Title
The consequences of transposable element and DNA methylation on plant genomes

Author
Roessler, Kyria Anna

Publication Date
2017

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA,
IRVINE

The consequences of transposable elements and DNA methylation on plant genomes

DISSERTATION

submitted in partial satisfaction of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

in Biological Sciences

by

Kyria Anna Roessler

Dissertation Committee:
Professor Brandon S. Gaut, Chair
Associate Professor Kevin R. Thornton
Assistant Professor J.J. Emerson

2017

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

I would like to express the deepest appreciation to my committee chair, Brandon S. Gaut, who has the attitude and was always supportive: he continually had an excitement in regard to research and teaching. Without his guidance and persistent help this dissertation would not have been possible.

I would like to thank my committee members, Professor Kevin Thornton and Professor J.J. Emerson, whose guidance during committee meetings was always encouraging and informative.

In addition, a thank you to Professor Ed Green of University of California, Santa Cruz, who took a chance on me and introduced me to the world of genomics.

I thank the UCI Mathematical, Computational, and Systems Biology gateway program for helping my dream of going to graduate school a reality, and the support its program provided. Support was provided by the National Institute of Biomedical Imaging and Bioengineering, National Research Service Award EB009418 from the University of California, Irvine, Center for Complex Biological Systems.

CURRICULUM VITAE

Kyria Anna Roessler

2007-11 B.S. in Biochemistry and Molecular Biology,
Minor Mathematics,
University of California, Santa Cruz

2012-13 Mathematical, Computational, and Systems Biology,
Gateway Program,
University of California, Irvine

2012-17 Ph.D. in Ecology and Evolutionary Biology,
University of California, Irvine

FIELD OF STUDY

Evolution of plant genomics

PUBLICATIONS

St John, John A., et al. "Sequencing three crocodilian genomes to illuminate the evolution of archosaurs and amniotes." *Genome biology* 13.1 (2012): 415.

Diez, Concepcion M., Kyria Roessler, and Brandon S. Gaut. "Epigenetics and plant genome evolution." *Current opinion in plant biology* 18 (2014): 1-8.

Roessler, Kyria, Shohei Takuno, and Brandon S. Gaut. "CG methylation covaries with differential gene expression between leaf and floral bud tissues of Brachypodium distachyon." *PloS one* 11.3 (2016): e0150002.

PRESENTATIONS

"Methylation and gene expression in leaf and floral tissues in *Brachypodium distachyon*." Presented at Society for Molecular Biology, June 11, 2014, San Juan, Puerto Rico

"The evolution of genomic content over generations of inbreeding maize lines" Presented at Society for Molecular Biology, July 2016, Gold Coast, Australia, *Poster Presentation*

"Methylation and gene expression in leaf and floral tissues in *Brachypodium distachyon*." Presented at Plant and Animal Genome Conference, January 17, 2017

AWARDS/GRANTS

2012-13	NIH Training Grant
2015	Federal Work Study Award
2016	Ayala School of Biological Sciences Graduate Award
2016-17	GAANN Fellowship

ABSTRACT OF THE DISSERTATION

The consequences of transposable element and DNA methylation on plant genomes

By

Kyria Anna Roessler

Doctor of Philosophy in Biological Sciences

University of California, Irvine, 2017

Brandon S. Gaut, Chair

Plant genomes are not static; they are constantly being transformed by nucleotide substitutions, the propagation of mobile DNA, and epigenetic modifications. In the three chapters of my dissertation, I show how plant genomes are shaped by transposable elements (TEs) and DNA methylation. In the first chapter, I test the hypothesis that DNA methylation is involved in differential gene expression between plant tissues. To explore this hypothesis, I measured whole genome DNA methylation and gene expression in leaf and floral bud tissue from *Brachypodium distachyon*. I found that differential CG methylation in the promoter region explains ~10% of the variation in gene expression between tissues. The second chapter examines the two modes that a plant uses to silence TEs, and I specifically question why both are necessary for efficient TE containment. I address this question by creating a mathematical model of ordinary differential equations that represents the interactions between TE propagation and epigenetic silencing, including DNA methylation. The model suggests that both modes are crucial for efficient silencing, and it also suggests that TE retention leads to more robust silencing. Finally, the

third chapter predicts that, because of their deleterious nature, TEs will be 'purged' from a lineage that has undergone inbreeding. To test this, I examined the properties of maize genomes that were subjected to inbreeding for six generations. Over a total of 11 inbred lines, I measured genome size with cell flow cytometry and characterized genome content by whole genome sequencing. The results revealed evidence that genome size decline is associated with TE loss in a subset of inbreeding lines and provided an opportunity to consider potential mechanisms for TE removal.

