Improving plant genome editing: CRISPR meets epigenetics

A DISSERTATION

SUBMITTED TO THE FACULTY OF THE

UNIVERSITY OF MINNESOTA

BY

Trevor John Weiss

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE

DEGREE OF DOCTOR OF PHILOSOPHY

Dr. Feng Zhang

December 2022

Acknowledgements

I would like to thank my advisor, Feng Zhang, for his support, patience, and encouragement during my time in his lab. I would also like to thank the members of my advisory committee: Nathan Springer, Dan Voytas, and Bob Stupar for input on my dissertation research and support throughout the process. I received much support from members of the Voytas lab, Zhang lab, and Springer lab throughout my time at University of Minnesota. Additionally, I would like to personally thank Colby Starker for maintaining a productive lab environment. Lastly, I would like to thank my Mom, Dad, and brothers Andrew and Zach for always supporting me through both the exciting and rough times.

Introduction

Crop improvement through traditional plant breeding has paved the way for a healthier and more productive society. Despite improvements to crop yield and nutritional value, food security remains one of society's most urgent and challenging problems. Therefore, improvements to the current technologies, such as CRISPR-Cas9 genome engineering, are necessary to meet the global food demands. Although the CRISPR gene editing system has been implemented into many research or plant breeding programs, the ability to quickly engineer a specific trait of interest can still require substantial time, resources, and effort. These challenges can be attributed, in part, to the incomplete understanding of the rules that govern CRISPR-Cas9 genome editing efficacy. To realize the full potential of gene editing, new strategies to increase multiplexed editing efficiency and precision are needed.

In the following three chapters I use the model organisms *Setaria viridis* and *Arabidopsis thaliana* to address these challenges. In Chapter I, I describe the development of a robust multiplexed editing and transformation pipeline for the emerging monocot model, *Setaria viridis*. To achieve the efficient creation of genetically modified plants, a protoplast system was used to rapidly test and optimize gene editing reagents. These optimized CRISPR-Cas9 editing reagents, capable of improved multiplexed genome editing, were then transformed into *S.viridis* using the standard tissue culture method. Highly efficient transformation enabled the creation of transgene-free plants harboring frameshift mutations at the target genes.

In Chapter II, I leverage our improved *Setaria viridis* gene editing pipeline to create a panel of mutants with perturbed DNA methylomes. To accomplish this, we designedbuilt, and transformed T-DNA constructs to individually target four different DNA methylation families: METHYLTRANSFERASE1 (MET1), DECREASE IN DNA METHYLATION1 (DDM1), CHROMOMETHYLASE (CMT), and DOMAINS REARRANGED METHYLTRANSFERASE (DRM). Currently, the DRM1 mutant has been successfully isolated. Chapter II focuses on the phenotypic and molecular characterization of this mutant.

In Chapter III, I report a series of experiments to better understand the rules that govern CRISPR-Cas9 genome editing. Two features have been implicated in CRISPR-Cas9 editing efficacy: (1) the DNA sequence of the CRISPR-Cas9 target site, and (2) nonsequence features such as epigenetic context at the CRISPR-Cas9 target site. Previous studies have suggested that the DNA sequence being targeted for gene editing is the primary determinant of CRISPR-Cas9 efficiency and precision. As such, several tools have been developed to predict gene editing efficiency and precision solely based on the gene's DNA sequence. However, the reliability of these tools varies and appears to translate poorly to plants. This observation strongly suggests that nonsequence factors, such as epigenetic context, could influence CRISPR gene editing in plants. In fact, this possibility has recently been demonstrated; however, these studies were limited by their inability to separate the effects of the gene's DNA sequence from its epigenetic features. To address the challenge of separating these two variables, we developed the strategy of editing identical DNA sequences that are located in different epigenetic contexts. This

strategy enabled me to systematically dissect the influences of nonsequence features on CRISPR-Cas9 efficacy, revealing the significant influence of distinct epigenetic features.

Table of Contents

List of Tables

CHAPTER I

CHAPTER II

CHAPTER III

List of Figures

CHAPTER I

CHAPTER II

CHAPTER III

CHAPTER I: Context Statement

Summary

In recent years, *Setaria viridis* has been developed as a model plant to better understand the C4 photosynthetic pathway in major crops. With the increasing availability of genomic resources for S. viridis research, highly efficient genome editing technologies are needed to create genetic variation resources for functional genomics. Here, we developed a protoplast assay to rapidly optimize the multiplexed CRISPR/Cas9 system in S. viridis. Targeted mutagenesis efficiency was further improved by an average of 1.4-fold with the exonuclease, Trex2. Distinctive mutation profiles were found in the Cas9_Trex2 samples with 94% of deletions larger than 10bp, and less than 1% of mutations being insertions. Further analyses indicated that 52.2% of deletions induced by Cas9_Trex2, as opposed to 3.5% by Cas9 alone, were repaired through microhomology-mediated end joining (MMEJ) rather than the canonical NHEJ DNA repair pathway. Combined with the robust agrobacterium-mediated transformation method with more than 90% efficiency, the multiplex CRISPR/Cas9_Trex2 system was demonstrated to induce targeted mutations in two tightly linked genes, svDrm1a and svDrm1b, at the frequency ranging from 73% to 100% in T0 plants. These mutations were transmitted to at least 60% of the transgene-free T1 plants with 33% of them containing bi-allelic or homozygous mutations in both genes. This highly efficient multiplex CRISPR/Cas9_Trex2 system makes it possible to create a large mutant resource for S. viridis in a rapid and high throughput manner, and has the potential to be widely applicable in achieving more predictable MMEJ-mediated mutations in many plant species.

Chapter I has been adapted from my work in the following publication: “Optimization of multiplexed CRISPR/Cas9 system for highly efficient genome editing in *Setaria viridis*”

Trevor Weiss, Chunfang Wang, Xiaojun Kang, Hui Zhao, Maria Elena Gamo, Colby G. Starker, Peter A. Crisp, Peng Zhou, Nathan M. Springer, Daniel F. Voytas, Feng Zhang (2020). The Plant Journal. doi: 10.1111/tpj.14949.

During the course of this work many authors contributed. In particular, Chunfang Wang and Hui Zhao performed the tissue culture transformation; Trevor Weiss and Peng Zhou created a bioinformatics script to quantify MMEJ events from NGS data; Xiaojun Kang performed protoplast transfection; Feng Zhang, Nathan Springer, and Peter Crisp identified the genes to target for editing; Trevor Weiss, Maria Elena Gamo and Colby Starker built the plasmids; Trevor Weiss and Meredith Song performed genotyping; Trevor Weiss and Feng Zhang analyzed the data and generated figures; Trevor Weiss, Feng Zhang, Nathan Springer, and Dan Voytas wrote the manuscript. I have removed contact information and acknowledgements as well as formatted figures and references to be consistent throughout my thesis.

CHAPTER I: Optimization of multiplexed CRISPR/Cas9 system for highly efficient genome editing in *Setaria viridis*

Introduction

Setaria viridis (green foxtail) is an annual diploid C4 panicoid grass with a small genome and the wild relative to *Setaria italica* (foxtail millet), an agriculturally important crop in parts of Africa and Asia (Lata et al., 2013). Although historically regarded as an invasive weed, *S. viridis* has recently been developed as an emerging monocot model species to study bioenergy feedstocks and panicoid food crops, such as maize, sorghum, sugarcane and switchgrass, and to better dissect the cellular and biochemical mechanisms of C4 photosynthesis (Brutnell et al., 2010). *S. viridis* has many features that make it an attractive model system including a short life cycle, compact stature, reproduction via self-pollination and the ability to generate a high number of seeds (Brutnell et al., 2010; Defelice, 2002). Furthermore, the expanding genetic and genomic resources, including diverse germplasm accessions, chemically induced mutant populations, high quality reference genome of the A10.1 variety and the resequenced genomes from more than 600 wild accessions, make it possible to conduct large-scale gene discovery and functional genomics in *S. viridis* (Bennetzen et al., 2012; Huang et al., 2019; Zhu et al., 2017). Lastly, as another key factor for a successful model plant system, an efficient agrobacterium-mediated transformation method has been reported in S. viridis indicating it is amenable to genetic engineering techniques (Huang et al., 2019; Nguyen et al., 2020; Van Eck, 2018).

Genome editing has significant potential to expedite gene discovery and functional genomics. A key characteristic of current genome editing technologies is the use of programmable nucleases, such as Meganucleases, Zinc finger nucleases (ZFNs), Transcriptional activator like effector nucleases (TALENs) or CRISPR/Cas9, to create double-stranded DNA breaks (DSBs, or single-stranded nicks in some applications) at targeted loci. The induced DSBs can be exploited to introduce a variety of genomic modifications, such as deletions, insertions and nucleotide substitutions, by using one of two main DNA repair pathways, end joining or homology-directed repair (HDR). The end joining pathways, including non-homologous end joining (NHEJ) and microhomology-mediated end joining (MMEJ), are mostly used to generate insertions/deletions (Indels) at targeted sites, while the HDR pathway is employed to precisely incorporate desired sequences into targeted loci by copying genetic information from co-transformed donor templates (Chen et al., 2019).

In recent years, the CRISPR/Cas9 system has become the reagent of choice to achieve efficient genome editing in many plant and animal species due to its simplicity, robust activity, versatility, and multiplexing capability (Yin et al., 2017). Using this system, several gene knockout resources have been created in rice and maize (H. Liu et al., 2020; Meng et al., 2017). To generate gene knock-out resources in a plant species, high mutagenesis frequency and multiplexing capability, i.e. targeting multiple loci simultaneously, are key factors. The CRISPR/Cas9 system often requires considerable optimization in vector construction, transgene expression, tissue culture and transformation efficiency when adopted in a new plant species (Yin et al., 2017). Additional strategies can be employed to further improve mutagenesis efficiency. For

example, it has been demonstrated that the use of plants with a deficiency in the NHEJ pathway, such as the Ku70/Ku80 and Ligase IV mutants, could significantly enhance the frequency of targeted mutagenesis (Qi et al., 2013). Moreover, the simultaneous expression of exonucleases, such as Trex2, with CRISPR/Cas9 has been shown to enhance the frequency of targeted mutagenesis up to 2.5-fold in tomato and barley (Čermák et al., 2017). As for improving the multiplexing capability of the CRISPR/Cas9 system, one of the most effective strategies thus far is to achieve multiplex CRISPR guide RNA (gRNA) expression from a single polycistronic cassette. To this end, expression of multiple CRISPR gRNAs is driven by a single promoter with each gRNA separated by ribozyme sites, Csy4 recognition sites, or transfer RNA (tRNA) sequences, which can be processed to release individual mature gRNAs for targeting (Tsai et al., 2014; Xie et al., 2015). However, several studies have indicated that these multiplexing systems need to be tested and optimized when used in a new species. The polycistronic cassettes may possess varied processing efficacy in different species, and the Csy4 system may result in cytotoxicity (Minkenberg et al., 2017; Shiraki & Kawakami, 2018).

Although one example of the CRISPR/Cas9 mediated gene knockouts has been described in S. viridis, a highly efficient, multiplexed gene editing system has yet to be reported (Huang et al., 2019). In this study, we developed a protoplast-based transient assay for rapidly testing and optimizing the multiplexed CRISPR/Cas9 system in S. viridis. This system was also used to test the strategy of co-expression of the Trex2 exonuclease to further improve targeted mutagenesis efficiency in S. viridis. Finally, the optimized system was validated in stable transgenic plants to achieve highly efficient and heritable knockouts in two tightly linked S. viridis genes. The applications of this highly

efficient, multiplexed CRISPR/Cas9_Trex2 system were discussed in creating a large genetic mutant resource for S. viridis and achieving unique mutations in plant species.

Results and Discussion

Development of multiplexed gene editing using *S. viridis* protoplasts

We sought to develop a protoplast-based assay for quickly assessing the CRISPR/Cas9 system in *Setaria viridis* (Figure S1). Protoplasts were isolated from young leaves of 14-day old S. viridis seedlings. Transformation efficiency was tested using the green fluorescent protein (GFP) reporter driven by two different promoters, the Cestrum yellow leaf curling virus (CmYLCV) promoter and the Ubiquitin 2 promoter from switchgrass (PvUbi2) (Figure S2A). Both constructs produced robust GFP expression in about 80% of protoplasts after 24-hours post transformation (hpt) and at nearly 100% frequency after 48 hpt (Figure S2B).

The S. viridis protoplast assay system was then used to test and optimize CRISPR/Cas9 constructs targeting endogenous S. viridis genes, the domains rearranged methylase 1a (Drm1a), domains rearranged methylase 1b (Drm1b), male sterile 26 (Ms26) and male sterile 45 (Ms45) genes, respectively (Figure 1). The coding sequences of each gene were obtained by BLAST searching the reference genome of S. viridis accession A10.1 with the sequences of their maize orthologs, zmDrm1a, zmDrm1b, zmMs26 and zmMs45 (Table S1). Targeted sequences were additionally verified by Sanger sequencing in S. viridis accession ME034v, the plant variety used in this study. CRISPR gRNAs were designed to target the 5' exons or the conserved domains in each gene using CRISPOR (Haeussler et al., 2016). Each target site contains a restriction

enzyme site overlapping the CRISPR/Cas9 cut site to facilitate the Cleaved Amplified Polymorphic Sequences (CAPS) assay for subsequent genotyping analysis (Figure 1).

To achieve multiplexed gene editing in S. viridis, we tested two polycistronic gRNA expression systems, the Csy-type (CRISPR system yersinia) ribonuclease 4 (Csy4)-based and the tRNA array-based systems, in protoplasts (Čermák et al., 2017; Xie et al., 2015). Constructs containing gRNAs targeting the Drm1a and Drm1b genes (Figure S3A) were each co-transformed with Cas9 plasmids into protoplasts. As depicted in Figure 2A, high indel mutation frequencies were observed at each target site, ranging from 46% to 82%, indicating that both Csy4 and tRNA-based systems worked effectively in S. viridis protoplasts.

The Trex2 exonuclease enhances targeted mutagenesis with unique mutation profiles

We chose the tRNA-based system for multiplexed genome editing in S. viridis for further development due to its proven efficiency and simplicity (Minkenberg et al., 2017; Xie et al., 2015). The multiplexing gene editing constructs, pMG198 and pMG199 (Figure S3B), were made containing the Cas9 expression cassette and the gRNA array that simultaneously target two genes, Ms26 and Ms45. When these constructs were tested in protoplasts, high indel mutation frequencies were observed at each target site estimated by Next Generation Sequencing (NGS), i.e. 45% to 60% for the Ms26 gRNA1 and gRNA2 sites and 35% to 37% for the Ms45 gRNA1 and gRNA2 sites, respectively (Figure 2B). To test whether the mutagenesis efficiency can be further improved through coexpression of the Trex2 exonuclease, these multiplexing CRISPR/Cas9 constructs were

modified by cloning the Trex2 coding sequences into the Cas9 expression cassette. The resulting Cas9_Trex2 plasmids, pMG201 and pMG202 (Figure S3B), were then transformed into protoplasts respectively. At each target site, an average of 1.4-fold increase in mutagenesis frequency, ranging from 1.1 to 1.7 fold, was observed from the samples with Trex2 as compared to those without Trex2 (Figure 2B). Thus, our results demonstrated that coexpression of the Trex2 exonucleases with CRISPR/Cas9 further improved mutagenesis frequency in S. viridis.

Increased deletion size was observed in tomato and barley when Trex2 was employed (Čermák et al., 2017). However, a thorough characterization of the mutations induced by the combination of Cas9 with the Trex2 exonuclease has yet to be reported using a large data set. To this end, the mutation profiles were analyzed from a total of 516,815 NGS reads, and compared between the samples with and without co-expression of Trex2 (Table S2). In the samples without Trex2, both insertional and deletional mutations were observed in all 4 targeted sites with 1.6% to 42.1% insertions, the majority of which were 1bp insertions, and 57.1% to 97.8% deletions (Figure 2C). Among these deletions, 97.2% to 98.9%, were smaller than 10 bp. Conversely, in the samples with Trex2, essentially no insertions were detected at any target site. On average, 94% of deletions, ranging from 92.3% to 96.6%, were larger than 10 bp with 12% of them extending over 100 bp (Figure 2C). When they were further plotted along each targeted region, the deletions from the samples without Trex2 were found clustered in 5' of the PAM sequences (PAM-distal regions) and within 10bp of the DSB site. In contrast, the sequences from the 4 targeted sites with Trex2 contained much larger deletions that

were symmetrically distributed on each side of PAM and that extended up to more than 100 bp (Figure S4).

Interestingly, in the samples with Trex2, some specific deletions appeared frequently, exemplified by the 48bp deletions (3.5% of all deletions) in the Ms26 gRNA2 sample and the 87bp deletions (7.4% of all deletions) in Ms45 gRNA 2 (Figure S5). Examination of these particular deletions uncovered 2, 3 and 4 bp microhomologies at the Ms26 gRNA 2 junction sites (Figure S6) and 2, 4, 5 and 6 bp microhomologies at Ms45 gRNA 2 junction sites, indicating that the microhomology-mediated end joining pathway was involved in creating these deletions. Although previous studies have reported that Cas9-induced DSBs can be repaired through both NHEJ and MMEJ pathway, a recent study indicated that co-expression of Trex2 with CRISPR/Cas9 predominately results in DSB repair via the NHEJ pathway in human cell lines (Ata et al., 2018; Bae et al., 2014; Taheri-Ghahfarokhi et al., 2018), (Allen et al., 2018). To investigate the contribution of these two major end joining pathways in our samples, over 150,000 NGS reads from the samples with and without co-expression of Trex2 were analyzed based on the presence/absence of microhomology at the deletion junction sites. As a result, in the samples with Trex2, a significant fraction of deletions, with an average of 52.2% (ranging from 39.9% to 68.4%) appeared to be repaired by MMEJ, while the samples without Trex2 exhibited an average of only 3.5% of deletions (ranging from 0.11% to 8.1%) repaired through MMEJ (Figure 2D). These data suggested that different organisms may invoke different end joining pathways to repair the Cas9_Trex2 induced DSBs. Further investigation is needed to better understand the mechanisms underlying these observations and the factors influencing the MMEJ efficiency. Nevertheless, our results

indicated that co-expression of the Trex2 exonuclease with CRISPR/Cas9 can be used as a general strategy to increase the efficiency of targeted deletions, and to create a large collection of mutation variants in plants. Moreover, at least in S. viridis, the high frequency of the MMEJ events may also increase the predictability of the mutation outcomes, which is of particular value when precise deletional mutations are needed for dissecting gene function or characterizing cis-regulatory elements (Allen et al., 2018; Rodríguez-Leal et al., 2017).

The previous study by Chari et al (Chari et al., 2015) suggested that Trex2 could increase off-targeting mutation frequency in human cells and that careful gRNA design was crucial to reduce this risk. When the gRNAs were designed in this study, the candidate gRNAs were selected using the program, CRIPSOR, with the lowest off-targeting scores possible (Haeussler et al., 2016). Furthermore, four top-ranked Ms26 gRNA1 potential off-target sites, two sites with 3 bp mismatches and two sites with 4 bp mismatches, were identified and evaluated for the potential off-targeting (Table S3). Each site was subjected to PCR amplification and NGS from the samples with and without Trex2, respectively. As a result, no significant mutation was detected at all 4 sites from any sample (Table S3). While we cannot rule out the possibility that Trex2 could increase off-targeting in plant cells, this finding confirmed that careful gRNA design is important to reduce the risk.

While the mechanisms of how Trex2 improves Cas9-induced mutation frequency and promote MMEJ repair in plants remain to be elucidated, this process likely involves at least two major steps. First, upon the DSB induction at a target site, the Trex2 protein likely displaces the CRISPR/Cas9 complex, allowing for resectioning of the broken DNA

ends in a 3'-to-5' manner (Mazur & Perrino, 1999). Next, the resected ends are rejoined through either NHEJ or MMEJ repair pathway (Figure S7). As a result, the Trex2-mediated resectioning makes this DNA repair process more error prone and gives rise to higher mutation rates with larger deletions. It is not clear why Trex2 promotes MMEJ over NHEJ repair in S. viridis but not in human cells. These findings suggested that different organisms may invoke different end-joining pathways to repair the resected DSBs. Further investigation is needed to understand the underlying mechanisms.

Highly efficient multiplexed genome editing in T0 transgenic plants

The multiplex CRISPR/Cas9_Trex2 system tested through the protoplast assay was then used to create heritable mutations in the two linked Drm1 genes. Three T-DNA constructs were made by assembling the tRNA-gRNA array cassette with the Cas9_Trex2 cassette through the Golden Gate assembly method (Čermák et al., 2017). In these T-DNA vectors, the tRNA-gRNA array contained up to 3 gRNAs targeting the Drm1a and Drm1b genes individually or collectively (Figure S7A). Stable transgenesis was carried out using agrobacterium-mediated transformation. A total of 85, 26, and 103 potential transgenic plants were regenerated from 86, 29 and 112 mature seed-derived calli in the transformation groups, pTW037, pTW044 and pTW045, respectively, exhibiting the high transformation efficiency (above 90%) for this S. viridis ME034v variety (Table 1). A subset of candidate plants was randomly picked from each group and further genotyped by genomic PCR using the primers for the hygromycin marker gene (Figure S7B). The presence of T-DNA constructs was confirmed in all 30 plants, indicating a very low escape rate in plant transformation. Notably, the overall

transformation efficiency reported in this study has been significantly higher than those from previous studies, i.e. 5-15% for the variety A10.1 (Nguyen et al., 2020; Van Eck, 2018). The difference between the observed transformation efficiencies could be attributed to the S. viridis variety, ME034v, used in this study. Similarly, high transformation efficiency for ME034v was also observed from other groups (Joyce Van Eck, personal communication). Further investigation will be required to understand the underlying mechanism(s).

Next, the transgenic plants were examined at the targeted regions using genomic PCR followed by restriction enzyme digestion, i.e. the CAPS assay (Figure 3A; Figure S8). As summarized in Table 1, the frequency of plants with mutations induced by the single gRNA T-DNA construct, pTW037, was 100% in the Drm1a target site. Similarly, 82% and 73% of plants with the double gRNA T-DNA plasmid, pTW044, carried mutations in the Drm1a and Drm1b genes; and 100% of plants with the triple gRNA T-DNA construct, pTW045, had mutations in all 3 target sites (Figure 3A and Table 1). Together, the multiplex CRISPR/Cas9_Trex2 system that was demonstrated to be highly efficient in the protoplast assay also induced high frequency mutagenesis in stable transgenic plants.

Inheritance of targeted mutations in T1 progenies

Plants with the triple gRNA T-DNA construct were further investigated to test heritability of the mutations induced in T0 plants. The CAPS genotyping assay indicated that all eleven T0 plants contained mutations in all three targeted sites at variable frequencies (Figure 3A). To quantify the mutagenesis frequency in each T0 plant, PCR

amplicons spanning each targeted region were sequenced using Next Generation Sequencing. Over 20,000 sequencing reads were generated from each PCR amplicon, and analyzed to estimate the indel mutation frequency (Figure 3B). Consistent with results from the CAPS genotyping assay, mutagenesis frequencies were observed in each T0 plant ranging from 3% to 99% in the Drm1a site, 10% to 99% in the Drm1b gRNA1 site and 11% to 99% in the Drm1b gRNA2 site, respectively. In general, the mutagenesis frequencies were positively correlated across all three target sites. For example, four out of eleven plants showing lower mutagenesis efficiency in the Drm1a target site (under 10%) also displayed lower mutagenesis efficiency in the two Drm1b target sites (Figure 3B).

Five T0 plants, 12, 15, 84, 86, and 94, showing mutation frequencies greater than 50% at all three target sites, were chosen to be self-pollinated and grown to maturity. Ten T1 progenies from each T0 plant were grown for further characterization. Using the CAPS genotyping assay, high frequencies of mutant plants were detected in these T1 populations, ranging from 50-100%, 60-90%, and 30-100% at three target sites, Drm1a, Drm1b gRNA1 and Drm1b gRNA2, respectively (Figure S9, Table S3). To further distinguish heritable mutations from somatic mutations in these T1 plants, genomic PCR was conducted to detect T-DNA transgene free plants using primers for the hygromycin marker gene. Four out of five T1 populations, 12, 15, 84 and 86, exhibited segregation of the transgene (Table 2, Figure S10). As the hygromycin marker gene is close to the left border (LB) region of the T-DNA, to rule out the possible presence of partial T-DNA sequences (Collier et al., 2018), the putative transgene free T1 plants were also subjected to a luciferase assay for detecting the expression of the luciferase reporter gene that close

to the right border (RB) region. Lack of fluorescent signal further confirmed the absence of T-DNA transgenes in these plants (Figure S12). Out of 50 T1 plants, 10 transgene-free plants were identified (Table 3). Among them, 6 plants (60%) showed mutations at at least one of the three target sites, with 1 plant (10%) having a mutation at two target sites and 2 plants (20%) having mutations at all three target sites (Table 3). These two transgene free plants, 12-1 and 12-9, with mutations in all three target sites were further characterized by NGS (Table 3, Figure 3C). Notably, while the CAPS assay suggested that heterozygous mutations occurred at the DRM1b gRNA2 site in these plants (Figure S9), the NGS data revealed bi-allelic mutations in both plants. Close examination of these mutations identified a 1bp deletion at this target site that did not disrupt the restriction enzyme (PflMI) recognition site used in the CAPS assay. This finding suggested that the CAPS assay used to screen the T1 plants may underestimate the mutagenesis frequency transmitted to the T1 populations. Additionally, as seen from the protoplast assay, MMEJ-mediated deletions were recovered in each plant (Figure 3C). Taken together, these results clearly demonstrated that the mutations are transmissible in the *S. viridis* mutant plants at high efficiency. This made it possible to recover homozygous mutants from a relatively small population of T1 plants and to generate multiple gene knock-out events simultaneously and rapidly, which is particularly useful in editing tightly linked genetic loci as shown in this study.

In summary, we developed a protoplast-based assay to rapidly test and optimize the multiplex CRISPR/Cas9 gene editing system in *S. viridis*. The mutagenesis frequency can be enhanced by co-expression with the Trex2 exonuclease resulting in a unique mutation profile with larger deletions, no insertions and a high frequency of MMEJ

repaired events. Further, the optimized multiplex CRISPR/Cas9_Trex2 system can induce targeted mutagenesis in stable transgenic plants at remarkably high efficiency. This system allowed us to generate heritable knock-outs in two tightly-linked *S. viridis* genes from a small number of transgenic plants (10 T0 plants) in a timeframe as short as three months (starting from plant transformation to T1 seedlings). With the efficient agrobacterium-mediated transformation method, this highly efficient pipeline makes it possible to create a large mutant collection of *S. viridis* in a rapid and high throughput manner. Moving forward, it would be interesting to test this CRISPR/Cas9_Trex2 system, or combinations of Trex2 with different Cas proteins, in other plants. These new systems have the potential to be widely applicable in achieving more predictable MMEJ-mediated mutations in many plant species.

Experimental procedures

Plant materials, seed germination and plant growth conditions

S. viridis variety ME034v was used in this study. To break dormancy and promote seedling germination, freshly harvested seeds were incubated at 29∘C for 24 hours in a 1.4 mM gibberellic acid and 30 mM potassium nitrate solution (Sebastian et al., 2014). After the 24-hour incubation, seeds were sterilized with 50% bleach for 10 minutes, followed by 5 water rinses and planted on germination media (0.5X MS, 0.5% sucrose, 0.4% PhytaGel, pH 5.7). Seedlings were transplanted to soil six days after germination and grown at 26oC/22oC (day/night) with a photoperiod of 16h/8h (day/night), under 30% relative humidity, a modified protocol from (Huang et al., 2019).

Guide RNA design and vector construction

The genomic sequences of each targeted gene were obtained by BLAST searching the S. viridis A10.1 reference genome from the phytozome database (https://phytozome.jgi.doe.gov). CRISPR gRNAs were designed to target exons in the 5' region of the gene or the conserved domains in each gene using CRISPOR (Haeussler et al., 2016). The targeted sequences were further verified by Sanger sequencing in the S. viridis variety ME034v. The conserved domains were identified by comparing the coding sequences from S. viridis with their orthologs from brachypodium, maize, and Arabidopsis.

The gRNA constructs were made by following the Golden Gate assembly method (Čermák et al., 2017). The backbone for the tRNA-based gRNA construct was pMOD_B2103, and the backbone of the Csy4-based gRNA construct was pMOD_B2303. The Cas9 constructs were pMOD_A1110 and pMOD_A1510 . The Cas9_Trex2 construct, pMOD_A1910, were made by cloning the Trex2 coding sequence into the codon-optimized Cas9 expression cassette, based on the codon usage from wheat (Triticum aestivum). The GFP reporter constructs, pMOD_C3003 and pMOD_C3013, were made by cloning the GFP coding sequences under the control of the CmYLCV and PvUbi promoters with the 35S terminator. All the constructs will be deposited to Addgene.

Protoplast isolation and transformation

Protoplast isolation and transformation were performed using a modified version of the PEG-mediated method (J. Li et al., 2016). In brief, leaves from 14-day young

seedlings were sliced into small pieces with a razor blade and digested with the enzyme solution (1.5% Cellulase, 0.75% Macerozyme, Kanematsu USA Inc.) for 4-5 hours on a shaker at 40 rpm. The digested tissues were filtered through a 70uM nylon filter (Fisher Scientific LLC) into W5 buffer (2mM MES with pH5.7, 154mM NaCl, 125mM CaCl2, 5mM KCl). Protoplasts were collected and resuspended in W5 buffer with a gentle centrifuge at 100xg for 5 minutes. The number of protoplasts was estimated using a hemocytometer. Roughly 200,000 protoplasts were mixed with DNA plasmids (15ug per construct) in 20% PEG buffer and incubated at room temperature in the dark for 48 hours. Transformation efficiencies were monitored by transforming protoplasts with a plasmid encoding GFP.

T-DNA transformation and tissue culture

Agrobacterium tumefaciens-mediated transformation of S. viridis was performed as previously described with a few modifications (Van Eck et al., 2017). Callus initiation was first performed by removing the seed coats and sterilizing seeds with a 10% bleach plus 0.1% tween solution for 5-10 minutes under gentle agitation. Seeds were plated on callus induction media with the embryos facing upward. The plates were placed at 24oC in the light for a week and then moved to dark for callus initiation. Embryogenic calli were collected after 4-7 weeks and inoculated with the AGL1 strain harboring the T-DNA construct. Inoculated calli were placed on co-culture medium and incubated in the dark at 20 oC for 5-7 days. Transformed calli were transferred to the selection medium with 50mg hygromycin for 4 weeks at 24oC, then the selected calli were subcultured on plant regeneration media with 20mg hygromycin under 16-hour light to allow the growth of the

transformed shoots. Elongated shoots were transferred to the rooting medium with 20mg hygromycin. Shoots with well-developed roots were transplanted to soil and grown to maturity.

Genotyping and mutant identification

Mutant identification and characterization were performed using two methods, genomic PCR with restriction enzyme digestion (CAPS assay) and Illumina paired-end read amplicon sequencing (NGS assay). PCR was performed with GoTaq Green Master Mix (Promega Inc.) according to the manufacturer's instructions with an annealing temperature of 58°C and an extension time of one minute. Amplicons were then subjected to restriction enzyme digestion using an enzyme that overlaps with the CRISPR/Cas9 cleavage site. PCR amplicons made with the corresponding primers were subjected to Illumina paired-end read amplicon sequencing by Genewiz Inc. The raw NGS reads were analyzed using CRISPResso2 (Clement et al., 2019). All the primers used in this study were listed in Table S4.

Characterization of mutation profiles

Mutations in the NGS reads were characterized into three categories, deletions, insertions and others (including substitutions and substitutions with insertions or deletions). The mutagenesis efficiency of each mutation type was estimated by dividing the total number of modified reads by the total number of reads. To minimize the problem caused by sequencing errors from NGS, mutation reads that only occurred once in the NGS data were not included in the calculation. To quantify the mutations derived from NHEJ or MMEJ repair pathways, each distinct deletion read was categorized into three

separate sequences: 1. the left flanking sequence, 2. the deleted sequence, and 3. the right flanking sequence. Mutation reads were considered as MMEJ products when greater than 2 bp of homology were identified at the junction site between left and right flanking sequences. Mutation reads without microhomology sequences at the junction sites were classified as NHEJ events.

Figures

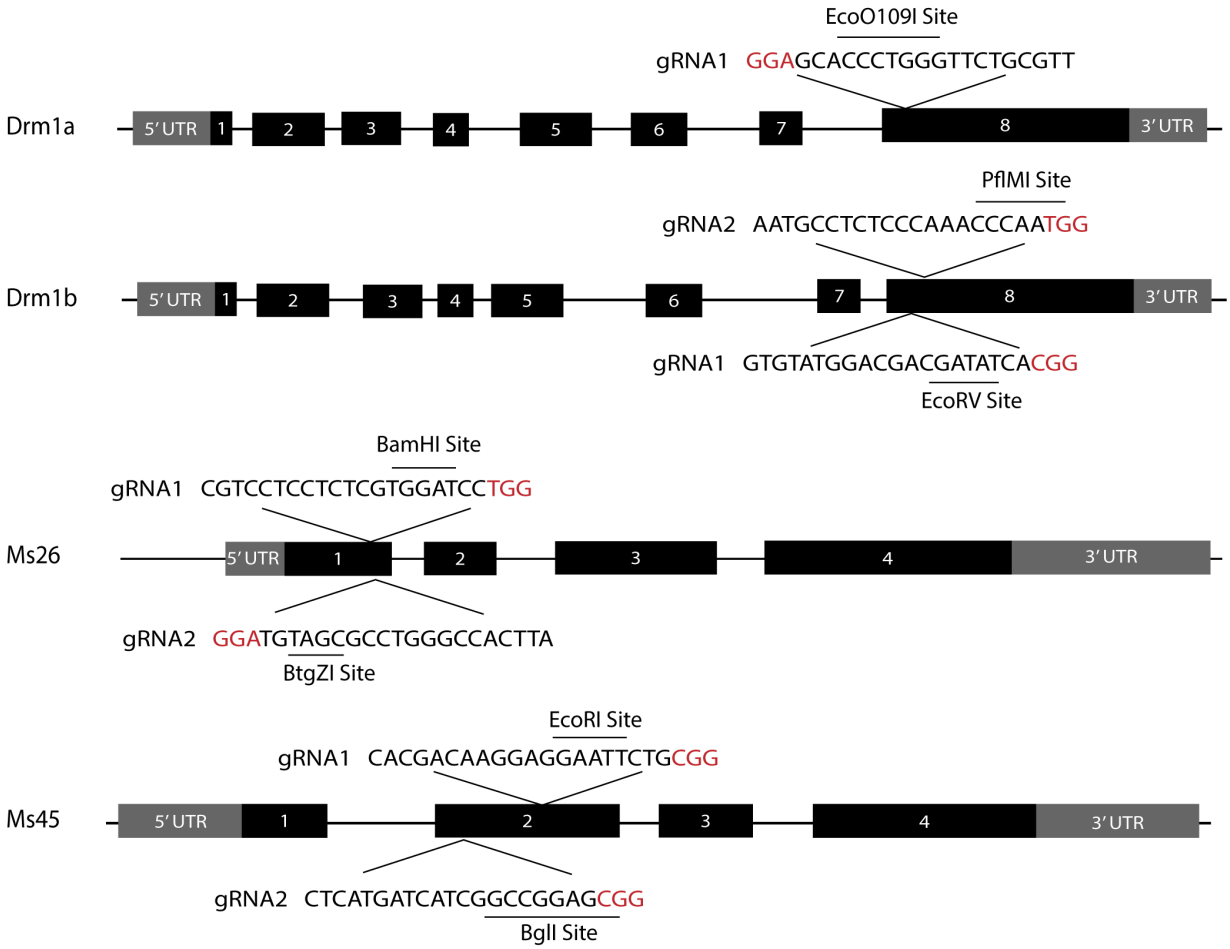


Figure 1. The schematic structures of the Drm1a, Drm1b, Ms26, and Ms45 genes. Each black box represented an exon, with gray boxes representing the 5' and 3' untranslated regions. Individual gRNA targeted sites were shown in each gene with the restriction enzyme sites underlined, and the Protospacer Adjacent Motif (PAM) in red.

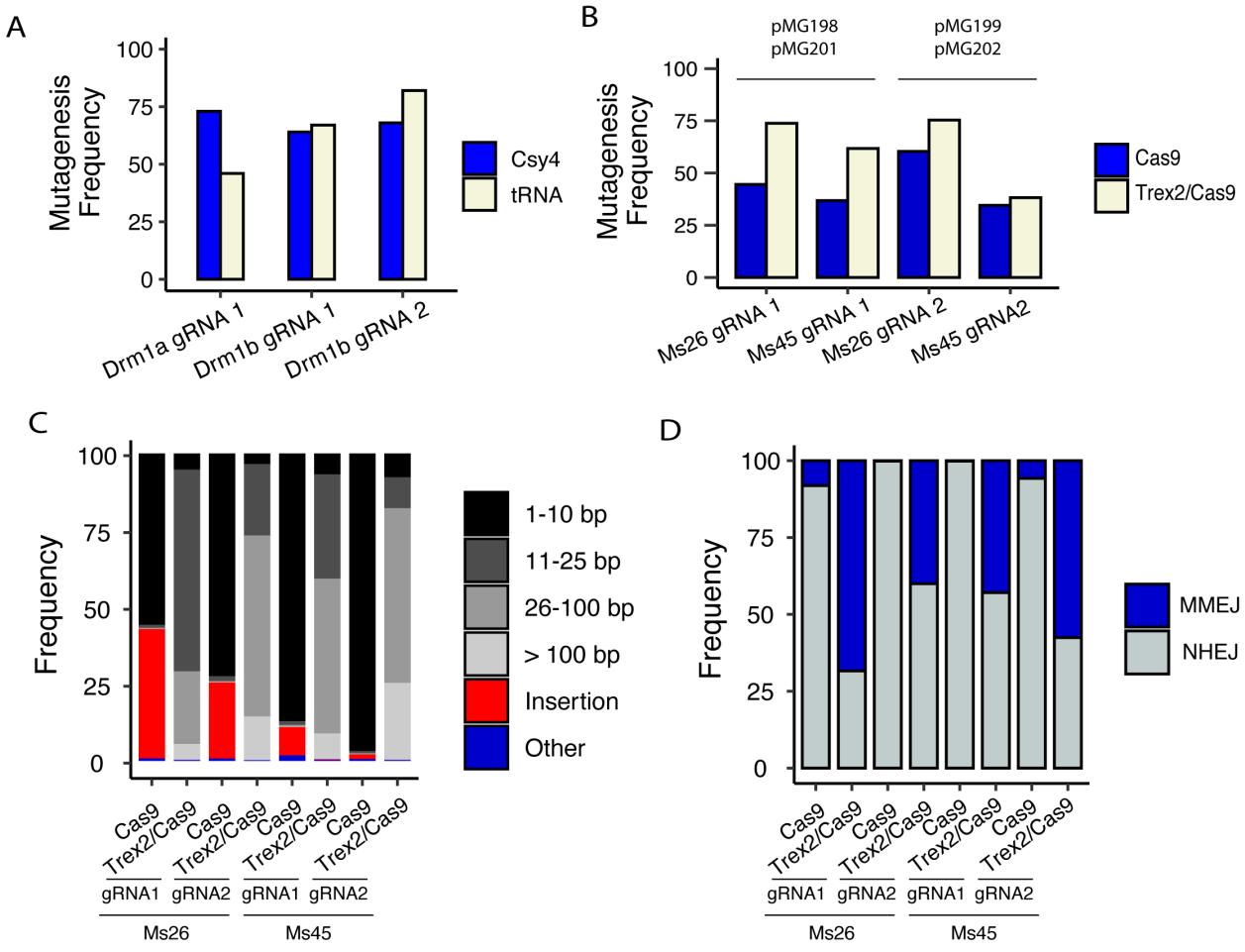


Figure 2. Analyses of mutagenesis frequencies and mutation profiles induced by the CRISPR/Cas9 systems. Mutagenesis frequency was calculated by dividing the total number of modified reads by the total number of reads. (A) Comparison of mutagenesis frequency mediated by the Csy4 (blue) and tRNA (beige) gRNA processing system. The gRNA sites, Drm1a gRNA 1 and Drm1b gRNAs 1 and 2 were targeted and analyzed by NGS. (B) Comparison of mutagenesis frequency induced by Cas9 (blue) and Cas9_Trex2 (beige). The gRNA sites, Ms26 gRNA 1, Ms26 gRNA 2, Ms45 gRNA1 and Ms45 gRNA 2, were targeted and analyzed by NGS. (C) Comparison of mutation profiles induced by Cas9 and Cas9_Trex2. The stacked bar graph was generated for each gRNA targeted site with either Cas9 or Cas9_Trex2. The deletions were represented on a

grayscale according to size, insertions were represented in red and all other reads (i.e. substitutions, substitutions plus deletions, substitutions plus insertions)were represented in blue. (D) Comparison of DNA repair outcomes induced by Cas9 and Cas9_Trex2. The frequencies of distinct DNA repair outcomes as either MMEJ (blue) or NHEJ (gray) were plotted in each sample treated with Cas9 or Cas9_Trex2.

