W).*

Next, we looked at histone H3 enrichment across our target regions (***Figure 60***). In leaves, seedlings and whole plant, H3-3 levels are reduced in the active thalianol cluster domain when compared to the inactive thalianol interacting region in seedlings and whole plant. The histone H3-3 variant has been implicated in active gene transcription[315], therefore, reduced levels at the thalianol cluster in leaves are as expected, and a higher number in the larger inactive domain where other genes are being expressed. Acetylation of H3 (H3ac) is also implicated in active gene transcription[316]. H3ac levels are indeed low in the thalianol cluster whereas the interacting region exhibits H3ac levels similar to the random control region. Surprisingly, H3ac levels remain stable in both roots and leaves.

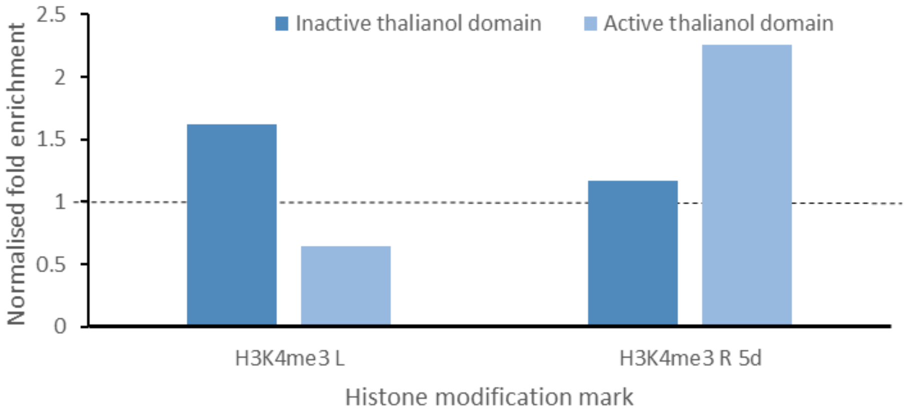


Figure 62: Enrichment of H3K4me3 marks across the core thalianol cluster and downstream interacting region. *Using data obtained from ReMap we analysed the number H3K4me3 marks across the five thalianol cluster genes (active thalianol domain) and the larger downstream region (inactive thalianol domain) which interacts with the cluster when it is repressed. Fold enrichment was determined against the histone variant enrichment across the 100 randomly selected regions for comparison, indicated by the dotted line. Plant material used for analyses is indicated by leaf (L) or root (R) and the time at which the sample was taken in days (d).*

We next analysed H3K4me3 levels. H3K4me3 is a canonically associated with gene activation.[317,318] Our analyses show that H3K4me3 is enriched at the thalianol cluster in roots, whereas its levels are reduced in leaves, ***Figure 61***.

In contrast to H3K4me3 levels, H3K27me3 is a well-known repressor of transcriptional activity. Indeed, it has been experimentally shown to be involved in metabolic gene cluster repression[161]. Multiple datasets shown an enrichment of this mark at the thalianol cluster leaves when the cluster is silenced, ***Figure 62***. In roots, when the cluster is expressed, a significant reduction in H3K27me3 levels can be observed. Interestingly, H3K27me3 levels in the large thalianol interacting region remain consistent across all time points and material types.

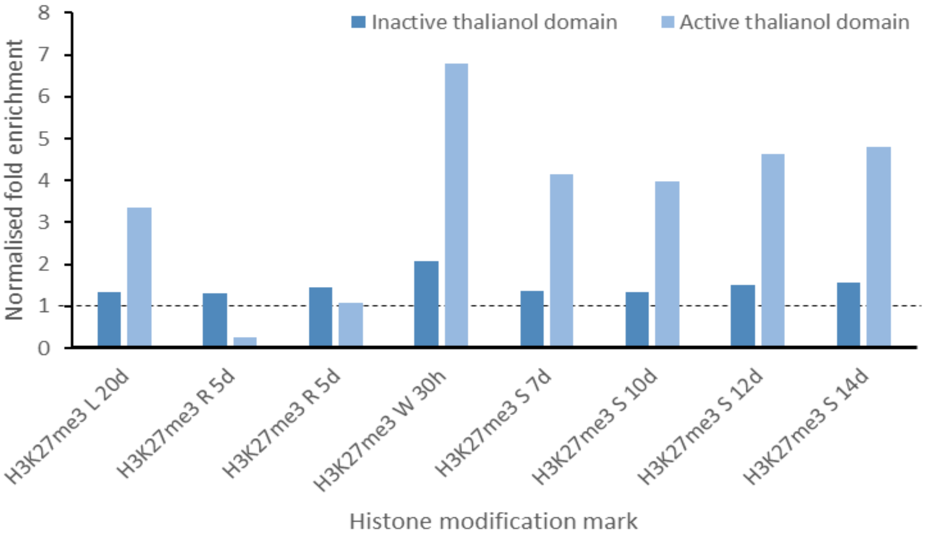


Figure 63: Enrichment of H3K27me3 marks across the core thalianol cluster and downstream interacting region. *Using data obtained from ReMap we analysed the number H3K27me3 across the five thalianol cluster genes (active thalianol domain) and the larger downstream region (inactive thalianol domain) which interacts with the cluster when it is repressed. Fold enrichment was determined against the histone variant enrichment across the 100 randomly selected regions for comparison, indicated by the dotted line. Plant material used for analyses is indicated by leaf (L), seedling (S), whole plant (W) or root (R) and the time at which the sample was taken in hours (h) or days (d).*

H3K9ac, H3K14ac, H3K23ac, H3K18ac and H3K27me1 have all been previously implicated as gene activators, while H3K9me2 acts as a gene repressor[318–321]. Here we have detected a lowered enrichment of H3K27me1 in seedlings in both thalianol regions that were analysed when compared to the random region, ***Figure 63***. We find that H3K9ac and H3K27ac levels are reduced in the thalianol cluster but enriched in the inactive thalianol interacting region. No H3K23ac signals can be detected at the thalianol cluster, while H3K23ac levels throughout the larger interacting region are comparable to the random control regions in seedlings. Interestingly, in the 12-day old seedling data for H3K18ac there is a large increase in marks at the active thalianol cluster region. The H3K4me2 mark shows similar enrichment levels for both investigated regions to the random control regions.

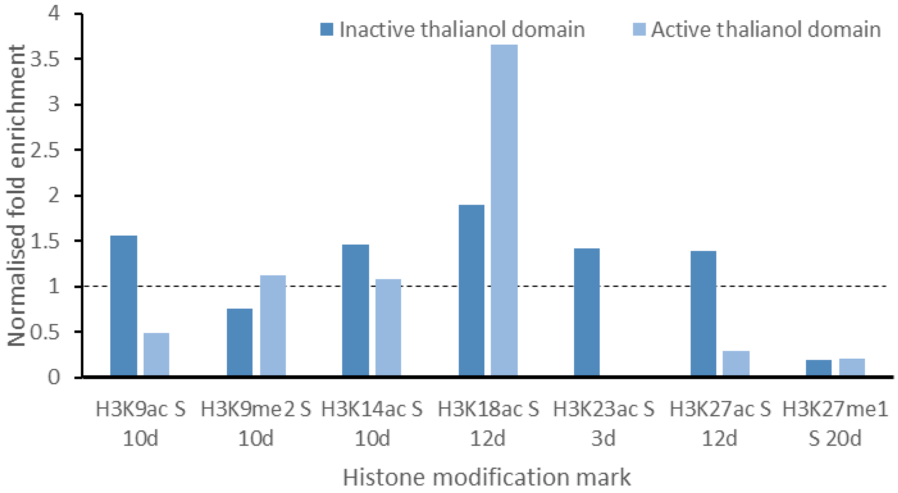


Figure 64: Enrichment of H3K9ac, H3K9me2, H3K14acm H3K18ac, H3K23ac, H3K27ac and H3K27me1 marks across the core thalianol cluster and downstream interacting region. *Using data obtained from ReMap we analysed the number H3K9ac, H3K9me2, H3K14acm H3K18ac, H3K23ac, H3K27ac and H3K27me1 marks across the five thalianol cluster genes (active thalianol domain) and the larger downstream region (inactive thalianol domain) which interacts with the cluster when it is repressed. Fold enrichment was determined against the histone variant enrichment across the 100 randomly selected regions for comparison, indicated by the dotted line. Plant material used for analyses is indicated by seedling (S) and the time at which the sample was taken in days (d).*

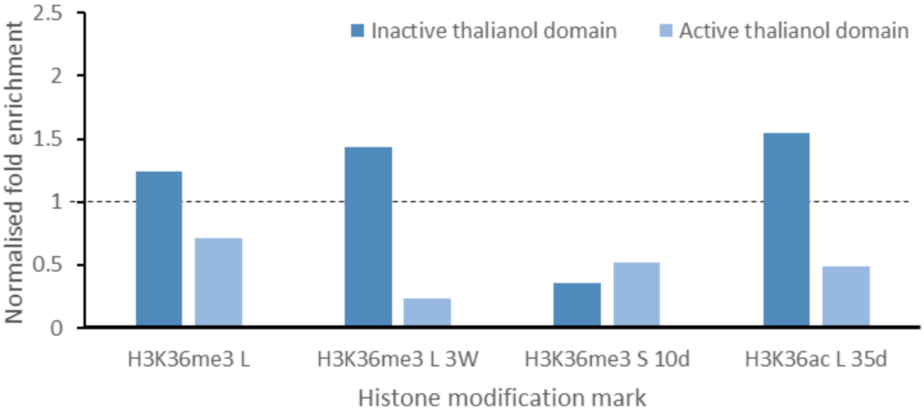


Figure 65: Enrichment of H3K36me3 and H3K36ac marks across the core thalianol cluster and downstream interacting region. *Using data obtained from ReMap we analysed the number H3K36me3 and H3K36ac across the five thalianol cluster genes (active thalianol domain) and the larger downstream region (inactive thalianol domain) which interacts with the cluster when it is repressed. Fold enrichment was determined against the histone variant enrichment across the 100 randomly selected regions for comparison, indicated by the dotted line. Plant material used for analyses is indicated by leaf (L) or seedling (S) and the time at which the sample was taken in days (d) or weeks (W).*

The activation mark, H3K36me3, has the potential to be an important activator of the thalianol cluster due to its antagonistic role .[322] In the leaves for the active thalianol domain there is reduced enrichment when compared to the randomly selected region and the inactive domain, indicating H3K36me3's role in activation, ***Figure 64***. Root specific data is not available; however, it is notable that seedling data, including both root and leave material, shows a reduction in H3K36me3 enrichment in the inactive domain compared to the active domain. The enrichment in both domains, however, is lower than in the randomly selected regions. The H4ac, H4K5ac, H4K8ac and H4K12ac marks have all been implicated active gene expression.[323] The data we obtained for these histone modification marks are all seedling specific, ***Figure 65***. All four marks showed an increased enrichment in the silencing inactive domain and decreased levels in the active domain. These marks have the potential to be involved in thalianol cluster regulation, however as we do not have separate root and leaf data it cannot be confirmed.

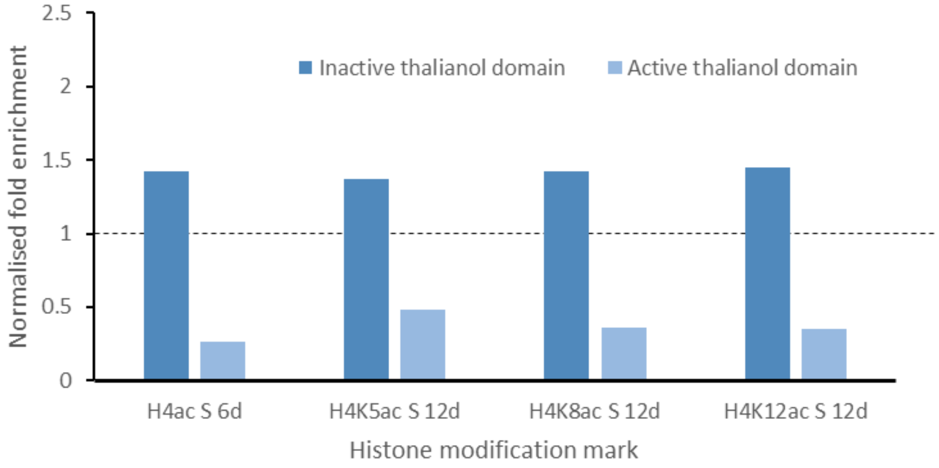


Figure 66: Enrichment of H4ac, H4K5ac, H4K8ac and H4K12ac marks across the core thalianol cluster and downstream interacting region. *Using data obtained from ReMap we analysed the number H4ac. H4K5ac, H4K8ac and H4K12ac marks across the five thalianol cluster genes (active thalianol domain) and the larger downstream region (inactive thalianol domain) which interacts with the cluster when it is repressed. Fold enrichment was determined against the histone variant enrichment across the 100 randomly selected regions for comparison, indicated by the dotted line. Plant material used for analyses is indicated by seedling (S) and the time at which the sample was taken in days (d).*

4.2.2 Marks across all clusters

To understand if biosynthetic gene clusters in *Arabidopsis thaliana* have a signature histone modification code, histone modification mark data was also analysed for the marneral and arabidiol clusters.

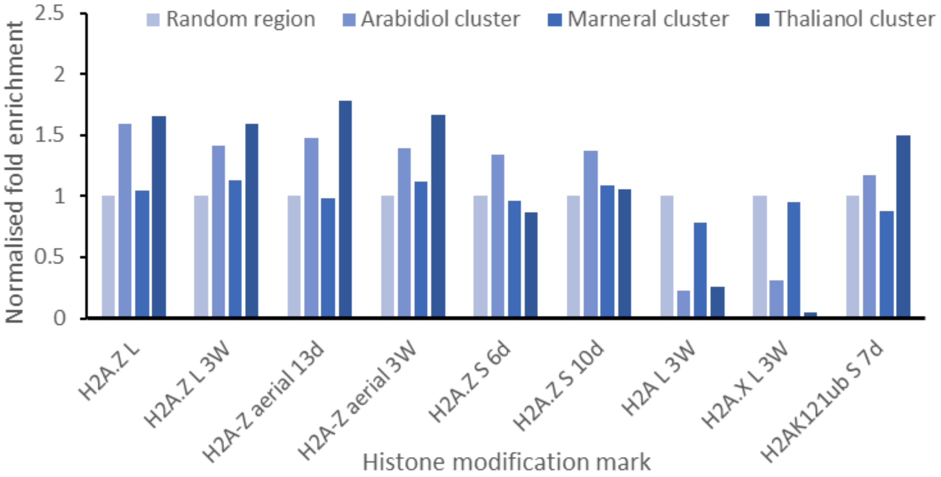


Figure 67: Enrichment of H2A variants across the arabidiol, marneral and thalianol cluster. *Using data obtained from ReMap we analysed the number H2A variants across the arabidiol, marneral and thalianol gene cluster. Fold enrichment was found compared to the whole genome and normalised to the 100 randomly selected regions for comparison, indicated by the dotted line. Plant material is indicated by leaf (L), seedling (S) or aerial part and the time at which the sample was taken in days (D) or weeks (W).*

Across all the marks shown in ***Figure 66*** it appears that the thalianol and arabidiol cluster follow a similar pattern for H2A histone modification marks. In leaves both the arabidiol and thalianol clusters show an enrichment in H2A.Z, no change in H2A.Z in seedlings and a decrease in H2A and H2A.X when compared to the randomly selected region. Interestingly, the marneral cluster showed the same pattern as the randomly selected region for these histone modification marks.

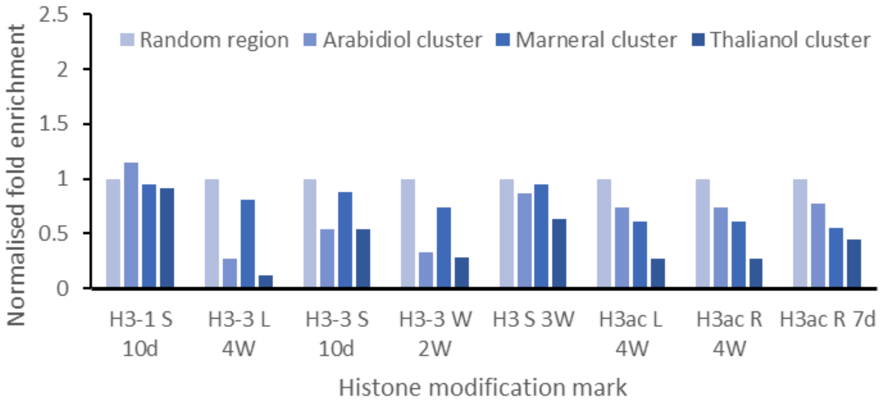


Figure 68: Enrichment of H3 variants and H3ac variants across the arabidiol, marneral and thalianol cluster. *Using data obtained from ReMap we analysed the number H2A variants across the arabidiol, marneral and thalianol gene cluster. Fold enrichment was found compared to the whole genome and normalised to the 100 randomly selected regions for comparison, indicated by the dotted line. Plant material is indicated by leaf (L), root (R), whole plant (W) or seedling (S) and the time at which the sample was taken in days (d) or weeks (W).*

As previously noted, the H3.1 histone variant did not show significant differences in enrichment at the thalianol cluster compared to our controls, ***Figure 67***. We detected this pattern consistently across all three clusters. The H3ac mark showed a lowered enrichment in the leaves and roots of all three clusters. We observed a similar trend for H3.3 marks in leaves, seedlings and whole plants. Notably, the marneral cluster showed least reduction across all three datasets.

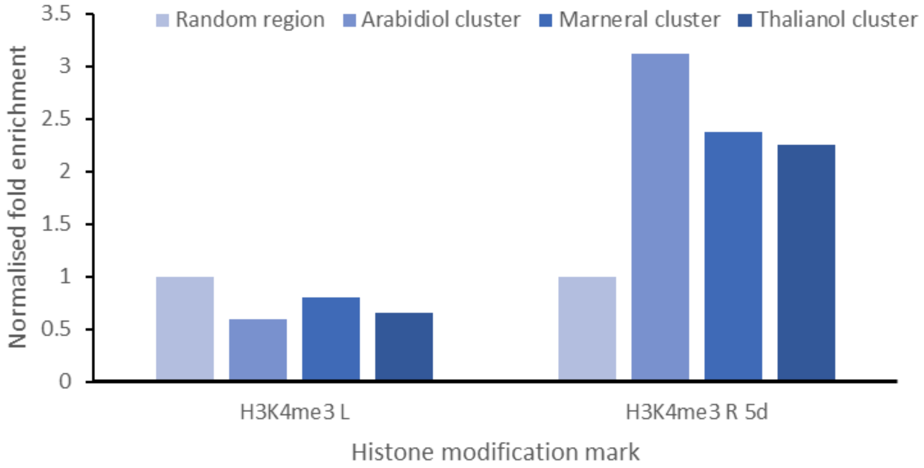


Figure 69: Enrichment of H3K4me3 marks across the arabidiol, marneral and thalianol cluster. *Using data obtained from ReMap we analysed the number H2A variants across the arabidiol, marneral and thalianol gene cluster. Fold enrichment was found compared to the whole genome and normalised to the 100 randomly selected regions for comparison, indicated by the dotted line. Plant material is indicated by leaf (L) and root (R) and the time at which the sample was taken in days (d).*

H3K4me3 chromatin modifications show strong enrichment across all three clusters in roots, ***Figure 68***. In leaves, the activation mark is slightly reduced when compared to the random region.

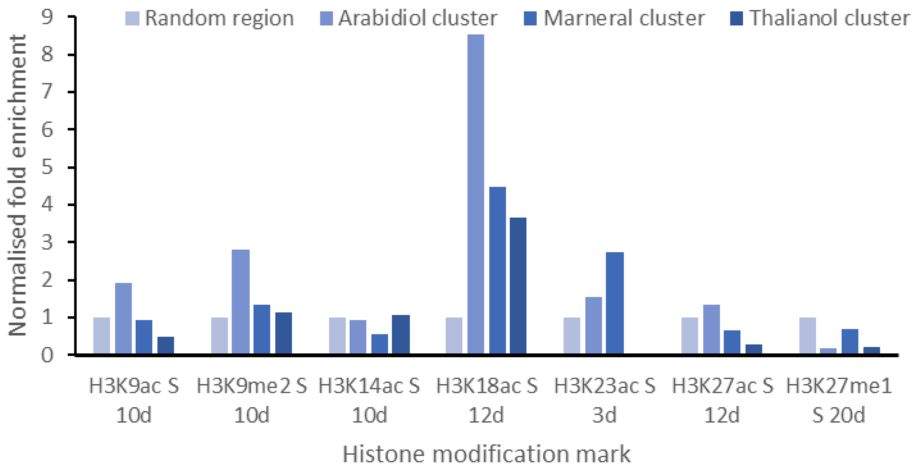


Figure 70: Enrichment of H3K9ac, H3K9me2, H3K14acm H3K18ac, H3K23ac, H3K27ac and H3K27me1 marks across the arabidiol, marneral and thalianol cluster. *Using data obtained from ReMap we analysed the number H2A variants across the arabidiol, marneral and thalianol gene cluster. Fold enrichment was found compared to the whole genome and normalised to the 100 randomly selected regions for comparison, indicated by the dotted line. Plant material is indicated by seedling (S) and the time at which the sample was taken in days (d).*

With the exception of H3K18ac, which showed consistent enrichment, we observed variable enrichment pattern at the clusters for the activating marks H3K9ac, H3K14ac, H3K23ac, and H3K27me1, ***Figure 69***. Interestingly, we saw significant enrichment at the arabidiol cluster for H3K9 ac an H3K9me3 and strong enrichment for H3K23ac at the marneral cluster.

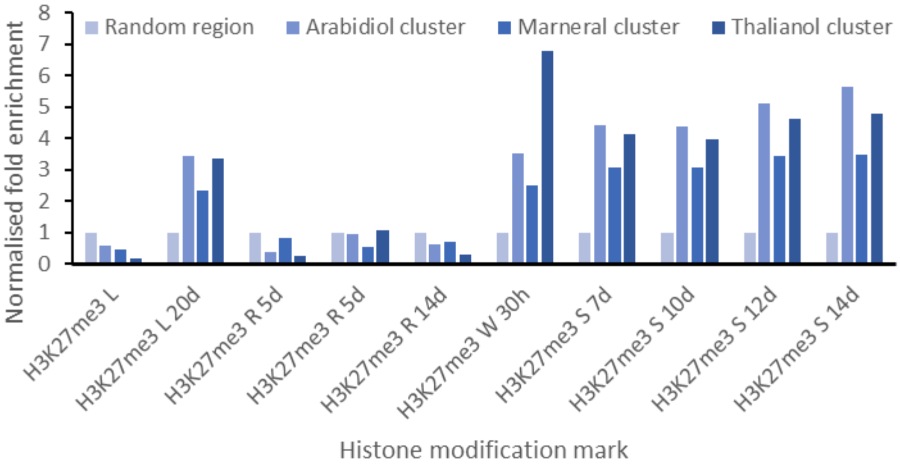


Figure 71: Enrichment of H3K27me3 across the arabidiol, marneral and thalianol cluster. *Using data obtained from ReMap we analysed the number H2A variants across the arabidiol, marneral and thalianol gene cluster. Fold enrichment was found compared to the whole genome and normalised to the 100 randomly selected regions for comparison, indicated by the dotted line. Plant material is indicated by leaf (L), root (R), whole plant (W) or seedling (S) and the time at which the sample was taken in hours (h) or days (D).*

The repressive chromatin mark H3K27me3 in contrast showed strong enrichment at all clusters in leave and seedling stages, ***Figure 70***.

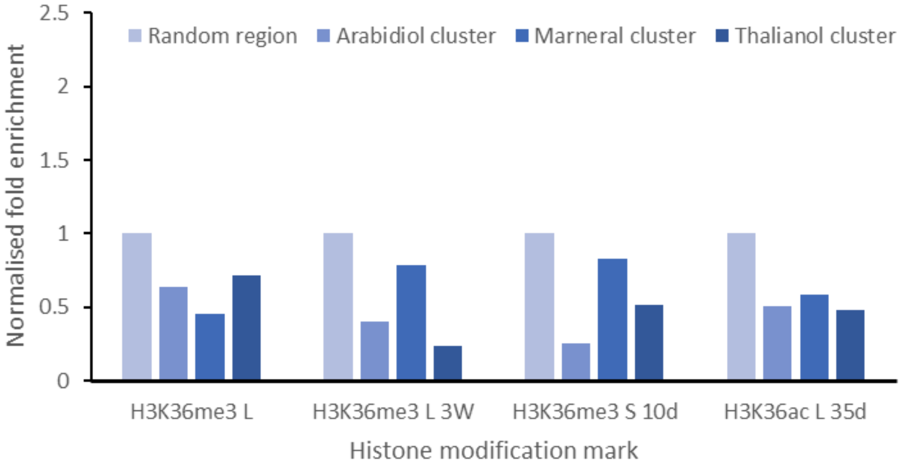


Figure 72: Enrichment of H3K36me3 marks across the arabidiol, marneral and thalianol cluster. *Using data obtained from ReMap we analysed the number H2A variants across the arabidiol, marneral and thalianol gene cluster. Fold enrichment was found compared to the whole genome and normalised to the 100 randomly selected regions for comparison, indicated by the dotted line. Plant material is indicated by leaf (L) or seedling (S) and the time at which the sample was taken in days (d) or weeks (W).*

The activation chromatin mark marker, H3K36me3, is decreased in all three clusters in the leaves and whole plant, ***Figure 71***.

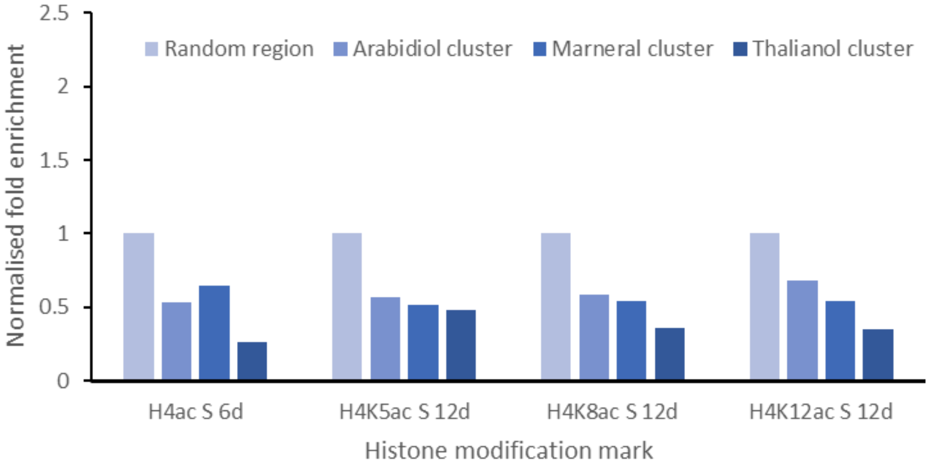


Figure 73: Enrichment of H4ac, H4K5ac, H4K8ac and H4K12ac marks across the arabidiol, marneral and thalianol cluster. *Using data obtained from ReMap we analysed the number H2A variants across the arabidiol, marneral and thalianol gene cluster. Fold enrichment was found compared to the whole genome and normalised to the 100 randomly selected regions for comparison, indicated by the dotted line. Plant material is indicated by seedling (S) and the time at which the sample was taken in days (d).*

For all three activating marks, H4ac, H4K5ac, H4K8ac and H4K12ac, we detected a stable depletion across all clusters, ***Figure 72***.

4.3 Selection of histone modification writer and eraser mutants

Based on our ReMap studies, four histone modification marks were selected for mutant analysis: H3K9, H3K4, H3K27 and H3K36. A literature screen was carried out (**Section 4.1**) to identify key enzymes responsible for the deposition or removal of the methyl group of these marks. Mutant lines for these key enzymes were then selected and analysed for their expression pattern in our regions of interest. The mutant lines and the reasoning for their selection will be described in the following section, ***Table 26***.

4.3.1 Histone modification mutant's transcript analysis

In the following section, we will describe the results of our comparative analysis of the thalianol cluster transcript levels in wild type and histone modifier mutant plants. Homozygous mutant lines were used throughout, shown in **4.3.13.**

Table 26: Proposed modifiers of the metabolic gene cluster histone modifications. *Top section is the histone modifiers we analysed; bottom is proposed modifiers to look at in future*

Mutant line	Protein mutated	Function of mutated protein	References
Salk_149692	*SDG25*	H3K36me2/3 transferase	Combinatorial functions of diverse histone methylations in Arabidopsis thaliana flowering time regulation (Shafiq et al, 2014)
Salk_003313	*JMJ32*	H3K27me3 demethylase	Jumonji demethylases moderate precocious flowering at elevated temperature via regulation of FLC in Arabidopsis (Han et al, 2014)
Salk_092672	*JMJ27*	H3K9me1/2 demethylase	*JMJ*27, an Arabidopsis H3K9 histone demethylase, modulates defense against Pseudomonas syringae and flowering time (Dutta et al, 2017)

SAIL_535_F09	*JMJ16*	H3K4me3 and H3K9 demethylase	The Histone H3K4 Demethylase *JMJ*16 Represses Leaf Senescence in Arabidopsis (Liu et al, 2019)
Salk_021008	*SDG2*	H3K4 transferase	Arabidopsis SET DOMAIN GROUP2 Is Required for H3K4 Trimethylation and Is Crucial for Both Sporophyte and Gametophyte Development (Berr et al, 2010)
Salk_117262	*ATX2*	H3K4 transferase	*ATX*3, *ATX*4, and *ATX*5 Encode Putative H3K4 Methyltransferases and Are Critical for Plant Development (Chen et al, 2017)
Salk_034869	*LDL1*	H3K4 demethyase	Two Arabidopsis Homologs of Human Lysine-Specific Demethylase Function in Epigenetic Regulation of Plant Defense Responses (Noh, 2014)
Salk_135831	*LDL2*	H3K4 demethyase	Two Arabidopsis Homologs of Human Lysine-Specific Demethylase Function in Epigenetic Regulation of Plant Defense Responses (Noh, 2014)
GK_128H01	*ATX3*	H3K4 demethyase	*ATX*3, *ATX*4, and *ATX*5 Encode Putative H3K4 Methyltransferases and Are Critical for Plant Development (Chen et al, 2017)
Salk_029530	*JMJ16*	H3K4 demethyase	The Histone H3K4 Demethylase *JMJ*16 Represses Leaf Senescence in Arabidopsis (Liu et al, 2019)
GK_229F03	*JMJ15*	H3K4 demethyase	Over-expression of histone H3K4 demethylase gene *JMJ*15 enhances salt tolerance in Arabidopsis (Shen et al, 2014)
GK_143H01	*ATX3*	H3K4 demethyase	*ATX*3, *ATX*4, and *ATX*5 Encode Putative H3K4 Methyltransferases and Are Critical for Plant Development (Chen et al, 2017)
GK_663C11	*JMJ15*	H3K4 demethyase	Over-expression of histone H3K4 demethylase gene *JMJ*15 enhances salt tolerance in Arabidopsis (Shen et al, 2014)
Salk_128444	*SDG4*	H3K4 demethyase	The Arabidopsis SDG4 contributes to the regulation of pollen tube growth by methylation of histone H3 lysines 4 and 36 in mature pollen (Cartagena et al, 2008)
Salk_135712	*JMJ14*	H3K4 demethylase	*JMJ*14 encoded H3K4 demethylase modulates immune responses by regulating defence gene expression and pipecolic acid levels
Salk_067880	*ino80*	Chromatin remodeller	COMPASS functions as a module of the INO80 chromatin remodeling complex to mediate histone H3K4 methylation in Arabidopsis (Shang et al, 2021)
Salk_060156	*ATX4*	H3K4 transferase	COMPASS functions as a module of the INO80 chromatin remodeling complex to mediate histone H3K4 methylation in Arabidopsis (Shang et al, 2021)
CS831182	*ATX5*	H3K4 transferase	COMPASS functions as a module of the INO80 chromatin remodeling complex to mediate histone H3K4 methylation in Arabidopsis (Shang et al, 2021)

Salk_021260	*JMJ24*	demethylase	*JMJ*24 antagonizes histone H3K9 demethylase IBM1/*JMJ*25 function and interacts with RNAi pathways for gene silencing (Audennet et al, 2017)
Salk_149002	*ATX1*	H3K9 demethylase	*ATX*3, *ATX*4, and *ATX*5 Encode Putative H3K4 Methyltransferases and Are Critical for Plant Development (Chen et al, 2017)
Salk_135712	*JMJ14*	H3K4 demethylase	*JMJ*14 encoded H3K4 demethylase modulates immune responses by regulating defence gene expression and pipecolic acid levels
Salk_11455	*JMJ14*	H3K4 demethylase	*JMJ*14 encoded H3K4 demethylase modulates immune responses by regulating defence gene expression and pipecolic acid levels

4.3.2 Genotyping

To confirm homozygosity of T-DNA insertional mutants, the mutant line was genotyped by PCR using leaf DNA using T-DNA line specific left and right primers (LP and RP, respectively). An insertion-specific primer (LBb1.3) and the RP was used to confirm the insertion, a homozygous mutant will only have a positive result for LBb1.3 and RP, a heterozygous mutant will have a positive result for both primers and a wild-type will have only a positive result for LP and RP. All amplification products were within the expected size ranges of 410-710 bp for the mutant lines and 900-1100 bp for the wild type, as listed in http://signal.Salk.edu/tdnaprimers.2.html. We identified at least one homozygous mutant for 15 lines, ***Figure 73***, which will be used in subsequent gene expression analysis.