INTRODUCTION

In the classic paper from 1950, "The origin and behavior of mutable loci in maize", Barbara McClintock challenged the preconceived notions of a static and stable genome (McClintock 1950). She had discovered that gene-like DNA could be mobile and change its position within a chromosome; these new genomic components were termed transposable elements (TEs). TEs were discovered because they could alter gene expression if inserted near or in a gene. By the 1970s TEs had been found in other organisms such as viruses and bacteria (Ravindran 2012). Now, in the age of genomics, we know that TEs compose nearly 65% of the human genome (Lander et. al 2001) and 85% of the maize genome (Schnable et. al. 2009). At first, because their exact function was unknown, TEs were described as repetitive 'junk' DNA. That view is, however, changing. As high-throughput sequencing continues to improve and as genomes are assembled more accurately, the tools to study TEs have also improved. This has allowed for better analysis of TE accumulation in genomes and of their evolutionary consequences.

TEs can be classified into two classes by the mechanisms they use for mobility. Class I elements require an RNA intermediate and reverse transcriptase to 'copy and paste' in the genome. Class II elements do not employ reverse transcriptase, but use transposase for direct transposition to 'cut and paste' into a new part of the genome (Fedoroff 1989). The proliferation of class I and class II elements can have a number of impacts on their host genome and gene function. For example, a TE can affect a gene by directly inserting into the coding region, disrupting gene expression (Schwarz-Sommer et. al. 1987; White et. al 1994). A TE can also insert into an intron and affect splice sites or transcription factor

binding sites, potentially affecting function (Bradley et. al 1993; Ortiz et. al. 1990). In addition, because of their repetitive nature, TEs can be the template for inter-element recombination, which may generate inversions, deletions and duplications (Robbins et. al. 1989; Lister et. al 1993; Flavell et. al. 1994). It is apparent that TEs can have significant mutational effects and thus pose a potential challenge to their host's fitness (McDonald 1993). On a larger scale, the proliferation of large numbers of TEs can greatly increase genome size (Gregory and Hebert 1999), and overly large genomes may be deleterious (Knight et al., 2005).

Given the potential for deleterious effects, natural selection may act to limit TE proliferation and copy numbers. For instance, newly inserted TEs are found at a very low frequency in populations (Charlesworth and Langley 1991; Biemont 1992), suggesting that natural selection against them does not allow them to rise to higher frequency. This selection could result from two processes: either selection against newly inserted TEs (because of the mutations it causes to nearby genes) or selection against ectopic recombination that causes large chromosomal rearrangements (Charlesworth and Langeley 1991). Either process is expected to remove deleterious TEs from the population. This removal of deleterious TEs from the population helps purge TEs from the population, retaining population fitness. These processes are of intrinsic interest in plants, because plant genomes vary widely in size and size correlates with repeat content (Michael 2014; Tenaillon et. al. 2011). Hence, the processes that purge TEs may be variable among populations and species. For example, the difference in genome size between *A. thaliana* and *A. lyrata* – two species that diverged 10 million years ago - can be attributed to hundreds of thousands of small deletions, mostly in noncoding DNA and TEs (Hu et. al.

2011). Why is it that one species purges TEs and the other retains them? One possible reason for this divergence is a shift in mating strategies. Plant populations with higher rates of selfing have been found to have a smaller genome size (Price 1976; Govindaraju and Cullis 1991). Indeed, in the example above, *A. thaliana* has a high rate of selfing while *A. lyrata* is an outcrosser. However, mating system is not the only difference between species, and cannot explain that vast majority of TE content and genome size variation. In addition, purging of individual TE insertions may not be sufficient to keep the remaining TEs from propagating, so other mechanisms from the host are needed to keep TEs under control.

