METHODS FOR PLANT-BASED GENOME AND EPIGENOME EDITING

by

WILLIAM THOMAS JORDAN

(Under the Direction of Robert J. Schmitz)

ABSTRACT

An understanding of gene expression is essential to elucidate how altered phenotypes arise in eukaryotic organisms. Importantly, modifications not directly altering the DNA sequence of genes can alter gene expression in profound ways and can result in the emergence of novel phenotypes. Such modifications often are caused by changes to the epigenome, or through the alteration of cis-regulatory elements (CREs). DNA methylation, a modification to the DNA base cytosine, is a primary component of the epigenome, and can serve to alter gene expression of genes, particularly when found in promoter sequences. Alterations to cis-regulatory elements can result in altered transcription factor binding at the promoters of target genes, resulting in altered expression patterns and the emergence of novel phenotypes.

My PhD research focuses on creating new methods to intentionally alter cis-regulatory elements and DNA methylation, with the goal of creating novel phenotypes from otherwise genetically identical individuals. To this end, I was the lead molecular biologist for the development of epimutagenesis, a method whereby the expression of a mammalian demethylase protein in *Arabidopsis thalian*a stochastically removes DNA methylation throughout the genome, and results in the emergence of novel phenotypes, some of which are stably inherited over generational time. Next, I sought to create a simplified method for creating genetic variation

at CREs by leveraging the multiplexing ability of Cas12a proteins. To this end, I created a heat-shock protocol for the efficient transmission of mutated alleles through the germline of *A. thaliana*, plants and successfully applied this method for multiplex engineering of a CRE with altered DNA sequences. Finally, I sought to leverage the development of new CRISPR RNA-editing technology for the application of silencer element detection in a genome-wide fashion. Taken together, my research has resulted in the establishment and successful application of two methods for both DNA demethylation and CRE editing in *A. thaliana*.

INDEX WORDS: Epigenetics, CRISPR, DNA Methylation, Arabidopsis, Genome Engineering, Epigenome Engineering

METHODS FOR PLANT-BASED GENOME AND EPIGENOME EDITING

by

WILLIAM THOMAS JORDAN

B.S, Hofstra University, 2015

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

DOCTOR OF PHILOSOPHY

ATHENS, GEORGIA

2020

METHODS FOR PLANT-BASED GENOME AND EPIGENOME EDITING

by

WILLIAM THOMAS JORDAN

Major Professor:	Robert Schmitz
Committee:	Zachary Lewis
	Wayne Parrott
	Douglas Menke

Electronic Version Approved:

Ron Walcott
Dean of the Graduate School
The University of Georgia
December 2020

TABLE OF CONTENTS

Page

LIST OF TABLES

Page

LIST OF FIGURES

Page

CHAPTER 1

A REVIEW OF EPIGENOME AND GENOME EDITING IN PLANTS

The creation of engineered strains of commercially important crops such as maize, squash, and cotton has resulted in vastly increased crop yields as well as resistance to pests and herbicides. While these hybrids and cultivars are ideal for sustaining a rapidly expanding global population, the initial prioritization of homogeneity and productivity over reliable and diversified food production resulted in a severe loss of genetic variation by the mid 20th century, which is estimated at nearly 75% by the UN Sustainable Development Knowledge Program [1]. Such loss of genetic diversity resulted in a narrow genetic base of traditional staple crops. Such genetic vulnerability limits the ability of crops to adapt to stressful environments and predisposes them to new strains of diseases and other pests. Traditional approaches for reintroducing variation into crop species involve crossbreeding with landraces and undomesticated wild ancestral species, which possess large amounts of genetic diversity. This approach has been widely implemented since the 1940s, with an estimated 80% of traits conferred from crossbreeding providing resistance to pests and disease. However, locating germplasm with desired traits can be a costly and time-consuming process, and crossbreeding with genetically distant plants can be very difficult. Furthermore, such species are often ignored in the plant research community due their lack of known agriculturally advantageous traits. Additionally, crossbreeding with wild relatives can bring in large amounts of unadapted alleles that traditionally were difficult to breed out, resulting in linkage drag, i.e., the incorporation of unwanted traits from the wild ancestor that may decrease agricultural yield, reduce quality, or even produce unknown disease

sensitivity. While the introduction of single or a few transgenes with desired traits by genetic engineering circumvents this problem, these Genetically Modified Organisms (GMOs), have recently lost favor in the consumer market. The development of methods to introduce variation into crop species without the need for crossbreeding or genetic engineering would potentially enable the creation of non-transgenic crop lines suitable for agricultural production that could thrive in changing environmental conditions and resist outbreaks of pathogens and pests.

Emerging technologies for genome and epigenome engineering without the incorporation of "foreign" transgenes into crops, have the potential to introduce considerable phenotypic variation into genetically homogenous plant populations and represents an additional opportunity for crop improvement. Here, I describe the current state of genome and epigenome engineering tools in plants.

DNA Methylation and demethylation systems in *Arabidopsis thaliana*

DNA methylation is a covalent modification to cytosine nucleotides and is a key component of a wide array of biological processes in plants including: genomic imprinting, establishment and maintenance of transposon silencing, and genome stability [2]. Alterations to patterns in DNA methylation have the potential to greatly affect plant phenotypes, such as known abnormalities in the mantling of oil palm, the improper ripening of tomatoes, and the height of rice plants [3,4,5]. As these phenotypes are solely the result of alterations of DNA methylation and not DNA sequence, biological pathways to ensure the proper deposition and maintenance of DNA methylation are critical for plant fitness and fecundity.

In plants, DNA methylation can be separated into three distinct sequence contexts; the sequence context of a given cytosine determines the biological pathway used to further maintain methylation after subsequent cell divisions. Methylated cytosines occurring in a 5′ CG 3′

sequence context is maintained by the combined actions of DNA METHYLTRANSFERASE 1 (MET1) and the VARIANT IN METHYLATION (VIM*)* family of methylcytosine binding proteins [6,7]. As the 5′ CG 3′ sequence context of methylation is symmetrical across both DNA strands, proper methylation of newly synthesized daughter DNA strands is accomplished by 'reading' the corresponding methylation state of the parent strand. CG methylation is the prevalent sequence context of DNA methylation across the *A. thaliana* genome, occurring in genic as well as heterochromatic regions, as up to 20% of 5′ CG 3′ sequence contexts genome-wide can be methylated in leaf tissue [8].

Methylation occurring in a 5′ CHG 3′ sequence, where H is A, T, or C, contexts is maintained by the combined actions of CHROMOMETHYLASE 3 (CMT3) and SUPPRESSOR OF VARIEGATION 3-9 HOMOLOGUE 4 (SUVH4) histone lysine methyltransferase [9,10]. The recruitment of CMT3 to DNA is highly dependent on levels of methylation on the ninth lysine residue on histone H3 (H3K9me2), which is the enzymatic product of SUVH4 [11]. As SUVH4 is recruited to DNA by the presence of CHG methylation, the combined actions of CMT3 and SUVH4 form a feed-forward loop, ensuring proper maintenance of CHG methylation and H3K9me2 through cellular divisions. Methylation occurring in a 5′ CHH 3′ context is maintained by two distinct pathways. Methylated cytosines in a 5′ CWA 3′ context found in constitutive heterochromatin is maintained by CHROMOMETHYLASE 2 (CMT2) [12]. Similar to the actions of CMT3, the combined actions of CMT2 and SUVH4 result in a feed-forward loop, ensuring proper methylation of cytosines at 5′ CWA 3′ sequence contexts in constitutive heterochromatin [13].

The establishment of methylation in all sequence contexts and maintenance of 5′ CHH 3′ methylation at genomic regions which are not targeted by CMT2, such as euchromatin, is

maintained and deposited by the *de novo* cytosine methyltransferase DOMAINS REARRANGED METHYLTRANSFERASE 2 (DRM2) [14]. The recruitment of DRM2 to DNA is accomplished through an intricate plant-specific pathway known as RNA-directed DNA methylation (RdDM). Briefly, plant-specific RNA polymerases (Pol IV, Pol V) produce long single-stranded RNA transcripts, which serve as templates for RNA-DEPENDENT POLYMERASE 2 (RDR2*)* to produce dsRNA [15]. The dsRNA products of RDR2 are cleaved by the endoribonuclease DICER-LIKE 3 (DCL3*)*, to produce 24 nucleotide siRNA transcripts [16]. These transcripts are deposited into ARGONAUTE 4 (AGO4*)*, which guide DRM2 to deposit *de novo* methylation at a given genomic regions complementary to these siRNAs [16].

The active removal of DNA methylation in *A. thaliana* is accomplished via a family of bifunctional DNA glycosylases-apurinic/apyrimidinic lyases, which specifically remove 5-methylcytosines irrespective of sequence context via a two-step mechanism resulting in the formation of an abasic site [17]. Four known bifunctional 5-mC DNA glycosylases exist in *Arabidopsis*, REPRESSOR OF SILENCING 1 (ROS1), DEMETER (DME), and DEMETER-LIKE PROTEIN 2 and 3 (DML2 and DML3) [17,18]. The expression and demethylating activity of DME is limited to the central egg cell and vegetative pollen cells, while that of ROS1, DML2 and DML3 occur in somatic tissue [19].

Importance of Epialleles to phenotypes

As plants inherit cytosine methylation meiotically as well as mitotically, a Mendelian pattern of inheritance occurs at methylated cytosines. Importantly, the rate of epimutation, or failure of proper inheritance of methylated regions over generational time, has been measured to be as low as 0.002% per generation in *A. thaliana* [20]. Due to the stability of cytosine methylation over generational time as well as its importance for proper gene regulation, the alteration of

cytosine methylation has the potential to create novel phenotypic variation without changes in any underlying DNA sequences.

Altered plant phenotypes resulting from differential methylation patterns have been documented for at least 275 years, as the *peloric* mutant in *Linaria vulgaris*, first documented by Carl Linnaeus in 1749, produces a highly altered flower morphology. *Peloric* was found to result from hypermethylation and subsequent silencing of transcription factor responsible for flower organ symmetry [21]. The dwarf phenotype associated with the *epi-d1* dwarf rice line, which has served as an important breeding line for over 100 years, was found to be the result of hypermethylation and subsequent silencing of the *D1* gene [5]. Examples of agriculturally significant epialleles include the colorless non-ripening (*cnr*) mutant in tomato [4]. Caused by hypermethylation and subsequent silencing of the *CNR* locus, *cnr* tomato lines fail to produce lycopene and form mealy pericarps (i.e., mealy fruit). The mantled fruit phenotype, which causes massive losses in the yield of oil from oil palm, is due to DNA hypomethylation of a LINE transposable element located within the *DEFICIENS* locus [3]. Improper expression of the LINE element results in improper splicing and premature termination of the DEFICIENS protein, rendering it non-functional and resulting in deficient fruit set. In spite of the extremely low spontaneous rate of epiallele formation in plants, the above examples indicate the importance of DNA methylation to proper plant development. The deliberate alteration of DNA methylation in plants could serve to create increased phenotypic variation in genetically homogenous populations and would represent an additional tool for the continuous improvement of agriculturally significant plant cultivars.

Induced epigenetic alterations

As changes to DNA methylation have previously been implicated in agronomically important phenotypes as described previously, methods for purposeful induction of epialleles, including chemical treatment as well as the creation of gene-knockouts in methyltransferases, could serve as important agronomic tools for creating phenotypic variation in plants. Treatment of whole plants with cytosine analogs such as azacytidine or zebularine result in genome-wide demethylation, without the deliberate alteration of DNA sequence [22]. As the incorporation of cytidine analogs during DNA replication occurs stochastically, the levels of demethylation observed genome-wide are similar in pericentromeric regions and chromosome arms. Successful introduction of agriculturally significant phenotypes via chemical DNA demethylation have been previously obtained for rice and rapeseed cultivars, as well as in strawberry. Stable progeny of azacytidine treated rice were observed to have increased resistance to the rice pathogen *Xanthomonas oryzae*, due to the demethylation and subsequent misexpression of *Xa21G* resistance gene [23]. Some strawberry plants resulting from azacytidine treatment were observed to have altered time to flowering compared to non-treated lines, and the increased variation in time to flowering due to demethylation was stable over multiple generations [24]. But in summary, cytosine analogs produce stochastic untargeted changes in DNA methylation, the phenotypes obtained are random, and subsequent breeding is needed to fix the desired epialleles and traits.

Alternative approaches for introducing more defined epigenetic variation involve the use of more reliable approaches such as transgenesis in an attempt to increase efficiency and throughput of induced epiallele creation. One such technique involves the creation of epigenetic recombinant inbred lines (epiRILs). The ability to stably propagate methyltransferase mutants with large genome-wide reductions in methylation such as *met1* and *ddm1* in *Arabidopsis* has

enabled the creation epiRILs [25,26]. Created by crossing wild-type individuals to *met1* or *ddm1* mutants, F1 plants derived from this cross are heterozygous for *met1* or *ddm1* mutations but contain one wild-type and one hypomethylated copy of each chromosome. F1 individuals are self-fertilized, and individuals containing only wild-type copies of *met1* or *ddm1* are further self-fertilized for several additional generations. Thus, epiRIL lines possess mosaic methylome patterns depending on chromosomal inheritance patterns, while remaining nearly genetically identical to wild-type parental lines.

The creation and characterization of *ddm1* epiRILs have revealed increased phenotypic diversity of traits such as time to flowering, plant height, stem height, and fruit size [26]. The number of putative epiQTLs identified for a majority of traits profiled indicated polygenic inheritance, supporting the stable inheritance of demethylated regions genome-wide for multiple generations in *ddm1* epiRIL populations.

EpiRIL lines created with a primary cross with *met1*, termed *met1* epiRILs, have additionally been observed to have increased phenotypic diversity with respect to salt stress, biomass, time to flowering, and pathogen resistance compared to wild-type populations [25]. Importantly, the variation observed in epiRIL lines, which are genetically identical, phenocopy nearly 60% of the variation found naturally across *A. thaliana* ecotypes globally, with respect to time to flowering and plant height [25]. Thus, the modulation of DNA methylation in the absence of genetic variation has the potential to drastically increase variation in genetically homogenous plant populations.

As chemical treatment with cytosine analogs and epiRIL creation alter methylation globally in a random or non-targeted manner, they are useful for identifying both the contribution and potential of DNA methylation to alter relevant plant phenotypes, as no *a-priori* knowledge of

specific genomic regions is required. Thus, the identification of genomic regions where DNA methylation is important for relevant phenotypes can be assessed and characterized.

However, the use of chemical demethylation treatments and epiRIL creation for large-scale induction of epigenetic variation are infeasible as compared to chemical mutagenesis. Cytosine analogs such as azacytidine and zebularine, while potent DNA demethylating agents, are additionally genotoxic, and have been observed to cause base-pair substitution mutations, sister chromatid exchange, chromosomal aberrations and gene mutations in eukaryotic organisms [27]. Thus, the induction of genetic mutations, in addition to widespread demethylation, would serve to confound the potential phenotypic effects of epiallele creation. The generation of epiRIL lines in *Arabidopsis* is possible due to the previous characterization of *met1* and *ddm1* mutants, which are fertile despite large reductions in DNA methylation genome wide. The isolation of homologues of *met1* and *ddm1* in plant species such as maize, rice, or tomato has not been described to date, perhaps due to the larger fraction of transposable element content and heterochromatin present in plant genomes other than *Arabidopsis* [28]. Thus, new techniques for altering DNA methylation in an untargeted manner genome-wide are necessary for large-scale, untargeted epiallele induction in plant populations. Chapter 2 of my thesis attempts to address this limitation by employing the mammalian ten–eleven translocation methylcytosine dioxygenase 1 (TET1) to generate large-scale genome-wide epimutation in plants.

Legacy of targeted genome-editing approaches in plants

The modification of genetic sequence in a targeted, predictable manner has the potential to greatly accelerate the development of crop varieties when compared to traditional breeding approaches or the use of chemical mutagens, which may require multiple generations for the successful introgression of beneficial traits and removal of undesired alleles or modifications.

Since first utilized in 2005, the use of transgenic approaches for delivery of targeted genome-editing in plant systems has rapidly expanded, due to the development of numerous gene-editing systems including meganucleases, zinc-finger nucleases (ZFN), transcription activator-like effector nucleases (TALENs) and bacterial Cas site-directed nucleases, including Cas9 and Cas12a. The use of these gene-editing systems has enabled breeders to create edited crops exhibiting phenotypes of interest in crops including rice, wheat, maize, soybean, and tomato, and will continue to facilitate crop improvement in a rapid and targeted manner.

The use of meganucleases for genome-editing represented the first attempt to use a transgenic approach for targeted modification [29]. Similar to restriction enzymes, meganucleases recognize a unique recognition site in dsDNA. The small size of meganucleases, coupled with their relatively large recognition site sequence motifs of between 15-40 base pairs, enables their use for precise cleavage of pre-determined target sites of DNA in plant genomes. The use of meganucleases has successfully been used to create maize-sterile maize plants by targeting and disruption of the *MS26* locus, as well as the targeted insertion of herbicide-resistance genes in cotton [30,31]. However, modifying meganuclease recognition sites to target specific genomic regions of interest is a significant challenge, as the alteration of recognition sites significantly decreases cleavage activity [32]. Thus, the use of meganuclease technology for routine genome-editing plant systems is infeasible, due to the high degree of difficulty needed to engineer specific recognition and cleavage of desirable custom target sites.

Zinc finger nucleases (ZFN) have been more extensively applied for plant genome-editing than meganucleases and have successfully been applied to develop edited varieties of maize, soybean, petunia, and rapeseed. Harnessing the binding specificity of zinc finger domains of the murine transcription factor *Zif268*, zinc finger arrays can additionally be used for fusion

protein localization and have been used to target methylation and demethylation in *Arabidopsis* [33,34]. As nuclease activity from zinc finger nucleases (ZFN) is the result of a fusion to a homodimeric restriction enzyme active site, two zinc finger monomers are required to bind within an 18 base-pair window to facilitate cleave of the target DNA sequence, potentially increasing target specificity. ZFNs have been utilized for targeted insertion of herbicide resistance genes in maize, as well as knocking out several dicer-like genes in soybean simultaneously [35,36]. However, despite the increased programmability of ZFNs compared to meganucleases, the regions which can be targeted in a genome is limited and require extensive testing and validation to ensure low levels of off-target DNA binding [36].

Similar to ZFNs, Transcription activator-like effector nucleases (TALENs) are fusion proteins consisting of TALE repeat domains and a homodimeric restriction enzyme active site [37]. However, programming TALENs for site-specific DNA binding is vastly simplified when compared to ZFNs, as the alteration of two consensus amino acids, known as repeat variable di-residues (RVDs) within each TALE domain are responsible for localization to specific nucleotide contexts. Thus, the targeting of TALE proteins can be accomplished by assembling TALE monomers containing RVD domains that match a genomic target site of interest. To achieve sequence specificity, TALEs are typically constructed with 15-20 RVDs, enabling the targeting of a 30-base-pair target site. Due to their ease of programmability compared to ZFNs and meganucleases, TALENs have been widely utilized for plant genome-editing, and have been used to create modified rice, wheat, maize, sugarcane, soybean, potato, tomato, and *Brassica oleracea* varieties [38,39,40,41]. However, the large size of TALENs relative to ZFNs, in combination with the highly repetitive nature of TALE repeats, poses considerable challenges for construction and delivery of engineered TALENs.

CRISPR-Cas based systems for genome-engineering

The discovery and development of CRISPR-Cas systems for genome engineering represent the most versatile tool used for plant genome-editing to date. Due to its ease of use and simplified construction, the adoption of Cas nucleases such as Cas9 for genome-editing purposes in plants is nearly ubiquitous. Unlike previous genome-editing technologies, which require protein engineering to bind and cut desired DNA targets of interest, Cas nucleases are RNA-guided DNA endonucleases [42]. Thus, Cas proteins have no intrinsic DNA target, and must be programmed to search for genomic targets via expression of a guide RNA (gRNA). However, Cas proteins recognize a specific nucleotide motif, termed the protospacer adjacent motif (PAM), which are essential for DNA cleavage. While PAMs differ between Cas nucleases, suitable target sequences for disruption of gene function can be found for many loci when utilizing Cas nucleases such as Cas9 and Cas12a for genome-editing [43].

In addition to producing dsDNA cleavage at target sites, the use of endonuclease inactivated Cas nucleases, such as dCas9, enable the use of Cas nucleases for the site-specific location of DNA-interacting proteins via attachment to the N or C terminus of dCas9. This property has been exploited for a range of fusion partners, including transcriptional activators, repressors, as well as DNA demethylases and methylases [33,44]. Additionally, the use of nuclease-deficient Cas proteins has enabled the creation of base-editing technologies, whereby the targeting of nucleotide deaminases directly introduces base substitutions without dsDNA cleavage [45].

To date, CRISPR-Cas mediated genome-editing has been successfully applied to a wide range of plant species, including *Arabidopsis*, maize, wheat, rice, soybean, tomato, potato, orange, canola and grape [39,46,47,48,49,50,51,66]. Due to their ease of use, Cas nucleases have been

utilized to engineer and improve numerous agriculturally important traits such as pathogen resistance, herbicide resistance, overall yield, and as well as alterations in plant metabolism. Accomplished primarily by the generation of gene knockouts, CRISPR-Cas modified plants represent an important avenue for crop improvement, and further improvements for creation of knock-in plant lines via CRISPR-based technologies may enable the high-throughput creation of designer alleles [52].

As the DNA-binding of Cas nucleases is entirely dependent on the gRNA sequence introduced, selecting gRNA sequences that have high predicted efficiency for the target gene of interest, as well as low off-target activity, is essential. Numerous tools, some of which are plant-specific, have been created to aid in gRNA selection. Such tools, including CHOPCHOP and CRISPR-P, enable the rank and identification of suitable target sequences in a graphical user interface format [53,54]. The computational identification of putative off-target sites for target sequences is relatively accurate, particularly in smaller genomes. However, effectively predicting target site cleavage efficiency remains a challenge, as numerous factors including chromatin accessibility, expression levels of transgenic constructs and cell-type specificity cannot be accurately represented in current prediction tools [55]. Thus, the design and selection of several gRNAs of interest for each unique target region is greatly advantageous for desired DNA targeting of Cas proteins.

Due to the requirement for absolute sequence specificity, the expression of gRNAs of interest, particularly when using Cas9-based systems, requires the use of specialized Pol-III promoters that prevent RNA-modifications such as polyadenylation and 5′ capping [56]. Such promoters, typically those of snRNA U6 or U3, have been identified and successfully applied for gRNA expression in numerous plant species, including *Arabidopsis*, rice, and *Medicago* [57].

However, the use of Pol-III promoters for multiplex gRNA expression results in the creation of repetitive promoter-gRNA arrays, which increase the difficulty of assembly and result in interference effects among different linked gRNA sequences [58].

To circumvent the limitations of Pol-III-based expression of gRNA arrays, adaptations of both exogenous and endogenous RNA-processing machinery have been applied in plant-based systems for Cas9 genome-editing purposes [59]. The use of such RNA-processing machinery enables multiple gRNAs to be transcribed on a single RNA molecule, facilitating the use of Pol-II promoters. The advantages of Pol-II expression of gRNAs include tissue specificity, increased transcript levels, cytoplasmic export, and increased transcript length [60]. Endogenous RNA processing machineries utilized for gRNA maturation include the use of RNase P and RNase Z, which cleave pre-tRNA molecules to release mature tRNA molecules from polycistronic transcripts [61]. The use of maize tRNA units flanking gRNA units has successfully been applied for the processing of eight gRNAs from a single transcript in rice and tomato and resulted in a 100% increase in mutation efficiency in tomato when compared to individual promoter-gRNA arrays [59]. The use of ribozymes, or self-cleaving RNAs for processing of multiple gRNAs from single transcripts has also successfully been applied for genome-editing in plant-based systems, resulting in reduced but detectable editing efficiency when compared to tRNA spacers [59]. The introduction of exogenous RNA processing machinery such as the CRISPR-associated RNA endoribonuclease 4 (*CSY4*), while requiring the introduction and expression of an additional protein, has been shown to result in the highest efficiency for multiplex genome-editing when compared to tRNA or ribozyme-based systems. In addition to increased efficiency, the smaller size of the Csy4 hairpin (20-bp) when compared to a tRNA spacer (77 bp), facilitates construction and delivery of gRNA arrays.

Alternatively, the use of other CRISPR-Cas proteins, such as Cas12a, greatly simplifies the introduction and targeting of multiple genomic regions simultaneously. In addition to possessing RNA-guided DNA endonuclease properties like Cas9, Cas12a additionally possesses specific RNase activity for immature gRNA scaffold sequences and has distinct domains for RNA and DNA processing [62]. Thus, the use of a single CRISPR-Cas enzyme can simultaneously facilitate mature gRNA processing and target editing at numerous targets [63].

To date, numerous Cas12a orthologs including *Acidaminococcus sp*. Cas12a (AsCas12a) and *Lachnospiraceae bacterium* Cas12a (LbCas12a) have been utilized for plant genome-editing in species such as rice and *Arabidopsis* [64,65]. Previous studies have utilized gRNA processing machinery such as ribozymes and tRNA spacers for Cas12a-based multiplex mutagenesis, achieving comparable levels of editing as seen using Cas9. However, the application Cas12a for massively multiplex genome-editing approaches in plants remains unexplored. Chapter 3 of my thesis explores and advances multiplex genome-editing using Cas12a.

References:

1. Esquinas-Alcázar, J. Protecting crop genetic diversity for food security: political, ethical and technical challenges. *Nat. Rev. Genet.* **6**, 946–953 (2005).

2. Zhang, H., Lang, Z. & Zhu, J.-K. Dynamics and function of DNA methylation in plants. *Nat. Rev. Mol. Cell Biol.* **19**, 489–506 (2018).