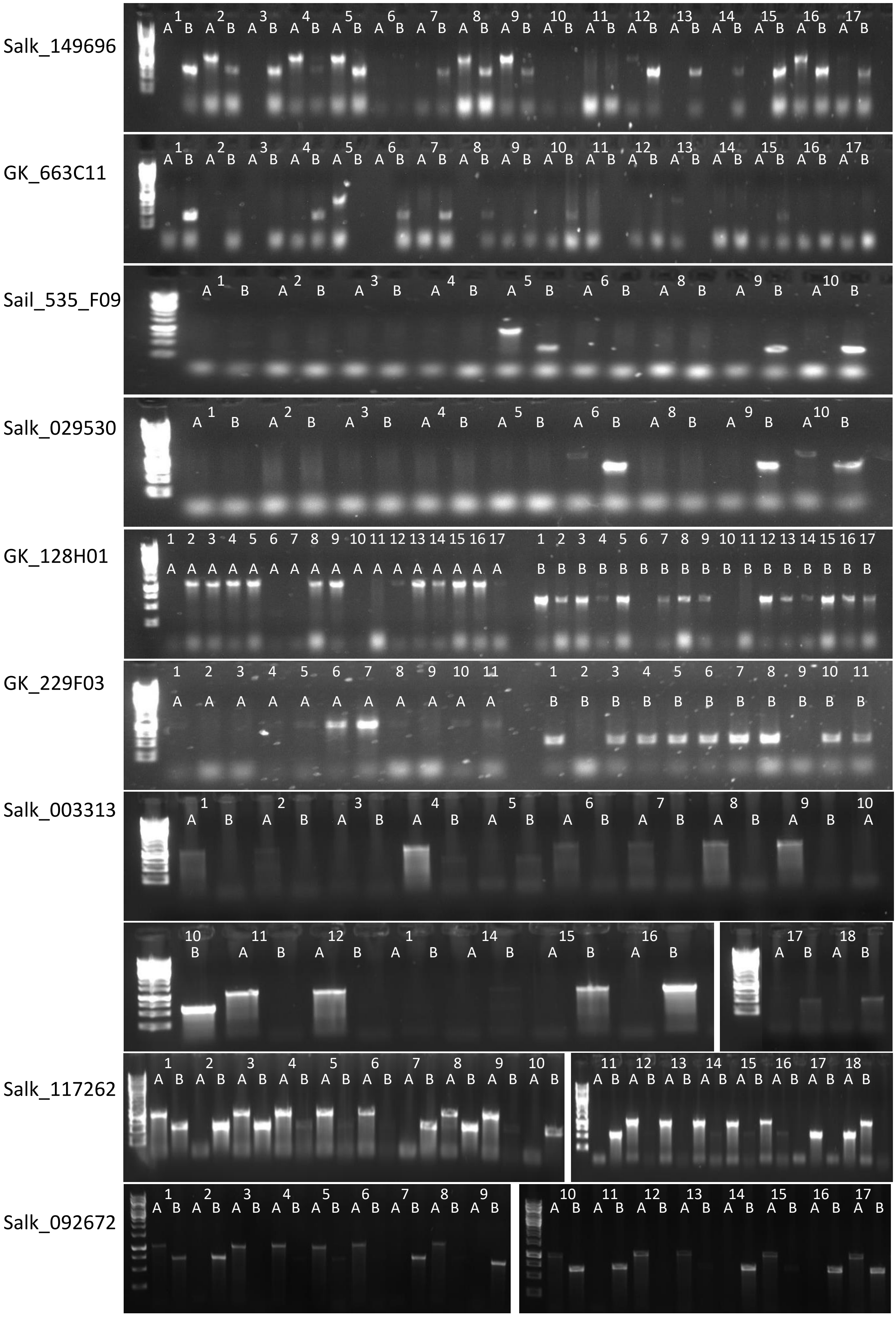

Salk_149696
GK_663C11
Sail_535_F09
Salk_029530
GK_128H01
GK_229F03
Salk_003313
Salk_117262
Salk_092672

Figure 73 continued…

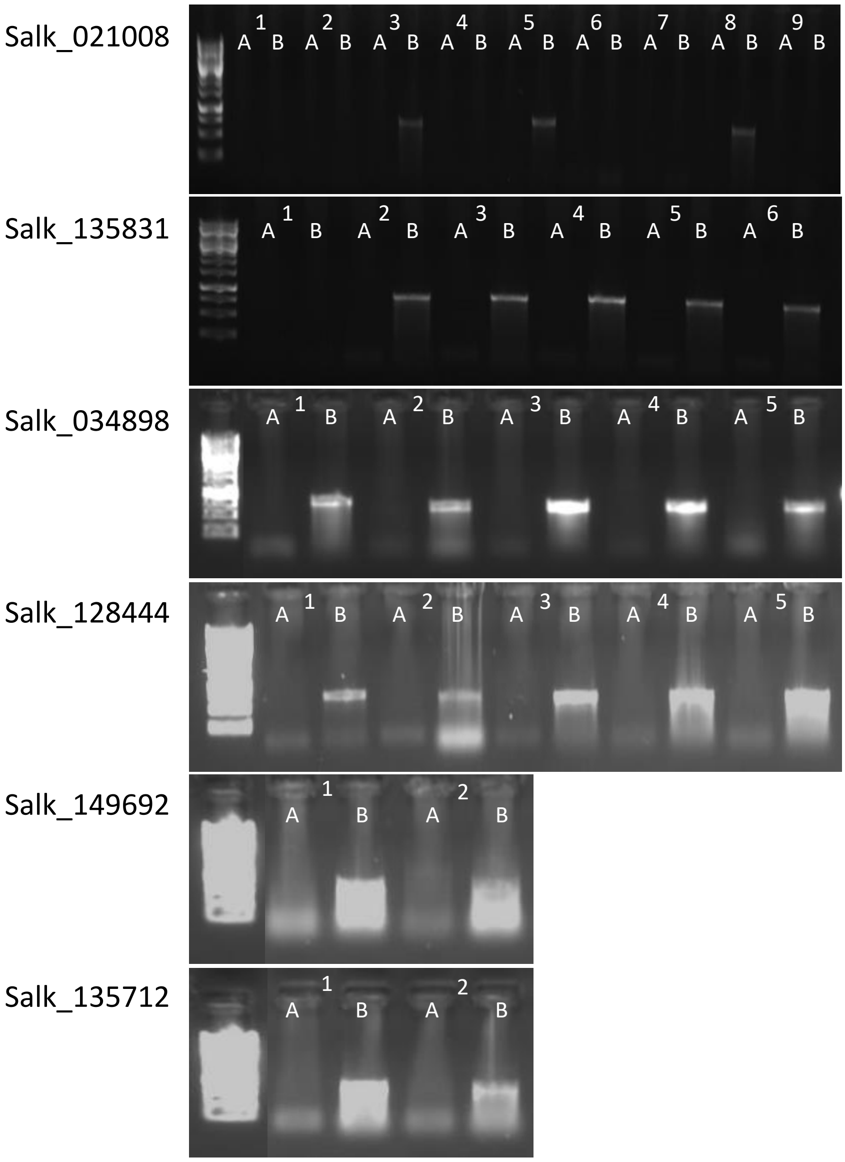


Figure 74: Agarose gel of genotyping T-DNA mutant lines. *To ensure homozygous mutants, T-DNA insertion mutants were genotyped. Name of respective T-DNA line is indicated on the left. Numbers indicate individually tested plant for each line., A refers to the LP + RP PCR reaction which will only amplify a product in wild-type DNA, B refers to the LP + LBb1.3 PCR reaction which will only amplify a product in the mutant DNA. A single band in A is wild-tpe DNA, a band in both A and B indicates a heterozygous mutant and a single band in B is a homozygous T-DNA mutant. Expected band size: 500 – 750 bp (A), 1000- 1500 bp (B).*

4.3.2.1 Transcript analysis of histone modification enzymes in the thalianol cluster

To assess if the mutated enzymes identified in ***Table 26*** are responsible for the regulation of the metabolic gene clusters in *A. thaliana* we will analyse the change in transcript expression in each T-DNA mutant line. First, we analysed the change in expression pattern for *THAD*, then if we deemed the mutant line of interest, we analysed the five thalianol cluster genes and the flanking, not co-regulated gene, *GFA2.*

4.3.2.1.1 Loss of H3K9 demethylase misregulates the thalianol gene cluster

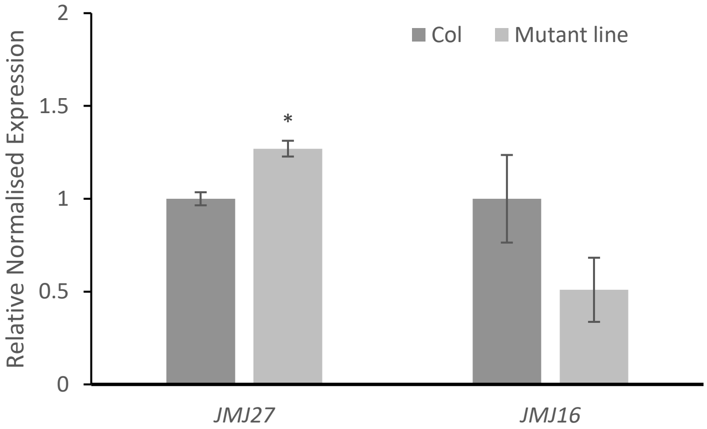


Figure 75: Transcript analysis of two H3K9 demethylase T-DNA mutants, JMJ17 and JMJ16-1. *Gene expression of THAD (At5g47990) was measured by qPCR and compared to the wild-type (Col-0) transcript level and normalised to 1. An internal control was used, PP2AA3 (AT5G13320). Error bars indicate standard error of the mean for three biological replicates. Statistical significance (paired t-test): *P < 0.05, **P < 0.01.*

H3K9me1/2 is a known repressive histone modification mark in *A. thaliana,* and are enriched in areas of heterochromatin.[22,324] *JMJ27* is a H3K9me1/2 demethylase and *JMJ16* exhibits H3K9 demethylation activity. As expected, mutation in *JMJ*27 using Salk_092672 resulted in significant (P < 0.05) increased gene expression, ***Figure 74***, of *THAD* compared to the wild type. This may implicate that *JMJ27* is acting to regulate the thalianol gene cluster by maintaining the repression of the cluster, as a mutation in this gene affects the *THAD* gene expression.

H3K9me3 and H3K4me3 have both been implicated as transcriptional activators in *A. thaliana*[23,39] in which *JMJ16*-1 has been implicated in the demethylation of both[285]. Although non-significant, the *JMJ16*-1 mutant line, SAIL_535_F09, shows a trend for downregulation of *THAD*, suggesting a removal of the activation markers (potentially H3K9me3 and H3K4me3) at this site. To further investigate the impact of the loss of *JMJ16*-1 on cluster expression levels, we analysed transcript levels across the entire thalianol cluster and its flanking genes, ***Figure 75***.

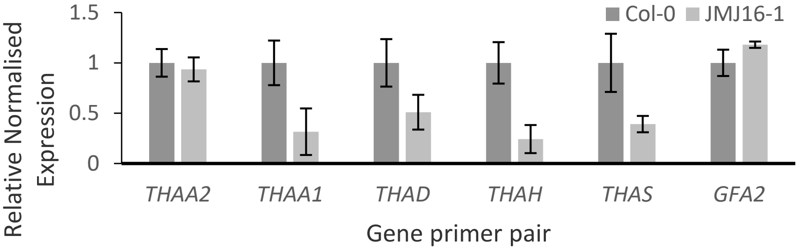


Figure 76: Transcript analysis of a H3K9 demethylase T-DNA mutant, JMJ16-1, across the thalianol gene cluster and flanking gene. *Gene expression of THAD (At5g47990) was measured by qPCR and compared to the wild-type (Col-0) transcript level and normalised to 1. An internal control was used, PP2AA3 (AT5G13320). Error bars indicate standard error of the mean for two biological replicates. Statistical significance (paired t-test): *P < 0.05, **P < 0.01.*

We detected a consistent reduction in transcript levels across the four core cluster genes *THAA1, THAD, THAH, THAS*. However, the changes to transcript levels were not significant due to the high variation between biological replicates and the use of two biological replicates, *THAH* (P = 0.09) and *THAS* (P-0.07). Notably, neither the non-core cluster gene, *THAA2,* nor the cluster flanking gene, *GFA2,* showed changes in transcript levels. This suggests that *JMJ16* could be acting as a regulator of the four core thalianol cluster genes by depositing the activation markers H3K4me3 and H3K9me3.

4.3.2.1.2 H3K4me3 demethylases show varied transcript changes

In our ReMap analysis, sections **4.2.1/4.2.2**, we identified H3K4me3 to be enriched in the thalianol gene cluster in roots. To further investigate the potential role of H3K4me3 in gene cluster regulation, we selected 9 *A. thaliana* lines with individual mutations in H3K4me3 demethylase encoding genes, ***Figure 76***. We tested each line for the expression of the *THAD* gene in comparison to the wild type. Three of the nine lines tested showed no significant difference in *THAD* gene expression, indicating that *LDL1*, *LDL2* and *JMJ16*, are not involved in thalianol gene cluster regulation. As we would predict for demethylase mutants, two of the mutant lines showed a trend for increased *THAD* expression, with *SDG4* showing significantly increased *THAD* transcript levels (P < 0.01). *JMJ14* showed a moderate, yet non-significant increase of *THAD* expression. *JMJ14* and *JMJ16* will be further investigated.

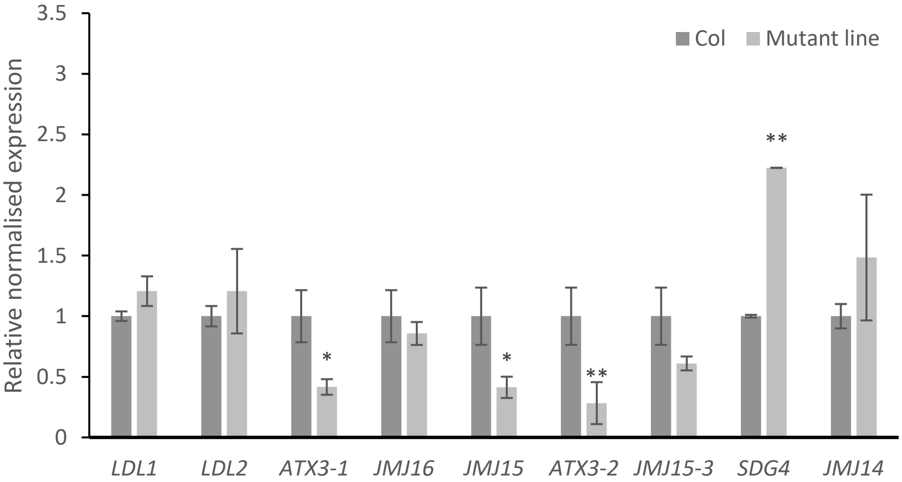


Figure 77: Transcript analysis of nine H3K4me3 demethylase T-DNA mutants on THAD expression. *Gene expression of the five thalianol cluster genes, and the flanking gene, GFA2, was measured by qPCR and compared to the wild-type (Col-0) transcript level and normalised to 1. An internal control was used, PP2AA3 (AT5G13320). Error bars indicate standard error of the mean for three biological replicates. Statistical significance (paired t-test): *P < 0.05, **P < 0.01.*

Interestingly and counterintuitively for a loss of H3K4me3 demethylation activity, we detected a decrease in *THAD* expression for four mutant lines, harbouring *ATX3* and *JMJ15* mutations. Three of the four lines showed a significant (P < 0.05) downregulation of *THAD*. These lines were chosen to be investigated further.

Of the lines that showed no changes to *THAD* transcript levels, we analysed the whole thalianol cluster expression levels for *JMJ*16, Salk_029530, ***Figure 77***. None of the four core cluster genes showed significant changes to transcript levels compared to the wild type. Interestingly, transcript levels for both flanking genes were significantly increased (P < 0.05). This may indicate regulatory role of *JMJ16* on the flanking genes rather than the thalianol cluster.

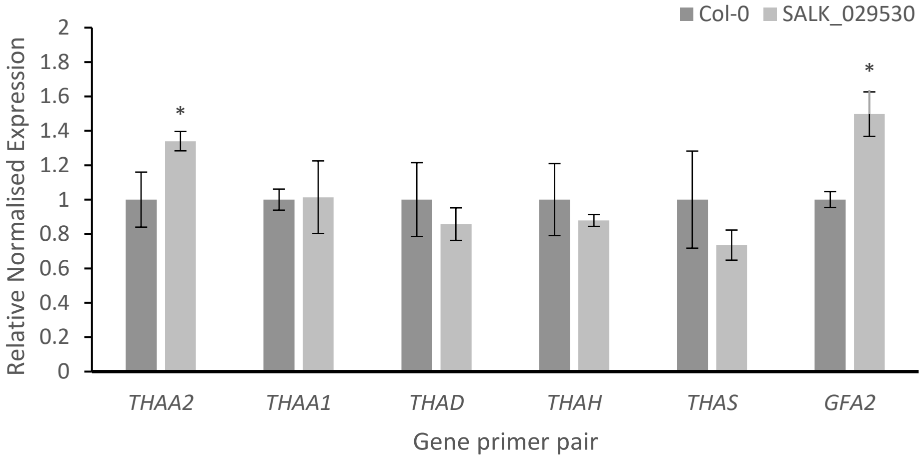


Figure 78: Transcript analysis of a H3K4me3 demethylase T-DNA mutant, JMJ16-3, across the thalianol gene cluster and flanking gene. *Gene expression of the five thalianol cluster genes, and the flanking gene, GFA2, was measured by qPCR and compared to the wild-type (Col-0) transcript level and normalised to 1. An internal control was used, PP2AA3 (AT5G13320). Error bars indicate standard error of the mean for three biological replicates. Statistical significance (paired t-test): *P < 0.05, **P < 0.01.*

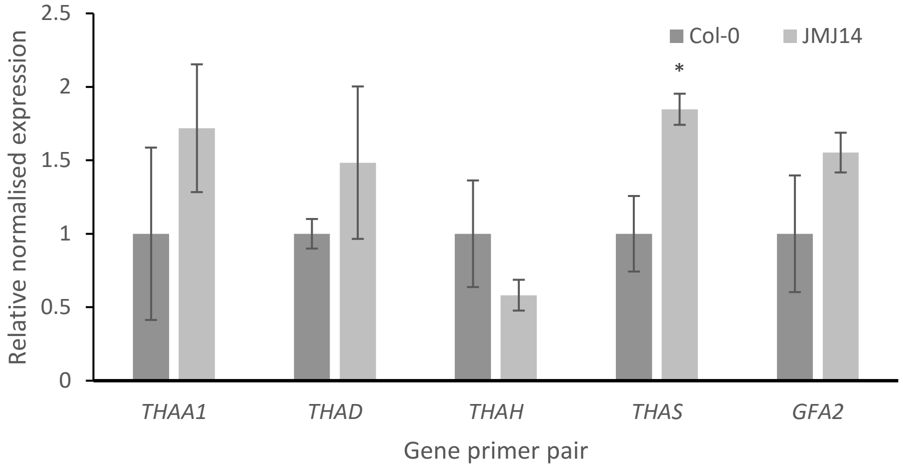


Figure 79: Transcript analysis of a H3K4me3 demethylase T-DNA mutant, JMJ14, across the thalianol gene cluster and flanking gene. *Gene expression of the five thalianol cluster genes, and the flanking gene, GFA2, was measured by qPCR and compared to the wild-type (Col-0) transcript level and normalised to 1. An internal control was used, PP2AA3 (AT5G13320). Error bars indicate standard error of the mean for three biological replicates. Statistical significance (paired t-test): *P < 0.05, **P < 0.01.*

The *JMJ14* H3K4me3 demethylase mutant[325], shows an unexpected increase in three of the thalianol cluster genes, *THAA1, THAD* and *THAS,* and the flanking gene, *GFA2,* where only *THAS* increase was significant due to the variation between biological replicates (P < 0.05), ***Figure 78***. We also detected a decrease in transcripts for *THAD,* which would be expected for this line, indicating that this mutation potentially misregulates the thalianol gene cluster, however the target is unlikely to be by H3K4me3.

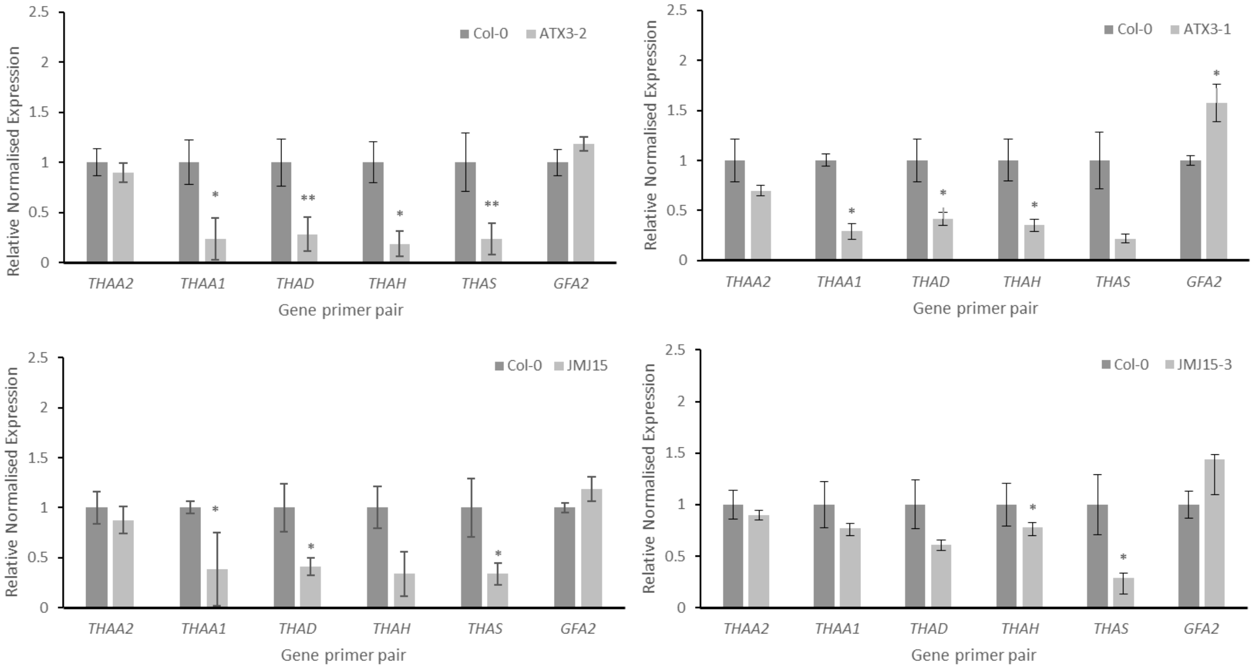


Figure 80: Transcript analysis of two H3K4me3 demethylases T-DNA mutant, JMJ14, across the thalianol gene cluster and flanking gene. *Gene expression of the five thalianol cluster genes, and the flanking gene, GFA2, was measured by qPCR and compared to the wild-type (Col-0) transcript level and normalised to 1. An internal control was used, PP2AA3 (AT5G13320). Error bars indicate standard error of the mean for three biological replicates. Statistical significance (paired t-test): *P < 0.05, **P < 0.01.*

Next, the four mutant lines, *ATX3* and *JMJ15*, that showed reduction in *THAD* (three of which were significant) transcripts were analysed further, ***Figure 79.*** Three of the lines showed a consistent trend for lowered transcript levels across the four core thalianol cluster genes. Most of these were significantly reduced (P < 0.05), aside from *ATX3-1 THAS* (P = 0.07) and *JMJ15 THAH* (P = 0.057). Interestingly, *JMJ15-3* only showed a significant reduction in *THAH* and *THAS* (P < 0.05) and no change in the other genes tested, suggesting only a partial mutation in the demethylase. The two flanking genes, *THAA2* and *GFA2,* showed no significant change in gene expression in all four mutant lines compared to the wild type, indicating a cluster specific regulatory effect.

4.3.2.2 Unexpected changes in gene expression

4.3.2.2.1 Mutation in activation mark transferase leads to up minor regulation

Our ReMap analysis indicated that H3K4me3 is enriched at the thalianol, marneral and arabidiol/baruol gene clusters. This conserved pattern indicates a role of H3K4me3 in cluster regulation. Therefore, we decided to assess transcript levels of an exemplar cluster gene *THAD* in two H3K4 methyl transferase mutants, *SDG2* (Salk_021008) and *ATX2* (Salk_117262). We predicted decreased expression levels of *THAD* as a loss of methyltransferase activity would lead to reduced H3K4me3 levels. However, in both the *ATX2* and *SDG2* mutant plants, we detected slightly elevated non-significant transcript levels for the *THAD* gene, ***Figure 80***. This indicates that both *SDG2* and *ATX2* could act in a non-canonical way leading to upregulation of *THAD.*

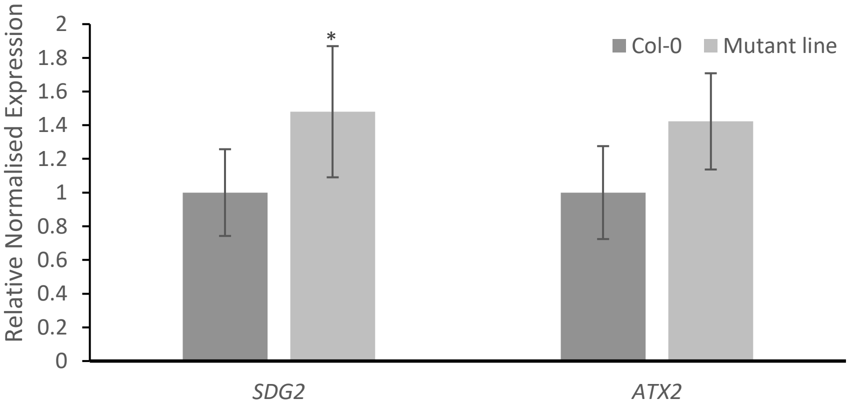


Figure 81: THAD transcript analysis in two H3K4 methyltransferase T-DNA mutants, SDG2 and ATX2. *Relative transcript levels of THAD (At5g47990) were measured by qPCR and compared to the wild-type (Col-0) transcript level and normalised to the internal controlPP2AA3 (AT5G13320). Error bars indicate standard error of the mean for three biological replicates. Statistical significance (paired t-test): *P < 0.05, **P < 0.01.*

4.3.2.2.2 *SGD25* mutant lead to *THAD* upregulation

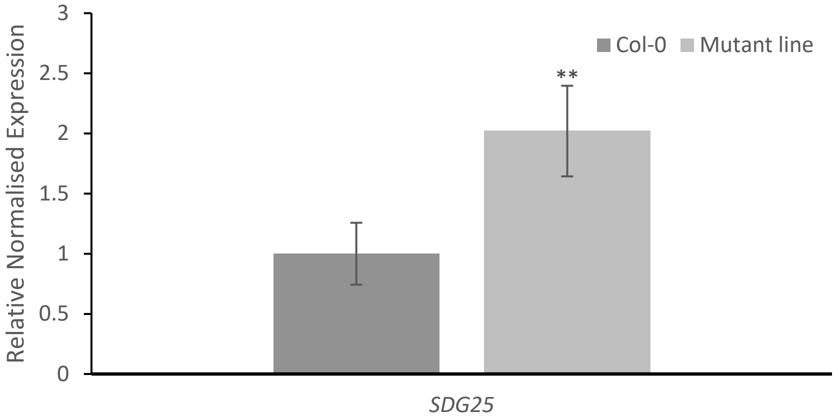


Figure 82: Transcript analysis of a H3K36me3 methyltransferase T-DNA mutants, SDG25, on THAD expression. *Gene expression of the thalianol cluster gene, THAD, was measured by qPCR and compared to the wild-type (Col-0) transcript level and normalised to 1. An internal control was used, PP2AA3 (AT5G13320). Error bars indicate standard error of the mean for three biological replicates. Statistical significance (paired t-test): *P < 0.05, **P < 0.01.*

Our ReMaP data identified low levels of H3K36me3 in the thalianol active domain in leaves. As there was no root data available, we decided to analyse a H3K36 methyltransferase mutant to assess the potential impact in cluster expression. The methyltransferase *SDG25* is associated with H3K36me2/3, a mark likely linked with active gene transcription.[29] To our surprise, *THAD* transcript levels significantly (P < 0.01) increased in *SDG25*, ***Figure 81.*** Thus, we hypothesis a non-canonical role for either *SDG25* or H3K36me3 at this location.

4.3.2.2.3 *JMJ*32 mutation leads to *THAD* upregulation

H3K27me3 has already been implicated as a repressor of the thalianol cluster. To identify potential key regulatory enzymes of H3K27me3 modifications at the cluster we analysed the *JMJ32* demethylase mutant line.[140] The *JMJ32* gene encodes for a demethylase implicated in the removal for methyl residues of H3K27me3. Similarly, to our results obtained for the *SDG25* mutant line, we detected a counterintuitive expression pattern for *THAD* in the *JMJ32* mutant line. Loss of demethylation activity should result in enhanced H3K27me3 methylation and more pronounced gene silencing. However, our transcript analyses show a mild yet significant increase in *THAD* transcript levels in *JMJ32* compared to the wild type, ***Figure 82***. This could be due to an increase in H3K27me2, as the methylation if being removed, the function of which is not fully understood.[326]

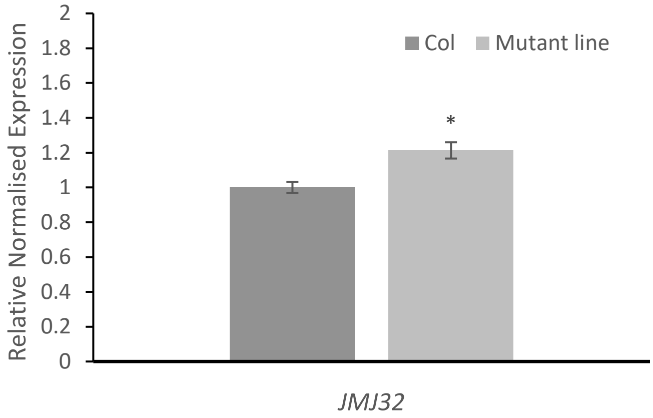


Figure 83: Transcript analysis of a H3K27me3 demethylase T-DNA mutants, JMJ32, on THAD expression. *Gene expression of the thalianol cluster gene, THAD, was measured by qPCR and compared to the wild-type (Col-0) transcript level and normalised to 1. An internal control was used, PP2AA3 (AT5G13320). Error bars indicate standard error of the mean for three biological replicates. Statistical significance (paired t-test): *P < 0.05, **P < 0.01.*

Next, we will analyse the histone modification mark changes in one of the histone modification mutants, using ChIP as an exemplar of how downstream experiments could be carried out on all mutants of interest.

4.3.2.3 Chromatin immunoprecipitation sonication optimisation

To experimentally validate our bioinformatic analyses of cluster specific chromatin marks combined with our findings for changes of expression in a mutant line across the thalianol gene cluster, we set out to perform gene specific chromatin immunoprecipitation (ChIP) experiments. To realise this, we firstly had to establish the ChIP protocol in our lab. A crucial step in this method establishment was the optimisation of chromatin sonication conditions. To analyse chromatin mark enrichment by ChIP, it is important to shear chromatin into segments of 250 – 500 bp. Chromatin shearing is performed by chromatin sonication. To optimise the sonication protocol, we carried out multiple test sonications on Col-0 *A. thaliana* separated roots and leaves. In ***Table 27*** all tested sonication conditions are shown. By running the sheared DNA obtained from the test sonication runs, on an agarose gel we were the able to identify the optimal conditions for use in chromatin immunoprecipitation experiments. The first set of sonication test conditions indicated that the best condition for shearing root derived DNA was five cycles of 10 seconds on and 90 seconds off or eight cycles of 5 seconds on and 90 seconds off. Unfortunately, DNA concentrations were low and banding pattern not fully visible. For leaf derived DNA, all conditions showed insufficient shearing with significant levels of DNA banding from 500 to 1000 bp. Due to the inconclusive experimental results, we repeated the sonication tests with additional conditions. The second round of sonication experiments identified five cycles of 10 seconds on and 90 seconds off as the optimal shearing DNA condition for root samples and being eight cycles of 15 seconds on and 90 seconds off for leave samples. These conditions resulted in strong signals at the target size range between 250 and 500 bp and no enrichment of larger DNA segments, ***Figure 83***. These conditions were applied for all subsequent ChIP experiments.

Table 27: Sonication conditions trialled for shearing chromatin.

Sample	On (seconds)	Off (seconds)	Cycles
Root	5	90	5
	10		5
	15		5
	5		2
	5		5
	5		8
leaf	10	90	2
	10		8
	10		5
	5		5
		None	
	3	90	3
	3		6
	5		5
	10		2
	3		3

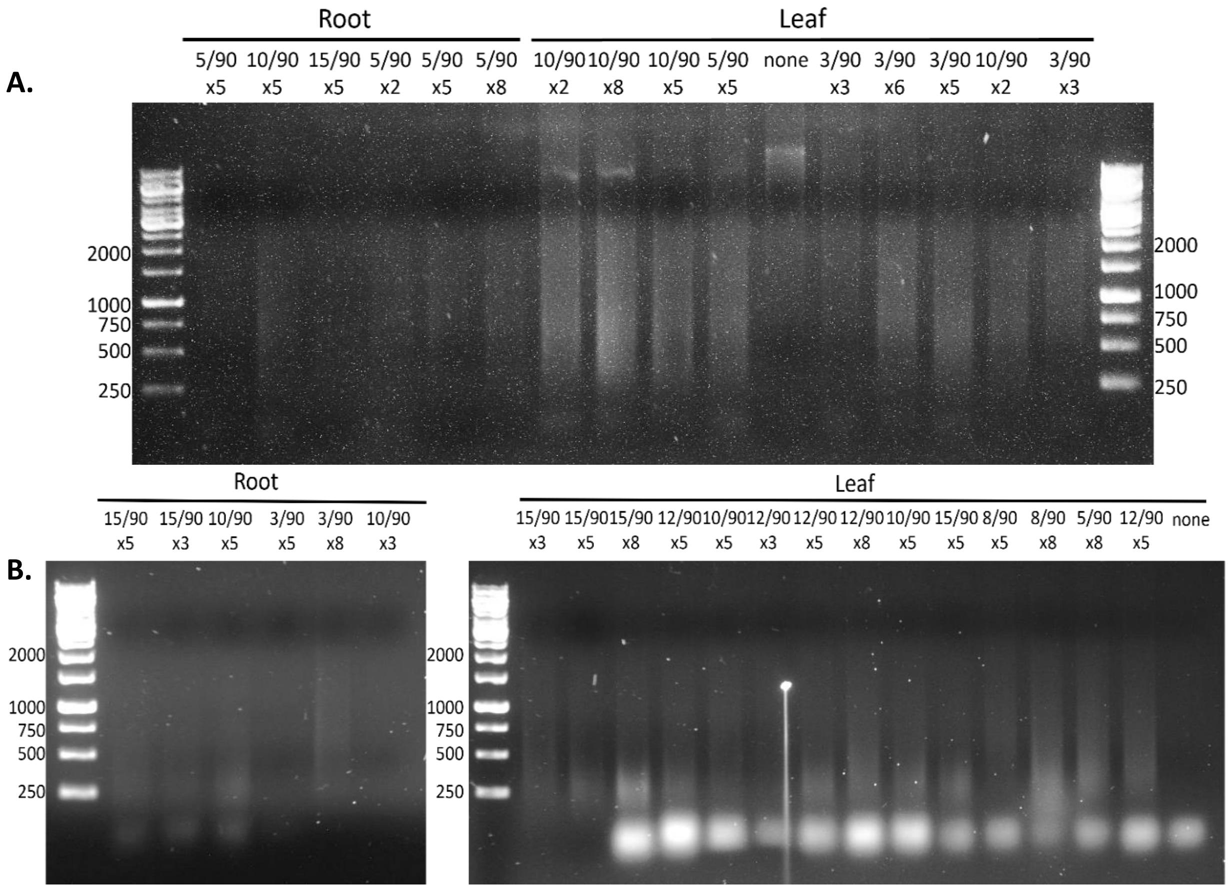


Figure 84: Agarose gel of sonicated chromatin under different shearing conditions and in root and leaf material. *The first set of sonication conditions (A) did not show strong banding for sheared DNA, however in the second set of conditions (B) we could identify banding around 250 and 500 bp. Therefore, identifying conditions to be used in subsequent experiments. Size of bands confirmed using a 1kb ladder, sizes of bands indicated to left of gels.*

4.3.2.4 H3K4me3 is more prevalent across the thalianol gene cluster in Col-0

Following ChIP we carried out qPCR analysis on the DNA fragments remaining in the sample to identify those enriched with H3K4me3, compared to the control H3, ***Figure 85***. We included three control genes in our analyses: the ubiquitously expressed *Acn* and *PP2AA3* genes, GFA2 cluster neighbouring gene and the *At5g47970* cluster intervening gene. H3K4me3 is known to be enriched at transcriptional start sites, so primers were designed to occupy the start of the gene of interest, however some primers were also located towards the end of the gene for comparison.[327] Intergenic regions were also included due to data obtained in section **3.2.3**. Two biological replicated were used for leaves and two for roots. Due to the low number of replicates, we decided not to perform statistical significance analyses.

First, we compared H3K4me3 levels across genes in the wild-type, Col-0, between, in roots (where the thalianol gene cluster is active) and leaves (where the cluster is inactive), ***Figure 85A***. There was a dramatic increase in the percentage change of H3K4me3 enrichment in the control genes, *Acn*, *PP2AA3* start site and *GFA2* in leaves and lesser so in roots, which is expected as they are constitutively expressed, this could indicate more efficient chromatin extraction in roots than leaves. The thalianol cluster had more transcripts for roots than leaves at the gene start sites, and an equal level in intergenic regions.

Next, we compared the transcripts from the *JMJ14* H3K4me3 demethylase mutant leaves and roots ***Figure 85B***. Surprisingly, there was a very dramatic increase in the percentage change of H3K4me3 enrichment in the leaves of the control genes than in the roots, but low H3K4me3 across the whole thalianol cluster, aside from the *At5G47970* gene.

To see the change between Col-0 and *JMJ14* mutant, the percentage change between leaves and then for roots were compared. In leaves we can identify that those changes in the control genes are drastic in the *JMJ14* mutant compared to the Col-0. While in the root the percentage change appears similar but an increase in the *JMJ14* mutant.

To assess this change between roots and leaves, a normalised (to enrichment of H3K4me3 at *PP2AA3* start site) fold enrichment for Col-0 and *JMJ14* mutant determined, ***Figure 85E***. This shows the extent of the increase in H3K4me3 enrichment across the thalianol cluster genes in the *JMJ14* mutant in the roots, as expected by a H3K4me3 demethylase. The change in marks between leaf and root in both Col-0 and *JMJ14* are very similar in the control genes, surprisingly the fold change is high at the end of *PP2AA3*. In our transcript analysis for *JMJ14*, section **4.4.2.3**, we noted an increase in expression for *THAA1, THAD, and THAS* which can be accounted for by the potential increase of H3K4me3 at these sites in the mutant. However, interestingly, there is an increase at *THAH,* whereas we noted a decrease in expression for this gene.