Although deleterious TE insertions are often removed by natural selection, the plant also limits TE proliferation by modifying them epigenetically. There are two steps plant hosts take to the epigenetically silence TEs. The first step is post-transcriptional regulation. In this step, the plant host degrades mRNAs from TEs into 21-22 nt small-interfering RNAs (siRNAs), and these 21-22nt siRNAs trigger the 'initiation' of the RNA-directed DNA methylation (RdDM) pathway. The second pathway uses 24 nt siRNAs that guides epigenetic modifications and lead to pre-transcriptional silencing of TEs (Bousios and Gaut 2016). Pre-transcriptional silencing in higher plants includes DNA methylation of TEs, which occurs in three sequence contexts: CG, CHG, and CHH, where H = A, C or T (Lister et. al. 2008). This modification is heritable, and it is associated with additional epigenetic markers such as histone modifications and nucleosome positioning (Bernatavichute et. al 2008; Chodavarapu et. al. 2010). DNA methylation acts to limit the transcription and proliferation of TEs and is therefore a defense against potential deleterious mutations caused by TE insertions (Lippman et. al. 2004). This means that once methylation occurs, the host genome is largely in control of TE proliferation. Most TEs are

found in this epigenetically silenced state. So far these mechanisms have been documented extensively in *A. thaliana* (Slotkin et. al. 2007), but other studies have shown both that lower plants lack RdDM genes and that there exist substantial differences in DNA methylation across plant species (Ma et. al. 2015; Takuno et. al. 2016). Differences in epigenetic silencing mechanisms could lead to varying host interactions in response to TE propagation and thus lead to differing TE content and potential genome size.

Since methylated TEs are found throughout the genome, they have long been seen as a potential mechanism for controlling gene expression. Studies of methylation mutants have revealed that they have the ability to regulate the expression of nearby genes. For example, the gene *FWA*'s expression is controlled by an upstream TE and the gene is upregulated when the element is demethylated (Soppe et. al. 2000). The advent of new high-throughput methods, such as bisulfite sequencing (BSseq), allows for characterizing the genome wide distribution of DNA methylation. Using a similar approach, a genome wide study revealed that methylated TEs inserted near genes are associated with reduced gene expression (Hollister et. al. 2011; Diez et. al. 2014), generalizing the example of *FWA*. Furthermore, these experiments suggest that there is variation of methylation between individuals and tissues (Schmitz et. al. 2013). Even though DNA methylation is primarily in noncoding repetitive DNA, could this variation near genes lead to varying gene expression between tissues? A few studies have investigated this question, using high-throughput sequencing technologies, but their results have mainly been mixed - either finding differentially methylated regions have an association with RNA-seq expression or finding higher variation between individuals than tissues (Schmitz et. al. 2013; Song et al. 2013).

More studies are needed to unmask the true relationship between differential methylation between tissues and its effects on gene expression.

From these examples about genome size and TE silencing, we can see that TEs can have large effects on plant genomes. The goal of my dissertation is to try to understand some of the genomic consequences of TEs and DNA methylation. The first chapter explores whether there are differences in DNA methylation between tissues and whether it is associated with differential gene expression. To answer this question, we performed whole genome BSseq and RNA-seq on leaf and floral tissue in *Brachypodium distachyon*. We found that methylation could explain ~10% of the variation of differentially expressed genes between tissues, depending on the methylation context and location, The second chapter introduces a new mathematical model to explain the dynamics between TE propagation and host silencing in plant genomes. We use the model to explore the different mechanisms of methylation and to consider the parameters that are are ideal for TE invasion. Interestingly, we find that retention of silenced TEs is important for robust methylation, and it could be a possible reason for high TE content and large genome sizes. In the third chapter we examine the transition period from outcrossing to selfing in the first few generations in maize (*Zea mays* spp. *mays*) and its effects on genome size. We predicted that if there is a decrease in genome size, it is related to the purging of TEs and other repetitive elements. We find this to be true in certain lines of maize, and attempt to explain possible mechanisms for this TE loss.

With these three projects I hope to further illuminate the interesting aspects of TE silencing in the context of plant genome evolution. I have used multiple contemporary methods -- such as next-generation sequencing and computational modeling -- to

understand some of the interactions among genome size, methylation, and gene expression. These evolutionary processes are currently affecting plants today, including crucial crop species, and this research can help understand how these genomes are being shaped.

Chapter 1: CG methylation covaries with differential gene expression between leaf and floral bud of *Brachypodium distachyon*

The term 'epigenetics' refers to processes beyond (*epi*-) genetics and, more concretely, to heritable chromosomal modifications that have the potential to vary during development and stress (Law 2010; Dowen et. al. 2012). Epigenetic modifications include histone variants but also DNA methylation. In plants, the methylation of cytosines occurs in three sequence contexts: CG, CHG and CHH, where H = A, C or T. All three contexts are usually methylated in repetitive sequences, which serves to limit the transcription and proliferation of transposable elements (TEs) (Lippman et. al. 2004). Genes are often also methylated, but typically only in the CG context (Cokus et. al. 2008; Lister et. al. 2008; Greaves et. al. 2012). The function of this genebody methylation (gbM) is not yet clear, but potential functions include the exclusion of histone H2A.Z (Coleman-Derr and Zilberman 2012), control of aberrant intragenic expression (Zilberman 2007), protection from transposable element (TE) insertion (Regulski et. al. 2013), and facilitation of intron-exon splicing (Lorincz et. al. 2004; Luco et. al. 2010; Shukia et. al. 2011).