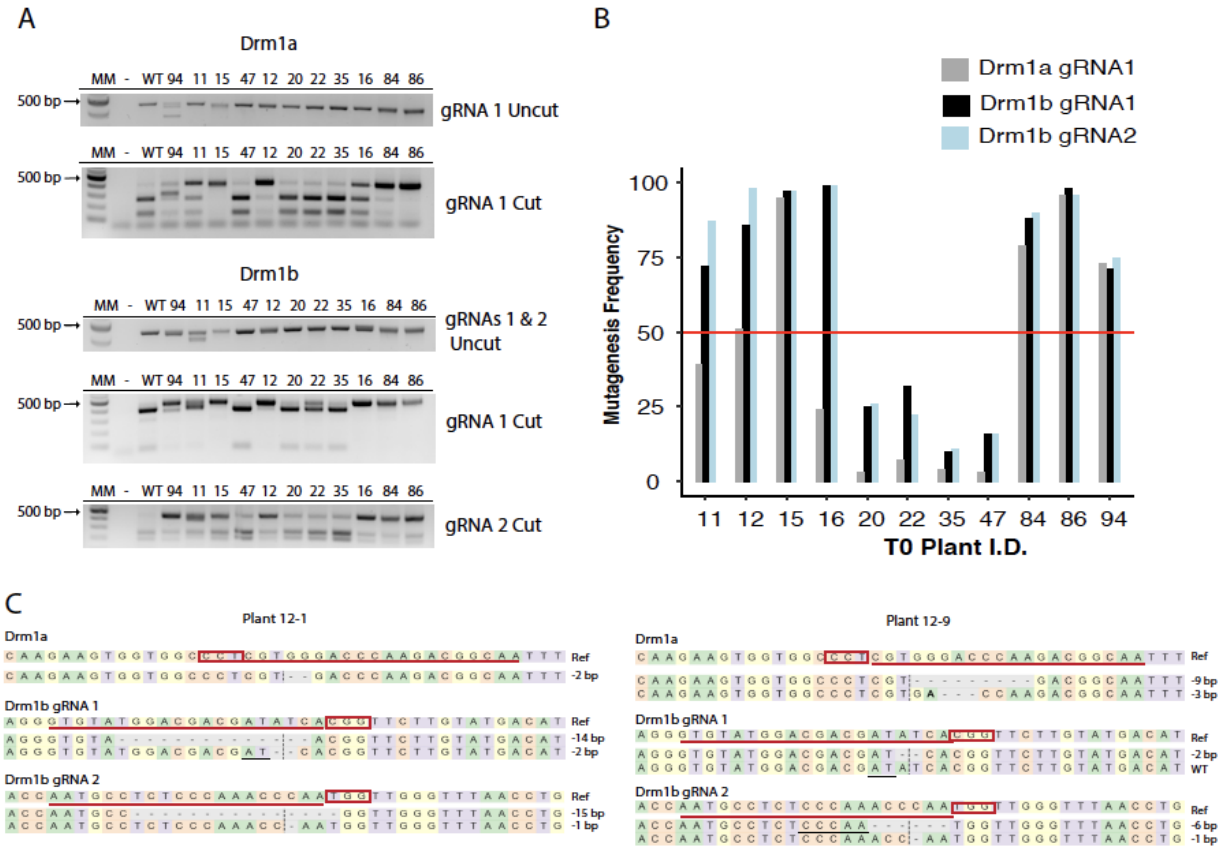


Figure 3. Characterization of mutations induced by the multiplex CRISPR/Cas9_Trex2 system in T0 and T1 plants. (A) Genomic PCR and CAPS genotyping assay. Eleven plants (numbered lanes) were genotyped by genomic PCR and the CAPS assay, with a 1kb ladder, no genomic DNA control (-), and wild type DNA control (WT). The results were presented as the PCR amplicons before restriction enzyme digestion (labeled as uncut), and after the digestion (labeled as cut) for each gRNA targeted site. (B) Mutagenesis frequency was determined for each T0 plant by NGS. The mutagenesis frequency was calculated by dividing the total number of modified reads by the total number of reads. The threshold of mutagenesis frequency with 50% was highlighted by the red line. (C) Genotypes of transgene-free T1 plants across all targeted sites. Sequence alignment was shown between the wild type reference sequences, indicated as Ref, and

the mutant sequences from Plant 12-1 and 12-9. The gRNA targeted sites were highlighted by the red line with the PAM sequences indicated in the red boxes and the cleavage sites indicated by the vertical dotted lines. All mutations found in these T1 plants were simple deletions. The deleted sequences were indicated by the dashed lines with the size of deletions indicated on the right. The microhomology sequences were highlighted by the black lines.

Supplemental Figures

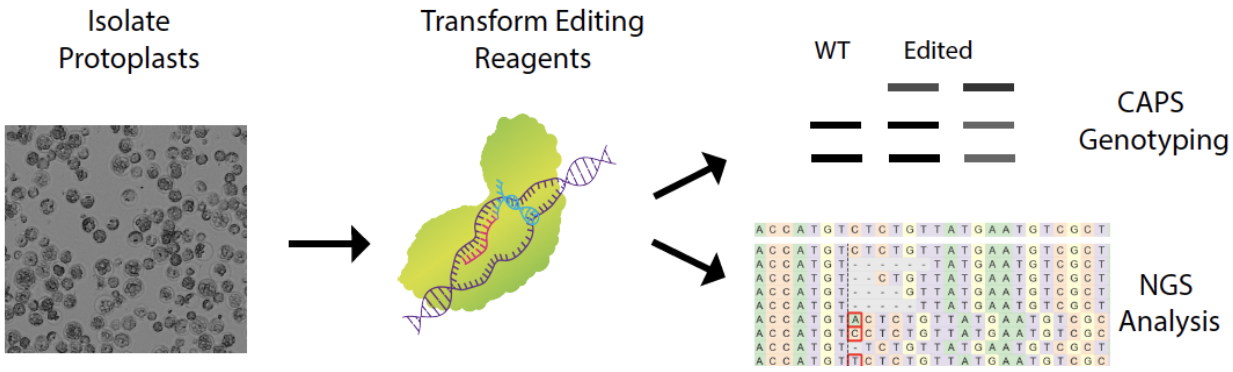


Figure S1. *Setaria viridis* protoplast assay pipeline to test genome editing reagents. Following protoplast isolation from young leaves of 14-day old plants, genome editing reagents can then be transformed into protoplasts to evaluate editing efficacy. At 48 hours post transformation (hpt), genomic DNA is extracted, subjected to Cleaved Amplified Polymorphic Sequences (CAPS) or Next Generation Sequencing (NGS) analysis.

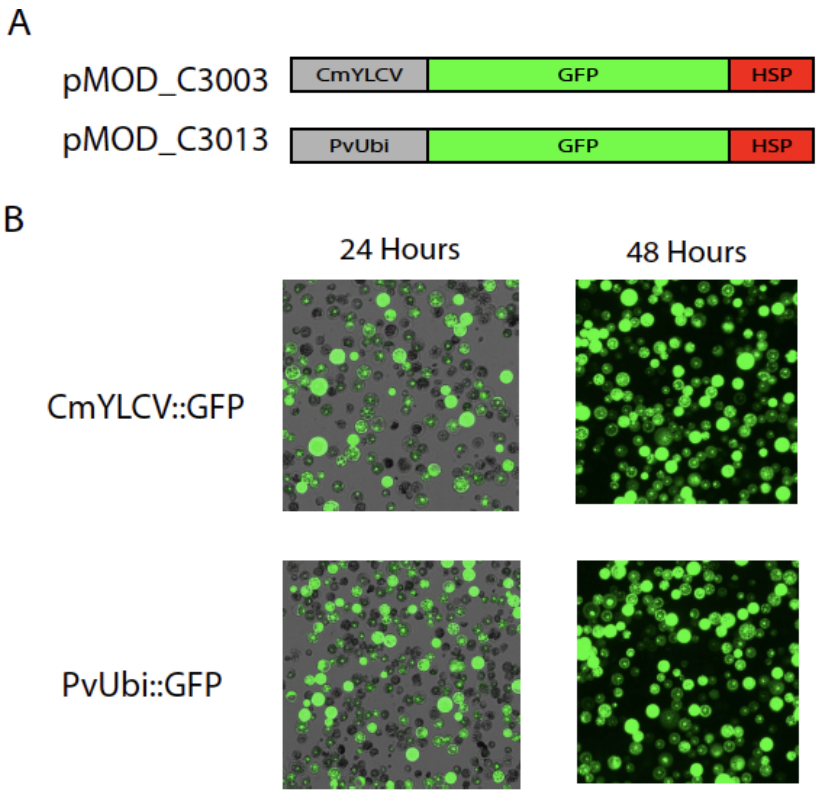


Figure S2. (A) The schematic structure of the GFP reporter plasmids used in the protoplast. The GFP coding sequence (green boxes) was driven by either the CmYLCV (plasmid ID pMOD_C3003) or the PvUbi2 promoter (plasmid ID pMOD_C3103) labeled as the gray boxes with the HSP terminator (red boxes). The illustration was not to scale. (B) Mesophyll protoplast cells isolated from young leaves and transformed with the GFP reporter plasmids, pMOD_C3003 and pMOD_C3013. GFP expression was assayed at 24 and 48 hpt.

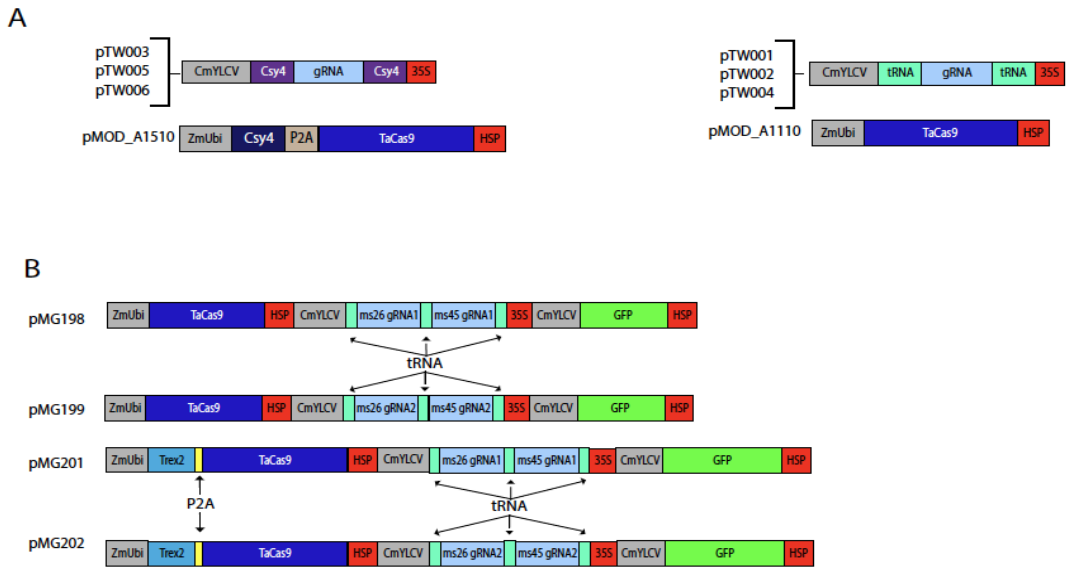


Figure S3. (A) The schematic structures of the plasmids to test the Csy and tRNA-based gRNA processing systems. In the Csy4 system, the Cas9 expressing plasmid, pMOD_A1510, contained the Cys4 coding sequence (dark blue box) with the Cas9 coding sequence (blue box) separated by the P2A sequence (yellow box) under the ZmUbi promoter (grey box). The gRNA expressing plasmids, pTW003, pTW005and pTW006, contained a single gRNA sequence (light blue box) flanked by the Csy4 recognition sites (purple boxes) under the control of the CmYLCV promoter and the 35S terminator (red box). In the tRNA system, the Cas9 expressing plasmid, pMOD_A1110, contained only the Cas9 coding sequences driven by the ZmUbi promoter. The gRNA expression plasmids, pTW001, pTW002 and pTW004, contained a single gRNA sequence red(light blue box) flanked by the tRNA sequences (light green boxes). (B) The schematic structures of the plasmids to the multiplexed Cas9 and Cas9_Trex2 systems. The plasmids, pMG198, pMG199, pMG201and pMG202, were constructed to contain three components, the Cas9 or Cas9_Trex2 expression cassette, the tRNA based multiplexing gRNA cassette, and the GFP reporter. The illustration was not to scale.

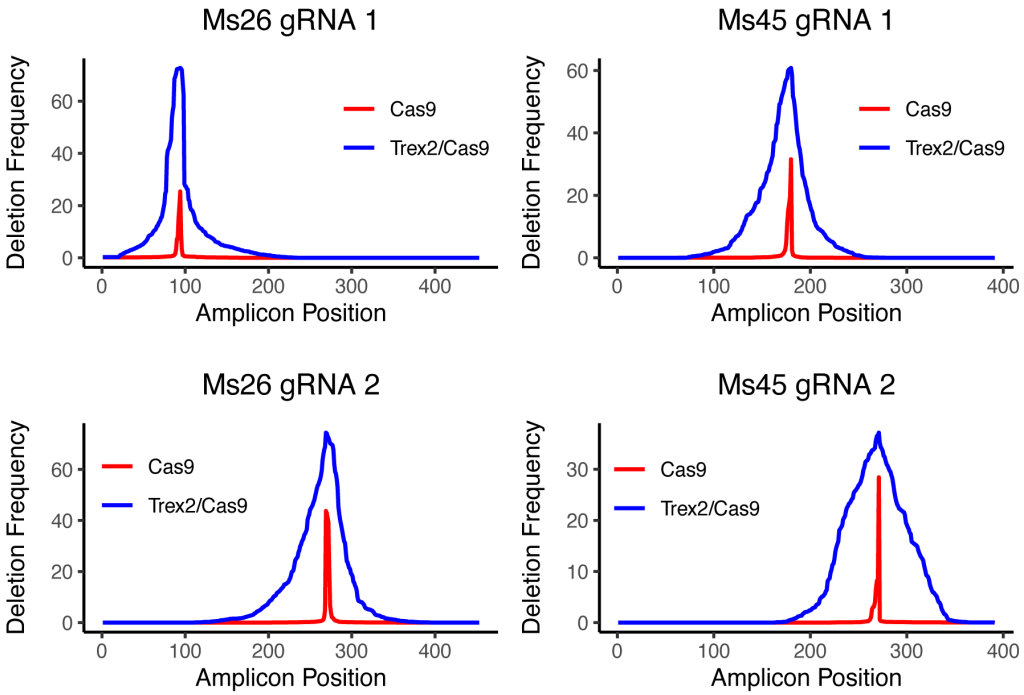


Figure S4. Distribution of deletions across the 400bp - 500 bp of the gRNA targeted regions induced by either Cas9 (red) and Cas9_Trex2 (blue). The Deletion frequency was calculated by dividing the total number of deletions at each nucleotide by the total number of deletions reads.

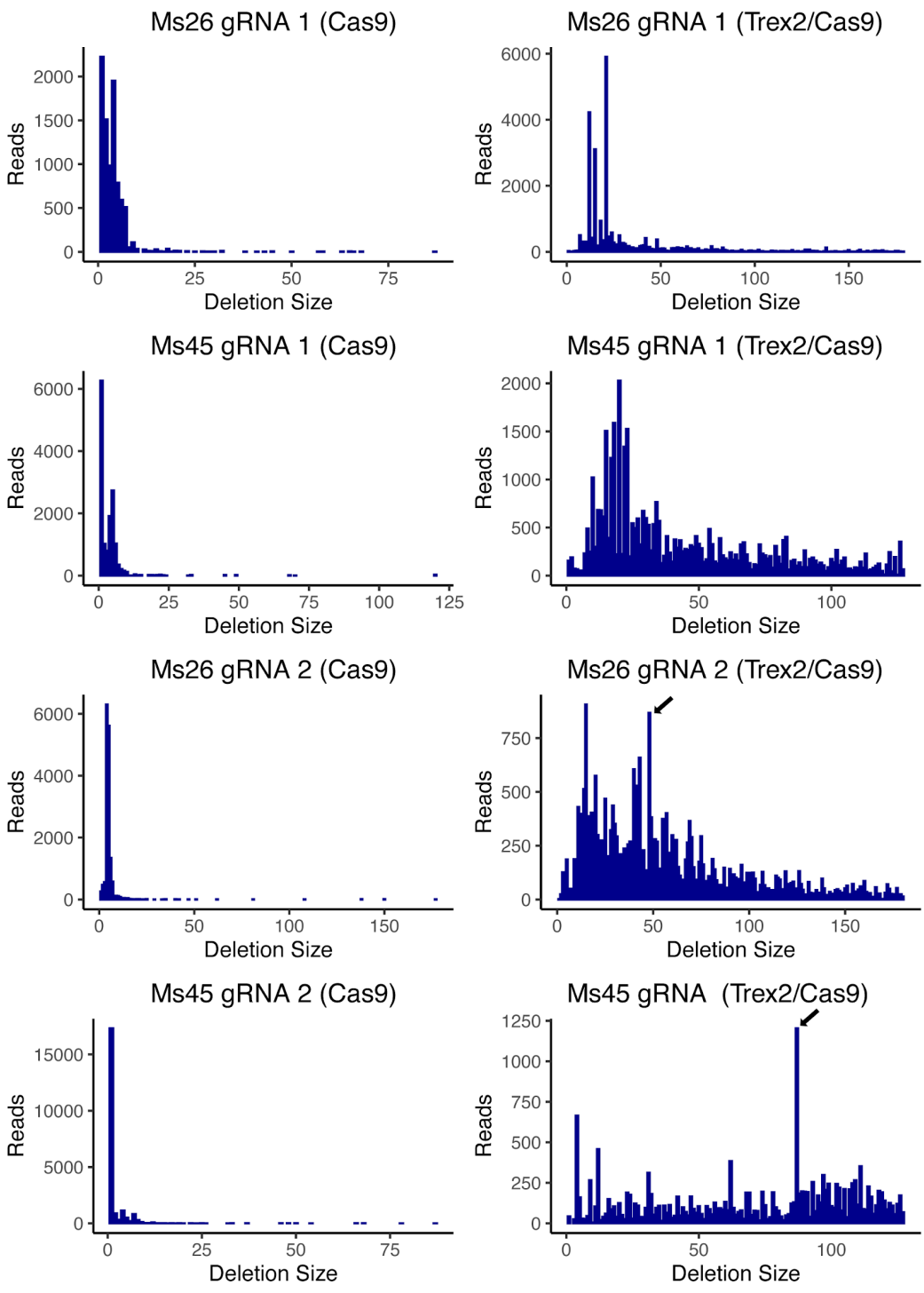


Figure S5. Comparison of the size distribution from deletions induced by Cas9 or Cas9_Trex2. The number of reads for each deletion size was estimated for 4 gRNA sites

from the Ms26 and Ms45 genes, respectively. The examples of MMEJ-mediated deletions were indicated by the black arrows.

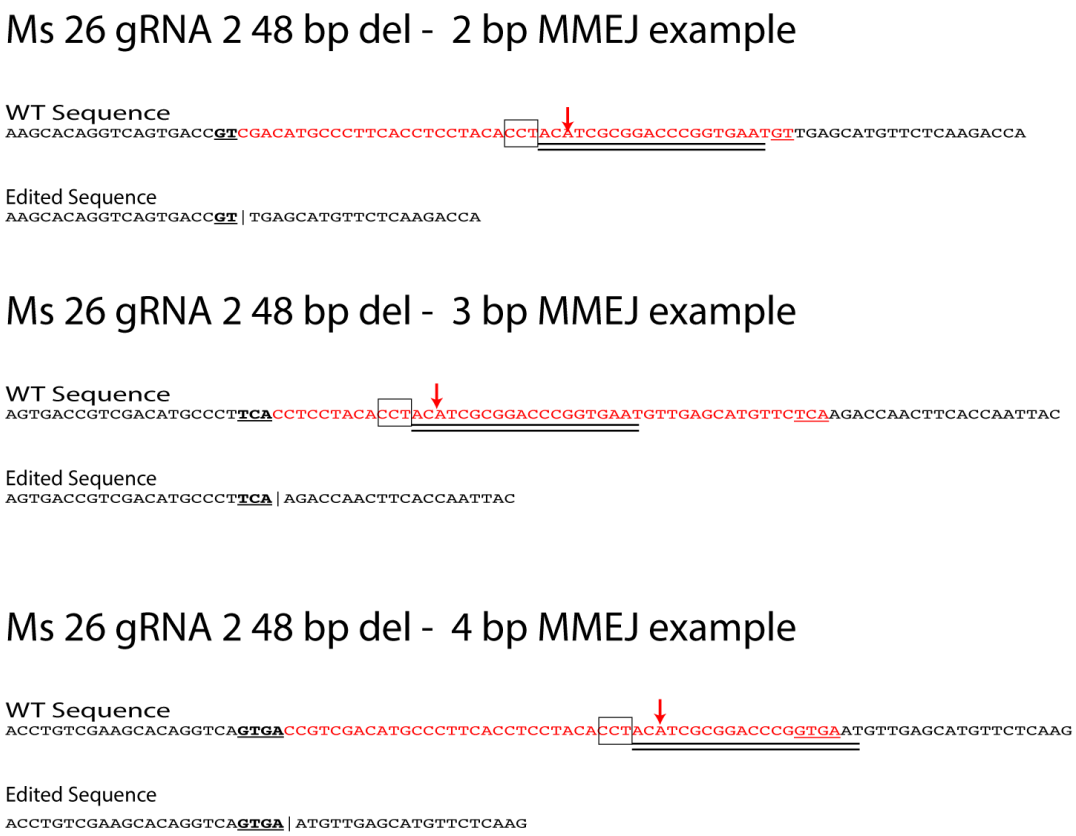


Figure S6. Example of the MMEJ-mediated deletions. The CRISPR gRNA target sites were indicated by the black double lines with the PAM sequences outlined with the black boxes. The CRISPR/Cas9 cleavage sites were pointed by the red arrows. The microhomology sequences were underlined in the wild type sequences with the deleted sequences highlighted in red.

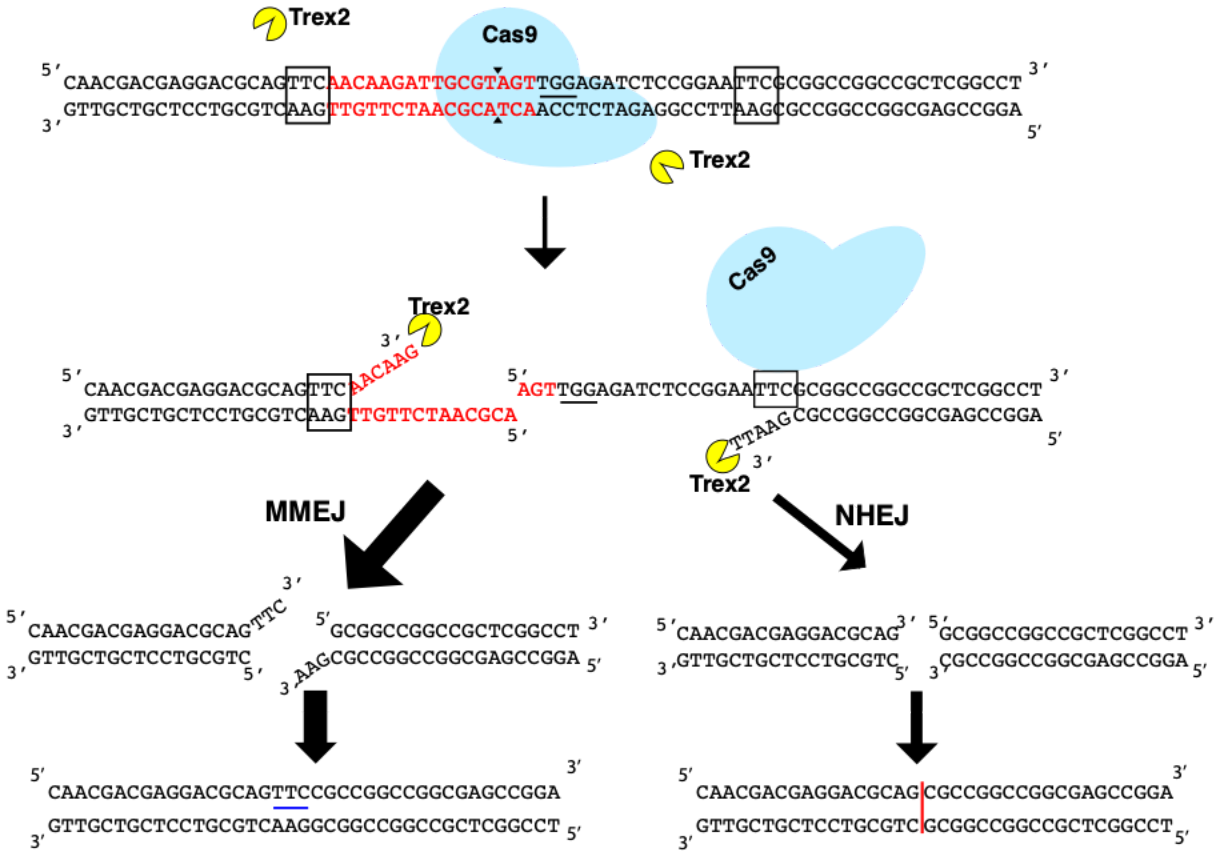


Figure S7. Model for DNA repair after a CRISPR/Cas9_Trex2-induced DSB. After Cas9 (light blue circles) binds the target genomic DNA and cleaves DNA creating a DSB, the Trex2 protein (yellow circles) then resections the exposed DNA 3' to 5'. The resected DSBs are repaired by either MMEJ, as indicated by the heavy-weighted arrow heads, or the NHEJ pathway, as indicated by the light-weighted arrow heads. The gRNA targeted site is highlighted in red, with the 3-bp PAM sequence underlined in black. The microhomology sequences are indicated by the black boxes. The junction site in the MMEJ repaired sequence is indicated by the blue line, whereas the junction in the NHEJ repaired sequence is indicated by the vertical red line.

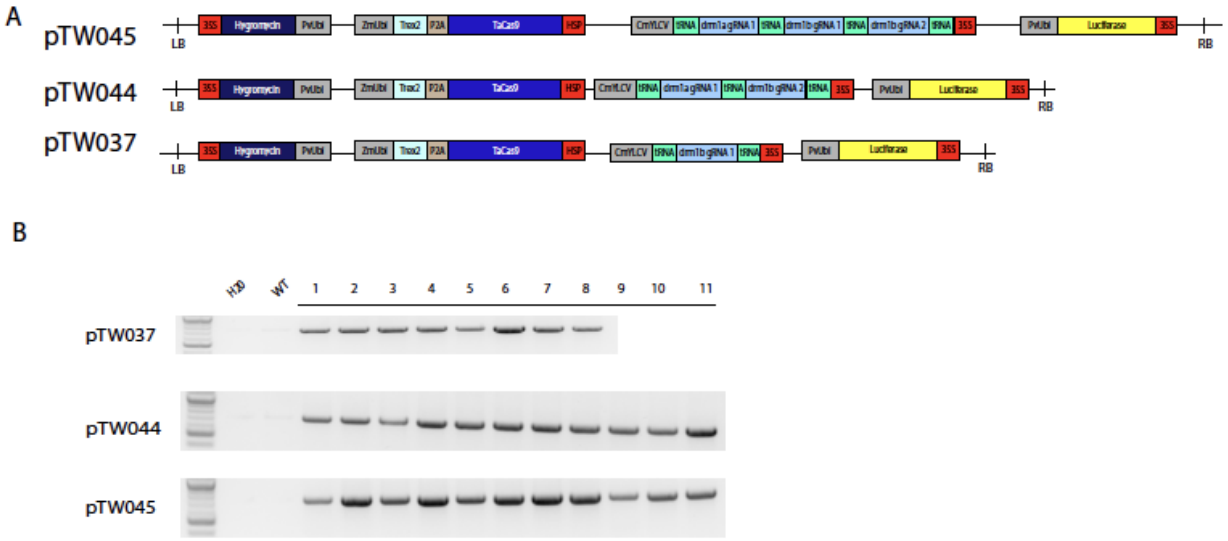


Figure S8. (A) The schematic structures of T-DNA binary plasmids for stable transgenesis. Each T-DNA plasmid contained three components, the Cas9_Trex2 expression cassette, the tRNA based multiplexing gRNA cassette, and the luciferase reporter. Within each construct, pTW037, contained one gRNA targeting Drm1b, pTW044 contained one gRNA targeting Drm1a and one gRNA targeting Drm1b, and pTW045 contained one gRNA targeting Drm1a and two gRNAs targeting Drm1b. (B) Genomic PCR genotyping to detect the presence of the T-DNA in T0 plants. Two controls were included in this experiment, one without genomic DNA (indicated as H2O) and the other with wild type S. virid genomic DNA (indicated as WT).

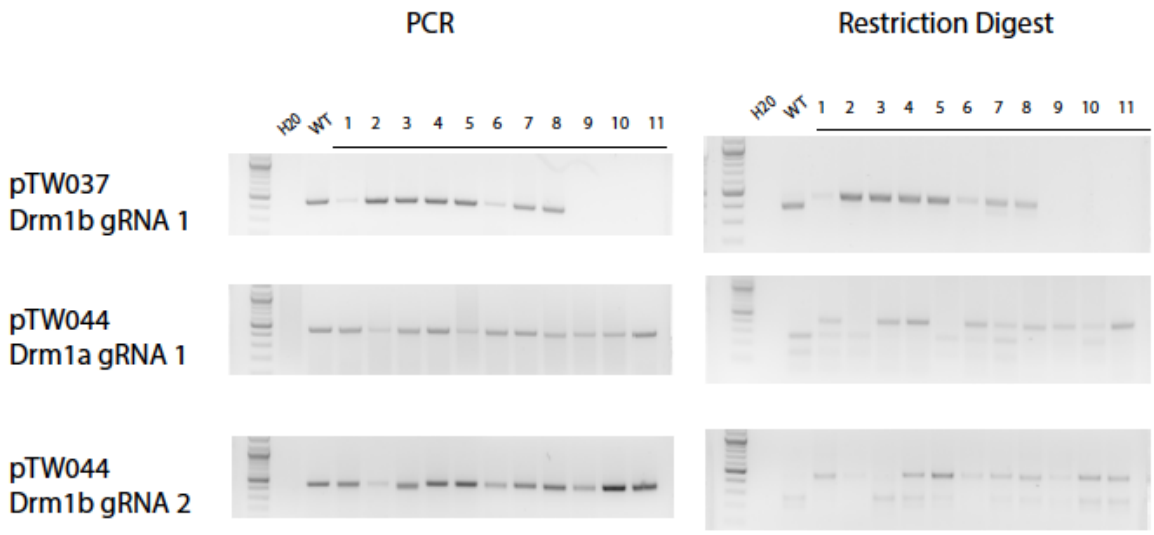


Figure S9. Genomic PCR genotyping of T0 plants with the CAPS assay. The samples from left to right were a 1 kb ladder, a no-genomic DNA control (indicated as H20), a control with wild type genomic DNA (indicated as WT) and dividual T0 samples.

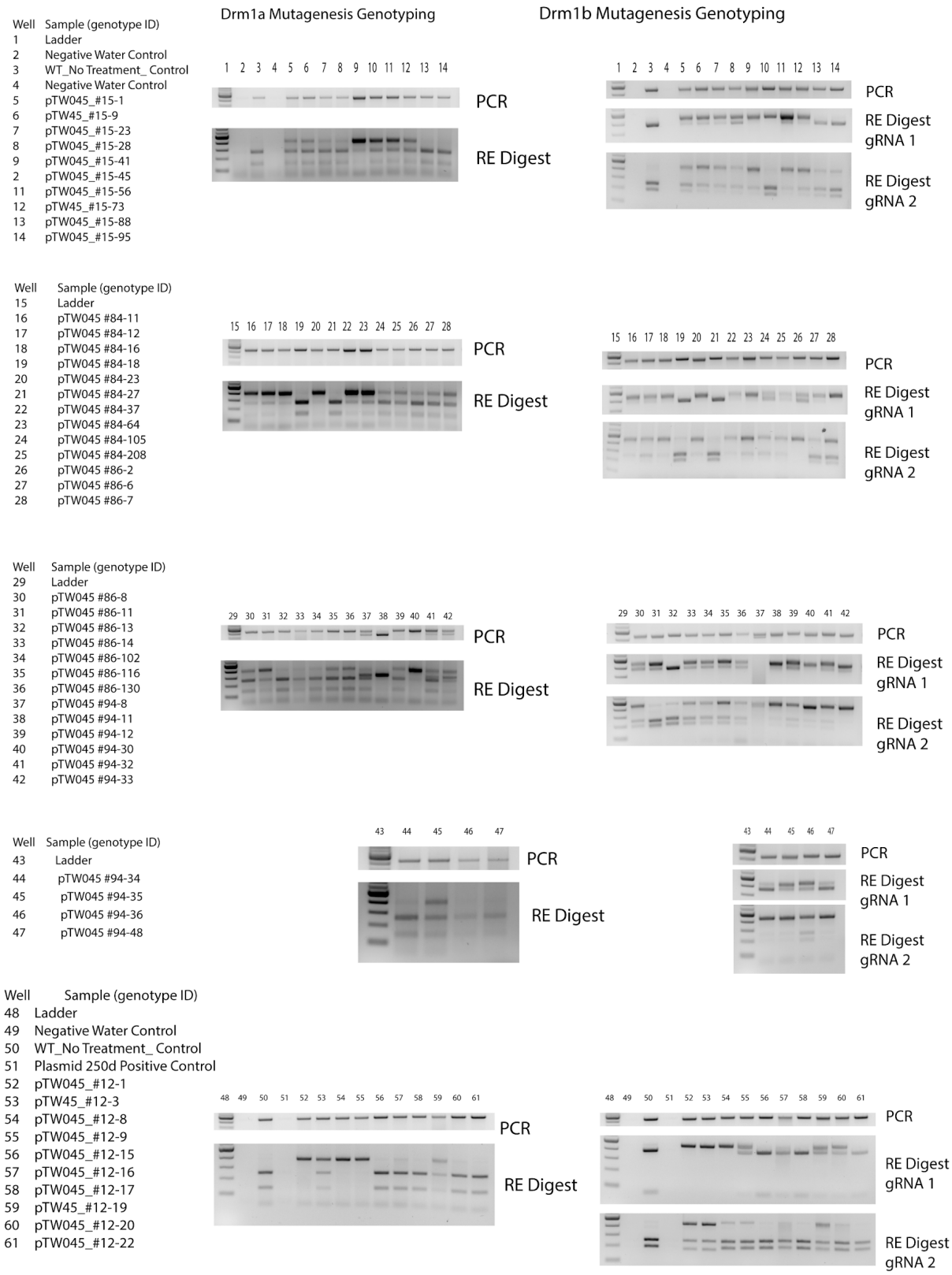
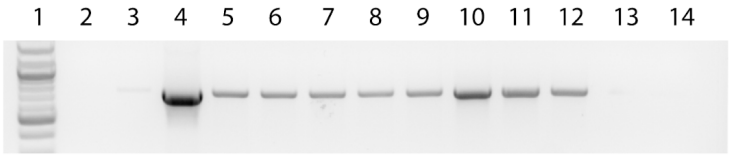


Figure S10. Genomic PCR genotyping of T1 plants with the CAPS assay. The samples

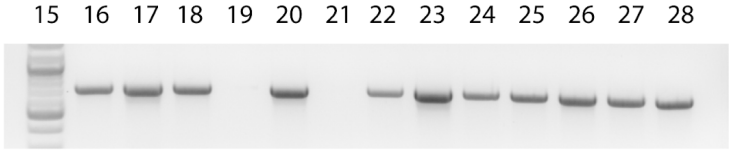
from left to right were 1 kb ladder, no-genomic DNA control (H20) and wild type genomic DNA control (WT). The remaining samples corresponded to individual T1 plants.

Hygromycin PCR

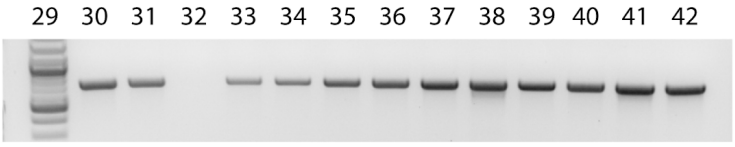
Well	Sample (genotype ID)
1	Ladder
2	Negative Water Control
3	WT_No Treatment_ Control
4	Plasmid 250d Positive Control
5	pTW045_#15-1
6	pTW45_#15-9
7	pTW045_#15-23
8	pTW045_#15-28
9	pTW045_#15-41
2	pTW045_#15-45
11	pTW045_#15-56
12	pTW45_#15-73
13	pTW045_#15-88
14	pTW045_#15-95



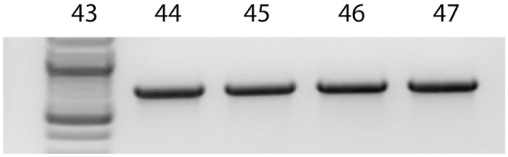
Well	Sample (genotype ID)
15	Ladder
16	pTW045 #84-11
17	pTW045 #84-12
18	pTW045 #84-16
19	pTW045 #84-18
20	pTW045 #84-23
21	pTW045 #84-27
22	pTW045 #84-37
23	pTW045 #84-64
24	pTW045 #84-105
25	pTW045 #84-208
26	pTW045 #86-2
27	pTW045 #86-6
28	pTW045 #86-7



Well	Sample (genotype ID)
29	Ladder
30	pTW045 #86-8
31	pTW045 #86-11
32	pTW045 #86-13
33	pTW045 #86-14
34	pTW045 #86-102
35	pTW045 #86-116
36	pTW045 #86-130
37	pTW045 #94-8
38	pTW045 #94-11
39	pTW045 #94-12
40	pTW045 #94-30
41	pTW045 #94-32
42	pTW045 #94-33



Well	Sample (genotype ID)
43	Ladder
44	pTW045 #94-34
45	pTW045 #94-35
46	pTW045 #94-36
47	pTW045 #94-48



Well	Sample (genotype ID)
48	Ladder
49	Negative Water Control
50	WT_No Treatment_ Control
51	Plasmid 250d Positive Control
52	pTW045_#12-1
53	pTW45_#12-3
54	pTW045_#12-8
55	pTW045_#12-9
56	pTW045_#12-15
57	pTW045_#12-16
58	pTW045_#12-17
59	pTW45_#12-19
60	pTW045_#12-20
61	pTW045_#12-22

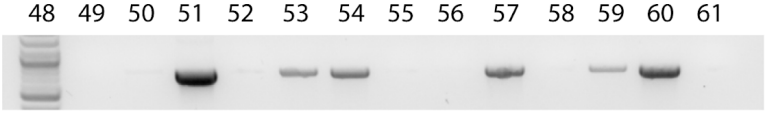


Figure S11. Genomic PCR genotyping for segregation of the T-DNA transgenes in T1 plants. The samples from left to right were 1 kb ladder, no-genomic DNA control (H20), and wild type genomic DNA control (WT) and dividual T1 samples.