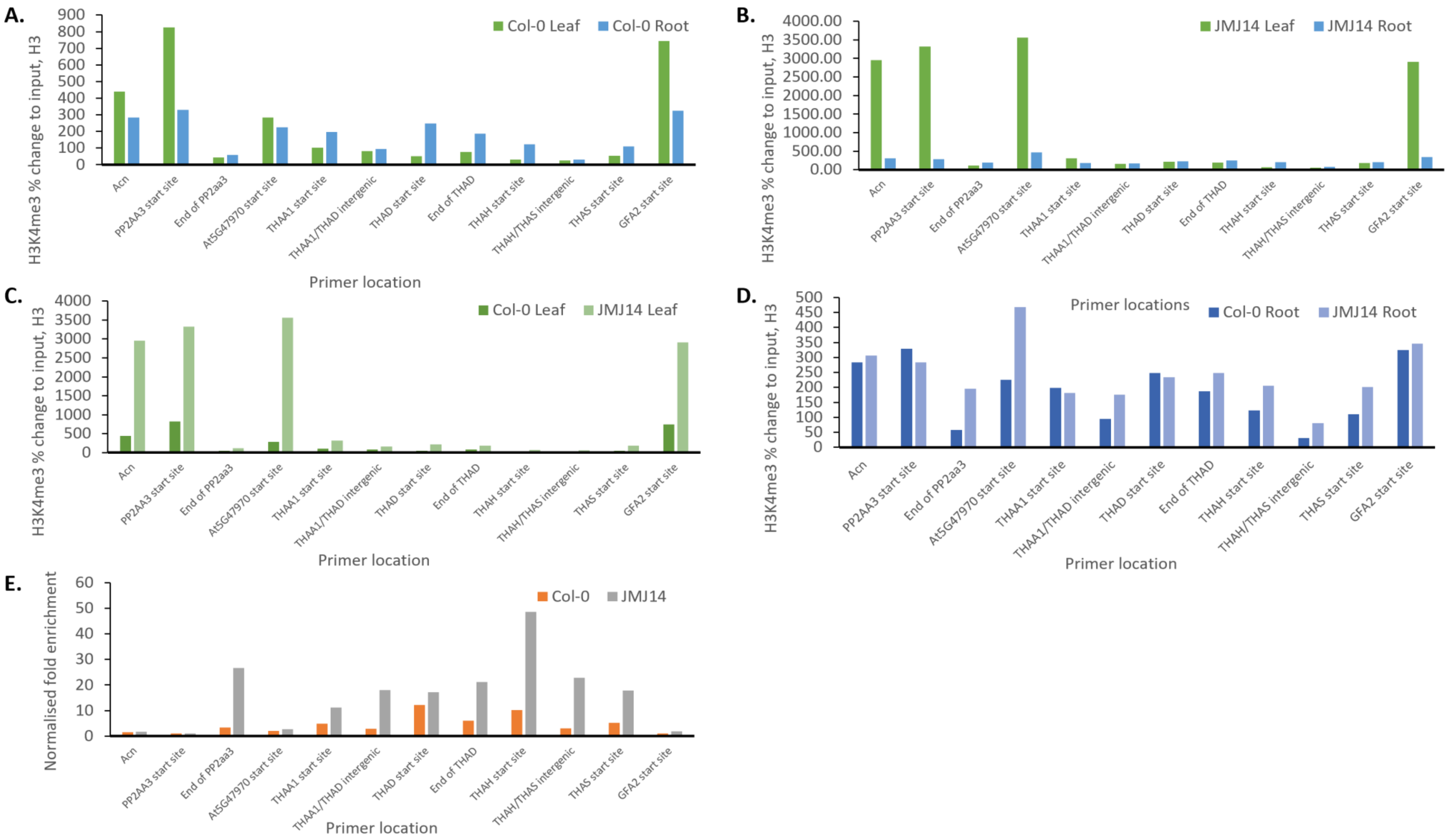


Figure 85: ChIP transcript analysis. *A. is the H3K4me3 percentage change compared to the input, H3 in Col-0 leaf and root, B is JMJ14 mutant leaf and root, C is the Col-0 leaf compared to the JMJ14 leaf and D the Col-0 root and JMJ14 mutant root. E. shows the normalised fold enrichment change between root and leaf in Col-0 and JMJ14 mutant. Primers were used covering the thalianol gene cluster and the control regions, Acn, PP2AA3 and GFA2.*

4.4 Discussion

4.4.1 Histone modification marks

4.4.1.1 H2A variants

In *A. thaliana* there are four main types of H2A variants, H2A, H2A.Z, H2A.X, and H2A.W each have distinct functions in regulating the chromatin structure, here we focus on H2A, H2A.Z and H2A.X due to data availability. H2A.X is associated with heterochromatin whereas H2A.Z and H2A are largely linked with euchromatin.[328] In *Arabidopsis* leaves at the three biosynthetic clusters analysed there is a reduced enrichment of H2A and H2A.X but an increased enrichment of H2A.Z. The thalianol, marneral and arabidiol clusters have been found to be enriched in the activation mark, H2A.Z. When the mark is depleted or defective at these regions then the cluster genes are strongly downregulated, in a cluster-specific pattern, not including the flanking genes.[161] Although we do find a slight increased enrichment of the H2A.Z mark in leaf material in both the active thalianol cluster domain, which is atypical due to the clusters not being active in leaves, we also find a reduction in both H2A and H2A.X. *A. thaliana* nucleosomes are typically homotypic, thus only contain a single H2A variant, thus if H2A.X and H2A are depleted, H2A.Z will be deposited at his region.[42] H2A.Z has also been implicated to have repressive properties depending on its localisation within the genes in response to environmental stimuli.[329] H2A.X has mainly been found to play a role in the DNA damage response, and there is lack of evidence in its function in the regulation of gene transcription, [330] thus it is unsurprising to see a reduction of this mark across the clusters. To better understand H2A enrichment at metabolic gene clusters, it would be important to obtain enrichment data for roots. However, as H2A.Z levels remain high in roots, we predict that H2A levels remain low at expressed clusters.

4.4.1.2 H3K4me3

Previously, studies investigating histone modification marks across the biosynthetic gene clusters deemed that the H3K4me3 mark did not have a strong enrichment across the activated genes.[163] However, our data analysis shows a clear increase in H3K4me3 in roots where the cluster is activated, and a reduction in leaves where the cluster is silenced, this pattern also extends to the marneral and arabidiol gene clusters. Studies have shown that H3K4me3 is increased at genes which are associated with the response to hormones, such as

ABA, JA and ethylene.[327] The thalianol gene cluster has been shown to be enhanced in *A. thaliana* roots in response to JA treatment.[331] In addition, H3K4me3 has been found to be enriched in metabolic gene clusters in rice genomes and in *Fusarium* and *Aspergillus.*[332] In section **3.3.1.4**, we identified how disruption of a region near a proposed super-enhancer region lead to misregulation of the thalianol gene cluster[224], in mammalian cells, super-enhancers have been linked with a strong association to H3K4me3 for activation.[333] We suggest that this mark could be a key regulator of metabolic gene clusters, thus should be further investigated.

4.4.1.3 H3K27me3 and H3K18ac

The repressive marker H3K27me3, has been implicated in the silencing of metabolic gene clusters.[167] Accordingly, we have identified strong H3K27me3 enrichment at all clusters in datasets associated with the silenced pathway. It was found that these regions of the genomes were characterised as belonging to B compartments in wild-type plants[334]. H3K27me3 is already a known mark across the three metabolic gene cluster, thalianol, marneral and arabidiol, as well as in clusters in rice and maize.[163] Studies in wild-type *A. thaliana* and a H3K27me3-depleted mutant, identified that an intrachromosomal interaction between the marneral and thalianol gene cluster was strongly elevated in the wildtype compared to the mutant. Indicating that H3K27me3 plays a role in connecting clusters in 3D space.[163] H3K27me3 is typically associated with facultative heterochromatin, and in repressed TADs in *Drosophilia*.[335] In the fungus, *Neurospora crassa* loss of H3K27me2/3 altered the genome architecture, in particularly at the sub telomeric regions which seems to play a role in maintaining overall genome conformation.[336] The loss of H3K27me3 marks lead to re-localisation of sub telomeric regions towards the nuclear interior and up-regulation of its target genes.[336]

4.4.1.4 H3K18ac

In our ReMap analysis we identified an enrichment of the activation mark, H3K18ac in seedlings across the three metabolic gene cluster, thalianol, marneral and arabidiol, also an increase in the thalianol active domain compared to the inactive domain.[337] H3K18ac also acts as a prerequisite for DNA demethylation[338,339], by opening up the chromatin the demethylase can access the DNA, and is enriched at transcriptional start sites for transcription activation. In *A. thaliana* both H3K27me3 and H3K18ac have been found to co-localise at the same region to form novel bivalent chromatin.[337] Studies have shown that both modification marks are required to determine the timing of gene expression and metabolite accumulation in the early stages of a response to stress stimuli. The clustered metabolic pathway, camalexin[340], is rapidly produced in response to the pathogen *Pseudomonas syringae*. This pathway is enriched with both H3K27me3 and H3K18ac.[341] The function of having bivalent chromatin allows for more rapid, tighter control on the activation and repression of the region. When one of the two marks were repressed, the accumulation of camalexin changed accordingly, confirming that histone modification marks can control the timing of expression of genes *Arabidopsis thaliana*. Interestingly, when the specialised metabolic genes were removed from the analysis there was a decrease in the correlation between H3K27me3 and H3K18ac indicating the genes play a role in the two marks.[341] This is a clear example of a metabolic pathway that relies on a tightly orchestrated activation and repression of genes for the formation of camalexin. Our data shows a clear enrichment of the H3K18ac mark at all three metabolic gene clusters suggesting there could be a bivalent chromatin system occurring with H3K27me3 at these sites. Future research using ChIP-qPCR should be carried out to determine whether there is a bivalent chromatin with these two histone modification marks at these locations. A quantitative PCR could also be carried out on a H3K18ac and H3K27me3 deficient single and double mutant lines to analyse if there is a change in expression levels of the genes within the clusters.

4.4.1.5 H3K36me3

H3K36me3 is associated with transcriptional activation[342], our H3K36me3 enrichment data from leaves showed an overall depletion of this mark across all three biosynthetic gene clusters. For future experimentation it would be beneficial to obtain data for the H3K36me3 enrichment of these biosynthetic gene clusters in root material. We predict that the mark would show an enrichment. Previous studies have shown that in plant immunity H3K36me3 deposition by *SDG8* is crucial for the induction of transcription in JA/ET-inducible genes in response to infection by a fungi.[337] As the thalianol cluster is responsive to hormones, such as jasmonic acid, we hypothesise that the H3K36me3 activation marker could be a key histone marker for regulating the activation of the biosynthetic clusters.[158] In the future, we suggest to perform H3K36me3 ChIP experiments in both expressing and non-expressing tissues. These experiments should be accompanied by hormone treatment of plants as well as cluster expression analyses in *SDG8* and related H3K36me3 mutant plants.

Another study suggested that H3K36me3 could be an opposing mark to H3K27me3, as an anticorrelation between the two marks was observed during cold exposure.[322] It was shown that the two histone modification marks cannot colocalise on the same histone tail and that the loss of H3K36me3 leads to H3K27me3 spreading across the locus. *SDG8* mutant data further shows that levels of H3K27me3 increased genome-wide, thus, supporting the antagonistic properties of H3K36 methylation.[322] Indeed, the data presented here shows opposing enrichment levels of H3K27me3 and H3K36me3 at metabolic gene clusters in tissues with low gene expression levels. Although mutually exclusive labelling of histones with H3K36me3 and H3K27me3 has only been associated with the response to changing temperatures, it would be valuable to investigate the H3K36me3/H3K27me3 dynamics at metabolic gene clusters.

4.4.1.6 H3

The histone protein H3 has two variants which differ by a few amino acid residues. All variants show distinct expression patterns.[343–345] Active expression throughout the cell cycle is associated with H3.3[346] whereas H3.1 is linked to DNA replication. It is suggested that H3.3 is required for the deposition or maintenance of DNA methylation over gene bodies. This has been corroborated by findings that mutant lines with depleted H3.3 levels exhibit an overall reduction in DNA methylation over gene bodies. This depletion was not associated with the loss of expression of DNA demethylases or methyltransferases, but overexpression of H3.1, H2A, H2B, and H4 resulting in a lower density of nucleosomes in the chromatin. Specifically, H1 is implicated in the reduction of chromatin density and preventing the access of DNA methyltransferases. H3K4me3 appeared to be unaffected by the change in H3.3 levels, due to H3K4me3 deposition at the 5' end of genes, whereas H3K36me3 levels were reduced at promoter regions.[347] In our data we present reduced H3.3 levels in the leaves in all three biosynthetic gene clusters. In light of the previous reports, this may indicate reduced accessibility for DNA methyltransferases at gene clusters. It would be interesting to analyse cluster expression and DNA methylation levels in H3.3 knockout mutants to establish if reduction in the marker causes an increase in activation in the leaf tissue or reduction in roots. This could be combined with H3.3 ChIP analyses to investigate relative H3.3 enrichment in both transcriptionally active and inactive clusters.

4.4.1.7 H3K9me2

H3K9me2 is a well-known repressive mark which is involved in silencing transposable elements and repetitive DNA in plants.[324] Both the marneral and thalianol cluster have been found to be significantly enriched in transposable elements, which could've contributed to the gene cluster assembly.[151] In our data we identify a basal level of H3K9me2 throughout the thalianol gene cluster, but less enrichment in the larger inactive domain, similar results were found for marneral but an increase in enrichment for arabidiol. Arabidiol is a larger gene cluster, thus will contain more transposable elements, leading to an increase of H3K9me2 enrichment. In section **3.2.4.2** we also noted that the arabidiol cluster also undergoes more genome rearrangements between accessions, indicative of the TEs present. In plants H3K9me1 and H3K9me2 is associated with heterochromatin and H3K9me3 associated with euchromatin.[348] In response to abscisic acid H3K9me2 is reduced at abiotic stress-response genes indicating that this mark could be associated with regulating stress response.[349] The thalianol gene cluster is upregulated in response to jasmonic acid, thus we would expect a reduction of H3K9me3 in the roots than in the leaves. We do not have root and leaf data for this mark; however, it would be of interest in future research.

4.4.1.8 Next steps

The data used for our histone modification analysis allowed for a broad overview of histone modification marks across the metabolic gene clusters. However, this experimentation was carried out by different laboratories that will have applied different procedures. This will inevitably lead to discrepancies in data quality and depth. This also means we only obtained one data set for each mark and condition, so we couldn't account for any biological or technical replicates. In addition to this, the large region that was covered and only having one replicate means that the hypergeometric test we carried out could only indicate if the data was or wasn't significantly different, thus there might be some marks that were on the boundary of being significant that could've been of interest.

Ideally, experiments for all marks would have to be repeated in the same lab under the same conditions for both expressing and non-expressing tissues. As these were genome-wide histone modifications there is not any bias for the gene clusters in the ReMap data. By using 100 random regions in which we normalised the data to, this took into account the experimental differences that might have been carried out.

We selected four histone modifications, H3K9, H3K4, H3K27 and H3K36, for further studies. We selected these modifications either because they have been previously implicated in cluster regulation (H3K27me3) or because their role on cluster regulation has not been investigated before.

4.4.2 Histone modification mutants

4.4.2.1 Importance of *JMJ*14

*JMJ*14 is a H3K4me1/2/3 demethylase[275], which modulates the immune response in *A. thaliana* by regulating defence gene expression. The transcript analysis of this mutant showed misregulation of the thalianol gene cluster, and ChIP, showed an increase of H3K4me3 marks across this region. We identified a trend for increased levels of gene expression in the *JMJ14* mutant line across three of the four core thalianol cluster genes, aside from *At5g48000* that non-significantly showed a decrease level of transcription presenting some limitations of our approach due to variation between biological samples. However, this indicates that the relationship between the thalianol cluster and the activation by H3K4me3 is more complex than originally thought. Interestingly, the largest fold change in the *JMJ14* mutant between leaf and root was at the *THAH* gene and the intergenic region between *THAH* and *THAS*, where previously a super-enhancer was identified.[350] Super enhancers have been linked to the enrichment of H3K4me3[333], thus *JMJ14* could be a potential regulator of this region. However, as the change across the whole cluster was not dramatic, we hypothesis that other regulators could also be vital for this region. This change could be due to *JMJ14* not only demethylating H3K4me3, abut also di- and mono-methylation marks.

Plants defective in *JMJ14* have also been shown to have reduced levels of H3K9K14ac levels.[351] This also leads to reduced polymerase II occupancy, reduced transgene transcription and increased methylation at the promotor region. This could explain why *THAH* showed decreased levels of gene transcription. The removal of *JMJ14* has also been shown to allow other H3K4 demethylases to reduce transcription levels, hence why there was an increased expression of the thalianol cluster genes, but non-significantly. [351]

This mutant was used as an exemplar of how downstream analyses of our histone modification analyses could be carried out. Future experiments should include the larger downstream inactive region of the thalianol cluster alongside the other gene clusters to assess if *JMJ14* is a regulator across clusters. To better understand the role of histone methylation on cluster expression and the interplay between different demethylases at the thalianol cluster, further experimentation including a more comprehensive screening of demethylase mutants and higher-order mutants as well as expression analyses under different conditions will be necessary. To identify a more general role for specific demethylases in cluster regulation will require expression analyses for additional clusters in *A. thaliana* and other plant species.

4.4.2.2 *JMJ*15 demethylase

*JMJ*C proteins are described as histone demethylases. Here we discuss the results we obtained for *JMJ15* which has been shown to demethylate H3K4me2/3 at target genes[281]. Expression data suggests that *JMJ15* is ubiquitously expressed in all organs and during the life cycle of *A. thaliana*[279]*.* In gain of function mutants for *JMJ15* stress-responsive genes were mis-expressed and salt tolerance was enhanced. In loss-of-function mutants stress resistance was reduced[281]. In our experiment we analysed a loss-of-function mutant to investigate the effect of loss of *JMJ15* on our region of interest. Our expression analyses showed reduced transcript levels across the core genes of the thalianol gene cluster. As for our results for *JMJ16*, this expression pattern is counterintuitive as loss of *JMJ15* should result in higher levels of the activating H3K4me2/3 marks and thus increased expression. As we have discussed before, indirect regulatory effects and multiple target histone amino acid sites may explain this pattern. It is intriguing that *JMJ15* activity has been strongly associated with the regulation of stress-related genes. The thalianol cluster is known to be important for the maintenance of the *A. thaliana* root microbiome. As such, it will have a crucial function in the response to root-specific environmental stress conditions. We may therefore speculate that a regulatory role for *JMJ15* in thalianol cluster expression may connect cluster expression to other stress response pathways.

4.4.2.3 JMJ16

The methylation of histone marks can be reversed by an oxidative demethylation reaction catalysed by demethylase enzymes. Lysine demethylase 1 (KDM1) was the first protein discovered to remove the methyl group from mono- and demethylated lysine 4 of histone H3[259]. Further studies identified a large family of Jumonji C (*JMJ*C) domain-containing proteins as being histone demethylases[352]. 30 *JMJ*C domain-containing proteins were identified in humans, 10 in *Oryza sativa* and 21 in *A. thaliana*[271]. In *A. thaliana* six *JMJ*C domain-containing proteins were homologues to human KDM1 in sequence similarity and domain structure, indicating they are likely to be active histone demethylases: *JMJ14, JMJ15, JMJ16, JMJ17, JMJ18*, and *JMJ19* [271]. Previously, *JMJ14, JMJ15 and JMJ18* have been characterised as regulating the transition to flowering by H3K4 demethylation. *JMJ16* has been shown to be a specific H3K4 demethylase to negatively regulate leaf senescence through demethylation. Unfortunately, prior studies on *JMJ16* provided leaf-specific expression data only. To find out whether loss of *JMJ16* would result in changes to thalianol cluster expression, we analysed cluster expression in the *JMJ16-1* mutant in *A. thaliana* roots. Interestingly, we found a significant reduction in transcripts for the four core thalianol cluster genes in the *JMJ16-1* mutant compared to wild type transcript levels. This was a surprising result as a loss of demethylase activity should result in increased H3K4me3 levels and thus elevated expression levels. However, as the prior genome-wide expression studies showed before, loss of *JMJ*16 leads to similar numbers of genes with increased and decreased expression levels. While counterintuitive, decreased expression levels are likely to be an indirect consequence of an altered histone modification pattern. For example, increased expression of a subset of genes due to elevated H3K4me3 levels may result in the downregulation of genes negatively connected to these genes.

Furthermore, *JMJ*16 demethylation activity may not be linked to H3K4me3 only. Indeed, a recent study showed that upon interaction with the *MMD1* (male meiocyte death 1) protein

JMJ16 substrate specificity is broadened to H3K9me3[285]. *MMD1*, typically expressed in developing male meiocytes, is required for the regulation of microtubules organisation, cell cycle transitions and meiotic genes[353]. In *mmd1* and *JMJ16* single mutants, H3K9me3 levels were found to be significantly increased compared to the wild type.[354] As a repressive mark, increased H3K9me3 levels are likely to result in decreased expression levels of target genes. While *MMD1* acts specifically during meiosis and is not expressed in roots, we speculate that root specific proteins similar to *MMD1* interact with *JMJ16* in and also broaden its target specificity in roots. In such a scenario, loss of *JMJ16* may result in elevated levels of the repressive H3K9me3 mark and subsequently in lowered thalianol gene cluster expression.

To further investigate the impact of the loss of *JMJ16* activity on cluster expression, we suggest measuring H3K4me3 and H3K9me3 levels at the thalianol cluster in wild type and mutant lines. Deviating enrichment pattern may inform us about the target methylation mark of *JMJ16* at the thalianol cluster. Additional experiments may include the identification of *JMJ16* protein interaction partners in roots and leaves and their potential involvement in *JMJ16* target specificity. To monitor direct binding of *JMJ16* to the cluster a *JMJ16*-specific ChIP experiments will be informative.

4.4.2.4 ATX

Lysine on the tail of histone H3 can be modified by the addition of a methyl group with methyl transferase activity.[272] A large protein family of methyltransferases contain a SET (suppressor of variegation) domain. In *A. thaliana* 49 of these proteins have been identified and are known as the SET domain group (SDG).[296] Only one of the five classes of SDG proteins is suggested catalyse the methylation of histone H3 lysine 4. This class III subfamily of proteins contains five proteins named *ATX1* through 5 in *A. thaliana*.[305] *ATX* 1-5 are homologues of Trithorax proteins found in eukaryotes. The different biochemical properties and gene specificity in *ATX1* and 2 is due to their slight differences in sequence identity, as *ATX1* specifically has H3K4me3 activity while *ATX2* works on H3K4me2.[306]

So far, functional studies on class III SGD proteins have been largely restricted to aerial parts of the plants. This may been explained by the relatively lower expression of SDG class III genes in roots, which however, does not rule out their functional importance in roots[305].

Here, we show that all four thalianol core cluster genes show reduced expression in the *ATX3* mutant line, suggesting a role of *ATX3* methyltransferase in root-specific expression of the metabolic gene cluster. In the past, it had been shown that *ATX3* functions redundantly with *ATX4* and *ATX5*. Future experiments should therefore be carried out with the *ATX3/4/5* triple mutant as well as the *ATX4* and *ATX5* single mutant to confirm if all three are involved in the regulation of the thalianol cluster. Interestingly, preliminary analyses of data provided by Chen et al[305], over half of H3K4me2 sites within the thalianol cluster show a reduction in the *ATX3/ATX4/ATX5* triple mutant compared to the wild type in leaves. No reduction of H3K4me3 levels was detected. Like H3K4me3, H3K4me2 has been associated with actively transcribed gene. The pattern observed in Chen et al. may suggest that loss of *ATX3* activity results in lower H3K4me2 levels at the thalianol cluster and thus in reduced gene transcription. To support this speculation, a detailed analysis of H3K4me2 levels at the thalianol cluster should be carried out including *ATX3* mutant and wild type lines. In addition, these mutations should be investigated for their impact on the other *A. thaliana* metabolic gene clusters to identify if *ATX3* is a conserved regulator of metabolic gene cluster activity.

4.4.2.5 LDL1/2

Lysine-specific demethylase 1-like (LDL) proteins *LDL1* and *LDL2* belong to a group of four proteins with demethylation activity against histone H3 lysines. *LDL1* is well described as regulating genes within roots such as repressing the expression of *FLC*. In contrast, *LDL2* has been shown to be involved in seed dormancy[267]. The LDL1/*LDL2* double mutant has been implicated in the immune respone[267]. Both, *LDL1* and *LDL2* mutants exhibit higher levels of H3K4me1 and H3K4me2 than the wild type but no difference in H3K4me3 levels, indicating that *LDL1* and *LDL2* regulate H3K4 mono and dimethylation[267,355]. Our thalianol gene cluster expression analysis did not detect significant differences in *ldl1* and *ldl2* mutant lines compared to our wild type controls. This may indicate that these histone modifiers are not involved in cluster regulation. To further exclude a role for *LDL1* and *LDL2* in cluster expression, the double mutant should be tested, and cluster expression monitored under various test conditions.

5 Final conclusions

Using a combination of different technologies, such as readily available T-DNA insertions, online histone modification data and customisable CRISPR/Cas9 we were able to identify key regulatory elements that aid in the regulation of the thalianol gene cluster.

T-DNA insertional mutants cause disruptions to genome architecture. They may result in genes no longer to function and being transcribed and may disrupt regulatory sites important for 3D genome organisation. Our first T-DNA analysis aimed to identify potential regulatory regions of the thalianol gene cluster by selecting a range of insertion lines in the intergenic and genic regions covering the active and inactive thalianol domains. Interestingly, insertional mutants that were only a few hundred base pairs apart from one another displayed differing expression patterns of the thalianol gene cluster. We observed a region downstream of the *THAA2* gene which led to downregulation of the whole thalianol cluster which could potentially interrupt intrachromosomal interactions, leading to the misregulation of the gene cluster. The *THAA2* gene is often debated for its inclusion in the thalianol gene cluster as it is not directly adjacent, interrupted by two intervening genes. We show that disruption to this gene, causing its downregulation, does not impact the remaining thalianol cluster genes. Interestingly, in seven of eight *A. thaliana* accessions analysed we find that there is an inversion leading to the *THAA2* gene now being directly adjacent to the core thalianol cluster genes. For the four accessions we analysed for cluster expression levels, we detected differing patterns of expression. However, in the three accessions with reduced *THAA2* expression, a trend for upregulation of the four core thalianol genes can be observed. Interestingly, the opposite occurs in the accession when *THAA2* is upregulated, the remaining genes are downregulated. This may indicate that the orientation and genomic positioning of the genes is more important than the disruption of the gene itself to co-regulate the gene cluster. For future work, it would be interesting to carry out ChIP and Capture Hi-C experimentation on the accessions alongside thalianol production analysis to analyse the 3D conformation of the accessions and how the histone modification marks differ to aid the change in gene regulation. Another interesting experiment would be to analyse the change in the root microbiome in response to the change in gene expression. As we already know that overexpression of thalianol leads to dwarfing in the plants[106], it would be imperative to look at the phenotypic response of the plants in regards to the various mutants and changes in gene expression.

We identified that disruption of *THAH* led to upregulation of *THAD,* and a potential regulatory region in an intron of a gene, a phenomenon known as intron-mediated enhancement. Disruption of this region may have led to interruption of a gene loop, thus causing misregulation of the cluster gene. Similarly, for the T-DNA lines in *THAS,* the disruption of the eighth exon*,* led to misregulation whereas disruption of the ninth and twelfth exon didn't, indicating a regulatory element in eighth exon of the gene.

The large interacting domain which aids the silencing of the gene cluster could be regulated by a region nearly 100kb downstream of the cluster, as a T-DNA insertion disrupting this region led to misregulation of the thalianol gene cluster.

It would be interesting to introduce deletions of these regions using gene editing technology and carry out ChIP and Capture Hi-C to analyse the histone modification changes and 3D organisational changes. As we are mainly interested in the interactions within the cluster capture Hi-C would be the most appropriate method for analysing how this interactions may change with genome editing of specific regions.[90] It would also be necessary to expand this experimentation to more T-DNA lines in the thalianol gene cluster, alongside the marneral and arabidiol gene cluster.

We attempted to insert the *THAS* regulatory regions into different areas of the genome, however due to time constraints it was not deemed if the insertions were in the correct location or if the mutants were homozygous. Future experiments should confirm this. However, we did create a successful cloning and CRISPR/Cas9 protocol that has been proven to work for insertions and deletions. Thus, we analysed if the flanking regions would affect the expression of the reporter elements. We found that with the flanking regions of the constitutively expressed *PP2AA3,* reporter expression was universally expressed in both roots in leaves. However, with the cluster flanking regions, *MRN* and *THAS*, reporter expression was restricted to the roots. Interestingly, only the latter two lines showed a gap between the root tip and the rest of the root with no expression. It would be interesting to carry out qPCR and

ChIP on these mutant plants to analyse the changes in gene expression that is occurring, specifically in the root material where the clusters are expressed. Although we show that the deletion system is working, we did not manage to obtain any large deletions, however, we believe, with more time for screening that is achievable.

We showed an exemplar ChIP reaction for H3K4me3 enrichment in a H3K4me3 demethylase, *JMJ14*. In the future, it would be interesting to carry out this experiment on all the key histone modification mutants we analysed, along with their corresponding histone mark to identify those that cause the most dramatic changes in thalianol gene expression. Once those potential regulators of the thalianol cluster are identified, this experimentation should be carried out for the arabidiol and marneral gene clusters to identify conserved functions of these regulatory enzymes. Based on our current data, including our ReMap bioinformatic analysis, H3K18ac, H3K4me3, H3K9me2, H3K27me3 and H3K36me3 are the most likely signature marks of metabolic gene clusters. As the ReMap data was independently carried out in separate labs, the ChIP-seq experimentation would need to be repeated under the same conditions, separating leaf and root material. Alongside this, we suggest *that JMJ16, SDG4, JMJ15* and *ATX3* are key regulators of the chromatin structure across the thalianol gene cluster and important for its precise expression in roots.

By further pinpointing key regulators of the thalianol cluster and expanding this to the other metabolic gene clusters in *A. thaliana* we can attempt to extrapolate our findings to other metabolic gene clusters. This could aid in the fight against crop pathogens, food production or by contributing to the production of medicinal products. For an example of how understanding metabolic gene clusters can aid in food production, the gene cluster which produces the tomato steroidal glycoalkaloid α-tomatine will be discussed[139]. Neurotoxic proteins can be produced in unripe fruits, such as α-tomatine which gets converted into esculeoside A during ripening in tomatoes.[340] α-tomatine can also act as a chemical barrier against pathogens as it is a steroidal alkaloid.[356] With the aim to produce tomatoes that ripen quicker, the α-tomatine biosynthesis pathway could be genetically modified in order to convert more α-tomatine into esculeoside . However, alteration of the metabolic pathway by silencing one of the key metabolic genes in the cluster which is downregulated during ripening, is known to cause toxicity to the plant cell leading to marked developmental defects including reduction in growth.[356] By having a greater understanding of the fundamental regulatory mechanisms of metabolic gene clusters, this gene cluster could be modified in such a way to quicken the ripening process, which could be through up or downregulation of key genes by histone modification marks, introducing specific mutations or altering regulatory sites. This can also be extrapolated onto plant pathogen defence mechanisms and medicine production, such as penicillin from fungi[123], to upregulate or even transfer these mechanisms to other species, allowing for the increase in crop yield or medicine production.

Genome editing technology is used for a range of purposes in various organisms in biology, from gene therapy to cure human disease[357,358], understanding protein function[359] and agriculture[360]. However, the implications of altering an organisms' genome is not fully understood, nor studied.[359] This is partly due to not fully understanding the complex mechanisms in which genes can be regulated.[361] Here, we show that regulatory elements do not solely exist in the a genes promoter and terminator regions and that the three-dimensional architecture and environment is of importance for regulation. We show that it is vital to understand the impact of either inserting genes into other loci and/or causing mutations in the genome and the potential effect it could have of neighbouring genes. Another way in which we can alter the expression of genes indirectly would be to identify a signature histone modification mark and/or the regulator of the mark at the specific genomic region. Then, the eraser or depositor of the histone mark could be altered to reduce or increase the gene of interest, depending on the context.[362] For example based on our findings, in the context of the thalianol gene cluster to reduce its expression we could knockout *JMJ16*, which we showed led to a decrease in expression of the thalianol gene cluster but not the surrounding non-cluster genes. This method of altering gene expression without altering the gene of interest could aid in reducing the selective pressures which lead to pathogen resistance.[363] By switching genes on and off when necessary through histone modification marks, the resistance genes will not be overused, thus reducing these selective pressures for the pathogen, particularly in agriculture where pesticides are the main defence mechanism against pathogens.[363] To further reduce the need for genome editing and in turn increasing the specificity of changing the gene of interest expression, the regulatory protein of interest can be tagged with a degrader. For an example, if you want to reduce the gene expression a histone methyltransferase for H3K4me3 could be tagged with the degradation molecule, and a chemical added to reduce its activity by degrading the methyltransferase at a precise timepoint.[364] This method is therefore rapid and reversible and can be used in a wide array of contexts.

Aside from the importance of understanding the regulation of metabolic gene clusters for human need, it is also important to get a better understanding of metabolic gene clusters for the fundamental basics in science. If we can understand the key regulators of the thalianol gene cluster, we can attempt to explain the evolution of gene clusters and how genes are positioned within the genome. We know this is of importance as our ecotype data showed a change in gene order altered gene expression, but we do not know, yet, why or how this is of importance. This could also aid our understanding of how these clusters are maintained under different selective pressures, as each ecotype comes from an environmental niche.

We conclude that our work provides key insights into the understanding all aspects of the regulation of metabolic gene clusters that can aid in the future direction of experiments with the thalianol gene cluster, and further expanding to other metabolic gene clusters. By understanding how and why these gene clusters are ordered in this manner and how they are maintained and regulated can provide insights into much larger aspects of science from evolution to the manufacture of genetically modified crops. We have shown how gene order is of importance and how disruption of specific regions can lead it misregulation of the surrounding genes, that is of great importance when genetically modifying a genome.

There are still lots of questions yet to be answered, such as why some metabolic pathways are clustered and why some are not? What elements in the non-coding region of the genome are vital for the regulation of genes and how do these alter the three-dimensional chromatin structure? How important is the three-dimension chromatin structure in gene regulation and how do insertional mutations or deletions change it? Overall, this thesis has provided vital direction for future experiments as well as insight into answering some of these important questions which has multiple applications across many different fields of study.