DNA methylation has long been hypothesized to have a direct effect on gene regulation during development (Bird 1997; Richards 1997). With the growing availability of single base resolution methylation data, like bisulfite sequencing (BSseq) data, this hypothesis has been tested directly. In humans, for example, DNA methylation varies dramatically throughout development, and this variation is often correlated with gene expression (Franklin et. al. 1996; Lister et. al. 2009; Hawkins et. al. 2010; Reik 2007). In plants, the available evidence suggests that methylation levels vary for highly specialized

tissues, such as the endosperm and the pollen vegetative nucleus (Lauria et. al 2004; Hsieh et. al 2009; Ibarra et. al. 2012; Zemach et. al. 2010), and that this methylation variation likely contributes to genetic imprinting and trans-generational silencing of TEs (Slotkin et. al. 2009; Choi et. al. 2002; Gehring et. al. 2009).

Outside of these few specialized tissues, a clear picture has not yet emerged as to whether methylation commonly varies among plant tissues and, if so, whether methylation variation contributes to tissue-specific gene expression (GE). Some evidence suggests that most plant tissues do not vary substantially in DNA methylation. For example, genome-wide profiling in rice (*Oryza sativa L.)* identified few DNA methylation differences between shoot and root, and only a few additional differences in CHH methylation between these two tissues and the embryo (Zemach et. al. 2010). Moreover, a survey of several *A. thaliana* accessions found that tissue-specific variation in methylation was much less pronounced than genetic variation, leading the authors to conclude that "...DNA methylation is less dynamic than gene expression patterns in plants and only plays a role during specific stages of development or cell type, such as companion cells" (Schmitz et. al. 2013).

In contrast to these studies, there is some emerging evidence that differential methylation may play a role in tissue specific GE. For example, researchers have detected ~2000 differentially methylated regions (DMRs) among four soybean tissues, and a subset of these DMRs correlate with tissue-specific GE of ~60 genes (Song et. al. 2013). Similarly, analysis of tissue-specific DNA methylation patterns in *Sorghum bicolor* (Zhang et. al. 2011), *Populus tichocarpa* (Vining et. al. 2012) and maize (*Zea mays* ssp. *mays*) (Clandaele et. al. 2014) hint that epigenetic variation among vegetative tissues correlates with tissuespecific expression. However, not all of these studies have measured methylation at

single-base resolution, which greatly limits the ability to draw firm conclusions; the number of contrasts of methylation between plant tissues is growing, but such studies remain rare.

Methodological differences among studies have also made conclusions difficult. For example, few methylation studies have employed biological replication, and thus it is usually unclear whether methylation variation between tissues exceeds the statistical variation expected from within a single sampled tissue. Even when the data are at single-base resolution, studies have used different summaries of the data as the basis for inferences, and this has caused confusion. Some studies have focused on summarizing methylation for genomic features like genes and TEs (Feng et. al. 2010; Zemach et. al. 2010). Other studies have focused on DMRs as a summary of the data. DMRs were initially defined as stretches of DNA sequence for which methylation differences between samples were higher than expected at random (Lister et. al. 2009). While the initial definition of DMRs is straightforward and meaningful, more recent studies have used empirical means, such as sliding windows, to identify DMRs, and these empirical definitions vary from study to study (Song et. al. 2013; Schmitz et. al. 2013). As a result, the interpretation and meaning of DMRs varies among studies, compromising the value of inferences.

This study is focused ultimately on the question of whether DNA methylation and GE covary between tissues. To that end, we have measured both DNA methylation and gene expression between two tissues (leaves and floral buds) of *Brachypodium distachyon* (brachypodium), a grass species that has served as a model for genomic studies (Vogel et. al. 2010). While our ultimate goal is to assess methylation and GE, our proximal goals include an empirical assessment of the effects of replication both on inferring methylation

differentiation between tissues and on the impact of summary methods (i.e., DMRs vs. single-base metrics) on inferences. Overall, we find the two tissue samples to be significantly different in DNA methylation patterns, but we also find that the false positive rate without replication is high (>50%). In all respects, DMRs are less useful than single-base or regional measures in our empirical analyses. Altogether, we find that CG methylation and GE covary between tissues, explaining up to 9% of variation in gene expression.