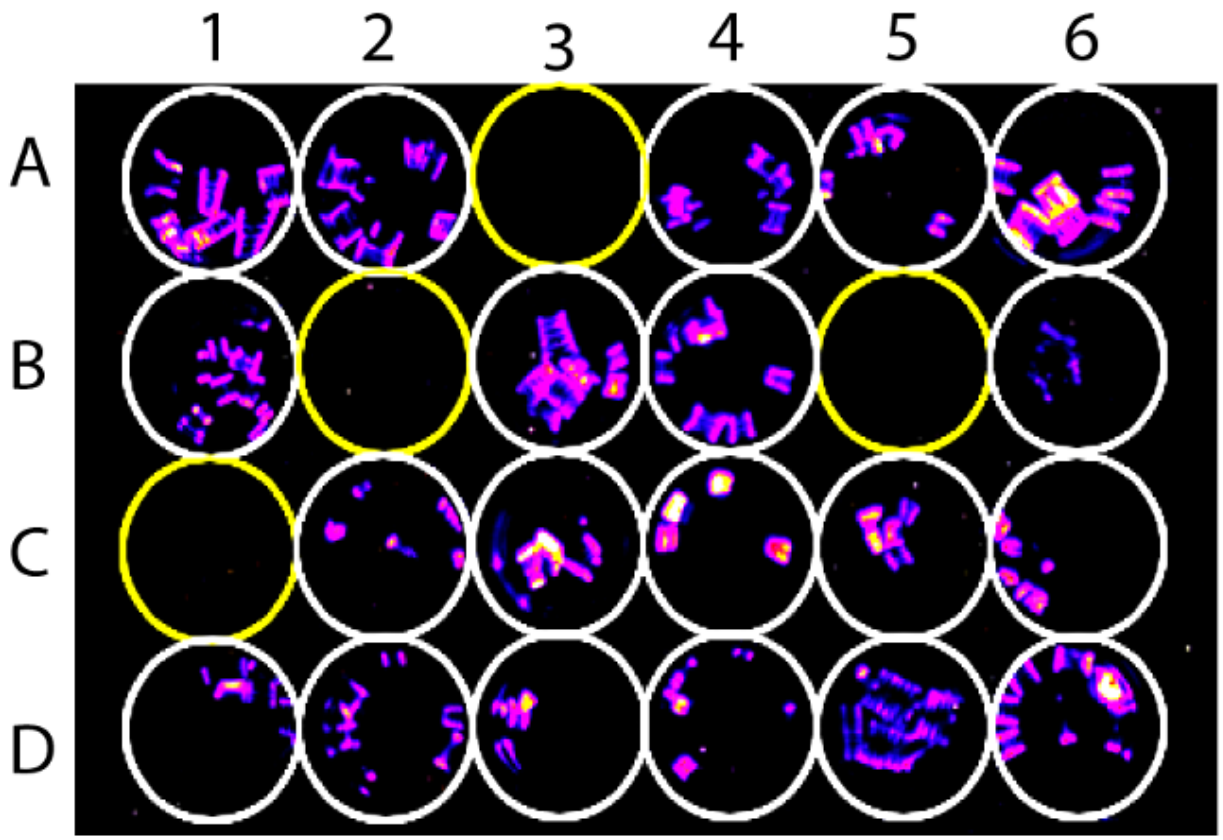


Figure S12. The luciferase assay for transgene-free plant screening. In this 24-well plate, each well contained the leaf samples collected from individual T1 plants. The plate layout was organized as A1 and A2, transgenic controls: A3, WT transgene-free plant; A4, 15-28; A5, 15-41; A6, 15-45; B1, 15-56; B2, 15-95; B3, 12-16; B4, 86-130; B5, 84-18; B6, 84-12; C1, 86-13; C2, 94-33; C3, 94-32; C4, 94-8; C5, 86-6; C6, 94-34; D1, 94-35; D2, 94-36; D3, 94-48; D4, 86-7; D5, 86-2; D6, 12-20. T1 plant samples without the luciferase activity, 15-95, 84-18 and 86-13 were highlighted by yellow circles.

Tables

Table 1. Summary of T0 plant characterization.

Characterization of T0 plants						
				# Plants with Mutations (Frequency)		
T-DNA Construct	# Calli	# Plants Regenerated	# Plants Genotyped	Drm1a (gRNA1)	Drm1b (gRNA1)	Drm1b (gRNA2)
pTW037	86	85	8 (100%)†	NA	8 (100%)	NA
pTW044	29	26	11 (100%)†	9 (82%)	NA	8 (73%)
pTW045	112	103	11(100%)†	11 (100%)	11 (100%)	11 (100%)

† Frequency of transgenic plants indicated in parentheses was confirmed by PCR

Table 2. Summary of T1 plant characterization.

			Number of Plants with mutations (Frequency)		
T0 Plant I.D.	# T1 Plants	# Transgene-free T1	Drm1a (gRNA1)	Drm1b (gRNA1)	Drm1b (gRNA2)
12	10	5	5 (50%)	6 (60%)	3 (30%)
15	10	2	8 (80%)	8 (80%)	9 (90%)
84	10	2	8 (80%)	8 (80%)	9 (90%)
86	10	1	10 (100%)	9 (90%)	10 (100%)
94	10	0	7 (70%)	8 (80%)	10 (100%)

Table 3. Summary of transgene-free T1 plant characterization.

Genoype Characterization of Transgene-free T1 Plants

Plant ID	*Drm1a*	*Drm1b* target 1	*Drm1b* target 2
12-1	homozygous (-2bp/-2bp)†	bi-allelic (-14bp/-2bp)†	bi-allelic (-15bp/-1bp)†
12-9	bi-allelic (-9bp/-3bp)†	heterozygous (-2bp/WT)†	bi-allelic (-6bp/-1bp)†
12-15	WT‡	WT	WT
12-17	WT	WT	WT
12-22	WT	WT	WT
15-88	WT	WT	heterozygous
15-95	WT	WT	heterozygous
84-18	WT	WT	heterozygous
84-27	WT	WT	WT
86-13	heterozygous	WT	heterozygous

† Genotypes were confirmed by Next Generation Sequencing. The size of deletion was indicated in the parentheses.
‡ WT meant the wild type sequence without mutations.

Supplemental Tables

Table S1. Summary of the targeted genes.

	Ms26	*Ms45*	*Drm1a*	*Drm1b*
Gene Name	Sevir.9G530800.1	Sevir.9G459500.1	Sevir.9G574800.1	Sevir.9G496200.1
Location	Chr_09:52841414..52843565	Chr_09:47642669..47644857	Chr_09:55520371..55525254	Chr_09:50382794..50387481
Sequence Similarity †	96.10%	87.10%	81.10%	76.37%
† DNA sequence similarity was estimated by comparing with the maize orthologous genes.				

Table S2. Summary of Next Generation Sequencing reads for each targeted site.

		Total		Unique	
Gene (gRNA)	Trex2 Co-expression	Reads	Modified Reads	Modified Read	Deletions
Ms26 (gRNA 1)	No	32,088	25,802	991	978
Ms26 (gRNA 2)	No	55,292	33,382	1,684	1,658
Ms45 (gRNA 1)	No	63,134	23,230	1,367	1,354
Ms45 (gRNA 2)	No	80,344	27,766	1,436	1,427
Ms26 (gRNA 1)	Yes	60,676	44,793	3,766	3,751
Ms26 (gRNA 2)	Yes	60,578	45,669	3,633	3,620
Ms45 (gRNA 1)	Yes	96,953	59,862	3,803	3,709
Ms45 (gRNA 2)	Yes	67,750	25,872	2,027	2,023

Table S3. Summary of off-targeting analyses on MS26 gRNA 1 using NGS

gRNA ID	# SNPs	gRNA Sequence†	Genomic Position	# Total Reads		# Modified Reads	
				Cas9	Trex2_Cas9	Cas9	Trex2_Cas9
Off-target site #1	3	CATCCTCCTCTAGTGGATCACGG	Chr_02 10.44 Mbp	822	862	0	0
Off-target site #2	3	CATCGTCCTGTCGTGGATCCTGG	Chr_08 23.77 Mbp	15079	15547	0	1
Off-target site #3	4	CATCGTCTTGTCGTGGATCCTGG	Chr_02 5.42 Mbp	14533	15098	0	1
Off-target site #4	4	CATCGTCATGTCGTGGATCCTGG	Chr_09 37.84 Mbp	814	619	0	0

† Sequecnes of off target gRNAs with the mismatched nucleotides in red and 3 bp PAM sequences underlined.

Table S4. Summary of T1 plant genotypes.

		T1 Genome Editing Characterization		
Plant ID	Transgenic	*Drm1a*	*Drm1b* target 1	*Drm1b* target 2
12-1	no	bi-allelic or homozygous	bi-allelic or homozygous	heterozygous
12-3	yes	heterozygous	bi-allelic or homozygous	heterozygous
12-8	yes	bi-allelic or homozygous	bi-allelic or homozygous	no mutation
12-9	no	bi-allelic or homozygous	heterozygous	no mutation
12-15	no	no mutation	no mutation	no mutation
12-16	yes	no mutation	no mutation	no mutation
12-17	no	no mutation	no mutation	no mutation
12-19	yes	heterozygous	heterozygous	heterozygous
12-20	yes	no mutation	heterozygous	no mutation
12-22	no	no mutation	no mutation	no mutation
15-1	yes	heterozygous	heterozygous	heterozygous
15-9	yes	heterozygous	heterozygous	heterozygous
15-23	yes	heterozygous	heterozygous	heterozygous
15-28	yes	heterozygous	heterozygous	heterozygous
15-41	yes	heterozygous	bi-allelic or homozygous	heterozygous
15-45	yes	heterozygous	bi-allelic or homozygous	no mutation
15-56	yes	heterozygous	heterozygous	heterozygous
15-73	yes	heterozygous	heterozygous	heterozygous
15-88	no	no mutation	no mutation	heterozygous
15-95	no	no mutation	no mutation	heterozygous
84-11	yes	heterozygous	heterozygous	heterozygous
84-12	yes	heterozygous	heterozygous	heterozygous
84-16	yes	heterozygous	heterozygous	heterozygous
84-18	no	no mutation	no mutation	heterozygous
84-23	yes	bi-allelic or homozygous	bi-allelic or homozygous	heterozygous
84-27	no	no mutation	no mutation	no mutation
84-37	yes	heterozygous	heterozygous	heterozygous
84-64	yes	heterozygous	heterozygous	heterozygous
84-105	yes	heterozygous	heterozygous	heterozygous
84-208	yes	heterozygous	heterozygous	heterozygous
86-2	yes	heterozygous	heterozygous	bi-allelic or homozygous
86-6	yes	heterozygous	heterozygous	heterozygous
86-7	yes	heterozygous	bi-allelic or homozygous	heterozygous
86-8	yes	heterozygous	heterozygous	heterozygous
86-11	yes	heterozygous	heterozygous	heterozygous
86-13	no	heterozygous	no mutation	heterozygous
86-14	yes	heterozygous	heterozygous	heterozygous
86-102	yes	heterozygous	heterozygous	heterozygous
86-116	yes	heterozygous	heterozygous	heterozygous
86-130	yes	heterozygous	heterozygous	heterozygous
94-8	yes	heterozygous	heterozygous	heterozygous
94-11	yes	bi-allelic or homozygous	no mutation	heterozygous
94-12	yes	heterozygous	heterozygous	heterozygous
94-30	yes	bi-allelic or homozygous	no mutation	heterozygous
94-32	yes	heterozygous	heterozygous	heterozygous
94-33	yes	heterozygous	heterozygous	bi-allelic or homozygous
94-34	yes	no mutation	heterozygous	bi-allelic or homozygous
94-35	yes	heterozygous	heterozygous	bi-allelic or homozygous
94-36	yes	no mutation	heterozygous	heterozygous
94-48	yes	no mutation	heterozygous	bi-allelic or homozygous

Table S5. Summary of the primer information.

Oligo Name	Sequence	Purpose
48_HYG	cgcgacgtctgtcgagaagtt	
51_HYG	gtctgctgctccatacaagcca	PCR for hygromycin selection marker
17_DRM	GCTTCAGAAGATGGTTGCTATGGC	
18_DRM	TTCCTGGCAGCCGCACATAA	PCR genotyping Drm1b
82_DRM	TCAAAGTTCTTCTGTGCCGCTG	
83_DRM	CGAGAGGTGATATGCAACAGTG	PCR genotyping Drm1a
445_MS	ATGGTGGAAGCTCATGCCAC	
446_MS	GAGCAGCACATCCATGTAGGA	PCR Genotyping Ms26 gene
462_MS	GTATCTCATGCTCGTGTCGGT	
471_MS	TGGAAATTAGTTCCGAAGACGT	PCR Genotyping MS45 gene
MS26-1-OT-1-F	ACCGTTGCAGTAAATCAGGCCA	
MS26-1-OT-1-R	GGATGCACCGCATGACCAAAG	Ms26 gRNA1 site #1 off-target analysis
MS26-1-OT-2-F	TTCATCTTCCGGCAGAACTC	
MS26-1-OT-2-R	AGAGCTTGGTCCCAGTCGT	Ms26 gRNA1 site #2 off-target analysis
MS26-1-OT-3-F	AGAACCAAGATCCGGTCCTC	
MS26-1-OT-3-R	AGAGCTTGGTCCCAGTCGT	Ms26 gRNA1 site #3 off-target analysis
MS26-1-OT-4-F	GTCGCACAGCTCCACTACC	
MS26-1-OT-4-R	TTCATCTTCCGGCAGAACTC	Ms26 gRNA1 site #4 off-target analysis

CHAPTER II: Context Statement

Summary

The Domains Rearranged Methyltransferases (DRMs) are crucial for RNA-directed DNA methylation (RdDM) in plant species. *Setaria viridis* is a model monocot species with a relatively compact genome that has limited transposable element content. CRISPR-based genome editing approaches were used to create loss-of-function alleles for the two putative functional DRM genes in S. viridis to probe the role of RdDM. Double mutant (*drm1ab*) plants exhibit some morphological abnormalities but are fully viable. Whole-genome methylation profiling provided evidence for wide-spread loss of methylation in CHH sequence contexts, particularly in regions with high CHH methylation in wild-type plants. Evidence was also found for locus-specific loss of CG and CHG methylation, even in some regions that lack CHH methylation. Transcriptome profiling identified genes with altered expression in the *drm1ab* mutants. However, the majority of genes with high levels of CHH methylation directly surrounding the transcription start site or in nearby promoter regions in wild-type plants do not have altered expression in the *drm1ab* mutant even when this methylation is lost, suggesting limited regulation of gene expression by RdDM. Detailed analysis of the expression of transposable elements identified several transposons that are transcriptionally activated in *drm1ab* mutants. These transposons likely require active RdDM for maintenance of transcriptional repression.

Chapter II has been adapted from my work in the following publication: “Genome-wide loss of CHH methylation with limited transcriptome changes in *Setaria viridis* domains rearranged methyltransferase (DRM) mutants”

Andrew Read*, **Trevor Weiss***, Peter A Crisp, Zhikai Liang, Jaclyn Noshay, Claire C Menard, Chunfang Wang, Meredith Song, Candice N Hirsch, Nathan M Springer, Feng Zhang (2022). The Plant Journal. doi: 10.1111/tpj.15781.

During the course of this work many authors contributed. In particular, Trevor Weiss, Meredith Song, Chunfang Wang, and Feng Zhang created the plant materials used in this study; Andrew Read, Peter Crisp, Zhikai Liang, Jaclyn Noshay, Claire Menard, Candice Hirsch, and Nathan Springer performed bioinformatics analysis; Trevor Weiss, Andrew Read, Nathan Springer and Feng Zhang wrote the manuscript. I have removed contact information and acknowledgements as well as formatted figures and references to be consistent throughout my thesis.

CHAPTER II: Genome-wide loss of CHH methylation with limited transcriptome changes in *Setaria viridis* domains rearranged methyltransferase (DRM) mutants

Introduction

DNA methylation is a common chromatin modification in many plant genomes. Cytosine methylation is the result of post-replication modification that adds a methyl group to the 5' carbon. While virtually all plants that have been assessed contain DNA methylation, there are differences in the levels and context-specific patterns of methylation in different species (Niederhuth et al., 2016). The majority of our knowledge about the molecular mechanisms that control DNA methylation and the functions of DNA methylation are based on studies in Arabidopsis thaliana (Arabidopsis) due to the viability of plants with highly reduced DNA methylation (Law & Jacobsen, 2010; Matzke & Mosher, 2014). However, studies in other plants have suggested differences in the patterns and control of DNA methylation (Niederhuth et al., 2016; Springer et al., 2016).

DNA methylation in plant genomes involves several distinct methyltransferases that create or maintain DNA methylation and these can be distinguished by the local sequence context (Law & Jacobsen, 2010). CG methylation is often present at high levels in plant genomes and is maintained following DNA replication due to the preference of MET1 and orthologous genes for hemimethylated sites (Law & Jacobsen, 2010). CHG (H = A, T or C) methylation is also quite common and is catalyzed by chromomethylase enzymes in a feed-forward loop with H3K9me2 (Du et al., 2012;

Johnson et al., 2007). CHH methylation occurs at non-symmetrical genomic sites and requires specific targeting mechanisms. Evidence from Arabidopsis suggests that the RNA-directed DNA methylation (RdDM) pathway is responsible for much of the CHH methylation (Cao & Jacobsen, 2002a; Stroud et al., 2014; Zemach et al., 2013). In this pathway, RNA Polymerase IV and V (PolIV and PolV) generate and utilize 24nt sRNAs to initiate or maintain CHH methylation at their target sequences. (Matzke & Mosher, 2014). Most of the RdDM activity is focused on either small TEs or the edges of longer TEs (Zemach et al., 2013). In maize, the RdDM activity seems to be particularly high at the edges of TEs near expressed genes (Gent et al., 2013; Q. Li et al., 2015). There is also evidence that some CHH methylation in Arabidopsis, particularly the CHH methylation found within internal regions of longer TEs, does not depend upon the DRM genes but instead requires activity of the CMT2 chromomethylase (Stroud et al., 2014; Zemach et al., 2013).

In plants, the DOMAINS REARRANGED METHYLTRANSFERASES (DRM) genes were identified as putative relatives of the mammalian de novo methyltransferase Dnmt3 with a unique rearrangement for the order of the methyltransferase domains (Cao et al., 2000). There are two putative functional DRM genes in Arabidopsis, DRM1 and DRM2, that are present as tandem duplicates. Evidence suggests that DRM2 is responsible for the bulk of RdDM in Arabidopsis but most studies utilize the drm1/drm2 double mutant to ensure complete loss of function (Cao & Jacobsen, 2002a; Stroud, Greenberg, et al., 2013b). Arabidopsis also encodes a third DRM-like gene, DRM3, that encodes a non-functional methyltrasnferase that plays a role in targeting or regulation of DRM1/2 (Costa-Nunes et al., 2014; Henderson et al., 2010). Most plant species have

several putative functional DRM genes as well as orthologs of the non-catalytic DRM3. Studies in Arabidopsis indicated the drm1/drm2 mutants had reduced CHH methylation and were compromised for silencing of some genes and TEs (Cao et al., 2003; Cao & Jacobsen, 2002b; Chan et al., 2004; Stroud, Greenberg, et al., 2013b; Tran et al., 2005). However, there are no substantial developmental or morphological abnormalities in Arabidopsis plants that lack DRM1/2 (Cao & Jacobsen, 2002b; Chan et al., 2006). Combining the drm mutant with a loss of function for CHROMOMETHYLASE3 (CMT3) results in significant phenotypic impacts suggesting partially redundant control of gene silencing and asymmetric methylation by DRM and CMT genes (Chan et al., 2006; Henderson & Jacobsen, 2008; Stroud et al., 2014). In Arabidopsis drm1/drm2 mutants there are substantial reductions of CHH methylation at many loci and there is partial reduction of CHG at these same regions that is completely reduced in drm1 drm2 cmt3 mutants suggesting combined control of CHH and CHG methylation by DRM and CMT at these sites (Stroud et al., 2014).

CHH methylation triggered by RNA directed DNA methylation (RdDM) appears to play a limited role in regulating expression (Cao & Jacobsen, 2002a; Stroud et al., 2014). There are certainly some endogenous loci and transgenes that are silenced by DRM or other components of the RdDM machinery. However, the number of genes or transposons that are activated in a drm1/drm2 mutant is limited, and it seems that there is substantial redundancy between DRM and CMT pathways to maintain silencing in Arabidopsis (Stroud et al., 2014). In contrast, rice plants with loss of function in DRM genes exhibit pleiotropic phenotypes. Loss-of-function for the rice DRM ortholog OsDRM2 results in pleiotropic phenotypes as well as aberrant expression of some

transposons and genes (Hu et al., 2021; Moritoh et al., 2012; Tan et al., 2016). Combining loss-of-function for OsDRM2 with CMT genes results in drastic changes in methylation and phenotype in rice (Hu et al., 2021).

In this study we sought to understand the roles of the DRM genes in *Setaria viridis*. *S. viridis* is an emerging model C4 grass that has a relatively small genome and a short generation time (Bennetzen et al., 2012; Brutnell et al., 2010; Mamidi et al., 2020; Thielen et al., 2020). Significant progress has been made to develop genetic and genomic resources for S. viridis, including two high quality reference genomes (A10 and ME034V accessions) and pan-genome sequence resources for 598 diverse genotypes (Mamidi et al., 2020; Thielen et al., 2020). Transposable elements (TEs) account for 46% of the genome and have accounted for structural variation impacting traits (Mamidi et al., 2020; Thielen et al., 2020). The proportion of the genome and distribution of TEs in S. viridis is similar to that observed for rice (40%) but much less than observed in the maize genome (85%) (Akakpo et al., 2020; Schnable et al., 2009).To date, epigenomic resources are largely lacking for S. viridis. In this work, we first develop a whole genome DNA methylation map for S. viridis accession ME034V, and use CRISPR-Cas9 to create loss-of-function alleles for the two putative functional DRM orthologs, DRM1a and 1b. Our results provide insight into how DNA methylation is controlled in monocot species, and present evidence of subtle differences in DNA methylation mechanisms in dicot and monocot species.

Results

Isolation of DRM loss-of-function alleles

The *Setaria viridis* genome has two genes that encode putatively functional orthologs of the Arabidopsis DRM1/2 genes; Drm1a - Sevir.9G574800 (A10) / Svm9G0069770 (ME034V) and Drm1b - Sevir.9G496200 (A10) / Svm9G0060600 (ME034V). These two genes are located ~5 Mb apart on chromosome 9 and represent a duplication event in Setaria that is not observed in maize or rice (Figure S1A). The two protein sequences have 67.1% identity and 76.5% similarity. Expression atlas data for both accession A10 and ME034v suggest that Drm1a is much more highly expressed than Drm1b in leaf tissue (Figure S1B-C). This suggests that Drm1a may have more functional relevance but there is potential redundancy for these two genes. There is also a DRM3-like gene, Sevir.3G052500 (A10) / Svm3G0006370 (ME034v), that lacks critical residues in the catalytic domain and is unlikely to provide functional methyltransferase activity. This is likely an ortholog of the Arabidopsis DRM3, which encodes a catalytically inert protein that appears to be required as a cofactor for proper CHH methylation at some loci (Henderson et al., 2010; Costa-Nunes et al., 2014)

A total of three guide RNAs (gRNAs) targeting Drm1a and Drm1b were designed as described in our previous study (Weiss et al., 2020). To generate S. viridis mutant plants with double gene knockouts, the T-DNA construct (pTW045) expressing Cas9_Trex2 with all three gRNA sequences was transformed through agrobacterium-mediated transformation into the transformable S. viridis genotype ME034V (Figure S2A; see methods for details). T0 plants with edits at both targeted genes were identified and selected for self-pollination. Progeny were screened, and two transgene free T1 plants were selected for further propagation: one containing edits at drm1a and drm1b (T1_12-9), and one plant with WT alleles (T1_84-27) (Weiss et al.

2020). To identify homozygous progenies with frame-shift mutations at both genes, two generations of self pollination were performed. T3 plants with the edits at both genes were identified (hereafter referred to as *drm1ab*), which included a 3 bp deletion in Drm1a that introduces an early stop codon as well as a 2bp and a 6bp deletion in Drm1b that results in a frameshift mutation (Figure 1A, Figure S2B). The predicted proteins produced by these mutant alleles both lack critical domains that are necessary for methyltransferase activity. The *drm1ab* plants are fully viable (Figure 1B). The *drm1ab* plants are reduced in stature and have reduced leaf length (Figure 1B-C). In addition, the *drm1ab* plants exhibit delayed flowering relative to wild-type. The severity of the change in stature and flowering time were variable in different growth conditions.

Characterization of methylation domains within the *Setaria viridis* genome

Whole genome DNA methylation profiles were generated for a single replicate sample of wild-type S. viridis ME034V as well as three biological replicates from plants whose parent (T1_84-27) was regenerated from tissue culture and three biological replicates of transgene-free *drm1ab* plants. All samples were collected from seedling leaf tissue at a developmental stage in which there are no phenotypic differences between the mutant and wild-type plants. Enzymatic conversion rates (based on alignments to the chloroplast genome) for all samples ranged from 99.43 to 99.81%. The genome-wide DNA methylation levels for wild-type ME034V plants (Figure 2A, S3A) are quite similar to reported levels for S. viridis accession, A10 (Niederhuth et al., 2016). Prior studies have suggested some changes in DNA methylation induced due to tissue culture in rice and maize (Stroud, Ding, et al., 2013; Han et al., 2018). We did not observe significant

differences in the overall DNA methylation levels between wild-type and tissue-culture derived samples (Figure S3A). The genome was divided into 100 bp tiles, each of which was classified based on the levels of CG, CHG and CHH methylation (Figure S3B-C; see Methods for details). The wild-type and tissue culture derived plants had very similar proportions of the genome classified as high CG and CHG (~28% of genome), CG-only (~15% of genome), or high CHH (~1.4% of genome) (Figure S3B). For the analysis of *drm1ab* we focused on contrasts between the three biological replicates of tissue-culture derived plants and the loss-of-function lines to ensure that the differences we detect would not be solely due to tissue-culture induced changes in methylation.

Widespread loss of CHH methylation in *drm1ab* mutant plants

Context-specific DNA methylation levels were evaluated genome-wide (Figure 2A) and using metaprofiles over genes or TEs (Figure 2B-C). This revealed substantial loss of CHH methylation in *drm1ab* relative to the control. CHH methylation is lost in regions that exhibit elevated levels of methylation including the regions surrounding genes as well as within TEs (Figure 2B-C). However, there is still CHH methylation remaining in the *drm1ab* plants, especially within TEs. A visualization of several genomic regions revealed that regions of high CHH methylation in the tissue culture control plants can be divided into regions that require DRM (DRM-dependent) and regions that have CHH methylation that is not dependent upon DRM (DRM-independent) (Figure 2D). The proportion of CHH methylation that is lost in *drm1ab* was assessed for genomic regions with varying levels of CHH methylation in the control. The majority (79.7%) of genomic loci with high (>20%) CHH methylation in the control exhibit at

least 80% loss of this methylation in *drm1ab* plants. In contrast, regions with intermediate (5%-20%) or low (2-5%) levels of CHH methylation only lose methylation at 43.3% or 11.8% of loci, respectively. Thus, the loss of DRM activity results in loss of CHH methylation at regions with high CHH methylation but rarely affects the large number of regions with low levels of CHH methylation.

In order to gain a better understanding of CHH methylation changes, and how these are related to CG and CHG methylation, we identified all 100bp tiles with >20% CHH methylation in the control sample. The methylation levels of these tiles were assessed in *drm1ab* to identify DRM-dependent tiles (>80% methylation loss in *drm1ab*), DRM-intermediate tiles (20-80% methylation loss in *drm1ab*), and DRM-independent CHH tiles (<20% methylation loss in *drm1ab*) (Table 1). The CG, CHG and CHH methylation levels in both control and *drm1ab* were evaluated at these regions (Figure 3). The majority (79%) of regions with CHH levels >20% in wild-type are DRM-dependent with only 4% that are DRM-independent (Table 1). We observed that DRM-dependent CHH methylation is often accompanied by high levels of CG and CHG methylation (Figure 3). In *drm1ab*, the CHG methylation is lost in the vast majority of these regions and the CG methylation is reduced at some loci, but not at others (Figure 3). The regions of the genome where we observed DRM-independent CHH methylation generally have high levels of CG and CHG methylation that are not dependent upon DRM (Figure 3).

DRM is expected to function in the RdDM pathway and to primarily contribute to maintenance of CHH methylation. However, genome-wide levels of CG and CHG methylation show 2-6% reductions (Figure 2A-B). Similar findings were reported for rice DRM mutants (Hu et al. 2021). Methylation levels in all contexts were determined at

each 100 bp genomic tile in the control and *drm1ab* samples. Differentially methylated tiles were classified as either hypermethylated (higher methylation in *drm1ab*) or hypomethylated (lower methylation in *drm1ab*) (Figure S4). In all three contexts there are many more examples of hypomethylated tiles in *drm1ab* relative to the control (Figure S4). All genomic regions with high (>40%) CG or CHG methylation levels in the control sample were evaluated to determine what proportion are DRM-dependent (Table 1). In contrast to high CHH regions in which the majority are dependent on DRM, only a small proportion of these high CG or CHG methylated regions are affected in *drm1ab*. However, since the number of genomic regions with high CG or CHG methylation vastly outnumbers the regions with elevated CHH, there are more total CG or CHG hypomethylated tiles genome-wide (Table 1).

The *drm1ab* mutant might be expected to affect CG and CHG methylation in regions with high CHH methylation, but we did not expect substantial changes in CG and CHG methylation at regions without CHH methylation. We sought to determine whether the changes in CG and CHG methylation in *drm1ab* co-occurred with changes in CHH methylation or whether some CG and CHG methylation losses occurred in regions without CHH methylation. We found many examples of CG or CHG DRM-dependent hypomethylation at tiles with low or no CHH methylation (Figure 3). A comparison of the regions with CG and CHG methylation loss found some examples of dual loss in both contexts as well as many with specific loss in CG or CHG (Figure 4A). The CG, CHG or CG/CHG methylation losses were then evaluated to assess what proportion overlap, are within 300bp of, or are greater than 300bp from a tile with reduced CHH methylation (including both DRM-dependent and CHH DRM-intermediate tiles). A portion of the

DRM-dependent CG (9.7%) or CHG (24%) methylation is found in regions that have moderate or high levels of CHH that require DRM (Figure 4A). In addition, another 4-10%% of the CG or CHG dependent methylation occurs within 300 bp of a region with CHH methylation loss in *drm1ab* (Figure 4B). This suggests that the loss of a small region of CHH methylation can be associated with broader loss of CG and/or CHG methylation at some loci (example in Figure S5). The mechanisms leading to losses of CG or CHG methylation in *drm1ab* in regions distal to CHH methylated regions are unclear.

Limited changes in gene expression in DRM mutants

Transcriptome profiling was performed using RNAseq on the same seedling tissue samples used for WGBS with the addition of multiple replicates for wild-type ME034V. Principal component analysis (PCA) showed limited variation between wild-type ME034V and tissue culture derived ME034V plants, and, as expected based on the PCA result, there are very few differentially expressed genes between these samples (Figure S6, 5A). However, PCA clustering and differential gene expression analysis finds evidence for hundreds of gene expression changes in *drm1ab* plants (Figure 5A-B). The *drm1ab* plants have more genes that are up-regulated compared to the control samples, including 136genes with >10-fold up-regulation (Figure 5A-B). The observed changes in expression in *drm1ab* plants may represent direct effects of changes in DNA methylation on gene expression or could represent secondary effects due to a small number of direct targets that influence expression of other genes. In order to identify potential direct effects of loss of RdDM we initially focused on the subset of genes with

CHH methylation immediately over the TSS region. We identified 1,043 genes with CHH methylation (>20%) in the region immediately surrounding the annotated transcription start site (TSS) in tissue culture control plants. Many (529) of these genes that contain high CHH immediately at or surrounding the TSS are expressed in the control, suggesting that the presence of CHH at or near the TSS is not necessarily silencing gene expression. While the vast majority of these genes (98%) are hypomethylated in *drm1ab*, only 3.5% are differentially expressed (24 up in *drm1ab*, 12 down in *drm1ab*) (Figure 5C). Over 95% of the genes with elevated CHH methylation surrounding the TSS do not exhibit changes in expression in *drm1ab*. These observations suggest that there are relatively few genes that are direct targets for silencing by RdDM in seedling leaf tissue of *Setaria viridis*.

Previous analysis of DNA methylation in monocots has shown that methylated CHH regions (mCHH islands) are often found upstream of highly expressed genes (Niederhuth et al., 2016; Li et al., 2015; Gent et al., 2013). It is unclear whether the mCHH islands influence nearby gene expression or if open chromatin associated with gene expression enables DRM-dependent methylation of these regions. We sought to determine if losses of CHH methylation at mCHH islands in *drm1ab* resulted in changes in expression of these genes. We classified 5,424 genes as having an mCHH island (requires at least one 100bp tile with >20% CHH methylation in the 1kb promoter region). Many (3,171) of these genes are expressed in control conditions and these genes tend to be expressed higher than genes without mCHH islands as previously observed in maize (Li et al., 2015; Gent et al., 2013). The genes containing mCHH islands exhibit only slightly higher proportions of DEGs than in all genes and there are similar

proportions of up- and down-regulated genes in *drm1ab* relative to control (Figure 5C). Over 95% of the genes with an mCHH island do not show altered expression in the *drm1ab* plants, suggesting that the presence of mCHH islands in gene promoters has limited functional significance for gene expression levels in seedling leaf tissue.

Identification of transposable elements that are up-regulated in *drm1ab* mutants

RdDM has been shown to play important roles in maintaining silencing of transposable elements in Arabidopsis (Cao et al., 2003; Tran et al., 2005; Stroud, Greenberg, et al., 2013; Chan et al., 2004; Cao and Jacobsen, 2002a). However, in Arabidopsis and other plants it has been shown that there is often redundant control of transposable element (TE) silencing through multiple DNA methylation pathways and sole loss of CHH methylation only results in limited TE activation (Chan et al., 2006; Henderson and Jacobsen, 2008; Stroud et al., 2014). We sought to investigate whether there are Setaria TEs that are transcriptionally activated in the *drm1ab* plants. The Extensive de-novo TE annotator (EDTA) (Ou et al., 2019) pipeline was used to perform both a structural and a homology based annotation of TEs in the S. viridis ME034V genome. Two distinct approaches were used to monitor expression of TEs. The first approach assessed uniquely aligned RNAseq reads that align to regions annotated as TEs using the structural annotation of intact ME034V TEs. There were 454 TEs with detectable expression (>5 uniquely mapping reads) in either tissue culture or *drm1ab* samples and 33 of these were differentially expressed (padj<.05 and greater than 2-fold change) between tissue culture and *drm1ab* samples (Figure 6A). The majority (25/33) of the differentially expressed TEs were up-regulated in *drm1ab*, and 14 of these had little

or no expression in the control plants indicating the requirement for RdDM to maintain effective silencing of these TEs. The up-regulated TEs include 11 class I retrotransposons as well as 14 class 2 terminal inverted repeat and helitron DNA transposons (Table S1). It is worth noting that this approach was useful to identify TEs that are up-regulated, but has two significant limitations. First, the reliance on uniquely mapping reads limits detection of TE families with multiple highly similar elements which might be common in TE families regulated by RdDM. Second, in many cases the expression that was detected for TEs likely reflects partial transcripts rather than expression of the full-length TE.

The second approach to monitor TE expression was implemented using a de novo transcriptome assembly of the RNAseq reads from the wild-type and *drm1ab* plants. This enables the identification of TEs that generate potential full-length transcripts. De novo transcriptome assembly does not rely upon alignment to the reference genome and therefore can potentially identify transcripts that arise from repetitive sequences. The RNAseq reads from the wild-type, tissue culture wild-type, and *drm1ab* samples were aligned to de novo assembled transcripts that were greater than 1kb in length to identify 103 up-regulated transcripts in *drm1ab* (minimum 2-fold change and padj<0.05). These transcripts include both gene and TE sequences. To focus on putative TEs, we removed any of the 103 *drm1ab* up-regulated transcripts with >50% overlap of an annotated gene based on alignment of the transcripts to the genome. There were 29 up-regulated transcripts that do not align to annotated genes. The analysis of conserved domains within the putative ORFs of these transcripts identified five of these transcripts that contain TE-associated domains and three additional transcripts that overlap structurally annotated

TEs. We refer to these eight transcripts as DRM Silenced TEs (DSTs) (Table S2). The eight DSTs were aligned to the genome to identify the best matching genomic sequence and we assessed the presence of LTR/TIRs and target site duplications. Only DST1 contained intact structural features that would be necessary for active transposition and we focused on further characterization of this element and related family members.

The DST1 de novo assembled transcript is 8.7kb in length and is >8 fold up-regulated in drm1ab (Table S2). Alignment of the de novo assembled DST1 transcript to the genome assembly of ME034V revealed 15 highly similar sequences (Table S3). Four of the DST1 elements are identified as TEs in the structural annotation and another 10 are at least partially annotated as TEs in the homology-based annotation. These annotated TEs were not identified as differentially expressed based on alignments of RNAseq reads to the genome, likely due to the repetitive nature of the family and limited uniquely mapping reads. The internal (non-LTR) sequences typically have 96-98% identity between members of this family with one element (DST1-12) having lower similarity, indicating that this may be the oldest member of the family. A phylogeny of the DST1 family based on the internal alignable sequences does not reveal any subgroups with very close relationships that would indicate on-going movement of this set of elements (Figure 6B). The majority of these (14/15) have intact LTR sequences on the 5' and 3' end with 90-96% identity of the two LTRs suggesting that these elements do not represent particularly recent transposition events. However, there is substantial variability in the LTR sequences between different elements of the family. Only DST1-2/DST1-11 and DST1-4/DST1-6 share similar LTRs. A comparison among all other elements revealed that there is conservation (90-95%) identity at the first 250-400bp and in the last

1.2kb of the ~2kb LTRs. The middle region of 300-800bp is highly variable and can not be aligned among family members. The DST1-1 element only contains homology for the first 250bp region. DST1 is an intriguing LTR family with highly conserved internal sequences but a variable region within the LTR reminiscent of observations for the Tnt1 TE family in tobacco (Casacuberta et al., 1997).

Alignment of RNAseq reads to the DST1 transcript reveals an ~8-fold increase in expression in the *drm1ab* individuals. However, many of the aligned reads contain SNPs relative to the DST1 assembled transcript, potentially reflecting expression of multiple family members with slight sequence variation. The transcripts that arise from DST1 elements were assessed to determine if the activation in *drm1ab* occurs at a single element or reflects coordinate activation of several members of the family. Based on SNPs within the DST1 sequence there is evidence for the expression of 11 members of the DST1 family and five of the DST1 elements are only detected in the *drm1ab* mutant (Figure 6B). The DST1 family members are often highly methylated with high CHH levels (>75%) at, or near, the LTRs. Six of the DST1 elements have loss of CHH methylation at or within 1 kb of the annotated element in *drm1ab* (Figure 6C) and several of these are the elements that exhibit strong transcriptional activation in the mutant.