6 References

1. Daubin V, Szöllősi GJ. Horizontal Gene Transfer and the History of Life. *Cold Spring Harb Perspect Biol*. 2016;8(4):a018036. doi:10.1101/cshperspect.a018036
2. A.S. Quina CL. Chromatin structure and epigenetics. *Biochem Pharmacol*. 2006;72(11):1563-1569. doi:10.1016/j.bcp.2006.06.016
3. Luger K, Mäder AW, Richmond RK, Sargent DF, Richmond TJ. Crystal structure of the nucleosome core particle at 2.8 A resolution. *Nature*. 1997;389(6648):251-260. doi:10.1038/38444
4. Rogge RA, Kalashnikova AA, Muthurajan UM, Porter-Goff ME, Luger K, Hansen JC. Assembly of nucleosomal arrays from recombinant core histones and nucleosome positioning DNA. *J Vis Exp*. 2013;(79). doi:10.3791/50354
5. Szerlong HJ, Prenni JE, Nyborg JK, Hansen JC. Activator-dependent p300 acetylation of chromatin in vitro: enhancement of transcription by disruption of repressive nucleosome-nucleosome interactions. *J Biol Chem*. 2010;285(42):31954-31964. doi:10.1074/jbc.M110.148718
6. E H. Das heterochromatin der moose. *Jahrb Wiss Bot*. 1928;69:762-818.
7. Huisinga KL, Brower-Toland B, Elgin SCR. The contradictory definitions of heterochromatin: transcription and silencing. *Chromosoma*. 2006;115(2):110-122. doi:10.1007/s00412-006-0052-x
8. van Holde K, Zlatanova J. Chromatin Higher Order Structure: Chasing a Mirage? *J Biol Chem*. 1995;270(15):8373-8376. doi:10.1074/jbc.270.15.8373
9. Allfrey VG, Faulkner R, Mirsky AE. ACETYLATION AND METHYLATION OF HISTONES AND THEIR POSSIBLE ROLE IN THE REGULATION OF RNA SYNTHESIS. *Proc Natl Acad Sci*. 1964;51(5):786-794. doi:10.1073/pnas.51.5.786
10. Thorne AW, Kmiciek D, Mitchelson K, Sautiere P, Crane-Robinson C. Patterns of histone acetylation. *Eur J Biochem*. 1990;193(3):701-713. doi:10.1111/j.1432-1033.1990.tb19390.x
11. Turner BM. Reading signals on the nucleosome with a new nomenclature for modified histones. *Nat Struct Mol Biol*. 2005;12(2):110-112. doi:10.1038/nsmb0205-110
12. Hyun K, Jeon J, Park K, Kim J. Writing, erasing and reading histone lysine methylations. *Exp Mol Med*. 2017;49(4):e324-e324. doi:10.1038/emm.2017.11

13. Zhang X, Clarenz O, Cokus S, et al. Whole-genome analysis of histone H3 lysine 27 trimethylation in Arabidopsis. *PLoS Biol*. 2007;5(5):e129. doi:10.1371/journal.pbio.0050129

14. You Y, Sawikowska A, Neumann M, et al. Temporal dynamics of gene expression and histone marks at the Arabidopsis shoot meristem during flowering. *Nat Commun*. 2017;8:15120. doi:10.1038/ncomms15120

15. Grossniklaus U, Paro R. Transcriptional silencing by polycomb-group proteins. *Cold Spring Harb Perspect Biol*. 2014;6(11):a019331. doi:10.1101/cshperspect.a019331

16. Crevillén P. Histone Demethylases as Counterbalance to H3K27me3 Silencing in Plants. *iScience*. 2020;23(11):101715. doi:10.1016/j.isci.2020.101715

17. Margueron R, Reinberg D. The Polycomb complex PRC2 and its mark in life. *Nature*. 2011;469(7330):343-349. doi:10.1038/nature09784

18. Breiling A, Turner BM, Bianchi ME, Orlando V. General transcription factors bind promoters repressed by Polycomb group proteins. *Nature*. 2001;412(6847):651-655. doi:10.1038/35088090

19. Wiles ET, Selker EU. H3K27 methylation: a promiscuous repressive chromatin mark. *Curr Opin Genet Dev*. 2017;43:31-37. doi:10.1016/j.gde.2016.11.001

20. Dellino GI, Schwartz YB, Farkas G, McCabe D, Elgin SC., Pirrotta V. Polycomb Silencing Blocks Transcription Initiation. *Mol Cell*. 2004;13(6):887-893. doi:10.1016/S1097-2765(04)00128-5

21. Martens JHA, O'Sullivan RJ, Braunschweig U, et al. The profile of repeat-associated histone lysine methylation states in the mouse epigenome. *EMBO J*. 2005;24(4):800-812. doi:10.1038/sj.emboj.7600545

22. Xu L, Jiang H. Writing and Reading Histone H3 Lysine 9 Methylation in Arabidopsis. *Front Plant Sci*. 2020;11(May):1-10. doi:10.3389/fpls.2020.00452

23. Becker JS, Nicetto D, Zaret KS. H3K9me3-Dependent Heterochromatin: Barrier to Cell Fate Changes. *Trends Genet*. 2016;32(1):29-41. doi:10.1016/j.tig.2015.11.001

24. Charron J-BF, He H, Elling AA, Deng XW. Dynamic landscapes of four histone modifications during deetiolation in Arabidopsis. *Plant Cell*. 2009;21(12):3732-3748. doi:10.1105/tpc.109.066845

25. Bentsink L, Jowett J, Hanhart CJ, Koornneef M. Cloning of DOG1, a quantitative trait locus controlling seed dormancy in Arabidopsis. *Proc Natl Acad Sci U S A*.

2006;103(45):17042-17047. doi:10.1073/pnas.0607877103

26. Wang Q, Liu P, Jing H, et al. JMJ27-mediated histone H3K9 demethylation positively regulates drought-stress responses in Arabidopsis. *New Phytol*. 2021;232(1):221-236. doi:10.1111/nph.17593

27. Strahl BD, Grant PA, Briggs SD, et al. Set2 is a nucleosomal histone H3-selective methyltransferase that mediates transcriptional repression. *Mol Cell Biol*. 2002;22(5):1298-1306. doi:10.1128/MCB.22.5.1298-1306.2002

28. Li B, Howe L, Anderson S, Yates JR, Workman JL. The Set2 histone methyltransferase functions through the phosphorylated carboxyl-terminal domain of RNA polymerase II. *J Biol Chem*. 2003;278(11):8897-8903. doi:10.1074/jbc.M212134200

29. Wagner EJ, Carpenter PB. Understanding the language of Lys36 methylation at histone H3. *Nat Rev Mol Cell Biol*. 2012;13(2):115-126. doi:10.1038/nrm3274

30. Barrera LO, Li Z, Smith AD, et al. Genome-wide mapping and analysis of active promoters in mouse embryonic stem cells and adult organs. *Genome Res*. 2008;18(1):46-59. doi:10.1101/gr.6654808

31. Bell O, Conrad T, Kind J, Wirbelauer C, Akhtar A, Schübeler D. Transcription-coupled methylation of histone H3 at lysine 36 regulates dosage compensation by enhancing recruitment of the MSL complex in Drosophila melanogaster. *Mol Cell Biol*. 2008;28(10):3401-3409. doi:10.1128/MCB.00006-08

32. He Y. Control of the transition to flowering by chromatin modifications. *Mol Plant*. 2009;2(4):554-564. doi:10.1093/mp/ssp005

33. Deng W, Ying H, Helliwell CA, Taylor JM, Peacock WJ, Dennis ES. FLOWERING LOCUS C (FLC) regulates development pathways throughout the life cycle of Arabidopsis. *Proc Natl Acad Sci U S A*. 2011;108(16):6680-6685. doi:10.1073/pnas.1103175108

34. Li Y, Brooks M, Yeoh-Wang J, et al. SDG8-Mediated Histone Methylation and RNA Processing Function in the Response to Nitrate Signaling. *Plant Physiol*. 2020;182(1):215-227. doi:10.1104/pp.19.00682

35. Pajoro A, Severing E, Angenent GC, Immink RGH. Histone H3 lysine 36 methylation affects temperature-induced alternative splicing and flowering in plants. *Genome Biol*. 2017;18(1):102. doi:10.1186/s13059-017-1235-x

36. Santos-Rosa H, Schneider R, Bannister AJ, et al. Active genes are tri-methylated at K4 of histone H3. *Nature*. 2002;419(6905):407-411. doi:10.1038/nature01080

37. Heintzman ND, Stuart RK, Hon G, et al. Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. *Nat Genet*. 2007;39(3):311-318. doi:10.1038/ng1966

38. Park S, Kim GW, Kwon SH, Lee J. Broad domains of histone H3 lysine 4 trimethylation in transcriptional regulation and disease. *FEBS J*. 2020;287(14):2891-2902. doi:10.1111/febs.15219

39. Howe FS, Fischl H, Murray SC, Mellor J. Is H3K4me3 instructive for transcription activation? *Bioessays*. 2017;39(1):1-12. doi:10.1002/bies.201600095

40. Talbert PB, Henikoff S. Histone variants at a glance. *J Cell Sci*. 2021;134(6). doi:10.1242/jcs.244749

41. Talbert PB, Henikoff S. Histone variants--ancient wrap artists of the epigenome. *Nat Rev Mol Cell Biol*. 2010;11(4):264-275. doi:10.1038/nrm2861

42. Osakabe A, Lorković ZJ, Kobayashi W, et al. Histone H2A variants confer specific properties to nucleosomes and impact on chromatin accessibility. *Nucleic Acids Res*. 2018;46(15):7675-7685. doi:10.1093/nar/gky540

43. Talbert PB, Henikoff S. Environmental responses mediated by histone variants. *Trends Cell Biol*. 2014;24(11):642-650. doi:10.1016/j.tcb.2014.07.006

44. Lang J, Smetana O, Sanchez-Calderon L, et al. Plant γH2AX foci are required for proper DNA DSB repair responses and colocalize with E2F factors. *New Phytol*. 2012;194(2):353-363. doi:10.1111/j.1469-8137.2012.04062.x

45. Lorković ZJ, Park C, Goiser M, et al. Compartmentalization of DNA Damage Response between Heterochromatin and Euchromatin Is Mediated by Distinct H2A Histone Variants. *Curr Biol*. 2017;27(8):1192-1199. doi:10.1016/j.cub.2017.03.002

46. Seo J, Kim SC, Lee H-S, et al. Genome-wide profiles of H2AX and γ-H2AX differentiate endogenous and exogenous DNA damage hotspots in human cells. *Nucleic Acids Res*. 2012;40(13):5965-5974. doi:10.1093/nar/gks287

47. Eleuteri B, Aranda S, Ernfors P. NoRC Recruitment by H2A.X Deposition at rRNA Gene Promoter Limits Embryonic Stem Cell Proliferation. *Cell Rep*. 2018;23(6):1853-1866. doi:10.1016/j.celrep.2018.04.023

48. Celeste A, Petersen S, Romanienko PJ, et al. Genomic instability in mice lacking histone H2AX. *Science*. 2002;296(5569):922-927. doi:10.1126/science.1069398

49. Giaimo BD, Ferrante F, Herchenröther A, Hake SB, Borggrefe T. The histone variant

H2A.Z in gene regulation. *Epigenetics Chromatin*. 2019;12(1):37. doi:10.1186/s13072-019-0274-9

50. Horikoshi N, Arimura Y, Taguchi H, Kurumizaka H. Crystal structures of heterotypic nucleosomes containing histones H2A.Z and H2A. *Open Biol*. 2016;6(6). doi:10.1098/rsob.160127

51. Horikoshi N, Sato K, Shimada K, et al. Structural polymorphism in the L1 loop regions of human H2A.Z.1 and H2A.Z.2. *Acta Crystallogr D Biol Crystallogr*. 2013;69(Pt 12):2431-2439. doi:10.1107/S090744491302252X

52. Adam M, Robert F, Larochelle M, Gaudreau L. H2A.Z is required for global chromatin integrity and for recruitment of RNA polymerase II under specific conditions. *Mol Cell Biol*. 2001;21(18):6270-6279. doi:10.1128/MCB.21.18.6270-6279.2001

53. Goldman JA, Garlick JD, Kingston RE. Chromatin remodeling by imitation switch (ISWI) class ATP-dependent remodelers is stimulated by histone variant H2A.Z. *J Biol Chem*. 2010;285(7):4645-4651. doi:10.1074/jbc.M109.072348

54. Dann GP, Liszczak GP, Bagert JD, et al. ISWI chromatin remodellers sense nucleosome modifications to determine substrate preference. *Nature*. 2017;548(7669):607-611. doi:10.1038/nature23671

55. Li B, Pattenden SG, Lee D, et al. Preferential occupancy of histone variant H2AZ at inactive promoters influences local histone modifications and chromatin remodeling. *Proc Natl Acad Sci U S A*. 2005;102(51):18385-18390. doi:10.1073/pnas.0507975102

56. Raisner RM, Hartley PD, Meneghini MD, et al. Histone variant H2A.Z marks the 5’ ends of both active and inactive genes in euchromatin. *Cell*. 2005;123(2):233-248. doi:10.1016/j.cell.2005.10.002

57. Sadeghi L, Bonilla C, Strålfors A, Ekwall K, Svensson JP. Podbat: a novel genomic tool reveals Swr1-independent H2A.Z incorporation at gene coding sequences through epigenetic meta-analysis. *PLoS Comput Biol*. 2011;7(8):e1002163. doi:10.1371/journal.pcbi.1002163

58. Wan Y, Saleem RA, Ratushny A V, et al. Role of the histone variant H2A.Z/Htz1p in TBP recruitment, chromatin dynamics, and regulated expression of oleate-responsive genes. *Mol Cell Biol*. 2009;29(9):2346-2358. doi:10.1128/MCB.01233-08

59. Millar CB, Xu F, Zhang K, Grunstein M. Acetylation of H2AZ Lys 14 is associated with genome-wide gene activity in yeast. *Genes Dev*. 2006;20(6):711-722.

doi:10.1101/gad.1395506

60. March-Díaz R, García-Domínguez M, Lozano-Juste J, León J, Florencio FJ, Reyes JC. Histone H2A.Z and homologues of components of the SWR1 complex are required to control immunity in Arabidopsis. *Plant J*. 2008;53(3):475-487. doi:10.1111/j.1365-313X.2007.03361.x

61. Carter B, Bishop B, Ho KK, et al. The Chromatin Remodelers PKL and PIE1 Act in an Epigenetic Pathway That Determines H3K27me3 Homeostasis in Arabidopsis. *Plant Cell*. 2018;30(6):1337-1352. doi:10.1105/tpc.17.00867

62. Bönisch C, Hake SB. Histone H2A variants in nucleosomes and chromatin: more or less stable? *Nucleic Acids Res*. 2012;40(21):10719-10741. doi:10.1093/nar/gks865

63. The histone variant H2A.W cooperates with chromatin modifications and linker histone H1 to maintain transcriptional silencing of transposons in Arabidopsis. doi:https://doi.org/10.1101/2022.05.31.493688

64. Talbert PB, Henikoff S. Histone variants on the move: substrates for chromatin dynamics. *Nat Rev Mol Cell Biol*. 2017;18(2):115-126. doi:10.1038/nrm.2016.148

65. Ingouff M, Rademacher S, Holec S, et al. Zygotic resetting of the HISTONE 3 variant repertoire participates in epigenetic reprogramming in Arabidopsis. *Curr Biol*. 2010;20(23):2137-2143. doi:10.1016/j.cub.2010.11.012

66. Akiyama T, Suzuki O, Matsuda J, Aoki F. Dynamic replacement of histone H3 variants reprograms epigenetic marks in early mouse embryos. *PLoS Genet*. 2011;7(10):e1002279. doi:10.1371/journal.pgen.1002279

67. Hödl M, Basler K. Transcription in the absence of histone H3.3. *Curr Biol*. 2009;19(14):1221-1226. doi:10.1016/j.cub.2009.05.048

68. Goldberg AD, Banaszynski LA, Noh K-M, et al. Distinct factors control histone variant H3.3 localization at specific genomic regions. *Cell*. 2010;140(5):678-691. doi:10.1016/j.cell.2010.01.003

69. Wollmann H, Stroud H, Yelagandula R, et al. The histone H3 variant H3.3 regulates gene body DNA methylation in Arabidopsis thaliana. *Genome Biol*. 2017;18(1):94. doi:10.1186/s13059-017-1221-3

70. Mishiba K, Nishihara M, Nakatsuka T, et al. Consistent transcriptional silencing of 35S-driven transgenes in gentian. *Plant J*. 2005;44(4):541-556. doi:10.1111/j.1365-313X.2005.02556.x

71. McKittrick E, Gafken PR, Ahmad K, Henikoff S. Histone H3.3 is enriched in covalent modifications associated with active chromatin. *Proc Natl Acad Sci U S A*. 2004;101(6):1525-1530. doi:10.1073/pnas.0308092100

72. Hake SB, Allis CD. Histone H3 variants and their potential role in indexing mammalian genomes: The "H3 barcode hypothesis." *Proc Natl Acad Sci*. 2006;103(17):6428-6435. doi:10.1073/pnas.0600803103

73. Koessler H, Doenecke D, Albig W. Aberrant Expression Pattern of Replication-Dependent Histone H3 Subtype Genes in Human Tumor Cell Lines. *DNA Cell Biol*. 2003;22(4):233-241. doi:10.1089/104454903321908629

74. Wu RS, Bonner WM. Separation of basal histone synthesis from S-phase histone synthesis in dividing cells. *Cell*. 1981;27(2):321-330. doi:10.1016/0092-8674(81)90415-3

75. Lieberman-Aiden E, van Berkum NL, Williams L, et al. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science*. 2009;326(5950):289-293. doi:10.1126/science.1181369

76. Liu Y, Nanni L, Sungalee S, et al. Systematic inference and comparison of multi-scale chromatin sub-compartments connects spatial organization to cell phenotypes. *Nat Commun*. 2021;12(1):2439. doi:10.1038/s41467-021-22666-3

77. Zheng H, Xie W. The role of 3D genome organization in development and cell differentiation. *Nat Rev Mol Cell Biol*. 2019;20(9):535-550. doi:10.1038/s41580-019-0132-4

78. Quinodoz SA, Ollikainen N, Tabak B, et al. Higher-Order Inter-chromosomal Hubs Shape 3D Genome Organization in the Nucleus. *Cell*. 2018;174(3):744-757.e24. doi:10.1016/j.cell.2018.05.024

79. Nuebler J, Fudenberg G, Imakaev M, Abdennur N, Mirny LA. Chromatin organization by an interplay of loop extrusion and compartmental segregation. *Proc Natl Acad Sci U S A*. 2018;115(29):E6697-E6706. doi:10.1073/pnas.1717730115

80. Dixon JR, Jung I, Selvaraj S, et al. Chromatin architecture reorganization during stem cell differentiation. *Nature*. 2015;518(7539):331-336. doi:10.1038/nature14222

81. Dixon JR, Selvaraj S, Yue F, et al. Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature*. 2012;485(7398):376-380. doi:10.1038/nature11082

82. Shen Y, Yue F, McCleary DF, et al. A map of the cis-regulatory sequences in the mouse genome. *Nature*. 2012;488(7409):116-120. doi:10.1038/nature11243

83. Rao SSP, Huntley MH, Durand NC, et al. A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell*. 2014;159(7):1665-1680. doi:10.1016/j.cell.2014.11.021

84. Liu C, Cheng Y-J, Wang J-W, Weigel D. Prominent topologically associated domains differentiate global chromatin packing in rice from Arabidopsis. *Nat plants*. 2017;3(9):742-748. doi:10.1038/s41477-017-0005-9

85. Nora EP, Lajoie BR, Schulz EG, et al. Spatial partitioning of the regulatory landscape of the X-inactivation centre. *Nature*. 2012;485(7398):381-385. doi:10.1038/nature11049

86. Sexton T, Yaffe E, Kenigsberg E, et al. Three-dimensional folding and functional organization principles of the Drosophila genome. *Cell*. 2012;148(3):458-472. doi:10.1016/j.cell.2012.01.010

87. Fudenberg G, Imakaev M, Lu C, Goloborodko A, Abdennur N, Mirny LA. Formation of Chromosomal Domains by Loop Extrusion. *Cell Rep*. 2016;15(9):2038-2049. doi:10.1016/j.celrep.2016.04.085

88. Guo Y, Xu Q, Canzio D, et al. CRISPR Inversion of CTCF Sites Alters Genome Topology and Enhancer/Promoter Function. *Cell*. 2015;162(4):900-910. doi:10.1016/j.cell.2015.07.038

89. Heger P, Marin B, Bartkuhn M, Schierenberg E, Wiehe T. The chromatin insulator CTCF and the emergence of metazoan diversity. *Proc Natl Acad Sci U S A*. 2012;109(43):17507-17512. doi:10.1073/pnas.1111941109

90. Ouyang W, Xiong D, Li G, Li X. Unraveling the 3D Genome Architecture in Plants: Present and Future. *Mol Plant*. 2020;13(12):1676-1693. doi:10.1016/j.molp.2020.10.002

91. Wang C, Liu C, Roqueiro D, et al. Genome-wide analysis of local chromatin packing in Arabidopsis thaliana. *Genome Res*. 2015;25(2):246-256. doi:10.1101/gr.170332.113

92. Dong P, Tu X, Chu P-Y, et al. 3D Chromatin Architecture of Large Plant Genomes Determined by Local A/B Compartments. *Mol Plant*. 2017;10(12):1497-1509. doi:10.1016/j.molp.2017.11.005

93. Kadauke S, Blobel GA. Chromatin loops in gene regulation. *Biochim Biophys Acta*. 2009;1789(1):17-25. doi:10.1016/j.bbagrm.2008.07.002

94. Bonev B, Cavalli G. Organization and function of the 3D genome. *Nat Rev Genet*.

2016;17(11):661-678. doi:10.1038/nrg.2016.112

95. Xu W, Xu H, Li K, et al. The R-loop is a common chromatin feature of the Arabidopsis genome. *Nat plants*. 2017;3(9):704-714. doi:10.1038/s41477-017-0004-x

96. Tan-Wong SM, Zaugg JB, Camblong J, et al. Gene loops enhance transcriptional directionality. *Science*. 2012;338(6107):671-675. doi:10.1126/science.1224350

97. Crevillén P, Sonmez C, Wu Z, Dean C. A gene loop containing the floral repressor FLC is disrupted in the early phase of vernalization. *EMBO J*. 2013;32(1):140-148. doi:10.1038/emboj.2012.324

98. Wang H, Li S, Li Y, et al. MED25 connects enhancer-promoter looping and MYC2-dependent activation of jasmonate signalling. *Nat plants*. 2019;5(6):616-625. doi:10.1038/s41477-019-0441-9

99. Ricci MA, Manzo C, García-Parajo MF, Lakadamyali M, Cosma MP. Chromatin fibers are formed by heterogeneous groups of nucleosomes in vivo. *Cell*. 2015;160(6):1145-1158. doi:10.1016/j.cell.2015.01.054

100. Hurst LD, Pál C, Lercher MJ. The evolutionary dynamics of eukaryotic gene order. *Nat Rev Genet*. 2004;5(4):299-310. doi:10.1038/nrg1319

101. Kruglyak S, Tang H. Regulation of adjacent yeast genes. *Trends Genet*. 2000;16(3):109-111. doi:10.1016/S0168-9525(99)01941-1

102. Cohen BA, Mitra RD, Hughes JD, Church GM. A computational analysis of whole-genome expression data reveals chromosomal domains of gene expression. *Nat Genet*. 2000;26:183-186. doi:10.1038/79896

103. Trinklein ND, Aldred SF, Hartman SJ, Schroeder DI, Otillar RP, Myers RM. An abundance of bidirectional promoters in the human genome. *Genome Res*. 2004;14(1):62-66. doi:10.1101/gr.1982804

104. Dávila López M, Martínez Guerra JJ, Samuelsson T. Analysis of gene order conservation in eukaryotes identifies transcriptionally and functionally linked genes. *PLoS One*. 2010;5(5):e10654. doi:10.1371/journal.pone.0010654

105. Pettitt J, Philippe L, Sarkar D, et al. Operons are a conserved feature of nematode genomes. *Genetics*. 2014;197(4):1201-1211. doi:10.1534/genetics.114.162875

106. Field B, Osbourn AE. Metabolic diversification--independent assembly of operon-like gene clusters in different plants. *Science (80-)*. 2008;194(April):543-547.

107. Osbourn A. Secondary metabolic gene clusters: evolutionary toolkits for chemical

innovation. *Trends Genet*. 2010;26(10):449-457. doi:10.1016/j.tig.2010.07.001

108. Vieux-Rochas M, Fabre PJ, Leleu M, Duboule D, Noordermeer D. Clustering of mammalian Hox genes with other H3K27me3 targets within an active nuclear domain. *Proc Natl Acad Sci U S A*. 2015;112(15):4672-4677. doi:10.1073/pnas.1504783112

109. Freund CL MR. Guidebook to the Homeobox Genes. *Am J Hum Genet*. 1995;57(3):736-737.

110. Wagner GP, Amemiya C, Ruddle F. Hox cluster duplications and the opportunity for evolutionary novelties. *Proc Natl Acad Sci U S A*. 2003;100(25):14603-14606. doi:10.1073/pnas.2536656100

111. Chambeyron S, Bickmore WA. Chromatin decondensation and nuclear reorganization of the HoxB locus upon induction of transcription. *Genes Dev*. 2004;18(10):1119-1130. doi:10.1101/gad.292104

112. Soshnikova N, Duboule D. Epigenetic temporal control of mouse Hox genes in vivo. *Science*. 2009;324(5932):1320-1323. doi:10.1126/science.1171468

113. Andrey G, Montavon T, Mascrez B, et al. A switch between topological domains underlies HoxD genes collinearity in mouse limbs. *Science*. 2013;340(6137):1234167. doi:10.1126/science.1234167

114. Tarchini B, Duboule D. Control of Hoxd genes' collinearity during early limb development. *Dev Cell*. 2006;10(1):93-103. doi:10.1016/j.devcel.2005.11.014

115. Nützmann H-W, Scazzocchio C, Osbourn A. Metabolic Gene Clusters in Eukaryotes. *Annu Rev Genet*. 2018;52(1):159-183. doi:10.1146/annurev-genet-120417-031237

116. Schilmiller AL, Last RL, Pichersky E. Harnessing plant trichome biochemistry for the production of useful compounds. *Plant J*. 2008;54(4):702-711. doi:10.1111/j.1365-313X.2008.03432.x

117. De Luca V, Salim V, Atsumi SM, Yu F. Mining the biodiversity of plants: A revolution in the making. *Science (80-)*. 2012;336(6089):1658-1661. doi:10.1126/science.1217410

118. Grimholt U. MHC and evolution in teleosts. *Biology (Basel)*. 2016;5(1). doi:10.3390/biology5010006

119. Huang A, Osbourn A, Centre JI. Plant metabolic clusters - from genetics to genomics. 2017;211(3):771-789. doi:10.1111/nph.13981.Plant

120. Knowles SL, Raja HA, Wright AJ, et al. Mapping the Fungal Battlefield: Using in situ Chemistry and Deletion Mutants to Monitor Interspecific Chemical Interactions

Between Fungi. *Front Microbiol*. 2019;10. doi:10.3389/fmicb.2019.00285

121. Keller NP. Fungal secondary metabolism: regulation, function and drug discovery. *Nat Rev Microbiol*. 2019;17(3):167-180. doi:10.1038/s41579-018-0121-1

122. Aharonowitz Y, Cohen G, Martin JF. Penicillin and cephalosporin biosynthetic genes: structure, organization, regulation, and evolution. *Annu Rev Microbiol*. 1992;46:461-495. doi:10.1146/annurev.mi.46.100192.002333

123. Laich F, Fierro F, Cardoza RE, Martin JF. Organization of the gene cluster for biosynthesis of penicillin in Penicillium nalgiovense and antibiotic production in cured dry sausages. *Appl Environ Microbiol*. 1999;65(3):1236-1240. doi:10.1128/AEM.65.3.1236-1240.1999

124. König CC, Scherlach K, Schroeckh V, et al. Bacterium induces cryptic meroterpenoid pathway in the pathogenic fungus Aspergillus fumigatus. *Chembiochem*. 2013;14(8):938-942. doi:10.1002/cbic.201300070

125. Peñalva MA, Tilburn J, Bignell E, Arst HN. Ambient pH gene regulation in fungi: making connections. *Trends Microbiol*. 2008;16(6):291-300. doi:10.1016/j.tim.2008.03.006

126. Macheleidt J, Mattern DJ, Fischer J, et al. Regulation and Role of Fungal Secondary Metabolites. *Annu Rev Genet*. 2016;50(1):371-392. doi:10.1146/annurev-genet-120215-035203

127. Lee I, Oh J-H, Shwab EK, Dagenais TRT, Andes D, Keller NP. HdaA, a class 2 histone deacetylase of Aspergillus fumigatus, affects germination and secondary metabolite production. *Fungal Genet Biol*. 2009;46(10):782-790. doi:10.1016/j.fgb.2009.06.007

128. Orekhova AS, Rubtsov PM. Bidirectional promoters in the transcription of mammalian genomes. *Biochemistry (Mosc)*. 2013;78(4):335-341. doi:10.1134/S0006297913040020

129. Lomvardas S, Barnea G, Pisapia DJ, Mendelsohn M, Kirkland J, Axel R. Interchromosomal Interactions and Olfactory Receptor Choice. *Cell*. 2006;126(2):403-413. doi:10.1016/j.cell.2006.06.035

130. Geyer PK, Green MM, Corces VG. Tissue-specific transcriptional enhancers may act in trans on the gene located in the homologous chromosome: the molecular basis of transvection in Drosophila. *EMBO J*. 1990;9(7):2247-2256. doi:10.1002/j.1460-2075.1990.tb07395.x

131. Ong C-T, Corces VG. Enhancer function: new insights into the regulation of tissue-

specific gene expression. *Nat Rev Genet*. 2011;12(4):283-293. doi:10.1038/nrg2957

132. Clapier CR, Cairns BR. The biology of chromatin remodeling complexes. *Annu Rev Biochem*. 2009;78:273-304. doi:10.1146/annurev.biochem.77.062706.153223

133. Andrulis ED, Neiman AM, Zappulla DC, Sternglanz R. Perinuclear localization of chromatin facilitates transcriptional silencing. *Nature*. 1998;394(6693):592-595. doi:10.1038/29100

134. Pott S, Lieb JD. What are super-enhancers? *Nat Genet*. 2015;47(1):8-12. doi:10.1038/ng.3167

135. Grosveld F, van Assendelft GB, Greaves DR, Kollias G. Position-independent, high-level expression of the human beta-globin gene in transgenic mice. *Cell*. 1987;51(6):975-985. doi:10.1016/0092-8674(87)90584-8

136. Magram J, Chada K, Costantini F. Developmental regulation of a cloned adult beta-globin gene in transgenic mice. *Nature*. 315(6017):338-340. doi:10.1038/315338a0

137. Kioussis D, Vanin E, DeLange T, Flavell RA, Grosveld FG. Beta-globin gene inactivation by DNA translocation in gamma beta-thalassaemia. *Nature*. 306(5944):662-666. doi:10.1038/306662a0

138. Driscoll MC, Dobkin CS, Alter BP. Gamma delta beta-thalassemia due to a de novo mutation deleting the 5' beta-globin gene activation-region hypersensitive sites. *Proc Natl Acad Sci U S A*. 1989;86(19):7470-7474. doi:10.1073/pnas.86.19.7470

139. Nützmann H-W, Huang A, Osbourn A. Plant metabolic clusters - from genetics to genomics. *New Phytol*. 2016;211(3):771-789. doi:10.1111/nph.13981

140. Nützmann HW, Doerr D, Ramírez-Colmenero A, et al. Active and repressed biosynthetic gene clusters have spatially distinct chromosome states. *Proc Natl Acad Sci U S A*. 2020. doi:10.1073/pnas.1920474117

141. Marszalek-Zenczak M, Satyr A, Wojciechowski P ZM. Copy number variations shape the structural diversity of Arabidopsis metabolic gene clusters and are associated with the climatic gradient. *bioRxiv*. 2022.

142. Ghosh S. Triterpene Structural Diversification by Plant Cytochrome P450 Enzymes. *Front Plant Sci*. 2017;8. doi:10.3389/fpls.2017.01886

143. Phillips DR, Rasbery JM, Bartel B, Matsuda SP. Biosynthetic diversity in plant triterpene cyclization. *Curr Opin Plant Biol*. 2006;9(3):305-314. doi:10.1016/j.pbi.2006.03.004

144. Misra RC, Maiti P, Chanotiya CS, Shanker K, Ghosh S. Methyl Jasmonate-Elicited

Transcriptional Responses and Pentacyclic Triterpene Biosynthesis in Sweet Basil. *Plant Physiol*. 2014;164(2):1028-1044. doi:10.1104/pp.113.232884

145. Delis C, Krokida A, Georgiou S, et al. Role of lupeol synthase in Lotus japonicus nodule formation. *New Phytol*. 2011;189(1):335-346. doi:10.1111/j.1469-8137.2010.03463.x

146. Kemen AC, Honkanen S, Melton RE, et al. Investigation of triterpene synthesis and regulation in oats reveals a role for β-amyrin in determining root epidermal cell patterning. *Proc Natl Acad Sci*. 2014;111(23):8679-8684. doi:10.1073/pnas.1401553111

147. Papadopoulou K, Melton RE, Leggett M, Daniels MJ, Osbourn AE. Compromised disease resistance in saponin-deficient plants. *Proc Natl Acad Sci*. 1999;96(22):12923-12928. doi:10.1073/pnas.96.22.12923

148. Moses T, Pollier J, Thevelein JM, Goossens A. Bioengineering of plant (tri)terpenoids: from metabolic engineering of plants to synthetic biology in vivo and in vitro. *New Phytol*. 2013;200(1):27-43. doi:10.1111/nph.12325

149. Osbourn A, Goss RJM, Field RA. The saponins – polar isoprenoids with important and diverse biological activities. *Nat Prod Rep*. 2011;28(7):1261. doi:10.1039/c1np00015b

150. Sawai S, Saito K. Triterpenoid biosynthesis and engineering in plants. *Front Plant Sci*. 2011;2:25. doi:10.3389/fpls.2011.00025

151. Field B, Fiston-Lavier AS, Kemen A, Geisler K, Quesneville H, Osbourn AE. Formation of plant metabolic gene clusters within dynamic chromosomal regions. *Proc Natl Acad Sci U S A*. 2011;108(38):16116-16121. doi:10.1073/PNAS.1109273108/-/DCSUPPLEMENTAL/PNAS.201109273SI.PDF

152. Xiong Q, Wilson WK, Matsuda SPT. An Arabidopsis oxidosqualene cyclase catalyzes iridal skeleton formation by Grob fragmentation. *Angew Chem Int Ed Engl*. 2006;45(8):1285-1288. doi:10.1002/anie.200503420

153. Field B, Fiston-Lavier AS, Kemen A, Geisler K, Quesneville H, Osbourn AE. Formation of plant metabolic gene clusters within dynamic chromosomal regions. *Proc Natl Acad Sci U S A*. 2011;108(38):16116-16121. doi:10.1073/pnas.1109273108

154. Boutanaev AM, Moses T, Zi J, et al. Investigation of terpene diversification across multiple sequenced plant genomes. *Proc Natl Acad Sci*. 2015;112(1). doi:10.1073/pnas.1419547112

155. Sohrabi R, Huh JH, Badieyan S, et al. In planta variation of volatile biosynthesis: An

alternative biosynthetic route to the formation of the pathogen-induced volatile homoterpene DMNT via triterpene degradation in arabidopsis roots. *Plant Cell*. 2015. doi:10.1105/tpc.114.132209

156. Sohrabi R, Ali T, Harinantenaina Rakotondraibe L, Tholl D. Formation and exudation of non-volatile products of the arabidiol triterpenoid degradation pathway in Arabidopsis roots. *Plant Signal Behav*. 2017;12(1):e1265722. doi:10.1080/15592324.2016.1265722

157. Huang AC, Jiang T, Liu YX, et al. A specialized metabolic network selectively modulates Arabidopsis root microbiota. *Science (80-)*. 2019;364(6440). doi:10.1126/science.aau6389

158. Bai Y, Fernández-Calvo P, Ritter A, et al. Modulation of Arabidopsis root growth by specialized triterpenes. *New Phytol*. 2021;230(1):228-243. doi:10.1111/nph.17144

159. Liu Z, Cheema J, Vigouroux M, et al. Formation and diversification of a paradigm biosynthetic gene cluster in plants. *Nat Commun*. 2020;11(1):1-11. doi:10.1038/s41467-020-19153-6

160. Freeling M, Lyons E, Pedersen B, Alam M, Ming R, Lisch D. Many or most genes in Arabidopsis transposed after the origin of the order Brassicales. *Genome Res*. 2008;18(12):1924-1937. doi:10.1101/gr.081026.108

161. Yu N, Nützmann HW, Macdonald JT, et al. Delineation of metabolic gene clusters in plant genomes by chromatin signatures. *Nucleic Acids Res*. 2016;44(5):2255-2265. doi:10.1093/nar/gkw100

162. Osbourn A, Nützmann H-W. Regulation of metabolic gene clusters in Arabidopsis thaliana. *New Phytol*. 2015;205:503-510.

163. Nützmann HW, Doerr D, Ramírez-Colmenero A, et al. Active and repressed biosynthetic gene clusters have spatially distinct chromosome states. *Proc Natl Acad Sci U S A*. 2020;117(24):13800-13809. doi:10.1073/pnas.1920474117

164. Krämer U. Planting molecular functions in an ecological context with Arabidopsis thaliana. *Elife*. 2015;4. doi:10.7554/eLife.06100

165. Koornneef M, Meinke D. The development of Arabidopsis as a model plant. *Plant J*. 2010;61(6):909-921. doi:10.1111/j.1365-313X.2009.04086.x

166. Szymańska R, Gabruk M, Kruk J. [Arabidopsis thaliana accessions - a tool for biochemical and phylogentical studies]. *Postepy Biochem*. 2015;61(1):102-113. http://www.ncbi.nlm.nih.gov/pubmed/26281359.