RESULTS AND DISCUSSION

DNA Methylation within and between tissue samples: To assess methylation variation, we utilized BSseq data from a previous study (Takuno and Gaut 2013) that generated reads from three biological replicates of two tissues: leaf and immature flower buds. We denoted the leaf replicates as L1, L2 and L3 and the floral bud replicates as F1, F2 and F3. The data had conversion error rates of <1.3% for each replicate (Takuno and Gaut 2013; S1.1 Table). Following the previous study, we mapped BSseq reads to the *B. distachyon* genome and tallied only uniquely mapping reads. Each replicate yielded ~15X of mapped coverage, such that each tissue had ~45X coverage per base, on average (Takuno and Gaut 2013).

To our knowledge, no plant DNA methylation papers have assessed whether tissue-specific variation exceeds that expected from proper biological replication. To assess this question, we first tested for a signal of differentiation between two BSseq datasets at single nucleotide sites, which we call Differentially Methylated Sites (DMSs). To identify DMSs, we required a minimum coverage of 3 reads for each site in each tissue and then applied Fisher's Exact Test (FET) (Lister et. al. 2009; see Methods). There were many DMSs

between two biological replicates from the same tissue. For example, there were 218,631 DMSs between L1 and L2 and an average of ~250,000DMSs between two leaf replicates (Fig 1A). However, DMSs were more abundant between replicates from different tissues, with an average of ~324,000 differential sites (Figure 1.1A). The average number of DMSs was significantly higher for between-tissue vs. intra-tissue comparisons (permutation, $p<0.01$), indicating that the tissue samples were significantly differentiated.

We also inferred DMSs by combining the three replicates within each tissue and then comparing the combined leaf data (L1+L2+L3) to the combined floral data (F1+F2+F3). Using the combined replicates from each tissue, FET analyses identified 500,245 DMSs. Following a previous study (Ziller et. al. 2014), we assumed these 500,245 DMSs to be our best estimate of "true" DMSs between tissues and found that this true set rarely overlapped in location with DMSs that were identified between replicates within a tissue; typically <5% of within-tissue DMSs overlapped with the true set (Figure 1.1A). In contrast, the overlap was more significant, at 15.6% on average, between the true set and DMSs identified between replicates from different tissues (Figure 1.1A). These percentages define genetic distances between two BSseq replicates that can then be used for clustering analyses. A neighbor-joining analysis clearly separated replicates from different tissues (Figure 1.1B), further supporting the contention that the two tissue samples differ in DNA methylation beyond that expected from sampling.

The 15.6% average amount of overlap between the 'true' set of DMSs and the DMSs based on single replicates likely reflects, in part, lower statistical power for the single replicates. It is interesting to note, however, that comparisons between single-replicates

also yield high false-positive rates (FPRs), which cannot be an artifact of higher statistical power with combined data. For example, the comparison of L3 and F3 replicates yielded 368,643 DMSs (Figure 1.1A). Of these, 92,989 (or 18.6%) overlapped with the set of true DMSs; hence, 81.4% of the DMSs identified between these two replicates were not supported by the larger, combined data set. In other words, had we relied on a single replicate from each tissue for this study, >80% of our inferences would have been incorrect relative to more extensive data. In theory, the high FPR can be reduced either by increasing the stringency of the FET or by adjusting the FET for multiple-tests, but these adjustments do not help in this case. For example, when we focused on the L3 vs. F3 comparison and applied a false discovery rate (FDR) correction at $q < 0.05$, we found 10,435 DMSs compared to 368,643 without FDR adjustment (Figure 1.1A). However, none of these DMSs overlapped with the true set, yielding an FPR of 100%.

A potential advantage of studying DMRs, as opposed to DMSs, is that they summarize signals over contiguous sites, and it is thus possible that they reduce the FPR. As we have noted, the definition of a DMR varies widely among studies; here we focused on the original definition to define DMRs as a region of non-random differentiation between samples (Lister et. al. 2009). To determine expectations under 'randomness', we permutated cytosine methylation states throughout the genome (see Materials & Methods); permutations indicated that ≥ 5 DMSs in a row were 1.3% of those expected at random (Figure S1.1). Accordingly, we defined a DMR as ≥ 5 DMSs in a row that had a consistent direction of methylation bias (i.e., hypomethylated in one or the other tissue). These DMRs contain 2,672 DMSs, which is a small percentage (0.5%) of the total set of 500,245 DMSs

(Table 1.1), indicating that DMSs are rarely clustered across the genome more than expected at random.