3. Ong-Abdullah, M. *et al.* Loss of Karma transposon methylation underlies the mantled somaclonal variant of oil palm. *Nature* **525**, 533–537 (2015).

4. Manning, K. *et al.* A naturally occurring epigenetic mutation in a gene encoding an SBP-box transcription factor inhibits tomato fruit ripening. *Nat. Genet.* **38**, 948–952 (2006).

5. Miura, K. *et al.* A metastable DWARF1 epigenetic mutant affecting plant stature in rice. *Proc. Natl. Acad. Sci.* **106**, 11218 (2009).

6. Ronemus, M. J., Galbiati, M., Ticknor, C., Chen, J. & Dellaporta, S. L. Demethylation-Induced Developmental Pleiotropy in Arabidopsis. *Science* **273**, 654 (1996).

7. Woo, H. R., Dittmer, T. A. & Richards, E. J. Three SRA-Domain Methylcytosine-Binding Proteins Cooperate to Maintain Global CpG Methylation and Epigenetic Silencing in Arabidopsis. *PLOS Genet.* **4**, e1000156 (2008).

8. Lister, R. *et al.* Highly Integrated Single-Base Resolution Maps of the Epigenome in Arabidopsis. *Cell* **133**, 523–536 (2008).

9. Bartee, L., Malagnac, F. & Bender, J. Arabidopsis cmt3 chromomethylase mutations block non-CG methylation and silencing of an endogenous gene. *Genes Dev.* **15**, 1753–1758 (2001).

10. Jackson, J. P., Lindroth, A. M., Cao, X. & Jacobsen, S. E. Control of CpNpG DNA methylation by the KRYPTONITE histone H3 methyltransferase. *Nature* **416**, 556–560 (2002).

11. Du, J. *et al.* Dual Binding of Chromomethylase Domains to H3K9me2-Containing Nucleosomes Directs DNA Methylation in Plants. *Cell* **151**, 167–180 (2012).

12. Gouil, Q. & Baulcombe, D. C. DNA Methylation Signatures of the Plant Chromomethyltransferases. *PLOS Genet.* **12**, e1006526 (2016).

13. Stroud, H. *et al.* Non-CG methylation patterns shape the epigenetic landscape in Arabidopsis. *Nat. Struct. Mol. Biol.* **21**, 64–72 (2014).

14. Cao, X. & Jacobsen, S. E. Role of the Arabidopsis DRM Methyltransferases in De Novo DNA Methylation and Gene Silencing. *Curr. Biol.* **12**, 1138–1144 (2002).

15. Jia, Y. *et al.* Loss of RNA–Dependent RNA Polymerase 2 (RDR2) Function Causes Widespread and Unexpected Changes in the Expression of Transposons, Genes, and 24-nt Small RNAs. *PLOS Genet.* **5**, e1000737 (2009).

16. Havecker, E. R. *et al.* The Arabidopsis RNA-Directed DNA Methylation Argonautes Functionally Diverge Based on Their Expression and Interaction with Target Loci. *Plant Cell* **22**, 321 (2010).

17. Gong, Z. *et al.* ROS1, a Repressor of Transcriptional Gene Silencing in Arabidopsis, Encodes a DNA Glycosylase/Lyase. *Cell* **111**, 803–814 (2002).

18. Choi, Y. *et al.* DEMETER, a DNA Glycosylase Domain Protein, Is Required for Endosperm Gene Imprinting and Seed Viability in Arabidopsis. *Cell* **110**, 33–42 (2002).

19. Park, K. *et al.* DNA demethylation is initiated in the central cells of Arabidopsis and rice. *Proc. Natl. Acad. Sci.* **113**, 15138 (2016).

20. Hofmeister, B. T., Lee, K., Rohr, N. A., Hall, D. W. & Schmitz, R. J. Stable inheritance of DNA methylation allows creation of epigenotype maps and the study of epiallele inheritance patterns in the absence of genetic variation. *Genome Biol.* **18**, 155 (2017).

21. Weigel, D. & Colot, V. Epialleles in plant evolution. *Genome Biol.* **13**, 249 (2012).

22. Griffin, P. T., Niederhuth, C. E. & Schmitz, R. J. A Comparative Analysis of 5-Azacytidine- and Zebularine-Induced DNA Demethylation. *G3 GenesGenomesGenetics* **6**, 2773 (2016).

23. Akimoto, K. *et al.* Epigenetic Inheritance in Rice Plants. *Ann. Bot.* **100**, 205–217 (2007).

24. Xu, J., Tanino, K. K., Horner, K. N. & Robinson, S. J. Quantitative trait variation is revealed in a novel hypomethylated population of woodland strawberry (Fragaria vesca). *BMC Plant Biol.* **16**, 240 (2016).

25. Reinders, J. *et al.* Compromised stability of DNA methylation and transposon immobilization in mosaic Arabidopsis epigenomes. *Genes Dev.* **23**, 939–950 (2009).

26. Johannes, F. *et al.* Assessing the Impact of Transgenerational Epigenetic Variation on Complex Traits. *PLOS Genet.* **5**, e1000530 (2009).

27. Christman, J. K. 5-Azacytidine and 5-aza-2′-deoxycytidine as inhibitors of DNA methylation: mechanistic studies and their implications for cancer therapy. *Oncogene* **21**, 5483–5495 (2002).

28. Hu, L. *et al.* Mutation of a major CG methylase in rice causes genome-wide hypomethylation, dysregulated genome expression, and seedling lethality. *Proc. Natl. Acad. Sci.* **111**, 10642 (2014).

29. Marton, I. *et al.* Nontransgenic Genome Modification in Plant Cells. *Plant Physiol.* **154**, 1079 (2010).

30. Djukanovic, V. *et al.* Male-sterile maize plants produced by targeted mutagenesis of the cytochrome P450-like gene (MS26) using a re-designed I–CreI homing endonuclease. *Plant J.* **76**, 888–899 (2013).

31. D'Halluin, K. *et al.* Targeted molecular trait stacking in cotton through targeted double-strand break induction. *Plant Biotechnol. J.* **11**, 933–941 (2013).

32. Gao, H. *et al.* Heritable targeted mutagenesis in maize using a designed endonuclease. *Plant J.* **61**, 176–187 (2010).

33. Gallego-Bartolomé, J. *et al.* Targeted DNA demethylation of the Arabidopsis genome using the human TET1 catalytic domain. *Proc. Natl. Acad. Sci.* **115**, E2125 (2018).

34. Gallego-Bartolomé, J. *et al.* Co-targeting RNA Polymerases IV and V Promotes Efficient De Novo DNA Methylation in Arabidopsis. *Cell* **176**, 1068-1082.e19 (2019).

35. Ainley, W. M. *et al.* Trait stacking via targeted genome editing. *Plant Biotechnol. J.* **11**, 1126–1134 (2013).

36. Curtin, S. J. *et al.* Targeted Mutagenesis of Duplicated Genes in Soybean with Zinc-Finger Nucleases. *Plant Physiol.* **156**, 466 (2011).

37. Zhang, Y. *et al.* Transcription Activator-Like Effector Nucleases Enable Efficient Plant Genome Engineering. *Plant Physiol.* **161**, 20 (2013).

38. Kelliher, T. *et al.* MATRILINEAL, a sperm-specific phospholipase, triggers maize haploid induction. *Nature* **542**, 105–109 (2017).

39. Du, H. *et al.* Efficient targeted mutagenesis in soybean by TALENs and CRISPR/Cas9. *J. Biotechnol.* **217**, 90–97 (2016).

40. Jung, J. H. & Altpeter, F. TALEN mediated targeted mutagenesis of the caffeic acid O-methyltransferase in highly polyploid sugarcane improves cell wall composition for production of bioethanol. *Plant Mol. Biol.* **92**, 131–142 (2016).

41. Nicolia, A. *et al.* Targeted gene mutation in tetraploid potato through transient TALEN expression in protoplasts. *J. Biotechnol.* **204**, 17–24 (2015).

42. Feng, Z. *et al.* Efficient genome editing in plants using a CRISPR/Cas system. *Cell Res.* **23**, 1229–1232 (2013).

43. Zhong, Z. *et al.* Improving Plant Genome Editing with High-Fidelity xCas9 and Non-canonical PAM-Targeting Cas9-NG. *Mol. Plant* **12**, 1027–1036 (2019).

44. Gentzel, I. N. *et al.* A CRISPR/dCas9 toolkit for functional analysis of maize genes. *Plant Methods* **16**, 133 (2020).

45. Kang, B.-C. *et al.* Precision genome engineering through adenine base editing in plants. *Nat. Plants* **4**, 427–431 (2018).

46. Nekrasov, V. *et al.* Rapid generation of a transgene-free powdery mildew resistant tomato by genome deletion. *Sci. Rep.* **7**, 482 (2017).

47. Li, J.-F. *et al.* Multiplex and homologous recombination–mediated genome editing in Arabidopsis and Nicotiana benthamiana using guide RNA and Cas9. *Nat. Biotechnol.* **31**, 688–691 (2013).

48. Gil-Humanes, J. *et al.* High-efficiency gene targeting in hexaploid wheat using DNA replicons and CRISPR/Cas9. *Plant J.* **89**, 1251–1262 (2017).

49. Wang, X. *et al.* CRISPR/Cas9-mediated efficient targeted mutagenesis in grape in the first generation. *Plant Biotechnol. J.* **16**, 844–855 (2018).

50. Andersson, M. *et al.* Efficient targeted multiallelic mutagenesis in tetraploid potato (Solanum tuberosum) by transient CRISPR-Cas9 expression in protoplasts. *Plant Cell Rep.* **36**, 117–128 (2017).

51. Macovei, A. *et al.* Novel alleles of rice eIF4G generated by CRISPR/Cas9-targeted mutagenesis confer resistance to Rice tungro spherical virus. *Plant Biotechnol. J.* **16**, 1918–1927 (2018).

52. Rozov, S. M., Permyakova, N. V. & Deineko, E. V. The Problem of the Low Rates of CRISPR/Cas9-Mediated Knock-ins in Plants: Approaches and Solutions. *Int. J. Mol. Sci.* **20**, (2019).

53. Montague, T. G., Cruz, J. M., Gagnon, J. A., Church, G. M. & Valen, E. CHOPCHOP: a CRISPR/Cas9 and TALEN web tool for genome editing. *Nucleic Acids Res.* **42**, W401–W407 (2014).

54. Liu, H. *et al.* CRISPR-P 2.0: An Improved CRISPR-Cas9 Tool for Genome Editing in Plants. *Mol. Plant* **10**, 530–532 (2017).

55. Naim, F. *et al.* Are the current gRNA ranking prediction algorithms useful for genome editing in plants? *PLOS ONE* **15**, e0227994 (2020).

56. Xing, H.-L. *et al.* A CRISPR/Cas9 toolkit for multiplex genome editing in plants. *BMC Plant Biol.* **14**, 327 (2014).

57. Jacobs, T. B., LaFayette, P. R., Schmitz, R. J. & Parrott, W. A. Targeted genome modifications in soybean with CRISPR/Cas9. *BMC Biotechnol.* **15**, 16 (2015).

58. Angulo, J. *et al.* Targeted mutagenesis of the Arabidopsis GROWTH-REGULATING FACTOR (GRF) gene family suggests competition of multiplexed sgRNAs for Cas9 apoprotein. *bioRxiv* 2020.08.16.253203 (2020) doi:10.1101/2020.08.16.253203.

59. Čermák, T. *et al.* A Multipurpose Toolkit to Enable Advanced Genome Engineering in Plants. *Plant Cell* **29**, 1196 (2017).

60. Zhong, G., Wang, H., Li, Y., Tran, M. H. & Farzan, M. Cpf1 proteins excise CRISPR RNAs from mRNA transcripts in mammalian cells. *Nat. Chem. Biol.* **13**, 839–841 (2017).

61. Xie, K., Minkenberg, B. & Yang, Y. Boosting CRISPR/Cas9 multiplex editing capability with the endogenous tRNA-processing system. *Proc. Natl. Acad. Sci.* **112**, 3570 (2015).

62. Zetsche, B. *et al.* Multiplex gene editing by CRISPR–Cpf1 using a single crRNA array. *Nat. Biotechnol.* **35**, 31–34 (2017).

63. Campa, C. C., Weisbach, N. R., Santinha, A. J., Incarnato, D. & Platt, R. J. Multiplexed genome engineering by Cas12a and CRISPR arrays encoded on single transcripts. *Nat. Methods* **16**, 887–893 (2019).

64. Bernabé-Orts, J. M. *et al.* Assessment of Cas12a-mediated gene editing efficiency in plants. *Plant Biotechnol. J.* **17**, 1971–1984 (2019).

65. Malzahn, A. A. *et al.* Application of CRISPR-Cas12a temperature sensitivity for improved genome editing in rice, maize, and Arabidopsis. *BMC Biol.* **17**, 9 (2019).

66. Zhai, Y. *et al*. CRISPR/Cas9-mediated genome editing reveals differences in the contribution of INDEHISCENT homologues to pod shatter resistance in Brassica napus L. Theor. Appl. Genet. **132**, 2111–2123 (2019)

CHAPTER 2

TET-MEDIATED EPIMUTAGENESIS OF THE ARABIDOPSIS THALIANA METHYLOME[1]

[1] William T. Jordan, Lexiang Ji, Xiuling Shi, Lulu Hu, Chuan He and Robert J. Schmitz (2018) Nature Communications March;9(895)

Reprinted here with permission of the publisher.

ABSTRACT

DNA methylation in the promoters of plant genes sometimes leads to transcriptional repression, and the loss of DNA methylation in methyltransferase mutants results in altered gene expression and severe developmental defects. However, many cases of naturally occurring DNA methylation variations have been reported, whereby altered expression of differentially methylated genes is responsible for agronomically important traits. The ability to manipulate plant methylomes to generate epigenetically distinct individuals could be invaluable for breeding and research purposes. Here, we describe "epimutagenesis," a method to rapidly generate DNA methylation variation through random demethylation of the *Arabidopsis thaliana* genome. This method involves the expression of a human ten–eleven translocation (TET) enzyme, and results in widespread hypomethylation that can be inherited to subsequent generations, mimicking mutants in the maintenance of DNA methyltransferase met1. Application of epimutagenesis to agriculturally significant plants may result in differential expression of alleles normally silenced by DNA methylation, uncovering previously hidden phenotypic variations.

INTRODUCTION

Our ability to develop novel beneficial crop traits has significantly improved over the last 100 years, although the ability to maintain this trajectory is limited by allelic diversity. While genetic variation has been heavily exploited for crop improvement, utility of epigenetic variation has yet to be efficiently implemented. Epigenetic variation arises not from a change in the DNA sequence, but by changes in modifications to DNA such as cytosine methylation. This variation can result in the emergence of novel and stably inherited phenotypes, as well as unique patterns of gene expression.

In plant genomes, cytosine methylation occurs at three major sequence contexts: CG, CHG, and CHH (where H = A, C, or T) (Law and Jacobsen 2010). Methylation at these different contexts is coordinated by distinct maintenance mechanisms during DNA replication. The methylation of DNA in all three contexts is essential for transcriptional silencing of transposons, repeat sequences, and certain genes. Genes regulated by this mechanism are stably repressed throughout the soma and represent an untapped source of hidden genetic variation if transcriptionally re-activated, as revealed from pioneering studies in the model plant *A. thaliana* (Johannes et al. 2009; Reinders et al. 2009; Cortijo et al. 2014a). However, the impact of this variation is not observed in wild-type plants, as genes silenced by DNA methylation are not expressed. This novel source of genetic variation was uncovered by creating epigenetic recombinant inbred lines (epiRILs) from crosses between a wildtype individual and a mutant defective in maintenance of DNA methylation (Johannes et al. 2009; Reinders et al. 2009; Cortijo et al. 2014a). EpiRILs, while genetically wild type, contain mosaic DNA methylomes dependent on chromosomal inheritance patterns, as DNA methylation is meiotically inherited in *A. thaliana* (Johannes et al. 2009; Cortijo et al. 2014b; Bewick et al. 2016; Hofmeister et al. 2017). Phenotypic characterization of epiRILs has revealed extensive morphological variation with respect to traits such as flowering time, root length, and resistance to bacterial infection (Johannes et al. 2009; Reinders et al. 2009; Cortijo et al. 2014a). The morphological variation generated by the creation of epiRILs has revealed extensive hidden genetic variation in plant genomes that can be observed due to expression of newly unmethylated regions. However, the creation of epiRILs requires one founding parent to be a null mutant in the maintenance DNA methylation pathway. Unfortunately, unlike in *A. thaliana*, the loss of DNA methylation maintenance activity often results in lethality in crops (Hu et al. 2014; Li et al. 2014b).

Therefore, novel methodologies are required to realize the potential of these hidden epialleles in crop genomes.

Epimutagenesis is an alternative method to generate epiRILs. Instead of relying on the genome-wide demethylation of one of the two founding parents, epimutagenesis introduces random methylation variation via the introduction of a transgene. Here, we describe a novel epimutagenesis approach in *A. thaliana* using a human ten–eleven translocation methylcytosine dioxygenase 1 (TET1) (Tahiliani et al. 2009; Ito et al. 2010; Pastor et al. 2011; Hollwey et al. 2016), which catalyzes the conversion of 5- methylcytosine (5mC) to 5-hydroxymethylcytosine (5hmC). Although TET enzymes and their primary product 5hmC are not found in plant genomes (Erdmann et al. 2014), ectopic expression of a human TET enzyme resulted in widespread DNA demethylation and induced phenotypic variation in *A. thaliana*.

RESULTS

Overexpressing TET1 in *Arabidopsis* hypomethylates the genome. Transgenic *A. thaliana* plants were generated expressing the catalytic domain (residues 1418–2136) of the human TET1 protein (hTET1cd) under the control of the CaMV35S promoter. To assess the impact of hTET1cd expression on the *A. thaliana* methylome, whole-genome bisulfite sequencing (WGBS) was performed on two independently derived transgenic plants (35S:TET1-1 and 35S:TET1-2). WGBS revealed a global reduction of CG methylation from 18.2% in two wild-type individuals to 8.9% in 35S:TET1-1 and 6.9% in 35S:TET1-2 (compared to 0.5% in met1-3). The effects of hTET1cd expression on CHG and CHH methylation were not as severe compared to CG methylation (**Figure 3.1a**). Importantly, different degrees of CG hypomethylation were observed in different events. This result has important implications for epimutagenesis in economically and agriculturally significant plant species, as it appears feasible

to control the degree of DNA hypomethylation by screening for plants with desired levels of demethylation. Taken together, these results show that the expression of hTET1cd results in intermediate CG methylation levels when compared to wild-type and met1 individuals. The primary product of TET1 oxidation is 5hmC, which is indistinguishable from 5mC by WGBS. We therefore performed Tet-assisted bisulfite sequencing (TAB-seq) to profile 5hmC levels in 35S:TET1 plants (Yu et al. 2012a). No detectable levels of 5hmC were found in the transgenic lines assayed (**Supplementary Figure 3.1a, b**). Thus, the widespread loss of CG DNA methylation observed may result from a failure to maintain methylation at CG sites that possess 5hmC, or through active removal of 5hmC or further oxidized products via the base excision repair pathway.

To better understand the effects of hTET1cd expression, we determined changes in the *A. thaliana* methylome at the chromosomal and local levels. Plotting methylation levels across all five chromosomes revealed a strong depletion of CG methylation at the pericentromeric region (**Figure 3.1b**). CG hypomethylation occurred at both gene body methylated (gbM) and select RNA-directed DNA methylated (RdDM) loci. (**Figure 3.1c, d**). To further quantify the observed hypomethylation, metaplots were created for genes and transposons, respectively (**Figure 3.1e, f and Supplementary Figure 3.1c–f**). A strong reduction of mCG and a mild reduction of mCHG/mCHH were observed at both genes and transposons. On average, 97.9% of gbM genes and 56.7% of methylated transposons (where these regions have at least 50% mCG in wild type) lost at least half of their CG methylation in epimutagenized lines. Collectively, these results indicate that hypomethylation was more severe in genes than transposons, possibly the result of de novo methylation by the RdDM pathway, which is primarily active at transposons.

TET1-mediated DNA demethylation mimics *met1* mutants.

An analysis of differentially methylated regions (DMRs) was then carried out to assess the genome-wide impact of hTET1cd expression. A total of 56,283 CG DMRs ranging in size from 6– 20,286 base pairs (bp) were identified (**Figure 3.1g**). Of these, 38.7% were located in intergenic sequences, 53.7% overlapped with genes, and 7.6% were located in promoter regions (≤1 kb upstream of a gene). As also seen in met1 mutants, the predominant effect of hTET1cd expression is CG hypomethylation (12,641 and 20,601 DMRs lost more than 50% mCG in 35S:TET1-1 and 35S:TET1-2, respectively; no region gained more than 50% mCG). However, the extent of CG methylation loss caused by hTETcd expression is lower than in met1: 31.8 Mb of the genome significantly lost CG methylation in met1, whereas 9.9 Mb and 18.0 Mb were lost in 35S:TET1-1 and 35S:TET1-2, respectively.

Previous studies of the *met1* methylome have revealed a loss of mCHG/mCHH methylation in a subset of CG-hypomethylated regions (Stroud et al. 2013b). At these loci, DNA methylation is stably lost, in contrast to regions where DNA methylation is re- established by de novo methylation pathways. These loci are ideal targets of epimutagenesis, as the co-existence of all three types of methylation is more frequently correlated with transcriptional repression of genes than CG methylation alone. This, coupled with the long-term stability of hypomethylation, may facilitate inherited transcriptional changes. An analysis of the interdependence of the loss of CG methylation on non-CG methylation levels revealed that 39.7 Kb and 931.5 Kb of CHG methylated sequences lost significant amounts of methylation in two independent epimutagenized lines, compared to 4.0 Mb of sequence in met1 mutants. A similar analysis for the loss of CHH methylation revealed losses of 23.3 Kb and 492.5 Kb in epimutagenized individuals, compared to 1.1 Mb lost in met1 mutants. Of the 56,283 identified CG DMRs, 10,491 overlapped regions that contained at least 20% CHG methylation and 7214 overlapped

regions that contained at least 10% CHH methylation in wild-type individuals. To determine how many of these regions are susceptible to losing non-CG methylation if CG methylation is first depleted, we created a frequency distribution of mCHG and mCHH levels in wild-type and epimutagenized individuals (**Figure 3.1h, i**). In total, 2341 and 3447 regions lost more than 10% CHG methylation in 35S:TET1-1 and 35S:TET1-2, respectively, whereas 2475 and 3379 regions lost more than 5% CHH methylation in 35S:TET1-1 and 35S:TET1-2, respectively. Regions that are susceptible to losses of CG and non-CG methylation in lines expressing hTET1cd share a substantial overlap with regions that lose non-CG methylation in met1 (**Supplementary Figure 3.1g, h**). In total, 1708 (73.0%) and 2386 (69.2%) regions that have lost more than 10% mCHG in 35S:TET1-1 and 35S:TET1-2 have reduced levels in met1, whereas 2013 (81.3%) and 2563 (75.9%) regions that have lost more than 5% mCHH in 35S:TET1-1 and 35S:TET1-2 have reduced levels in met1. As crop genomes have a greater number of loci targeted for silencing by CG, CHG, and CHH methylation when compared to *A. thaliana*, ectopic expression of hTET1cd is likely a viable approach for the creation epiRILs (Niederhuth et al. 2016).

TET1-mediated variation of CHG methylation.

Mutations in met1 also result in hypermethylation of CHG sites in gene bodies due to the loss of methylation in the seventh intron of the histone 3 lysine 9 (H3K9) demethylase, increase in bonsai methylation 1 (IBM1) (Lister et al. 2008; Saze et al. 2008; Miura et al. 2009; Rigal et al. 2012). This results in alternative splicing of IBM1, ultimately producing a non-functional gene product (IBM1-S), which results in ectopic accumulation of di-methylation of H3K9 (H3K9me2) throughout the genome (Rigal et al. 2012). As in met1, the seventh intron of IBM1 was hypomethylated in 35S:TET1-1, 35S:TET1-2, and an additional two lines, 35S:TET1-2T5 and 35S:TET1-2T6, which were propagated for an additional two and three generations,

respectively (**Figure 3.2a**). The increased abundance of IBM1-S transcript was confirmed by RT-qPCR in line 35S:TET1-2T6 (**Supplementary Figure 3.2a**), leading to CHG hypermethylation at gbM loci (**Figure 3.2b**). Further quantitative analysis revealed extensive variation in genome-wide gains and losses of CHG methylation in these two lines, ~1.8 Mb and 2.3 Mb of additional CHG methylation, respectively (**Figure. 3c and Supplementary Figure 3.2b, c**). To test the impact of a reduction in functional IBM1 on H3K9me2, we performed chromatin immunoprecipitation (ChIP) against H3K9me2 in 35S:TET1-2T6, which revealed a subtle increase in H3K9me2 in gbM loci that possessed CHG hypermethylation (**Figure 3.2d**).

To further characterize regions of differential CHG methylation, identified CHG DMRs in line 35S:TET1-2T5 were categorized into discrete groups based on their DNA methylation status in wild-type individuals. Of the 9917 CHG DMRs identified, 1460 were in loci that are defined as gbM in wildtype individuals, 584 were in unmethylated regions, and 6940 of them were in RdDM-like regions (**Figure 3.2e–g**). Interestingly, in line 35S:TET1-2T5, 1409 (96.5%) of the CHG DMRs in gbM-like loci gained CHG hypermethylation, whereas 2680 (38.6%) of the CHG DMRs in RdDM-like regions lost CHG, in contrast to 825 (11.9%) RdDM-like regions that gained CHG methylation. Lastly, there were 503 (86.1%) loci that are unmethylated in wild-type individuals that gain CHG methylation as well as CG and CHH methylation in the epimutagenized lines (**Figure 3.2e–g**). These results reveal that methods for epimutagenesis can result in both losses and gains in DNA methylation genome wide.