Discussion

The DRM methyltransferases are critical for RdDM in plants. However, there are variable consequences for the loss of functional RdDM among different plant species. While there are some gene and TE expression differences in Arabidopsis thaliana, there are limited phenotypic differences (Cao and Jacobsen, 2002a; Chan et al., 2006). In

contrast, rice and maize mutants lacking functional RdDM exhibit developmental abnormalities (Moritoh et al., 2012; Sidorenko et al., 2009; Alleman et al., 2006). In this study we isolated S. viridis plants with loss-of-function mutations in both catalytically active DRM genes. While there are some phenotypic differences such as reduced stature, the plants are fully viable with normal inflorescences. This suggests that a functional RdDM pathway is dispensable under standard growth conditions. However, it is worth noting that we have only maintained the *drm1ab* mutants in the homozygous mutant condition for several generations. One important function of RdDM might be to ensure the faithful maintenance and inheritance of DNA methylation patterns. Loss of fidelity in maintaining heterochromatin may have growing phenotypic consequences after many generations in the absence of RdDM activity.

A detailed analysis of the seedling leaf methylome in *drm1ab* plants revealed significant changes in CHH methylation, as expected. In particular, we find loss of CHH methylation at the vast majority of genomic loci with high (>20%) CHH methylation in wild-type plants. In contrast, many genomic regions that had lower, but detectable, levels of CHH methylation (5-10%) are not changed in a *drm1ab* mutant. It seems that the RdDM pathway is responsible for high levels of CHH methylation that is often found at the edges of TEs, especially near genes. In contrast, the lower levels of CHH methylation are often found within larger TEs and this methylation is likely the result of CMT2 or other chromomethylases similar to observations in other plant species (Zemach et al., 2013). We noted that the distribution of CHH methylation in the *drm1ab* mutant is quite similar to the distribution of CHG methylation, potentially supporting a role of CMT genes in contributing to the remaining CHH methylation.

The analysis of CG and CHG methylation patterns in *drm1ab* plants revealed some unexpected findings. First, we found that CHG methylation was rarely maintained at loci that had lost CHH methylation. The regions with high CHH methylation typically had CHG methylation in wild-type plants and both CHH and CHG methylation were lost in the *drm1ab* mutant. This suggests widespread failure to maintain CHG methylation at RdDM targets in the absence of functional DRM. Second, we found numerous examples of CG and/or CHG methylation loss in regions that were not located at or near loci with high CHH methylation. It is not clear why functional DRM is necessary for maintenance of CG/CHG methylation at these loci but similar results have been reported for the rice OsDRM2 mutant (Hu et al. 2021). One possible explanation is that these regions have elevated CHH at other developmental stages and the loss of active RdDM at that stage results in the loss of CG/CHG methylation which is observed in leaf tissue. It is also possible that these sites are targets of active DNA demethylation and require RdDM for maintenance.

There are relatively few changes in gene expression in *drm1ab* plants. We hypothesized that the subset of genes with high CHH methylation over the TSS would be up-regulated in *drm1ab*. However, we found that many of these genes are already expressed in wild-type plants that contain methylated TSS regions and very few of these genes have altered expression when the methylation is lost. This observation suggests a complex relationship between RdDM activity and gene expression. It is possible that many of these genes are redundantly silenced by RdDM activity as well as CMT2/3 and MET1 maintenance activities. It is possible that loss of CHH methylation may destabilize silencing, but does not lead to activation.

A small set of TEs that are silenced in wild-type plants exhibit transcriptional activity in *drm1ab* plants. These TEs appear to require RdDM for full silencing. In particular, we identified a novel TE family that exhibits coordinated activation of multiple loci in *drm1ab* mutants. This family will be of particular interest in future studies of potentially active TEs in the Setaria genome.

Methods

Plant material and growth conditions

Setaria viridis variety ME034V was used in this study. The tissue culture wild-type control plants and the *drm1ab* plants were derived from the same T0 transgenic plant as described in the previous study (Weiss et al., 2020). In the process of mutant screening, seed dormancy was broken by incubating freshly harvested seeds at 29°C for 24 h in a 1.4 mM gibberellic acid and 30 mM potassium nitrate solution (Sebastian et al., 2014). Seeds were then sterilized with 50% bleach for 10 min, rinsed five times with water, and then planted on germination media [0.5X MS, 0.5% sucrose, 0.4% Phytagel (Sigma, St Louis, USA), pH 5.7]. 6 days after germination, seedlings were transplanted to soil and grown under a 16:8 h light/dark photocycle at 26°C/22°C (day/night) and 30% relative humidity, according to a modified protocol (Huang et al., 2019).

Guide RNA design and vector construction

The genomic sequences of DRM1a and DRM1b were obtained prior to the publication of the ME034V genome and were identified by BLAST searching the S. viridis A10.1 reference (Mamidi et al., 2020) using the phytozome database

(https://phytozome.jgi.doe.gov). CRISPR gRNAs were designed to target the conserved domains in each gene using CRISPR (Haeussler et al., 2016). Conserved domains were identified by aligning the coding sequences from *S.viridis* with orthologs from brachypodium, maize and Arabidopsis. Construction of the T-DNA construct, pTW045, was described previously using the Golden Gate assembly method (Čermák et al., 2017; Weiss et al., 2020)

T-DNA transformation and tissue culture

Agrobacterium tumefaciens-mediated transformation of S. viridis ME034V was performed as described previously with a few modifications (Van Eck et al., 2017; Weiss et al., 2020). Callus initiation was first performed by removing the seed coats and sterilizing seeds with a 10% bleach plus 0.1% Tween solution for 5–10 minutes under gentle agitation. Seeds were then placed on callus induction media with the embryos facing upward at 24°C in the light for 1 week and then moved to dark for callus initiation. Embryogenic calli were collected after 4–7 weeks and inoculated with the AGL1 strain harboring the T-DNA construct pTW045 (Weiss et al., 2020). Inoculated calli were placed on a co-culture medium and incubated in the dark at 20°C for 5–7 days. Transformed calli were transferred to selection medium with 50 mg L-1 hygromycin for 4 weeks at 24°C. Selected calli were subcultured on plant regeneration media with 20 mg L–1 hygromycin with 16-h light to allow the growth of transformed shoots. Elongated shoots were transferred to rooting medium with 20 mg L-1 hygromycin. Shoots were transplanted to soil and grown to maturity.

Genotyping and *drm1ab* identification

The *drm1ab* and wild-type tissue culture control plants were identified using genomic PCR with restriction enzyme digestion (CAPS assay) followed by Sanger sequencing. PCR was performed with GoTaq Green Master Mix (Promega Corp., Madison, WI, USA) in accordance with the manufacturer's instructions, with an annealing temperature of 58°C and an extension time of 1 min. Amplicons were then subjected to restriction enzyme digestion using an enzyme that overlaps with the CRISPR-Cas9 cleavage site. PCR amplicons made with the corresponding primers were subjected to Sanger sequencing. T-DNA transgene detection was conducted using two methods: genomic PCR amplification of the hygromycin gene that is close to the T-DNA left border and a luciferase assay to detect the expression of the luciferase reporter gene that is next to the T-DNA right border. The luciferase assay procedure was conducted using the Bio-GloTM Luciferase Assay System (Promega Corp.) in accordance with the manufacturer's instructions. All the primer sequences used in the present study can be found in Table S4.

Methylome profiling

DNA was extracted from tissue collected from the 3rd and 4th leaf of 2.5 week old S. viridis plants grown in a growth chamber with 31°C/21°C 12-hour/12-hour day/night conditions. For each sample, tissue from 3-4 plants were pooled prior to CTAB DNA extraction. In total, seven samples from pooled tissue were converted and analyzed: one from wild type plants, three biological replicates of unedited plants regenerated from tissue culture, and three biological replicates of *drm1ab* edited plants. The samples were

converted for sequencing with the NEBNext enzymatic methyl-seq kit (NEB) and sequenced at the University of Minnesota Genomics Center. All samples were multiplexed in a full Novaseq S1 lane with 150bp paired-end sequencing. Sequencing reads were trimmed with Trim galore! version 0.4.3, powered by cutadapt v1.8.1 (Martin, 2011) and fastqc v0.11.5 and aligned to the ME034V reference genome (Thielen et al., 2020) with bsmap v2.74 using the following parameters: -v 5 -r 0 -p 8 -q 20 (Xi and Li, 2009). Sample conversion rates were calculated for each sample based on the ratio of predicted unconverted to converted cytosines in the un-methylated chloroplast genome. Methyl-seq alignment metrics for each sample are summarized in Table S5.

The bsmap alignments for the three replicates from the tissue culture and *drm1ab* were each merged using samtools merge (Danecek et al. 2021). The genome was divided into adjoining 100-bp tiles and, using the merged data, each tile was classified as one of: "missing data" (including "no data" and "no sites"), "CHH > 15%", "CG/CHG", "CG-only", "unmethylated", or "intermediate" following the classifiers and hierarchy outlined in (Crisp et al., 2020). The ratio of tiles in each category is displayed in Figure S2. Tissue culture and *drm1ab* replicates showed little variance with R-squared of CG methylation values greater than 0.975 within each treatment. Significantly hypomethylated and hypermethylated 100-bp tiles were identified from the merged replicates using the DSS library (Park and Wu 2016) in R using a false positive adjusted p-value of 0.05 and a 2-fold change as cut-offs (Figure S4).

We further classified tiles as DRM-dependent, DRM-intermediate, or DRM-independent. Tiles classified as "missing data" in either the tissue culture or

drm1ab mutant samples were omitted from this analysis. For all remaining tiles, we first asked whether the tile was methylated in one or more contexts (>40% for mCG or mCHG, >20% for mCHH) in the tissue-culture control plant. If yes, the percentage of methylation loss in *drm1ab* was determined ((mC tissue control - mC *drm1ab*)/mC tissue control)*100. Loss of >80% methylation was classified as DRM-dependent, loss of 20-80% was classified as DRM-intermediate, and a loss of <20% as DRM-independent (Table 1).

In order to determine the relationship of CHH methylation with gene expression, all genes were classified as either CHH-TSS genes and/or CHH-Island genes. CHH-TSS genes have a tile with >20% mCHH in the tissue culture samples that is overlapping, or within one tile, of the annotated transcriptional start site. CHH-Island genes have at least one tile with >20% mCHH in the tissue culture sample that is between 100 and 1000-bp upstream of the TSS. It is possible for a gene to be both a CHH-TSS gene and a CHH-Island gene. In addition to association with gene expression, we were interested in whether DRM-dependent CG and/or CHG methylated tiles were often found at or near CHH DRM-dependent and CHH DRM-intermediate tiles. For this, we classified each CG, CHG, and CG/CHG DRM-dependent hypomethylated tile as overlapping, within 300-bp (proximal), or greater than 300-bp (distal) from a CHH DRM-dependent/intermediate tile (Fig 4).

Transcriptome profiling

Prior to DNA extraction, a portion of the ground tissue used for methylome profiling was saved for RNA extraction with the QIAGEN RNeasy mini kit (QIAGEN).

For RNAseq, two additional wild-type biological replicates were included. RNA was submitted to the UMGC facility for 150 bp cDNA paired-end library preparation and run on the Illumina NovaSeq 6000. Sequencing reads were trimmed as described in the above methylation profiling section. The ME034V genome was indexed using the –runMode geneomeGenerate command of STAR 2.7.1 (Dobin et al., 2013) using an annotation file that included the primary transcript for each gene as well as all structural TEs. The trimmed reads were aligned to the indexed genome with the –quantMode GeneCounts feature of STAR 2.7.1 (Dobin et al., 2013). RNAseq metrics for each sample are summarized in Table S6.

Read counts were imported into R v4.1.1 (R Core Team, 2021). Normalization (median of ratios) and differential expression was determined using DEseq2 (Love et al., 2014). A gene or structural TE was determined to be differentially expressed if the absolute value of the log2 fold change was greater than 1 and the adjusted P-value was less than 0.05. EnhancedVolcano (https://github.com/kevinblighe/EnhancedVolcano) was used to visualize differentially expressed genes and TEs.

Additional RNAseq datasets were used to compare our observed expression ratios of Drm1a to Drm1b and Drm3 with previously published data (Fig S1). The data for S. viridis cultivar A10 were downloaded from the Phytozomev13 portal (Goodstein et al., 2012). The data for ME034v are from (Thielen et al., 2020).

De novo Transcript assembly and analysis

A de novo transcriptome assembly was generated using pooled trimmed RNAseq reads from all nine samples with Trinity version 2.10.0 (Haas et al., 2013). The

minimum contig length was set to 200 bp. The de novo transcripts were indexed with the 'gmap_build' command of gmap version 2015-09-26 (Wu and Watanabe, 2005). Default 'gmap' parameters were used to map the de novo transcripts back to the ME034V reference genome.

Next, the RNAseq reads from each sample were mapped to the transcriptome assembly with Salmon version 1.2.1 (Patro et al., 2017) using default parameters in order to determine transcripts per million (TPM) in individual biological samples. Differential transcript expression was determined as previously described using the Salmon TPM data as input.

To find transcripts from the transcriptome assembly that may represent TEs that are upregulated in the *drm1ab* mutant, we first filtered our DE transcript list to only include transcripts greater than 1-kb, reasoning that any shorter transcripts would not encode functional TEs. Next, we set a DE threshold of 2 fold upregulation in the mutant with an adjusted P-value < 0.05. Transcripts with at least 50% of their length annotated as genes based on overlap of coordinates from the gmap alignment and the current ME034V gene annotation were removed. The remaining transcripts were subjected to conserved domain BLAST (Lu et al., 2020). Transcripts without TE-associated domains (e-value < 0.01) were omitted from further analyses. Finally, the transcript plus the 3-kb flanking sequence on either side was submitted for a self vs. self BLAST to determine if putative LTR or TIR sequences could be observed.

Of the five transcripts that had evidence for upregulation, one had detectable LTR sequences (now referred to as DST1). DST1 family members were identified in the ME034V genome using BLAST (Altschul et al., 1990) with parameters 'blastn

-perc_identity 75 -qcov_hsp_perc 75'. A total of 15 similar sequences were identified and classified as a TE family using the 80-80-80 rule (Seberg and Petersen, 2009). To determine the phylogenetic relationship between the DST1 copies, the internal element sequence of the DST1 elements was first trimmed with trimal (Capella-Gutiérrez et al., 2009) with parameter '-automated1'. Trimmed internal sequences were aligned with MUSCLE (Edgar, 2004) using default settings with manual inspection and a tree was generated with RaxML (Stamatakis, 2014) with settings '-m GTRGAMMA -p 12345 -x 12345 -# autoMRE'. To examine expression of individual DST1 elements, diagnostic SNPs were identified for each member of the DST1 family from the multiple sequence alignment and expression per individual element was quantified by requiring a minimum of four RNAseq reads supported by a diagnostic SNP for each sample.

Annotation of TEs

The ME034V repetitive elements were previously identified using a homology based repeat masking approach (Thielen et al., 2020). We were interested in monitoring potentially active TEs that have intact structural elements, and therefore performed a TE annotation using the EDTA software (Ou et al., 2019). This approach was implemented using EDTA v1.9.6 with "--species others --sensitive 1 --anno 1" and all remaining parameters as default. This produced an initial structural annotation of 9,459 intact elements. Simple repeats ('target_site_duplication, 'repeat_region', 'long_terminal_repeat') were removed, resulting in a filtered structural annotation of 6,369 intact elements accounting for 16.9 Mb (referred to as "structural annotation"). These structural elements were then used for a homology search, which identified an

additional 188,707 elements (115 Mb) that have similarity to structural TEs, but lack intact structural features.

Data summary statement: The sequences used to profile DNA methylation (EM-Seq) and gene expression (RNA-seq) are available at the National Center for Biotechnology Information (NCBI) BioProject PRJNA787965. Transposable element annotations used in this manuscript are available at https://hdl.handle.net/11299/225624.

Figures

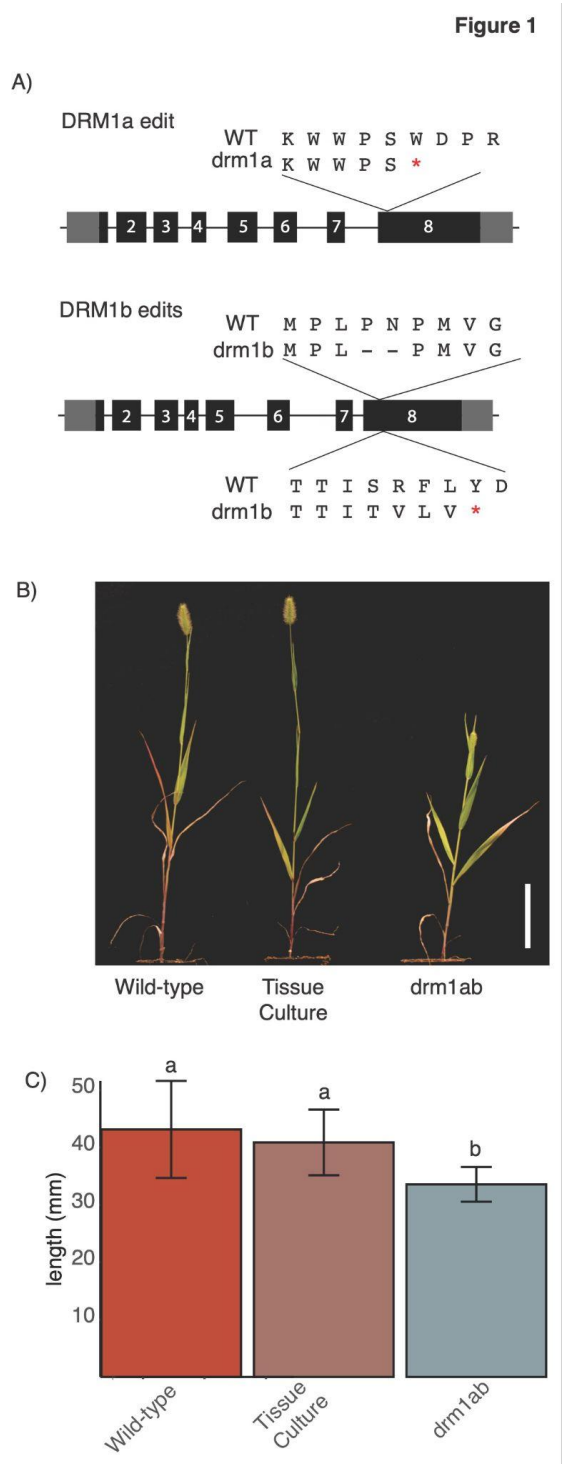


Figure 1
A)
DRM1a edit
WT K W W P S W D P R
drm1a K W W P S *
2 3 4 5 6 7 8
DRM1b edits
WT M P L P N P M V G
drm1b M P L - - P M V G
2 3 4 5 6 7 8
WT T T I S R F L Y D
drm1b T T I T V L V *
B)
Wild-type
Tissue Culture
drm1ab
C)
length (mm)
50
40
30
20
10
a
a
b
Wild-type
Tissue Culture
drm1ab

Figure 1. Isolation of loss-of-function alleles for DRM genes in *Setaria viridis*. (A) Sequencing of transgene-free plants derived from transgenic parents expressing gRNAs targeting the Drm1a and Drm1b gene identified individuals that are homozygous for mutations at both target genes. The schematic indicates the position and sequence change at each locus. (B) Images showing wild-type ME034V and *drm1ab* double mutant plants grown in growth chamber conditions. The white scale bar represents 9 cm (C) Leaf length is reduced in greenhouse grown *drm1ab* mutant plants relative to wild-type controls or the progeny of plants derived from tissue-culture. Letters indicate significant differences ($p<.05$).

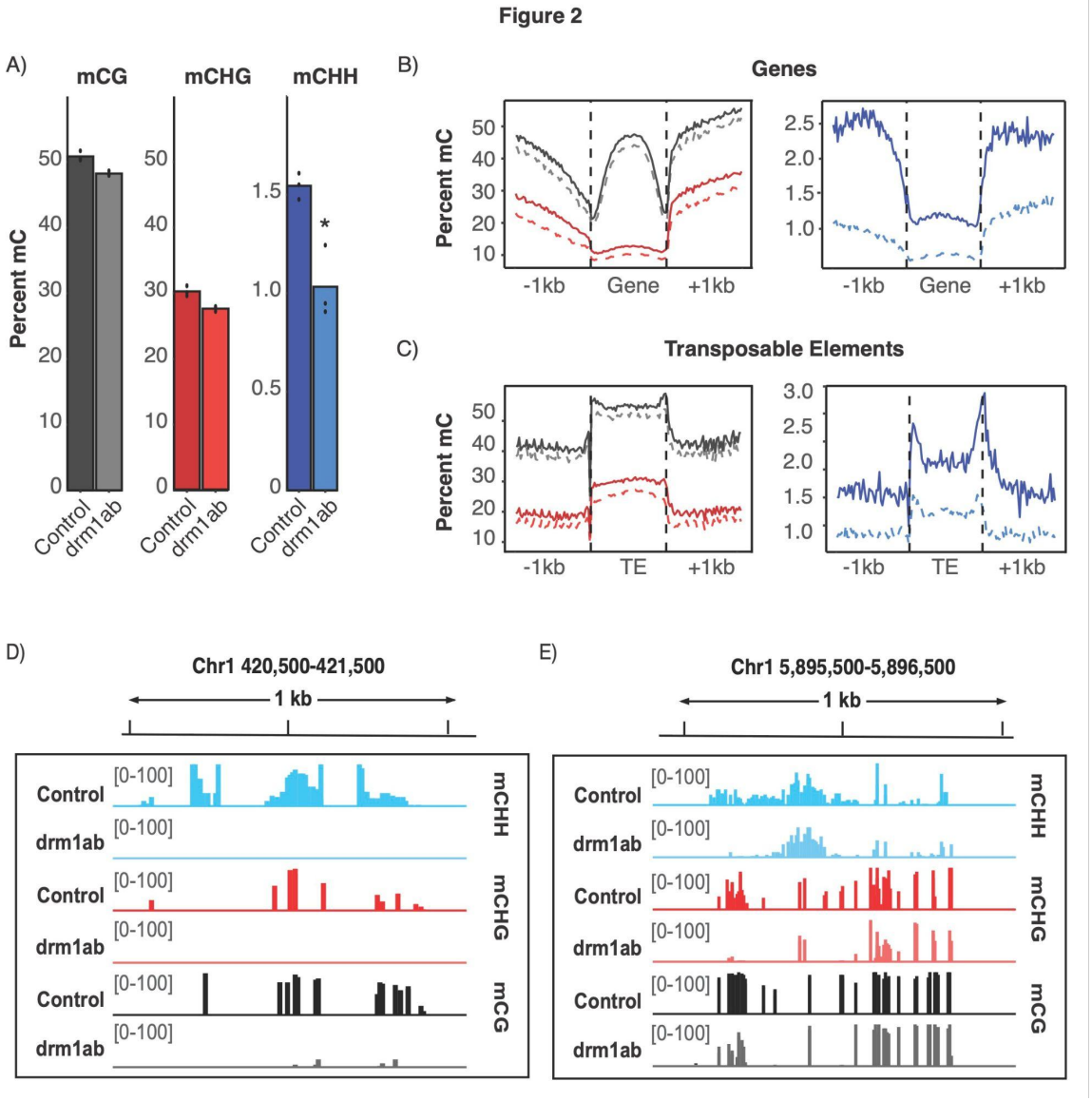


Figure 2. DNA methylation changes in *drm1ab* plants. (A) Genome-wide mean mCG, mCHG and mCHH levels were assessed in three biological replicates of non-edited tissue-culture control plants and *drm1ab* plants. Asterisk indicates significantly lower levels of mCHH methylation in *drm1ab*. (B) Metaplots of mCG, mCHG or mCHH levels in genic regions. Solid lines show the profile in tissue-culture control plants, dashed lines show the levels in *drm1ab*. A different scale is used for mCHH due to overall lower methylation levels in this context. Genes are all oriented 5' to 3' and the dashed lines

indicate the genic region normalized to the same length. The region to the left or right of the dashed vertical lines include 1kb of upstream or downstream sequences. (C) Similar metaprofiles of mC levels within and surrounding structurally annotated transposable elements. Genome viewer snapshots showing an example of (D) DRM-dependent mC loss and (E) DRM-independent maintenance of mC.

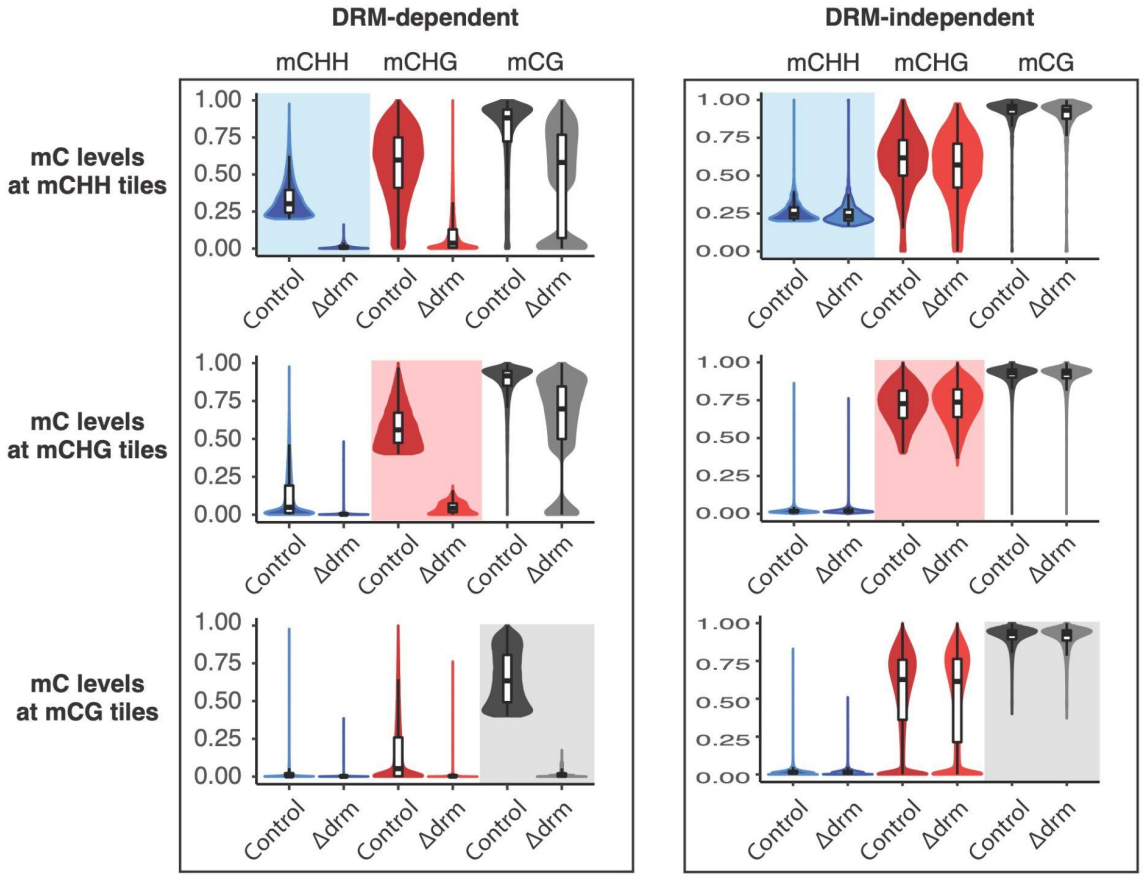


Figure 3. Comparisons of context-specific mC levels at DRM-dependent and DRM-independent loci. 100 bp tiles classified as methylated (>= 40% for CG or CHG, >= 20% for CHH) were examined to see if methylation was lost in *drm1ab* plants. DRM-dependent tiles lost >80% mC while DRM-independent tiles lost <20% mC in *drm1ab*. Each violin plot includes context-specific DNA methylation levels in both mutant and control plants for the set of DRM-dependent or -independent tiles. Shaded boxes highlight the mC context used to select the subset of tiles included in each plot. Details of tile classification can be found in table 1.

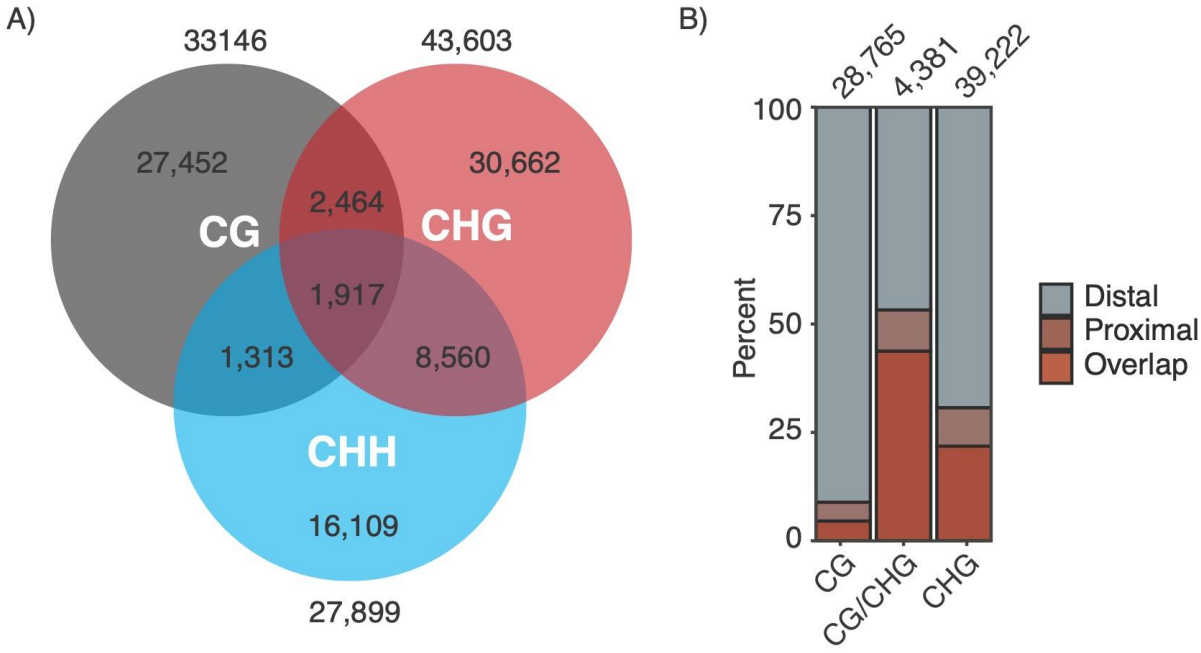


Figure 4. Relationship of DRM-dependent mCG and mCHG losses to mCHH hypomethylated tiles. (A) Venn diagram of the overlap of tiles with mCG and/or mCHG DRM-dependent methylation with tiles with mCHH DRM-dependent or -intermediate loss. (B) The proportion of mCG, mCHG, or mCG/CHG DRM-dependent tiles that overlap, are proximal to (within 300 bp), or distal to (greater than 300 bp) mCHH hypomethylated tiles. The number of hypomethylated tiles in each mC context can be found in Figure S4.

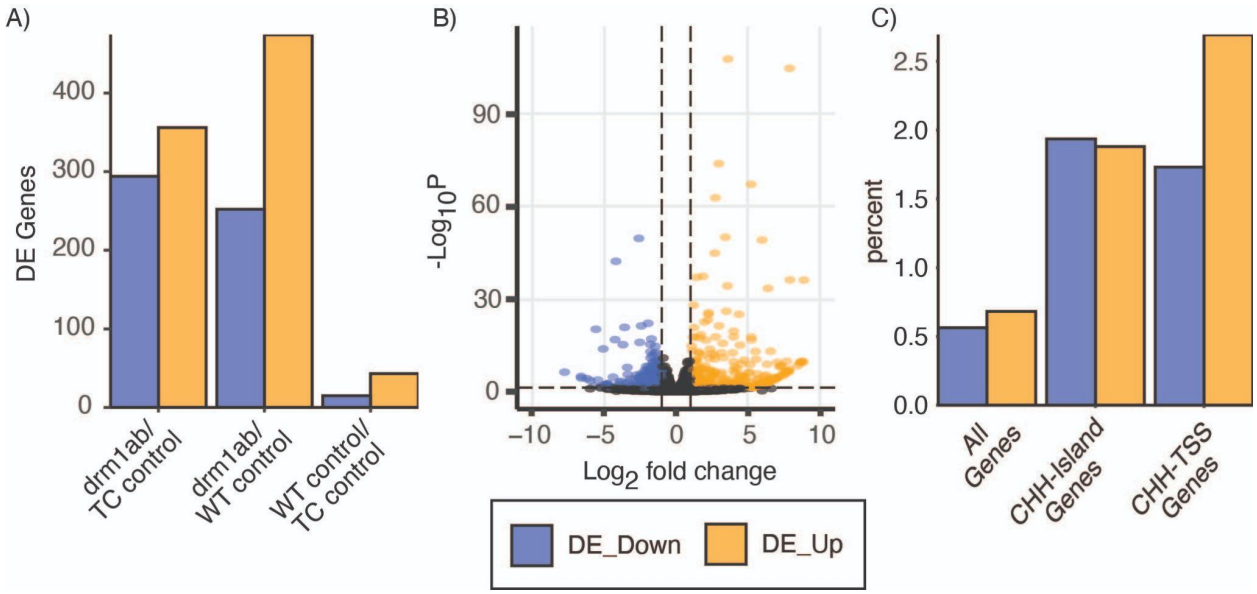


Figure 5. Transcriptome changes in *drm1ab* plants. (A) The number of genes with significant differences in expression (padj<0.05; >2x fold-change) was determined for all contrasts. (B) A volcano plot showing magnitude and padj for differentially expressed genes in the *drm1ab*:tissue culture comparison. Significant differences are indicated using blue and orange data points. (C) The proportion of genes that are up- (orange) or down-regulated (blue) is shown for all genes, the sub-set of 5,424 genes that contain a tile of >20% CHH within 1kb of the TSS (CHH-Island genes), or the 1,043 gene with >20% CHH methylation in the 100bp tile directly over the TSS or the two adjacent tiles (CHH-TSS genes).

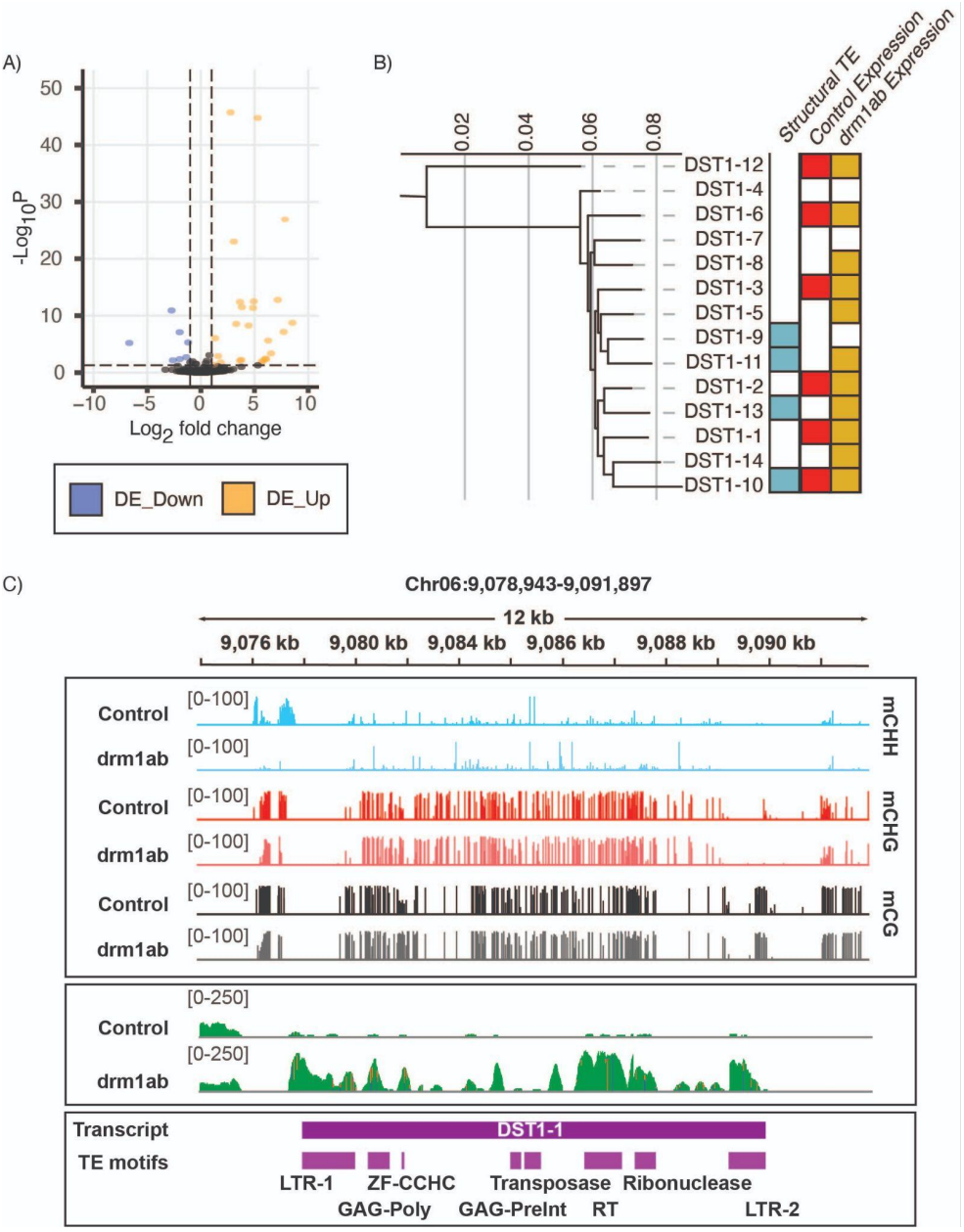


Figure 6. Expression of structurally intact TEs and analysis of DST-1-like TEs. (A) A volcano plot showing magnitude and padj for differentially expressed structurally annotated TEs in the tissue culture control:*drm1ab* comparison. Significant differences are indicated using blue and orange data points. (B) A maximum likelihood relatedness tree generated from an alignment of the putative transcripts of the 14 DST1-like TEs. DSTs that: overlap a structurally annotated TE, have expression in control, and/or

drm1ab are indicated. (C) Genome browser snapshot of the DST1-1 region showing mC levels in tissue-culture and *drm1ab* plants, (D) RNAseq coverage from tissue-culture control and *drm1ab* plants, and (E) the location of the predicted DST1-1 transcript and TE-associated domains identified by CD-BLAST.