167. Nützmann H, Doerr D, Ramírez-Colmenero A, et al. Active and repressed biosynthetic gene clusters have spatially distinct chromosome states. *Proc Natl Acad Sci*. 2020;117(24):13800-13809. doi:10.1073/pnas.1920474117

168. Patron NJ, Waller RF, Cozijnsen AJ, et al. Origin and distribution of epipolythiodioxopiperazine (ETP) gene clusters in filamentous ascomycetes. *BMC Evol Biol*. 2007;7(1):174. doi:10.1186/1471-2148-7-174

169. Wu D, Jiang B, Ye C-Y, Timko MP, Fan L. Horizontal transfer and evolution of the biosynthetic gene cluster for benzoxazinoids in plants. *Plant Commun*. 2022;3(3):100320. doi:10.1016/j.xplc.2022.100320

170. Chu HY, Wegel E, Osbourn A. From hormones to secondary metabolism: The emergence of metabolic gene clusters in plants. *Plant J*. 2011;66(1). doi:10.1111/j.1365-313X.2011.04503.x

171. Hurst LD, Pál C, Lercher MJ. The evolutionary dynamics of eukaryotic gene order. *Nat Rev Genet*. 2004;5(4):299-310. doi:10.1038/nrg1319

172. Mylona P, Owatworakit A, Papadopoulou K, et al. Sad3 and Sad4 Are Required for Saponin Biosynthesis and Root Development in Oat. *Plant Cell*. 2008;20(1):201-212. doi:10.1105/tpc.107.056531

173. Scholl RL, May ST, Ware DH. Seed and molecular resources for Arabidopsis. *Plant Physiol*. 2000;124(4):1477-1480. doi:10.1104/pp.124.4.1477

174. Miki D, Zhang W, Zeng W, Feng Z, Zhu JK. CRISPR/Cas9-mediated gene targeting in Arabidopsis using sequential transformation. *Nat Commun*. 2018;9(1):1-9. doi:10.1038/s41467-018-04416-0

175. Chèneby J, Ménétrier Z, Mestdagh M, et al. ReMap 2020: A database of regulatory regions from an integrative analysis of Human and Arabidopsis DNA-binding sequencing experiments. *Nucleic Acids Res*. 2020;48(D1):D180-D188. doi:10.1093/nar/gkz945

176. Fousse L, Hanrot G, Lefèvre V, Pélissier P, Zimmermann P. MPFR: A multiple-precision binary floating-point library with correct rounding. *ACM Trans Math Softw*. 2007;33(2):13. doi:10.1145/1236463.1236468

177. Wickham H, Averick M, Bryan J, et al. Welcome to the Tidyverse. *J Open Source Softw*. 2019;4(43):1686. doi:10.21105/joss.01686

178. Potter KC, Wang J, Schaller GE, Kieber JJ. Cytokinin modulates context-dependent

chromatin accessibility through the type-B response regulators. *Nat Plants*. 2018;4(12):1102-1111. doi:10.1038/s41477-018-0290-y

179. Fojtová M, Peška V, Dobšáková Z, Mozgová I, Fajkus J, Sýkorová E. Molecular analysis of T-DNA insertion mutants identified putative regulatory elements in the AtTERT gene. *J Exp Bot*. 2011;62(15):5531-5545. doi:10.1093/jxb/err235

180. Hsu PD, Lander ES, Zhang F. Development and applications of CRISPR-Cas9 for genome engineering. *Cell*. 2014;157(6):1262-1278. doi:10.1016/j.cell.2014.05.010

181. Zhang F, Wen Y, Guo X. CRISPR/Cas9 for genome editing: Progress, implications and challenges. *Hum Mol Genet*. 2014;23(R1):40-46. doi:10.1093/hmg/ddu125

182. Urnov FD, Rebar EJ, Holmes MC, Zhang HS, Gregory PD. Genome editing with engineered zinc finger nucleases. *Nat Rev Genet*. 2010;11(9):636-646. doi:10.1038/nrg2842

183. Bhakta MS, Henry IM, Ousterout DG, et al. Highly active zinc-finger nucleases by extended modular assembly. *Genome Res*. 2013;23(3):530-538. doi:10.1101/gr.143693.112

184. Joung JK, Sander JD. TALENs: a widely applicable technology for targeted genome editing. *Nat Rev Mol Cell Biol*. 2013;14(1):49-55. doi:10.1038/nrm3486.TALENs

185. Boch J, Bonas U. Xanthomonas AvrBs3 Family-Type III Effectors: Discovery and Function . *Annu Rev Phytopathol*. 2010;48(1):419-436. doi:10.1146/annurev-phyto-080508-081936

186. Martin Jinek, Krzysztof Chylinski, Ines Fonfara, Michael Hauer, Jennifer A. Doudna, Emmanuelle Charpentier. A Programmable Dual-RNA–Guided DNA Endonuclease in Adaptive Bacterial Immunity. *Science (80-)*. 2012;337(6096):816-821. doi:10.1126/science.1138140

187. Wiedenheft B, Sternberg SH, Doudna JA. RNA-guided genetic silencing systems in bacteria and archaea. *Nature*. 2012. doi:10.1038/nature10886

188. Jansen R, Van Embden JDA, Gaastra W, Schouls LM. Identification of genes that are associated with DNA repeats in prokaryotes. *Mol Microbiol*. 2002. doi:10.1046/j.1365-2958.2002.02839.x

189. Ran, F. Hsu PWJAVSDZ f. Genome engineering using the CRISPR-Cas9 system. *Nat Protoc*. 2013;8(11):2281-2308. doi:10.1038/nprot.2013.143.Genome

190. Shah SA, Erdmann S, Mojica FJM, Garrett RA. Protospacer recognition motifs: mixed

identities and functional diversity. *RNA Biol*. 2013;10(5):891-899. doi:10.4161/rna.23764

191. Deltcheva E, Chylinski K, Sharma CM, Gonzales K. CRISPR RNA maturation by trans-encoded small RNA and host factor RNase III. *Nature*. 2011;471(7340):602-607. doi:10.1038/nature09886.CRISPR

192. Garneau JE, Dupuis MÈ, Villion M, et al. The CRISPR/cas bacterial immune system cleaves bacteriophage and plasmid DNA. *Nature*. 2010;468(7320):67-71. doi:10.1038/nature09523

193. Cong L, Ran FA, Cox D, et al. Multiplex Genome Engineering Using CRISPR/Cas Systems. *Science (80-)*. 2013;339(6121):819-823. doi:10.1126/science.1231143.Multiplex

194. Mali P, Yang L, Esvelt KM, et al. RNA-guided human genome engineering via Cas9. *Science (80-)*. 2013. doi:10.1126/science.1232033

195. Castel B, Tomlinson L, Locci F, Yang Y, Jones JDG. Optimization of T-DNA architecture for Cas9-mediated mutagenesis in Arabidopsis. *PLoS One*. 2019;14(1):1-20. doi:10.1371/journal.pone.0204778

196. Chang HHY, Pannunzio NR, Adachi N, Lieber MR. Non-homologous DNA end joining and alternative pathways to double-strand break repair. *Nat Rev Mol Cell Biol*. 2017;18(8):495-506. doi:10.1038/nrm.2017.48

197. Goodarzi AA, Yu Y, Riballo E, et al. DNA-PK autophosphorylation facilitates Artemis endonuclease activity. *EMBO J*. 2006;25:3880-3889. doi:10.1038/sj.emboj.7601255

198. Dynan WS, Yoo S. Interaction of Ku protein and DNA-dependent protein kinase catalytic subunit with nucleic acids. *Nucleic Acids Res*. 1998;26:1551-1559. doi:10.1093/nar/26.7.1551

199. Mohapatra S, Yannone SM, Lee SH, et al. Trimming of damaged 3’ overhangs of DNA double-strand breaks by the Metnase and Artemis endonucleases. *DNA Repair (Amst)*. 2013;12:442-432. doi:10.1016/j.dnarep.2013.03.005

200. Aparicio T, Baer R, Gautier J. DNA double-strand break repair pathway choice and cancer. *DNA Repair (Amst)*. 2014;19:169-175. doi:10.1016/j.dnarep.2014.03.014

201. Shahar OD, Ram EVSR, Shimshoni E, Hareli S, Meshorer E, Goldberg M. Live imaging of induced and controlled DNA double-strand break formation reveals extremely low repair by homologous recombination in human cells. *Oncogene*. 2012;31(30):3495-3504. doi:10.1038/onc.2011.516

202. Chen K, Wang Y, Zhang R, Zhang H, Gao C. CRISPR/Cas Genome Editing and Precision Plant Breeding in Agriculture. *Annu Rev Plant Biol*. 2019;70(1):667-697. doi:10.1146/annurev-arplant-050718-100049

203. Khumsupan P, Donovan S, McCormick AJ. CRISPR/Cas in Arabidopsis: overcoming challenges to accelerate improvements in crop photosynthetic efficiencies. *Physiol Plant*. 2019;166(1):428-437. doi:10.1111/ppl.12937

204. Feng Z, Mao Y, Xu N, et al. Multigeneration analysis reveals the inheritance, specificity, and patterns of CRISPR/Cas-induced gene modifications in Arabidopsis. *Proc Natl Acad Sci U S A*. 2014;111(12):4632-4637. doi:10.1073/pnas.1400822111

205. Mao Y, Zhang Z, Feng Z, Wei P, Zhang H, Botella JR ZJ. Development of germ-line-specific CRISPR-Cas9 systems to improve the production of heritable gene modifications in Arabidopsis. *Plant Biotechnol*. 2016;14(2):519-532. doi:10.1016/j.physbeh.2017.03.040

206. Jana Ordon, Mauro Bressan, Carola Kretschmer, Luca Dall'Osto, Sylvestre Marillonnet, Roberto Bass JS. Optimized Cas9 expression systems for highly efficient Arabidopsis genome editing facilitate isolation of complex alleles in a single generation. *Funct Integr Genomics*. 2018;2:1-12. doi:10.1074/jbc.M109.058990

207. Pease LRJPANV. In vitro synthesis of novel genes : Mutagenesis and recombination by PCR. *Genome Res*. 1995;(May 2014):123-130. doi:10.1101/gr.4.3.S123

208. Geu-Flores F, Nour-Eldin HH, Nielsen MT, Halkier BA. USER fusion: A rapid and efficient method for simultaneous fusion and cloning of multiple PCR products. *Nucleic Acids Res*. 2007;35(7):0-5. doi:10.1093/nar/gkm106

209. Bo Salomonsen, Uffe H. Mortensen BAH. USER-Derived Cloning Methods and Their Primer Design. In: *DNA Cloning and Assembly Methods*. ; 2013:59-72.

210. Lebedenko EN, Birikh KR, Plutalov O V., Berlin YA. Method of artificial DNA spling by directed ligation (SDL). *Nucleic Acids Res*. 1991;19(24):6757-6761. doi:10.1093/nar/19.24.6757

211. Lu LLWJY. A Modified Gibson Assembly Method for Cloning Large DNA Fragments with High GC Contents. In: *Methods in Molecular Biology*. Vol 1671. ; 2018:203-209. doi:10.1007/978-1-4939-7295-1_5

212. Rashtchian A, Thornton CG, Heidecker G. A novel method for site-directed mutagenesis using PCR and uracil DNA glycosylase. *Genome Res*. 1992;2(2):124-130.

doi:10.1101/gr.2.2.124

213. Zapata L, Ding J, Willing E-M, et al. Chromosome-level assembly of Arabidopsis thaliana Ler reveals the extent of translocation and inversion polymorphisms. *Proc Natl Acad Sci U S A*. 2016;113(28):E4052-60. doi:10.1073/pnas.1607532113

214. Bharadwaj R, Kumar SR, Sharma A, Sathishkumar R. Plant Metabolic Gene Clusters: Evolution, Organization, and Their Applications in Synthetic Biology. *Front Plant Sci*. 2021;12:697318. doi:10.3389/fpls.2021.697318

215. Shimada TL, Shimada T, Hara-Nishimura I. A rapid and non-destructive screenable marker, FAST, for identifying transformed seeds of Arabidopsis thaliana. *Plant J*. 2010;61(3). doi:10.1111/j.1365-313X.2009.04060.x

216. Grandi FC, Modi H, Kampman L, Corces MR. Chromatin accessibility profiling by ATAC-seq. *Nat Protoc*. 2022;17(6):1518-1552. doi:10.1038/s41596-022-00692-9

217. de Nooijer S, Wellink J, Mulder B, Bisseling T. Non-specific interactions are sufficient to explain the position of heterochromatic chromocenters and nucleoli in interphase nuclei. *Nucleic Acids Res*. 2009;37(11):3558-3568. doi:10.1093/nar/gkp219

218. Fang Y, Spector DL. Centromere Positioning and Dynamics in Living Arabidopsis Plants. *Mol Biol Cell*. 2005;16(12):5710-5718. doi:10.1091/mbc.e05-08-0706

219. O'Malley RC, Barragan CC, Ecker JR. A user's guide to the Arabidopsis T-DNA insertion mutant collections. *Methods Mol Biol*. 2015;1284:323-342. doi:10.1007/978-1-4939-2444-8_16

220. Pucker B, Kleinbölting N, Weisshaar B. Large scale genomic rearrangements in selected Arabidopsis thaliana T-DNA lines are caused by T-DNA insertion mutagenesis. *BMC Genomics*. 2021;22(1):599. doi:10.1186/s12864-021-07877-8

221. Wang Q, Bao X, Chen S, et al. AtHDA6 functions as an H3K18ac eraser to maintain pericentromeric CHG methylation in Arabidopsis thaliana. *Nucleic Acids Res*. 2021;49(17):9755-9767. doi:10.1093/nar/gkab706

222. Molina C, Grotewold E. Genome wide analysis of Arabidopsis core promoters. *BMC Genomics*. 2005;6:25. doi:10.1186/1471-2164-6-25

223. Yadav V, Kundu S, Chattopadhyay D, et al. Light regulated modulation of Z-box containing promoters by photoreceptors and downstream regulatory components, COP1 and HY5, in Arabidopsis. *Plant J*. 2002;31(6):741-753. doi:10.1046/j.1365-313x.2002.01395.x

224. Zhao H, Yang M, Bishop J, et al. Identification and functional validation of super-enhancers in Arabidopsis thaliana. *Proc Natl Acad Sci*. 2022;119(48). doi:10.1073/pnas.2215328119

225. Rose AB, Elfersi T, Parra G, Korf I. Promoter-proximal introns in Arabidopsis thaliana are enriched in dispersed signals that elevate gene expression. *Plant Cell*. 2008;20(3):543-551. doi:10.1105/tpc.107.057190

226. Moabbi AM, Agarwal N, El Kaderi B, Ansari A. Role for gene looping in intron-mediated enhancement of transcription. *Proc Natl Acad Sci U S A*. 2012;109(22):8505-8510. doi:10.1073/pnas.1112400109

227. Huang AC, Jiang T, Liu YX, et al. A specialized metabolic network selectively modulates Arabidopsis root microbiota. *Science (80-)*. 2019;364(6440). doi:10.1126/science.aau6389

228. Stuart T, Eichten SR, Cahn J, Karpievitch Y V, Borevitz JO, Lister R. Population scale mapping of transposable element diversity reveals links to gene regulation and epigenomic variation. *Elife*. 2016;5. doi:10.7554/eLife.20777

229. Shaaban M, Palmer JM, El-Naggar WA, El-Sokkary MA, Habib E-SE, Keller NP. Involvement of transposon-like elements in penicillin gene cluster regulation. *Fungal Genet Biol*. 2010;47(5):423-432. doi:10.1016/j.fgb.2010.02.006

230. Bolle C, Koncz C, Chua NH. PAT1, a new member of the GRAS family, is involved in phytochrome A signal transduction. *Genes Dev*. 2000;14(10):1269-1278. http://www.ncbi.nlm.nih.gov/pubmed/10817761.

231. Pysh LD, Wysocka-Diller JW, Camilleri C, Bouchez D, Benfey PN. The GRAS gene family in Arabidopsis: sequence characterization and basic expression analysis of the SCARECROW-LIKE genes. *Plant J*. 1999;18(1):111-119. doi:10.1046/j.1365-313x.1999.00431.x

232. Lee M-H, Kim B, Song S-K, et al. Large-scale analysis of the GRAS gene family in Arabidopsis thaliana. *Plant Mol Biol*. 2008;67(6):659-670. doi:10.1007/s11103-008-9345-1

233. Hirsch S, Oldroyd GED. GRAS-domain transcription factors that regulate plant development. *Plant Signal Behav*. 2009;4(8):698-700. doi:10.4161/psb.4.8.9176

234. Bolle C. The role of GRAS proteins in plant signal transduction and development. *Planta*. 2004;218(5):683-692. doi:10.1007/s00425-004-1203-z

235. Feng S, Martinez C, Gusmaroli G, et al. Coordinated regulation of Arabidopsis thaliana development by light and gibberellins. *Nature*. 2008;451(7177):475-479. doi:10.1038/nature06448

236. de Lucas M, Davière J-M, Rodríguez-Falcón M, et al. A molecular framework for light and gibberellin control of cell elongation. *Nature*. 2008;451(7177):480-484. doi:10.1038/nature06520

237. Wang Z, Wong DCJ, Wang Y, et al. GRAS-domain transcription factor PAT1 regulates jasmonic acid biosynthesis in grape cold stress response. *Plant Physiol*. 2021;186(3):1660-1678. doi:10.1093/plphys/kiab142

238. Hu Y, Jiang Y, Han X, Wang H, Pan J, Yu D. Jasmonate regulates leaf senescence and tolerance to cold stress: crosstalk with other phytohormones. *J Exp Bot*. 2017;68(6):1361-1369. doi:10.1093/jxb/erx004

239. Gomi K. Jasmonic Acid: An Essential Plant Hormone. *Int J Mol Sci*. 2020;21(4). doi:10.3390/ijms21041261

240. Yoshida H, Hirano K, Sato T, et al. DELLA protein functions as a transcriptional activator through the DNA binding of the indeterminate domain family proteins. *Proc Natl Acad Sci U S A*. 2014;111(21):7861-7866. doi:10.1073/pnas.1321669111

241. Wasternack C, Hause B. Jasmonates: biosynthesis, perception, signal transduction and action in plant stress response, growth and development. An update to the 2007 review in Annals of Botany. *Ann Bot*. 2013;111(6):1021-1058. doi:10.1093/aob/mct067

242. Gasperini D, Chételat A, Acosta IF, et al. Multilayered Organization of Jasmonate Signalling in the Regulation of Root Growth. *PLoS Genet*. 2015;11(6):e1005300. doi:10.1371/journal.pgen.1005300

243. Zhou W, Lozano-Torres JL, Blilou I, et al. A Jasmonate Signaling Network Activates Root Stem Cells and Promotes Regeneration. *Cell*. 2019;177(4):942-956.e14. doi:10.1016/j.cell.2019.03.006

244. Heyman J, Cools T, Canher B, et al. The heterodimeric transcription factor complex ERF115–PAT1 grants regeneration competence. *Nat Plants*. 2016;2(11):16165. doi:10.1038/nplants.2016.165

245. Doench JG, Hartenian E, Graham DB, et al. Rational design of highly active sgRNAs for CRISPR-Cas9-mediated gene inactivation. *Nat Biotechnol*. 2014;32:1262-1267. doi:10.1038/nbt.3026

246. Doench JG, Fusi N, Sullender M, et al. Optimized sgRNA design to maximize activity and minimize off-target effects of CRISPR-Cas9. *Nat Biotechnol*. 2016;34:184-191. doi:10.1038/nbt.3437

247. Ren X, Yang Z, Xu J, et al. Enhanced specificity and efficiency of the CRISPR/Cas9 system with optimized sgRNA parameters in Drosophila. *Cell Rep*. 2014;9(3):1151-1162. doi:10.1016/j.celrep.2014.09.044

248. Chen B, Gilbert LA, Cimini BA, et al. Dynamic Imaging of Genomic Loci in Living Human Cells by an Optimized CRISPR/Cas System. *Cell*. 2013;155(7):1479-1491. doi:10.1016/j.cell.2013.12.001.Dynamic

249. Li X, Jiang DH, Yong K, Zhang DB. Varied transcriptional efficiencies of multiple Arabidopsis U6 small nuclear RNA genes. *J Integr Plant Biol*. 2007;49(2):222-229. doi:10.1111/j.1744-7909.2007.00393.x

250. Haberle V, Stark A. Eukaryotic core promoters and the functional basis of transcription initiation. *Nat Rev Mol Cell Biol*. 2018;19(10):621-637. doi:10.1038/s41580-018-0028-8

251. O'Connor CM, Perl A, Leonard D, Sangodkar J, Narla G. Therapeutic targeting of PP2A. *Int J Biochem Cell Biol*. 2018;96(June 2017):182-193. doi:10.1016/j.biocel.2017.10.008

252. Ordon J, Gantner J, Kemna J, et al. Generation of chromosomal deletions in dicotyledonous plants employing a user-friendly genome editing toolkit. *Plant J*. 2017. doi:10.1111/tpj.13319

253. Durr J, Papareddy R, Nakajima K, Gutierrez-Marcos J. Highly efficient heritable targeted deletions of gene clusters and non-coding regulatory regions in Arabidopsis using CRISPR/Cas9. *Sci Rep*. 2018;8(1):1-11. doi:10.1038/s41598-018-22667-1

254. Wu R, Lucke M, Jang Y, et al. An efficient CRISPR vector toolbox for engineering large deletions in Arabidopsis thaliana. *Plant Methods*. 2018;14(1):65. doi:10.1186/s13007-018-0330-7

255. Minard ME, Jain AK, Barton MC. Analysis of epigenetic alterations to chromatin during development. *Genesis*. 2009;47(8):559-572. doi:10.1002/dvg.20534

256. Jambhekar A, Dhall A, Shi Y. Roles and regulation of histone methylation in animal development. *Nat Rev Mol Cell Biol*. 2019;20(10):625-641. doi:10.1038/s41580-019-0151-1

257. Liu X, Wang C, Liu W, et al. Distinct features of H3K4me3 and H3K27me3 chromatin

domains in pre-implantation embryos. *Nature*. 2016;537(7621):558-562. doi:10.1038/nature19362

258. Bannister AJ, Kouzarides T. Regulation of chromatin by histone modifications. *Cell Res*. 2011;21(3):381-395. doi:10.1038/cr.2011.22

259. Shi Y, Lan F, Matson C, et al. Histone demethylation mediated by the nuclear amine oxidase homolog LSD1. *Cell*. 2004;119(7):941-953. doi:10.1016/j.cell.2004.12.012

260. Klose RJ, Kallin EM, Zhang Y. JmjC-domain-containing proteins and histone demethylation. *Nat Rev Genet*. 2006;7(9):715-727. doi:10.1038/nrg1945

261. Aravind L, Iyer LM. The SWIRM domain: a conserved module found in chromosomal proteins points to novel chromatin-modifying activities. *Genome Biol*. 2002;3(8):RESEARCH0039. doi:10.1186/gb-2002-3-8-research0039

262. Yu C-W, Chang K-Y, Wu K. Genome-Wide Analysis of Gene Regulatory Networks of the FVE-HDA6-FLD Complex in Arabidopsis. *Front Plant Sci*. 2016;7:555. doi:10.3389/fpls.2016.00555

263. Krichevsky A, Zaltsman A, Kozlovsky S V, Tian G-W, Citovsky V. Regulation of root elongation by histone acetylation in Arabidopsis. *J Mol Biol*. 2009;385(1):45-50. doi:10.1016/j.jmb.2008.09.040

264. Forneris F, Binda C, Vanoni MA, Battaglioli E, Mattevi A. Human histone demethylase LSD1 reads the histone code. *J Biol Chem*. 2005;280(50):41360-41365. doi:10.1074/jbc.M509549200

265. Krichevsky A, Gutgarts H, Kozlovsky S V, et al. C2H2 zinc finger-SET histone methyltransferase is a plant-specific chromatin modifier. *Dev Biol*. 2007;303(1):259-269. doi:10.1016/j.ydbio.2006.11.012

266. Zhao M, Yang S, Liu X, Wu K. Arabidopsis histone demethylases LDL1 and LDL2 control primary seed dormancy by regulating DELAY OF GERMINATION 1 and ABA signaling-related genes. *Front Plant Sci*. 2015;6:159. doi:10.3389/fpls.2015.00159

267. Noh SW, Seo R-R, Park HJ, Jung HW. Two Arabidopsis Homologs of Human Lysine-Specific Demethylase Function in Epigenetic Regulation of Plant Defense Responses. *Front Plant Sci*. 2021;12(June). doi:10.3389/fpls.2021.688003

268. Ishihara H, Sugimoto K, Tarr PT, et al. Primed histone demethylation regulates shoot regenerative competency. *Nat Commun*. 2019;10(1):1786. doi:10.1038/s41467-019-09386-5

269. Singh V, Roy S, Singh D, Nandi AK. Arabidopsis FLOWERING LOCUS D influences systemic-acquired-resistance-induced expression and histone modifications of WRKY genes. *J Biosci*. 2014;39(1):119-126. doi:10.1007/s12038-013-9407-7

270. Gan E-S, Xu Y, Ito T. Dynamics of H3K27me3 methylation and demethylation in plant development. *Plant Signal Behav*. 2015;10(9):e1027851. doi:10.1080/15592324.2015.1027851

271. Lu F, Li G, Cui X, Liu C, Wang XJ, Cao X. Comparative analysis of JmjC domain-containing proteins reveals the potential histone demethylases in arabidopsis and rice. *J Integr Plant Biol*. 2008;50(7):886-896. doi:10.1111/j.1744-7909.2008.00692.x

272. Liu C, Lu F, Cui X, Cao X. Histone Methylation in Higher Plants. *Annu Rev Plant Biol*. 2010;61(1):395-420. doi:10.1146/annurev.arplant.043008.091939

273. Lafos M, Kroll P, Hohenstatt ML, Thorpe FL, Clarenz O, Schubert D. Dynamic Regulation of H3K27 Trimethylation during Arabidopsis Differentiation. Kakutani T, ed. *PLoS Genet*. 2011;7(4):e1002040. doi:10.1371/journal.pgen.1002040

274. Searle IR, Pontes O, Melnyk CW, Smith LM, Baulcombe DC. JMJ14, a JmjC domain protein, is required for RNA silencing and cell-to-cell movement of an RNA silencing signal in Arabidopsis. *Genes Dev*. 2010;24(10):986-991. doi:10.1101/gad.579910

275. Lu F, Cui X, Zhang S, Liu C, Cao X. JMJ14 is an H3K4 demethylase regulating flowering time in Arabidopsis. *Cell Res*. 2010;20(3):387-390. doi:10.1038/cr.2010.27

276. Jeong J-H, Song H-R, Ko J-H, et al. Repression of FLOWERING LOCUS T chromatin by functionally redundant histone H3 lysine 4 demethylases in Arabidopsis. *PLoS One*. 2009;4(11):e8033. doi:10.1371/journal.pone.0008033

277. Li D, Liu R, Singh D, Yuan X, Kachroo P, Raina R. JMJ14 encoded H3K4 demethylase modulates immune responses by regulating defence gene expression and pipecolic acid levels. *New Phytol*. 2020;225(5):2108-2121. doi:10.1111/nph.16270

278. Le Masson I, Jauvion V, Bouteiller N, Rivard M, Elmayan T, Vaucheret H. Mutations in the Arabidopsis H3K4me2/3 demethylase JMJ14 suppress posttranscriptional gene silencing by decreasing transgene transcription. *Plant Cell*. 2012;24(9):3603-3612. doi:10.1105/tpc.112.103119

279. Yang H, Mo H, Fan D, Cao Y, Cui S, Ma L. Overexpression of a histone H3K4 demethylase, JMJ15, accelerates flowering time in Arabidopsis. *Plant Cell Rep*. 2012;31(7):1297-1308. doi:10.1007/s00299-012-1249-5

280. Deleris A, Greenberg MVC, Ausin I, et al. Involvement of a Jumonji-C domain-containing histone demethylase in DRM2-mediated maintenance of DNA methylation. *EMBO Rep*. 2010;11(12):950-955. doi:10.1038/embor.2010.158

281. Shen Y, Conde E Silva N, Audonnet L, Servet C, Wei W, Zhout DX. Over-expression of histone H3K4 demethylase gene JMJ15 enhances salt tolerance in Arabidopsis. *Front Plant Sci*. 2014;5(JUN):1-10. doi:10.3389/fpls.2014.00290

282. Lu F, Li G, Cui X, Liu C, Wang X-J, Cao X. Comparative analysis of JmjC domain-containing proteins reveals the potential histone demethylases in Arabidopsis and rice. *J Integr Plant Biol*. 2008;50(7):886-896. doi:10.1111/j.1744-7909.2008.00692.x

283. Ay N, Irmler K, Fischer A, Uhlemann R, Reuter G, Humbeck K. Epigenetic programming via histone methylation at WRKY53 controls leaf senescence in Arabidopsis thaliana. *Plant J*. 2009;58(2):333-346. doi:10.1111/j.1365-313X.2008.03782.x

284. Brusslan JA, Rus Alvarez-Canterbury AM, Nair NU, Rice JC, Hitchler MJ, Pellegrini M. Genome-wide evaluation of histone methylation changes associated with leaf senescence in Arabidopsis. *PLoS One*. 2012;7(3):e33151. doi:10.1371/journal.pone.0033151

285. Liu P, Zhang S, Zhou B, et al. The histone H3K4 demethylase JMJ16 represses leaf senescence in arabidopsis. *Plant Cell*. 2019;31(2):430-443. doi:10.1105/tpc.18.00693

286. Wang TJ, Huang S, Zhang A, et al. JMJ17–WRKY40 and HY5–ABI5 modules regulate the expression of ABA-responsive genes in Arabidopsis. *New Phytol*. 2021;230(2):567-584. doi:10.1111/nph.17177

287. Huang S, Zhang A, Jin JB, et al. Arabidopsis histone H3K4 demethylase JMJ17 functions in dehydration stress response. *New Phytol*. 2019;223(3):1372-1387. doi:10.1111/nph.15874

288. Chen H, Lai Z, Shi J, Xiao Y, Chen Z, Xu X. Roles of arabidopsis WRKY18, WRKY40 and WRKY60 transcription factors in plant responses to abscisic acid and abiotic stress. *BMC Plant Biol*. 2010;10(1):281. doi:10.1186/1471-2229-10-281

289. Dutta A, Choudhary P, Caruana J, Raina R. JMJ27, an Arabidopsis H3K9 histone demethylase, modulates defense against Pseudomonas syringae and flowering time. *Plant J*. 2017;91(6):1015-1028. doi:10.1111/tpj.13623

290. Gan ES, Xu Y, Wong JY, et al. Jumonji demethylases moderate precocious flowering at elevated temperature via regulation of FLC in Arabidopsis. *Nat Commun*. 2014;5:1-13.

doi:10.1038/ncomms6098

291. Jiang D, Wang Y, Wang Y, He Y. Repression of FLOWERING LOCUS C and FLOWERING LOCUS T by the Arabidopsis Polycomb Repressive Complex 2 Components. Dilkes BP, ed. *PLoS One*. 2008;3(10):e3404. doi:10.1371/journal.pone.0003404

292. Angel A, Song J, Dean C, Howard M. A Polycomb-based switch underlying quantitative epigenetic memory. *Nature*. 2011;476(7358):105-108. doi:10.1038/nature10241

293. Law JA, Jacobsen SE. Establishing, maintaining and modifying DNA methylation patterns in plants and animals. *Nat Rev Genet*. 2010;11(3):204-220. doi:10.1038/nrg2719

294. Wang Y, Jia S. Degrees make all the difference: the multifunctionality of histone H4 lysine 20 methylation. *Epigenetics*. 2009;4(5):273-276. doi:10.4161/epi.4.5.9212

295. Springer NM, Napoli CA, Selinger DA, et al. Comparative analysis of SET domain proteins in maize and Arabidopsis reveals multiple duplications preceding the divergence of monocots and dicots. *Plant Physiol*. 2003;132(2):907-925. doi:10.1104/pp.102.013722

296. Zhou H, Liu Y, Liang Y, et al. The function of histone lysine methylation related SET domain group proteins in plants. *Protein Sci*. 2020;29(5):1120-1137. doi:10.1002/pro.3849

297. Xiao J, Lee US, Wagner D. Tug of war: adding and removing histone lysine methylation in Arabidopsis. *Curr Opin Plant Biol*. 2016;34:41-53. doi:10.1016/j.pbi.2016.08.002

298. Fletcher JC. State of the Art: trxG Factor Regulation of Post-embryonic Plant Development. *Front Plant Sci*. 2017;8:1925. doi:10.3389/fpls.2017.01925

299. Thorstensen T, Grini PE, Aalen RB. SET domain proteins in plant development. *Biochim Biophys Acta*. 2011;1809(8):407-420. doi:10.1016/j.bbagrm.2011.05.008

300. Tripoulas N, LaJeunesse D, Gildea J, Shearn A. The Drosophila ash1 gene product, which is localized at specific sites on polytene chromosomes, contains a SET domain and a PHD finger. *Genetics*. 1996;143(2):913-928. doi:10.1093/genetics/143.2.913

301. Jacob Y, Feng S, LeBlanc CA, et al. ATXR5 and ATXR6 are H3K27 monomethyltransferases required for chromatin structure and gene silencing. *Nat Struct Mol Biol*. 2009;16(7):763-768. doi:10.1038/nsmb.1611

302. Cheng X, Collins RE, Zhang X. Structural and sequence motifs of protein (histone) methylation enzymes. *Annu Rev Biophys Biomol Struct*. 2005;34:267-294.

doi:10.1146/annurev.biophys.34.040204.144452

303. Alvarez-Venegas R, Avramova Z. Evolution of the PWWP-domain encoding genes in the plant and animal lineages. *BMC Evol Biol*. 2012;12:101. doi:10.1186/1471-2148-12-101

304. Pien S, Fleury D, Mylne JS, et al. ARABIDOPSIS TRITHORAX1 dynamically regulates FLOWERING LOCUS C activation via histone 3 lysine 4 trimethylation. *Plant Cell*. 2008;20(3):580-588. doi:10.1105/tpc.108.058172

305. Chen LQ, Luo JH, Cui ZH, et al. ATX3, ATX4, and ATX5 encode putative H3K4 methyltransferases and are critical for plant development. *Plant Physiol*. 2017;174(3):1795-1806. doi:10.1104/pp.16.01944

306. Ding Y, Avramova Z, Fromm M. The Arabidopsis trithorax-like factor ATX1 functions in dehydration stress responses via ABA-dependent and ABA-independent pathways. *Plant J*. 2011;66(5):735-744. doi:10.1111/j.1365-313X.2011.04534.x

307. Shang JY, Lu YJ, Cai XW, et al. COMPASS functions as a module of the INO80 chromatin remodeling complex to mediate histone H3K4 methylation in Arabidopsis. *Plant Cell*. 2021;33(10):3250-3271. doi:10.1093/plcell/koab187

308. Jiang D, Kong NC, Gu X, Li Z, He Y. Arabidopsis COMPASS-like complexes mediate histone H3 lysine-4 trimethylation to control floral transition and plant development. *PLoS Genet*. 2011;7(3):e1001330. doi:10.1371/journal.pgen.1001330

309. Berr A, McCallum EJ, Ménard R, et al. Arabidopsis SET DOMAIN GROUP2 is required for H3K4 trimethylation and is crucial for both sporophyte and gametophyte development. *Plant Cell*. 2010;22(10):3232-3248. doi:10.1105/tpc.110.079962

310. Cartagena JA, Matsunaga S, Seki M, et al. The Arabidopsis SDG4 contributes to the regulation of pollen tube growth by methylation of histone H3 lysines 4 and 36 in mature pollen. *Dev Biol*. 2008;315(2):355-368. doi:10.1016/j.ydbio.2007.12.016

311. Baumbusch LO, Thorstensen T, Krauss V, et al. The Arabidopsis thaliana genome contains at least 29 active genes encoding SET domain proteins that can be assigned to four evolutionarily conserved classes. *Nucleic Acids Res*. 2001;29(21):4319-4333. doi:10.1093/nar/29.21.4319

312. Jenuwein T. The epigenetic magic of histone lysine methylation. *FEBS J*. 2006;273(14):3121-3135. doi:10.1111/j.1742-4658.2006.05343.x

313. Shafiq S, Berr A, Shen W. Combinatorial functions of diverse histone methylations in A rabidopsis thaliana flowering time regulation. *New Phytol*. 2014;201(1):312-322.

doi:10.1111/nph.12493

314. Hu G, Cui K, Northrup D, et al. H2A.Z Facilitates Access of Active and Repressive Complexes to Chromatin in Embryonic Stem Cell Self-Renewal and Differentiation. *Cell Stem Cell*. 2013;12(2):180-192. doi:10.1016/j.stem.2012.11.003