With this definition, we again observed more DMRs between tissues (187, on average) than within tissues (59, on average)(Figure 1.1C), and this difference was again statistically significant (permutation, $p<0.01$). By combining data (L1+L2+L3 vs. F1+F2+F3), we inferred a 'true' set of 448 DMRs with an average size of 38.4 bp, a minimum length of 5.00 bp and a maximum length of 522 bp. For DMRs, 1.0% of within-tissue DMRs overlapped on average with the true set, whereas 6.9% of between-tissue DMRs overlapped with the true set on average. These distances again separated the tissues in clustering analyses (Figure 1.1D), verifying significant tissue differentiation. However, the between-tissue FPR based on single BSseq replicates was consistently higher for DMRs than for DMSs, with a *minimum* FPR of 89.2%. In other words, ~80% or more of our DMR inferences were incorrect for contrasts between single replicates relative to the more extensive dataset.

This raises the question as to why the FPR is so high and whether our observations are unique. To answer the latter question, to our knowledge only one other study of plants has used biological replication for BSseq data to compare methylation between *Arabidopsis thaliana* and *A. lyrata* (Seymour et. al. 2014). [At least one other paper replicated their control but not experimental samples (Stroud et. al. 2013)]. In the *Arabidopsis* paper, the authors used replication to help filter the number sites for testing, thus reducing the multiple test problem and increasing statistical power. They did not, however, explicitly report on the level of within vs. between tissue differences. To address the former question, the FPR may be high for technical, statistical and/or biological reasons. Technically, BSseq

data are subject to conversion error, but conversion errors are unlikely to explain our observations because coverage is high and conversion error is low. Statistically, it is easy to envision that the FET may signal numerous false-positives, but the FPR remains high for DMSs when the FET is FDR corrected (Becker et. al. 2011), as noted above. Finally, biological variation among replicates may contribute to the FPR, both because tissue samples are likely to include mixtures of different cell types that vary in proportion among replicates (Ji et. al. 2015) and also because it is likely that there is heterogeneity in methylation levels even among cells of a single type (Smallwood et. al. 2014). We thus suspect, but cannot prove, that the largest contribution to variation among replicates is biological in origin.

Our FPR calculations deserve two further comments. First, the FPR calculations rely on the assumption that the set of 'true' DMSs and DMRs are defined by our analysis of combined data. This assumption cannot hold fully because there must be false-positives in the combined data, but the FPR rate of the combined data are difficult to assess. Second, we note that as levels of methylation differentiation become more pronounced, the signal:noise ratio will also increase. Thus, our data reflect the importance of replication for contrasts between tissues in the same species; however, it may not be as useful to replicate data that are designed to summarize broad-scale differences in methylation patterns between distantly related species (Feng et. al. 2010; Zemach et. al. 2010).

Differences in methylation patterns between tissues: Given that we found tissue-specific differentiation between leaves and floral buds, we sought to categorize the pattern of methylation differences, both in terms of cytosine contexts and genomic

locations. For these results, we based all analyses on combined leaf (L1+L2+L3) and floral (F1+F2+F3) data, following (Ziller et. al. 2014).

Our first finding was that the two tissue samples were more similar than different in their methylation patterns. To reach this conclusion, we identified sites with conserved methylation between the two tissues–i.e., Conserved Methylated Sites (CMSs). To be a CMS, a site required the support of a binomial test (Lister et. al. 2008) at a *p*-value of 0.05 in both tissue samples. Overall, we found that 18,780,682 cytosines were methylated in both tissues (Figure 1.2A; Table 1.1), representing 18.7% of the 100,229,480 genomic cytosines in the proper context for methylation (i.e., CG, CHG or CHH). Among CMSs, most (62.7%) were in the CG context, with an appreciable minority in the CHG context (29.8%) and relatively few in the CHH context (7.5%). Overall, the set of CMSs and DMSs were mutually exclusive, and there were 37-fold more CMSs (Figure 1.2A; Table 1.1). Other studies have also found more similarities than differences among plant tissue samples (Zemach et. al. 2010; Zhang et. al. 2011).