Supplemental Figures

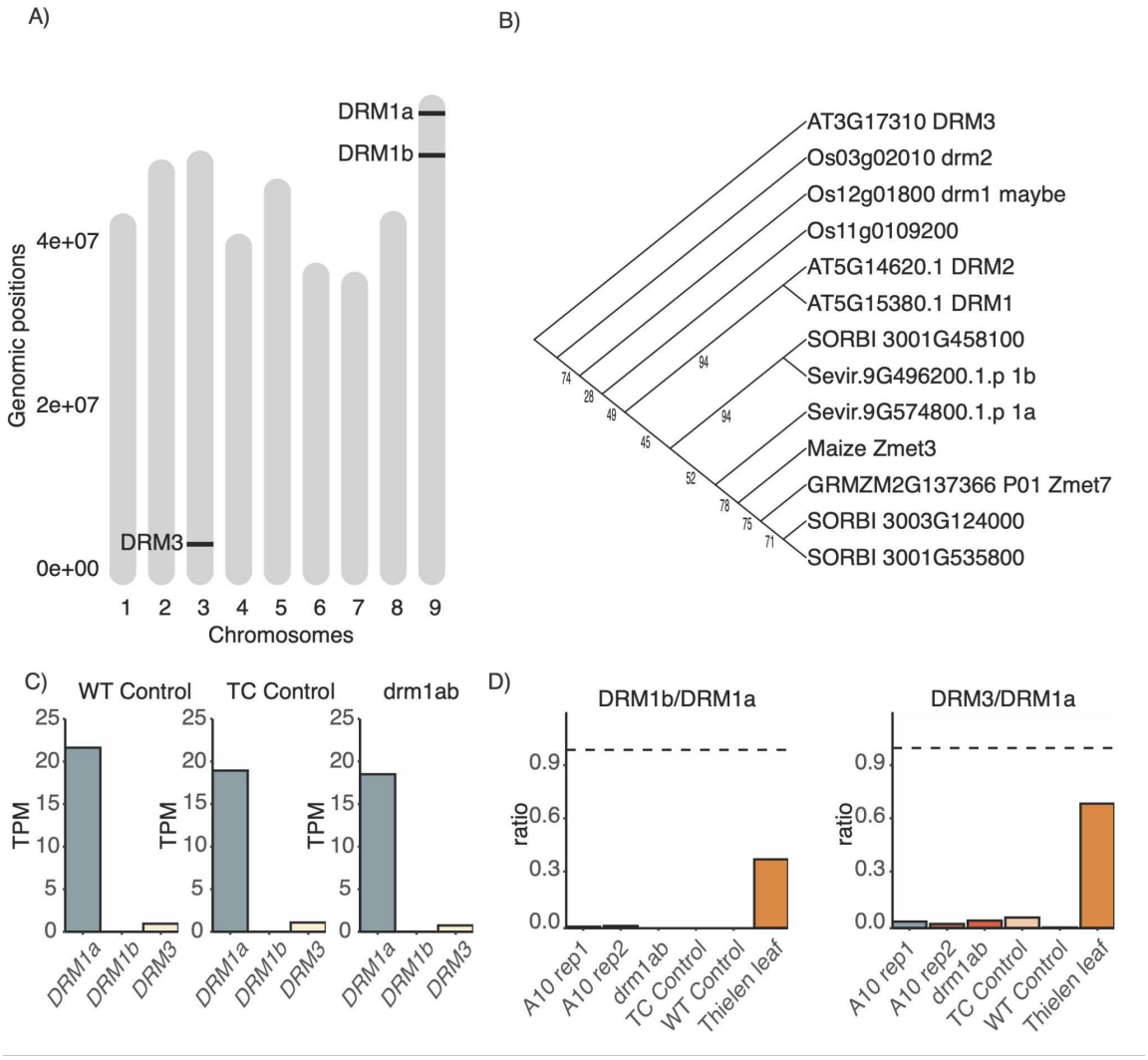


Figure S1. DRM genes in *Setaria viridis*. (A) The physical location of DRM genes in the *Setaria viridis* ME034V genome is shown. DRM1a and DRM1b encode putatively functional DRM genes and are located in linked positions on chromosome 7. A DRM3-like gene on chromosome 3 has mutations in the catalytic domain predicted to disrupt function. (B) The relative expression (tags per million - TPM) of the three DRM genes was determined in seedling leaf tissue for wild-type plants, plants derived from

tissue culture and the double mutant *drm1ab*. (C) The relative expression of DRM1b or DRM3 compared to DRM1a in leaf tissue from several experiments is shown. Dashed line indicates an expression level equal to that of DRM1a. In most experiments, the expression level of DRM1a is much higher than the other two genes. A10 Transcriptome data is from Thielen et al. 2020. ME034V data is from this study and(Bennetzen et al., 2012; Brutnell et al., 2010; Mamidi et al., 2020; Thielen et al., 2020).

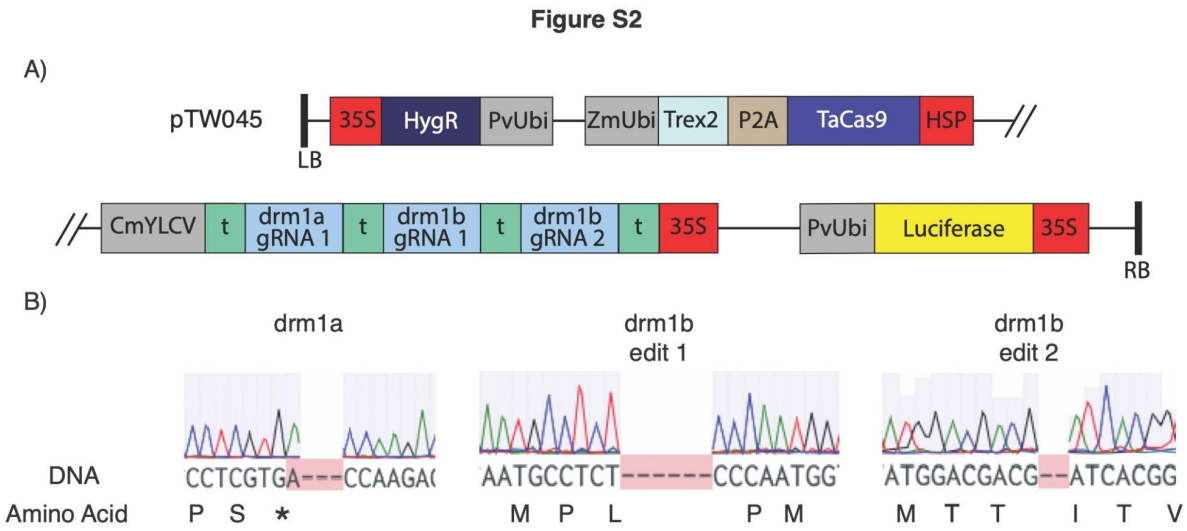


Figure S2. Genome editing reagents and genotyping data. (A) map of the T-DNA used in transformation to generate edited ME034V. The T-DNA encodes Hygromycin resistance, a wheat codon optimized Cas9 protein encoded as a polyprotein with Trex2 followed by a P2A protein cleavage site, an array of guide RNAs separated by tRNA cleavage sites, and a luciferase reporter. (B) Sanger sequencing results and amino acid translation of transgene negative edited *drm1ab* plants. Pink shading indicates genomic edits relative to the wild-type sequence.

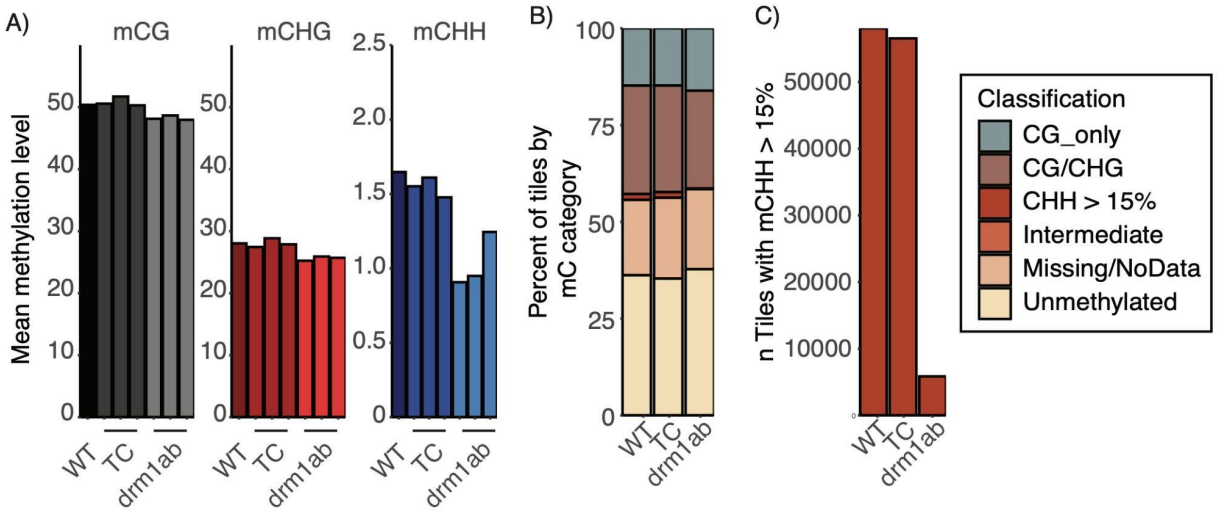


Figure S3. Comparisons of DNA methylation levels in *drm1ab* and unedited plants. (A) Genome-wide DNA methylation levels for a single replicate of wild-type as well as each of the three replicates of tissue-culture derived and *drm1ab* plants. (B) Each 100bp tile (n=3,968,817) of the ME034V genome was classified based on the relative levels of mCG, mCHG and mCHH as described in Crisp et al., 2020. The proportion of tiles classified as mCG only, mCG and mCHG, mCHH, intermediate, missing/no data, or unmethylated. (C) The number of tiles classified as mCHH in wild-type, tissue-culture and *drm1ab* plants.

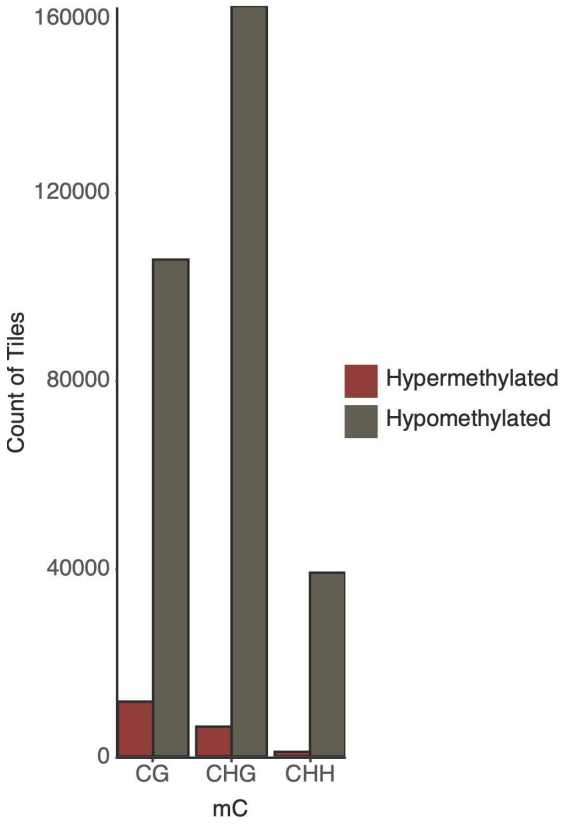


Figure S4. Methylation changes in the *drm1ab* edited line. Number of hyper- and hypo-methylated mCG, mCHG, and mCHH tiles when comparing tiles from Tissue Culture to *drm1ab*.

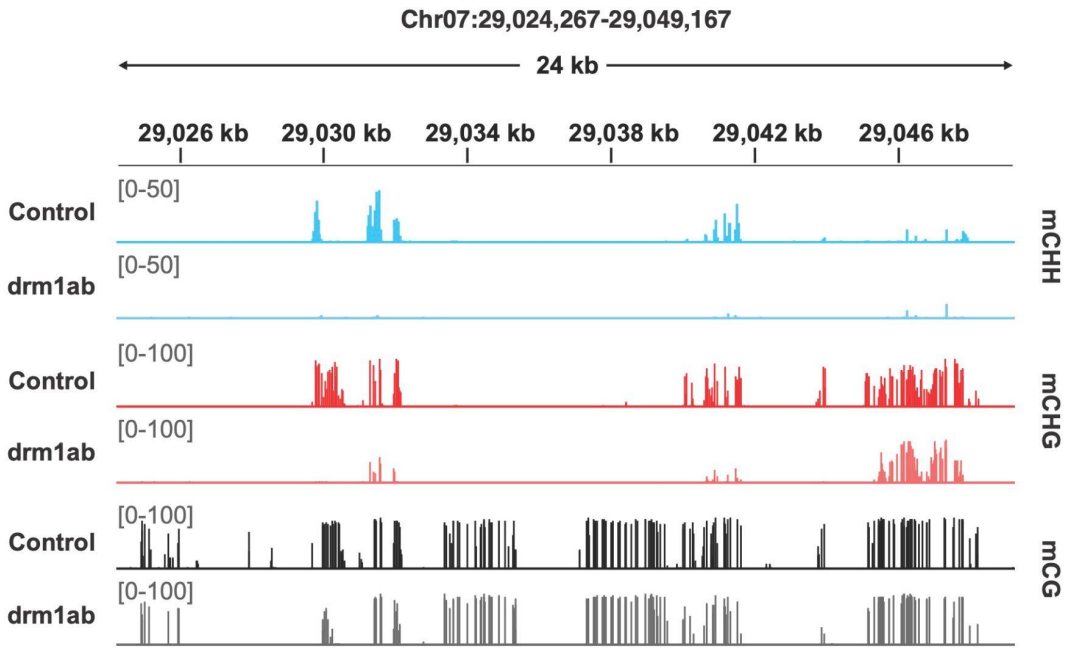


Figure S5. DRM-dependent loss of mCG and mCHG at regions demarcating edges between high and low mC levels. A snapshot of a genomic region with a near complete loss of mCHH accompanied by a reduction of mCHG and mCG at, and near, the locations of drm-dependent mCHH.

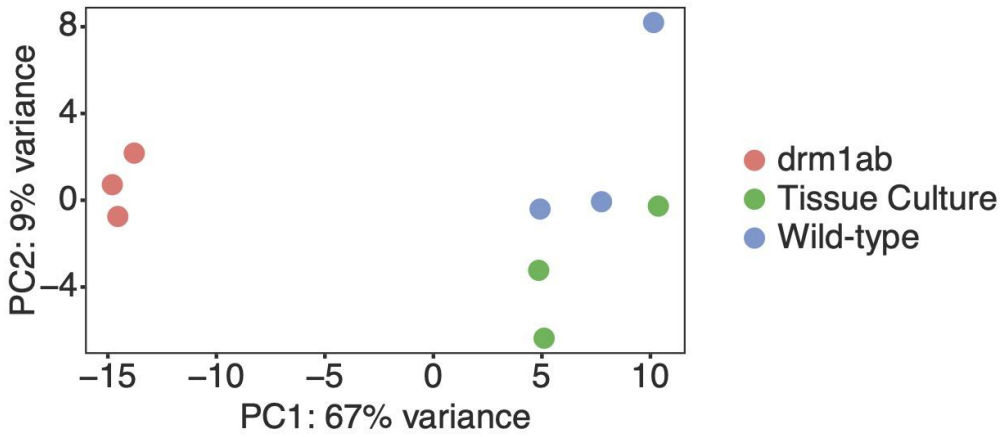


Figure S6. A principal component analysis was used to compare expression profiles of *drm1ab* mutant with wild-type and tissue culture controls. The *drm1ab* edited plants form a unique cluster distinct from unedited plants, regardless of whether the unedited plants have been through the tissue-culture process.

Tables

Table 1. Count of DRM-dependent, DRM-intermediate, and DRM-independent methylated tiles in all contexts.

	DRM_Dependent (≥80% loss in *drm1ab*)	**DRM_Intermediate (20-80% loss in *drm1ab*)**	**DRM_Independent (<20% loss in *drm1ab*)**	**Total**
CG ≥ 40%	33,146 (2%)	93,268 (6%)	1,520,749 (92%)	1,647,163
CHG ≥ 40%	43,603 (4%)	148,452 (13%)	986,556 (92%)	1,178,611
CHH ≥ 20%	22,883 (79%)	5,016 (17.25%)	1,174 (4%)	29,073

Supplemental Tables

Table S1. Significantly differentially expressed structurally annotated TEs

TEs upregulated in Δdrm1a/b mutant

	EDTA_ID	Classification	mean_drm1ab_expression	mean_TCcontrol_expression	log2FoldChange	padj
DNA Elements	DHHSvm1G00623	Helitron	73.5	0	-8.5	0
	DTCSvm1G00097	CACTA_TIR_transposon	136.8	1	-7.2	0
	DTCSvm1G00729	CACTA_TIR_transposon	9	0	-6.1	0.004
	DTMSvm1G00191	Mutator_TIR_transposon	10.5	0	-5.6	0.015
	DHHSvm1G00072	Helitron	271	8	-5.3	0
	DHHSvm1G00661	Helitron	97	3	-4.9	0
	DTMSvm1G00519	Mutator_TIR_transposon	69.5	5.7	-3.8	0
	DTHSvm1G00154	PIF_Harbinger_TIR_transposon	163.3	14.7	-3.8	0.006
	DTHSvm1G00696	PIF_Harbinger_TIR_transposon	143.3	10	-3.7	0
	DHHSvm1G00635	Helitron	17.3	1.7	-3.6	0.008
	DHHSvm1G00831	Helitron	28	7	-1.9	0.015
	DHHSvm1G00604	Helitron	75.3	26	-1.6	0.001
	DHHSvm1G00548	Helitron	386.5	145	-1.3	0
	DTTSvm1G01107	Tc1_Mariner_TIR_transposon	183.8	43.7	-1.2	0.042
RNA Elements	RLCSvm1G00134	Copia_LTR_retrotransposon	373	1.7	-7.8	0
	RLGSvm1G00023	Gypsy_LTR_retrotransposon	66	0.3	-7.7	0
	RLGSvm1G00212	Gypsy_LTR_retrotransposon	17.3	0	-6.5	0
	RLGSvm1G00175	Gypsy_LTR_retrotransposon	60.5	0.7	-6.3	0
	RLGSvm1G00211	Gypsy_LTR_retrotransposon	11.5	0	-6	0.003
	RLGSvm1G00088	Gypsy_LTR_retrotransposon	7.3	0	-5.8	0.008
	RLCSvm1G00292	Copia_LTR_retrotransposon	81.3	2.7	-4.9	0
	RLCSvm1G00027	Copia_LTR_retrotransposon	58.5	3	-4.5	0
	RLCSvm1G00007	Copia_LTR_retrotransposon	119	12.7	-3.3	0
	RLCSvm1G00110	Copia_LTR_retrotransposon	210.3	25.7	-3.1	0
	RLCSvm1G00129	Copia_LTR_retrotransposon	577.3	85.7	-2.8	0

TEs downregulated in Δdrm1a/b mutant

	EDTA_ID	Classification	meanMUT	meanTC	log2FoldChange	padj
DNA Elements	DTCSvm1G00004	CACTA_TIR_transposon	1	39.7	6.7	0
	DTCSvm1G00143	CACTA_TIR_transposon	28.3	157.7	2.7	0
	DTHSvm1G00421	PIF_Harbinger_TIR_transposon	5.5	17.7	2.6	0.007
	DTMSvm1G00401	Mutator_TIR_transposon	20.5	83	2	0
	DTTSvm1G00030	Tc1_Mariner_TIR_transposon	16.8	59	2	0.004
	DTASvm1G00155	hAT_TIR_transposon	24.8	74	1.3	0.002
	DTMSvm1G00009	Mutator_TIR_transposon	94.3	217.7	1.2	0

Table S2. TRINITY de novo transcripts that passed filters for: length, expression, and overlap % of annotated genes. Highlighted rows are categorized as putative DRM silenced TEs (DSTs)

Transcript	ALIAS	Location	drm1ab_TP	TC_control_TPM	log2foldchange	pval	% annotated as gene	TE-motifs?
TRINITY_DN7017_c0_g1_i3	DST-1	Chr06:9077678-9092414	5.78	0.71	3.998	0.031	0	RNAse, gag, zf, RVT,transposase
TRINITY_DN4252_c0_g4_i1	DST-2	Chr02:1303627-1312901	7.95	0.38	5.484	0.002	0	RT_nLTR
TRINITY_DN5401_c0_g1_i3	DST-3	Chr01:6898796-6909485	7.56	0.22	5.356	0.003	4.82	zfRVT and RT
TRINITY_DN399_c0_g1_i9	DST-4	Chr04:37658708-37667232	19.42	0.3	6.796	0	0	gag - duf4219
TRINITY_DN8847_c0_g2_i3	DST-5	Chr02:28431042-28439608	17.57	0	6.631	0	0	overlaps structural TE
TRINITY_DN6583_c0_g1_i5	DST-6	Chr07:17838190-17846698	25.76	0	7.213	0	0	overlaps structural TE
TRINITY_DN7280_c0_g1_i5	DST-7	Chr08:33636314-33643501	8.91	0	5.609	0.001	18.2	partial overlaps structural TE
TRINITY_DN5207_c0_g1_i2	DST-8	Chr09:32371451-32382719	33.26	0.09	7.58	0	14.84	Transposase associated domain
TRINITY_DN580_c0_g1_i5		Chr08:38157449-38166077	23.55	0.13	7.092	0	0	-
TRINITY_DN4338_c0_g1_i6		Chr07:14350129-14357768	21.46	0.14	6.942	0	0	-
TRINITY_DN3849_c0_g1_i9		Chr08:12976947-12984964	15.96	0	6.487	0	8.78	-
TRINITY_DN399_c1_g1_i1		Chr04:32219211-32232840	14.74	0.47	6.389	0	35.39	-
TRINITY_DN3729_c1_g2_i1		Chr02:9957396-9965277	6.93	0.52	5.212	0.005	15.63	-
TRINITY_DN6599_c0_g1_i1		Chr02:786809-794273	6.81	0.03	5.212	0.005	15.37	-
TRINITY_DN3951_c0_g1_i1		Chr03:21280390-21287987	6.24	0.12	5.053	0.008	32.12	-
TRINITY_DN3314_c0_g1_i1		Chr08:11894292-11901552	5.81	0	4.96	0.011	0	-
TRINITY_DN11532_c0_g1_i2		Chr01:38215958-38224190	5.65	0.04	4.868	0.014	0	-
TRINITY_DN4027_c0_g1_i3		Chr01:34532800-34540236	4.62	0	4.653	0.042	24.23	-
TRINITY_DN9435_c0_g1_i1		Chr08:11896514-11904163	22.67	1.03	4.577	0	0	-
TRINITY_DN8157_c0_g2_i1		Chr07:23717079-23724131	4.74	0.09	4.545	0.033	36.22	-
TRINITY_DN8802_c0_g1_i1		Chr02:10968535-10975712	3.89	0.09	4.426	0.044	0	-
TRINITY_DN5957_c0_g2_i1		Chr08:29928561-29936671	3.95	0.12	4.424	0.045	49.76	-
TRINITY_DN7972_c0_g1_i1		Chr07:32418214-32431890	19.82	3.46	2.796	0.001	31.96	-
TRINITY_DN3845_c0_g1_i6		Chr07:32402790-32411421	14.74	2.61	2.713	0.011	5.93	-
TRINITY_DN237_c0_g1_i4		Chr02:13317033-13345155	13.27	2.06	2.801	0.022	35.53	-

Table S3. DST1-like elements identified in the ME034V genome assembly

DST-1 Family Member	Location	mean_drm1ab_reads	mean_TCcontrol_reads	structural support (-f 0.50 -r)	homology support(-f -r 0.5)	long terminal repeats?
DST1-1	Chr06:9080942-9089897	2306	137			yes
DST1-2	Chr06:32667398-32678901	201	12			yes
DST1-3	Chr06:8878529-8890102	280	30			yes
DST1-4	Chr05:10673619-10684053	430	32		yes- fully covered, copia LTR 876	yes
DST1-5	Chr05:6287860-6299080	233	24		yes- fully covered, copia LTR 876	yes
DST1-6	Chr05:29836693-29848080	36	2		yes- fully covered, copia LTR 876	yes
DST1-7	Chr05:39705044-39716602	6	3		yes- fully covered, copia LTR 876	yes
DST1-8	Chr08:11896250-11907711	956	14			yes
DST1-9	Chr04:5338236-5349776	49	2	yes- fully covered, copia_same, 686		yes
DST1-10	Chr04:32019841-32031391	659	82	yes- fully covered, copia_same, 686		yes
DST1-11	Chr01:35139743-35151209	61	5	yes- fullly covered, copia LTR, 689		yes
DST1-12	Chr01:28176578-28188093	7	2		yes- fully covered, copia LTR	yes
DST1-13	Chr09:40596655-40608252	100	5	yes- fully covered LTR		yes
DST1-14	Chr09:34796753-34809229	2	2			no
DST1-15	Chr02:45062959-45074634	71	11			yes

Table S4. oligos used for genotyping *drm1ab* plants

Oligo Name	Sequence	Purpose
48_HYG	cgcgacgtctgtcgagaagtt	PCR for hygromycin selection marker
51_HYG	gtctgctgctccatacaagcca	
17_DRM	GCTTCAGAAGATGGTTGCTATGGC	PCR genotyping Drm1b
18_DRM	TTCCTGGCAGCCGCACATAA	
82_DRM	TCAAAGTTCTTCTGTGCCGCTG	PCR genotyping Drm1a
83_DRM	CGAGAGGTGATATGCAACAGTG	

Table S5. BIS metrics

	WT1	TC1	TC2	TC3	drm1	drm2	drm3
Total Reads (fastqc)	169244126	91006658	125504462	126033248	112192420	184647519	107299174
pairs (bsmapped)	100305510 (59%)	53881611 (59%)	75623856 (60%)	74060607 (59%)	65739850 (59%)	109240146 (59%)	62745255 (59%)
average cytosine coverage	26.87 fold	16.48 fold	21.47 fold	21.15 fold	19.05 fold	28.99 fold	18.36 fold
cytosine conversion rate	99.63%	99.77%	99.57%	99.80%	99.81%	99.78%	99.43%

Table S6. RNA-Seq Metric

	WT1	WT2	WT3		TC1	TC2	TC3		drm1	drm2	drm3
Number of input reads	23602239	27728125	27163038		23144069	24126567	23658897		22041686	26146694	24780877
Average input read length	292	291	293		292	292	292		291	291	292
UNIQUE READS:											
Uniquely mapped reads number	22165757	26196014	25692385		22226979	23333760	22724881		21241767	25125078	23733447
Uniquely mapped reads %	93.91%	94.47%	94.59%		96.04%	96.71%	96.05%		96.37%	96.09%	95.77%
Average mapped length	291.55	290.72	292.13		291.83	291.4	292.04		290.78	290.99	291.9
Number of splices: Total	24199864	29260719	28737439		24905805	26126849	24808358		23556558	27478151	26122622
Number of splices: Annotated (sjdb)	19293749	23369707	22998195		19903223	20836719	19762389		18772310	21937897	20794986
Number of splices: GT/AG	23854241	28836501	28327473		24549983	25745177	24456399		23216651	27088765	25753017
Number of splices: GC/AG	312274	386672	372638		323016	348157	319911		308236	353129	336379
Number of splices: AT/AC	5572	6132	6057		5973	6202	5524		5402	5786	5685
Number of splices: Non-canonical	27777	31414	31271		26833	27313	26524		26269	30471	27541
Mismatch rate per base, %	0.10%	0.10%	0.10%		0.10%	0.10%	0.10%		0.10%	0.10%	0.10%
Deletion rate per base	0.00%	0.00%	0.00%		0.00%	0.00%	0.00%		0.00%	0.00%	0.00%
Deletion average length	1.55	1.52	1.51		1.5	1.53	1.52		1.5	1.54	1.54
Insertion rate per base	0.00%	0.00%	0.00%		0.00%	0.00%	0.00%		0.00%	0.00%	0.00%
Insertion average length	1.32	1.31	1.28		1.32	1.31	1.32		1.3	1.31	1.32
MULTI-MAPPING READS:											
Number of reads mapped to multiple loci	380938	450384	441412		399187	396587	411388		355432	427833	408901
% of reads mapped to multiple loci	1.61%	1.62%	1.63%		1.72%	1.64%	1.74%		1.61%	1.64%	1.65%
Number of reads mapped to too many loci	115096	151997	129225		75327	87299	87340		76597	73382	136468
% of reads mapped to too many loci	0.49%	0.55%	0.48%		0.33%	0.36%	0.37%		0.35%	0.28%	0.55%
UNMAPPED READS:											
Number of reads unmapped: too many mismatches	0	0	0		0	0	0		0	0	0
% of reads unmapped: too many mismatches	0.00%	0.00%	0.00%		0.00%	0.00%	0.00%		0.00%	0.00%	0.00%
Number of reads unmapped: too short	915931	900906	882358		430959	295389	421648		356149	508596	481731
% of reads unmapped: too short	3.88%	3.25%	3.25%		1.86%	1.22%	1.78%		1.62%	1.95%	1.94%
Number of reads unmapped: other	24517	28824	17658		11617	13532	13640		11741	11805	20330
% of reads unmapped: other	0.10%	0.10%	0.07%		0.05%	0.06%	0.06%		0.05%	0.05%	0.08%
CHIMERIC READS:											
Number of chimeric reads	0	0	0		0	0	0		0	0	0
% of chimeric reads	0.00%	0.00%	0.00%		0.00%	0.00%	0.00%		0.00%	0.00%	0.00%

CHAPTER III: Context Summary

Summary

CRISPR-Cas9-mediated genome editing has been widely adopted for basic and applied biological research in eukaryotic systems. While many studies consider DNA sequences of CRISPR target sites as the primary determinant for CRISPR mutagenesis efficiency and mutation profiles, increasing evidence reveals the substantial role of chromatin context. Nonetheless, most prior studies are limited by the lack of sufficient epigenetic resources and/or by only transiently expressing CRISPR-Cas9 in a short time window. In this study, we leveraged the wealth of high-resolution epigenomic resources in Arabidopsis (*Arabidopsis thaliana*) to address the impact of chromatin features on CRISPR-Cas9 mutagenesis using stable transgenic plants. Our results indicated that DNA methylation and chromatin features could lead to substantial significant variations in mutagenesis efficiency by up to 250-fold. Low mutagenesis efficiencies were mostly associated with repressive heterochromatic features. This repressive effect appeared to persist through cell divisions but could be alleviated through substantial reduction of DNA methylation at CRISPR target sites. Moreover, specific chromatin features, such as H3K4me1, H3.3, and H3.1, appear to be associated with significant variation in CRISPR-Cas9 mutation profiles mediated by the non-homologous end joining repair pathway. Our findings provide strong evidence that specific chromatin features could have significant and lasting impacts on both CRISPR-Cas9 mutagenesis efficiency and DNA double strand break repair outcomes.

Chapter III has been adapted from my work in the following publication: “Epigenetic features drastically impact CRISPR-Cas9 efficacy in plants”

Trevor Weiss, Peter A Crisp, Krishan M Rai, Meredith Song, Nathan M Springer, Feng Zhang (2022). Plant Physiology. doi.org/10.1093/plphys/kiac285

During the course of this work many authors contributed. In particular, Trevor Weiss and Peter A Crisp identified the multicopy CRISPR sites; Trevor Weiss, Peter Crisp, and Krishan M Rai characterized the epigenetic context at the multicopy CRISPR sites; Meredith Song assisted with genomic DNA isolation and genotyping; Trevor Weiss, Feng Zhang, and Nathan Springer wrote the manuscript. I have removed contact information and acknowledgements as well as formatted figures and references to be consistent throughout my thesis.

CHAPTER III: Epigenetic features drastically impact CRISPR-Cas9 efficacy in plants

Introduction

CRISPR-Cas based genome editing technologies have greatly advanced both basic and applied biological research. Among them, CRISPR-Cas9, the class II bacterial CRISPR system, has been the first and the most widely adopted in eukaryotes (Jinek et al., 2012). The key steps in CRISPR-Cas9 mediated genome editing involve searching, binding and then cleaving a 20 nucleotide target site directed by a guide RNA (gRNA). The resulting cleavage product with double-strand breaks (DSBs) can then be repaired by either error-prone DNA repair pathways, such as classical non-homologous end-joining (NHEJ) or microhomology-mediated end-joining (MMEJ), or by a template DNA-dependent pathway, i.e. homology directed repair (HDR) (Chen et al., 2019). Thus, specific mutations, including insertions, deletions or point mutations, can be introduced by employing distinct DNA repair machineries (Chen et al., 2019).

Previous studies indicated that the CRISPR targeted sequence is the primary determinant for mutagenesis efficiency and mutation profile (Allen et al., 2018; Lazzarotto et al., 2020). A number of tools have been developed to predict the efficiency and mutation outcomes solely based on CRISPR gRNA sequences (Allen et al., 2018; Concordet & Haeussler, 2018; Xiang et al., 2021). However, the predictability of these tools, primarily based on the large dataset from human cells, often varies and appears to translate poorly to other species, such as plants (Naim et al., 2020). This observation

suggested that non-sequence features could influence CRISPR-Cas9 mediated mutagenesis.

Increasing evidence revealed negative correlations between CRISPR-Cas9 mutagenesis rates and heterochromatic signatures or low chromatin accessibility in multiple systems, such as yeast, zebrafish, mouse, human and rice (Daer et al., 2017; G. Liu et al., 2019; Uusi-Mäkelä et al., 2018; X. Wu et al., 2014; Yarrington et al., 2018). However, most of these studies were conducted at various genomic locations, making it difficult to separate the effects of chromatin contexts from those of DNA sequences. Recently two studies investigated the impact of chromatin features by using over 1,000 copies of integrated reporter sequences to fix the sequence variables (Gisler et al., 2019; Schep et al., 2021). Their findings confirmed the previous observations that heterochromatin has a negative impact on CRISPR-Cas9 mutagenesis efficacy. Notably, specific chromatin features were also identified to impact both efficiency and mutation outcomes (Gisler et al., 2019; Schep et al., 2021). Nevertheless, these studies were limited by two factors: 1) the bias of the integration sites in certain genomic locations, and 2) the ambiguity from whether the newly integrated sequences can quickly and faithfully adopt the local chromatin context. Furthermore, all of these previous studies were conducted in human cell lines within a short time window (usually less than 72 hours) using transiently expressed CRISPR-Cas9. It is still unclear whether the heterochromatic features have a long-lasting effect on CRISPR-Cas9 mutagenesis, or only delay it as suggested by (Kallimasioti-Pazi et al., 2018).

In this study, we leveraged the high resolution epigenomic resources in the model plant species, *Arabidopsis thaliana*, to address the impact of chromatin features on both

CRISPR-Cas9 mutagenesis efficiency and mutation outcomes. To fix the sequence variable, the Arabidopsis genome was scanned to identify identical CRISPR target sites located in various chromatin contexts. By using stable CRISPR-Cas9 transgenic plants targeting 15 chromosomal regions, we systematically characterized mutagenesis efficiency and mutation outcomes with 23 distinct DNA methylation and chromatin features using a Next Generation Sequencing (NGS) approach. Our findings provided insight into the influences of chromatin features on CRISPR-Cas9 mutagenesis and mutation outcomes. This could help develop better technologies for more efficient and precise genome editing.

Results

Identification of identical CRISPR-Cas9 sites in diverse chromatin contexts

To identify identical CRISPR target sites in various chromatin contexts, the Arabidopsis Col-0 reference genome was scanned for 20 bp CRISPR-Cas9 recognition sequences with 3 bp NGG (the PAM sequence) at the 3' end. Out of 7,376,476 distinct target sites, 19,161 were identified as repeating 7-25 times across the genome (Figure 1A). A series of filters were then applied to remove the sites with one of the following features: simple repeats, GC content outside the range of 40-60%, matching sequences in mitochondria or chloroplast genomes, or containing no overlapping restriction enzyme site for subsequent mutation genotyping. The remaining 7,971 sequences, representing 92,117 total genomic sites, were assessed for three key chromatin features using 100 bp windows: chromatin accessibility indicated by ATAC-seq scores (Lu et al. 2016), DNA methylation patterns categorized as a DNA methylation domain (RdDM,

heterochromatin, CG-only, unmethylated and intermediate) (Crisp et al. 2017; Springer and Schmitz 2017), and nine chromatin states (Supplemental Dataset 1) (Sequeira-Mendes et al. 2014). Seven multicopy CRISPR sites (MCsites) were identified with individual sequences in each family having highly diverse chromatin contexts, including both open and closed chromatin, at least 3 different DNA methylation domains, and at least 2 distinct chromatin states (Figure 1B-C; Supplemental Dataset 2).

Differential CRISPR-Cas9 efficiency is associated with distinct chromatin features

Next, we evaluated CRISPR-Cas9 efficacy for each of the seven MCsite families. T-DNA constructs, containing a CRISPR-Cas9 expression cassette, a firefly luciferase reporter and the bialaphos resistance (BAR) selection marker gene, were made to target each MCsite (Supplemental Figure S1A). Each construct contained two gRNA expression cassettes, one targeting the MCsite and the other as the CRISPR mutagenesis control targeting a single-copy endogenous gene located in unmethylated and accessible chromatin, the Cheletase2 (CHLl2) gene, as reported previously (Figure 2A) (Mao et al. 2013). It is worth noting that we intentionally chose the CaMV 35S promoter to drive expression of the Cas9 and gRNA sequences because this promoter has much lower activity in Arabidopsis embryos than leaves (Wang et al. 2015; Yan et al. 2015). By reducing the mutagenesis potential in early embryo development stages, we were able to capture more independent mutation events in somatic cells during leaf development.

After transformation of each T-DNA construct, the resulting CRISPR-Cas9 transgenic plants (T1) were first assessed for mutagenesis efficiency at the CHLl2 control site using the Cleaved Amplified Polymorphic Sequences (CAPS) method. Individual

plants with detectable mutagenesis at the control site were then analyzed at each of the seven MCsites using a CAPS or NGS assay. We were able to identify two MCsites, MCsite4 and MCsite5, that produced notable mutagenesis for at least one of the CRISPR target sites (Supplemental Figure S1B). MCsite4 and 5, totaling 15 individual sites, were found spanning all 5 chromosomes (Supplemental Figure S1C). Among the 15 sites, four seemed to overlap with Arabidopsis genes, and only 1 site, MCsite5.8, was detected in transcribed RNA sequences (Supplemental Dataset 2).

When CRISPR-Cas9 mutagenesis efficiency was assessed at the individual sequences of MCsite4 and 5 using NGS, substantial variations were observed across individual target sites, consistent with the CAPS data (Supplemental Figure S2A). To control for variation in CRISPR-Cas9 expression levels between plants, the mutation rate at each targeted site was normalized to the CHLI2 control. As a result, the normalized frequencies at MCsite4 and MCsite5 sites ranged from an average 0.61-152.28% and 9.17%-69.17%, respectively (Figure 2B-C). Pairwise comparisons indicated mutagenesis frequencies of MCsite4 sites can be categorized into three distinct groups (p value < 0.001): the high editing group (group H: MCsite4.8, 152.28%), the moderate editing frequency group (group M: MCsite4.4 with 17.04% and 4.7 with 16.20%), and the low editing frequency group (group L: MCsite4.1, 4.2, 4.3, 4.5, and 4.9, ranging from 0.61% to 1.71%) (Figure 2B). Comparison between the highest and lowest MCsite4 edited sites revealed a 249.64-fold difference. Similarly, mutagenesis frequencies at the MCsite5 sequences can also be grouped into the high, moderate and low editing groups (p value < 0.01) with 7.54-fold differences between the highest and lowest edited sites (Figure 2C). We then characterized the local sequence context surrounding each target site (25 bp from each

side). High sequence similarities were found in these extended sequences for both MCsite4 and 5 families (Supplemental Figure S2B and C). Phylogenetic analyses within each target site family found no evident associations between sequence similarity and the mutagenesis frequency groups (Supplemental Figure S2B and C). Thus, the local sequence context could not explain differential mutagenesis frequencies observed from individual target sites.

The initial association analysis of indel frequencies with DNA methylation domains and chromatin accessibility indicated unmethylated and accessible sites generally had higher mutagenesis levels than the methylated and inaccessible sites (Figure 2B-C). Notably, negative associations between mutagenesis frequency and DNA methylation levels at both MCsite4 and 5 sites could also be observed when the cytosine methylation status was examined at the single nucleotide level of individual CRISPR target sites (Supplemental Figure S3). To systematically investigate the relationship between mutagenesis efficiency and chromatin features, we further characterized individual MCsites using all 23 chromatin features, including distinct histone modifications and histone variants (Supplemental Dataset 3) (Y. Liu et al. 2018). When hierarchical cluster analysis was performed, the lowly edited sites (group L) tended to cluster with the heterochromatic features, such as H2A.W, H3K9me1, H3K9me2, H3K27me1, and hyper DNA methylation, for the MCsite4 and 5 targets with exceptions observed for MCsite5.6 and MCsite5.8 (Figure 2D-E). While they are in the lowly edited group, these two sites appeared to be associated with unmethylated and accessible chromatin features (Figure 2C and E). On the contrary, the highly and moderately edited groups (group H and M) appeared to be associated with accessible, active chromatin

features such as histone acetylation, H3K36me3, and H3K4 methylation (Figure 2D-E) (Roudier et al. 2011). To examine the impact of individual features, we performed correlation analyses by plotting mutagenesis efficiency at all 15 target sites with each chromatin and DNA methylation feature (Supplemental Figure S4). Consistent with the hierarchical cluster analysis, strong positive correlations were observed between mutagenesis frequency and euchromatin-related features such as H3K56ac ($R = 0.85$, $p = 1.8e-13$), H3K9ac ($R = 0.82$, $p = 8.6e-12$), H3K36ac ($R = 0.75$, $p = 4.2e-9$), H3K27ac ($R = 0.71$, $p = 4.1e-8$), H3K36me3 ($R = 0.71$, $p = 3.5e-8$), H3K4me3 ($R = 0.65$, $p = 1.6e–6$), and accessibility ($R = 0.51$, $p = 0.00038$) measured with ATAC-Seq data (Supplemental Figure S4). Moreover, strong negative correlations were found between mutagenesis frequencies and the heterochromatin-related features H2A.W ($R = -0.57$, $p = 4.1e-5$), H3K9me1 ($R = -0.55$, $p = 0.00011$), H3K9me2 ($R = -0.55$, $p = 1e-4$) and cytosine DNA methylation ($R=-0.45$, $p = 0.0022$) (Supplemental Figure S4).