To characterize the effect of hTET1cd-induced methylome changes on gene expression, we performed RNA-sequencing (RNA-seq) on leaf tissue of wild-type, 35S:TET1-1 and 35S:TET1-2. Compared to wild-type plants, 629 and 736 upregulated genes were identified in 35S:TET1-1 and 35S:TET1-2, respectively, with 176 and 260 genes overlapping with identified

CG DMRs. A total of 1277 and 1428 downregulated genes were identified and 268 and 324 of them overlapped with CG DMRs. There was a high level of overlap in transcriptome changes seen in 35S:TET1-1 and 35S:TET1-2 compared to met1 and ibm1 (**Figure 3.2h, i**). Of the genes upregulated in met1, 36.8 and 38.7% overlapped with upregulated genes in 35S:TET1-1 and 35S:TET1-2, respectively (**Supplementary Figure 3.2d**). An even greater overlap was observed with downregulated genes in met1, as 60.1 and 65.2% overlapped with downregulated genes in 35S:TET1-1 and 35S:TET1-2, respectively (**Supplementary Figure 3.2e**). These results reveal that hTET1cd expression in *A. thaliana* is a viable approach for accessing hidden sources of allelic variation by inducing expression variation.

TET1 expression leads to a delay in the floral transition.

In the transgenic plants that were used for WGBS, we observed a delay in the developmental transition from vegetative to reproductive growth (**Figure 3.3a, b**). We hypothesized that the observed late flowering phenotype was associated with the demethylation of the FLOWERING WAGENINGEN (FWA) locus, as is observed in met1 mutants (Finnegan et al. 1996; Soppe et al. 2000). A closer inspection of the DNA methylation status of this locus revealed that DNA methylation was completely abolished, as was methylation at adjacent CHG and CHH sites (**Figure 3.3c**). As in met1, the loss of methylation at the FWA locus was associated with an increase in FWA expression (**Figure 3.3d**), which is known to cause a delay in flowering by restricting the movement of the florigen signal, FT, to the shoot apex (Ikeda et al. 2007). These results demonstrate that expression of hTET1cd leads to phenotypic variation by abolishing methylation at some regions in all sequence contexts (CG, CHG, and CHH sites).

TET1-mediated demethylation is transgenerationally inherited. To assess the stability and inheritance of TET1-mediated demethylation, T1 individuals expressing 35S:TET1 were self-

fertilized, allowing for the loss of the hTET1cd transgene due to allelic segregation. Unexpectedly, transgene-free 35S:TET1-1 T2 individuals (35S:TET1-1.3-TET1, 35S:TET1-1.4-TET1, and 35S:TET1- 1.5-TET1) exhibited a reversion to a normal flowering phenotype, and genome-wide methylation levels closely resembled that of wild-type individuals (**Figure 3.4a**). WGBS on T2 individuals retaining the transgene revealed similar levels of CG methylation as the T1 parent (35S:TET1-1.1+TET1 and 35S:TET1-1.2+TET1). The methylation level of CGs at genes and transposons revealed that demethylation of gbM loci was partially inherited in transgene-free T2 individuals, whereas active remethylation was found at transposons in the same individuals (**Figure 3.4b–f**). These results indicate that an active process likely in the meristem and/or germline is counteracting the activity of the hTET1cd transgene (Baubec et al. 2014). To quantify how many regions were susceptible to the loss of non-CG methylation, a DMR analysis was conducted for each 35S:TET1-1 T2 individual. A total of 655 and 659 CHG hypomethylated regions were identified in T2 lines retaining the transgene. In contrast, 211, 155 and 199 hypomethylated regions were identified in three transgene-free T2 individuals, respectively (**Figure 3.4g**).

To determine if increased expression of hTET1cd in meristematic tissue would increase the likelihood of germline transmittance of demethylation to transgene-free progeny, transgenic *A. thaliana* plants were generated expressing a previously described superfolder GFP (sfGFP) hTET1cd fusion, under control of the *A. thaliana* ACTIN 2 (ACT2) promoter (ACT2:sfGFP-hTET1cd), which is known to have activity in all tissues of juvenile plants, including meristematic tissue (An et al. 1996). Translation and nuclear localization of the sfGFP-TET1cd fusion protein was confirmed in young cotyledons using confocal microscopy (**Supplementary Figure 3.3a**). T1 populations transformed with ACT2:sfGFP-hTET1cd exhibited a 27-fold

increase in later flowering compared to 35S:hTET1cd, indicating high activity of the sfGFP-TET1cd fusion protein in *A. thaliana* (**Figure 3.5a**). To assess the variation between lines, we performed WGBS on four independent ACT2:TET1 T1 lines (**Supplementary Figure 3.4a–c**). Differential levels of global CG demethylation were observed in these four lines, further confirming that plants subjected to epimutagenesis can possess different degrees of demethylation. The expression of the sfGFPhTET1cd transgene was also confirmed by RT-qPCR in select lines (**Supplementary Figure 3.4d**). Subsequent DMR analysis revealed 68,260 CG DMRs, 9235 CHG DMRs, and 2793 CHH DMRs between these lines, a drastic increase in DMRs compared to those within siblings in 35S:TET1 lines (**Figure 3.4g**).

To further assess the inheritance of the demethylation pattern as a result of epimutagenesis, we selected T2 progeny of a late flowering ACT2:TET1-1.2-TET1 individual containing and lacking the transgene. WGBS data from these two individuals was used to confirm the presence/absence of the sfGFP-hTET1cd transgene (**Supplementary Figure 3.3d, e**). Individuals retaining and lacking the transgene due to allelic segregation both exhibited a late flowering phenotype (31 and 28 leaves upon flowering in ACT2:TET1-1.1+TET1 and ACT2:TET1-1.2-TET1, respectively), ectopic expression, and loss of DNA methylation of FWA (**Supplementary Figure 3.3b, c**). WGBS on these two individuals revealed a reduction in CG methylation that was maintained and stably inherited irrespective of transgene presence, confirming the high activity of demethylation in ACT2:TET1 lines compared to 35S:TET1 lines (**Figure 3.5b-f**). It is unclear exactly why the late flowering phenotype was stably inherited in the T2 individuals without the transgene in the ACT2-driven lines versus the 35S-driven lines, although it is likely a combination of promoter strength and cell type specificity. For stable inheritance of demethylation and the late-flowering phenotype, TET1 activity would be required

in the meristematic and/or germline cells. Collectively, these results demonstrate the stable inheritance of TET1-mediated demethylation and a delayed floral transition in the absence of the transgene.

DISCUSSION

The discovery that expression of the catalytic domain of the human TET1 protein in *A. thaliana* leads to widespread loss of CG methylation enables the creation epimutants without the need for methyltransferase mutants, which often causes lethality in crops. In addition to epimutagenesis, hTET1cd could be used in combination with sequence-specific DNA-binding proteins such as dCas9 to direct DNA demethylation in plant genomes, as has been demonstrated in mammalian systems (Maeder et al. 2013; Mendenhall et al. 2013; Choudhury et al. 2016; Liu et al. 2016; Vojta et al. 2016). The stable meiotic inheritance of DNA methylation states in flowering plant genomes provides a stark contrast to the inheritance of DNA methylation in mammalian genomes, where genome-wide erasure of DNA methylation and reprogramming occurs each generation (Heard and Martienssen 2014). This property of flowering plant genomes makes them ideal targets of induced epialleles, as once a new methylation state occurs it is often inherited in subsequent generations. Application of epimutagenesis and the use of TET-mediated engineering of DNA methylation states in economically and agriculturally significant plant species will be an interesting area of future investigation.

METHODS

Synthesis and cloning of the human TET1 catalytic domain. A human TET1 catalytic domain (hTET1cd) sequence (residues 1418–2136) was synthesized by GenScript, and moved to a plant transformation compatible vector (pMDC32) using LR clonase from Life Technologies as per the manufacturer's instructions (catalog #11791100). ACT2:sfGFP-

hTET1cd was subcloned by Genscript in the pMDC32 vector background using the sfGFP-TET1cd fragment from Addgene plasmid #82561. The ACT2 promoter sequence was kindly provided by Dr. Richard Meagher.

Plant transformation and screening. The hTET1cd sequence in the pMDC32 vector was transformed into *Agrobacterium tumefaciens* strain C58C1 and plated on LB- agar supplemented with kanamycin (50 μg/mL), gentamicin (25 μg/mL), and rifampicin (50 μg/mL). A single kanamycin-resistant colony was selected and used to start a 250- mL culture in LB Broth Miller liquid media supplemented with gentamicin (25 μg/mL), kanamycin (50 μg/mL), and rifampicin (50 μg/mL), which was incubated for 2 days at 30 °C. Bacterial cells were pelleted by centrifugation at 4000 rpm for 30 min and the supernatant decanted. The remaining bacterial pellet was re-suspended in 200 mL of 5% sucrose with 0.05% Silwet L77. Plant transformation was performed using the floral dip method (Clough and Bent 1998). Seeds were collected upon senescenceat maturity and transgenic plants were identified via selection on 1/2 LS plates supplemented with Hygromycin B (25 μg/mL). 35S:TET1-1 is a T1 individual, 35S:TET1-2 is a T3 plant, 35S:TET1-3 is a T4 plant. All transgenic individuals chosen for analysis contain independent insertions of hTET1cd and are not the result of single-seed decent unless otherwise noted.

DNA and RNA isolation. *A. thaliana* leaf tissue was flash-frozen and finely ground to a powder using a mortar and pestle. DNA extraction was carried out on all samples using the DNeasy Plant Mini Kit (Qiagen), and the DNA was sheered to ~200 bp by sonication. RNA was isolated from finely ground flash-frozen leaf tissue using Trizol (Thermo Scientific). For RT-qPCR, RNA was further treated with TURBO™ DNase (Thermo Scientific) according to the manufacturer's instructions. One microgram of RNA was subsequently reverse transcribed with M-MuLV

reverse transcriptase according to the manufacturer's instructions (NEB). RT-qPCR was used to analyze cDNA populations using PP2AA-3 (AT1G13320) as an endogenous control, and was performed on a Roche LIghtCycler 480 instrument using SYBR Green detection chemistry. The genes assayed by this method were IBM1-S, IBM1-L, FWA and sfGFP-hTET1cd. Primers used for RT-qPCR were designed using PrimerQuest from Integrated DNA Technologies (www.idtdna.com/PrimerQuest/). Primer sequences used for RT-qPCR: PP2AA-3-F: 5′ -AATGAGGCAGAAGTTCGGATAG-3′ , PP2AA-3-R: 5′ -CAGGGAAGAATGTGCTGGATAG-3′ , ibm1s-F: 5′ -TCTTTCTTCTAAGTCTGTCCATTCT-3′ , ibm1s-R: 5′ -GTGACCGATTAGGAAATGGTATCT-3′ , ibm1L-F: 5′ -CCGAAGCCAAAGTGGAGATA-3′ , ibm1L-R: 5′ -CTTCCTCTTCCGTAGACTTCTTT-3′ , FWA-F: 5′ -CAAGATGGTGGAAGGATGAGAA-3′ , FWA-R: 5′ -CTCTGTTCTTCAGTGGGATGAG-3′ , sfGFP-hTET1cd-F: 5'-CAAAGATGACGGGACCTACAA-3', sfGFP-hTET1cd-R: 5'-GTACTCGAGTTTGTGTCCAAGA-3'.

Library construction. Genomic DNA libraries were prepared following the MethylC-seq protocol without use of the bisulfite conversion step. MethylC-seq libraries were prepared as previously described in (Urich et al. 2015). Briefly, genomic DNA was sonicated to 200 bp using a Covaris S-series focused ultrasonicator, and end-repaired using End-It DNA end-repair kit (Epicentre). End-repaired DNA was subjected to A- tailing using Klenow 3'-5' exo− (NEB) and ligated to methylated adapters using T4 DNA ligase (NEB). Ligated DNA was subsequently bisulfite-converted using the EZ DNA methylation-Gold kit as per the manufacturer’s instructions and amplified using KAPA HiFi uracil + Readymix Polymerase. RNAseq libraries

were constructed using Illumina TruSeq Stranded RNA LT Kit (Illumina, San Diego, CA) following the manufacturer's instructions with limited modifications. The starting quantity of total RNA was adjusted to 1.3 μg, and all volumes were reduced to a third of the described quantity.

TAB-seq libraries were prepared as previously described in ref. 16. Briefly, genomic DNA was glucosylated and oxidized using T4-βGT (NEB) and recombinant mTET1. Standard bisulfite treatment is then performed using the MethylCode Bisulfite Conversion kit (Thermo Fisher Scientific).

ChIP library preparation. Leaves were treated with formaldehyde to covalently link protein to DNA, washed several times with distilled water, patted dry, and ground into fine powder in liquid nitrogen. Chromatin was extracted with a series of extraction buffers and sonicated. The final chromatin solution was incubated overnight with anti- H3K9me2 antibody (Cell Signaling Technology, 9753S)-coated Dynabeads protein A (Life Technologies, 10002D) to precipitate the immune complex. After a few washes, the immune complex was eluted and incubated at 65° in the presence of a high concentration of NaCl in a water-bath overnight. After degrading the proteins with proteinase K, DNA was recovered by phenol/chloroform/isoamyl alcohol extraction followed by ethanol precipitation. The DNA pellet was then dissolved in 30 μl of nuclease-free water.

Sequencing. Illumina sequencing was performed at the University of Georgia Genomics Facility using an Illumina NextSeq 500 instrument. For MethylC-seq and TAB-seq, raw reads were trimmed for adapters and preprocessed to remove low-quality reads using Ccutadapt 1.9.dev1 (Martin 2011). For RNA-seq and ChIP-seq, these processes were carried out by Trimmomatic

v0.32 (Bolger et al. 2014). The mutant allele of met1 is met1-3, and methylome data was downloaded under accession GSE39901. The mutant allele of *ibm1* is *ibm1-6*.

MethylC-seq data processing. Qualified reads were aligned to the *A. thaliana* TAIR10 reference genome as described in ref. (Schmitz et al. (2013a). Chloroplast DNA (which is fully unmethylated) was used as a control to calculate the sodium bisulfite reaction nonconversion rate of unmodified cytosines. All conversion rates were >99%. The list of gbM genes used in this study was previously curated[18]. Heat maps were clustered by complete linkage method conducted by R (https://www.r-project.org). All methylation levels reported in all analyses are presented as differences in absolute values, including defining DMRs and calculating hyper/hypomethylated regions. The only exception is in the comparison of mCG loss between gbM, where we used a percentage difference.

RNA-seq data processing. Qualified reads were aligned to the *A. thaliana* TAIR10 reference genome using TopHat v2.0.13 (Kim et al. 2013). Gene expression values were computed using Cufflinks v2.2.1 (Trapnell et al. 2010). Genes determined to have at least twofold log2 expression changes by Cufflinks and passed tests were identified as differentially expressed genes. The Col-0 wild-type transcriptomes were downloaded using data from accession GSE75071.

TAB-seq data processing. Qualified reads were aligned to the *A. thaliana* TAIR10 reference genome using Methylpy as described in (Schmitz et al. (2013a). A modified lambda DNA sequence was used to assess the quality of prepared libraries. In the added lambda sequence, all non-CG cytosines are unmethylated and CG cytosines are methylated to 5mC. The "non-conversion" rate is used to measure the rate of non-CG cytosines failing to be converted to

thymines after bisulfite treatment. The "5mC nonconversion" is used to estimate the 5mCG cytosines failing to be converted to thymines after TET treatment.

ChIP-seq data analysis. Qualified reads were aligned to the *A. thaliana* TAIR10 reference genome using Bowtie 1.1.1 with following parameters: bowtie -m 1 -v2 --best - -strata --chunkmbs 1024 -S (Langmead et al. 2009). Aligned reads were sorted using SAMtools v 1.2 and clonal duplicates were removed using SAMtools version 0.1.19 (Li et al. 2009).

Metaplot analysis. For metaplot analyses, twenty 50-bp bins were created for both upstream and downstream regions of gene bodies/TEs. Gene bodies/TE regions were evenly divided into 20 bins. Weighted methylation levels were computed for each bin (Schultz et al. 2012).

DMR analysis. Identification of DMRs was performed as described in ref. (Schultz et al. 2015) and adjusted p-value (Benjamini–Hochberg correction) 0.05 was adopted as the cutoff. Only DMRs with at least five DMSs (differential methylated sites) and a 10% absolute methylation level difference within each DMR were reported and used for subsequent analysis. For coverage calculations, each sample was combined with two Col-0 WT replicates to identify DMRs. Each sample was compared with both Col-0 WT replicates separately, and for a DMR to be identified, it must have been identified in both comparisons. Absolute methylation differences of ± (50% for CG, 10% for CHG and CHH) were defined as hyper/hypomethylation, respectively. DMRs overlapping regions with mCG≥5%, mCHG and mCHH ≥1% in both two Col-0 WT replicates were defined as RdDM-like regions. DMRs overlapping regions with mCG ≥5%, mCHG and mCHH <1% in both two Col-0 WT replicates were defined as gbM regions. DMRs overlapping regions with all three contexts less methylated at less than 1% in both Col-0 WT replicates were defined as unmethylated regions. Overlap comparisons were performed using bedtools v2.26.0 (Quinlan and Hall 2010).

Data availability. The data generated from this study have been uploaded to Gene Expression Omnibus (GEO) database and can be retrieved through accession number GSE93024. This chapter contains supplementary tables online at https://doi.org/10.1038/s41467-018-03289-7.

ACKNOWLEDGEMENTS

We thank Nathan Springer for comments and discussions on this study as well as the Georgia Genomics & Bioinformatics Core and the Georgia Advanced Computing Resource Center for technical support. This work was supported by the National Science Foundation (MCB-1650331), by The Pew Charitable Trusts and by the Office of the Vice President of Research at UGA to R.J.S. C.H. was supported by the National Institutes of Health (NIH HG006827) and is an investigator of the Howard Hughes Medical Institute. W.T.J. was supported by a Scholars of Excellence Fellowship from the University of Georgia and National Institute of General Medical Sciences of the National Institutes of Health award number T32GM007103. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

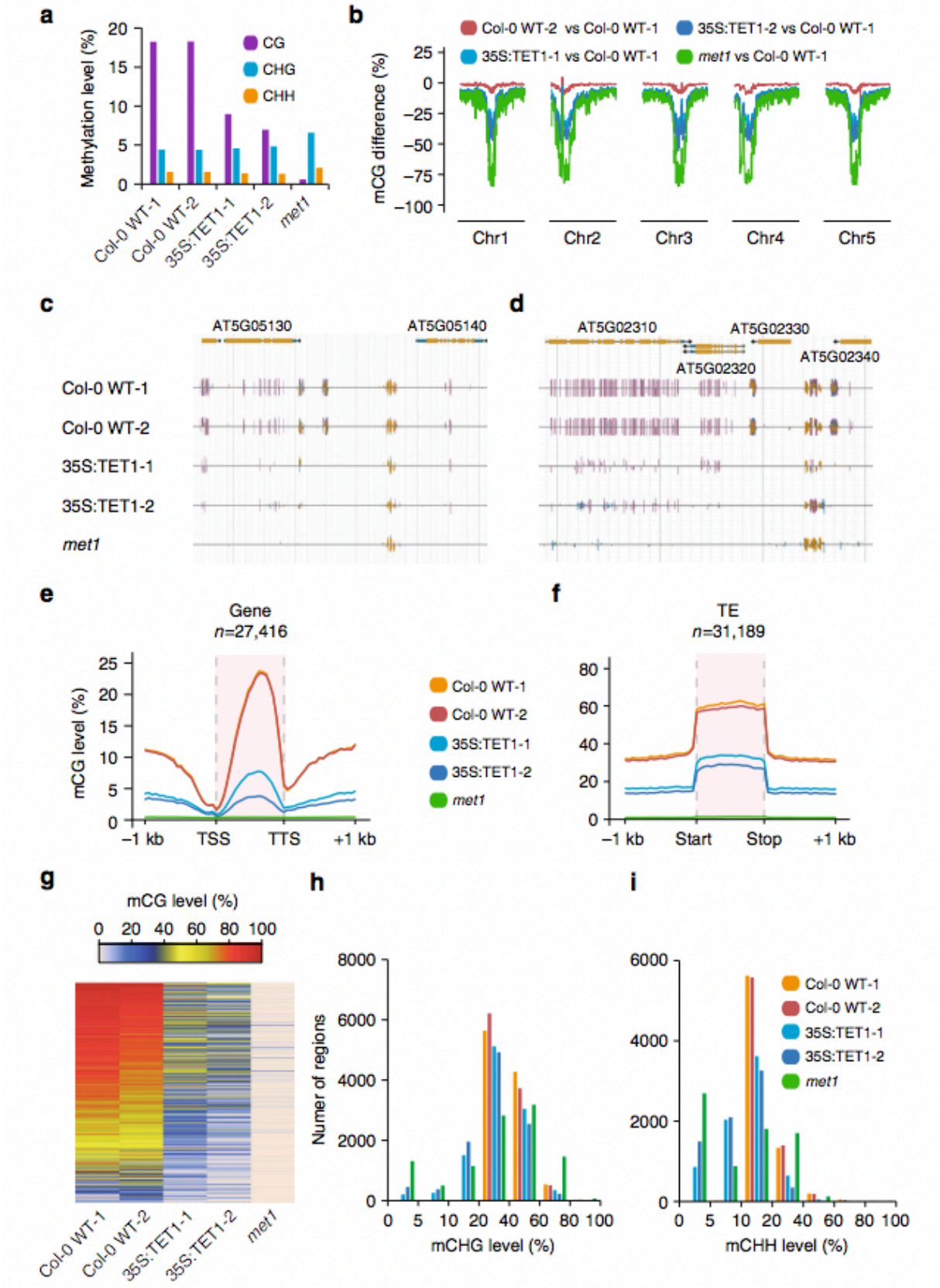

a
Methylation level (%)
CG
CHG
CHH
Col-0 WT-1
Col-0 WT-2
35S:TET1-1
35S:TET1-2
met1
b
Col-0 WT-2 vs Col-0 WT-1
35S:TET1-2 vs Col-0 WT-1
35S:TET1-1 vs Col-0 WT-1
met1 vs Col-0 WT-1
mCG difference (%)
Chr1
Chr2
Chr3
Chr4
Chr5
c
AT5G05130
AT5G05140
d
AT5G02310
AT5G02330
AT5G02320
AT5G02340
e
Gene
n=27,416
mCG level (%)
−1 kb
TSS
TTS
+1 kb
f
TE
n=31,189
Start
Stop
g
mCG level (%)
h
Numer of regions
mCHG level (%)
i
mCHH level (%)

Figure 2.1. Overexpression of hTET1cd-induced global CG demethylation in *A. thaliana*. a Bar graph of global methylation levels in two Col-0 WT replicates, two 35S:TET1 transgenic individuals, and *met1*. **b** Metaplot of CG methylation levels (100-kb windows) across five *A. thaliana* chromosomes. Methylation level differences were defined relative to Col-0 WT-1, and Col-0 WT-2 was used to assess background interference. Genome browser view of methylome profile of two regions (**c**, **d**) of the *A. thaliana* genome (purple vertical lines = CG methylation, blue vertical lines = CHG methylation, and gold vertical lines = CHH methylation). Metagene plots of CG methylation level across **e** gene bodies and **f** transposable elements. **g** Heat map of CG methylation level of CG DMRs. Bar plots of CHG (**h**) and CHH (**i**) methylation levels of CG DMRs that possess non-CG methylation in wild-type individuals.

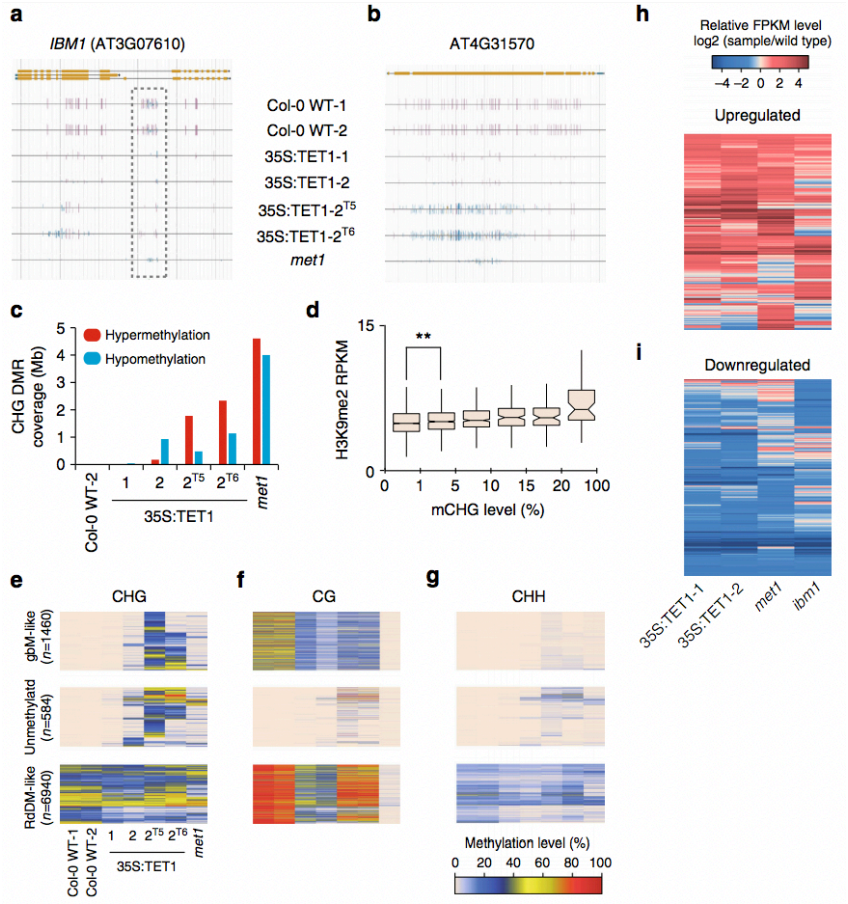


Figure 2.2. Global fluctuation of CHG methylation in 35S:TET1 plants. a Genome browser view of *IBM1* (AT3G07610) in Col-0 WT, four 35S:TET1 transgenic plants, and *met1*. A decrease in CG methylation from coding regions was accompanied by an increase in non-CG methylation. Both CG and non-CG methylation were lost from the large intron (purple vertical lines = CG methylation, blue vertical lines = CHG methylation, and gold vertical lines = CHH methylation). **b** Genome browser view of a representative CHG hypermethylated region. **c** The amount of the genome affected by differential CHG methylation. These DMRs were defined relative to Col-0 WT-1, as Col- 0 WT-2 DMRs were used to assess background interference. **d** Boxplot of H3K9me2 distribution in gbM loci (**t-test, p value <0.01). **e** Heat map of CHG methylation displaying CHG DMRs. Corresponding CG and CHH methylation levels are shown

in **f** and **g**. Heat maps showing log2 transformed FPKM profiles of upregulated genes (**h**) and downregulated genes (**i**) in two 35S:TET1 transgenic individuals, met1, and ibm1 mutants relative to WT.