315. Orsi GA, Couble P, Loppin B. Epigenetic and replacement roles of histone variant H3.3 in reproduction and development. *Int J Dev Biol*. 2009;53(2-3):231-243. doi:10.1387/ijdb.082653go

316. Santos-Rosa H, Schneider R, Bannister AJ, et al. Active genes are tri-methylated at K4 of histone H3. *Nature*. 2002;419(6905):407-411. doi:10.1038/nature01080

317. Regha K, Sloane MA, Huang R, et al. Active and Repressive Chromatin Are Interspersed without Spreading in an Imprinted Gene Cluster in the Mammalian Genome. *Mol Cell*. 2007;27(3):353-366. doi:10.1016/j.molcel.2007.06.024

318. Beacon TH, Delcuve GP, López C, et al. The dynamic broad epigenetic (H3K4me3, H3K27ac) domain as a mark of essential genes. *Clin Epigenetics*. 2021;13(1):1-17. doi:10.1186/s13148-021-01126-1

319. Zhang C, Du X, Tang K, et al. Arabidopsis AGDP1 links H3K9me2 to DNA methylation in heterochromatin. *Nat Commun*. 2018;9(1):4547. doi:10.1038/s41467-018-06965-w

320. Ferrari KJ, Scelfo A, Jammula SG, et al. Polycomb-Dependent H3K27me1 and H3K27me2 Regulate Active Transcription and Enhancer Fidelity. *Mol Cell*. 2014;53(1):49-62. doi:10.1016/j.molcel.2013.10.030

321. Ma L, Yuan L, An J, Barton MC, Zhang Q, Liu Z. Histone H3 lysine 23 acetylation is associated with oncogene TRIM24 expression and a poor prognosis in breast cancer. *Tumor Biol*. 2016;37(11):14803-14812. doi:10.1007/s13277-016-5344-z

322. Yang H, Howard M, Dean C. Antagonistic roles for H3K36me3 and H3K27me3 in the cold-induced epigenetic switch at Arabidopsis FLC. *Curr Biol*. 2014;24(15):1793-1797. doi:10.1016/j.cub.2014.06.047

323. Lindeman LC, Winata CL, Håvard A, Sinnakaruppan M, Aleström P, Collas P. Chromatin states of developmentally-regulated genes revealed by DNA and histone methylation patterns in zebrafish embryos. *Int J Dev Biol*. 2010;54(5):803-813. doi:10.1387/ijdb.103081ll

324. Wendte JM, Schmitz RJ. Specifications of Targeting Heterochromatin Modifications in Plants. *Mol Plant*. 2018;11(3):381-387. doi:10.1016/j.molp.2017.10.002

325. Lu F, Cui X, Zhang S, Liu C, Cao X. JMJ14 is an H3K4 demethylase regulating flowering time in Arabidopsis. *Cell Res*. 2010;20(3):387-390. doi:10.1038/cr.2010.27

326. Juan AH, Wang S, Ko KD, et al. Roles of H3K27me2 and H3K27me3 Examined during Fate Specification of Embryonic Stem Cells. *Cell Rep*. 2016;17(5):1369-1382. doi:10.1016/j.celrep.2016.09.087

327. Yan H, Liu Y, Zhang K, Song J, Xu W, Su Z. Chromatin State-Based Analysis of Epigenetic H3K4me3 Marks of Arabidopsis in Response to Dark Stress. *Front Genet*. 2019;10:306. doi:10.3389/fgene.2019.00306

328. Yelagandula R, Stroud H, Holec S, et al. The histone variant H2A.W defines heterochromatin and promotes chromatin condensation in Arabidopsis. *Cell*. 2014;158(1):98-109. doi:10.1016/j.cell.2014.06.006

329. Gómez-Zambrano Á, Merini W, Calonje M. The repressive role of Arabidopsis H2A.Z in transcriptional regulation depends on AtBMI1 activity. *Nat Commun*. 2019;10(1):2828. doi:10.1038/s41467-019-10773-1

330. Lorković ZJ, Berger F. Heterochromatin and DNA damage repair: Use different histone variants and relax. *Nucleus*. 2017;8(6):583-588. doi:10.1080/19491034.2017.1384893

331. Bai Y, Fernández-Calvo P, Ritter A, et al. Modulation of Arabidopsis root growth by specialized triterpenes. *New Phytol*. 2021;230(1):228-243. doi:10.1111/nph.17144

332. Gacek-Matthews A, Berger H, Sasaki T, et al. KdmB, a Jumonji Histone H3 Demethylase, Regulates Genome-Wide H3K4 Trimethylation and Is Required for Normal Induction of Secondary Metabolism in Aspergillus nidulans. Stukenbrock EH, ed. *PLOS Genet*. 2016;12(8):e1006222. doi:10.1371/journal.pgen.1006222

333. Cao F, Fang Y, Tan HK, et al. Super-Enhancers and Broad H3K4me3 Domains Form Complex Gene Regulatory Circuits Involving Chromatin Interactions. *Sci Rep*. 2017;7(1):2186. doi:10.1038/s41598-017-02257-3

334. Yang T, Wang D, Tian G, et al. Chromatin remodeling complexes regulate genome architecture in Arabidopsis. *Plant Cell*. 2022;34(7):2638-2651. doi:10.1093/plcell/koac117

335. Hou C, Li L, Qin ZS, Corces VG. Gene density, transcription, and insulators contribute to the partition of the Drosophila genome into physical domains. *Mol Cell*. 2012;48(3):471-484. doi:10.1016/j.molcel.2012.08.031

336. Klocko AD, Ormsby T, Galazka JM, et al. Normal chromosome conformation depends

on subtelomeric facultative heterochromatin in Neurospora crassa. *Proc Natl Acad Sci U S A*. 2016;113(52):15048-15053. doi:10.1073/pnas.1615546113

337. Zhao K, Kong D, Jin B, Smolke CD, Rhee SY. A novel bivalent chromatin associates with rapid induction of camalexin biosynthesis genes in response to a pathogen signal in Arabidopsis. *Elife*. 2021;10(3):1403-1414. doi:10.7554/eLife.69508

338. Kan Y, La H, Li X, Li S, Zhu X, Shi X. A histone acetyltransferase regulates active DNA demethylation in Arabidopsis. *Science (80-)*. 2012;336(6087):1445-1448. doi:10.1126/science.1219416.A

339. Wang L, Wang C, Liu X, et al. Peroxisomal β-oxidation regulates histone acetylation and DNA methylation in Arabidopsis. *Proc Natl Acad Sci U S A*. 2019;116(21):10576-10585. doi:10.1073/pnas.1904143116

340. Bharadwaj R, Kumar SR, Sharma A, Sathishkumar R. Plant Metabolic Gene Clusters: Evolution, Organization, and Their Applications in Synthetic Biology. *Front Plant Sci*. 2021;12. doi:10.3389/fpls.2021.697318

341. Zhao K, Kong D, Jin B, Smolke CD, Rhee SY. A novel form of bivalent chromatin associates with rapid induction of camalexin biosynthesis genes in response to a pathogen signal in arabidopsis. *Elife*. 2021;10:1-15. doi:10.7554/eLife.69508

342. Yang H, Howard M, Dean C. Antagonistic Roles for H3K36me3 and H3K27me3 in the Cold-Induced Epigenetic Switch at Arabidopsis FLC. *Curr Biol*. 2014;24(15):1793-1797. doi:10.1016/j.cub.2014.06.047

343. Talbert PB, Henikoff S. Histone variants on the move: Substrates for chromatin dynamics. *Nat Rev Mol Cell Biol*. 2017;18(2):115-126. doi:10.1038/nrm.2016.148

344. Filipescu D, Müller S, Almouzni G. Histone H3 variants and their chaperones during development and disease: contributing to epigenetic control. *Annu Rev Cell Dev Biol*. 2014;30:615-646. doi:10.1146/annurev-cellbio-100913-013311

345. Talbert PB, Ahmad K, Almouzni G, et al. A unified phylogeny-based nomenclature for histone variants. *Epigenetics and Chromatin*. 2012;5(1):1-19. doi:10.1186/1756-8935-5-7

346. Ahmad K, Henikoff S. The histone variant H3.3 marks active chromatin by replication-independent nucleosome assembly. *Mol Cell*. 2002;9(6):1191-1200. doi:10.1016/s1097-2765(02)00542-7

347. Wollmann H, Stroud H, Yelagandula R, et al. The histone H3 variant H3.3 regulates gene

body DNA methylation in Arabidopsis thaliana. *Genome Biol*. 2017;18(1):1-10. doi:10.1186/s13059-017-1221-3

348. Naumann K, Fischer A, Hofmann I, et al. Pivotal role of AtSUVH2 in heterochromatic histone methylation and gene silencing in Arabidopsis. *EMBO J*. 2005;24(7):1418-1429. doi:10.1038/sj.emboj.7600604

349. Chen L-T, Luo M, Wang Y-Y, Wu K. Involvement of Arabidopsis histone deacetylase HDA6 in ABA and salt stress response. *J Exp Bot*. 2010;61(12):3345-3353. doi:10.1093/jxb/erq154

350. Hainan Zhao, Mingyu Yang, Jade Bishop, Yuhan Teng, Yingxue Cao, Brandon D. Beall, Shuanglin Li, Tongxin Liu, Qingxi Fang, Chao Fang, Haoyang Xin, Hans-Wilhelm Nützmann, Anne Osbourn, Fanli Meng and JJ. Identification and functional validation of super-enhancers in Arabidopsis thaliana. *PNAS*.

351. Le Masson I, Jauvion V, Bouteiller N, Rivard M, Elmayan T, Vaucheret H. Mutations in the Arabidopsis H3K4me2/3 demethylase jmj14 suppress posttranscriptional gene silencing by decreasing transgene transcription. *Plant Cell*. 2012;24(9):3603-3612. doi:10.1105/tpc.112.103119

352. Tsukada YI, Fang J, Erdjument-Bromage H, et al. Histone demethylation by a family of JmjC domain-containing proteins. *Nature*. 2006;439(7078):811-816. doi:10.1038/nature04433

353. Andreuzza S, Nishal B, Singh A, Siddiqi I. The Chromatin Protein DUET/MMD1 Controls Expression of the Meiotic Gene TDM1 during Male Meiosis in Arabidopsis. *PLoS Genet*. 2015;11(9):1-22. doi:10.1371/journal.pgen.1005396

354. Wang J, Yu C, Zhang S, et al. Cell-type-dependent histone demethylase specificity promotes meiotic chromosome condensation in Arabidopsis. *Nat Plants*. 2020;6(7):823-837. doi:10.1038/s41477-020-0697-0

355. Hung FY, Chen FF, Li C, et al. The LDL1/2-HDA6 histone modification complex interacts with TOC1 and regulates the core circadian clock components in Arabidopsis. *Front Plant Sci*. 2019;10(February):1-10. doi:10.3389/fpls.2019.00233

356. Itkin M, Rogachev I, Alkan N, et al. GLYCOALKALOID METABOLISM1 Is Required for Steroidal Alkaloid Glycosylation and Prevention of Phytotoxicity in Tomato. *Plant Cell*. 2011;23(12):4507-4525. doi:10.1105/tpc.111.088732

357. Fletcher JC, Richter G. Human fetal gene therapy: moral and ethical questions. *Hum*

Gene Ther. 1996;7(13):1605-1614. doi:10.1089/hum.1996.7.13-1605

358. Grisch-Chan HM, Schwank G, Harding CO, Thöny B. State-of-the-Art 2019 on Gene Therapy for Phenylketonuria. *Hum Gene Ther*. 2019;30(10):1274-1283. doi:10.1089/hum.2019.111

359. Li H, Yang Y, Hong W, Huang M, Wu M, Zhao X. Applications of genome editing technology in the targeted therapy of human diseases: mechanisms, advances and prospects. *Signal Transduct Target Ther*. 2020;5(1):1. doi:10.1038/s41392-019-0089-y

360. Karavolias NG, Horner W, Abugu MN, Evanega SN. Application of Gene Editing for Climate Change in Agriculture. *Front Sustain Food Syst*. 2021;5. doi:10.3389/fsufs.2021.685801

361. Chow C-N, Tseng K-C, Hou P-F, Wu N-Y, Lee T-Y, Chang W-C. Mysteries of gene regulation: Promoters are not the sole triggers of gene expression. *Comput Struct Biotechnol J*. 2022;20:4910-4920. doi:10.1016/j.csbj.2022.08.058

362. Kelly TK, De Carvalho DD, Jones PA. Epigenetic modifications as therapeutic targets. *Nat Biotechnol*. 2010;28(10):1069-1078. doi:10.1038/nbt.1678

363. Manyi-Loh C, Mamphweli S, Meyer E, Okoh A. Antibiotic Use in Agriculture and Its Consequential Resistance in Environmental Sources: Potential Public Health Implications. *Molecules*. 2018;23(4):795. doi:10.3390/molecules23040795

364. Tovell H, Testa A, Maniaci C, et al. Rapid and Reversible Knockdown of Endogenously Tagged Endosomal Proteins via an Optimized HaloPROTAC Degrader. *ACS Chem Biol*. 2019;14(5):882-892. doi:10.1021/acschembio.8b01016

7 Supplementary data

7.1 Sequences for CRISPR/Cas9 construct assembly

>DD45 promoter (1020bp)

AAATGTTCCTCGCTGACGTAAGAAGACATTAGTAATGGTTATAATATATAGCTTTCTATGAATGTATG
GTGAGAAAATGTCTGTTCACTGATTTTGAGTTTGGAATAAAAGCATTTGCGTTTGGTTTATCATTGCG
TTTATACAAGGACAGAGATCCACTGAGCTGGAATAGCTTAAAACCATTATCAGAACAAAATAAACCA
TTTTTTGTTAAGAATCAGAGCATAGTAAACAACAGAAACAACCTAAGAGAGGTAACTTGTCCAAGAA
GATAGCTAATTATATCTATTTTATAAAAGTTATCATAGTTTGTAAGTCACAAAAGATGCAAATAACAG
AGAAACTAGGAGACTTGAGAATATACATTCTTGTATATTTGTATTCGAGATTGTGAAAATTTGACCAT
AAGTTTAAATTCTTAAAAAGATATATCTGATCTAGATGATGGTTATAGACTGTAATTTTACCACATGT
TTAATGATGGATAGTGACACACATGACACATCGACAACACTATAGCATCTTATTTAGATTACAACATG
AAATTTTTCTGTAATACATGTCTTTGTACATAATTTAAAAGTAATTCCTAAGAAATATATTTATACAAG
GAGTTTAAAGAAAACATAGCATAAAGTTCAATGAGTAGTAAAAACCATATACAGTATATAGCATAAA
GTTCAATGAGTTTATTACAAAAGCATTGGTTCACTTTCTGTAACACGACGTTAAACCTTCGTCTCCAAT
AGGAGCGCTACTGATTCAACATGCCAATATATACTAAATACGTTTCTACAGTCAAATGCTTTAACGTT
TCATGATTAAGTGACTATTTACCGTCAATCCTTTCCCATTCCTCCCACTAATCCAACTTTTTAATTACTC
TTAAATCACCACTAAGCTTCGAATCCATCCAAAACCACAATATAAAAACAGAACTCTCGTAACTCAAT
CATCGCAAAACAAAACAAAACAAAACAAAAACCCCAAAAAGAAAGAATAATGGCTTCTAACACAAG
T

>MRN LA (900bp)

GGACAACATAGAAATCATTTATGTTGTTGATTAAGATTTAATTTGTATGTTATTCAACAATAATAGAT

CTAGTTGACTGGCATTTTGTTTATATGCGATATGTCCTTTCAAAGTAATCAGATTCAGCTGCACAAAA

TCGGGAGAATTCTCATTATTGTAGATACGGTCCATTACTTAATAGAATTCTACGTAACCAAGTCGTTA

CGTATCAATTTATTATTACGGAAGCATGTGGAAAGATAATGTATAGAAAATAGTTTATGTGTAAGTC

CAAAAAAATGTGGATAATATAATATTATGCAATCGATGTTGGACTACTACAAATGATGAAGACCGTT

GATTTATATATTTATTTAATATAAAAATGTTAAACTTGTTGATTAACAAAAAGATAAATGACAATTTGT

GCAAAGTATTGGAAGTTGCAAGTAAACTCTCTGAAGAGTGTAGATGATCAAAACAAAACAAAACAA

TTAGGTTTAGACAAAAAAAACAAAAAAAAAAAAAAAAACAAAAAAAAAAACAATTAGGTTGATTTGCTA

CTTGGGAAATTCTTTTATTCTTCAAAACAATTTTATCCTGCTTCAAAATCTTAAAGTTTTACTTTTATGC

ATTTTCTTAGATACCCCAGTTTAATCATTTATTTTACATTCAATTTAGCACCGTTTTTGACCTACATTGT

AGCGTTCATTGCAAAATTTCCAACATGTCCATTATCTGGTCCTAGGAAAACTGACATAAGGCATTTAA

AAGGTATACTTAGCTATAATTTTGGGGCTATCATCAGAGATTATTTCGCTATATAGTCCAAAAATAGC

CCAACATACATTTTTGTAAACGACTAAAACAATGTGGCTATTGATTAACACATCGATTAAATCTTATG

TGCATACGACCCCAATC

>MRN RA (950bp)

GAATGGTATAATGATGTAGCACTTTAGGATCAAAACTATTAGGAACAATGCTTAATTCACTAGCAAT

GTATATATGCCATCATTTAATCTAAGCTAGAGAATATTGGTTTTGTTTTTTATAAAATTTAAGTAGAA

GACTAACAAAATTCAAAAACAAAGTGGGAAATCTACTAAACAGTTTTAAGCAATAGATAAAAGAATA

ATTGCTCAGAATGCCCCTACCTTTTTCAATTCTACCCTTTCCTATTTGCCATTTTTATTTTTATACCCTCT

TTAAATTTTTAATGACCATTTTACCCCTATTTGTAAATTGTTTTAGGTTAACAAAATGAAATATTAATA

TTTTTCACCTAAATTATTCCAAAATAAATTTTCCGCCAAAAATTTTCGAAACTTTTTTTGGGAAATTTG

CAAGAATAACTAACAAAAACTTATAAATTAAGAAAGTAACATAAGACAAAAAAAATAGCTATACTCT

CTATATTATATATAATAATGGCATATTTACCCCTCTCTCTCCACCTCATTCTAAGTCTCTCCTTCTCTCCT

TTCTCTCTCATCTCTTTTTCTTTTTTCTCTCGATCTCTTCTCTTCTTTTTTTCTGTATCACATAACAAACAA

ACGAGAAAGACAAAGTAGATTTTTCAGATACCACTTATGTTATATGGTTGCTTGGGTTACCAAGGTT

GTATTTTAGGGTATAGTTTTTCTATACATCCAAAATAATATAGTACAACATTTTAATTCTATCCAAAAT

AGTACATAATTTTACTTACAAACACTATTTACTATTTTATAAATTCAAATCCGATTTTAAAAATTCAAA

CTATTTTGTATGTAGAAAAAACATAGTTTTTTTTGTAATAAATTTTATAATTGGAATCTATTTTTTTGAA

AAAGATATATCGCAAAAAAAAAACAAATGGGTAAAAACTAAACTAAGTGTTACT

>*THAS* LA (913bp)

GTCTGACCCGTTGGGATGTTAATTTTCATCCACCATGAAATAGAATTTCTACAATTTTTTAATTTTATC

CATAGTAAGGCCATCCTTTCCGACCCATCTCTTATAAAATCAAAATTTTGGTTTTCTTCTACTCAATGA

CTTTTCGCAACTCAATAAAATCTATGAGCAAAACCAATCATTTCCAGATATTCTGTTTTAGAATAAGA

AAACTAATTATGCAAGCCCTCATTTCCTAAGGTCCGATAAAGCCTGCATCTGACTATCTGTTGGTTCG

TATTACCTCTCTATGATAACTGATTTCCATAATATTGAAATTCTAACATACCATTATGTCGAATTATAA

AATTAAGGGTTACAAAAGAAGCTGCCAATCTTAAGTTACCTCCGGTTTTTTCATGATTTCCAGTAAAC

TCATTAAAATCTCCAATCATTAGCTATGGTTTTTTCCGATTAGAACCTAGTAGTATTATTTTTTACCATA

CAATTTCGCGTTTTGTACCAATAAGATTGCCATAAATAAACGAAATAAACACCATTTGTTTTCTGAGT

AATGTTTTAACGTCAATGACCCGATTTCTTTCATAGAGAATCTCTCTATTTTTATATCAGTATTACAAA

ACTAAACTAAACTTTCACTAGCTTTGAGAGGATCAGCTGATTACAAATTATTATTACCAAATTAAGAC

TGAAATATTTGCTAATAAGAATAACATCTTTTAGTTTCTGATTTATTGTTCAAAAAATAAAATAAAAA

GTTTCTGATAAAACTAAAATATATGACTTGTACGTTTTCCATAAATCCATTATGTAATTTTCTATCATA

TCGGCCCCTACTCCCTAATCAAAATAAATCTTTCTAAAAAAAAGAAAAAAAAAAAAAAAATCATACA
ATACCGTTGTAACTAACGAGTAA

>*THAS* RA (933bp)

CGACGGCAATAACCTAAAACAATAACCCATTATTCTAAATAAATAGATAGGCATATAAAAACATTTTT
TGCTAAACACGTAACACTTTTTTTAATTTGGTTAGCACCTACCTCTTTTTTAATGTTAGTTTTATTTAAT
CAGTATTGATTACTGTTAAATGAGTCAAAGCTCATCGAAAGAGAATTTTTTTTTAAAAATTATTTCTTA
ATATCTTATCATATCATATTGGTATAAGATCAGTCATTGCACGTTATAGATTATAAAATACAACTTTAT
ATTTTGAATAATTTTCATATCGTAATCAGATCATATTACGATCTTCTTATATGATTCCAACGTTGATTTT
CCCATTATAAAATGCAACGTGGTTTTTTCTTATAGTTTTCAATGTGATATATTATAAAATGCAACCCTT
TTTACTTATTAACAGTTGACACTTAACGGGGGGTTAAATACTATTTGTAATATCAATATCAATTATAC
GATTAATTAACATTACCGAATAAAATGTCAACGAAAGGGAAAATGAGAAAAATATACTTCTATCGTT
TTTTTATAAAAATTACTTATATCTTAATAATCTGAATATATCTTATTATTATATTAATTTTTTTTGATCAT
TTTCTAAATCATCATTTCAAGTGATGAAACAACGTGAATATTTATGGGAATTTTGCAAAATCTGTTAT
ATTTTGCTTTCAATCTTATAATTTCTGCCATATTCTAACTCAGCTTACAAAATCTACCACATTTTATCAC
ATGTTGCAATTCGAGGACGAAAATACTCTTACGATGGGACAGAGACTCTGACTTAGATGGTGACAT
GACAGCTTTTGGCTGATGTGATGATACATGACGAAAAATCAGAGTAGAAACATAAAGGATTAGGAG
ACAGAATAGTGGAGAACAGCTGGCTTCTACTGTTGTTATTGG

>*PP2AA3* LA (945bp)

GTTCGTCAACAAAAGCAACTTCGAATTAACATGAGAAAAGATACTAGCTCGTGAACTTAACTATAAA
GTATTATTTAAAGGCTAATTGTGCATGTTAATTACTATTAGATTATTGTTTATTTTATTGACATGATCT
GATTTTGTGAGTTCAAGGTTTTTTTGATTCGCAATAAAGTAACTAGATCGGACAGAAATATATAGTTT
TCAGAGTTTTCTTATGATATTAACTAATTACATCTTTGATCAATGAGTAATGTCGTATAATTTTATATA
CACAAACTACGTTTTGACTAAAAGGGTTTGTTTTATACCTTTTTCTTTTTTCTTTTGTCGAAGAGGTTT
GTTTTATACCTAAAAATAGAAAATATAGCAATGACCAAAAACATCTTGGTAATAGTATATTTTTTGCC
TTCTTCCATTATTATATTTAATTCGAATTGTCTATTAATCTTTTTTCAATGAAATTGCCTATGAATATTA
TTATAGTATGGACATATATATTATCAAATAATGAAATTTTTGAATTGCATTTGAACTTCATAGACATTG
TTATTTGCTGATCACGAACTTTCTTCTACGCCAATGTTTTTCAATTCCCACATTATTAAAGTCCAATTAC
GATCAGTTGATGTATTAAATTAGTCGTAGGAGATTACATTGTACGTGTATAAAAGTTGGTCTTGTTTG
GCTCCAATTGGCGTGGTTTACGTGTGAACAGACACAGACAACTTGACGAATCAAAACGCATAAAAC
GCGTGGCAGAAAGTTTCTATTGCTCCAAAAACAACTTTTTTATTCTCTTTTTTTTTTGACAAAGAAATA

CAGTAACTTTTTTTTATTCTCTTAAAGAAATAAAATTCCAGAGTACTTTAGTTCAGTGTATATCTCTGA
TTTATTTAATTACTATTATAATGATTACATACAACACACCTGCTACTGCCACGC

>*PP2AA3* RA (928bp)

TTTTATTCGGTAAGAGAGATCTTTATTTTCAACTATCCCAAATATCTATAAATTTAAAAATTAATTTTCT
GTACCAACAAAAAATATCTAATAATAAGCATTTTGGGCCGCCACTTGTGCCCGAGACGTTGAAACCA
AGGAAAATATTTGTGGTGCTGGTGCAGCTAAATTCCAAAAATGAATTTATGGGTTCATTTTCAGAGT
TTTGTTTTCTAGGACCGTTGTAATAAAAAATCTTTTGTTCTCGTTAAGTATAATAAGAATCTTGATAAT
AGCTCTTTGTAGTTATAGATGTGTATCTATTAGTCTTGGTCAATAAGGCAAAACTAATACATACTACT
AAGTTACATGAATTGTTCGAGAAAGTCTTGTCTTATAAACTCGTTTAGTTTTCTTTGAGAGTACTTTTT
TCTTTCTTTACTTTGAGAGTACTATGGTTACGTAGATAATAATCATCGTTATCGGCTTCTAATTTGTAT
AAGATAATAACTTTTATGCATTTTCTAACACGAGAACTTATCGAGTTATCTCAAACGTCACGAAAGAA
AAAGGACAATGAACATAATATATTATATTATCCATTTGTATATTTAGACATATATTTTAATTTATTAAC
TAAATCCGTTCATGTGACAAGATCGACCCTTATAAGAAGGAGAGATTCGATTTTGTCACGAATCCTT
ATCTGTCCGAAATTGTTTTTAGCAATATGCGATTATCGTAATAATTTTCTTCTTTATAATTAGTGTCAT
AAATCTATTACTTTCTATTTTGCTGTTTCCAAAATCTTTTTGTGTTTATGAACTGAACACTCATATAAAA
AAAAAATACTTGTAATTGCAACCGTTAAATGAATGTTATATTAGCAACAAACTTTGGTAGTTGTTGGA
ATGGAACTAAGCACGACATTATCCACGCACTTCAAACA

>*PP2AA3* promoter region (2031bp)

CTAACAACAATAATACAACAGATTGAAACAAAACTTGCAATATATATATATATATATATATATATATA
TATATAACCTGTATCAATCAATTTTTAAAACCTAGTCAATATCAATAGCTTCAAACGAAAAATTCATGT
GTAAACTTATAGCACAAATTGAAACCCAAGTGATTAAATTCCCAGATTCAGAATCATGTGAACATCA
CTGCGTACCTTAACCTAGACTCAATTTCAAATTCGAATTCTAATTTTTATGGTCGTATCTGTTTGTGAT
CTCCGTTTAGAATTAAAGCAAAAATGTAGAATCGCAGAGAAGTGCTTACCTCATTGACAAAGTTGAG
TACGATAGCTGCAGAAATAAAAAATTCAGGAAACAAGTAAATAAGAGAGAAATCAAAATGAGGAAT
AAGATCAAAGAGAACTAGCAGATTTTTTTCAGAGATTATATCATTATACCTTCGGTGTTATCCGATTT
AGGAGCCATTTTTGTCGAAATAATTGGCGGGAAATAAAAAACCTGAAGCGAGAGCAACTTTTTTCCT
TTTCCGATATTTTCTGAGGAAGAAGAAGAGAGTTACTGAGAAGCTTCCTTTATACGAGAGAGATATT
AAACGAAGTCGTATTTGTTGGATATATGGGCCTTTATGTATTATACATAAGCCCATTTAGTTCTTATCT
TTGACAGGCCCAATTATATGGACACAGCTTGAAGATGGTTTAGCTGCTGCGAAAGACGAGCTCCGG
TCTCATTTCTCGTTCTTCTGATTGATAGATCGCTCGGAACTTGGAAAGCAGCGTAATCGGTAAATTCT

CGATCATTTCTTCAATTTCGAAACATTTTCCAATTGTATATTGAGAATCTCAAATCGAATTATTGTTTA
GATCTTCTCTCTTTATTAGTCTAAGAATGGTAATAGTTTCTCTCTAGATTGGGACGGCACAAAGTTTC
AGTTCTCCATTCTAATCATTTTTTAGAGGCTAATGGGTCACCTTTGAATTATCTCAGAGGCGGAATTG
AGTTTAAAGTTTGTAACTTTATCGTCTTTTGTATTGATGGTGTCTATGTATGCATTGTGTTTTATTCTG
ACTTTGTGTACATTTGTGTTTGATTGAAGGTAGGGAGTGATTTGAGTTTTGGTGAGGATGTCTATGG
TTGATGAGCCTTTATACCCGATTGCTGTGCTTATCGACGAGCTAAAAAACGATGATATTCAGCGTAG
ATTGAACTCTATTAAACGGCTTTCTATCATTGCTCGTGCTCTTGGAGAGGAGAGGACAAGAAAAGAG
TTGATTCCATTTCTTAGTGAGAACAATGACGATGACGATGAGGTGCTTTTGGCTATGGCGGAAGAGT
TGGGGGGTTTTATTCTGTATGTAGGAGGGGTTGAGTATGCATATGTTCTGCTTCCACCTTTGGAGAC
TCTATCCACTGTTGAGGAAACTTGCGTGAGGGAGAAAGCTGTGGATTCACTTTGTAGAATTGGTGCT
CAGATGAGGGAGAGTGACTTGGTTGAGCATTTCACTCCTCTGGCTAAGGTTAGATAAGGCTGGTTTA
CTTATTTTATCAAATCTCTTTGGTCGATTTTATTGATTCCTTTCCATGCTCTGTCAAACTTAATATTATT
AAGCCACTAAAACAGAAGAACATAGATTTTTTGGGGTCCTTTGTCAATTGTCCCCCAAAGTTTGGAA
ATATGCTTTTCTTGTTCAGTGTTCCCCTTATAGGTGTTTTGGGAAATTGGTAGGTTGAAGTTGTTTAC
ATCCATTTTGGTTGATCTAAGAGAACATCGCTCTGGCTTTAGCGCTCTTTTGGGTTGTGGTTGTGTAT
GTGACCATACATACCTCGGTATAGGTGCTCGTAATGTGAAACCGACTTAAGCATTACTGAACACTGT
CATTGTGTAACAGCGACTTTCAGCTGGTGAATGGTTCACAGCCAGAGTATCAGCATGTGGGATTTTC
CATATTGCATACCCAAGTGCCCCAGATGTGCTAAAGACGGAGCTAAGATCAATATATGGTCAGCTTT
GTC

>*PP2AA3* terminator region (861bp)

TTGTGGAGAACATGATACGGCCATGCTTGGTGGAGCTAAGTGAAGACCCAGATGTTGATGTTCGGT
ATTTCGCAAATCAAGCTCTCCAATCTATTGACAATGTGATGATGTCTAGCTAAAAAAAGGTAAAGAA
GACAGCAACGAATTGTGTTTGGTTCTCATGAGATTTTGTAAACGATACTTTGTCGTGTGTTGTCTTTT
ACTTTACACGTACGTGACCATTGTTTCTCTCTGCTACTAATGTTAATGTTGGCTTCATGTTTTCTGTGA
TTTGTTCGTTGGGCGTATTTGCTTTTTGGTGCTTAATTTTGTTTAGTCCAAATAATTTACTTATCAAGT
GATTCGGCAACGTTTTTGCTCGAAGCATGAGTGTACAATTGGTCACCACTGGAATATTTTAATTGGTC
AACCAACTTATTGAGTCGTTTTCAAGCCCATGGGCTTTTATACTCCACATATCAGATTCAGGTCGTTG
TTGTCTTTTTAACTAGAAGACTATTTTATCAAAAGTCAAAACCTTGTCGTAGAAGGCACTTACGGAGA
ATTATTGACCTAGGAACTTCCCTCGAAAAGTTCAATATATATGTGTTATTGACGTTTTCTGTTTGATTA
TTTACAAATTTTGTTGATTATTTTTCACCACTCTTGCACACACCTCTCTCGCTATTTCAAAACTCTACGA
AATTTCCTTAATATTTCTATTTGATCTTCCAACAGTTTTCGACTCGCAATAGGATTAGGAAATTACAAA

TAATATCTGAATTATTTAAACGACAACACAAATATTATAGAGTCTTCTCTTGTTTCTCGGCTTCAGTTT
ATTTATGTTGTTTATGATTCTGGCTTAATAGTTACGTAACC

>*THAS* promoter region (3416bp)

CCACTCTCTCCTGGTAGATGTAAATAATTTATTAAAATTCAATAATACAACATATGAACATTTTTTGTA
ATAAATTATTGACTAATTTATTAAAATTCAACCGCATTATTCTAAATCTCTCGCTTTACATTTTTGAATA
TCGTAACTTATAATAAACAAAAATTATATCTAACAAAAAATATTTTAGTTGAATTATAATATGAATAT
AAACAATGTTTTTAGTGAAACCAAAAAAATATTATCATACAAATTTATCAGTAAATACAATGTAAATA
TAATAATATAACAGGATCGTCATATTTTGACTTGATGTCCAGTTTTGTAAATAGTAGACATTTTTTTGT
TTTTTTAAATAGTATTCCATCTGTTATTTTATTTTGGAAAAAAAAATTGAATAAAAAAAATTGAATTTT
TATATTTTCTATGCAATATTTATTAGTTATTATTCATAAATGGTAAACTTTAATAAAAATAATTAAATA
TTATTGGTGAATATTATTGGTTTAAAGTTATAAGAAATTGTAAATTACAAAATAATACATTTTAAATC
AACTTTAATATGTTTACCTTAATATGTGTGTTTTTCCTAAACAATCAAATAAAAAGGAACAGAGGAAG
TAAGACTTTTATATCTCGTTGCTTCTAATTCTTAACATTAAAAAGATACGCCCATCAACTTTCTTATTG
AAGTCTAGATAGCTTTTAATATTTTTGAATTGAAAAGTAATTGAAAACAATGAATATAAAACATTCAC
GCATCATAATCATACGACAATTAGTCAAATTTATGCCATGATGAAACTACCAAATCCAAAAATTTAAA
TTTGGAACCCGTTTGGAAAAAAAACAATCATAGAAAATATCCAACGGATCTTAGATTTAGGAATTTT
GCACATCGGATAATACTTGAACAAAACCCGAAAATAGTACAGTAAAAAACTCAAAGCTAAGTTAACT
TATTAACCATTTTTATATGGAAAAAGTTAAAAAAAATTTACGTACTTTTAAATTTGGAACGTTTAAATC
CTATGTAAGAAGATAAAAAAATCGTAATGCTTTTGTTAACTTTAAATTTAAATAAGAAAGTTTGGTTA
ACAAGCCATTTTAGTCATGACGTTATGTAAACTAACAGAAAACCTTTACGGTGTTAAAATCTCGTTAG
TCTTCTTTATTATTTCTAAACAACATCATTTATCGTATAGAGAAAACGACGTCATTTTGTCGATTTAAA
AAAACAAAAACCTTAATCCCTAAATCGAAACTCATTTTTCCCTAATCGATTTCTAATCCTAGAATCTTT
ATGTCTCTCTTTTGAAATTCCGATTGAAATGGTTACAACTCGTGGGGGGGGGGGGGAAGGGGGGA
AGGTTCGTCTTCATCGGTCACAGAGACTAAATGAAACAGCAGCCGTACAAAATGACGACACATGAT
GTGACAGAGCTTGAACTCATCAAAACATACTGAAAGGAAGAAGTCTCGAGCTAATTTCAACATAGA
GATCTTTGGGGATTGTGTTTTCGATTTGGGAAATTTTTAGCTTTCAGGGATTGTTTTTTCGATGTGGG
GATTTTGTCTATTTTCATTTATCAATACGACTTCGTTTAGTGTACAATAAACCGATGACGTTTAGCCAT
AACAGAAACGACTAACGGGATTTTAACACCGTAAAGTTTTTCTGTTAACCTACTTAACGACATGTCTA
AATTGGCTTGTCAACCAACATTGTCATTTAAATTTAAAGTTAACAAAAACTTTACGATTTTTTTGTTAT
CCTCGCAGTTCATGATTTAATCATCCCAAATTTGAAAGTTAGTGATTTTTTTGACATTTTCCCCATTTTT
ATATAGCCTTAAGTAATTAAGATATTTTCGAGTATTTTTGGATAGCTTTTTCTTATCAAGTTTTTCGGT