Our second finding was that most variation between tissues occurred at CHG sites. Cytosines were most commonly methylated in the CG context, but 56.5% (282,440 sites) of DMS sites occurred in the CHG context (Table 1.1; Figure 1.2A). To investigate further, we estimated the DMS 'rate' by comparing the observed number of DMSs to the available number of cytosines in a particular context. For example, there were 19,722,162 cytosines in the CHG context throughout the genome and a total of 282,440 DMSs, yielding a rate of 1.43% (Table 1.1). In contrast, CG and CHH methylation had lower rates, at 0.55% and 0.17%, respectively (Table 1.1). CHH methylation may not be as differentiated in part because the overall proportion of methylated CHH sites was much lower than CG or CHG

sites. Interestingly, the direction of DMSs was biased, because 57% were methylated in floral buds but not leaf, representing a deviation from the expectation of equality (binomial, $p<10^{-15}$).

Our observation that variability between tissues was highest at CHG sites is similar to comparisons among rice tissues (Zemach et. al. 2010) and among somaclonal variants of oil palm (Ong-Abdullah et. al. 2015). Similarly, in *Arabidopsis* species tissue-specific differences were attributable to CHH and CHG methylation changes within DMRs (Seymour et. al. 2014). However, CG methylation varies more than CHH or CHG methylation among tomato developmental stages (Zhong et. al. 2013) and also between generations of *A. thaliana* mutation accumulation lines (Becker et. al. 2011). Thus, the principle context of DNA methylation variability varies either as a function of species or the tissues sampled.

Having assessed the effect of context, we shifted our attention to three genomic features of interest: genes, promoters and transposable elements (TEs). Among the three features, the set of 68,264 non-genic, annotated TEs had the highest CMS rates (Figure 1.2B), as was expected from previous studies of plant genomes (Becker et. al. 2011; Gent et. al. 2013), with methylation levels of 95.3% at CG sites and 64.4% at CHG sites (Table 1.1; Figure 1.2B). That said, TEs also had the highest DMS rates, at 2.62% in the CHG context (Table 1.1; Figure 1.2C). CHH methylation levels were low (<5%) throughout TEs, as noted previously for the entire brachypodium genome (Takuno and Gaut 2013). In contrast to TEs, the 26,072 annotated genes had the lowest DMS rate at 0.18% (Table 1.1; Figure 1.2C), but this low rate may reflect the fact that genes were primarily methylated in the CG context, which had the lowest DMS rates. Promoter regions, which were defined as 1.0 kb 5' upstream of the 26,072 genes, had noticeably higher levels of

conserved CHG methylation between tissues (at 14.5%) than genes (3.46%), but were similar to genes in most other respects (Figure 1.2; Table 1.1).

Given that CMSs and DMSs were especially prominent in TEs, it was not surprising that the distribution of CMSs and DMSs across chromosomes mimicked the density of TEs (Figure 1.2 and Figure S1.2), and there was no obvious correlation between CMSs and DMSs with gene density (Figure 1.2G and Figure S1.2). Altogether, the analysis of single sites paints a clear picture: most methylation occurred in TEs and most variation between tissues was within TEs in the CHG context.

Finally, we examined the pattern and location of the 448 DMRs identified between tissues to assess whether they paralleled results based on single sites. First, 65% of DMRs were hypo-methylated in floral buds ($p < 0.01$), verifying increases in overall methylation in this tissue. Second, although most DMSs were found in the CHG context (Figure 1.2C), we found that 67% of the DMSs within all of our DMRs were sites in the CHH context. This observation suggests that there may be a spatial (clustered) context to the mechanisms that underlie CHH differences between tissues, consistent with the observation in maize that CHH sites tend to be clustered (Gent et. al 2013). Finally, the location of DMRs was biased: 39% of DMRs were found in unannotated regions of the genome, but 31% were found within TEs, 17% within genes and 13% within promoter regions. Given that the total number of cytosine sites within TEs and within genes was similar, at ~29 million bases (Table 1.1) each, the lower percentage in genes again indicates that genic methylation is more highly conserved between tissues than methylation of TEs.

Methylation & Gene Expression: The primary goal of this paper is to determine whether methylation differentiation between tissues covaries with GE. The idea that GE

and methylation covary traces back to the origin of epigenetics (Diez et. al. 2013) and seems to be upheld by weak signals from plant data (Song et. al. 2013; Zhong et. al. 2013.