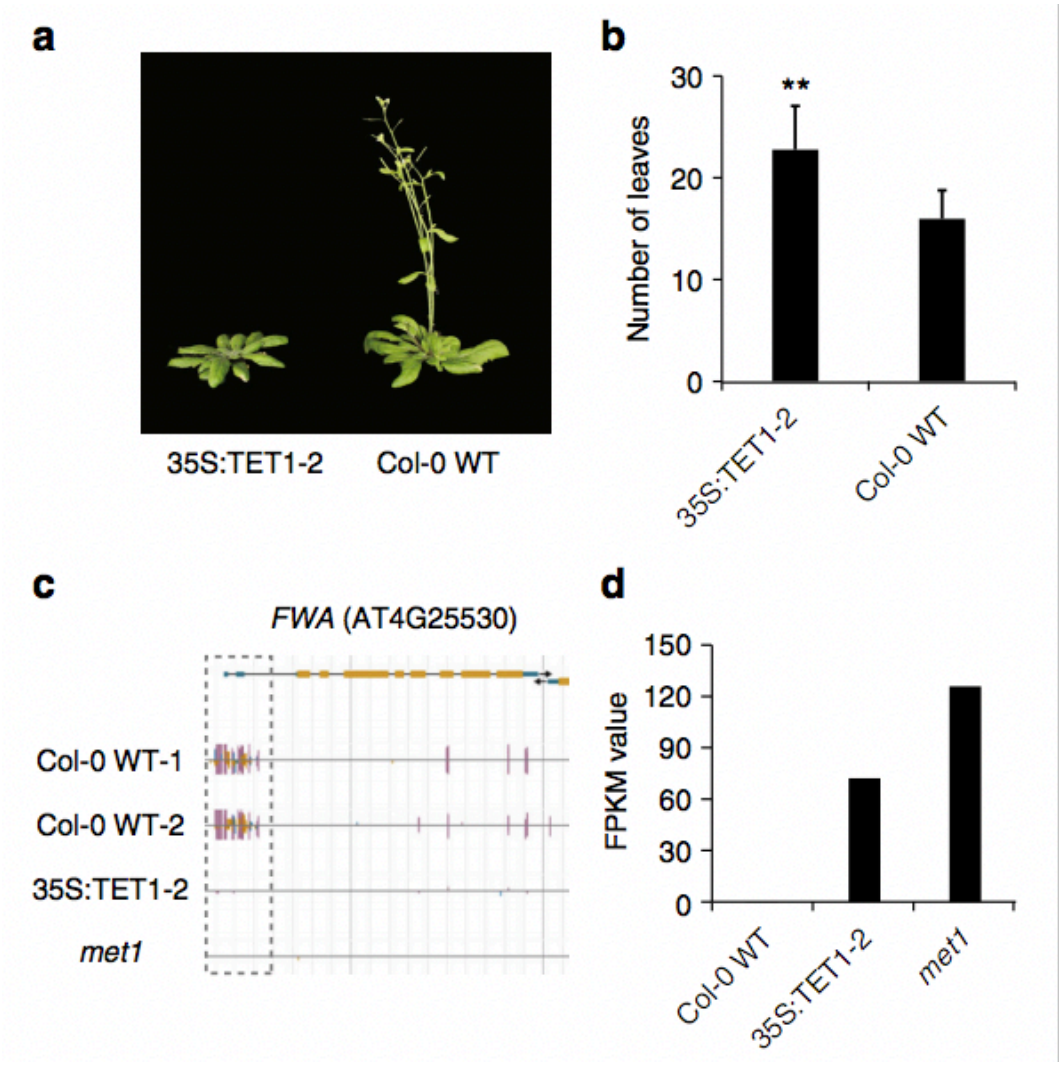


Figure 2.3. 35S:TET1 plants have a delayed flowering phenotype. **a** Photographs of one 35S:TET1-2 transgenic plant and Col-0 WT plant and **b** corresponding number of rosette leaves. Error bars indicate s.d. (**_t_-test, _p_ value <0.01). **c** Genome browser view of *FWA* (AT4G25530). Both CG and non-CG DNA methylation are depleted from the 5' UTR in 35S:TET1-2 plants (purple vertical lines = CG methylation, blue vertical lines = CHG methylation, and gold vertical lines = CHH methylation). **d** Expression level (FPKM) of *FWA*.

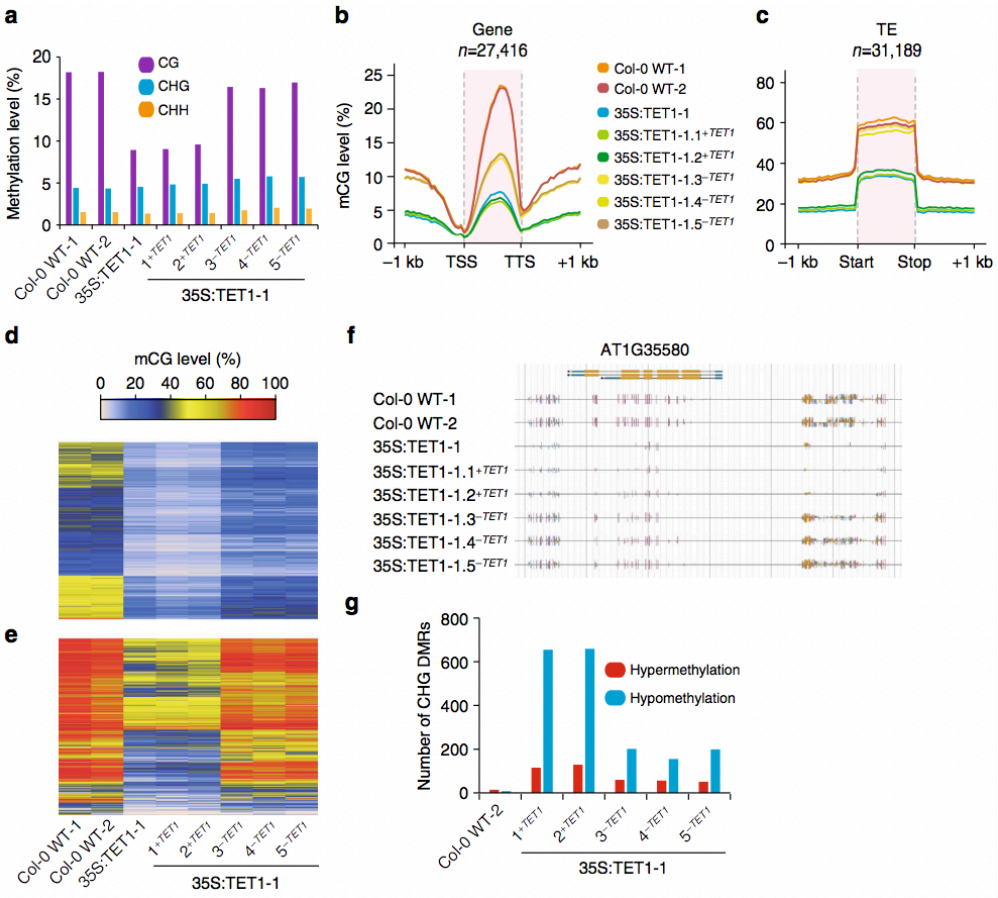


Figure 2.4. Transgenerational demethylation profile of 35S:TET1 individuals. **a** Bar plots of global methylation levels in two Col-0 WT replicates and 35S:TET1-1 plants. Metagene plots of mCG level across **b** gene bodies and **c** transposable elements. Heat map of mCG level of **d** all gbM genes and **e** transposable elements with >20% mCG in wild type. **f** Genome browser view of a methylome profile of a representative region of the *A. thaliana* genome in 35S:TET1-1 individuals. **g** Number of identified CHG DMRs in 35S:TET1 T2 plants. These DMRs were defined relative to Col-0 WT-1, as Col-0 WT- 2 DMRs were used to assess background.

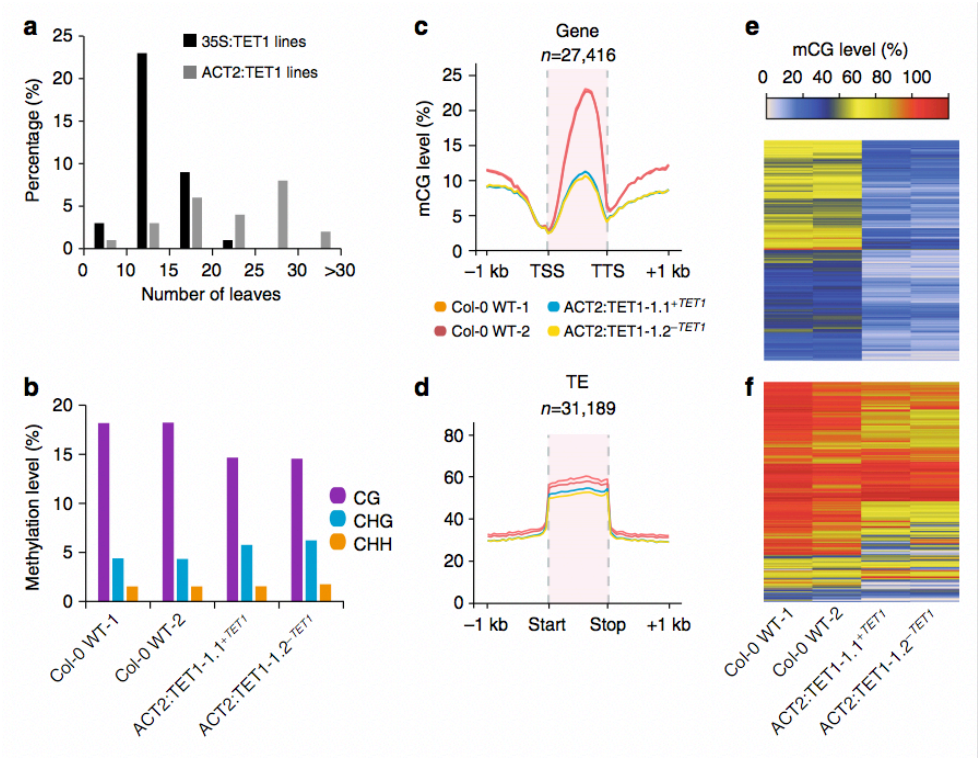


Figure 2.5. Demethylation profile of ACT2:TET1 individuals. a Bar plot of number of rosette leaves in 35S:TET1 and ACT2:TET1 T1 individuals upon flowering. **b** Bar plot of global methylation levels in two Col-0 WT replicates and two ACT2:TET1-1 T2 plants. Metagene plots of mCG level across **c** gene bodies and **d** transposable elements. Heat map of mCG level of **e** all gbM genes and **f** transposable elements with >20% mCG in wild type

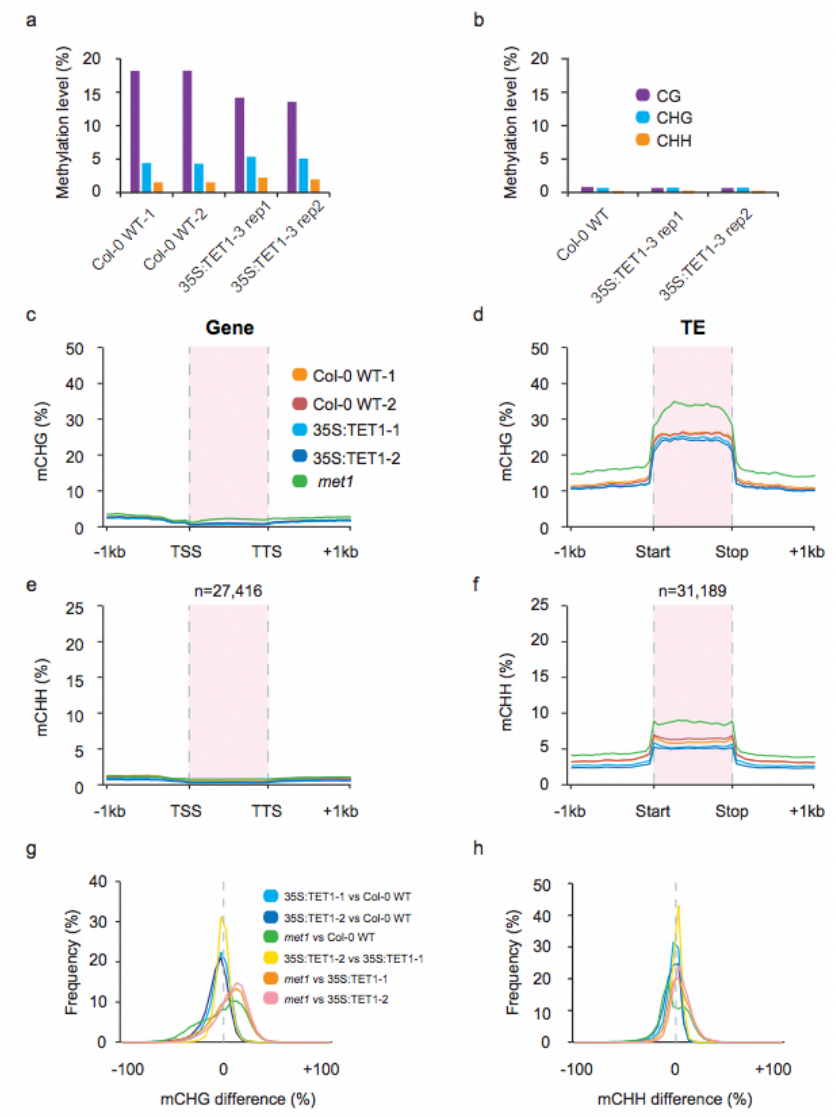


Figure 2.6. Global methylation (**a**) and 5hmC (**b**) levels of Col-0 WT plants and 35S:TET1 transgenic plants. Metagene plots of CHG (**c** and **d**) and CHH (**e** and **f**) methylation level across gene bodies and transposable elements. Frequency plots of mCHG (**g**) and mCHH (**h**) differences in CG DMRs that possess non-CG methylation in wild type. The average value of two Col-0 WT plants was defined as the Col-0 WT.

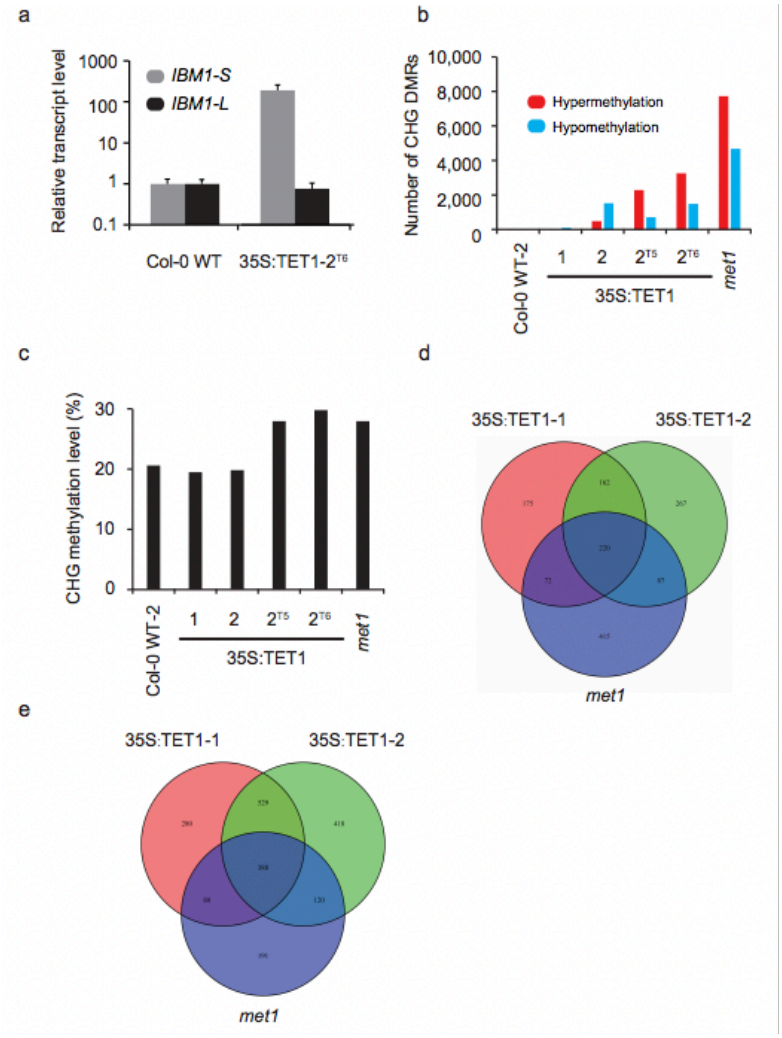


Figure 2.7. (**a**) Relative mRNA expression of *IBM1-L* and *IBM1-S* in Col-0 WT and 35S:TET1-2T6 examined by RT-qPCR. Error bars indicate s.d. (**b**) Number of identified CHG DMRs in 35S:TET1 individuals. These DMRs were defined relative to Col-0 WT- 1, as Col-0 WT-2 DMRs were used to assess background interference. (**c**) Bar plot of CHG methylation levels in identified CHG DMRs. Venn diagrams of upregulated genes (**d**) and downregulated genes (**e**) in two 35S:TET1 transgenic individuals and *met1*.

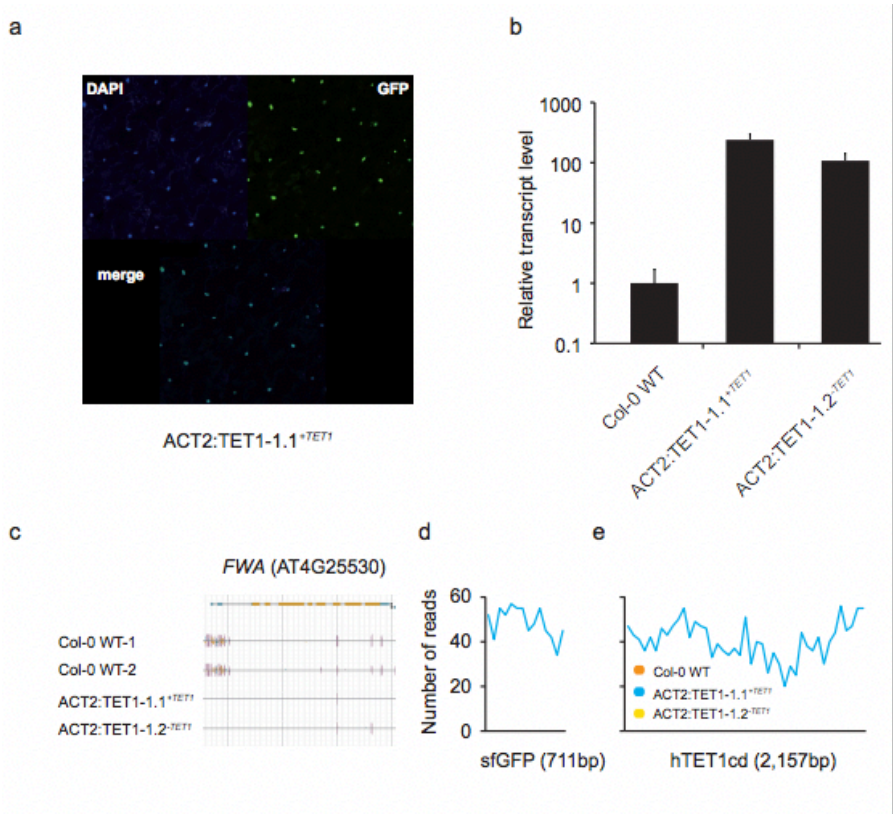


Figure 2.8. (**a**) Confocal microscopy of sfGFP-TET1cd fusion protein in young cotyledons. DAPI and GFP stains show the locations of nuclei and GFP signal, respectively. (**b**) Relative expression of *FWA* examined by RT-qPCR. Error bars indicate s.d. (**c**) Genome browser view of *FWA* (AT4G25530). Read distribution (50bp windows) in (**d**) sfGFP and (**e**) hTET1cd sequences indicate the absence of sfGFP-hTET1cd in Col- 0 WT and ACT2:TET1-1.2-*TET1*.

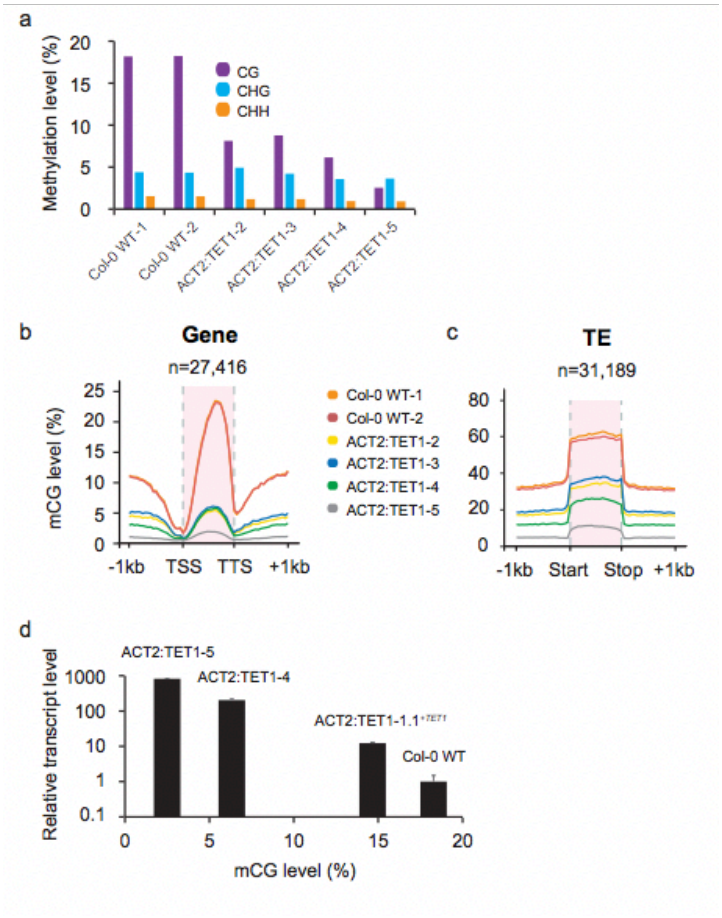


Figure 2.9. (**a**) Bar plot of global methylation levels in two Col-0 WT replicates and four ACT2:TET1 T1 plants. Metagene plots of CG methylation level across (**b**) gene bodies and (**c**) transposable elements. (**d**) Relative expression of sfGFP-hTET1cd examined by RT-qPCR and corresponding mCG level indicate a relationship between transgene expression level and mCG loss. Error bars indicate s.d.

Table 1.

Table S1. Methylome sequencing summary statistics

Sample	Length (bp)	Mapped reads	Non-conversion (%)	Genome coverage	Coverage per-strand
Col-0 WT-1	150	25,858,836	0.30%	32.5	16.3
Col-0 WT-2	150	20,708,437	0.17%	26.0	13.0
35S:TET1-1	150	11,483,070	0.17%	14.4	7.2
35S:TET1-2	150	25,300,675	0.12%	31.8	15.9
35S:TET1-3 rep1	150	15,239,451	0.16%	19.2	9.6
35S:TET1-3 rep2	150	13,976,359	0.13%	17.6	8.8
35S:TET1-2^{T5}	150	10,149,189	0.19%	12.8	6.4
35S:TET1-2^{T6}	75	46,968,226	0.12%	29.5	14.8
35S:TET1-1.1^{+TET1}	150	21,161,361	0.11%	26.6	13.3
35S:TET1-1.2^{+TET1}	150	24,145,320	0.09%	30.4	15.2
35S:TET1-1.3^{-TET1}	150	24,782,835	0.11%	31.2	15.6
35S:TET1-1.4^{-TET1}	150	21,725,685	0.30%	27.3	13.7
35S:TET1-1.5^{-TET1}	150	23,032,079	0.25%	29.0	14.5
ACT2:TET1-1.1^{+TET1}	150	25,564,705	0.16%	32.1	16.1
ACT2:TET1-1.2^{-TET1}	150	24,743,874	0.15%	31.1	15.6
ACT2:TET1-2	75	35,548,625	0.12%	22.3	11.2
ACT2:TET1-3	75	34,292,795	0.12%	21.6	10.8
ACT2:TET1-4	75	42,867,331	0.16%	26.9	13.5
ACT2:TET1-5	75	36,877,672	0.47%	23.2	11.6

Table 2.