TAATTCAAATATTTTGGAGTACTTTCGAATACTTATGAGTAGTTTGGATTTCTTTTTGACCAAATTTTC

CGGATGACAGATATTTGAATGCTCGCTTGGGTTAGCGGGAAGTCTAAGTAAAGAAGGAGGCAAGC

AAATGAAATTCGGATATGTTGTAGAGTATTTTTGATATTTCCTCATAGTTCTTTTAACCTTAAACATAT

AAAAAAAAACACAAAAAAATATCTGAAAGTTTGAACCCTATCTGAAACAAACGTAAATATTATAGAT

ACCCAAACGGGGCTATGAGCCTCACCCAAAAACCCGAATCGAGTCCAAACCCCGAACCCTAGTGAA

AAGCATTTGCAGTAAAAGATTATAAACATGAAGTAATAAATTAGTTGGAACTTATAATATTGATTTAT

ATCACCAGTTCCCCACACGCAACGTGGTTTCTTTCATCAATCTATATAGTATTATAAGTAGAATGTGT

ACTCTCTTTCCAAATAAACTAGGAGGATTCAAAATAGATGTGGAGGCTAAGACTTGGACCGAAGGC

CGGTGAGGATACTCGCTTGTTCACCACCAACAACTAATTAAGCCAGGAGGCAGATTTGGGTGTTGA

ATACCAACTCAGGATCTCCACAAGAAATTGCCGAGATAGAGAACGCTCGTCAGAATTTTTCAAACAA

CATGTCACGTTTCAAAACGATGCAGATATATATGTACCTTAAAAAATCCTAGATATACGCAACTTGCA

AACAAAGCAATAATTTGTTATCAAAGCTAGTTAGTTTTTTTTTTTTTGTAATTCCATATGCCTTGTTTTT

GAGATTAACAAGAAAACGACTTCTTTTTTTTCTTAAATTCTACTTTCCATTTCTCTCAACTCTCGGAATC

ACCTATGCTGATTCTACTCGCACATTGCATATAGTTCTTAATCTTAAACTAGGTACGGCGAACAACTT

TCCTATAGATGTCTACTTGGCTTATAATATTTTTGTATTGAAAAACTAATTAGAAAGATATGAATATC

GTTAAGCAAATATAAAACAATTACTAATACGTCAAGAAGTCAAAGTGATGACAATGCGAAACGTAG

TTGCTTACGAAATAAACGTTAAGGAATAAAATTATTTGAGAACTTATAATTTATATTCACACCCGTTTC

TCACATGCATCGTGTTTATTAACAACAATCTATATAATATTCTTTTCTACGTATTTTTGTGGTTGATGTT

TAAAATTTTTACAAAACAATTTTTTTTTTTTTAATTTTTGTTTTTCTCAAAACTAACACATTGTTTATAAT

CTATTATAACTCAAACAAACCAATAATAAAATCATTTACATAATAGAAATCATCAAAAATAAAAATAA

AAATTAATTGGTTTGTATTAAAATTTGAAAACATAATTTAATTGCAAGAAAAAAAAATTCCTAAAACA

TCGACTATAAAAAATACGAAGCGAGTATAAAGTATAATGTGTCCTCCTTGTCCAAGTAAACTTAGAG

CAACCTTACAAA

>*THAS* terminator region (1349bp)

CCATGTCACATCTATCTTCTTCACCGTTTTATATATCAATTCTGTTTTTGTTTTTATTTTCCTTTGCATGT

ATAATAAAAGAAATGTTATGTACTTCTATGAAATGAACTATGGTTATTGACAGATTACATATTAATTC

CACAAATTTATTTGTTGAATGTGTTCAAAGAGTACAAGTACTAAAAATTATAACCTACAAGTTGAAAA

TTTATAATTTATAATTAGTTTGATTTTTTTTTTTTTAGAAATAAATAGTTAATTATAATTTAGGTATTTA

AGTACTCATACAACTAAAAAATATGTAGTTTTTCTTTTCTTCTCAAATGAAGTATATTTTCCTAAATTTT

ATAGTTTAGTGTATTTAAAAAATTTCCGCCCATTGTTTTAACTAAGATTATGTGTCTTACCTAATGATA

ATAAGGTCATCTCCCTATCTATAAGTAGATGCTTTGCTCCATCACAAATGTATCATATTAATAAAAAC

AACACTTAAAGAGAAAGATTAGAGGTAAGGGAAGAGAAACCTTATCTTCGAATCAAAGTTATAAGA
GTTTCACTTTCAAGTCTACAAATCATTTAGGAAGAAAGCTTTACCATCAAAGGTAGCATCATTTGTAC
AATGAATGGAGGTAACTCTAAATCCTCTTTTTGATTCTATTATGTGATTGATATAGATGTCCCATATCT
TAGGGTTCATATCAAAAAATATGTTTAACAAAGCAATCTCGCTGAGTCTTTTGATGCGGTTTTAAAAG
ATGCACGGTAATTCCTCATCATTACAAAAACCGAGTACATACAGAGTACACTCGGGATGTGGTTTTG
GCAAATGTACCAGTCCAACTTCGAAAGGATGATATCGCTTGGCGTGCTCCAAATAAAAAATGCAGA
GTATGAGGTTAAGACAGTCGCCGGTTCATCTATCAAGAATATCTGCTCATGAAATCTTGCAAATGAA
TGTACAATCTGGAATGGATCACATGCAATGCAGGGCTTAATTAGGTCCACACACTTGTTAATGTGGA
ATATAGCTCTACTTTTAGAGGTGGCTTAGCGCAAAAGTTTCTAACCAGTACCAGACATTCCTGCATTC
AATATGTTGCATTTGGCGGCCATTTGAATACAAGGGAACGGCATAACTATTGCTTTTTATTCTTCTTT
GTAGTTGACCAATAAAAAAGTTTACACCGGTTCAAATGAAACAAATACAAGTCCTAAATTCAACCAA
ATTCTGAGTTCATCAACATCAGTTGCTTTAACAAAAGCAAAGATTTCAAAACTTCAGAATCAGCAGG
AAGTAGGAAATATTTGTCGACCCATGAATGCAACTGGTTCACGGCTGAACACTGAAA

>eGFP/Gus (2535bp)

ATGGTGAGCAAGGGCGAGGAGCTGTTCACCGGGGTGGTGCCCATCCTGGTCGAGCTGGACGGCGA
CGTAAACGGCCACAAGTTCAGCGTGTCCGGCGAGGGCGAGGGCGATGCCACCTACGGCAAGCTGA
CCCTGAAGTTCATCTGCACCACCGGCAAGCTGCCCGTGCCCTGGCCCACCCTCGTGACCACCCTGAC
CTACGGCGTGCAGTGCTTCAGCCGCTACCCCGACCACATGAAGCAGCACGACTTCTTCAAGTCCGCC
ATGCCCGAAGGCTACGTCCAGGAGCGCACCATCTTCTTCAAGGACGACGGCAACTACAAGACCCGC
GCCGAGGTGAAGTTCGAGGGCGACACCCTGGTGAACCGCATCGAGCTGAAGGGCATCGACTTCAA
GGAGGACGGCAACATCCTGGGGCACAAGCTGGAGTACAACTACAACAGCCACAACGTCTATATCAT
GGCCGACAAGCAGAAGAACGGCATCAAGGTGAACTTCAAGATCCGCCACAACATCGAGGACGGCA
GCGTGCAGCTCGCCGACCACTACAAGCAGAACACCCCCATCGGCGACGGCCCCGTGCTGCTGCCCG
ACAACCACTACCTGAGCACCCAGTCCGCCCTGAGCAAAGACCCCAACGAGAAGCGCGATCACATGG
TCCTGCTGGAGTTCGTGACCGCCGCCGGGATCACTCTCGGCATGGACGAGCTGTACGAACCCGGCA
TGTTACGTCCTGTAGAAACCCCAACCCGTGAAATCAAAAAACTCGACGGCCTGTGGGCATTCAGTCT
GGATCGCGAAAACTGTGGAATTGATCAGCGTTGGTGGGAAAGCGCGTTACAAGAAAGCCGGGCAA
TTGCTGTGCCAGGCAGTTTTAACGATCAGTTCGCCGATGCAGATTTTCGTAATTATGCGGGCAACGT
CTGGTATCAGCGCGAAGTCTTTATACCGAAAGGTTGGGCAGGCCAGCGTATCGTGCTGCGTTTCGAT
GCGGTCACTCATTACGGCAAAGTGTGGGTCAATAATCAGGAAGTGATGGAGCATCAGGGCGGCTAT
ACGCCATTTGAAGCCGATGTCACGCCGTATGTTATTGCCGGGAAAAGTGTACGTATCACCGTTTGCG

TGAACAACGAACTGAACTGGCAGACTATCCCGCCGGGAATGGTGATTACCGACGAAAACGGCAAGA

AAAAGCAGTCTTACTTCCATGATTTCTTTAACTATGCCGGAATCCATCGCAGCGTAATGCTCTACACC

ACGCCGAACACCTGGGTGGACGATATCACCGTGGTGACGCATGTCGCGCAAGACTGTAACCACGCG

TCTGTTGACTGGCAGGTGGTGGCCAATGGTGATGTCAGCGTTGAACTGCGTGATGCGGATCAACAG

GTGGTTGCAACTGGACAAGGCACTAGCGGGACTTTGCAAGTGGTGAATCCGCACCTCTGGCAACCG

GGTGAAGGTTATCTCTATGAACTGTGCGTCACAGCCAAAAGCCAGACAGAGTGTGATATCTACCCGC

TTCGCGTCGGCATCCGGTCAGTGGCAGTGAAGGGCGAACAGTTCCTGATTAACCACAAACCGTTCTA

CTTTACTGGCTTTGGTCGTCATGAAGATGCGGACTTGCGTGGCAAAGGATTCGATAACGTGCTGATG

GTGCACGACCACGCATTAATGGACTGGATTGGGGCCAACTCCTACCGTACCTCGCATTACCCTTACG

CTGAAGAGATGCTCGACTGGGCAGATGAACATGGCATCGTGGTGATTGATGAAACTGCTGCTGTCG

GCTTTAACCTCTCTTTAGGCATTGGTTTCGAAGCGGGCAACAAGCCGAAAGAACTGTACAGCGAAGA

GGCAGTCAACGGGGAAACTCAGCAAGCGCACTTACAGGCGATTAAAGAGCTGATAGCGCGTGACA

AAAACCACCCAAGCGTGGTGATGTGGAGTATTGCCAACGAACCGGATACCCGTCCGCAAGGTGCAC

GGGAATATTTCGCGCCACTGGCGGAAGCAACGCGTAAACTCGACCCGACGCGTCCGATCACCTGCG

TCAATGTAATGTTCTGCGACGCTCACACCGATACCATCAGCGATCTCTTTGATGTGCTGTGCCTGAAC

CGTTATTACGGATGGTATGTCCAAAGCGGCGATTTGGAAACGGCAGAGAAGGTACTGGAAAAAGA

ACTTCTGGCCTGGCAGGAGAAACTGCATCAGCCGATTATCATCACCGAATACGGCGTGGATACGTTA

GCCGGGCTGCACTCAATGTACACCGACATGTGGAGTGAAGAGTATCAGTGTGCATGGCTGGATATG

TATCACCGCGTCTTTGATCGCGTCAGCGCCGTCGTCGGTGAACAGGTATGGAATTTCGCCGATTTTG

CGACCTCGCAAGGCATATTGCGCGTTGGCGGTAACAAGAAAGGGATCTTCACTCGCGACCGCAAAC

CGAAGTCGGCGGCTTTTCTGCTGCAAAAACGCTGGACTGGCATGAACTTCGGTGAAAAACCGCAGC

AGGGAGGCAAACAATGA

>U6-26 promoter::sgRNA::U6-26 terminator (378bp)

CATCTTCATTCTTAAGATATGAAGATAATCTTCAAAAGGCCCCTGGGAATCTGAAAGAAGAGAAGCA

GGCCCATTTATATGGGAAAGAACAATAGTATTTCTTATATAGGCCCATTTAAGTTGAAAACAATCTTC

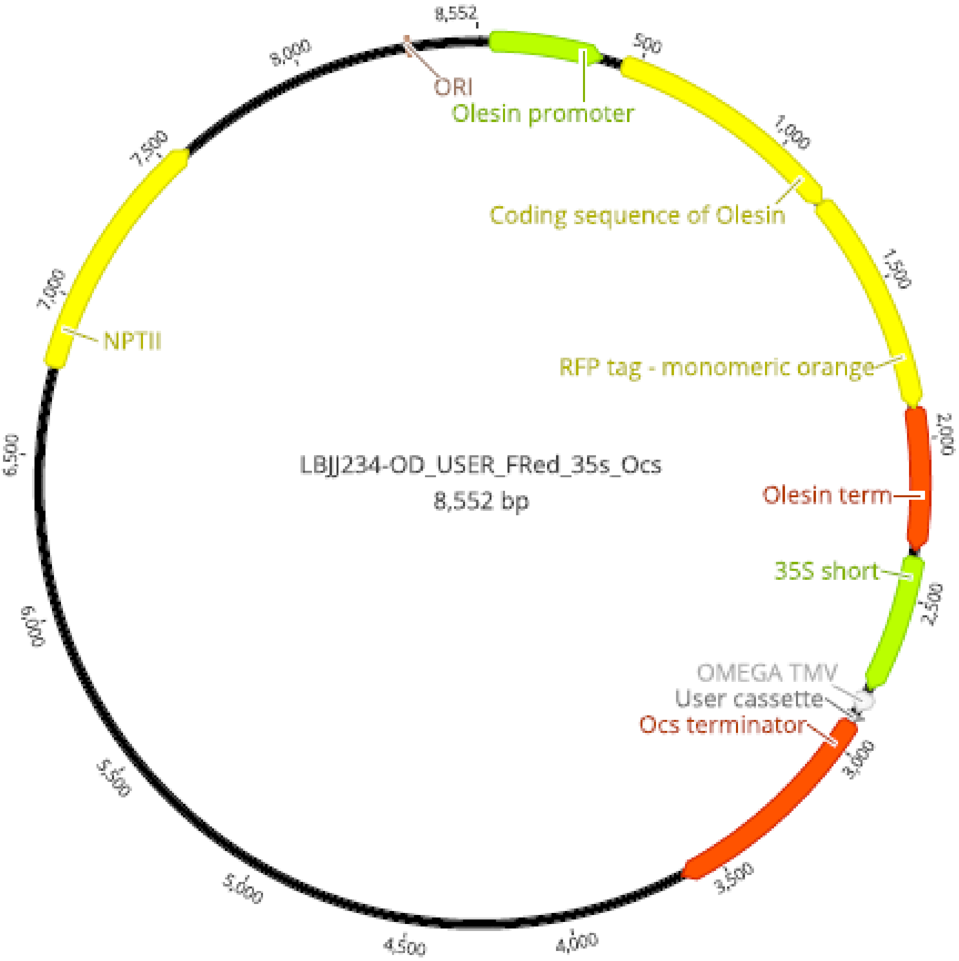
AAAAGTCCCACATCGCTTAGATAAGAAAACGAAGCTGAGTTTATATACAGCTAGAGTCGAAGTAGT

GATTGTTTAAGAGCTATGCTGGAAACAGCATAGCAAGTTTAAATAAGGCTAGTCCGTTATCAACTTG

AAAAAGTGGCACCGAGTCGGTGCTTTTTTTTGCAAAATTTTCCAGATCGATTTCTTCTTCCTCTGTTCT

TCGGCGTTCAATTTCTGGGGT

7.2 USER vector



Supplementary Figure 1: Map of the USER plasmid used for USER cloning.

7.3 Primers

Supplementary Table 1: Primers used for genotyping PCRs

T-DNA line	Primers
SALK_021008	TCAGACGATGCTTATGGTTCC
	TTATGAACCAAAGCAAAACCG
SALK_073442	TTCAGGTTTATGGTCACCCTG
	TAGAACCATCATTTCCCTCCC
SALK_057762	AATGTGGAGGCAACAACTCAC
	TCAGTCATTCAGCAGTTGTCG
SALK_034869	TGTTTCCGTGTAACTTCTGGG
	TTGTTCTTGACGACGACTGTG
SALK_149692	TCTTGTGACAGGTGCAACTTG
	AAACAAAGCTAGGCACAAGGC

SALK_135712	GAATTCTAAATTTCCCACCCG
	TGAAATGAAAAACGAATTGGG
SALK_128444	TGGTGAACTGTGGAAACCTAG
	TCCTAAAATCCAAATCCCATG
SALK_135831	AATTGCAGCATATTCCAAACG
	TAGTTGCTGGAGAAGCTGCTC
SALK_117262	GCAATTTGGCTGTACAGAAGC
	AAGTTGCTGCTATGATGCCTC
SALK_003313	CCGAGCTTTCTACCATGACTG
	TATTCGACGGTTGCGTTAGAC
SALK_013895	CTTTCTCGCAAGATCCATGAC
	TTTACATGCTTTGCCGGTTAC
SALK_035608	GCTGCTACCACTAGTTGCCAG
	ACTGCCACGATAATGAGGTTG
SALK_092672	ACGCAATGTATGTCTTGGGTC
	ATCCATTGCCAATTTTCACAG
SALK_026442	GCTGGGGGTTTATGTAGGAAG
	CACTGTCCAGTAAAAGCTGGC
GK-128H01	AGCATAGCAATTGGGCATACACTG
	GAATTTAGTTGCACTGCCTGGAAA
GK-143H01	TGTAAAATCTGCTTCCTGGTGTGT
	TTGTGGGTTCATGTAACATGTGCT
GK-229F03	GATAGTAGAGATCCCTGCAGACGG
	TCTGGAAAACTGAAACCTCTTTAATTG
GK-663C11	TATTACCTTGAAGAGTGGACCCAG
	TTGAAGCTGTTAAGTCCCTCTCAG
Salk_149002	AATGAAAGCATGCGGATACAC
	TCCGTGTTGACTGGAAAGATC
Salk_029530	AGGCGACATATGAACGAACAC
	GATGCAATTTCTTCTGGCAAG
Salk_065480	CCTTCATCGCAATCGTAAATC
	TTTTGCGCTAAACTAGTTGGG
Salk_037362	TTTTTGTTGGATGCGTTCTTC
	AGAGGAGATTCTCCGTCCATC
Salk_ 060156	ATCGTCTTGCTTCTTCTTCCC
	TGGCTTACTTGGTGGATTGAG
Sail_80_F03	TTCACGCAACAAAAAGGAAAC
	AGCACAATAATGGCGATGAAC
Sail_535_F09	TTCAAAAAGGTCAGGCAAATG
	TTGAGACGTTTGGGTTTGAAC

Supplementary Table 2: Primers used for genotyping CRISPR plants

Region	Primer
PP2AA3 insertion	CAGAGATTGTAGAGCCTACC
	CTCTCACTCTCTCAGCTTTG
MRN insertion	CTCGACATCTCAACGCAAC
	GTGTTGAATTGGAGGTGAG
THAS insertion	CTAGTAGGAACAGTCGGAGT
	GCTAGCTGGTAATTATGTC
GFP/Gus	GAACTTGTGGCCGTTTACG
	GACTTCAAGGAGGACGGCAA
Chr3 del 1	GATGGTCCTATGGGTTCAG
	GCAGATTGGCTAGACTCTG
Chr3 del 2	GCAAAGTGTGATGCTCCAC
	GTCAGCATCAGGCCAACCG
Chr3 del mid 1	CTGCGGATTTTGTACACGC
	CAAGGATGTGAAGACTCTT
Chr3 del mid 2	GCAAAGCTTGTGGAACTAG
	CATAATCCACAGGGACCTG
Chr3 del 1	CTCCGGCTGCGATTTTTAG
	GTCCTGGATCGTCTAATACG
Chr3 del 2	GAATCGGACCAAGTATCAGG
	CCCAATCATAAGCTGATCAC
Chr5 del 1	CAGGGTCGATGCGATTGATG
	GCCAGTCTTGCTAACACTT
Chr5 del 2	GTTTCTTCTCTCGCACTCCA
	CTGAAAAATGAGGACAGCG
Chr5 del mid 1	GCAGACGAGAAATGAGTGG
	GGTAACAAGCGACTAAATG
Chr5 del mid 2	GCATGTTTAAGATCGGTCG
	CCAATCAGATAAGCCGTTG
Chr5 del 1	GTTGTTTCCATTGGCTTGG
	CATCTAACGAGTCTTGAG
Chr5 del 2	GACGCGGCTGTAGATTGAT
	GGTCTTCTACCACCTACGA

Supplementary Table 3: Primers used for qPCR

qPCR primer location	Primer
PP2AA3	GTTGTGGAGAACATGATACGG
	GCTAGACATCATCACATTGTC
At5g47950	GCTGGATCTGTTTCTCCGA
	CAAGCAGTTCTGTGCTCTG
At5g47970	GGATACCAACGAGACCTGTG
	CATCACATCTTTGTAGCTGCATG
At5g47980	CGGCCATACTAACACTAGACT

	TGCCTACCCATACTTCAAGAC
At5g47990	ATCCTGATTTCTGGGAAGACC
	TGTTGCACCATCATTCCAATT
At5g48000	GAGGATTAATCAAAGGTTATCA
	GTTAGTAAGAAAGCCTGTCGC
At5g48010	CTTCAATCCACTATGGCACTC
	TAAAAATAATCACTCTTAGGGTCTTC
At5g48030	GACAGCGTGAACTGCTTGAG
	CACTGGGAAGATCCAGTTGC
Acn	GAGCAGGTCACAGTCATGAAG
	CGACAGACACGAATCATTTCG

7.4 Codes

7.4.1 Edit data, make readable in R

```
setwd("Data files/")
library("tidyverse")

##click on data file

dataname <- GSE25446.H3K27me3.Col.0_seedling_10d.ref6_peaks.broadPeak
names(dataname )[1]<- "Chromosome"
names(dataname )[2]<- "From"
names(dataname )[3]<- "To"
names(dataname )[4]<- "Accession"
names(dataname )[5]<- "Score"
names(dataname )[6]<- "/"
names(dataname )[7]<- "Signal Value"
names(dataname )[8]<- "pValue (-log10)"
names(dataname )[9]<- "qValue (-log10)"

##Change file name!
dataname$Size <- dataname$To - dataname$From
```

```
filename <- paste("H3.1.Col.0_seedling_10d_", nrow(dataname), "Sizes_", ncol(dataname),
"variables.csv", sep = "")
write.csv(dataname,filename )

##Change table name (will open called table)
table <- read.csv("H2A.Z_leaves_26406Sizes_10variables.csv", header = TRUE)
```

7.4.2 Select for ROI

```
##Regions of interest

Thalianol_cluster_large <- as.data.frame(list(5, 19200000, 19550000))
names(Thalianol_cluster_large )[1]<- "Chromosome"
names(Thalianol_cluster )[2]<- "From"
names(Thalianol_cluster_large )[3]<- "To"
rownames(Thalianol_cluster_large) <- "Thalianol_large"
ThalLarge <- 19550000-19200000

Thalianol_cluster_small <- as.data.frame(list(5, 19412077, 19462516))
names(Thalianol_cluster_small )[1]<- "Chromosome"
names(Thalianol_cluster_small )[2]<- "From"
names(Thalianol_cluster_small )[3]<- "To"
rownames(Thalianol_cluster_small) <- "Thalianol_small"
ThalSmall <-19462516-19412077

Marneral_cluster <- as.data.frame(list(5, 17000000, 17120000))
names(Marneral_cluster )[1]<- "Chromosome"
names(Marneral_cluster )[2]<- "From"
names(Marneral_cluster )[3]<- "To"
rownames(Marneral_cluster) <- "Marneral"
Marn  <- 17120000-17000000

Arabidol_cluster <- as.data.frame(list(4, 8700000, 8830000))
```

```
names(Arabidol_cluster )[1]<- "Chromosome"
names(Arabidol_cluster )[2]<- "From"
names(Arabidol_cluster )[3]<- "To"
rownames(Arabidol_cluster) <- "Arabidiol"
Arabidiol <- 8830000-8700000

Chr3_reg <- as.data.frame(list(3, 4250000, 4300000))
names(Chr3_reg )[1]<- "Chromosome"
names(Chr3_reg )[2]<- "From"
names(Chr3_reg )[3]<- "To"
rownames(Chr3_reg) <- "Chr3"
Chr3 <- 4300000-4250000

##Make table of all
table1 <- rbind(Thalianol_cluster, Marneral_cluster)
table2 <- rbind(table1, Arabidol_cluster)
table3 <- rbind(table2, Chr3_reg)

dim(table)
ROI <- rbind(table3,Thalianol_cluster_meg )

##Save table
write.csv(ROI, "Regions_of_interest.csv")
Regions <- read.csv("Regions_of_interest", header = TRUE)

## Bind with random region table
Random <- read.csv("Random_region", header = TRUE)
Random$X <- NULL
Random$Size <- NULL
All_regions <- rbind(Random, table3)

#Save table
```

```
write.csv(All_regions, "All_region1.csv")
All_region <- read.csv("All_region", header = TRUE)
```

7.4.3 Selecting random regions

```
##100 random chromosomes
Chromosome <- as.data.frame(sample(1:5, 100, replace = T))
view(Chromosome)

#100 random sample sizes
SampleSize <- as.data.frame (sample (20000:200000, 100, replace = T))

table <- cbind(Chromosome, SampleSize)
view(table)

##How many random regions in each chromosome
Chromosome_1 <- sum(Chromosome == 1)
Chromosome_2 <- sum(Chromosome == 2)
Chromosome_3 <- sum(Chromosome == 3)
Chromosome_4 <- sum(Chromosome == 4)
Chromosome_5 <- sum(Chromosome == 5)
view(Chromosome_1)

Chromosome_1_length <- 17438289
Chromosome_2_length <- 9907821
Chromosome_3_length <- 12830135
Chromosome_4_length <- 10011585
Chromosome_5_length <- 15235375
Genome_length <- 65423205

##Random starter number for Chromosome
Ch1_start <- sample(1:Chromosome_1_length, Chromosome_1, replace = T)
Ch2_start <- sample(1:Chromosome_2_length, Chromosome_2, replace = T)
```

```
Ch3_start <- sample(1:Chromosome_3_length, Chromosome_3, replace = T)
Ch4_start <- sample(1:Chromosome_4_length, Chromosome_4, replace = T)
Ch5_start <- sample(1:Chromosome_5_length, Chromosome_5, replace = T)
view(Ch2_start)

##save as data frame to bind
Ch1_start_1 <- as.data.frame(Ch1_start)
Ch2_start_1 <- as.data.frame(Ch2_start)
Ch3_start_1 <- as.data.frame(Ch3_start)
Ch4_start_1 <- as.data.frame(Ch4_start)
Ch5_start_1 <- as.data.frame(Ch5_start)

##merge rows together, can only do 2 at a time and need same column name
names(Ch1_start_1)[1]<- "From"
names(Ch2_start_1)[1]<- "From"
names(Ch3_start_1)[1]<- "From"
names(Ch4_start_1)[1]<- "From"
names(Ch5_start_1)[1]<- "From"



table1 <- rbind(Ch1_start_1, Ch2_start_1)
table2 <- rbind(table1, Ch3_start_1)
table3 <- rbind(table2, Ch4_start_1)
table4 <- rbind(table3, Ch5_start_1)
dim(table)
total_table <- cbind(table, table4)

#rename columns
names(total_table)[1] <- "Chromosome"
names(total_table)[2] <- "Size"

#Add end bp
```

```
total_table$To <- total_table$Size + total_table$From

##Save table
write.csv(total_table, "Random_region.csv")
Random <- read.csv("Random_region.csv", header = TRUE)

##Order by chromosome
table2 <- read.csv("Random_region.csv")
table2 <- with(table, table[order(Chromosome ) ,])
table2$X = NULL

write.csv(table2, "Random_region.csv")
```

Table S 1: Co-ordinates of the 100 randomly selected regions.

Chromosome	Size (bp)	From	To
1	167083	3934083	4101166
1	79034	4275854	4354888
1	64574	4147771	4212345
1	109773	4183042	4292815
1	62702	5524154	5586856
1	72169	316380	388549
1	157713	2323288	2481001
1	73463	339191	412654
1	107055	2907405	3014460
1	189531	8242420	8431951
1	143602	8060831	8204433
1	171473	4768892	4940365
1	79184	472503	551687
1	150293	4544123	4694416
1	47276	4525094	4572370
1	75390	3554987	3630377
1	105282	2820289	2925571
2	39824	3978652	4018476
2	133302	8034466	8167768
2	61367	7525293	7586660
2	131756	792722	924478
2	151905	10248480	10400385
2	125903	8756884	8882787

2	179652	10358358	10538010
2	20360	16230117	16250477
2	194889	16223073	16417962
2	74603	7271873	7346476
2	88788	3132824	3221612
2	92099	5992354	6084453
2	166500	1062661	1229161
2	129427	7729577	7859004
2	187172	8507587	8694759
2	109052	9001625	9110677
2	138078	4144595	4282673
2	72586	4302067	4374653
2	143148	3942167	4085315
2	65452	9075743	9141195
2	102837	6734263	6837100
2	158077	5195272	5353349
2	142205	541202	683407
2	165751	2479479	2645230
2	188469	1645201	1833670
2	117714	12021605	12139319
2	62086	7123142	7185228
2	29135	11547432	11576567
2	31442	3072884	3104326
3	177808	4506446	4684254
3	136267	1899495	2035762
3	118989	9029031	9148020
3	33317	9059491	9092808
3	107953	2123106	2231059
3	46550	2628498	2675048
3	75565	4910970	4986535
3	82540	14353308	14435848
4	30234	3062228	3092462
4	86986	8594414	8681400
4	124325	9224953	9349278
4	96858	3406375	3503233
4	163691	8452885	8616576
4	122254	7278350	7400604
4	150428	7536521	7686949
4	197572	3854480	4052052
4	56632	1262405	1319037
4	82667	7232835	7315502
4	126910	9943480	10070390
4	38008	4588560	4626568
4	188713	9146328	9335041
4	125825	5053651	5179476
4	23245	6310451	6333696

4	147065	5862477	6009542
4	94488	6147192	6241680
4	40225	4920036	4960261
4	167117	7222503	7389620
4	49822	6896087	6945909
5	45651	8357049	8402700
5	78085	13230401	13308486
5	101200	10837762	10938962
5	62496	12917232	12979728
5	129175	15206188	15335363
5	60769	10408378	10469147
5	110848	3743087	3853935
5	53891	1310479	1364370
5	73042	8857025	8930067
5	123053	4599183	4722236
5	142149	5146615	5288764
5	189881	7514097	7703978
5	151258	4038146	4189404
5	198314	6764316	6962630
5	159408	6474257	6633665
5	192270	11305911	11498181
5	102511	2778821	2881332
5	125165	8294514	8419679
5	156225	4380871	4537096
5	83365	14975244	15058609
5	52194	1317964	1370158
5	117546	3352695	3470241
5	102448	14959904	15062352
5	60474	9991963	10052437
5	172761	8918606	9091367
5	40061	9531445	9571506

7.4.4 Extracting values

```
###### Code for Jade########

####### Soheila Bayat sb3014@bath.ac.uk########

setwd("")

library("dplyr")

# make a list of all tables. path = "." means working directory

file_names<- list.files(path = ".")

# chech th length of the list
```

```
length(file_names)
#test <- read.csv("test_1.csv")
#names(test)

##Random regions code - change WD
random <- read.csv("Random_region.csv")
random$X = NULL

C <- 100
A <- random[C , 3]
B <- random[C , 4]

for (i in file_names) {
  #read each data frame
  table <- read.csv(i, row.names = 1)
  # filter each dataset
  #1      3934083          4101166
  table <- filter(table, Chromosome == "5" , From >= A, To <= B  )
  #rewmove these three columns
  table$X = NULL
  table$Accession = NULL
  table$X. = NULL
  #get the sum of other columns
  sum_of_size <- as.data.frame(t(apply(table, 2, sum)))
  #bind the sum_of_size with table and assign the name "sum_size_ as the row name of the
sum_of_size table
  table <- rbind(table, "sum_size" = sum_of_size)
  #in the table change the values of 1:7 column and sum_size row to "NA"
  table["sum_size", 1:7] <- "NA"
  # save each data frame as csv
  write.csv(table, file = paste("RandomReg_", C , i, sep = "_"))
}
```

```
table2 <- read.csv("All_region1.csv")

##Specific region code

Region <- read.csv("Regions_of_interest.csv")

#REMEMBER TO CHANGE CHROMOSOME MANUALLY

#Thalianol <- C
#A <- Region[1 , 3]
#B <- Region[1 , 4]

#Marneral <- C
#A <- Region[2 , 3]
#B <- Region[2 , 4]

#Arabidiol <- C
#A <- Region[3 , 3]
#B <- Region[3 , 4]

#Chr3_region <- C
A <- Region[4 , 3]
B <- Region[4 , 4]

#Thalainol_Meg <- C
A <- Region[5 , 3]
B <- Region[5 , 4]

file_names <- list.files(path = ".")

for (i in file_names) {
  #read each data frame
  table <- read.csv(i, row.names = 1)
  # filter each dataset
  #1      3934083          4101166
  table <- filter(table, Chromosome == "5" , From >= A, To <= B  )
  #rewmove these three columns
  table$X = NULL
  table$Accession = NULL
  table$X. = NULL
  #get the sum of other columns
  sum_of_size <- as.data.frame(t(apply(table, 2, sum)))
  #bind the sum_of_size with table and assign the name "sum_size_ as the row name of the
sum_of_size table
  table <- rbind(table, "sum_size" = sum_of_size)
  #in the table change the values of 1:7 column and sum_size row to "NA"
  table["sum_size", 1:7] <- "NA"
  # save each data frame as csv
  write.csv(table, file = paste("ThalianolMeg_" , i, sep = "_"))
}

##Total sizes
file_names<- list.files(path = ".")
for (i in file_names) {
  #read each data frame
  table <- read.csv(i, row.names = 1)
  #remove these three columns
  table$X = NULL
  table$Accession = NULL
  table$X. = NULL
  #get the sum of other columns
  sum_of_size <- as.data.frame(t(apply(table, 2, sum)))
  #bind the sum_of_size with table and assign the name "sum_size_ as the row name of the
sum_of_size table
  table <- rbind(table, "sum_size" = sum_of_size)
  #in the table change the values of 1:7 column and sum_size row to "NA"
  table["sum_size", 1:7] <- "NA"
  # save each data frame as csv
  write.csv(table, file = paste("Total_", i, sep = "_"))
}
```