Improving CRISPR-Cas9 mutagenesis efficiency through combined reduction of DNA methylation

Because of the strong negative association observed between the lowly edited sites and heterochromatic features, we hypothesized that altering these chromatin states could improve mutagenesis efficiency at the refractory sites. In this study, we chose to perturb DNA methylation at the lowly edited sites because high levels of DNA methylation are often associated with heterochromatic features (Crisp et al. 2020; Johnson et al. 2007; Zemach and Grafi 2003). We first sought to test the possible impact

of CHG methylation on mutagenesis efficiencies because CHG methylation is often associated with H3K9me2 and heterochromatin (Springer and Schmitz 2017). To this end, the chromomethylase3 (*cmt3-11t*) mutant was chosen due to the well-documented genome-wide reduction in CHG methylation (Stroud et al. 2013). Analysis of the single-base resolution DNA methylation profiles indicated that a substantial reduction in CHG methylation was confirmed at individual MCsite4 and 5 CRISPR targets in this mutant (Supplemental Figure S5). The CRISPR-Cas9 T-DNA constructs targeting MCsite4 and 5 sequences were transformed into the *cmt3-11t* mutant using the same procedure described above. As a result, transgenic *cmt3-11t* plants (T1) were identified with detectable mutagenesis activities at both MCsite4 and 5 target sites (Supplemental Figure S6). After normalizing to the CHLI2 control target site, we observed a nearly identical pattern in mutagenesis efficiency between WT and *cmt3-11t* plants for both MCsite4 and MCsite5 (Figure 3A-B). These results suggested that a reduction of CHG methylation alone is not sufficient to improve CRISPR-Cas9 mutagenesis at the refractory sites.

Although additional genetic mutant lines with perturbed DNA methylation, such as Methyltransferase 1 (*met1*), were available (Stroud et al. 2013), we have not been able to obtain transgenic CRISPR-Cas9 T-DNA events in these mutant lines. Therefore, we chose a chemical approach to reduce genome-wide DNA methylation in all 5mC contexts using 5-azacytidine (Griffin, Niederhuth, and Schmitz 2016). Analysis of the DNA methylation for 5-azacytidine treated plants revealed substantial reductions in DNA methylation at the highly methylated sequences at both MCsite4 and 5 (Supplemental Figure S7A) (Griffin, Niederhuth, and Schmitz 2016). We focused on MCsite4 in the

subsequent 5-azacytidine experiment due to the lower mutagenesis frequencies at the lowly edited sites. The T2 siblings with the CRISPR-Cas9 T-DNA targeting MCsite4 in the wildtype background were obtained from a self-pollinated T1 plant used above. After 2-weeks of growth with or without 100 μM 5-azacytidine, individual plants were subjected to mutagenesis analyses using the NGS assay (Supplemental Figure S7B). No significant differences were observed for the normalized mutagenesis frequency at each MCsite4 site between the untreated and treated samples (Figure 3C). Thus, partial reductions in all DNA methylation contexts using 5-azacytidine did not reveal changes in mutagenesis frequencies.

Lastly, we tested the impact of combined reductions in DNA methylation using both the *cmt3-11t* mutant and 5-azacytidine treatments. The T2 siblings with the MCsite4-targeting T-DNA were obtained from a self-pollinated T1 plant in the *cmt3-11t* mutant background, grown for 2 weeks with or without 100 μM 5-azacytidine, and subjected to NGS analysis using the same procedure (Supplemental Figure S7C). After normalizing to the CHLl2 control, we observed two distinct patterns for mutagenesis efficiency across individual MCsite4 targets. At the unmethylated and accessible sites (groups H and M), no significant differences in mutagenesis frequency were found between the 5-azacytidine treated mutants and the control group (Figure 3D). On the contrary, significant increases in mutagenesis frequency were observed at the inaccessible and hypermethylated sites (group L) when 5-azacytidine treatment was combined with the *cmt3-11t* mutant, i.e., MCsite4.9 (4.01-fold), MCsite4.3 (2.12-fold), MCsite4.1 (4.73-fold), MCsite4.5 (4.83-fold), and MCsite4.2 (3.87-fold) (Figure 3D). Together, these data indicated that strong reduction in DNA methylation in multiple contexts could

result in significant improvement of CRISPR-Cas9 mutagenesis efficiency at the hypermethylated refractory sites.

Differential CRISPR-Cas9 mutational profiles are associated with distinct chromatin features

CRISPR-Cas9 induced mutations are typically composed of either small deletions or 1 bp insertions (Allen et al. 2018). Small deletions are mainly derived from the NHEJ or MMEJ pathway through exonuclease-mediated end processing and ligation, while the 1 bp insertions were mostly generated from blunt-end or 1 bp staggered cleavage by Cas9 followed by DNA polymerase-mediated end filling (Gisler et al. 2019). To investigate the potential impact of chromatin context on CRISPR-Cas9 mutation outcomes, we examined the insertion and deletion profiles for both MCsite4 and 5. As expected, most mutations contained 1 bp insertions or small deletions (< 10 bp) for both MCsite4 and 5 (Figure 4A). MCsite4 was preferentially repaired as 1 bp insertions, while MCsite5 showed a strong bias towards small deletion outcomes (Figure 4A).

While the major mutation types at individual sites within each family seemed to be highly similar (Supplemental Figure S8A-D), further analysis revealed substantial variations for the insertion rate between individual sites. The rate of insertion outcomes ranged from 56.25-81.73% and 7.57-30.27% in MCsite4 and MCsite5 sites, respectively (Figure 4B-C). We then conducted a correlation analysis using the 23 epigenetic features with the insertion rates at all 15 sites. Three histone H3 related features, H3K4me1 ($R = -0.64$, $p = 0.01$), H3.3 ($R = -0.83$, $p = 3e\text{-}4$), and H3.1 ($R = -0.91$, $p = 2.8e\text{-}6$), were

identified with significant negative correlations with 1 bp insertional mutations (Figure. 4D; Supplemental Figure S9). Thus, these findings strongly suggested that chromatin features could not only affect CRISPR-Cas9 mutagenesis efficiency but also influence mutation outcomes.

Discussion

In recent studies, chromatin contexts have been demonstrated to have significant impacts on CRISPR-Cas9 mediated genome editing (Daer et al. 2017; G. Liu et al. 2019; Wu et al. 2014; Yarrington et al. 2018). Most of these findings indicated that heterochromatic features at the CRISPR target regions could impede CRISPR-Cas9 mutagenesis efficiency in multiple systems, such as yeast, rice, mouse, and human cell lines (Daer et al. 2017; G. Liu et al. 2019; Wu et al. 2014; Yarrington et al. 2018). However, it remained unclear whether heterochromatic features could only temporarily delay CRISPR-Cas9 mutagenesis (Kallimasioti-Pazi et al. 2018). In this study, we systematically characterized the impact of 23 distinct DNA methylation and chromatin features on CRISPR-Cas9 mutagenesis efficiency and mutation outcomes by investigating CRISPR-Cas9 transgenic Arabidopsis plants. Consistent with the previous studies, our results demonstrated that inaccessible and heterochromatic features were associated with low mutagenesis efficiency. Such repressive effects can be long-lasting through plant development leading up to a 250-fold difference between identical CRISPR target sites. However, it was worth noting that, although this observation broadly holds for both MCsite4 and 5, two lowly edited sites in MCsite5, MCsite5.6 and 5.8, appeared to be associated with open and active chromatin features. The weaker associations

observed here could be due to the overall higher mutagenesis frequencies at the lower edited MCsite5 targets than those at the MCsite4 targets. Furthermore, mutagenesis efficiency at the target sequences in accessible chromatin regions could also have significant variations ranging from 1.53-fold to 10-fold at MCsite5 and MCsite4, respectively (Figure 2B-C). These observations suggested that specific chromatin features other than just open or closed chromatin should be considered to account for CRISPR-Cas9 mutagenesis efficacy. Close examination of individual chromatin features identified several euchromatic marks, such as H3K9ac, H3K56ac, H3K36ac, H3K27ac, H3K4 methylation and H3K36m3, that were positively correlated with mutagenesis efficiency. Further investigation with a larger data set will be needed to dissect their impacts on CRISPR-Cas9 mutagenesis in greater detail.

In this study, we observed strong negative correlations between CRISPR-Cas9 mutagenesis efficiency and repressive chromatin features, such as DNA hypermethylation, low DNA accessibility, and H3K9 methylation. To test the hypothesis that modulating some of these features could improve mutagenesis frequency at the lowly edited sites, we sought to reduce DNA methylation by using both genetic and chemical approaches. Our results indicated that partial reduction of DNA methylation using either a mutant affecting CHG methylation or 5-azacytidine treatment alone was not sufficient to improve mutagenesis efficiency at any of the tested sites. When combining the CHG deficiency mutant with 5-azacytidine chemical treatment, 2.1 to 4.8-fold improvements in mutagenesis efficiency were found at the lowly edited sites but not at the highly and moderately edited sites. Combined reduction of DNA methylation in multiple contexts has been demonstrated to increase chromatin accessibility and even alter the higher order

3D chromatin organization in Arabidopsis (Zhong et al. 2021). Thus, mutagenesis efficiency improvement at lowly edited sites observed here could have resulted from increasing chromatin accessibility and/or changing 3D chromatin organization due to the substantial reduction of DNA methylation in multiple contexts; but this would need to be experimentally verified, for example using ATAC-Seq or DNase I hypersensitivity assays. Additionally, as discussed above, open/close chromatin structure alone could not explain all the variations in editing efficiencies. A systematic approach would be required to further dissect the relationship between epigenetic features and targeted mutagenesis efficiency. Furthermore, as reported previously, repressive chromatin features could act as the barriers to hinder CRISPR-Cas9 binding and cleavage (Wu et al., 2014). However, it is noted that the primary readout in this study, targeted mutation frequency, is an indirect measurement of CRISPR-Cas9 target binding and cleavage dependent on DSB repair. The mutation efficiency differences observed here could reflect differences in either CRISPR-Cas9 binding, cleaving, DNA repair, or a combination of these. Further investigation would be required to differentiate these possibilities.

In addition to the impacts on mutagenesis efficiency, chromatin features have been suggested to influence CRISPR-Cas9 mutation outcomes. For example, a recent study with more than 1,000 copies of identical insertion sites indicated the 1 bp insertions were found more prevalent in euchromatin than in heterochromatin, likely through recruiting different DNA repair machinery (Schep et al. 2021). However, conflicting results were also reported showing little impact of chromatin features on mutation outcomes (Gisler et al. 2019; Kallimasioti-Pazi et al. 2018). In this study, we observed significant variations for the 1 bp insertion rate in different chromatin contexts. Of the 23

chromatin features analyzed, three distinct histone H3 features, H3K4me1, H3.3, and H3.1, exhibited significantly strong negative correlations with the 1 bp insertion rate. Interestingly, H3K4me1 was also identified by Schep et al. as a marker to correlate with distinct mutation outcomes, while H3.1 and H3.3 were not included in their study (Schep et al. 2021). This suggested that the balance between 1 bp insertions and small deletions could be influenced by specific chromatin markers. Distinct chromatin features have been reported to recruit different DNA repair machinery and result in different repair outcomes in mammalian systems (Fnu et al. 2011; Jacquet et al. 2016; Luijsterburg et al. 2016). Further investigation is needed to address the potential roles of specific chromatin markers in determining DNA repair outcomes.

We propose a model to account for the impacts of chromatin features on CRISPR-Cas9 mutagenesis efficiency and mutation outcomes (Figure 5). In the first step, chromatin features are the key determinants for CRISPR-Cas9 recognition and binding efficiency (Figure 5). In general, heterochromatic features, such as H3K9 methylations, H3K27me1, H2A.W, and DNA hypermethylation, could substantially reduce chromatin accessibility and thus reduce the recognition and binding efficiency of CRISPR-Cas9. Such repressive effects could persist through cell division and development. After CRISPR-Cas9 locates and binds the genomic target site, it can introduce a DSB with either a 1-bp 5’ overhang (staggered cut), or blunt ends (Gisler et al. 2019). The cleaved product can be repaired to yield three outcomes: wild-type sequence by perfect ligation, 1 bp insertions, or small deletions (less than 10 bp) (Figure 5). It has been proposed that the staggered cut primarily leads to 1 bp insertion via template-dependent repair, facilitated by a DNA polymerase, while the blunt cut mainly results in small deletions through end

resection (Schmid-Burgk et al. 2020; Gisler et al. 2019). The blunt ends could occasionally be repaired by DNA polymerase, likely by members from DNA polymerase family X, without DSB end resection, resulting in template-independent 1 bp insertions (Gisler et al. 2019). In this study, we observed both templated 1 bp insertion and template-independent insertions, as exemplified in MCsite5 and MCsite4, respectively (Supplemental Figure S8A-B). Nevertheless, the balance between 1 bp insertion and small deletion products is primarily dependent on the repair choices between short-range DSB end resection and DNA polymerase end filling (Figure 5) (Schmid-Burgk et al. 2020; Lemos et al. 2018). While sequence features are a key determinant in mutation profile, our data indicated that chromatin features, such as H3K4me1, H3.3, and H3.1, could also significantly impact the balance between 1 bp insertion and small deletion outcomes. These specific chromatin features may exert their influences on the balance between the staggered and blunt cut, or through modulating the balance between DNA polymerase end filling and short-range DSB end resection during NHEJ. In fact, previous studies have demonstrated that distinct chromatin features, such as the H3.3 variant, could impact DSB end resection (Luijsterburg et al. 2016).

Recently, much effort has been made to develop computational tools to predict CRISPR-Cas9 editing efficiency and/or repair outcomes. To our knowledge, these tools primarily relied on sequence features (Allen et al. 2018; Concordet and Haeussler 2018; Xiang et al. 2021). Yet our results indicated that non-sequence features should also be taken into consideration for the prediction of both CRISPR-Cas9 efficiency and mutation outcomes. One implication from this study is that heterochromatin related features, such as low accessibility, H3K9me1, H3K9me2, H3K27me1, H2A.W and DNA

hypermethylation, should be avoided in order to design gRNAs with high efficiency. When it comes to predicting CRISPR-Cas9 mutation profiles, chromatin features such as H3K4me1, H3.3, and H3.1 should also be considered. Future studies using genetic mutants that result in altered states of histone modifications will be interesting to further dissect the impact of distinct chromatin features on CRISPR-Cas9 mutagenesis. A better understanding of the interplay between chromatin dynamics and CRISPR-Cas9 will enable the development of more precise and efficient genome engineering technologies.

Materials and Methods

Identification of multicopy CRISPR sites

To identify gRNAs that matched to multiple places in the genome, the Arabidopsis (*Arabidopsis thaliana*) Col-0 reference genome (TAIR10) was parsed to identify every NGG PAM site using seqkit locate (Shen et al. 2016) with the search motif NNNNNNNNNNNNNNNNNNNNGG. The number of occurrences of each non-redundant sequence in the genome was summarized using csvtk. The resulting distinct target sites were then filtered in R (v4.1.2) to retain gRNAs that had between 7 to 25 distinct matches. Sites were then eliminated if they had simple sequence motifs consisting of 5 As, Ts, Gs or Cs in a row; CG content outside 40%-60% or if the sequence was found in the chloroplast or mitochondria genome; to give a final list of 9,902 candidate gRNAs. The gRNA identification script is available at https://github.com/pedrocrisp/Weiss_et_al_gRNA_chromatin. The potential gRNAs were also screened to identify those with restriction enzyme recognition motifs (from a list commercially available enzymes from NEB) that overlapped position 17-18 of the gRNA (between position 3 and 4 bp upstream from the PAM) such that an indel mutation in this position would disrupt restriction enzyme recognition and cleave for efficient screening of edited transgenic lines. A final list of 7,971 gRNAs were then annotated with chromatin state information by overlapping the coordinates of the gRNA target sites with chromatin annotation files using bedtools (Quinlan and Hall 2010). Chromatin data included nine histone states (Sequeira-Mendes et al. 2014); chromatin accessibility (Lu et

al. 2016); DNA methylation (Crisp et al. 2017; Springer and Schmitz 2017). DNA methylation data was converted to methylation domains for each 100 bp non-overlapping window of the TAIR10 genome using the method detailed in (Crisp et al. 2020). Chromatin accessibility at the CRISPR target site was called either Open or Closed based on the presence of an overlapping ATAC-Seq peak or lack thereof, respectively, using the accessibility profiles in (Lu et al. 2016). Gene and transposable element annotations were downloaded from Araport v11.

T-DNA vector construction

The CRISPR-Cas9 constructs were created using the Golden Gate assembly method as outlined in Čermák et al. 2017. The gRNA sequences were first assembled into the pMOD_B2301 backbone containing the gRNA targeting the multicopy CRISPR site (MCsite) and the CHL2 positive control using the oligos listed in Supplemental Table S3. T-DNA constructs were assembled using pMOD_A0101 (Cas9), pMOD_B2301 containing the gRNA array, pMZ105 (luciferase reporter), and pTRANS230d (T-DNA backbone) using the Golden Gate method with AarI type 2 restriction enzyme. The modular components used to build the T-DNA plasmids can be found at https://www.addgene.org/browse/article/28189956/ (Čermák et al. 2017). The T-DNA constructs described in this study, and their associated sequence maps, are available at Addgene.

Plant materials and growth conditions

The *Arabidopsis thaliana* Columbia ecotype (Col-0) was used in these experiments. The cmt3-11t (stock CS16392) genotype was acquired from the Arabidopsis Biological Resource Center (ABRC). Floral dip transformation was performed according to the protocol as previously outlined (Zhang et al. 2006). Transgenic T1 seeds were sown on soil and exposed to BASTA selection to recover transgenic plants. Plants were grown in a growth chamber with the following conditions: 16/8 hours light/dark cycle, 22 °C, and 55% humidity.

Characterization of chromatin features

Previously analyzed datasets (BigWig files) for the 23 chromatin features were downloaded from Plant Chromatin State Database (PCSD) (Y. Liu et al. 2018). The datasets for wild type Arabidopsis with and without 100 μM 5-azacytidine treatment was downloaded from Griffin et al. (Griffin, Niederhuth, and Schmitz 2016). For each dataset, the values for the 1kb window (500 bp upstream and 500 bp downstream from the center of gRNA) were calculated using the Deeptools2 computeMatrix in the reference-point mode. Similarly, the values for each nucleotide of target sites were calculated using the scale-region mode at single base resolution. Each dataset was then normalized to allow for comparison on a scale of 0-1 with 1 indicating the highest level of that feature.

5-azacytidine treatment and luciferase screening

T2 seedlings from self-pollinated T1 Arabidopsis plants were grown on 1% (w/v) agar plates containing 0.5 Murashige and Skoog (PhytoTech Labs) and 100 μM 5-azacytidine (Griffin, Niederhuth, and Schmitz 2016). After two weeks, seedlings were

screened for presence of the transgene using a luciferase reporter. The luciferase assay procedure was performed using the Bio-GloTM Luciferase Assay System (Promega Corp., Madison, WI, USA) in accordance with the manufacturer's instructions.

Mutation genotyping and characterization of mutation profiles

Genotyping was performed using two methods: genomic PCR followed by restriction enzyme digestion (CAPS) and the NGS assay using Illumina paired-end read amplicon sequencing. All tissues for genotyping were collected at two weeks post germination for the CTAB-base genomic DNA extraction. PCR was performed using GoTaq Green Mastermix (Promega Corp., Madison, WI, USA) according to the manufacturer's instructions, with an annealing temperature of 55 °C (CHlL2 and MCsite4) or 60 °C (MCsite5) with an extension time of 1 minute. Primers to amplify CHLl2, MCsite4, and MCsite5 target sites can be found in Table S1. Amplicons were then subjected to restriction enzyme digestion using BsmAI (CHlL2), AluI (MCsite4), or DrdI (MCsite5) according to the manufacturer's instructions. PCR amplicons generated with the corresponding primers were subjected to Illumina paired-end read sequencing (Genewiz Inc., South Plainfield, NJ, USA). The raw NGS reads were analyzed using CRISPResso2 to estimate indel mutation rates (Clement et al. 2019). To analyze the mutation profiles for each sample, the NGS reads with indel mutations were extracted from the CRISPResso2 output files with a 2% threshold. The resulting output files were then loaded into R studio (version 4.1.0) for data visualization using ggplot. The total read counts for CRISPR-Cas9 editing frequency and repair profiles can be found in Table

S2. Normalized indel frequencies were calculated by dividing the indel frequency of each MCsite by the CHlL2 positive control indel frequency within each replicate.

Data availability

All sequencing data analyzed in this manuscript is available at the National Center for Biotechnology Information (NCBI) under BioProject Accession PRJNA795172 (Supplemental Dataset 4).

Figures

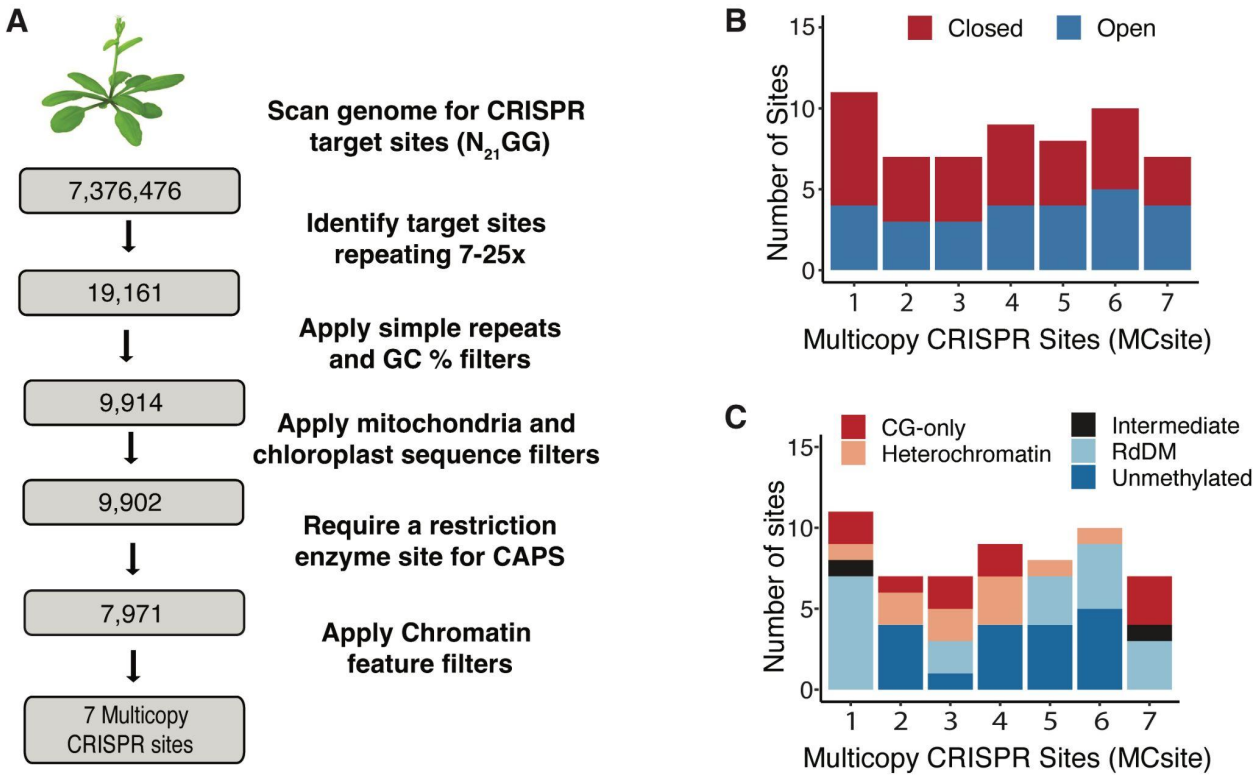


Figure 1 | Identification of multicopy CRISPR sites (MCsites) in various chromatin contexts. (A) Bioinformatic pipeline used to identify MCsites. (B) Characterization of the

chromatin accessibility as Closed (red) or Open (blue) at each MCsite using ATAC-Seq data. (C) Characterization of DNA methylation domains at each MCsite. CG-only (red) sites contain greater than 40% mCG; Heterochromatin (tan) sites contain greater than 40% mCG and CHG methylation; RdDM (light blue) sites contain mCG, mCHG, and mCHH, with at least 15% mCHH; Unmethylated (blue) sites contain less than 10% mCG, mCHG, and mCHH; Intermediate (black) sites are everything else with data that didn't meet one of the above criteria.

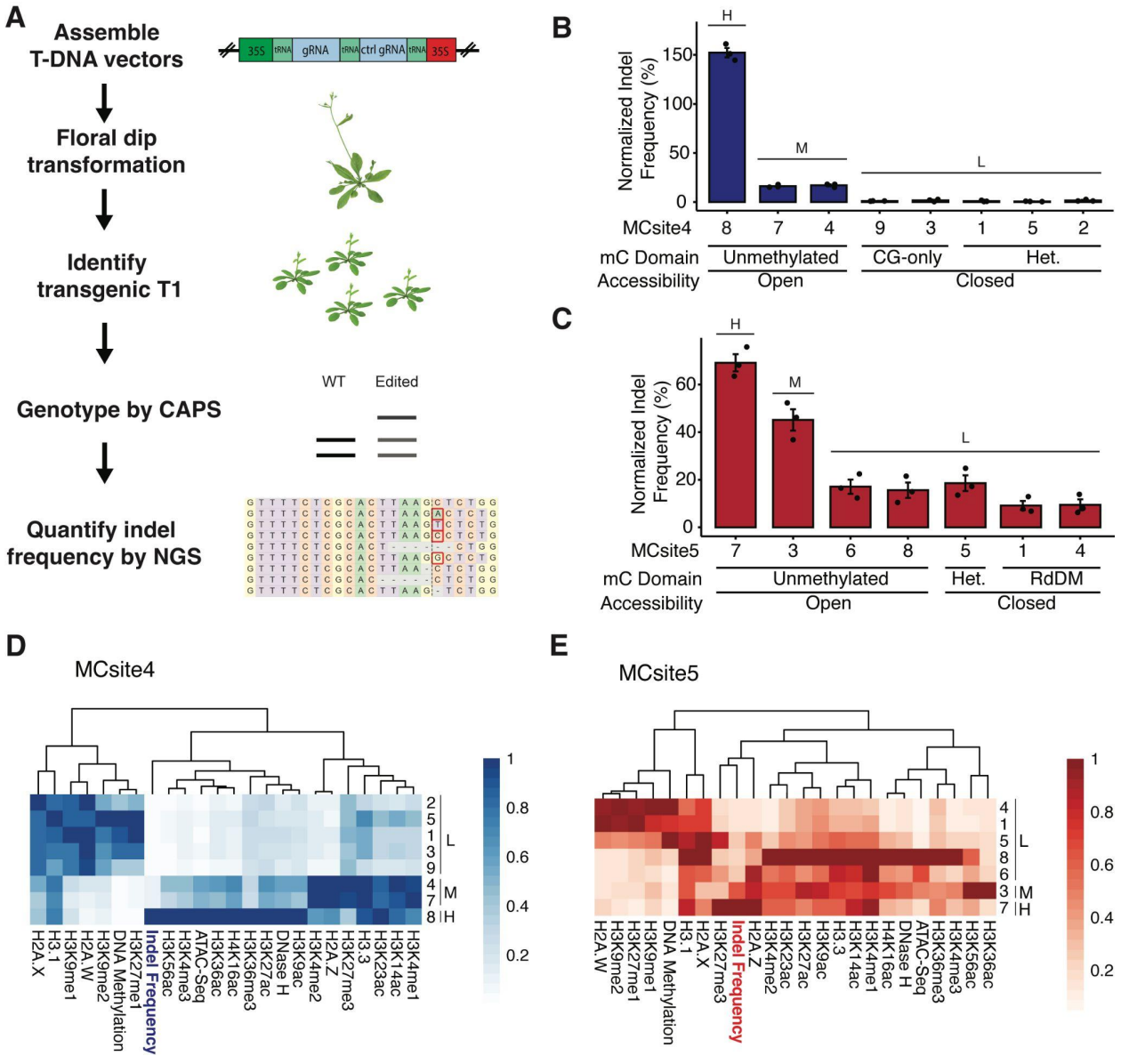


Figure 2 | Characterization of CRISPR-Cas9 mutation frequencies and chromatin features at MCsite4 and MCsite5 sites. (A) The NGS-based pipeline to generate CRISPR-Cas9 transgenic Arabidopsis plants for mutation efficiency and profile analysis. (B) Normalized indel mutation frequency for the eight MCsite4 sites. The standard error (SEM) is displayed for each target site with the black dots indicating replicates (n = 3). H (high), M (moderate), and L (low) correspond with the mutagenesis group that the CRISPR site is associated with. The DNA methylation domain and accessibility contexts

are listed below each target site. The Heterochromatin DNA methylation domain is abbreviated as Het. A one-way analysis of variance and Tukey's multiple comparison test was performed. (C) Bar graph displaying normalized indel frequency for the seven MCsite5 CRISPR targeted sites (red). The standard error (SEM) is displayed for each target site with the black dots indicating replicates (n = 3). H (high), M (moderate), and L (low) correspond with the mutagenesis group that the CRISPR site is associated with. The DNA methylation domain and accessibility contexts are listed below each target site. The Heterochromatin DNA methylation domain is abbreviated as Het. A one-way analysis of variance and Tukey's multiple comparison test was performed. (D) Hierarchical clustering heatmap of distinct chromatin features and CRISPR-Cas9 indel frequency (x-axis) for all eight MCsite4 CRISPR targets (y-axis). H (high), M (moderate), and L (low) correspond with the mutagenesis group that the CRISPR site is associated with. Each feature was normalized to a scale of 0-1, with 1 indicating high levels (dark blue) and 0 indicating low levels (white). The dendrogram indicates similarity clustering of the x-axis. (E) Hierarchical clustering heatmap of distinct chromatin features and CRISPR-Cas9 indel frequency (x-axis) for all eight MCsite4 CRISPR targets (y-axis). H (high), M (moderate), and L (low) correspond with the mutagenesis group that the CRISPR site is associated with. Each feature was normalized to a scale of 0-1, with 1 indicating high levels (dark red) and 0 indicating low levels (white). The dendrogram indicates similarity clustering of the x-axis.

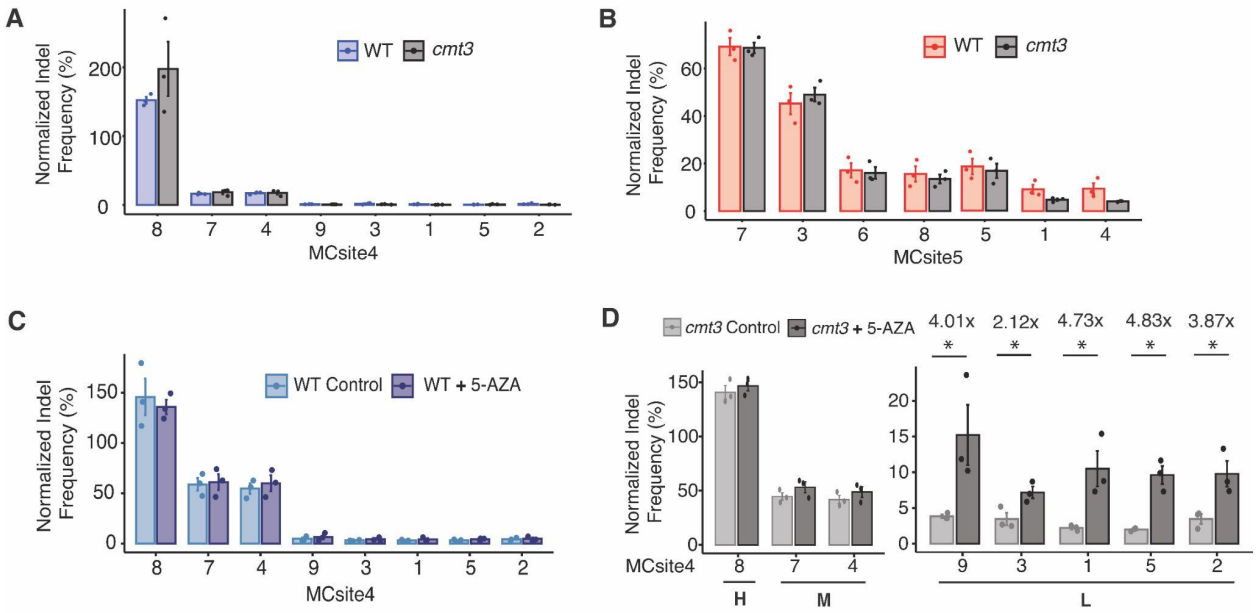


Figure 3 | CRISPR-Cas9 normalized mutagenesis frequencies. (A) Bar graph displaying normalized indel frequency for the eight MCsite4 CRISPR targeted sites for wild type (blue) and cmt3 (dark gray). The standard error (SEM) is displayed for each target site with the dots indicating replicates (n = 3). (B) Bar graph displaying normalized indel frequency for the eight MCsite5 CRISPR targeted sites for wild type (red) and cmt3 (dark gray). The standard error (SEM) is displayed for each target site with the dots indicating replicates (n = 3). (C) Bar graphs of the normalized mutagenesis frequencies for MCsite4 T2 wild type plants with (WT 5-AZA Treatment, dark blue) and without (WT Control, light blue) 5-azacytidine treatment. The standard error (SEM) is displayed for each target site with the dots indicating replicates (n = 3). (D) Bar graphs of the normalized mutagenesis frequencies for MCsite4.8, 7, and 4 T2 cmt3 plants with (cmt3 5-AZA Treatment, dark gray) and without (cmt3 Control, light gray) 5-azacytidine. The standard error (SEM) is displayed for each target site with the dots indicating replicates (n = 3). The x-axis displays the MCSite4 CRISPR target (MCsite4), DNA methylation domain

(mC domain), and the chromatin accessibility domain (Accessibility). (E) Bar graphs of the normalized mutagenesis frequencies for MCsite4.9, 3, 1, 5, and 2 T2 cmt3 plants with (cmt3 5-AZA Treatment, dark gray) and without (cmt3 Control, light gray) 5-azacytidine. The standard error (SEM) is displayed for each target site with the dots indicating replicates (n = 3). The x-axis displays the MCSite4 CRISPR target (MCsite4), DNA methylation domain (mC domain), and the chromatin accessibility domain (Accessibility). A single asterisk indicates a significant p-value ($p < 0.05$) according to a Mann-Whitney U test.

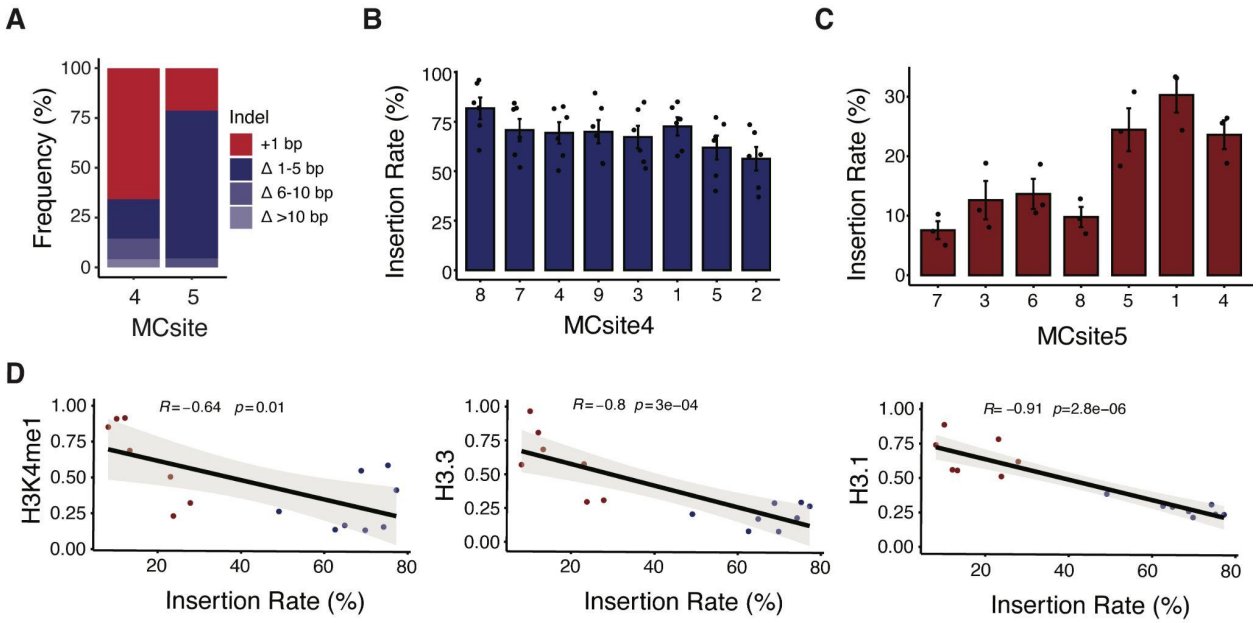


Figure 4 | CRISPR-Cas9-induced DSB repair analysis for MCsite4 and MCsite5 in wildtype samples. (A) Stacked bar graph showing the proportion of indel repair outcomes for MCsite4 and MCsite5. All 7 and 8 CRISPR targets are included for MCsite4 and MCsite5, respectively. 1 bp insertions are labeled red (+1 bp), 1-5 bp deletions are labeled dark blue (Δ 1-5 bp), 6-10 bp deletions are labeled blue (Δ 6-10 bp), and deletions >10 bp are labeled light blue (Δ >10 bp). (B) Bar graph for the insertion rate at MCsite4 CRISPR targets. Each blue bar represents a CRISPR target site; the standard error (SEM) is displayed for each target site with the dots indicating replicates (3 wild type T1 and 3 wild type T2 samples; n = 6). The asterisk indicates a significant p-value ($p < 0.05$) between MCsite4.8 and 4.2 according to a Kruskal-Wallis analysis of variance and Tukey's multiple comparison tests. The total number of edited NGS reads used for repair profile analysis at each CRISPR target site ranged from 837 at MCsite4.1 to 342,029 at MCsite4.8. (C) Bar graph for the insertion rate at MCsite5 CRISPR targets. Each red bar represents a CRISPR target site; the standard error (SEM) is displayed for

each target site with the dots indicating replicates (n = 3). The Kruskal-Wallis analysis of variance p-value of 0.016 is indicated in the upper left corner of the graph. The total number of edited NGS reads used for repair profile analysis at each CRISPR target site ranged from 2,495 at MCsite5.1 to 21,999 at MCsite5.3 (D) Correlation plots for the insertion rate with H3K4me1, H3.3 or H3.1. The blue dots represent MCsite4 and the red dots represent MCsite5. The trendline is black with gray indicating the standard error. The R value and p-values are indicated at the top of each correlation plot according to Spearman's rank correlation coefficient. Each chromatin feature was normalized using all 15 target sites to allow for comparison between the CRISPR target sites on a scale of 0 to 1, with higher values indicating more for the respective chromatin feature.