Table S2. Transriptome sequencing summary statistics

Sample	Mapped reads	Percent Mapped
Col-0 WT rep1	18,065,032	95.39%
Col-0 WT rep2	24,156,251	96.63%
Col-0 WT rep3	20,987,335	96.48%
35S:TET1-1	19,919,035	95.38%
35S:TET1-2	15,977,645	96.51%
met1	45,615,533	94.61%
ibm1 rep1	20,810,937	94.16%
ibm1 rep2	22,011,997	96.17%
ibm1 rep3	25,241,389	97.05%

Table 3.

Table S3. TAB-seq summary statistics

	Mapped reads	5mC non-conversion (%)	Non-conversion (%)	Genome coverage
Col-0 WT	15,044,614	3.97%	0.28%	18.9
35S:TET1-3 rep1	14,293,516	4.69%	0.28%	18.0
35S:TET1-3 rep2	16,145,140	4.68%	0.29%	20.3

Table 4.

Table S4. ChIP-seq summary statistics

	Uniquely mapped reads	Non-clonal reads
35S:TET1-2^{T6}	6,754,041	4,639,956

References:

1 Law, J. A. & Jacobsen, S. E. Establishing, maintaining and modifying DNA methylation patterns in plants and animals. *Nature reviews. Genetics* **11**, 204-220, doi:10.1038/nrg2719 (2010).

2 Johannes, F. *et al.* Assessing the impact of transgenerational epigenetic variation on complex traits. *PLoS Genet* **5**, e1000530, doi:10.1371/journal.pgen.1000530 (2009).

3 Reinders, J. *et al.* Compromised stability of DNA methylation and transposon immobilization in mosaic Arabidopsis epigenomes. *Genes Dev* **23**, 939-950, doi:23/8/939 [pii] 10.1101/gad.524609 (2009).

4 Cortijo, S. *et al.* Mapping the Epigenetic Basis of Complex Traits. *Science*, doi:10.1126/science.1248127 (2014).

5 Bewick, A. J. *et al.* On the origin and evolutionary consequences of gene body DNA methylation. *Proc Natl Acad Sci U S A* **113**, 9111-9116, doi:10.1073/pnas.1604666113 (2016).

6 Cortijo, S., Wardenaar, R., Colome-Tatche, M., Johannes, F. & Colot, V. Genome-Wide Analysis of DNA Methylation in Arabidopsis Using MeDIP-Chip. *Methods Mol Biol* **1112**, 125-149, doi:10.1007/978-1-62703-773-0_9 (2014).

7 Reinders, J. *et al.* Genome-wide, high-resolution DNA methylation profiling using bisulfite-mediated cytosine conversion. *Genome Res* **18**, 469-476, doi:gr.7073008 [pii] 10.1101/gr.7073008 (2008).

8 Li, Q. *et al.* Genetic Perturbation of the Maize Methylome. *Plant Cell*, doi:10.1105/tpc.114.133140 (2014).

9 Hu, L. *et al.* Mutation of a major CG methylase in rice causes genome-wide hypomethylation, dysregulated genome expression, and seedling lethality. *Proc Natl Acad Sci U S A* **111**, 10642-10647, doi:10.1073/pnas.1410761111 (2014).

10 Tahiliani, M. *et al.* Conversion of 5-methylcytosine to 5-hydroxymethylcytosine in mammalian DNA by MLL partner TET1. *Science* **324**, 930-935, doi:10.1126/science.1170116 (2009).

11 Pastor, W. A. *et al.* Genome-wide mapping of 5-hydroxymethylcytosine in embryonic stem cells. *Nature* **473**, 394-397, doi:10.1038/nature10102 (2011).

12 Ito, S. *et al.* Role of Tet proteins in 5mC to 5hmC conversion, ES-cell self-renewal and inner cell mass specification. *Nature* **466**, 1129-1133, doi:10.1038/nature09303 (2010).

13 Erdmann, R. M., Souza, A. L., Clish, C. B. & Gehring, M. 5-hydroxymethylcytosine is not present in appreciable quantities in Arabidopsis DNA. *G3 (Bethesda)* **5**, 1-8, doi:10.1534/g3.114.014670 (2014).

14 Yu, M. *et al.* Tet-assisted bisulfite sequencing of 5-hydroxymethylcytosine. *Nature protocols* **7**, 2159-2170, doi:10.1038/nprot.2012.137 (2012).

15 Niederhuth, C. E. *et al.* Widespread natural variation of DNA methylation within angiosperms. *Genome biology* **17**, 194, doi:10.1186/s13059-016-1059-0 (2016).

16 Saze, H., Shiraishi, A., Miura, A. & Kakutani, T. Control of genic DNA methylation by a jmjC domain-containing protein in Arabidopsis thaliana. *Science* **319**, 462-465, doi:319/5862/462 [pii]
10.1126/science.1150987 (2008).

17 Miura, A. *et al.* An Arabidopsis jmjC domain protein protects transcribed genes from DNA methylation at CHG sites. *EMBO J* **28**, 1078-1086, doi:emboj200959 [pii]

10.1038/emboj.2009.59 (2009).

18 Rigal, M., Kevei, Z., Pelissier, T. & Mathieu, O. DNA methylation in an intron of the IBM1 histone demethylase gene stabilizes chromatin modification patterns. *EMBO J* **31**, 2981-2993, doi:10.1038/emboj.2012.141 (2012).

19 Lister, R. *et al.* Highly integrated single-base resolution maps of the epigenome in Arabidopsis. *Cell* **133**, 523-536, doi:10.1016/j.cell.2008.03.029 (2008).

20 Soppe, W. J. *et al.* The late flowering phenotype of fwa mutants is caused by gain-of-function epigenetic alleles of a homeodomain gene. *Mol. Cell* **6**, 791-802 (2000).

21 Finnegan, E. J., Peacock, W. J. & Dennis, E. S. Reduced DNA methylation in Arabidopsis thaliana results in abnormal plant development. *Proceedings of the National Academy of Sciences of the United States of America* **93**, 8449-8454 (1996).

22 Ikeda, Y., Kobayashi, Y., Yamaguchi, A., Abe, M. & Araki, T. Molecular basis of late-flowering phenotype caused by dominant epi-alleles of the FWA locus in Arabidopsis. *Plant & cell physiology* **48**, 205-220, doi:10.1093/pcp/pcl061 (2007).

23 Choudhury, S. R., Cui, Y., Lubecka, K., Stefanska, B. & Irudayaraj, J. CRISPR-dCas9 mediated TET1 targeting for selective DNA demethylation at BRCA1 promoter. *Oncotarget*, doi:10.18632/oncotarget.10234 (2016).

24 Maeder, M. L. *et al.* Targeted DNA demethylation and activation of endogenous genes using programmable TALE-TET1 fusion proteins. *Nat Biotechnol* **31**, 1137-1142, doi:10.1038/nbt.2726 (2013).

25 Vojta, A. *et al.* Repurposing the CRISPR-Cas9 system for targeted DNA methylation. *Nucleic Acids Res*, doi:10.1093/nar/gkw159 (2016).

26 Mendenhall, E. M. *et al.* Locus-specific editing of histone modifications at endogenous enhancers. *Nat Biotechnol* **31**, 1133-1136, doi:10.1038/nbt.2701 (2013).

27 Liu, X. S. *et al.* Editing DNA Methylation in the Mammalian Genome. *Cell* **167**, 233-247 e217, doi:10.1016/j.cell.2016.08.056 (2016).

28 Heard, E. & Martienssen, R. A. Transgenerational epigenetic inheritance: myths and mechanisms. *Cell* **157**, 95-109, doi:10.1016/j.cell.2014.02.045 (2014).

29 Clough, S. J. & Bent, A. F. Floral dip: a simplified method for Agrobacterium-mediated transformation of Arabidopsis thaliana. *The Plant journal : for cell and molecular biology* **16**, 735-743 (1998).

30 Urich, M. A., Nery, J. R., Lister, R., Schmitz, R. J. & Ecker, J. R. MethylC-seq library preparation for base-resolution whole-genome bisulfite sequencing. *Nature protocols* **10**, 475-483, doi:10.1038/nprot.2014.114 (2015).

31 Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet. journal* **17**, pp. 10-12 (2011).

32 Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114-2120, doi:10.1093/bioinformatics/btu170 (2014).

33 Schmitz, R. J. *et al.* Epigenome-wide inheritance of cytosine methylation variants in a recombinant inbred population. *Genome research* **23**, 1663-1674, doi:10.1101/gr.152538.112 (2013).

34 Kim, D. *et al.* TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome biology* **14**, R36, doi:10.1186/gb-2013-14-4-r36 (2013).

35 Trapnell, C. *et al.* Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nature biotechnology* **28**, 511-515, doi:10.1038/nbt.1621 (2010).

36 Schultz, M. D., Schmitz, R. J. & Ecker, J. R. 'Leveling' the playing field for analyses of single-base resolution DNA methylomes. *Trends in genetics : TIG* **28**, 583-585, doi:10.1016/j.tig.2012.10.012 (2012).

37 Schultz, M. D. *et al.* Human body epigenome maps reveal noncanonical DNA methylation variation. *Nature* **523**, 212-216, doi:10.1038/nature14465 (2015).

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

CHAPTER 3

MULTIPLEX GENE-EDITING IN ARABIDOPSIS THALANA USING MB3CAS12A

Introduction:

The use of bacterial Cas nucleases for genome-editing has enabled facile manipulation of numerous plant genomes, including *Arabidopsis thaliana*, maize, tomato, soybean and rice [1-5]. Part of the CRISPR-Cas bacterial immune system, Cas enzymes such as Cas9 function as DNA endonucleases which recognize their cognate DNA by complementary base pairing with a tracrRNA and target-specific crRNA. For genome-editing purposes in eukaryotic organisms, the tracrRNA and crRNA are often fused into a single chimeric guide RNA (gRNA) [6]. Cas9 from *Streptococcus pyogenes* (SpyCas9) is by far the most widely adopted enzyme currently used for plant genome-editing, due to its high activity at room temperature in numerous plant species, acceptable rates of off-targeting, incorporation into numerous plant-compatible vectors, and the availability of several engineered variants [1, 3]. Through the use of sequence-specific chimeric single-guide RNA scaffold (sgRNA), sequences immediately adjacent to the protospacer adjacent motif (PAM) of 5' NGG 3' can be targeted for DNA binding and blunt-end cleavage [6].

As sgRNAs must match their DNA targets with sequence specificity, Pol-III promoters are typically required to express sgRNAs, due to the lack of 5' and 3' processing of Pol-III transcripts [2, 4]. However, Pol-III promoters suffer from a lack of tissue and environmental specificity, nuclear export of transcripts, and contain cryptic termination poly(T) termination sequences, which are present with the sgRNA scaffold and may severely limit expression of

sgRNA transcripts [7]. Additionally, while the three-base pair PAM of Cas9 occurs on average once every sixteen base-pairs, finding suitable gRNA sequences in AT-rich regions such as promoters and some genic regions can be challenging.

Due to the inherent lack of RNase activity in Cas9 nucleases, approaches for targeting of multiple genomic regions using Cas9 (multiplex genome-editing) require the use of promoter stacking for individual gRNA cassettes, or the recruitment and/or introduction of sequence-specific RNA processing machinery. While promoter stacking is commonly utilized in plants for multiplex-genome editing, construction of stacked constructs requires numerous cloning steps and frequently utilize a small number of repeated promoter sequences, resulting in repetitive clones that are prone to bacterial recombination.

Alternatives to promoter stacking using Pol-III promoters include the insertion of RNA sequences that result in RNA cleavage at desired locations, either through the recruitment of endogenous RNases or self-cleavage mechanisms. Successful recruitment of endogenous RNAses for gRNA processing is routinely accomplished by the introduction of tRNAs flanking gRNA cassettes, which serve to facilitate processing by Rnase P and Rnase Z. Alternatively, ribozyme sequences flanking gRNA cassettes can also be used, albeit with lower reported efficiency when compared to tRNA-based gRNA processing [8].

The use of exogenous RNA processing machinery, such as CRISPR-associated RNA endoribonuclease 4 (Csy4), has also been successfully adapted to process gRNA cassettes from a single transcript molecule in plants, with the addition of a flanking 20 bp Csy4 recognition sequence [8]. While the use of exogenous RNases requires expression of an additional protein, the use of Csy4 to process polycistronic gRNAs has been demonstrated to outperform the use of ribozymes and native tRNA processing machinery for obtaining genome-edited plants [8].

While Cas9 remains the most utilized CRISPR-Cas nuclease for genome-editing purposes, an alternative CRISPR-Cas based system for genome-editing, CRISPR-Cas12a (formerly called Cpf1), has been utilized for genome-editing purposes in animals, plants, and fungi. Cas12a enzymes employ unique mechanisms for RNA processing, DNA cleavage, as well as distinct PAM sequences [9]. Cas12a enzymes requiring only a short 20-bp crRNA scaffold for target recognition and can self-process multiple crRNAs from a single molecule, through the utilization of an RNase domain for specifically processing crRNA arrays [10]. Cas12a enzymes recognize a T-rich PAM, (typically 5' TTTV 3'), enabling easier targeting of AT-rich genomic regions when compared to Cas9, and additionally produce staggered cuts with a 5-bp overhang, often resulting in the formation of larger indels when compared to Cas9-based genome-editing [9]. Studies examining off-target genome-editing have shown reduced off-target activity with Cas12a enzymes in mammalian systems when compared to Cas9, a result further supported in studies involving rice [11, 12 13].

Several previous studies have shown the efficacy of LbCas12a based gene-editing in *A. thaliana* by introducing single or dual crRNAs per gene target using a double-ribozyme system [14, 15]. However, the successful application of alternative Cas12a orthologs, as well as the use of Cas12a for massively multiplexed genome-editing remains unexplored in plants. Here, we report the use of Mb3Cas12a for genome-editing in plants and demonstrate the one-step introduction of up to 13 crRNAs in a single ORF for massively multiplexed genome-editing. When coupled with a modified heat-shock protocol previously described in plants [16], we report the isolation of homozygous T2 populations with up to six different edits at unique target sites from a single T1 parent plant.

Results:

Mb3Cas12a is a functional RNase and DNase in *A. thaliana*

To test the functionality of Mb3Cas12a in plant systems, we constructed binary vectors based on the pKI1.1R vector backbone. The pKI1.1R backbone was selected due to the high efficiency of the *RIBOSOMAL PROTEIN 5A* (*RPS5A*) promoter for Cas9-based mutagenesis, which was used to drive expression of HuMb3Cas12a for all experiments. For crRNA array expression, the efficiency of Pol-II and Pol-III promoters was compared by expressing identical crRNA arrays under either the U6-26 Pol-III promoter or a Pol-II promoter. To identify suitable Pol-II promoters for crRNA expression, RNA-seq data from the shoot apical meristem (SAM) at various developmental timepoints was analyzed to identify candidate genes with the highest expression across developmental time [17]. The expression of AT3G16640, or *TRANSLATIONALLY CONTROLLED TUMOR PROTEIN1 (TCTP1)*, was to have high, consistent expression across developmental time in the SAM, with an average transcripts per million (TPM) count of 3,157 compared to TPMs of 1,309 and 325 for *RPS5A* and POLYUBIQUITIN 10 (*UBI10)*, respectfully. Additionally, the expression patterns of multiple *AtTCTP1* upstream region fragments as well as tissue-specificity of transgenic expression of *TCTP1* has been previously experimentally validated [18]. To assess the ability of Mb3Cas12a for genome-editing, a four-crRNA array consisting of two guides targeting the floral regulator genes *APATELLA1* (*AP1*; AT1G69120) and two targeting *CAULIFLOWER* (*CAL*, AT1G26310) using both *pU6-26* and *pTCTP* was transformed via floral dip and T1 transformants screened via hygromycin selection and grown at 22°C with a 16-hour photoperiod. A total of 64 and 63 plants containing Pol-III driven or Pol-II driven *AP1-CAL* arrays were analyzed for the presence of

mutant phenotypes and genotypes, respectively. Unexpectedly, no mutant phenotypes or genotypes were detected for both Pol-II and Pol-III transformed lines.

To investigate if temperature-sensitive activity of Mb3Cas12a was resulting in undetectable editing activity, T2 seeds from a single T1 transformant was split into three groups and subject to three heat treatments (**Figure 3.2A**), and the emergence of expected phenotypes was observed upon transition to reproductive growth. Plants subjected to a 30°C heat stress during the entirety of vegetative growth, in addition to individuals grown at room-temperature, failed to exhibit any observable mutant genotypes and phenotypes, while approximately 10% (1/10) of T2 plants transformed with a Pol-II driven array subject to repeated 37°C heat shocks exhibited observable *ap1* mutant phenotypes and genotypes (**Figure 3.2B**). The observed molecular genotypes and phenotypes were stable for at least one generation post heat-treatment, indicating that 37°C heat treatment is sufficient for detectable and inherited edits for Pol-II driven arrays (**Figure 2C**).

Assessment of editing of inflorescences using a single crRNA array

To investigate the efficiency and specificity of multiplex genome-editing, the floral regulator genes *APETALA 2* (*AP2)*, *PISTILLATA* (*PI)*, and *AGAMOUS* (*AG)* were selected for editing using a single Pol-II array. Successful editing of these genes in all floral meristem cells would induce distinct phenotypes, as well as indicate editing of the cells which give rise to the floral primordia. Additionally, the proportion and identification of heterozygous genotypes in T1 plants can be ascertained by phenotyping T2 progeny, as homozygous or biallelic mutations in *PI* and *AG* result in sterility.

Four unique guides were used to target the *AP2* and *PI* loci, whereas two overlapping guides containing a single mismatch at positions 13 and 18 were used to target the *AG* locus

(**Figure 3.3A**). Twelve independent T1 plants were subjected to 37°C heat shocks (**Figure 3.2A**), and the emergence of floral phenotypes were assessed upon transition to reproductive growth. Upon flowering, three plants exhibited sectoring with altered floral morphology, phenocopying previously described *ap2* and *pi* mutants, as well as *ap2 pi* double mutants (**Figure 3.3B**). Interestingly, no *ag*-like sectors were observed on any T1 plants, indicating reduced editing activity at the *ag* locus when compared to the *AP2* and *PI* loci.

The inheritance of induced mutations following 37°C heat treatments from these three lines were subsequently assessed by collecting and growing T2 seeds at a constant temperature of 22°C, to eliminate potential additional Mb3Cas12a endonuclease activity. Upon flowering, 16-38% of T2 plants exhibited *ap2*-like flower phenotypes (**Figure 3.3B**). Sanger sequencing performed at the *AP2* locus revealed homozygous deletions at a single crRNA target site ranging from –4bp to -18bp, supporting previous observations that Cas12a mediated genome-editing primarily results in the formation of deletions between 5 and 30 bp in size. Despite the greatly reduced fertility of *ap2* mutants, homozygous mutations were readily observed in T2 populations, indicating a high degree of inheritance of induced mutations, despite negative selective pressure.

To ascertain the level heterozygosity in T1 parent plants, T2 populations were assessed for editing at the *PI* and *AG* loci. As *pi* mutants are sterile due the lack of anther formation, and *ag* mutants fail to produce anthers and pistils, homozygous *pi* and/or *ag* mutants observed in T2 individuals would result from residual heterozygosity in their respective T1 parents. Out of 180 and 190 observed T2 plants for lines 1 and 2, only two individuals for each line were observed to have homozygous or biallelic mutations at the *PI* locus for each line, indicating a low degree of heterozygosity observed with 37°C heat treatments. Interestingly, Sanger sequencing of the *PI*

locus revealed deletions ranging from -4bp to -996 bp in *pi* T2 individuals, indicating Mb3Cas12a can induce a large range of insertion/deletion (indel) sizes in *A. thaliana*. A similar number of *ag* phenotypes were observed in T2 populations, indicating a similar rate of transmission of heterozygous *AG* and *PI* alleles in T1 plants. As *ag* mutant sectors were not observed in T1 individuals, the rate of Mb3Cas12a-induced editing at crRNA sites containing single mismatches is substantially reduced.

Efficacy of non-canonical PAM editing using Mb3Cas12a

Previous characterization of Mb3Cas12a activity in mammalian cell culture systems revealed the ability to recognize alternate or non-canonical PAM sequences for genome-editing. To determine the efficiency of non-canonical PAM sequences for editing in *A. thaliana*, four crRNAs targeting the *SHOOT APICAL MERISTEM ARREST 1 (SHA1)* locus with either TTN or CTN PAM motifs were transformed and T1 individuals subject to 37°C heat shocks (**Figure 3.2A**). Of 23 and 13 plants observed with TTN and CTN PAM arrays respectively, no detectable editing of the *SHA1* locus was observed in T1 individuals. To assess the potential of hemizygous editing in the SAM of T1 individuals, fifteen seeds from each T1 transformant was grown at a constant temperature of 22°C. No detectable editing was observed from any T2 individuals sampled, indicating low non-canonical PAM activity of Mb3Cas12a in *A. thaliana* at the *SHA1* locus. While minimal activity was observed at the *SHA1* locus, testing of non-canonical PAM editing at additional target sites is necessary to confirm low non-canonical PAM activity in *A. thaliana.*

Multiplexed Mb3Cas12a mutagenesis and off-target analysis using Mb3Cas12a

Previous studies conducted in plants have used CRISPR-Cas mutagenesis to rapidly engineer *cis*-regulatory element variation via Cas9-multiplexed gRNA systems [31]. To investigate

the use of Mb3Cas12a for multiplexed mutagenesis of *cis* regulatory regions, a crRNA array targeting 830 bp of the *FLOWERING LOCUS T* (*FT*) enhancer BlockE was constructed (**Figure 3.4A**). To maximize the likelihood of successful editing at multiple target sites, crRNAs were subject to target efficiency calculations using the CRISPR-DT webtool [28]. Thirteen crRNAs with a predicted efficiency of greater than 0.8 were selected and cloned as a single dsDNA block. Nineteen independent T1 lines were subjected to 37°C heat shocks as previously described, and editing efficiency was assessed by sequencing pools of T2 individuals from individual lines grown at a constant temperature of 22°C. Upon sequencing, three pools contained edits at the BlockE region. Of lines containing edits, approximately 60% of sequenced individuals contained edits at least one single crRNA site. In total, editing was detected at five unique crRNA targets, with deletions detected ranging in size from 6 to 581 bp (**Figure 3.4B**).

To assess potential off-target effects from the introduction of multiple crRNAs, whole genome-sequencing (WGS) was performed on seven T2 BlockE individuals from two independently transformed lines at an average sequencing depth of ~41X coverage ranging from ~19x to ~62x coverage. Putative off-target deletions were identified using DeepVariant 1.0. A total of 329 high-confidence deletions were detected across all sequenced individuals, with at least 96% of detected deletions shared between all individuals, indicating shared deletions inherited from T_0 parents (**Figure 3.5**). To assess the possibility of Cas12a off-target editing at uniquely detected deletions, these variants were filtered for deletion sizes typical for Cas12a-mediated genome-editing (between 3 and 25 bases), and sites further filtered for proximity to a Cas12a PAM site. No unique deletions containing a 3-25 bp deletion and within 20 bp of a Cas12a PAM site were observed in the seven single individuals sequenced, supporting previous

studies indicating the high-specificity of Cas12a-mediated genome-editing in eukaryotic systems [32].

Discussion:

In this study, we created and optimized a simplified CRISPR-Cas genome-editing system for multiplex targeting, using the Mb3Cas12a nuclease. As previous studies have indicated temperature-sensitivity for Cas12a-mediated DNase activity, we first investigated which temperature profile would enable efficient editing in *A. thaliana* [11]. Our results show no detectable editing at temperatures of 21°C and 30°C. Interestingly, previous studies have shown activity of LbCas12a at 30°C in *A. thaliana* and even 22°C in rice, suggesting that temperature sensitivity for plant-genome editing varies widely depending on the Cas12a ortholog and plants systems used [14]. When applying a modified heat-shock protocol as previously described in Leblanc et al 2016, we observed potent levels of editing in recovered somatic tissue, indicating the minimum temperature for DNase activity of Mb3Cas12a is between 31-37°C [16].

Using a U6-driven crRNA array, we first tested the ability of Mb3Cas12a to properly process multiple crRNAs from a single RNA transcript, by creating mutations at two distinct sites at the *AP1* locus. While we were able to observe somatic mutations in rosette leaf samples, poor germline transmission of edited alleles was observed using this system. As previous studies in mammalian systems indicated increased mutagenesis efficiency using Cas12a when crRNAs were driven by a Pol-II promoter, we performed an exhaustive search for promoters with high expression levels, specifically in the SAM of *A. thaliana* [22,27]. Upon the utilization of the 0.3 kb *AtTCTP1* promoter, we observed genome-editing in both somatic and germline tissue of T_2 plants. As *TCTP1* is one of the most highly expressed genes in dividing tissues, and particularly

the SAM, it is likely the increased germline mutagenesis observed when using a *pAtTCTP1*-driven crRNA expression system is due to increased abundance of crRNA molecules available for processing in the SAM.

We next investigated the efficiency of targeting multiple loci simultaneously using a Pol-II driven array. Upon flowering, we observed both single and double mutants targeting the *AP2* and *PI* loci. T_2 progeny from these edited plants exhibited high levels of homozygous editing of the *AP2* locus, indicating efficient germline transmission of edited alleles, despite reduced fitness of *ap2* flowers. Interestingly, we observed relatively few *pi* and *ap2 pi* T_2 plants. As *pi* mutants are sterile, this observation supports a high level of biallelic or homozygous editing in T_1 plants, which were unable to be transmitted through the germline due to sterility. To test the effects of single mismatches on editing rates, we simultaneously introduced two mismatched guides with a single mismatch targeting the *AG* locus. When compared to non-mismatched guides targeting *AP2*, editing rates at *AG* were 50-fold lower, indicating a severe penalty for mutagenesis using singly mismatched crRNAs.