7.4.5 Calculate percentage of marks and fold enrichement

```
##Percentages

sizebp <- read.csv("Size_of_regions.csv")
#peaks <- read.csv("Region_PeakSizes.csv")
peaks <- read.csv("New_RegionsPeakSizes.csv")

pROI <- NULL
pWhole <- NULL
enrich <- NULL
exp <- NULL

b <- 1

#(total marks across DNA)
m <- peaks[b, 8]

#(total size of DNA)
N <- sizebp[1 , 2]

n <-  N - m

#(size of region of interest)
```

```
k <- sizebp[1, 3]

#(marks across region of interest)
x <-  peaks[b, 7]

## Percent of marks across ROI
ROIperc <- ( x / k * 100)

## Percent of marks across whole genome
Genperc <- ( x / m * 100)

# expected number of red marbles
#k * m / (m + n)

#expmarks <- k * m / (m + n) * (m + n - k) / (m + n) * n / (m + n - 1)

#Fold enrichment
Enrichment <-  (x / k ) / (m / N)

## Random expectation
Expected <- m / N

pROI <- rbind(pROI , ROIperc)
pROI <- as.data.frame(pROI)
names(pROI) <- "New_Percent of marks across ROI_Thalianol"

pWhole <- rbind(pWhole , Genperc)
pWhole <- as.data.frame(pWhole)
names(pWhole) <- "New_Percent of marks across whole genome_Thalianol"

enrich <- rbind(enrich , Enrichment)
enrich <- as.data.frame(enrich)
```

```
names(enrich) <- "New_Fold enrichment_Thalianol"

exp <- rbind(exp , Expected)
exp <- as.data.frame(exp)
names(exp) <- "New_Expected_Thalianol"

#rownames(pROI) <- c("H2A.X_leaves_3w" , "H2A.Z_leaves_3w" , "H2A.Z_leaves" ,
"H2A.Z_seedling_10d_wt" , "H2A_leaves_3w" , "H2AK121ub_seedling_7d_1hp1" ,
"H2AK121ub_seedling_7d_atbmi" , "H2AK121ub_seedling_7d_clf.swn" ,
"H2AK121ub_seedling_7d_wt" , "H3ac_leaves_4w_wt" , "H3ac_roots_4w_wt" ,
"H3ac_roots_7d_hag1" , "H3ac_roots_7d_wt" , "H3K4me3_leaves" ,
"H3K4me3_roots_5d_mutant" , "H3K4me3_roots_5d_wt" , "H3K4me3_roots_14d_OTU5" ,
"H3K4me3_roots_14d_wt" , "H3K9me2_seedling_10d_wt" , "H3K14ac_seedling_10wt" ,
"H3K27me3_leaves_20d" , "H3K27me3_leaves" , "H3K27me3_roots_5d_mutant" ,
"H3K27me3_roots_5d_wt" , "H3K27me3_roots_14d_OTU5" , "H3K27me3_roots_14d_wt" ,
"H3K27me3_whole_30h" , "H3K36ac_leaves_35d_sdg8" , "H3K36ac_leaves_35d_wt" ,
"H3K36me3_leaves" , "H3K36me3_seedling_10d_wt")
#rownames(pWhole) <- c("H2A.X_leaves_3w" , "H2A.Z_leaves_3w" , "H2A.Z_leaves" ,
"H2A.Z_seedling_10d_wt" , "H2A_leaves_3w" , "H2AK121ub_seedling_7d_1hp1" ,
"H2AK121ub_seedling_7d_atbmi" , "H2AK121ub_seedling_7d_clf.swn" ,
"H2AK121ub_seedling_7d_wt" , "H3ac_leaves_4w_wt" , "H3ac_roots_4w_wt" ,
"H3ac_roots_7d_hag1" , "H3ac_roots_7d_wt" , "H3K4me3_leaves" ,
"H3K4me3_roots_5d_mutant" , "H3K4me3_roots_5d_wt" , "H3K4me3_roots_14d_OTU5" ,
"H3K4me3_roots_14d_wt" , "H3K9me2_seedling_10d_wt" , "H3K14ac_seedling_10wt" ,
"H3K27me3_leaves_20d" , "H3K27me3_leaves" , "H3K27me3_roots_5d_mutant" ,
"H3K27me3_roots_5d_wt" , "H3K27me3_roots_14d_OTU5" , "H3K27me3_roots_14d_wt" ,
"H3K27me3_whole_30h" , "H3K36ac_leaves_35d_sdg8" , "H3K36ac_leaves_35d_wt" ,
"H3K36me3_leaves" , "H3K36me3_seedling_10d_wt")
#rownames(enrich) <- c("H2A.X_leaves_3w" , "H2A.Z_leaves_3w" , "H2A.Z_leaves" ,
"H2A.Z_seedling_10d_wt" , "H2A_leaves_3w" , "H2AK121ub_seedling_7d_1hp1" ,
"H2AK121ub_seedling_7d_atbmi" , "H2AK121ub_seedling_7d_clf.swn" ,
"H2AK121ub_seedling_7d_wt" , "H3ac_leaves_4w_wt" , "H3ac_roots_4w_wt" ,
"H3ac_roots_7d_hag1" , "H3ac_roots_7d_wt" , "H3K4me3_leaves" ,
"H3K4me3_roots_5d_mutant" , "H3K4me3_roots_5d_wt" , "H3K4me3_roots_14d_OTU5" ,
"H3K4me3_roots_14d_wt" , "H3K9me2_seedling_10d_wt" , "H3K14ac_seedling_10wt" ,
"H3K27me3_leaves_20d" , "H3K27me3_leaves" , "H3K27me3_roots_5d_mutant" ,
"H3K27me3_roots_5d_wt" , "H3K27me3_roots_14d_OTU5" , "H3K27me3_roots_14d_wt" ,
"H3K27me3_whole_30h" , "H3K36ac_leaves_35d_sdg8" , "H3K36ac_leaves_35d_wt" ,
"H3K36me3_leaves" , "H3K36me3_seedling_10d_wt")
#rownames(exp) <- c("H2A.X_leaves_3w" , "H2A.Z_leaves_3w" , "H2A.Z_leaves" ,
"H2A.Z_seedling_10d_wt" , "H2A_leaves_3w" , "H2AK121ub_seedling_7d_1hp1" ,
"H2AK121ub_seedling_7d_atbmi" , "H2AK121ub_seedling_7d_clf.swn" ,
"H2AK121ub_seedling_7d_wt" , "H3ac_leaves_4w_wt" , "H3ac_roots_4w_wt" ,
"H3ac_roots_7d_hag1" , "H3ac_roots_7d_wt" , "H3K4me3_leaves" ,
"H3K4me3_roots_5d_mutant" , "H3K4me3_roots_5d_wt" , "H3K4me3_roots_14d_OTU5" ,
"H3K4me3_roots_14d_wt" , "H3K9me2_seedling_10d_wt" , "H3K14ac_seedling_10wt" ,
"H3K27me3_leaves_20d" , "H3K27me3_leaves" , "H3K27me3_roots_5d_mutant" ,
"H3K27me3_roots_5d_wt" , "H3K27me3_roots_14d_OTU5" , "H3K27me3_roots_14d_wt" ,
"H3K27me3_whole_30h" , "H3K36ac_leaves_35d_sdg8" , "H3K36ac_leaves_35d_wt" ,
"H3K36me3_leaves" , "H3K36me3_seedling_10d_wt")


#rownames(pROI) <- c("H1.Col-0seedling10d-H3-KD" , "H2A-Z.Col-0aerial-part12-13d-wt" ,
"H2A-Z.Col-0aerial-part12-14d-arp6" , "H2A-Z.Col-0aerial-part3w-clf-28" , "H2A-Z.Col-
0aerial-part3w-pie1-5" , "H2A-Z.Col-0aerial-part3w-pkl-1" , "H2A-Z.Col-0aerial-part3w-wt" ,
"H2A-Z.Col-0seedling6d-arp" , "H2A-Z.Col-0seedling6d-wt" , "H2AV.Col-0seedling12d-17c-
15min" , "H2AV.Col-0seedling12d-17c-1h" , "H2AV.Col-0seedling12d-27c-1h" , "H2AV.Col-
0seedling12d-37c-1h-exp" , "H2AV.Col-0seedling12d-37c-4h-exp" , "H2B.Col-0seedling12d-
17c-15min" , "H2B.Col-0seedling12d-17c-1h" , "H2B.Col-0seedling12d-27c-15min" , "H2B.Col-
0seedling12d-27c-4h" , "H3-1.Col-0leaves3w-A31T" , "H3-1.Col-0seedling10d" , "H3-3.Col-
0meristem-young-leaves4w" , "H3-3.Col-0old-leaves4w" , "H3-3.Col-0seedling10d" , "H3-
3.Col-0whole-plant2w-wt" , "H3.Col-0aerial-part3w-clf-28" , "H3.Col-0aerial-part3w-pie1-5" ,
"H3.Col-0aerial-part3w-pkl-1" , "H3.Col-0aerial-part6d-wt" , "H3.Col-0leaves-apical-
meristem5w-16c" , "H3.Col-0leaves-apical-meristem5w-25c" , "H3.Col-0rosette-leaves21d-
```

wt-watered" , "H3.Col-0seedling10-12d-wt" , "H3.Col-0seedling18d-wt" , "H3.Col-0seedling3w-wt" , "H3.Col-x-Lerendosperm20d-1" , "H3.Col-x-Lerendosperm20d-2" , "H3.Ler-x-Colendosperm20d-1" , "H3.Ler-x-Colendosperm20d-2" , "H3K14ac.Col-0seedling12d-wt" , "H3K18ac.Col-0seedling12d-wt" , "H3K23ac.Col-0seedling3d-wt-air" , "H3K27ac.Col-0seedling12d-wt" , "H3K27me1.Col-0leaves20d", "H3K27me1.Col-0rosette-leaves" , "H3K27me1.Col-x-Lerendosperm20d" , "H3K27me1.Ler-x-Colendosperm20d-2" , "H3K27me3.C24seedling10d-fen1" , "H3K27me3.C24seedling10d-pold2", "H3K27me3.C24seedling10d-wt" , "H3K27me3.C24seedling14d-clf" , "H3K27me3.C24seedling14d-wt" , "H3K27me3.Col-0aerial-part2w-atbmi" , "H3K27me3.Col-0aerial-part2w-atring" , "H3K27me3.Col-0aerial-part2w-clf" , "H3K27me3.Col-0aerial-part2w-lhp" , "H3K27me3.Col-0aerial-part2w-wt" , "H3K27me3.Col-0aerial-part3w-clf-28" , "H3K27me3.Col-0aerial-part3w-pie1-5" , "H3K27me3.Col-0aerial-part3w-pkl-1" , "H3K27me3.Col-0aerial-part3w-wt" , "H3K27me3.Col-0aerial-part6d-wt" , "H3K27me3.Col-0roots5d-wt" , "H3K27me3.Col-0seedling10d-ref6", "H3K27me3.Col-0seedling10d-wt" , "H3K27me3.Col-0seedling12d-wt" , "H3K27me3.Col-0seedling14d-clf" , "H3K27me3.Col-0seedling14d-wt" , "H3K27me3.Col-0seedling7d-atbmi" , "H3K27me3.Col-0seedling7d-clf-swn" , "H3K27me3.Col-0seedling7d-wt" , "H3K27me3.Col-x-C24seedling14d-clf" , "H3K27me3.Col-x-C24seedling14d-wt" , "H3K27me3.Col-x-Lerendosperm20d" , "H3K27me3.Lerpartial-inflorescence10d-mutant-Dex" , "H3K27me3.Lerpartial-inflorescence10d-mutant" , "H3K27me3.Lerrosette-leaves10d-wt" , "H3K27me3.Lerwhole-inflorescence10d-wt" , "H3K36me3.Col-0leaves3w-wt" , "H3K4me3.C24seedling10d-fen1" , "H3K4me3.C24seedling10d-pold2" , "H3K4me3.C24seedling10d-wt" , "H3K4me3.Col-0rosette-leaves21d-mutant-watered" , "H3K4me3.Col-0seedling10d-wt" , "H3K4me3.Lerpartial-inflorescence10d-mutant-Dex" , "H3K4me3.Lerpartial-inflorescence10d-mutant" , "H3K4me3.Lerrosette-leaves10d-wt" , "H3K4me3.Lerwhole-inflorescence10d-wt" , "H3K56ac.Col-0leaves" , "H3K9-14ac.Col-0seedling7d-water-3h" , "H3K9-14ac.Col-0seedling7d-water-6h" , "H3K9ac.Col-0rosette-leaves30d" , "H3K9ac.Col-0rosette-leaves34d" , "H3K9ac.Col-0rosette-leaves42d" , "H3K9ac.Col-0seedling10d-pwr2" , "H3K9ac.Col-0seedling10d-wt" , "H3K9ac.Col-0seedling12d-wt" , "H3K9ac.Col-0seedling3d-air" , "H3K9me2.C24seedling10d-pold2" , "H3K9me2.C24seedling10d-wt" , "H3K9me2.Col-0aerial-part6d-wt" , "H3K9me2.Col-x-Lerendosperm20d" , "H3K9me2.Ler-x-Colendosperm20d" , "H3T3ph.Col-0rosette-leaves21d-wt-watered" , "H4ac.Col-0seedling6d-

```
arp" , "H4ac.Col-0seedling6d-wt" , "H4K12ac.Col-0seedling12d-wt" , "H4K5ac.Col-
0seedling12d-wt" , "H4K5ac.undefseedling10d-wt" , "H4K8ac.Col-0seedling12d-wt" ,
"HTR12.Col-0seedling7-10d-At" , "HTR12.Col-0seedling7-10d-Lo" , "HTR12.Col-0seedling7-
10d-Zm")
#rownames(pWhole) <- c("H1.Col-0seedling10d-H3-KD" , "H2A-Z.Col-0aerial-part12-13d-wt"
, "H2A-Z.Col-0aerial-part12-14d-arp6" , "H2A-Z.Col-0aerial-part3w-clf-28" , "H2A-Z.Col-
0aerial-part3w-pie1-5" , "H2A-Z.Col-0aerial-part3w-pkl-1" , "H2A-Z.Col-0aerial-part3w-wt" ,
"H2A-Z.Col-0seedling6d-arp" , "H2A-Z.Col-0seedling6d-wt" , "H2AV.Col-0seedling12d-17c-
15min" , "H2AV.Col-0seedling12d-17c-1h" , "H2AV.Col-0seedling12d-27c-1h" , "H2AV.Col-
0seedling12d-37c-1h-exp" , "H2AV.Col-0seedling12d-37c-4h-exp" , "H2B.Col-0seedling12d-
17c-15min" , "H2B.Col-0seedling12d-17c-1h" , "H2B.Col-0seedling12d-27c-15min" , "H2B.Col-
0seedling12d-27c-4h" , "H3-1.Col-0leaves3w-A31T" , "H3-1.Col-0seedling10d" , "H3-3.Col-
0meristem-young-leaves4w" , "H3-3.Col-0old-leaves4w" , "H3-3.Col-0seedling10d" , "H3-
3.Col-0whole-plant2w-wt" , "H3.Col-0aerial-part3w-clf-28" , "H3.Col-0aerial-part3w-pie1-5" ,
"H3.Col-0aerial-part3w-pkl-1" , "H3.Col-0aerial-part6d-wt" , "H3.Col-0leaves-apical-
meristem5w-16c" , "H3.Col-0leaves-apical-meristem5w-25c" , "H3.Col-0rosette-leaves21d-
wt-watered" , "H3.Col-0seedling10-12d-wt" , "H3.Col-0seedling18d-wt" , "H3.Col-
0seedling3w-wt" , "H3.Col-x-Lerendosperm20d-1" , "H3.Col-x-Lerendosperm20d-2" , "H3.Ler-
x-Colendosperm20d-1" , "H3.Ler-x-Colendosperm20d-2" , "H3K14ac.Col-0seedling12d-wt" ,
"H3K18ac.Col-0seedling12d-wt" , "H3K23ac.Col-0seedling3d-wt-air" , "H3K27ac.Col-
0seedling12d-wt" , "H3K27me1.Col-0leaves20d", "H3K27me1.Col-0rosette-leaves" ,
"H3K27me1.Col-x-Lerendosperm20d" , "H3K27me1.Ler-x-Colendosperm20d-2" ,
"H3K27me3.C24seedling10d-fen1" , "H3K27me3.C24seedling10d-pold2",
"H3K27me3.C24seedling10d-wt" , "H3K27me3.C24seedling14d-clf" ,
"H3K27me3.C24seedling14d-wt" , "H3K27me3.Col-0aerial-part2w-atbmi" , "H3K27me3.Col-
0aerial-part2w-atring" , "H3K27me3.Col-0aerial-part2w-clf" , "H3K27me3.Col-0aerial-
part2w-lhp" , "H3K27me3.Col-0aerial-part2w-wt" , "H3K27me3.Col-0aerial-part3w-clf-28" ,
"H3K27me3.Col-0aerial-part3w-pie1-5" , "H3K27me3.Col-0aerial-part3w-pkl-1" ,
"H3K27me3.Col-0aerial-part3w-wt" , "H3K27me3.Col-0aerial-part6d-wt" , "H3K27me3.Col-
0roots5d-wt" , "H3K27me3.Col-0seedling10d-ref6", "H3K27me3.Col-0seedling10d-wt" ,
"H3K27me3.Col-0seedling12d-wt" , "H3K27me3.Col-0seedling14d-clf" , "H3K27me3.Col-
0seedling14d-wt" , "H3K27me3.Col-0seedling7d-atbmi" , "H3K27me3.Col-0seedling7d-clf-
```

```
swn" , "H3K27me3.Col-0seedling7d-wt" , "H3K27me3.Col-x-C24seedling14d-clf" , "H3K27me3.Col-x-C24seedling14d-wt" , "H3K27me3.Col-x-Lerendosperm20d" , "H3K27me3.Lerpartial-inflorescence10d-mutant-Dex" , "H3K27me3.Lerpartial-inflorescence10d-mutant" , "H3K27me3.Lerrosette-leaves10d-wt" , "H3K27me3.Lerwhole-inflorescence10d-wt" , "H3K36me3.Col-0leaves3w-wt" , "H3K4me3.C24seedling10d-fen1" , "H3K4me3.C24seedling10d-pold2" , "H3K4me3.C24seedling10d-wt" , "H3K4me3.Col-0rosette-leaves21d-mutant-watered" , "H3K4me3.Col-0seedling10d-wt" , "H3K4me3.Lerpartial-inflorescence10d-mutant-Dex" , "H3K4me3.Lerpartial-inflorescence10d-mutant" , "H3K4me3.Lerrosette-leaves10d-wt" , "H3K4me3.Lerwhole-inflorescence10d-wt" , "H3K56ac.Col-0leaves" , "H3K9-14ac.Col-0seedling7d-water-3h" , "H3K9-14ac.Col-0seedling7d-water-6h" , "H3K9ac.Col-0rosette-leaves30d" , "H3K9ac.Col-0rosette-leaves34d" , "H3K9ac.Col-0rosette-leaves42d" , "H3K9ac.Col-0seedling10d-pwr2" , "H3K9ac.Col-0seedling10d-wt" , "H3K9ac.Col-0seedling12d-wt" , "H3K9ac.Col-0seedling3d-air" , "H3K9me2.C24seedling10d-pold2" , "H3K9me2.C24seedling10d-wt" , "H3K9me2.Col-0aerial-part6d-wt" , "H3K9me2.Col-x-Lerendosperm20d" , "H3K9me2.Ler-x-Colendosperm20d" , "H3T3ph.Col-0rosette-leaves21d-wt-watered" , "H4ac.Col-0seedling6d-arp" , "H4ac.Col-0seedling6d-wt" , "H4K12ac.Col-0seedling12d-wt" , "H4K5ac.Col-0seedling12d-wt" , "H4K5ac.undefseedling10d-wt" , "H4K8ac.Col-0seedling12d-wt" , "HTR12.Col-0seedling7-10d-At" , "HTR12.Col-0seedling7-10d-Lo" , "HTR12.Col-0seedling7-10d-Zm")
#rownames(enrich) <- c("H1.Col-0seedling10d-H3-KD" , "H2A-Z.Col-0aerial-part12-13d-wt" , "H2A-Z.Col-0aerial-part12-14d-arp6" , "H2A-Z.Col-0aerial-part3w-clf-28" , "H2A-Z.Col-0aerial-part3w-pie1-5" , "H2A-Z.Col-0aerial-part3w-pkl-1" , "H2A-Z.Col-0aerial-part3w-wt" , "H2A-Z.Col-0seedling6d-arp" , "H2A-Z.Col-0seedling6d-wt" , "H2AV.Col-0seedling12d-17c-15min" , "H2AV.Col-0seedling12d-17c-1h" , "H2AV.Col-0seedling12d-27c-1h" , "H2AV.Col-0seedling12d-37c-1h-exp" , "H2AV.Col-0seedling12d-37c-4h-exp" , "H2B.Col-0seedling12d-17c-15min" , "H2B.Col-0seedling12d-17c-1h" , "H2B.Col-0seedling12d-27c-15min" , "H2B.Col-0seedling12d-27c-4h" , "H3-1.Col-0leaves3w-A31T" , "H3-1.Col-0seedling10d" , "H3-3.Col-0meristem-young-leaves4w" , "H3-3.Col-0old-leaves4w" , "H3-3.Col-0seedling10d" , "H3-3.Col-0whole-plant2w-wt" , "H3.Col-0aerial-part3w-clf-28" , "H3.Col-0aerial-part3w-pie1-5" , "H3.Col-0aerial-part3w-pkl-1" , "H3.Col-0aerial-part6d-wt" , "H3.Col-0leaves-apical-meristem5w-16c" , "H3.Col-0leaves-apical-meristem5w-25c" , "H3.Col-0rosette-leaves21d-
```

wt-watered" , "H3.Col-0seedling10-12d-wt" , "H3.Col-0seedling18d-wt" , "H3.Col-0seedling3w-wt" , "H3.Col-x-Lerendosperm20d-1" , "H3.Col-x-Lerendosperm20d-2" , "H3.Ler-x-Colendosperm20d-1" , "H3.Ler-x-Colendosperm20d-2" , "H3K14ac.Col-0seedling12d-wt" , "H3K18ac.Col-0seedling12d-wt" , "H3K23ac.Col-0seedling3d-wt-air" , "H3K27ac.Col-0seedling12d-wt" , "H3K27me1.Col-0leaves20d", "H3K27me1.Col-0rosette-leaves" , "H3K27me1.Col-x-Lerendosperm20d" , "H3K27me1.Ler-x-Colendosperm20d-2" , "H3K27me3.C24seedling10d-fen1" , "H3K27me3.C24seedling10d-pold2", "H3K27me3.C24seedling10d-wt" , "H3K27me3.C24seedling14d-clf" , "H3K27me3.C24seedling14d-wt" , "H3K27me3.Col-0aerial-part2w-atbmi" , "H3K27me3.Col-0aerial-part2w-atring" , "H3K27me3.Col-0aerial-part2w-clf" , "H3K27me3.Col-0aerial-part2w-lhp" , "H3K27me3.Col-0aerial-part2w-wt" , "H3K27me3.Col-0aerial-part3w-clf-28" , "H3K27me3.Col-0aerial-part3w-pie1-5" , "H3K27me3.Col-0aerial-part3w-pkl-1" , "H3K27me3.Col-0aerial-part3w-wt" , "H3K27me3.Col-0aerial-part6d-wt" , "H3K27me3.Col-0roots5d-wt" , "H3K27me3.Col-0seedling10d-ref6", "H3K27me3.Col-0seedling10d-wt" , "H3K27me3.Col-0seedling12d-wt" , "H3K27me3.Col-0seedling14d-clf" , "H3K27me3.Col-0seedling14d-wt" , "H3K27me3.Col-0seedling7d-atbmi" , "H3K27me3.Col-0seedling7d-clf-swn" , "H3K27me3.Col-0seedling7d-wt" , "H3K27me3.Col-x-C24seedling14d-clf" , "H3K27me3.Col-x-C24seedling14d-wt" , "H3K27me3.Col-x-Lerendosperm20d" , "H3K27me3.Lerpartial-inflorescence10d-mutant-Dex" , "H3K27me3.Lerpartial-inflorescence10d-mutant" , "H3K27me3.Lerrosette-leaves10d-wt" , "H3K27me3.Lerwhole-inflorescence10d-wt" , "H3K36me3.Col-0leaves3w-wt" , "H3K4me3.C24seedling10d-fen1" , "H3K4me3.C24seedling10d-pold2" , "H3K4me3.C24seedling10d-wt" , "H3K4me3.Col-0rosette-leaves21d-mutant-watered" , "H3K4me3.Col-0seedling10d-wt" , "H3K4me3.Lerpartial-inflorescence10d-mutant-Dex" , "H3K4me3.Lerpartial-inflorescence10d-mutant" , "H3K4me3.Lerrosette-leaves10d-wt" , "H3K4me3.Lerwhole-inflorescence10d-wt" , "H3K56ac.Col-0leaves" , "H3K9-14ac.Col-0seedling7d-water-3h" , "H3K9-14ac.Col-0seedling7d-water-6h" , "H3K9ac.Col-0rosette-leaves30d" , "H3K9ac.Col-0rosette-leaves34d" , "H3K9ac.Col-0rosette-leaves42d" , "H3K9ac.Col-0seedling10d-pwr2" , "H3K9ac.Col-0seedling10d-wt" , "H3K9ac.Col-0seedling12d-wt" , "H3K9ac.Col-0seedling3d-air" , "H3K9me2.C24seedling10d-pold2" , "H3K9me2.C24seedling10d-wt" , "H3K9me2.Col-0aerial-part6d-wt" , "H3K9me2.Col-x-Lerendosperm20d" , "H3K9me2.Ler-x-Colendosperm20d" , "H3T3ph.Col-0rosette-leaves21d-wt-watered" , "H4ac.Col-0seedling6d-

```
arp" , "H4ac.Col-0seedling6d-wt" , "H4K12ac.Col-0seedling12d-wt" , "H4K5ac.Col-
0seedling12d-wt" , "H4K5ac.undefseedling10d-wt" , "H4K8ac.Col-0seedling12d-wt" ,
"HTR12.Col-0seedling7-10d-At" , "HTR12.Col-0seedling7-10d-Lo" , "HTR12.Col-0seedling7-
10d-Zm")
#rownames(exp) <- c("H1.Col-0seedling10d-H3-KD" , "H2A-Z.Col-0aerial-part12-13d-wt" ,
"H2A-Z.Col-0aerial-part12-14d-arp6" , "H2A-Z.Col-0aerial-part3w-clf-28" , "H2A-Z.Col-
0aerial-part3w-pie1-5" , "H2A-Z.Col-0aerial-part3w-pkl-1" , "H2A-Z.Col-0aerial-part3w-wt" ,
"H2A-Z.Col-0seedling6d-arp" , "H2A-Z.Col-0seedling6d-wt" , "H2AV.Col-0seedling12d-17c-
15min" , "H2AV.Col-0seedling12d-17c-1h" , "H2AV.Col-0seedling12d-27c-1h" , "H2AV.Col-
0seedling12d-37c-1h-exp" , "H2AV.Col-0seedling12d-37c-4h-exp" , "H2B.Col-0seedling12d-
17c-15min" , "H2B.Col-0seedling12d-17c-1h" , "H2B.Col-0seedling12d-27c-15min" , "H2B.Col-
0seedling12d-27c-4h" , "H3-1.Col-0leaves3w-A31T" , "H3-1.Col-0seedling10d" , "H3-3.Col-
0meristem-young-leaves4w" , "H3-3.Col-0old-leaves4w" , "H3-3.Col-0seedling10d" , "H3-
3.Col-0whole-plant2w-wt" , "H3.Col-0aerial-part3w-clf-28" , "H3.Col-0aerial-part3w-pie1-5" ,
"H3.Col-0aerial-part3w-pkl-1" , "H3.Col-0aerial-part6d-wt" , "H3.Col-0leaves-apical-
meristem5w-16c" , "H3.Col-0leaves-apical-meristem5w-25c" , "H3.Col-0rosette-leaves21d-
wt-watered" , "H3.Col-0seedling10-12d-wt" , "H3.Col-0seedling18d-wt" , "H3.Col-
0seedling3w-wt" , "H3.Col-x-Lerendosperm20d-1" , "H3.Col-x-Lerendosperm20d-2" , "H3.Ler-
x-Colendosperm20d-1" , "H3.Ler-x-Colendosperm20d-2" , "H3K14ac.Col-0seedling12d-wt" ,
"H3K18ac.Col-0seedling12d-wt" , "H3K23ac.Col-0seedling3d-wt-air" , "H3K27ac.Col-
0seedling12d-wt" , "H3K27me1.Col-0leaves20d", "H3K27me1.Col-0rosette-leaves" ,
"H3K27me1.Col-x-Lerendosperm20d" , "H3K27me1.Ler-x-Colendosperm20d-2" ,
"H3K27me3.C24seedling10d-fen1" , "H3K27me3.C24seedling10d-pold2",
"H3K27me3.C24seedling10d-wt" , "H3K27me3.C24seedling14d-clf" ,
"H3K27me3.C24seedling14d-wt" , "H3K27me3.Col-0aerial-part2w-atbmi" , "H3K27me3.Col-
0aerial-part2w-atring" , "H3K27me3.Col-0aerial-part2w-clf" , "H3K27me3.Col-0aerial-
part2w-lhp" , "H3K27me3.Col-0aerial-part2w-wt" , "H3K27me3.Col-0aerial-part3w-clf-28" ,
"H3K27me3.Col-0aerial-part3w-pie1-5" , "H3K27me3.Col-0aerial-part3w-pkl-1" ,
"H3K27me3.Col-0aerial-part3w-wt" , "H3K27me3.Col-0aerial-part6d-wt" , "H3K27me3.Col-
0roots5d-wt" , "H3K27me3.Col-0seedling10d-ref6", "H3K27me3.Col-0seedling10d-wt" ,
"H3K27me3.Col-0seedling12d-wt" , "H3K27me3.Col-0seedling14d-clf" , "H3K27me3.Col-
0seedling14d-wt" , "H3K27me3.Col-0seedling7d-atbmi" , "H3K27me3.Col-0seedling7d-clf-
```

```
swn" , "H3K27me3.Col-0seedling7d-wt" , "H3K27me3.Col-x-C24seedling14d-clf" ,
"H3K27me3.Col-x-C24seedling14d-wt" , "H3K27me3.Col-x-Lerendosperm20d" ,
"H3K27me3.Lerpartial-inflorescence10d-mutant-Dex" ,
"H3K27me3.Lerpartial-inflorescence10d-mutant" , "H3K27me3.Lerrosette-leaves10d-wt" ,
"H3K27me3.Lerwhole-inflorescence10d-wt" , "H3K36me3.Col-0leaves3w-wt" , "H3K4me3.C24seedling10d-fen1" ,
"H3K4me3.C24seedling10d-pold2" , "H3K4me3.C24seedling10d-wt" ,
"H3K4me3.Col-0rosette-leaves21d-mutant-watered" , "H3K4me3.Col-0seedling10d-wt" ,
"H3K4me3.Lerpartial-inflorescence10d-mutant-Dex" ,
"H3K4me3.Lerpartial-inflorescence10d-mutant" , "H3K4me3.Lerrosette-leaves10d-wt" ,
"H3K4me3.Lerwhole-inflorescence10d-wt" , "H3K56ac.Col-0leaves" , "H3K9-14ac.Col-0seedling7d-water-3h" ,
"H3K9-14ac.Col-0seedling7d-water-6h" , "H3K9ac.Col-0rosette-leaves30d" ,
"H3K9ac.Col-0rosette-leaves34d" , "H3K9ac.Col-0rosette-leaves42d" , "H3K9ac.Col-0seedling10d-pwr2" ,
"H3K9ac.Col-0seedling10d-wt" , "H3K9ac.Col-0seedling12d-wt" ,
"H3K9ac.Col-0seedling3d-air" , "H3K9me2.C24seedling10d-pold2" , "H3K9me2.C24seedling10d-wt" ,
"H3K9me2.Col-0aerial-part6d-wt" , "H3K9me2.Col-x-Lerendosperm20d" ,
"H3K9me2.Ler-x-Colendosperm20d" , "H3T3ph.Col-0rosette-leaves21d-wt-watered" ,
"H4ac.Col-0seedling6d-arp" , "H4ac.Col-0seedling6d-wt" , "H4K12ac.Col-0seedling12d-wt" ,
"H4K5ac.Col-0seedling12d-wt" , "H4K5ac.undefseedling10d-wt" , "H4K8ac.Col-0seedling12d-wt" ,
"HTR12.Col-0seedling7-10d-At" , "HTR12.Col-0seedling7-10d-Lo" , "HTR12.Col-0seedling7-10d-Zm")

#write.csv(pROI, "Percent of marks across ROI_Thalianol.csv")
#write.csv(pWhole,"Percent of marks across whole genome_Thalianol.csv" )
#write.csv(enrich, "Fold enrichment_Thalianol.csv")
#write.csv(exp, "Expected_Thalianol.csv")
```

7.4.6 Fishers test

```
##Statistical tests
setwd("~/Remap data"")
library("dplyr")
library("Rmpfr"")
```

```
sizebp <- read.csv("Size_of_regions.csv")
peaks <- read.csv("Region_PeakSizes.csv")


##Fishers test

out <- NULL

#Thalianol
b <- 3

#Random


#(total marks across DNA)
m <- peaks[b, 6]

#(total size of DNA)
N <- sizebp[1 , 3]

n <-  N - m

#(size of region of interest)
k <- sizebp[1, 5]

#(marks across region of interest)
x <-  peaks[b, 2]

contingency.table <- data.frame(matrix(nrow=2, ncol=2))
rownames(contingency.table) <- c("predicted.target", "non.predicted")
colnames(contingency.table) <- c("class.member", "non.member")
```

```
## Assign the values one by one to make sure we put them in the right
## place (this is not necessary, we could enter the 4 values in a
## single instruction).
contingency.table["predicted.target", "class.member"] <- x ## Number of marked genes in the selection
contingency.table["predicted.target", "non.member"] <- k - x ## Number of non-marked genes in the selection
contingency.table["non.predicted", "class.member"] <- m - x ## Number of marked genes outside of the selection
contingency.table["non.predicted", "non.member"] <- n - (k - x) ## Number of non-marked genes in the selection

print(contingency.table)

##                class.member non.member
## predicted.target          19        40
## non.predicted            592     12937
## Print marginal sums
(contingency.row.sum <- apply(contingency.table, 1, sum))
## predicted.target    non.predicted
##             59           13529
(contingency.col.sum <- apply(contingency.table, 2, sum))
## class.member   non.member
##         611        12977
## Create a contingency table with marginal sums
contingency.table.margins <- cbind(contingency.table, contingency.row.sum)
contingency.table.margins <- rbind(contingency.table.margins, apply(contingency.table.margins, 2, sum))
names(contingency.table.margins) <- c(names(contingency.table), "total")
rownames(contingency.table.margins) <- c(rownames(contingency.table), "total")
print(contingency.table.margins)
```

```
##                class.member non.member total
## predicted.target          19        40    59
## non.predicted           592      12937 13529
## total                   611      12977 13588
## Check the total
print(sum(contingency.table)) ## The value should equal N, since every
## [1] 13588
## possible gene must be assigned to one
## cell of the contingency table.
print(N)
## [1] 13588
## Run Fisher's exact test
ftest.result <- fisher.test(x=contingency.table, alternative="greater", simulate.p.value=TRUE)
print(ftest.result)

print(x)

pvals <- fisher.test(x=contingency.table, alternative="greater",
simulate.p.value=TRUE)$p.value
pvals
c <- format.pval(pvals)
c <- as.data.frame.integer(c)
names(c) <- "ThalianolvRandom"
c


out <- rbind(out,c)
out <- as.data.frame(out)
names(out) <- "ThalianolvRandom"
```

```
rownames(out) <- c("H2A.X_leaves_3w" , "H2A.Z_leaves_3w" , "H2A.Z_leaves" ,
"H2A.Z_seedling_10d_wt" , "H2A_leaves_3w" , "H2AK121ub_seedling_7d_1hp1" ,
            "H2AK121ub_seedling_7d_atbmi" , "H2AK121ub_seedling_7d_clf.swn" ,
"H2AK121ub_seedling_7d_wt" , "H3ac_leaves_4w_wt"
            , "H3ac_roots_4w_wt" , "H3ac_roots_7d_hag1" , "H3ac_roots_7d_wt" ,
"H3K4me3_leaves" , "H3K4me3_roots_5d_mutant" ,
            "H3K4me3_roots_5d_wt" , "H3K4me3_roots_14d_OTU5" ,
"H3K4me3_roots_14d_wt" , "H3K9me2_seedling_10d_wt" , "H3K14ac_seedling_10wt" ,
"H3K27me3_leaves_20d" ,
            "H3K27me3_leaves" , "H3K27me3_roots_5d_mutant" ,
"H3K27me3_roots_5d_wt" , "H3K27me3_roots_14d_OTU5" , "H3K27me3_roots_14d_wt" ,
            "H3K27me3_whole_30h" , "H3K36ac_leaves_35d_sdg8" ,
"H3K36ac_leaves_35d_wt" , "H3K36me3_leaves" , "H3K36me3_seedling_10d_wt")

write.csv(out, "Thalianol_fishers.csv")

chisq.test(x=contingency.table)
?fisher.test()
```

7.5 FASTA files

Thalianol cluster sequences (file://DESKTOP-TUVGVRQ/Users/jb2980/Documents/Bioinformatics/Thalianol%20cluster%20sequences)

Marneral cluster sequences (file://DESKTOP-TUVGVRQ/Users/jb2980/Documents/Bioinformatics/Marneral%20cluster%20sequences)

Arabidol cluster sequences (file://DESKTOP-TUVGVRQ/Users/jb2980/Documents/Bioinformatics/Arabidol%20cluster%20sequences)