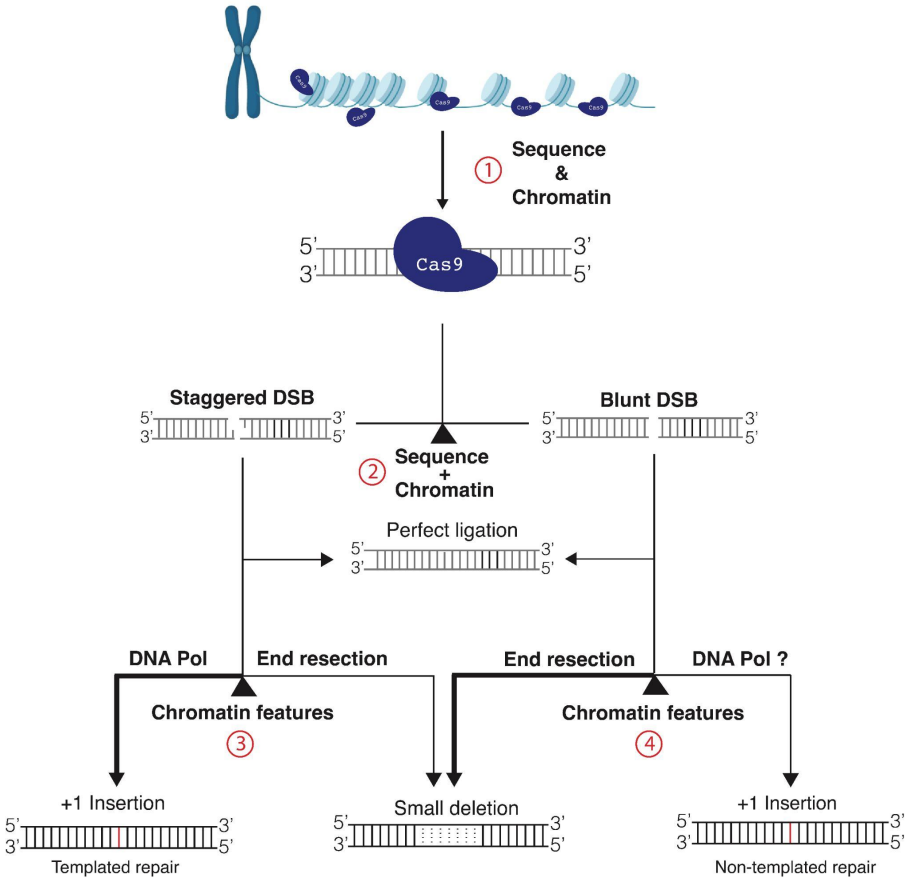


Figure 5 | Model depicting the features that influence CRISPR-Cas9 editing efficiency and DNA repair outcomes. (A) CRISPR-Cas9 editing efficiency. A chromosome cartoon zooming in to show individual nucleosomes along a strand of genomic DNA. Blue Cas9 cartoons attempt to bind the genomic DNA. The arrow is pointing downwards to Cas9 binding the genomic DNA with "Chromatin features" indicating a primary determinant of Cas9 (blue Cas9 cartoon) recognition and binding to genomic DNA (gray DNA cartoon). (B) CRISPR-Cas9-mediated repair of a DSB conveyed as a decision tree. Each triangle represents a node with the features that contribute to the decision listed underneath. The line thickness indicates the preference for that repair outcome to occur. "DNA Pol" and "End resection" indicate the contributing factors for the repair outcome.

Supplemental Figures

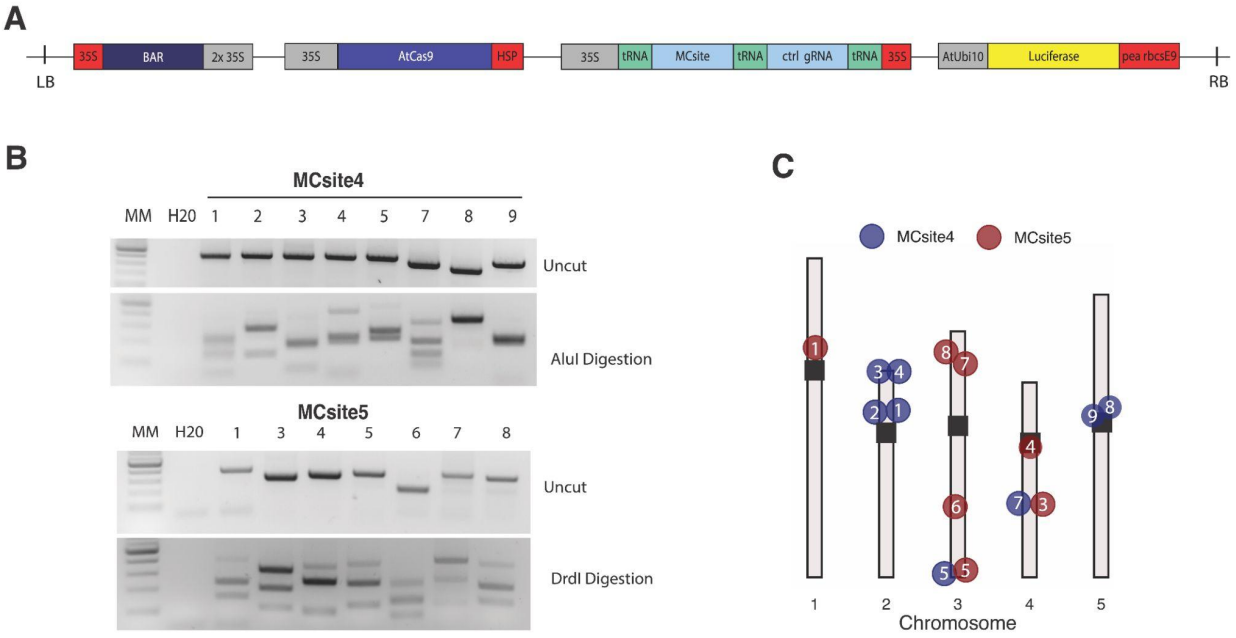


Figure S1. Characterization of multicopy CRISPR sites (MCsites) for CRISPR-Cas9 mutagenesis. (A) Illustration of the transfer DNA (T-DNA) constructs with left border (LB) and right border (RB) on each end . (B) Representative cleaved amplified polymorphic sequence (CAPS) genotyping images for MCsite4 and MCsite5. Samples were genotyped by genomic PCR (Uncut) and the CAPS assay (Alu Digestion for MCsite4 and DrdI digestion for MCsite5), with a 1-kb ladder (MM), and no genomic DNA control (H20). (C) Distribution of the CRISPR target sites of MCsite4 (blue) and MCsite5 (red). Gray bars represent each chromosome with the black box indicating the centromere. The white number inside of each colored circle corresponds to the CRISPR target for that MCsite. Site 6 from MCSite4 and Site 2 from MCSite5 were not amplifiable with the site-specific PCR primers. Thus, they were excluded from the CAPS and NGS assays.

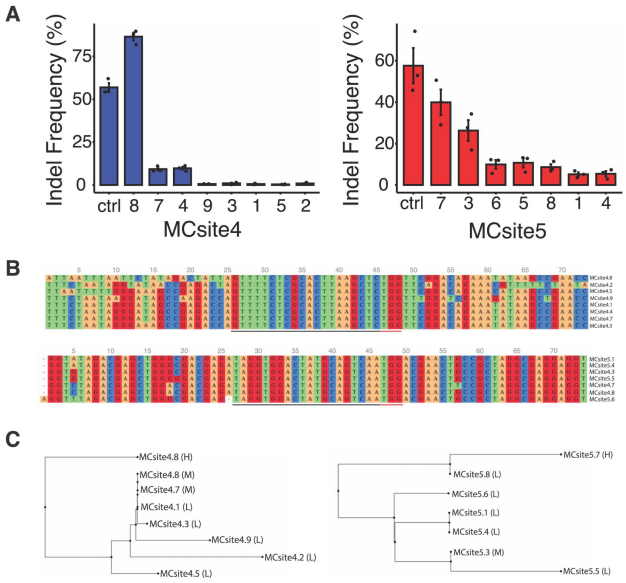


Figure S2. Non-normalized mutagenesis efficiency and sequence comparison for individual target sites in multicopy CRISPR site 4 (MCsite4) and multicopy CRISPR site 5 (MCsite5). (A) Bar graphs displaying the non-normalized mutagenesis frequencies at CRISPR targeted sites for MCsite4 (blue) and MCsite5 (red). Ctrl is abbreviated for the CHLl2 site. The non-normalized indel frequency was calculated by dividing the number of edited reads by the number of total reads, for each replicate. The standard error (SEM) is displayed for each target site with the dots indicating independent replicates (n = 3). (B) Sequence alignments of the CRISPR targets for MCsite4 and MCsite5. The sequence includes the 25 nucleotides to the left and to the right of the protospacer (underlined by a black bar) and PAM sequence (underlined by a red bar). The dendrogram indicates similarity between the sequences. Alignment was created using the MAFFT version 7 online tool with default settings (https://mafft.cbrc.jp/alignment/server/) (1). (C) Sequence similarity dendrograms were created using the Neighbor-Joining (NJ) method on MAFFT version 7 online tool with default settings

(https://mafft.cbrc.jp/alignment/server/). H (high), M (moderate), and L (low) next to each target site indicate the mutagenesis group that CRISPR site is associated with.

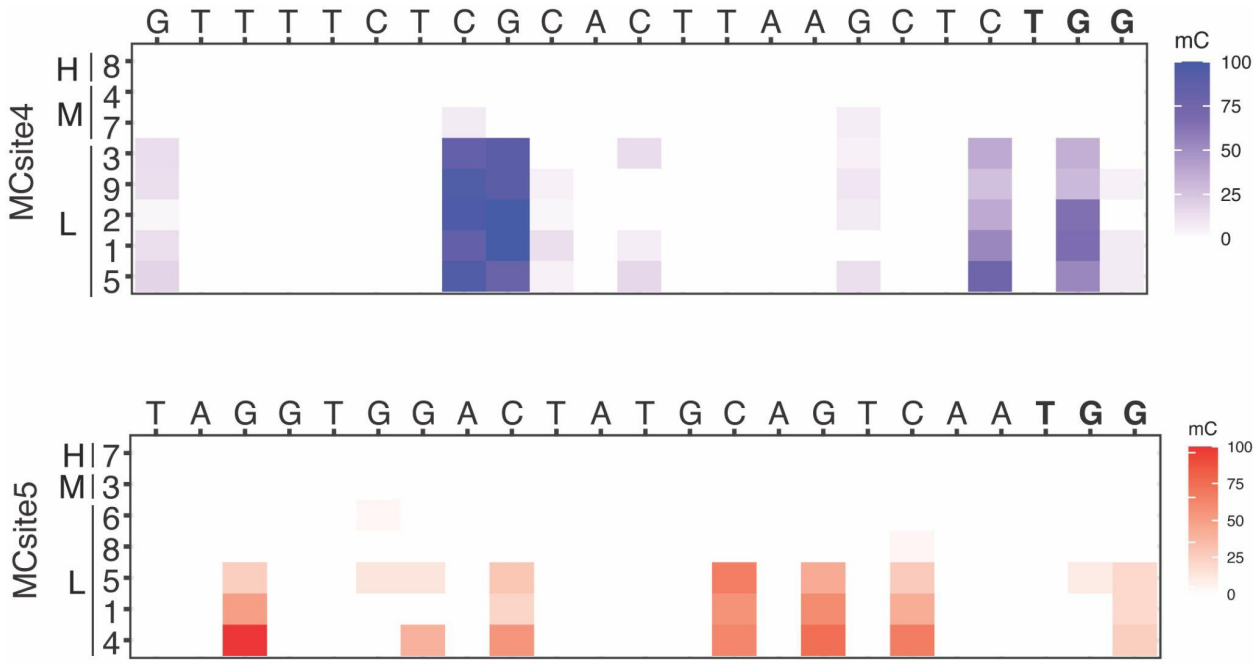


Figure S3. Single nucleotide heatmap of DNA methylation levels at multicopy CRISPR site 4 (MCsite4) (blue) and multicopy CRISPR site 5 (MCsite5) (red) protospacer and PAM (bold) sequences from 0 (unmethylated) to 100 (fully methylated). H (high), M (moderate), and L (low) along the y-axis indicate which mutagenesis group that CRISPR site is associated with. The PAM sequence is bold.

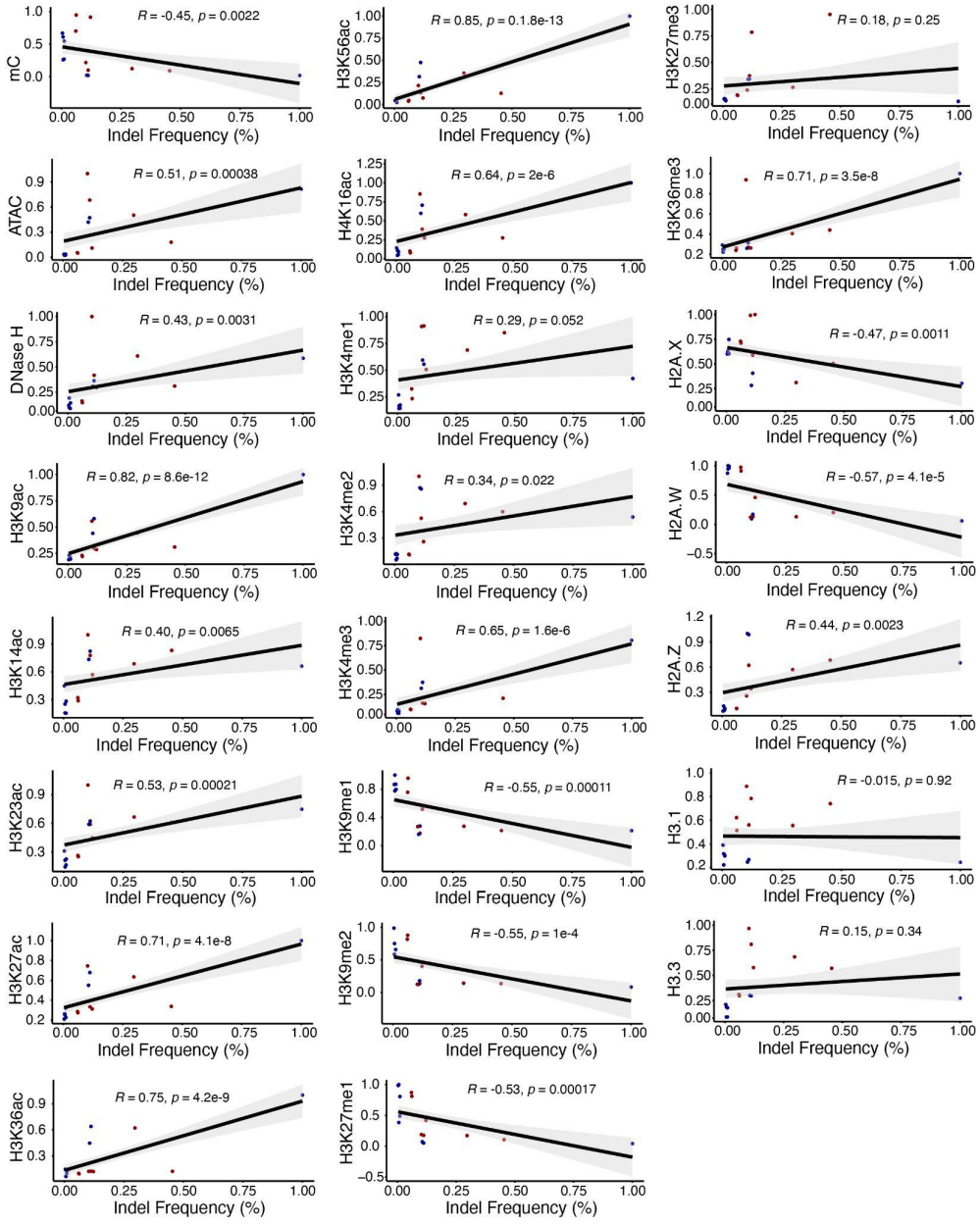


Figure S4. Correlation analysis for CRISPR-Cas9 mutagenesis frequencies and chromatin features. The blue dots represent MCsite4 and the red dots represent MCsite5. The trendline is black with gray indicating the standard error. The R value and p-values are indicated at the top of each correlation plot according to Spearman's rank correlation coefficient. Each feature was normalized using all 15 target sites on a scale of 0 to 1, with higher values indicating the higher levels for the respective feature.

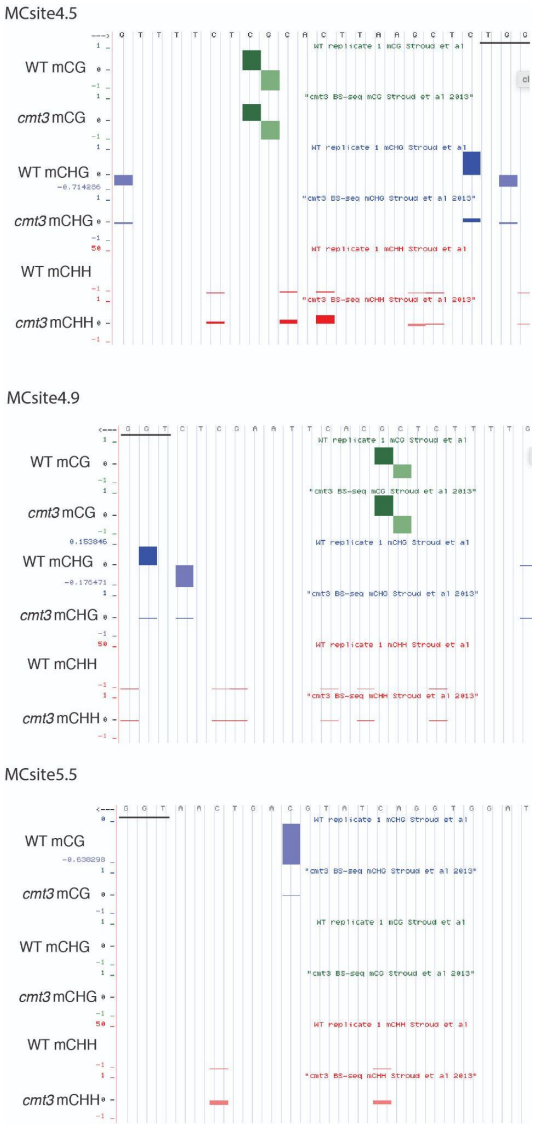


Figure S5. Characterization of the single-based DNA methylation status at MCsite4 and MCsite5 in the wild type and cmt3 mutant plants. Representative screenshots of MCsite4.5, MCsite4.9, and MCsite5.5 from the UCSC genome browser (2). The protospacer sequence is displayed along the top of each screenshot with the PAM underlined. 6 individual methylome tracks (wild type mCG context, cmt3 mCG context, WT mCHG context, cmt3 mCHG context, WT mCHH context, and cmt3 mCHH context) displaying the levels of DNA methylation at each individual nucleotide along the

protospacer and PAM genomic DNA sequence. Green corresponds with mCG, blue with mCHG and red with mCHH.

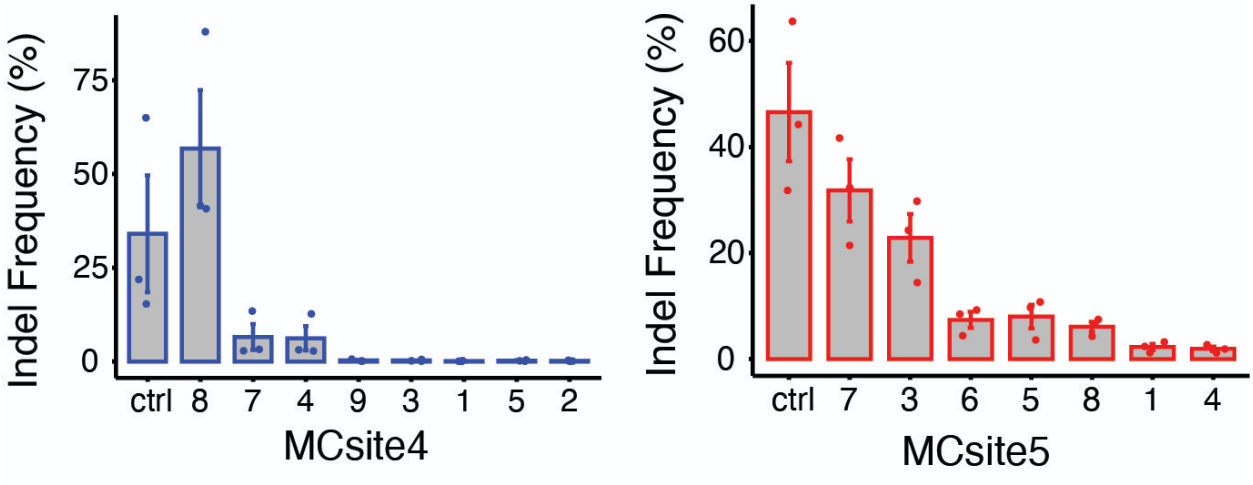


Figure S6. Unnormalized mutagenesis frequencies for MCsite4 (blue and gray) and MCsite5 (red and gray) in the cmt3 mutant plants. Ctrl is abbreviated for the CHL2 site. The standard error (SEM) is displayed for each target site with the dots indicating independent replicates (n = 3).

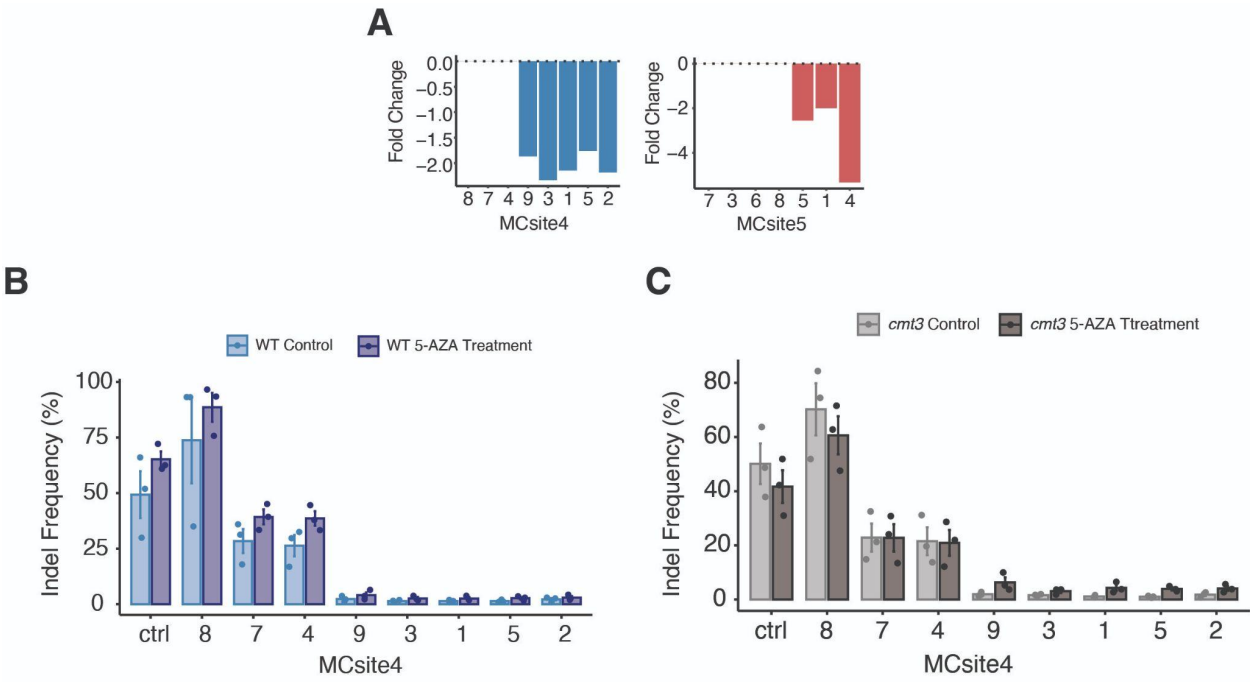


Figure S7. 5-azacytidine treatment of the wild type and cmt3 T2 seedlings. (A) DNA methylation reduction fold changes in a 1kb window (500 bp upstream and downstream from the CRISPR-Cas9 cut site) in the 100µM treated 5-azacytidine samples relative to the mock untreated samples from (3). Fold change was calculated by dividing the WT mock nontreated value by the 5-AZA 100 µM value, and then multiplied by -1. The dotted line at the top of each bar graph represents zero change, and each bar is color coded as either blue (MCsite4) or red (MCsite5). (B) and (C) Unnormalized mutagenesis frequencies for MCsite4 and the CHlL2 control (ctrl) in the wild type plants with and without 5-azacytidine treatment (B), and in the cmt3 mutant plants with and without 5-azacytidine treatment (C). The standard error (SEM) is displayed for each target site with the dots indicating replicates (n = 3).

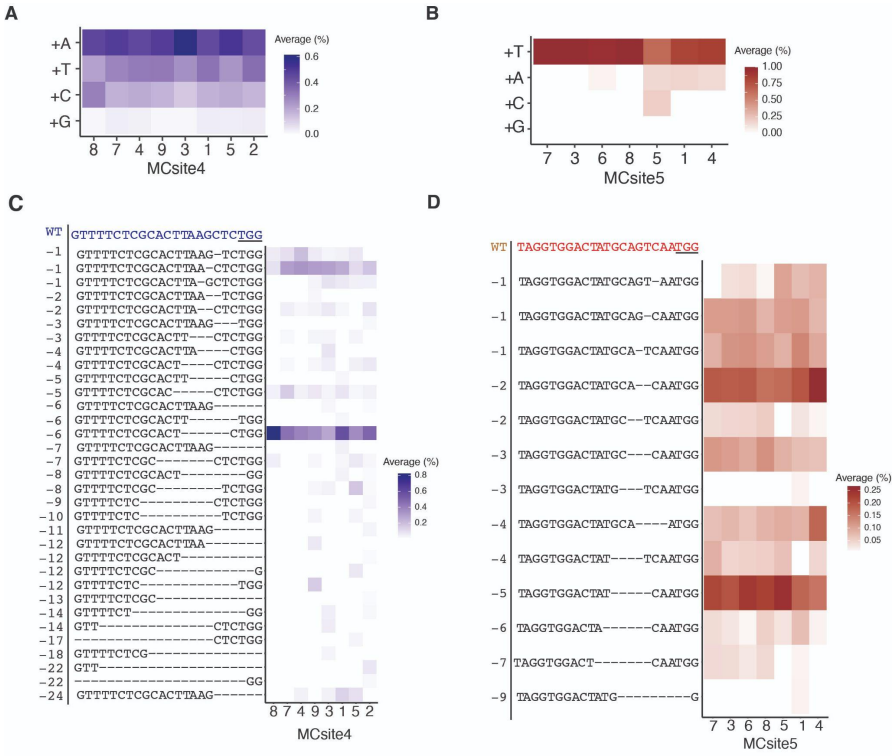


Figure S8. Characterization of mutation outcomes for MCsite4 (blue) and MCsite5 (red). (A) and (B) Heatmap displaying the frequency of 1 bp insertions that occurred at each MCsite4 (A) and MCsite5 (B) sites in wild type plants. In the MCsite4 sites, the majority of 1 bp insertions (A, T or C) was derived from template-independent DNA polymerase-mediated end filling. In the MCsite5 sites, the majority of 1 bp insertions (T) was derived from templated-dependent end filling. (C) and (D) Heatmap displaying the frequency of deletion outcome that occurred at each MCsite4 (C) and MCSite5 (D) sites in wild type plants. The wild type sequence is at the top of the y-axis with the PAM underlined. The number to the left of each deletion outcome indicates the size of the deletion. The frequency of each repair outcome was calculated by using the total number of reads with insertions or deletions divided by the total number of mutated reads for

each site. This was done for all three replicates. The average of the three replicates was then calculated and plotted as a heatmap.

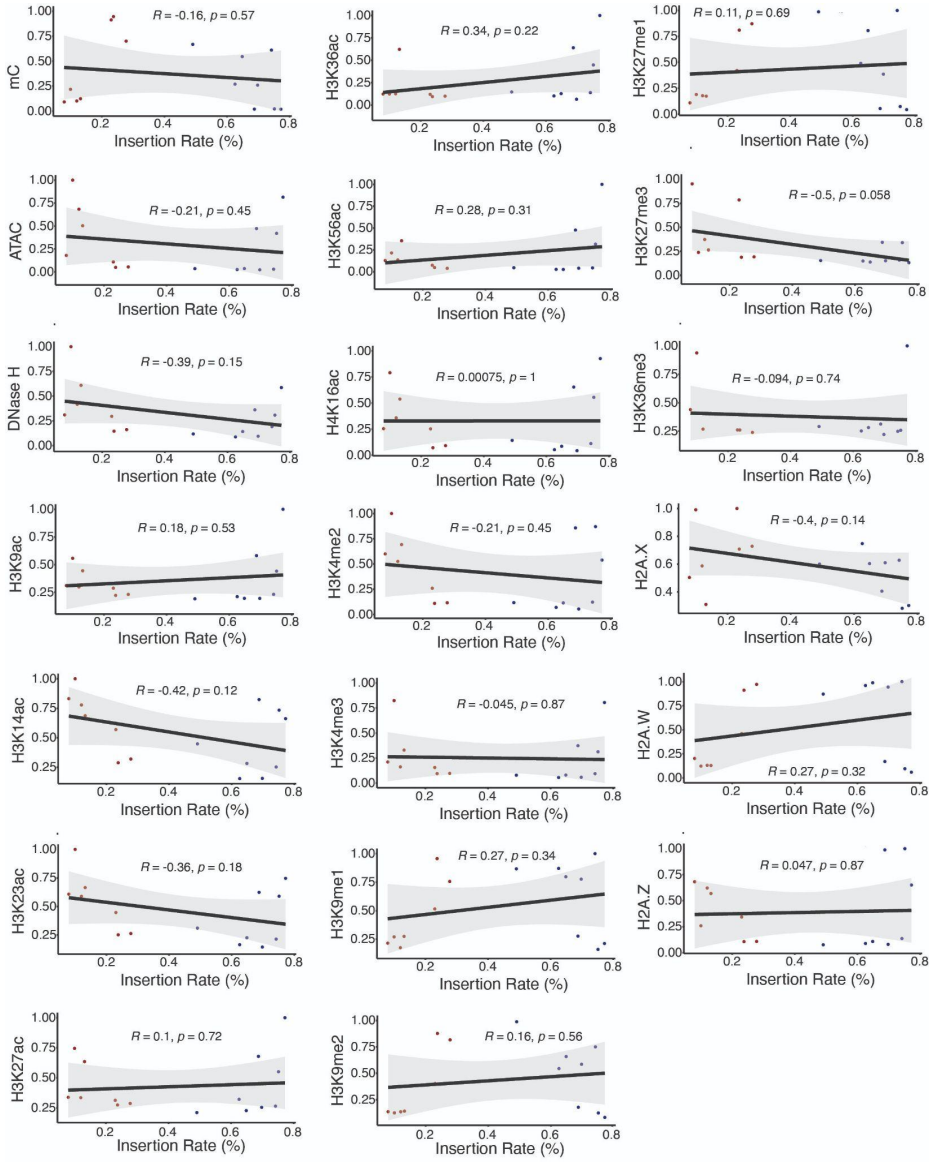


Figure S9. Correlation analysis for 1 bp insertion rate and chromatin features. The blue dots represent MCsite4, and the red dots represent MCsite5. The trendline is black with gray indicating the standard error. The R value and p-values are indicated at the top of each correlation plot according to Spearman's rank correlation coefficient. Each chromatin feature was normalized using all 15 target sites to allow for comparison

between the CRISPR target sites on a scale of 0 to 1, with higher values indicating the higher levels for the respective chromatin feature.

Supplemental Tables

Supplemental Table S1. Primer sequences to amplify each CRISPR target site analyzed in these experiments.

target site	primer name	primer sequence
CHLI2	119_F2_CHLI2	GTCCCATCTCTGCTTCGGAC
	120_R1_CHLI2	CACACCACCTATCTTCGGGT
MCsite4.1	330_F_1_gRNA_5907	ACGAATGGCTTGATCCCTTC
	334_R_1_gRNA_5907	TTCTGGGCTCTGTTTCAAGAATA
MCsite4.2	339_F_2_gRNA_5907	TCCTTGGAATGCTGTTCACC
	343_R_2_gRNA_5907	CCCTTGTAATGGCTACCGATC
MCsite4.3	348_F_3_gRNA_5907	AATACGAACTACGAATGGCTTGA
	351_R_3_gRNA_5907	TCCAGCAGGCTTATGTATGATTT
MCsite4.4	356_F_4_gRNA_5907	ACTGGGTACGTAGGCAATCC
	362_R_4_gRNA_5907	TAAAAGTATTCGGCGTGTACTGG
MCsite4.5	143_F_5_5907	GCAGATCGCCTATCGCAAGG
	374_R_5_gRNA_5907	TGGACTAGATGTTAGGTGCGA
MCsite4.7	146_F_7_5907	AATCCGACCTCTCCGTTAGAACT
	142_R_7_5907	AAAATAAGTGAGGCGGACTCAGCA
MCsite4.8	394_F_8_gRNA_5907	CGGTTTATGTTGATCGGATTTTA
	397_R_8_gRNA_5907	AAAGAGGGAGTTCTGGGCTC
MCsite4.9	403_F_9_gRNA_5907	GTCATTCTCAAGGATGCAGCTAT
	407_R_9_gRNA_5907	AGACTTAGCTCGGCTTTTTCAAT
MCsite5.1	195_F_1_gRNA 6169	AAGTGAGCTAGGCCGTCATT
	196_R_1_gRNA 6169	ACACCCAGAGACAGAGTAGC
MCsite5.3	201_F_3_gRNA 6169	GGCTGCCTACGTACCCTCAGAAA
	202_R_3_gRNA 6169	GCTCGTCTTCAAGCTCTTCCTCAG
MCsite5.4	197_F_4_gRNA 6169	GGTAGGTAGATGTGCGACGA
	198_R_4_gRNA 6169	GGACACCCAGAGAGAGTAGC
MCsite5.5	199_F_5_gRNA 6169	TAGGTCGGCAGGTAGTTGAC
	200_R_5_gRNA 6169	CTCGTTAGCTCGCTCCTCAT
MCsite5.6	203_F_6_gRNA 6169	CTGGATGAACAGTGGTAGAGAGGCG
	204_R_6_gRNA 6169	TCTTCAAGCTCTTCCCATCACTCGA
MCsite5.7	205_F_7_gRNA 6169	AGGTAGGTCGGCAGGTAGGT
	206_R_7_gRNA 6169	TTAGCTCGCTCCTCTTCGTCAAGC
MCsite5.8	207_F_8_gRNA 6169	TAGTCAAGCGAGACGACACGTAG
	208_R_8_gRNA 6169	AGCTCGTCTTCATGCTCCACTCAA

Supplemental Table S2. Summary of NGS reads count for each tested target site. Read counts used to characterize mutagenesis frequency and mutation outcomes were shown in the "edited reads count" column for each site. The numbers were derived from the sum of all three replicates.

Mcsite	CRISPR target	genotype	generation	5- Azacytidine treamtent?	edited reads count
MCsite4	1	WT	T1	No	397
MCsite4	2	WT	T1	No	1106
MCsite4	3	WT	T1	No	1049
MCsite4	4	WT	T1	No	19272
MCsite4	5	WT	T1	No	468
MCsite4	7	WT	T1	No	22338
MCsite4	8	WT	T1	No	253601
MCsite4	9	WT	T1	No	1240
MCsite5	1	WT	T1	No	2495
MCsite5	3	WT	T1	No	21999
MCsite5	4	WT	T1	No	5080
MCsite5	5	WT	T1	No	3772
MCsite5	6	WT	T1	No	6677
MCsite5	7	WT	T1	No	8075
MCsite5	8	WT	T1	No	14552
MCsite4	1	WT	T2	No	440
MCsite4	2	WT	T2	No	1482
MCsite4	3	WT	T2	No	755
MCsite4	4	WT	T2	No	8867
MCsite4	5	WT	T2	No	168
MCsite4	7	WT	T2	No	41693
MCsite4	8	WT	T2	No	88428
MCsite4	9	WT	T2	No	2340
MCsite4	1	WT	T2	Yes	1724
MCsite4	2	WT	T2	Yes	2749
MCsite4	3	WT	T2	Yes	3057
MCsite4	4	WT	T2	Yes	19237
MCsite4	5	WT	T2	Yes	1247
MCsite4	7	WT	T2	Yes	102362
MCsite4	8	WT	T2	Yes	104087
MCsite4	9	WT	T2	Yes	4517
MCsite4	1	CMT3	T1	No	149
MCsite4	2	CMT3	T1	No	152
MCsite4	3	CMT3	T1	No	511
MCsite4	4	CMT3	T1	No	4482
MCsite4	5	CMT3	T1	No	56
MCsite4	7	CMT3	T1	No	15676
MCsite4	8	CMT3	T1	No	94339
MCsite4	9	CMT3	T1	No	240
MCsite5	1	CMT3	T1	No	1795
MCsite5	3	CMT3	T1	No	20003
MCsite5	4	CMT3	T1	No	1991
MCsite5	5	CMT3	T1	No	2117
MCsite5	6	CMT3	T1	No	5290
MCsite5	7	CMT3	T1	No	7377
MCsite5	8	CMT3	T1	No	10643
MCsite4	1	CMT3	T2	No	1277
MCsite4	2	CMT3	T2	No	2138
MCsite4	3	CMT3	T2	No	2709
MCsite4	4	CMT3	T2	No	19925
MCsite4	5	CMT3	T2	No	661
MCsite4	7	CMT3	T2	No	71816
MCsite4	8	CMT3	T2	No	163323
MCsite4	9	CMT3	T2	No	2773
MCsite4	1	CMT3	T2	Yes	2355
MCsite4	2	CMT3	T2	Yes	2555
MCsite4	3	CMT3	T2	Yes	2934
MCsite4	4	CMT3	T2	Yes	10976
MCsite4	5	CMT3	T2	Yes	1539
MCsite4	7	CMT3	T2	Yes	41451
MCsite4	8	CMT3	T2	Yes	129139
MCsite4	9	CMT3	T2	Yes	7043

Supplemental Table S3. Oligos for cloning MCsite and CHLl2 into pMOD_B2301.