To test the ability of massively-multiplex editing of genomic regions, using a one-step cloning approach, we introduced thirteen crRNAs in a single array for targeting the BlockE region, adjacent to the *FT* locus into *A. thaliana*. BlockE has previously been shown to function as a downstream enhancer of *FT*, and artificial deposition of DNA methylation at BlockE has resulted in a delayed flowering phenotype [29]. T_2 plants recovered were observed to have deletions at numerous target sites, ranging in size from 6-581 bp. Interestingly, no delayed flowering phenotypes in BlockE-edited lines when grown in long-day conditions were observed. As the introduction of large numbers of crRNAs could potentially increase off-targeting rates, we profiled the presence of off-target editing in seven T_2 plants using WGS. As no off-target edits

were detected in this population of edited plants, the introduction of multiple crRNAs with predicted target specificity is unlikely to drastically affect off-targeting rates genome-wide.

The simplicity of construction of multiplex targeting using our system, combined with the efficiency of inducing mutations when using a high-temperature heat-shock protocol, enables a simplified, efficient massively multiplexed mutagenesis in *A. thaliana*, without the introduction or recruitment of endogenous RNases or exogenous RNA processing proteins. By reducing the size and repeat content of transgenic material needed for targeting of Cas proteins to multiple genomic sites, this Mb3Cas12a expression system enables more facile multiplex editing in plant-based systems.

Methods:

Plant Transformation, Selection and Heat Treatment:

Vectors were transformed into *Agrobacterium tumefaciens* strain C58C1 by electroporation, and transformation of Col-0 *A. thaliana* inflorescence tissue was performed using the floral dip method [19]. Seeds were collected upon plant senescence, and transgenic plants were identified via selection on ½ strength LS plates supplemented with Hygromycin B (25 µg/mL) as described previously [20]. Approximately 10-14 days following germination, resistant seeds were selected and planted into SunGro #3B Mix. Ten days following transplanting, plants were subjected to eight rounds of heat stress as described previously with minor modifications. Plants were subsequently recovered at room temperature during reproductive growth and grown at 21°C with a 16-hour photoperiod.

Cloning:

The ORF containing NLS-Mb3Cas12a-3xHA was subcloned from 35-pcDNA3-huMb3Cpf1 (a gift from Feng Zhang) into the pKI1.1R backbone by GenScript, creating Mb3Cas12a-pKI1.1R. To construct TCTP-Mb3Cas12a-pKI1.1R, 1 μg of Mb3Cas12a-pKI1.1R was doubly digested using 50 units of *Apa*I (New England Biolabs), and 2 units of EcoRI-HF (New England Biolabs). The native 0.3-kb TCTP Promoter and CaMV poly(A) signal cassette was synthesized as a single FragmentGENE DNA fragment (Genewiz), doubly digested with *Apa* I and *Eco* RI and ligated to the Mb3Cas12a-pKI1.1R using 3,000 units of T7 DNA Ligase (New England Biolabs). crRNA arrays were cloned by amplifying synthesized FragmentGENE or PriorityGene (Genewiz) fragments with primers crRNA F (5' GTAGTCGTAGTCGGTCTC 3') and crRNA R (5' GGACTCCGTGGATACAAA 3') using Q5 DNA Polymerase (New England Biolabs). The resulting amplicons were cleaned using a Monarch PCR & DNA Cleanup Kit (New England Biolabs). Approximately 300 ng of cleaned PCR product were digested with 12 units of BsaI-HFv2 (New England Biolabs), gel-purified using a Monarch PCR & DNA Cleanup Kit and inserted into *Aar* I-digested Mb3Cas12a-pKI1.1R with T7 DNA Ligase (New England Biolabs), and transformed into DH10B cells using electroporation. All clones were sequence verified using Sanger sequencing.

DNA Extraction and Analysis:

DNA for Sanger sequencing analysis was extracted from rosette leaves using as described in [21]. Candidate regions were amplified with Q5 DNA Polymerase and sequenced using Sanger sequencing (Macrogen USA). Mutated alleles were deconvoluted from Sanger sequencing traces using CRISP-ID [22].

DNA Extraction and Whole Genome Sequencing Analysis:

DNA for Illumina NGS analysis was extracted using a DNeasy Plant Mini Kit (Qiagen) following the manufacturer's instructions. Genomic DNA library preparation was carried out using the protocol outlined in ref. 30 and paired-end 150bp reads were sequenced using an Illumina NovaSeq 6000. Raw fastq reads were preprocessed using fastp v0.20.1, aligned to the TAIR10 genome using Bowtie 2.3.5.1 and duplicates removed using Sambamba v 0.6.6. Variants were called using DeepVariant v1.0, and gVCF files merged and joint variants called using GLnexus v1.2.7 [26]. Obtained merged gVCF files were filtered for deletion-specific variants using Jvarkit v20200206, and sites containing sufficient data for all sequenced lines and a minimum sequencing depth of 10 were filtered using VCFtools v0.1.16.

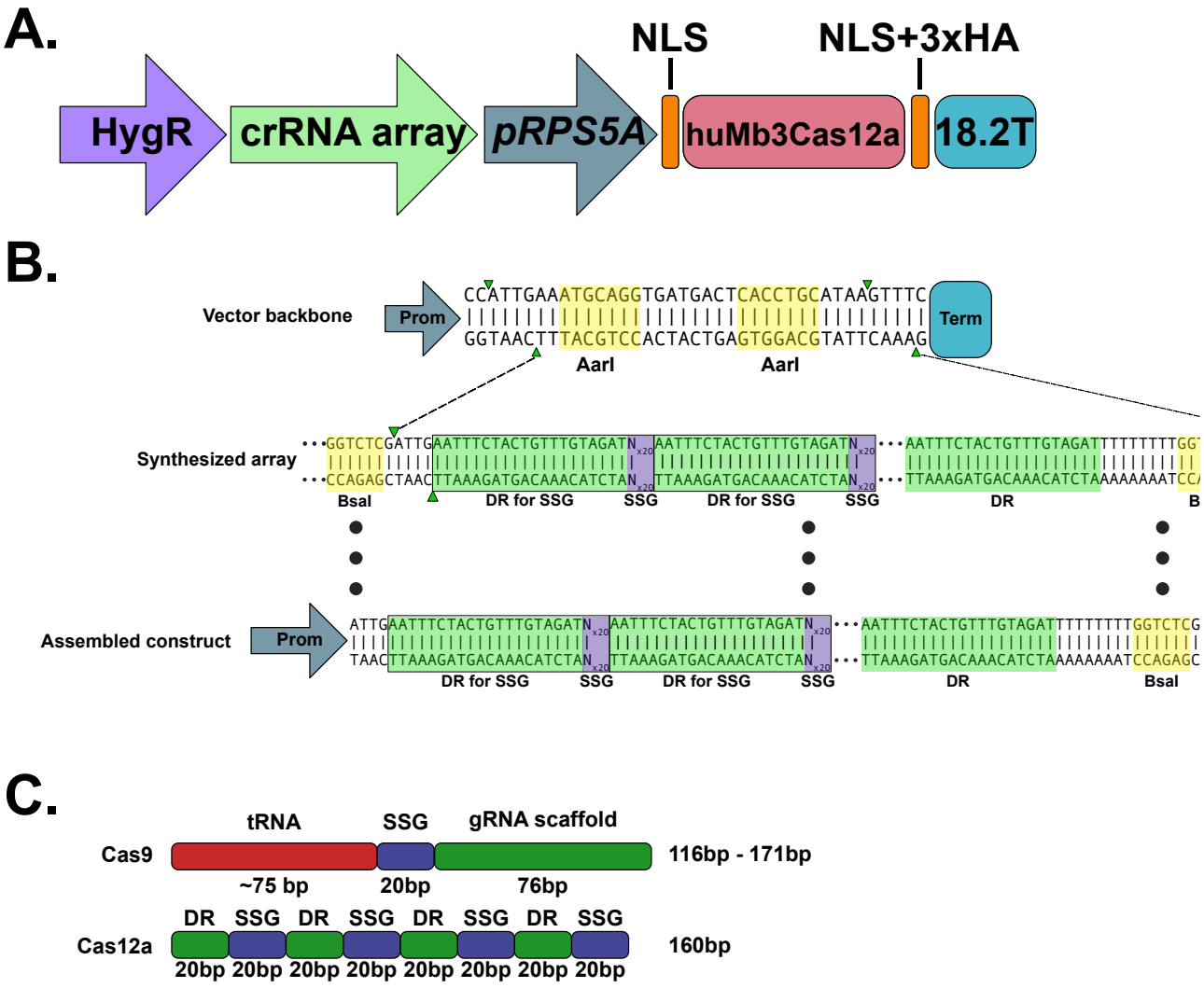


Fig 3.1. Construction and cloning protocol of Mb3Cas12a vectors

A. Vector design consisting of a hygromycin resistance cassette, crRNA expression cassette, and huMb3Cas12a driven by the At*RPS5A promoter*. NLS = nuclear localization signal; 18.2T = Hsp18.2 terminator from *Arabidopsis thaliana*.

B. Cloning strategy for insertion of custom crRNA arrays. Custom crRNA arrays are cloned in a single step using Golden-Gate Assembly via *AarI* restriction sites. DR = Mb3Cas12a direct repeat sequence; SSG = sequence-specific guide

C. Size comparison (bp) of SSG cassettes for Cas9 and Cas12a-based multiplexing systems in plants

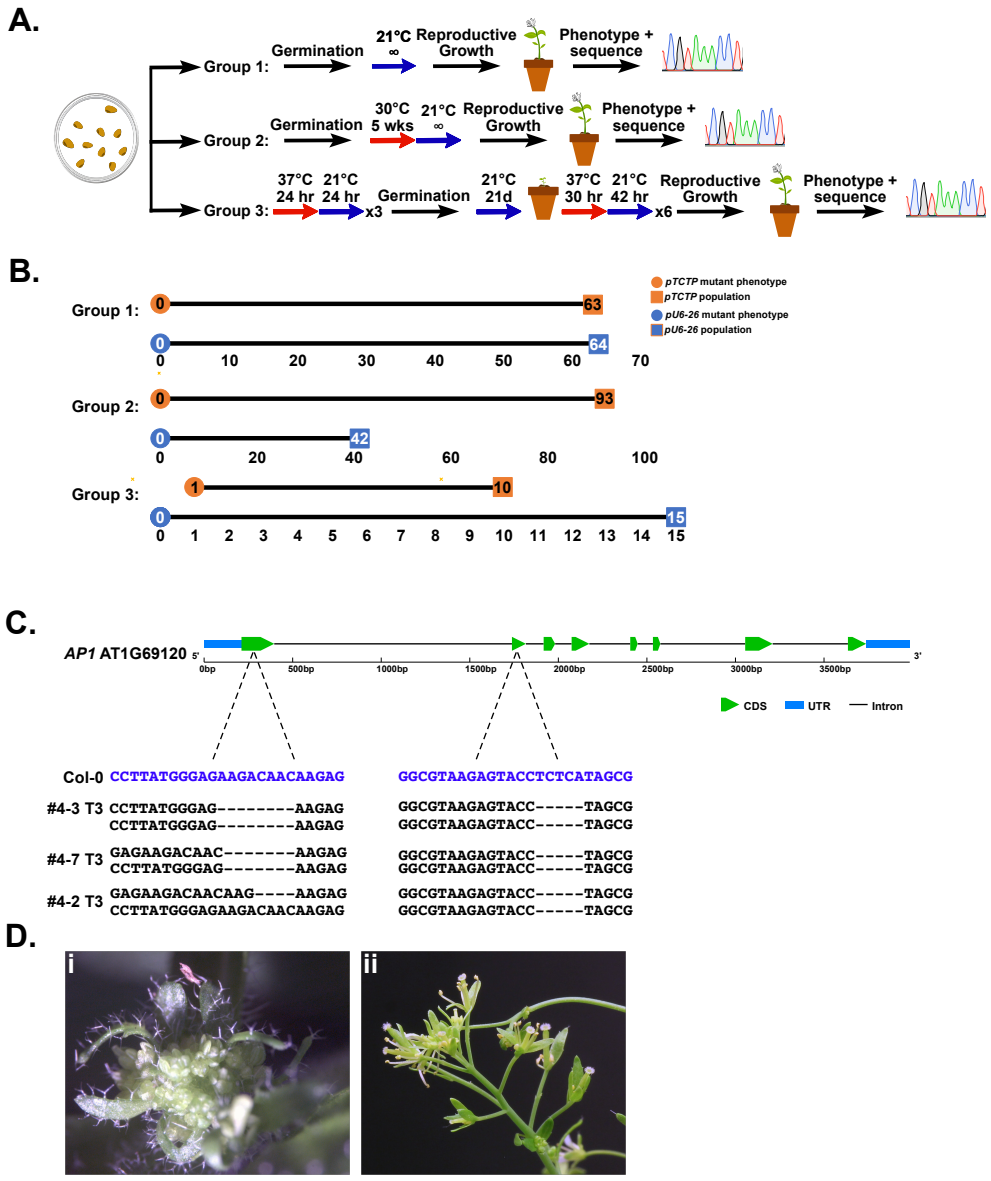


Fig 3.2. Assessment of Mb3Cas12a activity at different growth temperatures

A. Identical T2 seed pools were subjected to three temperature conditions: constant growth at 21°C **(1)**, growth at 30°C until flowering was observed **(2)**, or repeated heat shocks at 37°C **(3)**.

B. Assessment of mutant phenotypes observed in T2 pools using either *pU6-26* or *pAtTCTP* for crRNA array expression targeting AP1 and CAL using various temperature conditions. Pools

with mutant phenotypes were saved and propagated an additional generation to obtain stable T3 lines

C. Sanger sequencing of T3 lines reveal deletions crRNA target sites

D. Edited lines phenocopy *ap1* with respect to inflorescence meristem architecture **(i)** and floral architecture **(ii)**

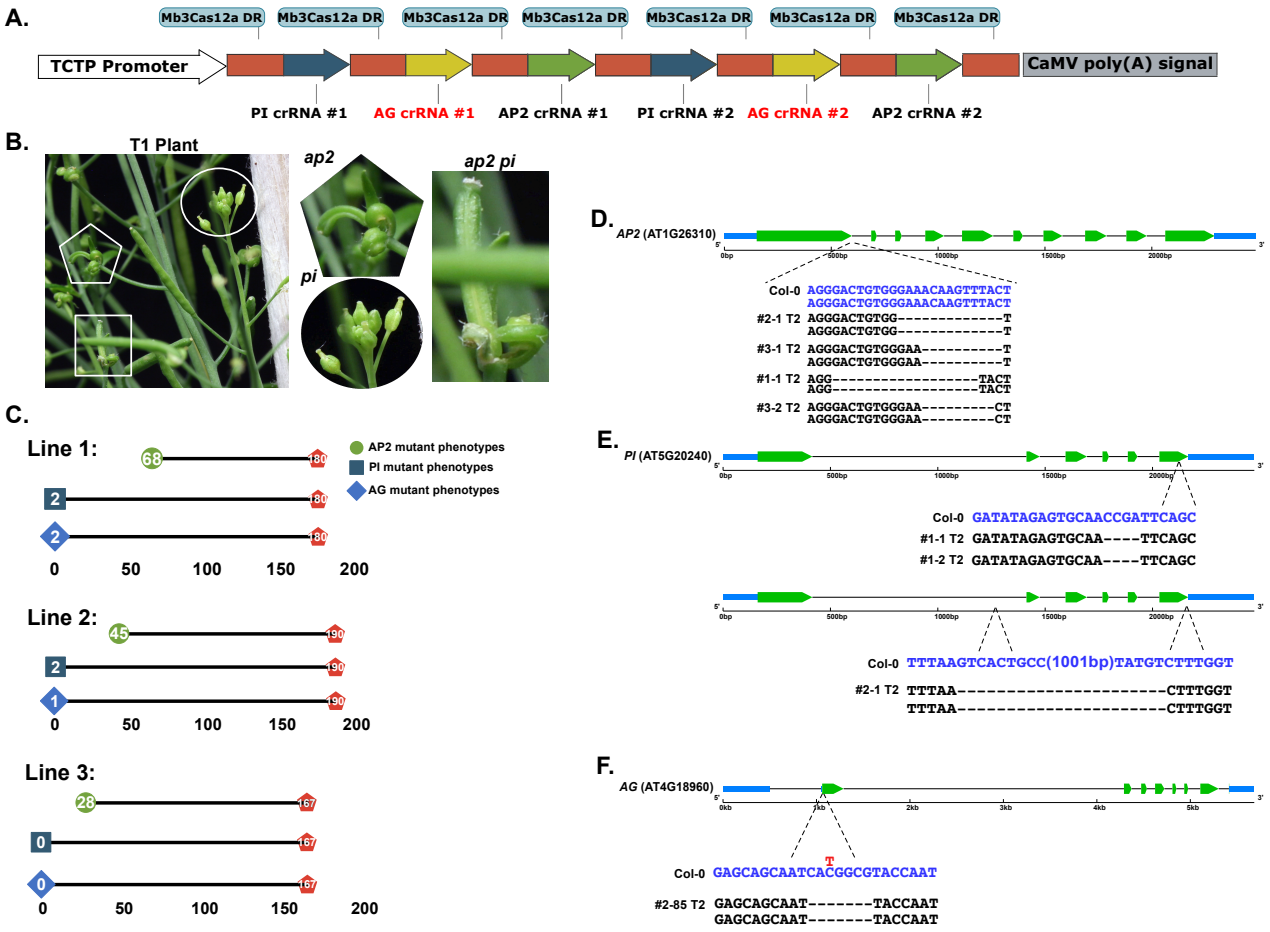


Fig 3.3 Assessment of multiplex gene-editing using Mb3Cas12a

A. Schematic of crRNA array targeting three floral regulator genes (**AP**ETALA**2**, **PI**STILLATA, **AG**AMOUS) expressed from the *AtTCTP* promoter. Guides targeting *AG* contain a SNP at positions 19 and 13, respectively.

B. Sectoring observed within T1 individuals phenocopy single and higher-order floral regulator mutations.

C. Assessment of mutant phenotypes observed in T2 pools from T1 lines exhibiting floral phenotypes. A lack of *PI* knockout phenotypes in the T2 likely represents low levels of heterozygosity from T1-edited plants.

D. E. F. Sanger sequencing of T2 lines exhibiting floral defects at *AP2*, *PI*, and *AG* loci.

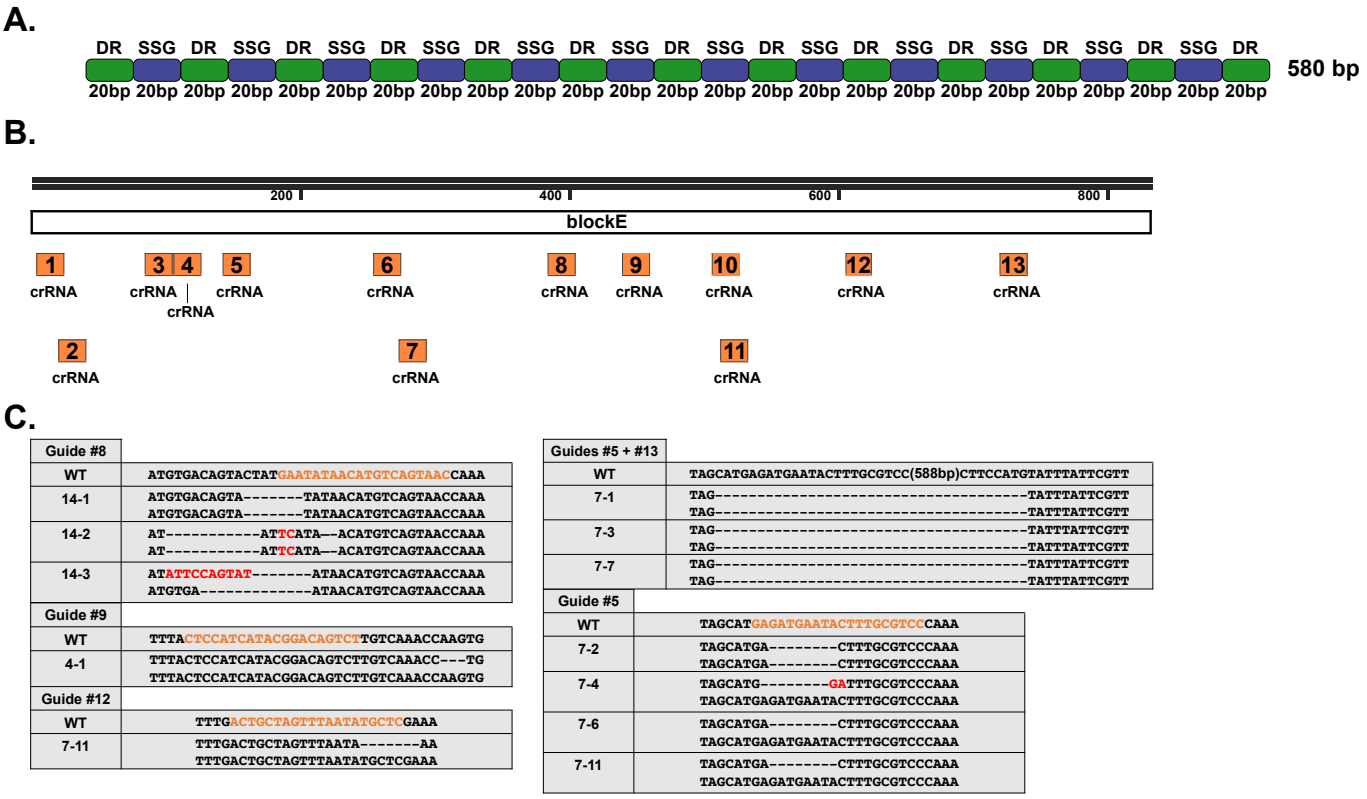


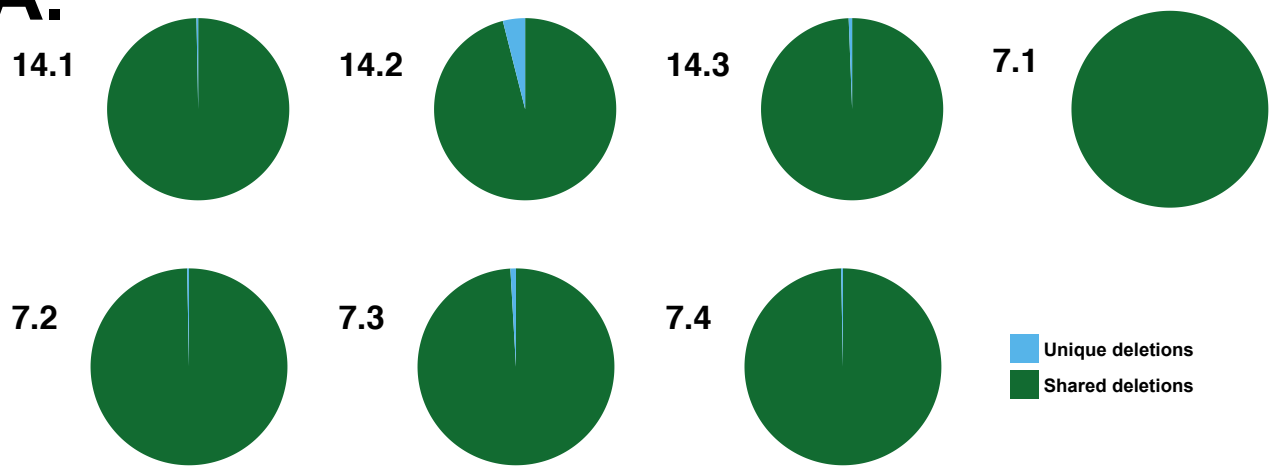
Fig 3.4. CRE mutagenesis using Mb3Cas12a

A. Schematic of a 13crRNA array targeting the *FT* regulatory element BlockE.

B. Location of high-quality crRNAs span the entire BlockE region of *FT* and allow for recovery of numerous edited alleles.

C. Edited alleles obtained from T_2 transgene-free individuals. Edits were observed at six target sites, resulting in indel patterns ranging from -3 to -588 bp.

A.



B.

Line	14.1	14.1	14.2	14.2	14.3	14.3	7.1	7.1	7.2	7.2	7.3	7.3	7.4	7.4
Deletions	Unique	Shared	Unique	Shared	Unique	Shared	Unique	Shared	Unique	Shared	Unique	Shared	Unique	Shared
Count	1	328	13	316	2	327	0	329	1	328	3	326	1	328

C.

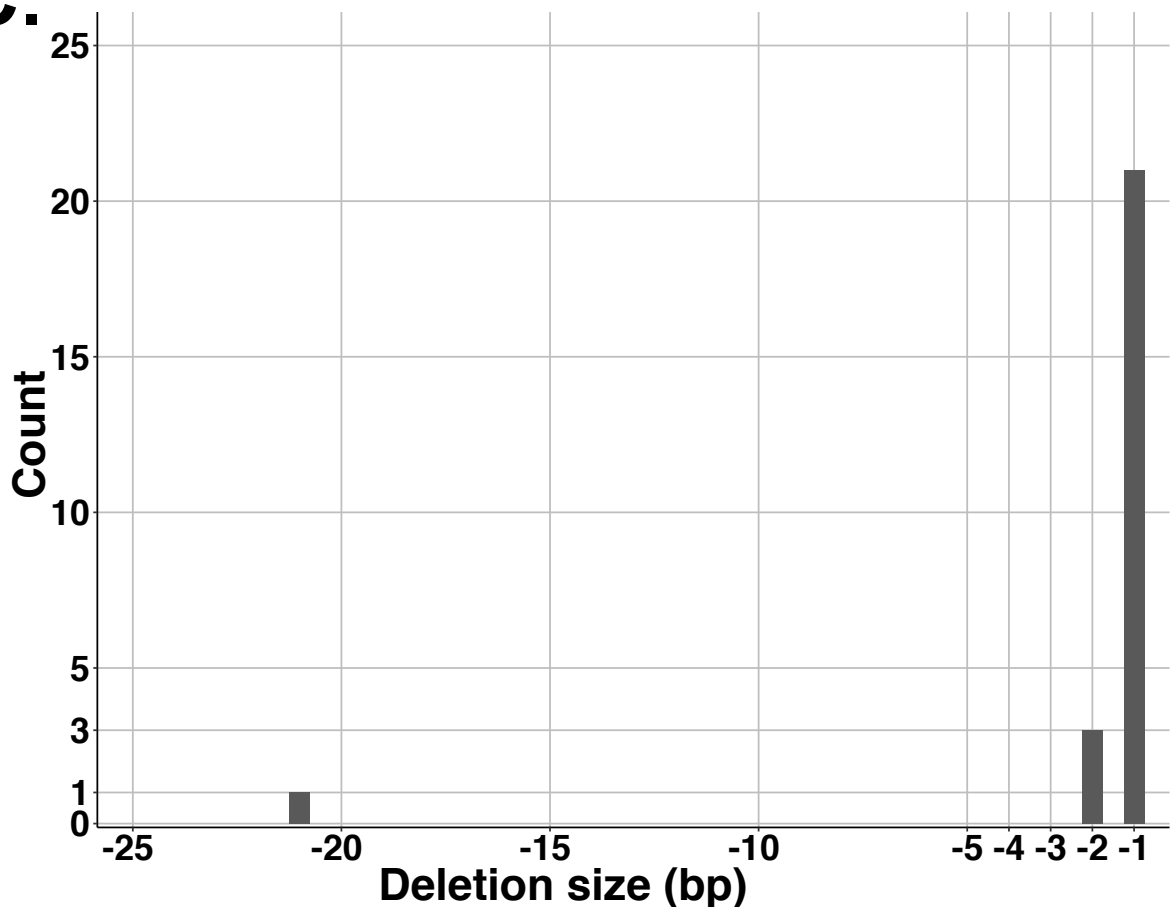


Fig 3.5. Detected variants in BlockE-targeted T_2 individuals

A. Number of unique and shared deletions in BlockE L7 and L14 T_2 individuals

B. Count of unique and shared deletions in BlockE L7 and L14 T_2 individuals

C. Size distribution of non-shared deletions in all BlockE individuals

References:

1. Li, J.-F. et al. Multiplex and homologous recombination-mediated genome editing in Arabidopsis and Nicotiana benthamiana using guide RNA and Cas9. *Nature biotechnology* **31**, 688-691 (2013).

2. Xing, H.-L. et al. A CRISPR/Cas9 toolkit for multiplex genome editing in plants. *BMC Plant Biology* **14**, 327 (2014).

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

4. Jacobs, T.B., LaFayette, P.R., Schmitz, R.J. & Parrott, W.A. Targeted genome modifications in soybean with CRISPR/Cas9. *BMC Biotechnology* **15**, 16 (2015).

5. Jiang, W. et al. Demonstration of CRISPR/Cas9/sgRNA-mediated targeted gene modification in Arabidopsis, tobacco, sorghum and rice. *Nucleic acids research* **41**, e188 (2013).

6. Jinek, M. et al. A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. *Science (New York, N.Y.)* **337**, 816-821 (2012).

7. Ui-Tei, K., Maruyama, S. & Nakano, Y. Enhancement of single guide RNA transcription for efficient CRISPR/Cas-based genomic engineering. *Genome* **60**, 537-545 (2017).

8. Čermák, T. et al. A Multipurpose Toolkit to Enable Advanced Genome Engineering in Plants. *The Plant Cell* **29**, 1196-1217 (2017).

9. Zetsche, B. et al. Cpf1 is a single RNA-guided endonuclease of a class 2 CRISPR-Cas system. *Cell* **163**, 759-771 (2015).

10. Zetsche, B. et al. Multiplex gene editing by CRISPR–Cpf1 using a single crRNA array. *Nature Biotechnology* **35**, 31 (2016).

11. Kleinstiver, B.P. et al. Genome-wide specificities of CRISPR-Cas Cpf1 nucleases in human cells. *Nature Biotechnology* **34**, 869 (2016).

12. Zhang, Q. et al. Potential high-frequency off-target mutagenesis induced by CRISPR/Cas9 in Arabidopsis and its prevention. *Plant molecular biology* **96**, 445-456 (2018).

13. Tang, X. et al. A large-scale whole-genome sequencing analysis reveals highly specific genome editing by both Cas9 and Cpf1 (Cas12a) nucleases in rice. *Genome Biology* **19**, 84 (2018).

14. Malzahn, A.A. et al. Application of CRISPR-Cas12a temperature sensitivity for improved genome editing in rice, maize, and Arabidopsis. *BMC Biology* **17**, 9 (2019).

15. Bernabé-Orts, J.M. et al. Assessment of Cas12a-mediated gene editing efficiency in plants. *Plant Biotechnology Journal* **17**, 1971-1984 (2019).

16. LeBlanc, C. et al. Increased efficiency of targeted mutagenesis by CRISPR/Cas9 in plants using heat stress. *The Plant Journal* **93**, 377-386 (2018).

17. Clough, S.J. & Bent, A.F. Floral dip: a simplified method for Agrobacterium -mediated transformation of Arabidopsis thaliana. *The Plant Journal* **16**, 735-743 (1998).

18. Harrison, S.J. et al. A rapid and robust method of identifying transformed Arabidopsis thaliana seedlings following floral dip transformation. *Plant Methods* **2**, 19 (2006).

19. LeBlanc, C. et al. Increased efficiency of targeted mutagenesis by CRISPR/Cas9 in plants using heat stress. *Plant J* **93**, 377-386 (2018).

20. Edwards, K., Johnstone, C. & Thompson, C. A simple and rapid method for the preparation of plant genomic DNA for PCR analysis. *Nucleic acids research* **19**, 1349-1349 (1991).

21. Dehairs, J., Talebi, A., Cherifi, Y. & Swinnen, J.V. CRISP-ID: decoding CRISPR mediated indels by Sanger sequencing. *Scientific Reports* **6**, 28973 (2016).

22. Klepikova, A.V., Logacheva, M.D., Dmitriev, S.E. & Penin, A.A. RNA-seq analysis of an apical meristem time series reveals a critical point in Arabidopsis thaliana flower initiation. *BMC Genomics* **16**, 466 (2015).

23. Han, Y.J., Kim, Y.M., Hwang, O.J. & Kim, J.I. Characterization of a small constitutive promoter from Arabidopsis translationally controlled tumor protein (AtTCTP) gene for plant transformation. *Plant Cell Rep* **34**, 265-275 (2015).

24. Kempin, S., Savidge, B. & Yanofsky, M. Molecular basis of the cauliflower phenotype in Arabidopsis. *Science (New York, N.Y.)* **267**, 522-525 (1995).

25. Urich, M.A., Nery, J.R., Lister, R., Schmitz, R.J. & Ecker, J.R. MethylC-seq library preparation for base-resolution whole-genome bisulfite sequencing. *Nature Protocols* **10**, 475-483 (2015).

26. Poplin, R. et al. A universal SNP and small-indel variant caller using deep neural networks. *Nature Biotechnology* **36**, 983-987 (2018).

27. Zhong, G., Wang, H., Li, Y., Tran, M. H. & Farzan, M. Cpf1 proteins excise CRISPR RNAs from mRNA transcripts in mammalian cells. *Nat. Chem. Biol.* **13**, 839–841 (2017).

28. Zhu, H., Liang C., CRISPR-DT: designing gRNAs for the CRISPR-Cpf1 system with improved target efficiency and specificity. *Bioinformatics* **35,** 2783–2789 (2019).

29. Zicola, J., Liu, L., Tänzler, P. et al. Targeted DNA methylation represses two enhancers of FLOWERING LOCUS T in Arabidopsis thaliana. *Nat. Plants* **5**, 300–307 (2019).

30. Ji, L., Jordan, W.T., Shi, X. et al. TET-mediated epimutagenesis of the Arabidopsis thaliana methylome. *Nat Commun* **9**, 895 (2018).

31. Rodríguez-Leal D, Lemmon ZH, Man J, et al. Engineering Quantitative Trait Variation for Crop Improvement by Genome Editing. *Cell* **171**, 470–480 (2017).

32. Strohkendl I, Saifuddin FA, Rybarski JR, et al. Kinetic Basis for DNA Target Specificity of CRISPR-Cas12a. *Mol Cell*. **71**, 816–824 (2018)

CHAPTER 4

DEVELOPMENT OF A MASSIVELY PARALLEL ASSAY FOR SILENCER DETECTION

Introduction:

The detection and identification of cis-regulatory elements (CREs) in large, eukaryotic genomes is essential for understanding biological processes such as transcription and gene regulation. Such features, termed either silencers or enhancers depending on their propensity to affect transcription, are typically associated with distinct chromatin features, such as increased accessibility, reduction in cytosine DNA methylation, and the presence of specific histone modifications in animals [1]. While the presence of chromatin features often correlate with regions possessing enhancer or silencer activity, chromatin features alone cannot be used to confirm enhancer or silencer activity [2]. Additionally, determination of the strength of transcriptional enhancement or silencing from candidate regions is not possible using chromatin profiling techniques [1].

Assays previously developed for determining silencer activity, such as RsSE, rely on the repression of an inhibitor, often constitutively expressed, to determine candidate element activity for transcriptional silencing [3]. Thus, the reduction of expression of a reporter is used to assess silencing activity, which presents challenges for the determination of activity compared to lack of detection power. Additionally, these and other silencer-detections assays rely on the ability to tightly control plasmid copy number in transformed cells, accomplished through the use of lentiviral vectors [4].

Traditionally, massively parallel reporter assays (MPRAs) such as STARR-seq have been utilized to determine the enhancer capacity of thousands of genomic regions simultaneously [5]. Through the creation and transfection of a plasmid library, the output of STARR-seq is a quantitative readout of the enhancer capacity of candidate regions. Regions which contain enhancer capacity, when placed downstream of a minimal promoter, serve to upregulate expression, resulting in increased RNA output compared to regions which lack enhancer capacity. By subsequently measuring the ratio of RNA/DNA from transfected cells, the enhancer capacity of all candidate regions can be assessed simultaneously.

However, as MPRAs used for regulatory element detection often require a positive readout for element identification, assays such as STARR-seq cannot be easily used to determine sequences that reduce expression of target genes (termed silencers), or prevent DNA:DNA interactions between enhancers and promoters (termed insulators). Additionally, the use of techniques such as RNA interference (RNAi) to eliminate regions which allow transcription result in trans-acting silencing of the plasmid backbone, if present in multiple copies in transfected cells. Thus, the creation of MPRA techniques specifically for silencer and/or insulator element detection are needed to characterize silencer and insulator elements in a genome-wide fashion.

Previous discovery of silencing elements in plants has been accomplished via individual cloning of candidate elements into plasmid vectors and has revealed the silencing ability of DNA fragments to recruit the Polycomb repressive complex (PRC2). These PRC2 silencing elements serve to directly recruit class I BPC and C1-2iD ZnF transcription factors (TFs), via their GA repeat and telobox motif sequences [6]. The binding of TFs to PRC2 silencing elements directly recruits PRC2 complex members, such as FIE, EMF2, and MSI1, which serve to deposit the

repressive histone mark H3K27me3, ultimately resulting in chromatin compaction and the silencing of gene expression [6].

Additional silencing elements serve to recruit other chromatin modifying elements, including histone chaperone proteins such as Histone Regulator A (HIRA) and histone deacetylases (HDACs). One well characterized example of target repression via HIRA and HDAC involves the silencing of KNOTTED1-like homeobox (*KNOX*) genes during the differentiation of stem cells to determinate lateral organs. The presence of specific MYB-binding sites in the promoters of *BREVIPEDICELLUS* and *KNAT2* serve to recruit the MYB-binding domain proteins *ASYMMETRIC LEAVES1* (*AS1)* and *AS2* [7]. The cooperative binding between *AS1* and *AS2* result in the recruitment of HIRA and HDACs, resulting in heterochromatin deposition at *BREVIPEDICELLUS* and *KNAT2* [7].

The discovery and characterization of RNA-targeting Cas proteins, such as Cas13d, has resulted in the availability of sequence-specific programmable RNases [8]. Similar in function to Cas12a enzymes, Cas13d enzymes are programmed to target a specific RNA sequence for degradation via expression of a sequence-specific gRNA. Possessing gRNA processing as well as target RNase activity, targeting numerous unique RNA targets using Cas13d is easily accomplished, in a manner identical to Cas12a multiplexing using a single mRNA molecule [9].

The development and adaptation of RNA-targeting Cas proteins represent an additional tool for potential identification of silencer and/or insulator elements; while maintaining high on-target activity of RNA cleavage of target sites, the use of Cas13d to degrade target RNA would not result in trans-acting effects on additional RNA molecules. Here, we present an all-in-one system for silencer detection, utilizing a Cas13d-guided knockdown approach, to enable a

positive readout of reporter expression directly correlated with silencer and/or insulator activities of candidate elements.

Results:

To explore whether Cas13d can be used for silencer identification, a multi-part vector system was created, to specifically detect the ability of candidate elements to repress expression of an open-reading frame (ORF) on the same DNA molecule. The system consists of two ORFs, which are both driven by constitutively expressing promoters. ORF #1 consists of codon optimized Cas13d, directly followed by a triplex stabilization sequence and a gRNA targeting a unique barcode sequence. To obtain unique, high-efficiency targets with minimal overlap with native soybean genomic transcripts, the T4 bacteriophage genome was binned into unique 2kb windows, which were run through the Cas13design v0.2 pipeline to identify high-quality candidate target sites. In total, 22,813 high-quality barcode sequences were identified for potential use.

As processing of the gRNA unit from mRNA will result in the loss of the 3' poly(A) tail, the addition of a triplex stabilization sequence protects poly(A)-lacking mRNA from degradation by cellular RNases, and additionally permits translation of mRNA molecules lacking a poly(A) tail.

ORF #2, orientated head-to-head when compared with ORF #1, consists of a reporter gene that can be optionally used to assess transformation efficiency, followed by a previously characterized genomic insulator sequence. To maximize the insulating effect observed using this system, the previously characterized strong insulator "1-kb Eco RI/Sal I fragment from bacteriophage lambda" (EXOB) was selected [10]. Immediately downstream of the EXOB

insulator is the location for library fragment cloning, followed by a unique barcode sequence targeted by Cas13d. Importantly, the unique barcode and barcoding gRNA must be identical on every plasmid molecule. Thus, the total length of sequence between the unique barcode and gRNA must be as small as possible to ensure the ability to synthesize both elements on a single oligonucleotide fragment.

As oligo synthesis technologies currently prohibit synthesis of oligonucleotides more than 300 bp in length, ORF #1 and ORF #2 must terminate using a single, bidirectional terminator to ensure synthesis of terminator elements, along matching gRNA targets and barcodes. As bi-directional terminators in plant-based systems have not been previously characterized, a custom bioinformatic pipeline was constructed to identify genomic regions with putative bi-directional terminator activity. Five putative bidirectional terminator regions <230 bp were identified from the *A. thaliana* genome and synthesized to test their ability to terminate ORF#1 and ORF#2 simultaneously. Surprisingly, all five regions exhibited robust termination activity for RFP and Cas13d, when measured by qRT-PCR (**Figure 4.10**).

Future experiments will include the testing of a candidate silencer region from the *Glycine max* sucrose binding protein locus (*GmSBP)* to ensure increased transcription of barcodes linked to putative silencer elements [11]. Concurrently, a candidate enhancer sequence from the *GmSBP* locus will be tested to ensure differentiation in RNA output levels between enhancer and silencer sequences.

Discussion:

The creation of a massively parallel reporter assay (MPRA) for the high-throughput identification of silencer and insulator elements will enable the discovery of silencer elements,

which often lack a distinct chromatin profile, genome-wide in plant systems. To circumvent the lack of lentivirus transfection, we chose to utilize the highly specific RNA-guided endonuclease Cas13d for the specific knockdown and degradation of RNA molecules containing DNA elements without silencing capacity. While enabling alternative means for DNA delivery, the use of a positive readout for silencing activity additionally removes the susceptibility for false negatives by allowing for the detection of weak and strong silencing elements.

To allow for simplified construction and cloning, we chose to discover bi-directional terminator elements to enable termination of head-to-head ORFs in the smallest possible sequence space, enabling the creation of oligo pools of matching crRNA sequences and unique barcode sequences. Five terminator elements, when characterized by qRT-PCR, greatly increased the levels of detectable RNA transcripts of ORFs #1 and #2. These results strongly support the functionality of the tested terminator elements and may serve to function in additional transient and stable transformation systems.

Methods:

To assemble the full-length Cas13d-silencing vector, two gBlocks (Integrated DNA Technologies) were synthesized, containing ORF #1 and ORF #2, respectively, and were assembled into the pHSG299 plasmid backbone (Clontech) using NEBuilder HiFi DNA Assembly (New England Biolabs). Sequence verification of this terminator-lacking construct was performed by restriction digest. To select putative bidirectional terminator regions for use in the silencer assay, we chose to search the *A.thaliana* genome due to the limited intergenic space found throughout chromosome arms. Previously published RNA-seq data was mapped to the *A. thaliana* transcriptome using STAR v 2.5.3, and transcript levels quantified using StringTie

2.1.1. Overlapping terminator sequences, were identified using BEDtools closest, and gene pairs with a median expression level <100 FPKM and a standard deviation of expression >2.5 were removed. Five elements were subsequently selected for manual testing, and the putative terminator regions were synthesized as gBlocks and cloned via HiFi DNA Assembly.

For protoplast isolation, the first two leaves of the 10-d seedling were cut into 5mm strips. The leaf strips were placed into digestive medium (:2 % (w/v) cellulase, 0,5% (w/v) pectolyase, 0,4 M mannitol, 10 mM $CaCl_2$, 1% (w/v) BSA, and 5 mM (N-morpholino)ethanesulfonic acid (MES) (pH 5.8) , and vacuum infiltrated for 30 mins and kept under incubation for 4 to 5 hours, at 30°C under low light with constant shaking (60 cycles/min). Protoplasts were recovered by centrifuging at 700 x g, washed with wash buffer (0,4 M mannitol, 2 mM $CaCl^2$, 0,1% (w/v) bovine serum albumin (BSA), and 5 mM N-2-hydroxyethylpiperazine-N'-2-ethanesulfonic acid (HEPES)-KOH (pH 7.0) twice followed by washing with electroporation buffer (25 mM HEPES-KOH (pH 7.2), 10 mM KCl, 15 mM $MgCl_2$ and 0.6 M mannitol). The protoplasts were kept on ice for 1 hour. Transient expression assays were performed by electroporation (2500 V, 250 µF) of 10 µg of the expression cassette DNA and 30 µg of sheared salmon sperm DNA into 2×10^5 - 5×10^6 protoplasts in a final volume of 800 µL. The protoplasts were kept on ice for 15 minutes and diluted into 8 ml of MS medium supplemented with 0.2 mg/ml 2, 4- dichlorophenoxyacetic acid and 0.6 M mannitol, pH 5.5. After 36 h of incubation in the dark, the protoplasts were washed and observed under the microscope and RNA extracted. RNA was isolated using a TRIzol-based lysis method, and cDNA transcribed using SuperScript II Reverse Transcriptase (Invitrogen).

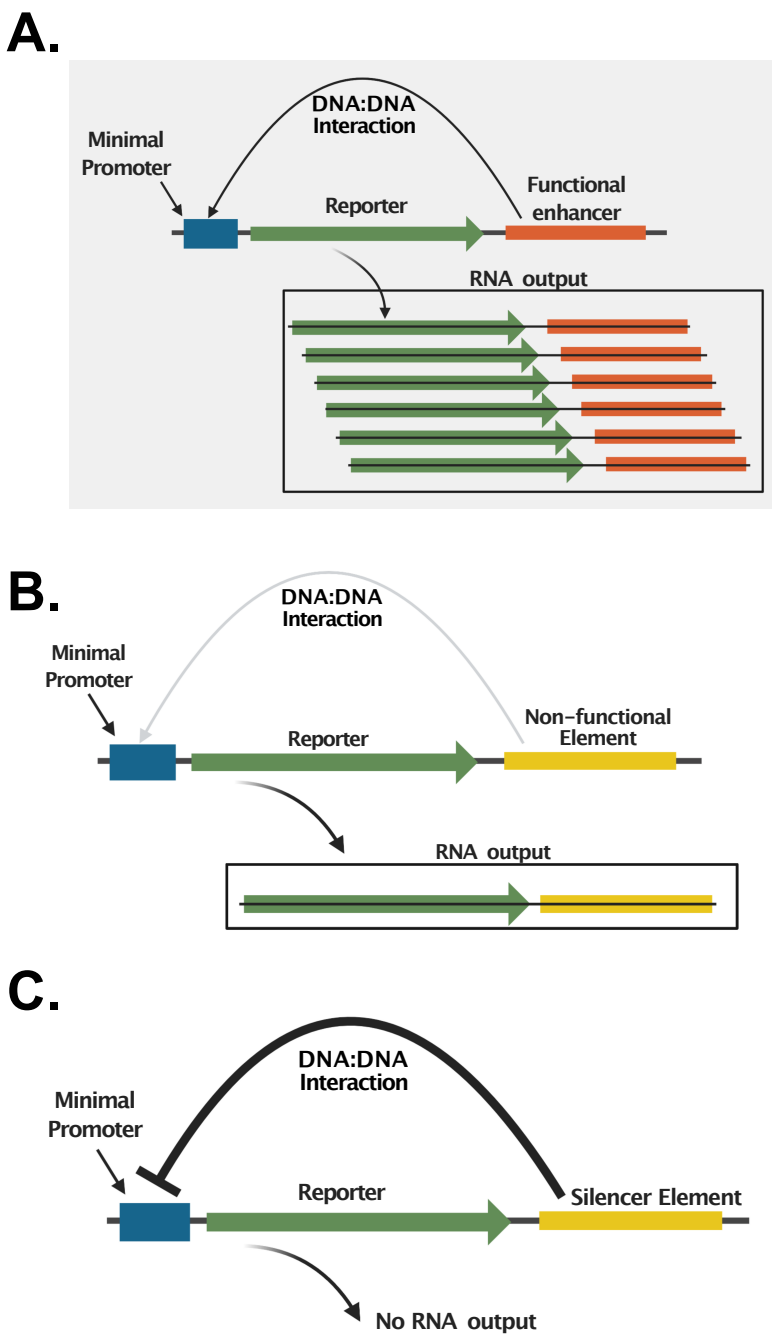


Figure 4.1. Previous applications of MPRAs such as STARR-seq can report enhancer activity **(A)**, but cannot distinguish non-functioning elements **(B)** from silencer elements **(C)**

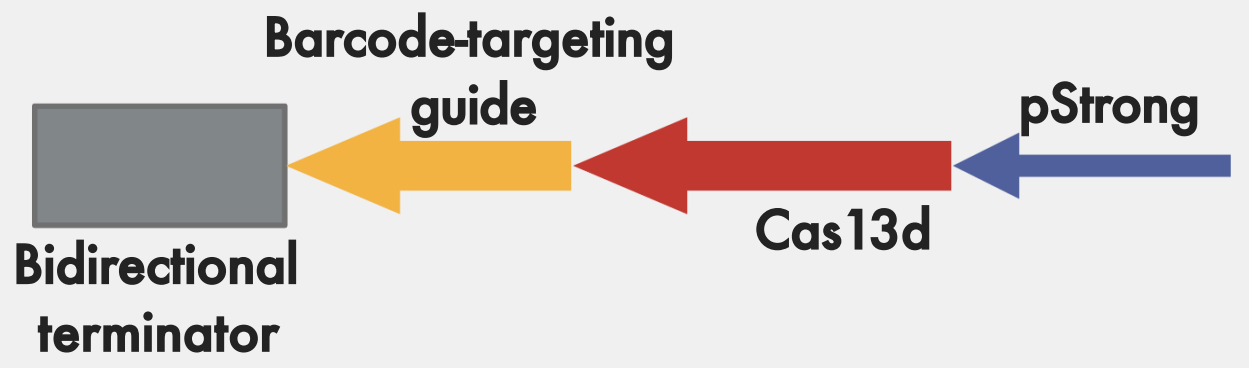


Figure 4.2. Composition of ORF #1. Driven by the Cassava mosaic virus promoter (pCsVMV), soybean codon-optimized RfxCas13d and the unique barcode-targeting gRNA are expressed as a single transcript. A Triplex sequence is located between Cas13d and the gRNA scaffold, to ensure translation of Cas13d after gRNA processing.