Oligo Name	Sequence 5' to 3'	Note
92_TRNA_5907	TCGTCTCCGTGCGAGAAAACTGCACCAGCCGGGAATCG	cloning pMOD_B2301 with MCsite4
93_REP_5907	TCGTCTCAGCACTTAAGCTCGTTTTAGAGCTAGAAATAGC	cloning pMOD_B2301 with MCsite4
94_TRNA_518	TCGTCTCCCCATAATGTTGTTGCACCAGCCGGGAATCG	cloning pMOD_B2301 with MCsite3
95_REP_518	TCGTCTCAATGGACAGTCCAGTTTTAGAGCTAGAAATAGC	cloning pMOD_B2301 with MCsite3
96_TRNA_3186	TCGTCTCCCTGCGTTTTGCGTGCACCAGCCGGGAATCG	cloning pMOD_B2301 with MCsite1
97_REP_3186	TCGTCTCAGCAGTCACCATTGTTTTAGAGCTAGAAATAGC	cloning pMOD_B2301 with MCsite1
98_TRNA_3606	TCGTCTCCCTTAAGTGCGAGTGCACCAGCCGGGAATCG	cloning pMOD_B2301 with MCsite2
99_REP_3606	TCGTCTCATAAGCTCTGGTTGTTTTAGAGCTAGAAATAGC	cloning pMOD_B2301 with MCsite2
100_TRNA_715	TCGTCTCCAAGTGTTCGGGTTGCACCAGCCGGGAATCG	cloning pMOD_B2301 with MCsite7
101_REP_715	TCGTCTCAACTTTTCCGGTTGTTTTAGAGCTAGAAATAGC	cloning pMOD_B2301 with MCsite7
102_TRNA_6169	TCGTCTCCATAGTCCACCTATGCACCAGCCGGGAATCG	cloning pMOD_B2301 with MCsite5
103_REP_6169	TCGTCTCACTATGCAGTCAAGTTTTAGAGCTAGAAATAGC	cloning pMOD_B2301 with MCsite5
104_TRNA_6334	TCGTCTCCTTCGTCCATTGATGCACCAGCCGGGAATCG	cloning pMOD_B2301 with MCsite6
105_REP_6334	TCGTCTCACGAACTGCCGCTGTTTTAGAGCTAGAAATAGC	cloning pMOD_B2301 with MCsite6
108_TRNA_CHLI2	TCGTCTCCTTATGAATGTCGTGCACCAGCCGGGAATCG	cloning pMOD_B2301 with CHLI2 and an Mcsite
109_REP_CHLI2	TCGTCTCAATAACAGAGACAGTTTTAGAGCTAGAAATAGC	cloning pMOD_B2301 with CHLI2 and an Mcsite
112_TRNA_term	TGCTCTTCTGACTGCACCAGCCGGGAATCG	universal oligo for cloning pMOD_B2301
113_o35s_prom	TGCTCTTCGCGCATGGAGTCAAAGATTCAA	universal oligo for cloning pMOD_B2301

Dataset S1. Characterization of the sequences, DNA methylation, chromatin accessibility, and chromatin states for the 7,971 candidate CRISPR target sites identified. Data can be retrieved at https://doi.org/10.1093/plphys/kiac285.

Dataset S2. Characterization of the 7 candidate MCsite chromosomal locations, DNA methylation domain, chromatin accessibility, chromatin state, gene annotation, and RNA detection. Data can be retrieved at https://doi.org/10.1093/plphys/kiac285.

Dataset S3. Characterization of the 1 kb region (500 bp upstream and 500 bp downstream) flanking the CRISPR-Cas9 cut site. For each dataset, the values for each individual nucleotide flanking the CRISPR cut site in the 1kb window (500 bp upstream and 500 bp downstream) were quantified by calculating the sum. Data can be retrieved at https://doi.org/10.1093/plphys/kiac285.

Dataset S4. Index for matching samples with the correct fastq files.

Conclusion

Throughout my dissertation I have described new insights and approaches to improve CRISPR-Cas9 genome editing in plants. I sought to understand the rules that govern CRISPR-Cas9 genome editing, and utilize this knowledge for improved editing efficacy. The following paragraphs will highlight the primary takeaways from each chapter, and explore additional avenues to better understand CRISPR-Cas9 genome editing.

Chapter I describes the optimization of a multiplexed gene editing and transformation pipeline to rapidly create genetically modified *Setaria viridis* plants. First, a robust protoplast transformation pipeline was developed to quickly test editing reagents. Using the protoplast assay, we demonstrated that coexpression of TREX2 can increase editing 1.4-fold. Additionally, TREX2 alters the double strand break repair profile, resulting in essentially no insertions and a large diversity of deletion outcomes. Closer examination of these deletion edits revealed a substantial shift in the DNA repair pathways used for repair. The presence of TREX2 strongly shifted the repair pathway from NHEJ to MMEJ. The Cas9_TREX2 system was then used for tissue culture transformation, resulting in the successful creation of transgenic plants. Plants with high levels of editing were advanced to the next generation, where we recovered a transgene-free plant with frameshift CRISPR-Cas9-mediated indels. This optimized multiplexed editing and transformation pipeline will enable the creation of new gene knockout materials for research. Further, the TREX2 system has the ability to create a large diversity of deletion repair outcomes, which may be beneficial for creating sequence variation in cis regulatory elements.

Chapter II focuses on the characterization of the epigenome and transcriptome of the *drm1ab* plant created in Chapter I. As expected, we observed a significant genome-wide decrease in the levels of CHH methylation. Further analysis of the methylome revealed two unique observations. First, CHG methylation was rarely maintained at loci that had lost CHH methylation. Second, we found numerous examples of CG and/or CHG methylation loss in regions that were not located at or near loci with high CHH methylation. Transcriptome analysis revealed relatively few changes in gene expression in *drm1ab* plants, suggesting a complex interaction between DNA methylation and gene expression. Future efforts will be aimed towards creating additional epigenetic mutants, such as *cmt*, *met1*, and *ddm1*. These genetic materials can then be leveraged to further dissect the relationship between DNA methylation and gene expression in a monocot model. Additionally, the isolation of additional epigenetic mutants can be a powerful resource to map epialleles.

In Chapter III, I characterized the influence of epigenetic features on CRISPR-Cas9 efficacy. Comparison of editing efficiencies and repair outcomes at identical target sites yet in diverse epigenetic contexts revealed that epigenetic features can significantly impact CRISPR-Cas9 genome editing. High editing efficiencies were associated with unmethylated and accessible chromatin, while low efficiencies correlated with hypermethylated heterochromatin. Further, we demonstrated that substantial reduction in DNA methylation using *cmt3* and 5-azacytidine led to increased editing at the lowly edited sites located in highly methylated heterochromatic regions. Lastly, correlation analysis revealed differences in the insertion rate that are associated with distinct histone markers. Future studies will aim to determine which specific features

dictate DNA repair outcomes after a CRISPR-Cas9-mediated double strand break. Knowledge gleaned from such studies can be leveraged to manipulate DNA repair outcomes to favor a specific desired mutation.

BIBLIOGRAPHY

Akakpo, R., Carpentier, M.-C., Ie Hsing, Y., & Panaud, O. (2020). The impact of transposable elements on the structure, evolution and function of the rice genome. The New Phytologist, 226(1), 44–49.

Alleman, M., Sidorenko, L., McGinnis, K., Seshadri, V., Dorweiler, J. E., White, J., Sikkink, K., & Chandler, V. L. (2006). An RNA-dependent RNA polymerase is required for paramutation in maize. Nature, 442(7100), 295–298.

Allen, F., Crepaldi, L., Alsinet, C., Strong, A. J., Kleshchevnikov, V., De Angeli, P., Páleníková, P., Khodak, A., Kiselev, V., Kosicki, M., Bassett, A. R., Harding, H., Galanty, Y., Muñoz-Martínez, F., Metzakopian, E., Jackson, S. P., & Parts, L. (2018). Predicting the mutations generated by repair of Cas9-induced double-strand breaks. Nature Biotechnology. https://doi.org/10.1038/nbt.4317

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. Journal of Molecular Biology, 215(3), 403–410.

Ata, H., Ekstrom, T. L., Martínez-Gálvez, G., Mann, C. M., Dvornikov, A. V., Schaefbauer, K. J., Ma, A. C., Dobbs, D., Clark, K. J., & Ekker, S. C. (2018). Robust activation of microhomology-mediated end joining for precision gene editing applications. PLoS Genetics, 14(9), e1007652.

Bae, S., Kweon, J., Kim, H. S., & Kim, J.-S. (2014). Microhomology-based choice of Cas9 nuclease target sites. Nature Methods, 11(7), 705–706.

Bennetzen, J. L., Schmutz, J., Wang, H., Percifield, R., Hawkins, J., Pontaroli, A. C., Estep, M., Feng, L., Vaughn, J. N., Grimwood, J., Jenkins, J., Barry, K., Lindquist, E., Hellsten, U., Deshpande, S., Wang, X., Wu, X., Mitros, T., Triplett, J., … Devos, K. M. (2012). Reference genome sequence of the model plant Setaria. Nature Biotechnology, 30(6), 555–561.

Brutnell, T. P., Wang, L., Swartwood, K., Goldschmidt, A., Jackson, D., Zhu, X.-G., Kellogg, E., & Van Eck, J. (2010). Setaria viridis: a model for C4 photosynthesis. The Plant Cell, 22(8), 2537–2544.

Cao, X., Aufsatz, W., Zilberman, D., Mette, M. F., Huang, M. S., Matzke, M., & Jacobsen, S. E. (2003). Role of the DRM and CMT3 methyltransferases in RNA-directed DNA methylation. Current Biology : CB, 13(24), 2212–2217.

Cao, X., & Jacobsen, S. (2002a). Role of the Arabidopsis DRM Methyltransferases in De Novo DNA Methylation and Gene Silencing. Current Biology: CB, 12(13), 1138–1144.

Cao, X., & Jacobsen, S. (2002b). Locus-specific control of asymmetric and CpNpG methylation by the DRM and CMT3 methyltransferase genes. Proceedings of the National Academy of Sciences of the United States of America, 99(90004), 16491–16498.

Cao, X., Springer, N., Muszynski, M., Phillips, R., Kaeppler, S., & Jacobsen, S. (2000). Conserved plant genes with similarity to mammalian de novo DNA methyltransferases. Proceedings of the National Academy of Sciences of the United States of America, 97(9), 4979–4984.

Capella-Gutiérrez, S., Silla-Martínez, J. M., & Gabaldón, T. (2009). trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics , 25(15), 1972–1973.
Casacuberta, J. M., Vernhettes, S., Audeon, C., & Grandbastien, M. A. (1997). Quasispecies in retrotransposons: a role for sequence variability in Tnt1 evolution. Genetica, 100(1-3), 109–117.

Čermák, T., Curtin, S. J., Gil-Humanes, J., Čegan, R., Kono, T. J. Y., Konečná, E., Belanto, J. J., Starker, C. G., Mathre, J. W., Greenstein, R. L., & Voytas, D. F. (2017). A Multipurpose Toolkit to Enable Advanced Genome Engineering in Plants. In The Plant Cell (Vol. 29, Issue 6, pp. 1196–1217). https://doi.org/10.1105/tpc.16.00922

Chan, S. W., Henderson, I. R., Zhang, X., Shah, G., Chien, J. S., & Jacobsen, S. E. (2006). RNAi, DRD1, and histone methylation actively target developmentally important non-CG DNA methylation in arabidopsis. PLoS Genetics, 2(6), e83.

Chan, S. W., Zilberman, D., Xie, Z., Johansen, L. K., Carrington, J. C., & Jacobsen, S. E. (2004). RNA silencing genes control de novo DNA methylation. Science , 303(5662), 1336.

Chari, R., Mali, P., Moosburner, M., & Church, G. M. (2015). Unraveling CRISPR-Cas9 genome engineering parameters via a library-on-library approach. Nature Methods, 12(9), 823–826.

Chen, K., Wang, Y., Zhang, R., Zhang, H., & Gao, C. (2019). CRISPR/Cas Genome Editing and Precision Plant Breeding in Agriculture. Annual Review of Plant Biology, 70, 667–697.

Clement, K., Rees, H., Canver, M. C., Gehrke, J. M., Farouni, R., Hsu, J. Y., Cole, M. A., Liu, D. R., Joung, J. K., Bauer, D. E., & Pinello, L. (2019). CRISPResso2 provides accurate and rapid genome editing sequence analysis. Nature Biotechnology, 37(3), 224–226.

Collier, R., Thomson, J. G., & Thilmony, R. (2018). A versatile and robust Agrobacterium-based gene stacking system generates high-quality transgenic Arabidopsis plants. The Plant Journal: For Cell and Molecular Biology, 95(4), 573–583.

Concordet, J.-P., & Haeussler, M. (2018). CRISPOR: intuitive guide selection for CRISPR/Cas9 genome editing experiments and screens. Nucleic Acids Research, 46(W1), W242–W245.
Costa-Nunes, P., Kim, J. Y., Hong, E., & Pontes, O. (2014). The cytological and molecular role of domains rearranged methyltransferase3 in RNA-dependent DNA methylation of Arabidopsis thaliana. BMC Research Notes, 7, 721.

Crisp, P. A., Ganguly, D. R., Smith, A. B., Murray, K. D., Estavillo, G. M., Searle, I., Ford, E., Bogdanović, O., Lister, R., Borevitz, J. O., Eichten, S. R., & Pogson, B. J. (2017). Rapid Recovery Gene Downregulation during Excess-Light Stress and Recovery in Arabidopsis. The Plant Cell, 29(8), 1836–1863.

Crisp, P. A., Marand, A. P., Noshay, J. M., Zhou, P., Lu, Z., Schmitz, R. J., & Springer, N. M. (n.d.). Stable unmethylated DNA demarcates expressed genes and their cis-regulatory space in plant genomes. https://doi.org/10.1101/2020.05.21.109744

Crisp, P. A., Marand, A. P., Noshay, J. M., Zhou, P., Lu, Z., Schmitz, R. J., & Springer, N. M. (2020). Stable unmethylated DNA demarcates expressed genes and their cis-regulatory space in plant genomes. Proceedings of the National Academy of Sciences of the United States of America, 117(38), 23991–24000.

Daer, R. M., Cutts, J. P., Brafman, D. A., & Haynes, K. A. (2017). The Impact of Chromatin Dynamics on Cas9-Mediated Genome Editing in Human Cells. ACS Synthetic Biology, 6(3), 428–438.

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

Defelice, M. S. (2002). Green Foxtail, Setaria viridis (L.) P. Beauv. Weed Technology: A Journal of the Weed Science Society of America, 16(1), 253–257.

Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., & Gingeras, T. R. (2013). STAR: ultrafast universal RNA-seq aligner. Bioinformatics , 29(1), 15–21.

Du, J., Zhong, X., Bernatavichute, Y. V., Stroud, H., Feng, S., Caro, E., Vashisht, A. A., Terragni, J., Chin, H. G., Tu, A., Hetzel, J., Wohlschlegel, J. A., Pradhan, S., Patel, D. J., & Jacobsen, S. E. (2012). Dual binding of chromomethylase domains to H3K9me2-containing nucleosomes directs DNA methylation in plants. Cell, 151(1), 167–180.

Edgar, R. C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Research, 32(5), 1792–1797.

Fnu, S., Williamson, E. A., De Haro, L. P., Brenneman, M., Wray, J., Shaheen, M., Radhakrishnan, K., Lee, S.-H., Nickoloff, J. A., & Hromas, R. (2011). Methylation of histone H3 lysine 36 enhances DNA repair by nonhomologous end-joining. Proceedings of the National Academy of Sciences of the United States of America, 108(2), 540–545.

Gent, J. I., Ellis, N. A., Guo, L., Harkess, A. E., Yao, Y., Zhang, X., & Dawe, R. K. (2013). CHH islands: de novo DNA methylation in near-gene chromatin regulation in maize. Genome Research, 23(4), 628–637.

Gisler, S., Gonçalves, J. P., Akhtar, W., de Jong, J., Pindyurin, A. V., Wessels, L. F. A., & van Lohuizen, M. (2019). Multiplexed Cas9 targeting reveals genomic location effects and gRNA-based staggered breaks influencing mutation efficiency. Nature Communications, 10(1), 1598.

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

Griffin, P. T., Niederhuth, C. E., & Schmitz, R. J. (2016). A Comparative Analysis of 5-Azacytidine- and Zebularine-Induced DNA Demethylation. G3 , 6(9), 2773–2780.

Haas, B. J., Papanicolaou, A., Yassour, M., Grabherr, M., Blood, P. D., Bowden, J., Couger, M. B., Eccles, D., Li, B., Lieber, M., MacManes, M. D., Ott, M., Orvis, J., Pochet, N., Strozzi, F., Weeks, N., Westerman, R., William, T., Dewey, C. N., … Regev, A. (2013). De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. Nature Protocols, 8(8), 1494–1512.

Haeussler, M., Schönig, K., Eckert, H., Eschstruth, A., Mianné, J., Renaud, J.-B., Schneider-Maunoury, S., Shkumatava, A., Teboul, L., Kent, J., Joly, J.-S., & Concordet, J.-P. (2016). Evaluation of off-target and on-target scoring algorithms and integration into the guide RNA selection tool CRISPOR. Genome Biology, 17(1), 148.

Han, Z., Crisp, P. A., Stelpflug, S., Kaeppler, S. M., Li, Q., & Springer, N. M. (2018). Heritable Epigenomic Changes to the Maize Methylome Resulting from Tissue Culture. Genetics. https://doi.org/10.1534/genetics.118.300987

Henderson, I. R., Deleris, A., Wong, W., Zhong, X., Chin, H. G., Horwitz, G. A., Kelly, K. A., Pradhan, S., & Jacobsen, S. E. (2010). The de novo cytosine methyltransferase DRM2 requires intact UBA domains and a catalytically mutated paralog DRM3 during RNA-directed DNA methylation in Arabidopsis thaliana. PLoS Genetics, 6(10), e1001182.

Henderson, I. R., & Jacobsen, S. E. (2008). Tandem repeats upstream of the Arabidopsis endogene SDC recruit non-CG DNA methylation and initiate siRNA spreading. Genes & Development, 22(12), 1597–1606.

Huang, P., Mamidi, S., Healey, A., Grimwood, J., Jenkins, J., Barry, K., Sreedasyam, A., Shu, S., Feldman, M., Wu, J., Yu, Y., Chen, C., Johnson, J., Sakakibara, H., Kiba, T., Sakurai, T., Rokhsar, D., Baxter, I., Schmutz, J., … Kellogg, E. A. (2019). The Setaria viridis genome and diversity panel enables discovery of a novel domestication gene. In bioRxiv (p. 744557). https://doi.org/10.1101/744557

Hu, D., Yu, Y., Wang, C., Long, Y., Liu, Y., Feng, L., Lu, D., Liu, B., Jia, J., Xia, R., Du, J., Zhong, X., Gong, L., Wang, K., & Zhai, J. (2021). Multiplex CRISPR-Cas9 editing of DNA methyltransferases in rice uncovers a class of non-CG methylation specific for GC-rich regions. The Plant Cell, 33(9), 2950–2964.

Jacquet, K., Fradet-Turcotte, A., Avvakumov, N., Lambert, J.-P., Roques, C., Pandita, R. K., Paquet, E., Herst, P., Gingras, A.-C., Pandita, T. K., Legube, G., Doyon, Y., Durocher, D., & Côté, J. (2016). The TIP60 Complex Regulates Bivalent Chromatin Recognition by 53BP1 through Direct H4K20me Binding and H2AK15 Acetylation. Molecular Cell, 62(3), 409–421.
Jinek, M., Chylinski, K., Fonfara, I., Hauer, M., Doudna, J. A., & Charpentier, E. (2012). A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. Science, 337(6096), 816–821.

Johnson, L. M., Bostick, M., Zhang, X., Kraft, E., Henderson, I., Callis, J., & Jacobsen, S. E. (2007). The SRA methyl-cytosine-binding domain links DNA and histone methylation. Current Biology : CB, 17(4), 379–384.

Kallimasioti-Pazi, E. M., Thelakkad Chathoth, K., Taylor, G. C., Meynert, A., Ballinger, T., Kelder, M. J. E., Lalevée, S., Sanli, I., Feil, R., & Wood, A. J. (2018). Heterochromatin delays CRISPR-Cas9 mutagenesis but does not influence the outcome of mutagenic DNA repair. PLoS Biology, 16(12), e2005595.

Lata, C., Gupta, S., & Prasad, M. (2013). Foxtail millet: a model crop for genetic and genomic studies in bioenergy grasses. Critical Reviews in Biotechnology, 33(3), 328–343.
Law, J. A., & Jacobsen, S. E. (2010). Establishing, maintaining and modifying DNA methylation patterns in plants and animals. Nature reviews.Genetics, 11(3), 204–220.

Lazzarotto, C. R., Malinin, N. L., Li, Y., Zhang, R., Yang, Y., Lee, G., Cowley, E., He, Y., Lan, X., Jividen, K., Katta, V., Kolmakova, N. G., Petersen, C. T., Qi, Q., Strelcov, E., Maragh, S., Krenciute, G., Ma, J., Cheng, Y., & Tsai, S. Q. (2020). CHANGE-seq reveals genetic and epigenetic effects on CRISPR–Cas9 genome-wide activity. Nature Biotechnology, 38(11), 1317–1327.

Lemos, B. R., Kaplan, A. C., Bae, J. E., Ferrazzoli, A. E., Kuo, J., Anand, R. P., Waterman, D. P., & Haber, J. E. (2018). CRISPR/Cas9 cleavages in budding yeast reveal templated insertions and strand-specific insertion/deletion profiles. Proceedings of the National Academy of Sciences of the United States of America, 115(9), E2040–E2047.

Li, J., Stoddard, T. J., Demorest, Z. L., Lavoie, P.-O., Luo, S., Clasen, B. M., Cedrone, F., Ray, E. E., Coffman, A. P., Daulhac, A., & Others. (2016). Multiplexed, targeted gene editing in Nicotiana benthamiana for glyco-engineering and monoclonal antibody production. Plant Biotechnology Journal, 14(2), 533–542.

Li, Q., Gent, J. I., Zynda, G., Song, J., Makarevitch, I., Hirsch, C. D., Hirsch, C. N., Dawe, R. K., Madzima, T. F., McGinnis, K. M., Lisch, D., Schmitz, R. J., Vaughn, M. W., & Springer, N. M. (2015). RNA-directed DNA methylation enforces boundaries between heterochromatin and euchromatin in the maize genome. Proceedings of the National Academy of Sciences of the United States of America, 112(47), 14728–14733.

Liu, G., Yin, K., Zhang, Q., Gao, C., & Qiu, J.-L. (2019). Modulating chromatin accessibility by transactivation and targeting proximal dsgRNAs enhances Cas9 editing efficiency in vivo. Genome Biology, 20(1), 145.

Liu, H., Jian, L., Xu, J., Zhang, Q., Zhang, M., Jin, M., Peng, Y., Yan, J., Han, B., Liu, J., Gao, F., Liu, X., Huang, L., Wei, W., Ding, Y., Yang, X., Li, Z., Zhang, M., Sun, J., … Yan, J. (2020). High-Throughput CRISPR/Cas9 Mutagenesis Streamlines Trait Gene Identification in Maize. The Plant Cell. https://doi.org/10.1105/tpc.19.00934

Liu, Y., Tian, T., Zhang, K., You, Q., Yan, H., Zhao, N., Yi, X., Xu, W., & Su, Z. (2018). PCSD: a plant chromatin state database. Nucleic Acids Research, 46(D1), D1157–D1167.
Love, M. I., Huber, W., & Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biology, 15(12), 550.

Luijsterburg, M. S., de Krijger, I., Wiegant, W. W., Shah, R. G., Smeenk, G., de Groot, A. J. L., Pines, A., Vertegaal, A. C. O., Jacobs, J. J. L., Shah, G. M., & van Attikum, H. (2016). PARP1 Links CHD2-Mediated Chromatin Expansion and H3.3 Deposition to DNA Repair by Non-homologous End-Joining. Molecular Cell, 61(4), 547–562.

Lu, S., Wang, J., Chitsaz, F., Derbyshire, M. K., Geer, R. C., Gonzales, N. R., Gwadz, M., Hurwitz, D. I., Marchler, G. H., Song, J. S., Thanki, N., Yamashita, R. A., Yang, M., Zhang, D., Zheng, C., Lanczycki, C. J., & Marchler-Bauer, A. (2020). CDD/SPARCLE: the conserved domain database in 2020. Nucleic Acids Research, 48(D1), D265–D268.

Lu, Z., Hofmeister, B. T., Vollmers, C., DuBois, R. M., & Schmitz, R. J. (2016). Combining ATAC-seq with nuclei sorting for discovery of cis-regulatory regions in plant genomes. Nucleic Acids Research, 45(6), e41–e41.

Mamidi, S., Healey, A., Huang, P., Grimwood, J., Jenkins, J., Barry, K., Sreedasyam, A., Shu, S., Lovell, J. T., Feldman, M., Wu, J., Yu, Y., Chen, C., Johnson, J., Sakakibara, H., Kiba, T., Sakurai, T., Tavares, R., Nusinow, D. A., … Kellogg, E. A. (2020). A genome resource for green millet Setaria viridis enables discovery of agronomically valuable loci. Nature Biotechnology, 38(10), 1203–1210.

Mao, Y., Zhang, H., Xu, N., Zhang, B., Gou, F., & Zhu, J.-K. (2013). Application of the CRISPR-Cas system for efficient genome engineering in plants. Molecular Plant, 6(6), 2008–2011.

Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet.journal, 17(1), 10–12.

Matzke, M. A., & Mosher, R. A. (2014). RNA-directed DNA methylation: an epigenetic pathway of increasing complexity. Nature reviews.Genetics, 15(6), 394–408.

Mazur, D. J., & Perrino, F. W. (1999). Identification and Expression of the TREX1 and TREX2 cDNA Sequences Encoding Mammalian 3′→5′ Exonucleases. In Journal of Biological Chemistry (Vol. 274, Issue 28, pp. 19655–19660). https://doi.org/10.1074/jbc.274.28.19655

Meng, X., Yu, H., Zhang, Y., Zhuang, F., Song, X., Gao, S., Gao, C., & Li, J. (2017). Construction of a Genome-Wide Mutant Library in Rice Using CRISPR/Cas9. Molecular Plant, 10(9), 1238–1241.

Minkenberg, B., Wheatley, M., & Yang, Y. (2017). CRISPR/Cas9-Enabled Multiplex Genome Editing and Its Application. Progress in Molecular Biology and Translational Science, 149, 111–132.

Moritoh, S., Eun, C.-H., Ono, A., Asao, H., Okano, Y., Yamaguchi, K., Shimatani, Z., Koizumi, A., & Terada, R. (2012). Targeted disruption of an orthologue of DOMAINS REARRANGED METHYLASE 2, OsDRM2, impairs the growth of rice plants by abnormal DNA methylation. The Plant Journal: For Cell and Molecular Biology, 71(1), 85–98.

Naim, F., Shand, K., Hayashi, S., O'Brien, M., McGree, J., Johnson, A. A. T., Dugdale, B., & Waterhouse, P. M. (2020). Are the current gRNA ranking prediction algorithms useful for genome editing in plants? PloS One, 15(1), e0227994.

Nguyen, D. Q., Van Eck, J., Eamens, A. L., & Grof, C. P. L. (2020). Robust and reproducible Agrobacterium-mediated transformation system of the C4 genetic model species Setaria viridis. Frontiers in Plant Science, 11, 281.

Niederhuth, C. E., Bewick, A. J., Ji, L., Alabady, M. S., Kim, K. D., Page, J. T., Li, Q., Rohr, N. A., Rambani, A., Burke, J. M., Udall, J. A., Egesi, C., Schmutz, J., Grimwood, J., Jackson, S. A., Springer, N. M., & Schmitz, R. J. (2016). Widespread natural variation

of DNA methylation within angiosperms. Genome Biology, http://dx.doi.org/10.1101/045880(1), 194.

Ou, S., Su, W., Liao, Y., Chougule, K., Agda, J. R. A., Hellinga, A. J., Lugo, C. S. B., Elliott, T. A., Ware, D., Peterson, T., Jiang, N., Hirsch, C. N., & Hufford, M. B. (2019). Benchmarking transposable element annotation methods for creation of a streamlined, comprehensive pipeline. Genome Biology, 20(1), 275.

Park, Y., & Wu, H. (2016). Differential methylation analysis for BS-seq data under general experimental design. Bioinformatics , 32(10), 1446–1453.

Patro, R., Duggal, G., Love, M. I., Irizarry, R. A., & Kingsford, C. (2017). Salmon provides fast and bias-aware quantification of transcript expression. Nature Methods, 14(4), 417–419.

Qi, Y., Zhang, Y., Zhang, F., Baller, J. A., Cleland, S. C., Ryu, Y., Starker, C. G., & Voytas, D. F. (2013). Increasing frequencies of site-specific mutagenesis and gene targeting in Arabidopsis by manipulating DNA repair pathways. In Genome Research (Vol. 23, Issue 3, pp. 547–554). https://doi.org/10.1101/gr.145557.112

Quinlan, A. R., & Hall, I. M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics , 26(6), 841–842.

R Core Team. (2021). R: A language and environment for statistical computing. R Foundation for Statistical Computing.

Rodríguez-Leal, D., Lemmon, Z. H., Man, J., Bartlett, M. E., & Lippman, Z. B. (2017). Engineering Quantitative Trait Variation for Crop Improvement by Genome Editing. Cell, 171(2), 470–480.e8.

Roudier, F., Ahmed, I., Bérard, C., Sarazin, A., Mary-Huard, T., Cortijo, S., Bouyer, D., Caillieux, E., Duvernois-Berthet, E., Al-Shikhley, L., Giraut, L., Després, B., Drevensek, S., Barneche, F., Dèrozier, S., Brunaud, V., Aubourg, S., Schnittger, A., Bowler, C., … Colot, V. (2011). Integrative epigenomic mapping defines four main chromatin states in Arabidopsis. The EMBO Journal, 30(10), 1928–1938.

Schep, R., Brinkman, E. K., Leemans, C., Vergara, X., van der Weide, R. H., Morris, B., van Schaik, T., Manzo, S. G., Peric-Hupkes, D., van den Berg, J., Beijersbergen, R. L., Medema, R. H., & van Steensel, B. (2021). Impact of chromatin context on Cas9-induced

DNA double-strand break repair pathway balance. Molecular Cell, 81(10), 2216–2230.e10.

Schmid-Burgk, J. L., Gao, L., Li, D., Gardner, Z., Strecker, J., Lash, B., & Zhang, F. (2020). Highly Parallel Profiling of Cas9 Variant Specificity. Molecular Cell, 78(4), 794–800.e8.

Schnable, P. S., Ware, D., Fulton, R. S., Stein, J. C., Wei, F., Pasternak, S., Liang, C., Zhang, J., Fulton, L., Graves, T. A., Minx, P., Reily, A. D., Courtney, L., Kruchowski, S. S., Tomlinson, C., Strong, C., Delehaunty, K., Fronick, C., Courtney, B., … Wilson, R. K. (2009). The B73 maize genome: complexity, diversity, and dynamics. Science , 326(5956), 1112–1115.

Sebastian, J., Wong, M. K., Tang, E., & Dinneny, J. R. (2014). Methods to promote germination of dormant Setaria viridis seeds. PloS One, 9(4), e95109.

Seberg, O., & Petersen, G. (2009). A unified classification system for eukaryotic transposable elements should reflect their phylogeny [Review of A unified classification system for eukaryotic transposable elements should reflect their phylogeny]. Nature Reviews. Genetics, 10(4), 276.

Sequeira-Mendes, J., Aragüez, I., Peiró, R., Mendez-Giraldez, R., Zhang, X., Jacobsen, S. E., Bastolla, U., & Gutierrez, C. (2014). The Functional Topography of the Arabidopsis Genome Is Organized in a Reduced Number of Linear Motifs of Chromatin States. The Plant Cell, 26(6), 2351–2366.

Shen, W., Le, S., Li, Y., & Hu, F. (2016). SeqKit: A Cross-Platform and Ultrafast Toolkit for FASTA/Q File Manipulation. PloS One, 11(10), e0163962.

Shiraki, T., & Kawakami, K. (2018). A tRNA-based multiplex sgRNA expression system in zebrafish and its application to generation of transgenic albino fish. In Scientific Reports (Vol. 8, Issue 1). https://doi.org/10.1038/s41598-018-31476-5

Sidorenko, L., Dorweiler, J. E., Cigan, A. M., Arteaga-Vazquez, M., Vyas, M., Kermicle, J., Jurcin, D., Brzeski, J., Cai, Y., & Chandler, V. L. (2009). A dominant mutation in mediator of paramutation2, one of three second-largest subunits of a plant-specific RNA polymerase, disrupts multiple siRNA silencing processes. PLoS Genetics, 5(11), e1000725.

Springer, N. M., Lisch, D., & Li, Q. (2016). Creating Order from Chaos: Epigenome Dynamics in Plants with Complex Genomes. The Plant Cell, 28(2), 314–325.

Springer, N. M., & Schmitz, R. J. (2017). Exploiting induced and natural epigenetic variation for crop improvement. Nature Reviews. Genetics, 18(9), 563–575.

Stamatakis, A. (2014). RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics , 30(9), 1312–1313.

Stroud, H., Ding, B., Simon, S. A., Feng, S., Bellizzi, M., Pellegrini, M., Wang, G. L., Meyers, B. C., & Jacobsen, S. E. (2013). Plants regenerated from tissue culture contain stable epigenome changes in rice. eLife, 2, e00354.

Stroud, H., Do, T., Du, J., Zhong, X., Feng, S., Johnson, L., Patel, D. J., & Jacobsen, S. E. (2014). Non-CG methylation patterns shape the epigenetic landscape in Arabidopsis. Nature Structural & Molecular Biology, 21(1), 64–72.

Stroud, H., Greenberg, M. V. C., Feng, S., Bernatavichute, Y. V., & Jacobsen, S. E. (2013a). Comprehensive Analysis of Silencing Mutants Reveals Complex Regulation of the Arabidopsis Methylome. Cell, 152(1), 352–364.

Stroud, H., Greenberg, M. V., Feng, S., Bernatavichute, Y. V., & Jacobsen, S. E. (2013b). Comprehensive analysis of silencing mutants reveals complex regulation of the Arabidopsis methylome. Cell, 152(1-2), 352–364.

Taheri-Ghahfarokhi, A., Taylor, B. J. M., Nitsch, R., Lundin, A., Cavallo, A.-L., Madeyski-Bengtson, K., Karlsson, F., Clausen, M., Hicks, R., Mayr, L. M., Bohlooly-Y, M., & Maresca, M. (2018). Decoding non-random mutational signatures at Cas9 targeted sites. Nucleic Acids Research, 46(16), 8417–8434.

Tan, F., Zhou, C., Zhou, Q., Zhou, S., Yang, W., Zhao, Y., Li, G., & Zhou, D.-X. (2016). Analysis of Chromatin Regulators Reveals Specific Features of Rice DNA Methylation Pathways. Plant Physiology, 171(3), 2041–2054.

Thielen, P. M., Pendleton, A. L., Player, R. A., Bowden, K. V., Lawton, T. J., & Wisecaver, J. H. (2020). Reference Genome for the Highly Transformable Setaria viridis ME034V. G3 , 10(10), 3467–3478.

Tran, R. K., Zilberman, D., de Bustos, C., Ditt, R. F., Henikoff, J. G., Lindroth, A. M., Delrow, J., Boyle, T., Kwong, S., Bryson, T. D., Jacobsen, S. E., & Henikoff, S. (2005). Genome Biology (Vol. 6, Issue 11, p. R90). https://doi.org/10.1186/gb-2005-6-11-r90

Tsai, S. Q., Wyvekens, N., Khayter, C., Foden, J. A., Thapar, V., Reyon, D., Goodwin, M. J., Aryee, M. J., & Joung, J. K. (2014). Dimeric CRISPR RNA-guided FokI nucleases for highly specific genome editing. Nature Biotechnology, 32(6), 569–576.

Uusi-Mäkelä, M. I. E., Barker, H. R., Bäuerlein, C. A., Häkkinen, T., Nykter, M., & Rämet, M. (2018). Chromatin accessibility is associated with CRISPR-Cas9 efficiency in the zebrafish (Danio rerio). PloS One, 13(4), e0196238.

Van Eck, J. (2018). The Status of Setaria viridis Transformation: Agrobacterium-Mediated to Floral Dip. In Frontiers in Plant Science (Vol. 9). https://doi.org/10.3389/fpls.2018.00652
Van Eck, J., Swartwood, K., Pidgeon, K., & Maxson-Stein, K. (2017). Agrobacterium tumefaciens-Mediated Transformation of Setaria viridis. In A. Doust & X. Diao (Eds.), Genetics and Genomics of Setaria (pp. 343–356). Springer International Publishing.

Wang, Z.-P., Xing, H.-L., Dong, L., Zhang, H.-Y., Han, C.-Y., Wang, X.-C., & Chen, Q.-J. (2015). Egg cell-specific promoter-controlled CRISPR/Cas9 efficiently generates homozygous mutants for multiple target genes in Arabidopsis in a single generation. Genome Biology, 16, 144.
Weiss, T., Wang, C., Kang, X., Zhao, H., Elena Gamo, M., Starker, C. G., Crisp, P. A., Zhou, P., Springer, N. M., Voytas, D. F., & Zhang, F. (2020). Optimization of multiplexed CRISPR/Cas9 system for highly efficient genome editing in Setaria viridis. The Plant Journal: For Cell and Molecular Biology, 104(3), 828–838.

Wickham, H. (2016). ggplot2: Elegant Graphics for Data Analysis. Springer, Cham.
Wu, T. D., & Watanabe, C. K. (2005). GMAP: a genomic mapping and alignment program for mRNA and EST sequences. Bioinformatics , 21(9), 1859–1875.

Wu, X., Scott, D. A., Kriz, A. J., Chiu, A. C., Hsu, P. D., Dadon, D. B., Cheng, A. W., Trevino, A. E., Konermann, S., Chen, S., Jaenisch, R., Zhang, F., & Sharp, P. A. (2014). Genome-wide binding of the CRISPR endonuclease Cas9 in mammalian cells. Nature Biotechnology, 32(7), 670–676.

Xiang, X., Corsi, G. I., Anthon, C., Qu, K., Pan, X., Liang, X., Han, P., Dong, Z., Liu, L., Zhong, J., Ma, T., Wang, J., Zhang, X., Jiang, H., Xu, F., Liu, X., Xu, X., Wang, J., Yang,

H., … Luo, Y. (2021). Enhancing CRISPR-Cas9 gRNA efficiency prediction by data integration and deep learning. Nature Communications, 12(1), 3238.

Xie, K., Minkenberg, B., & Yang, Y. (2015). Boosting CRISPR/Cas9 multiplex editing capability with the endogenous tRNA-processing system. Proceedings of the National Academy of Sciences of the United States of America, 112(11), 3570–3575.

Xi, Y., & Li, W. (2009). BSMAP: whole genome bisulfite sequence MAPping program. BMC Bioinformatics, 10, 232.

Yan, L., Wei, S., Wu, Y., Hu, R., Li, H., Yang, W., & Xie, Q. (2015). High-Efficiency Genome Editing in Arabidopsis Using YAO Promoter-Driven CRISPR/Cas9 System. Molecular Plant, 8(12), 1820–1823.

Yarrington, R. M., Verma, S., Schwartz, S., Trautman, J. K., & Carroll, D. (2018). Nucleosomes inhibit target cleavage by CRISPR-Cas9 in vivo. Proceedings of the National Academy of Sciences of the United States of America, 115(38), 9351–9358.

Yin, K., Gao, C., & Qiu, J.-L. (2017). Progress and prospects in plant genome editing. In Nature Plants (Vol. 3, Issue 8). https://doi.org/10.1038/nplants.2017.107

Zemach, A., & Grafi, G. (2003). Characterization of Arabidopsis thaliana methyl-CpG-binding domain (MBD) proteins. The Plant Journal: For Cell and Molecular Biology, 34(5), 565–572.

Zemach, A., Kim, M. Y., Hsieh, P. H., Coleman-Derr, D., Eshed-Williams, L., Thao, K., Harmer, S. L., & Zilberman, D. (2013). The Arabidopsis nucleosome remodeler DDM1 allows DNA methyltransferases to access H1-containing heterochromatin. Cell, 153(1), 193–205.

Zhang, X., Henriques, R., Lin, S.-S., Niu, Q.-W., & Chua, N.-H. (2006). Agrobacterium-mediated transformation of Arabidopsis thaliana using the floral dip method. Nature Protocols, 1(2), 641–646.

Zhong, Z., Feng, S., Duttke, S. H., Potok, M. E., Zhang, Y., Gallego-Bartolomé, J., Liu, W., & Jacobsen, S. E. (2021). DNA methylation-linked chromatin accessibility affects genomic architecture in Arabidopsis. Proceedings of the National Academy of Sciences of the United States of America, 118(5). https://doi.org/10.1073/pnas.2023347118

Zhu, C., Yang, J., & Shyu, C. (2017). Setaria Comes of Age: Meeting Report on the Second International Setaria Genetics Conference. Frontiers in Plant Science, 8, 1562.