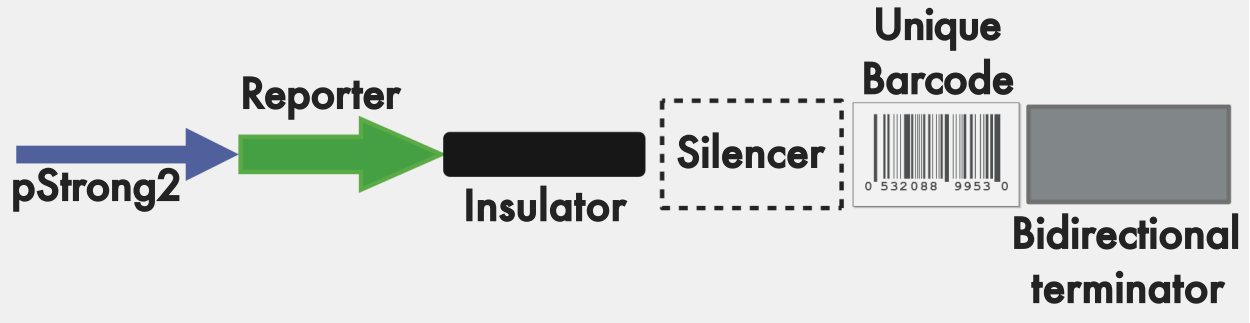


Figure 4.3. Composition of ORF #2. Driven by AtTCTP1 promoter (pTCTP), a florescent reporter followed by a the EXOB insulator, putative silencer element and unique barcode sequence are expressed. The EXOB insulator serves to prevent potential interactions between putative silencer elements and the TCTP1 promoter.

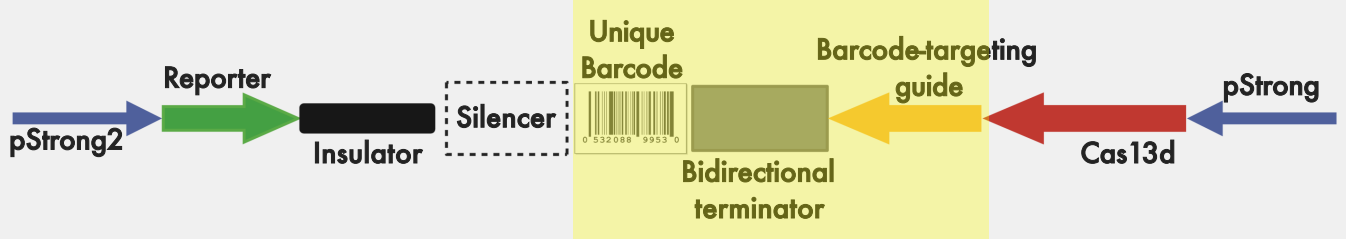


Figure 4.4. Overview of silencer assay system. Highlighted sections are synthesized in high-throughput oligo pools, to ensure matching of unique barcodes to barcode targeting guides on individual plasmid molecules.

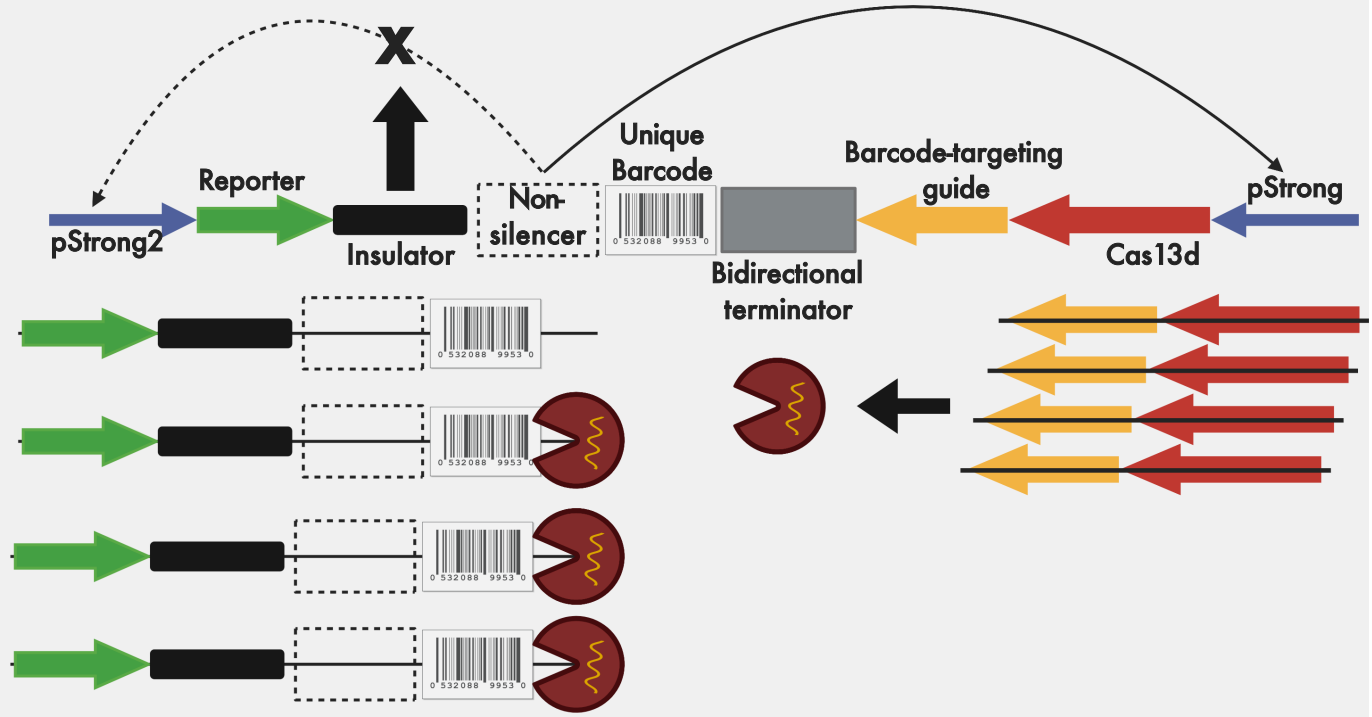


Figure 4.5. Predicted interactions of non-silencing elements. The candidate element does not reduce activity of pStrong. High levels of the unique barcode sequence along with Cas13d are transcribed, leading to degradation of RNA molecules containing the candidate element. The resulting RNA/DNA ratio for the candidate element will be low as a result.

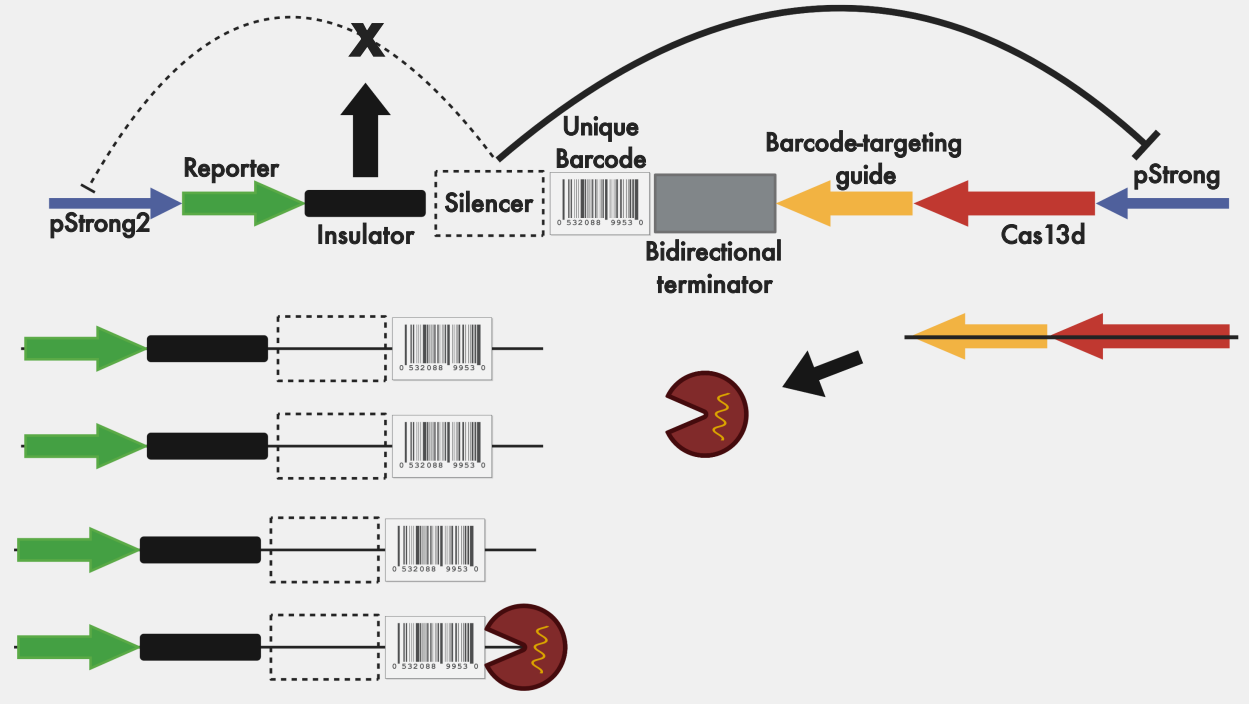


Figure 4.6. Predicted interactions of silencer elements. The candidate element reduces activity of pStrong. Lower levels of the unique barcode sequence along with Cas13d are transcribed, leading to decreased degradation of RNA molecules containing the candidate element. The resulting RNA/DNA ratio for the candidate element will be high as a result.

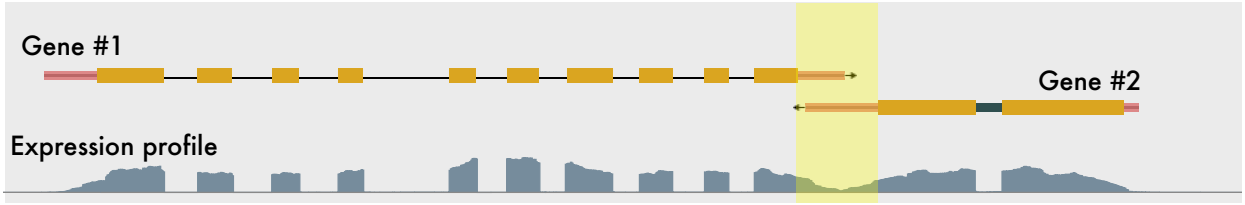


Figure 4.7. Example of validated bidirectional terminator element (highlighted yellow) and expression profile of expressed genes.

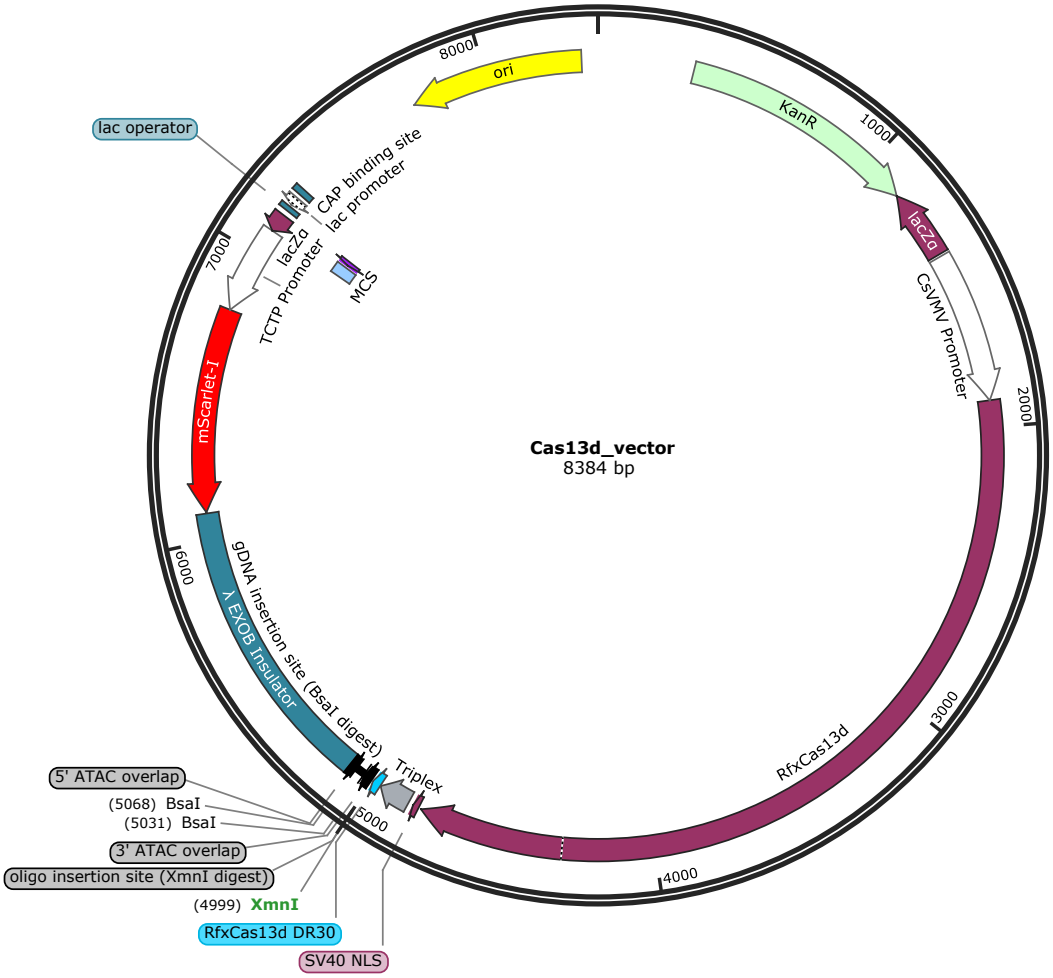


Figure 4.8. Vector map of fully assembled Cas13d-silencer vector backbone.

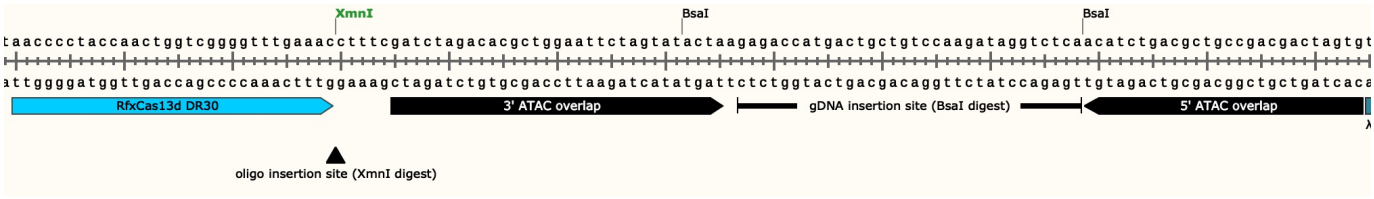


Figure 4.9. Sequence map of cloning sites for Cas13d-silencer vector. Bidirectional terminator elements, along with barcode and sequence-specific guide sequences, are cloned into the insertion sites (ATAC overlap, oligo insertion site).

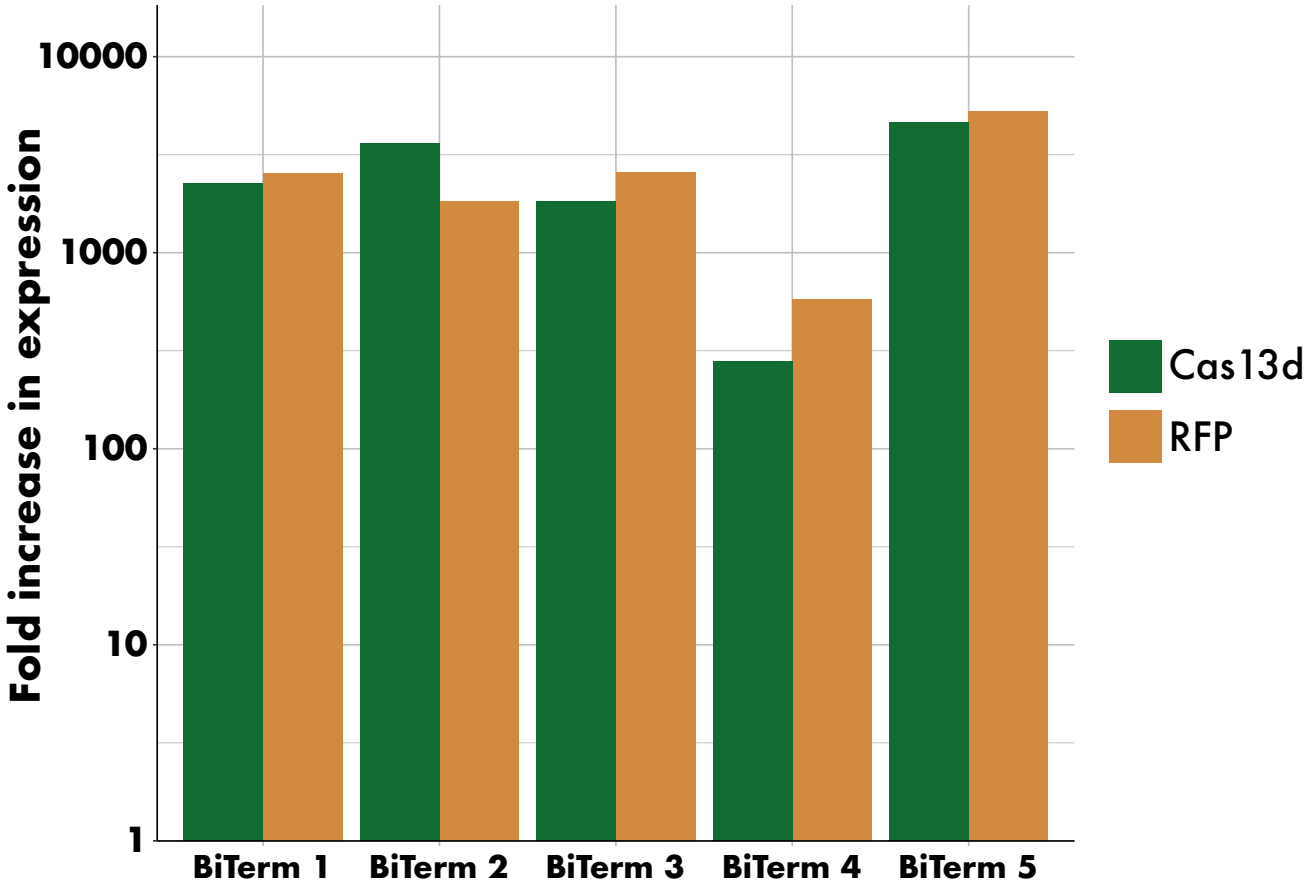


Figure 4.10. Putative bi-directional elements enable expression of ORFs. Fold change in expression level measured by RT-qPCR of Cas13d and RFP ORFs terminated by putative bi-directional terminators relative to no-terminator controls.

References:

1. Ricci, W. A. et al. Widespread long-range cis-regulatory elements in the maize genome. Nat. Plants 5, 1237–1249 (2019).

2. Muerdter, F., Boryń, Ł. M. & Arnold, C. D. STARR-seq — Principles and applications. Recent Adv. Funct. Assays Transcr. Enhanc. 106, 145–150 (2015).

3. Pang, B. & Snyder, M. P. Systematic identification of silencers in human cells. Nat. Genet. 52, 254–263 (2020).

4. Doni Jayavelu, N., Jajodia, A., Mishra, A. & Hawkins, R. D. Candidate silencer elements for the human and mouse genomes. Nat. Commun. 11, 1061 (2020).

5. Arnold, C. D. et al. Genome-Wide Quantitative Enhancer Activity Maps Identified by STARR-seq. Science 339, 1074 (2013).

6. Xiao, J. et al. Cis and trans determinants of epigenetic silencing by Polycomb repressive complex 2 in Arabidopsis. Nat. Genet. 49, 1546–1552 (2017).

7. Guo, M., Thomas, J., Collins, G. & Timmermans, M. C. P. Direct Repression of KNOX Loci by the ASYMMETRIC LEAVES1 Complex of Arabidopsis. Plant Cell 20, 48 (2008).

8. Konermann, S. et al. Transcriptome Engineering with RNA-Targeting Type VI-D CRISPR Effectors. Cell 173, 665-676.e14 (2018).

9. Yan, W. X. et al. Cas13d Is a Compact RNA-Targeting Type VI CRISPR Effector Positively Modulated by a WYL-Domain-Containing Accessory Protein. Mol. Cell 70, 327-339.e5 (2018).

10. Yang, Y., Singer, S. D. & Liu, Z. Evaluation and comparison of the insulation efficiency of three enhancer-blocking insulators in plants. Plant Cell Tissue Organ Cult. PCTOC 105, 405–414 (2011).

11. Waclawovsky, A. J., Freitas, R. L., Rocha, C. S., Contim, L. A. S. & Fontes, E. P. B. Combinatorial regulation modules on GmSBP2 promoter: A distal cis-regulatory domain confines the SBP2 promoter activity to the vascular tissue in vegetative organs. Biochim. Biophys. Acta BBA - Gene Struct. Expr. 1759, 89–98 (2006).

CHAPTER 5

CONCLUSIONS

Summary:

The understanding of and purposeful manipulation of gene expression has the potential to drastically impact plant phenotypes, and currently represents an underexplored avenue for crop improvement. The selection of genetic alterations of cis-regulatory elements, which serve to control the expression of target genes, has been responsible for drastic alterations of plant architecture in crops such as maize and tomato [1,2]. Recent advancements in transgenic technology, including the use of recombinant proteins for modulation of gene expression and genetic sequence, have the potential to drastically increase our ability to create and select causal variants that modulate gene expression by DNA sequence modification or epigenome editing. My dissertation attempted to create novel methods for both detection and editing of sequences involved in the modification of gene expression.

Chapter 2 describes a novel method for epimutagenesis, where we recombinantly expressed a mammalian enzyme (TET1) involved in DNA demethylation in *A. thaliana*. Upon TET1 expression, we observed an increase in time to flowering in multiple independent transgenic lines, which is indicative of demethylation of the *FWA* locus. Upon performing whole-genome bisulfite sequencing, we observed genome-wide demethylation across all five *A. thaliana* chromosomes, as well as in genic and transposable element genomic features. However, germline transmission of demethylated alleles was not initially observed with the original

transgenic construct. To increase levels of germline inheritance, we utilized a native promoter of *ACT2* from *A. thaliana*. Upon the use of the native *ACT2* promoter, we observed greatly increased rates of demethylation in transgenic populations, as well as instances of stable inheritance of alleles through the germline. This work represented the first known application of epimutagenesis in *A. thaliana* and showed that the overexpression of TET1 leads to heritable, genome-wide demethylation.

Chapter 3 focused on the development of a novel, Mb3Cas12a-based genome-editing system for multiplex editing of CREs. The use of previously utilized Cas nucleases, such as Cas9, for multiplex genome-editing presents issues relating to both DNA delivery and multiplex gRNA plasmid construction. I attempted to alleviate these limitations by harnessing the self-processing ability of Cas12a, to greatly decrease the size and complexity of transgenic DNA needed for multiplex genome-editing in *A. thaliana*. As Cas12a-mutagenesis is temperature sensitive, I utilized a modified heat-shock protocol previously applied in *A. thaliana* for genome-editing purposes [3]. To increase the efficiency of germline-editing in heat-shocked plants, I utilized a novel promoter for genome editing purposes, *pAtTCTP1*. The use of *pAtTCTP1* enabled increased germline inheritance of Mb3Cas12a-edited alleles when targeted to several genomic regions. Finally, I utilized this system to target the putative cis-regulatory element BlockE with 13 unique guide sequences. Edits ranging in size from 6-581 bp were obtained in edited plants, indicating the constructed system represents an efficient platform for the construction and editing of multiple genomic regions in *A. thaliana*. To profile for the potential of off-target editing when large number of guide sequences are introduced simultaneously into *A. thaliana*, whole-genome sequencing (WGS) was performed on seven BlockE-edited individuals.

Two putative off-target edits were detected from all seven profiled individuals, indicating high specificity of Mb3Cas12a genome editing in *A. thaliana*.

Chapter 4 describes the proof-of-concept creation of a massively parallel reporter assay (MPRA) specifically for the detection of silencer and insulator elements in plants. Previously developed MPRAs for silencer detection require the use of lentiviral transformation vectors, which are not available for plant-based systems [4,5]. To circumvent this limitation, I created a Cas13-based system, which utilizes the specific knockdown of unique barcode sequences to circumvent limitations in plant-based transformation. Interactions between candidate silencer elements are restricted to the promoter controlling the expression of Cas13d through the use of the EXOB insulator [6]. Thus, the absolute expression level of a silencer element is proportional to the ability of the element to decrease expression of Cas13d. By enabling a positive readout of unique silencer activity, the requirement for lentiviral delivery and/or tight regulation of plasmid copy number is eliminated.

To this end, I additionally identified five novel putative bidirectional terminator elements for use in this system using a novel bioinformatic pipeline, as well as identify 22,813 high-quality barcode sequences for use in this assay from the T4 bacteriophage genome. The future optimization and use of this system may serve to discovery thousands of previously uncharacterized silencer and insulator elements in plants genomes in an unbiased manner.

References:

1. Studer, A., Zhao, Q., Ross-Ibarra, J. & Doebley, J. Identification of a functional transposon insertion in the maize domestication gene tb1. *Nat. Genet.* **43**, 1160–1163 (2011).

2. Muños, S. *et al.* Increase in Tomato Locule Number Is Controlled by Two Single-Nucleotide Polymorphisms Located Near WUSCHEL. *Plant Physiol.* **156**, 2244 (2011).

3. LeBlanc, C. *et al.* Increased efficiency of targeted mutagenesis by CRISPR/Cas9 in plants using heat stress. *Plant J.* **93**, 377–386 (2018).

4. Doni Jayavelu, N., Jajodia, A., Mishra, A. & Hawkins, R. D. Candidate silencer elements for the human and mouse genomes. *Nat. Commun.* **11**, 1061 (2020).

5. Pang, B. & Snyder, M. P. Systematic identification of silencers in human cells. *Nat. Genet.* **52**, 254–263 (2020).

6. Yang, Y., Singer, S. D. & Liu, Z. Evaluation and comparison of the insulation efficiency of three enhancer-blocking insulators in plants. *Plant Cell Tissue Organ Cult. PCTOC* **105**, 405–414 (2011).