Gene expression data: To measure GE, we generated RNAseq data from leaf and floral tissues, using the same three plants and samples (biological replicates) that were used to generate BSseq data (see Methods). Each of the replicates had > 12 million RNAseq reads that mapped uniquely to the *B*. *distachyon* genome (Table S1.2). Out of 26,552 annotated protein-coding genes, we retained 26,072 that did not overlap with annotated TEs, of which 19,956 had evidence of expression in at least one tissue, as determined by a cutoff of FPKM > 0.02 (see Materials and Methods). Second, we identified differentially expressed genes between tissues at an FDR of $q < 0.01$ (Figure S1.3). A total of 7,704 genes were significantly differentially expressed between leaf and floral tissue; these exhibited no obvious clustering by chromosomal position (Figure 1.2H and Figure S1.2). GO analyses of differentially expressed genes suggested enrichment for functions in membrane and microtubule development (Table S1.3).

GE and DMRs: Since many studies have focused on DMRs (rather than annotation features) to assess correlations with GE, we began by testing for associations between GE and DMRs. If DMRs influence differential GE, we hypothesized that DMRs should be enriched around differentially expressed genes. To test this hypothesis, we measured the distance (in bp) between a DMR and the closest differentially expressed gene. We then tested whether the observed average distance from DMRs to genes was smaller than expected at random, as tested by permutation (see Materials & Methods). We found that on average a DMR was 18,710 bp from a differentially expressed gene, which was not significantly smaller than the random expectation of 17,572 bp ($p = 0.82$; Figure S1.4).

Based on this analysis, there is no evidence to suggest that DMRs are enriched near differentially expressed genes, as one might expect if DMRs help drive tissue-specific expression on a genome-wide scale.

Thinking that we may have missed an important signal by focusing on the entire genome, we delved into the three genomic features separately. For each feature, we focused on DMRs that were hyper-methylated in one vs. the other tissue. For example, we tallied DMRs within 25 genes that were hyper-methylated within floral buds. For this set of 25 genes, we predicted lower GE in floral than leaf tissue. Similarly, for the 19 genes that had a methylated DMR in leaf but not floral tissue, we predicted lower GE in leaf. These predictions were not upheld by the data, however (Figure 1.3A). In fact, the average level of differential expression did not vary among genes that had a hyper-methylated DMR in floral bud, a hyper-methylated DMR leaf or no DMR whatsoever (Figure 1.3A). We repeated this analysis for promoter regions of 1.0 kb 5' upstream of genes, and again found no discernible pattern (Figure 1.3B). Finally, because the methylation of TEs may effect the expression of nearby genes (Lippman et. al. 2004; Hollister and Gaut 2009), we also examined DMRs within annotated TEs closest to a gene. Again, there was no signal (Figure 1.3C). While the lack of signal may reflect low sample sizes, the presence of DMRs did not correlate with differential GE between the two tissues.

The proportion of converted reads: To investigate covariation between GE and methylation more thoroughly, we turned to a measure of DNA methylation that summarizes the proportion of non-converted reads over the total number of reads at cytosine residues in the proper contexts (CG, CHG or CHH) (Zemach et. al. 2010). This measure, which we call $prop_C$, can be applied to the entire genome, to specific genomic

features or to specific contexts (e.g., $prop_{CG}$, $prop_{CHG}$, $prop_{CHH}$). For example, over the entire genome, $prop_c$ was estimated to be 0.1815 for leaf tissue and 0.1823 for floral bud tissue throughout the entire genome, suggesting again (very) slightly higher levels of methylation in floral bud tissue. The *prop* measures provide an estimate of the methylation level for a region, but without a corresponding measure of significance. We focus on the use of these measures for the remainder of our analyses.

GE and Genic Methylation: To better understand patterns of methylation within genes, we first assessed the relationship among $prop_{CG}$, $prop_{CHG}$, and $prop_{CHH}$ within a tissue, using correlation analyses. In brief, all are significantly correlated with one another, with *r* values ranging from 0.43 to 0.61 (Table 1.2). However, $prop_{CG}$ and $prop_{CHG}$ were positively correlated in a somewhat striking pattern: CHG methylation was often present but rarely higher than CG methylation (i.e., in only 3,424 of 26,072 genes) (Figure 1.4A). This observation reaffirms that methylation in the CG context is predominant for genes (Cokus et. al. 2008; Lister et. al. 2008; Zhang et. al. 2006) but also illustrates that genic methylation is not limited to the CG context (Takuno and Gaut 2012).

Given that CG methylation is the primary component of genic methylation, we compared $prop_{CG}$ to GE within a tissue. Previous work has shown that the relationship between GE and gene body methylation is complex (Lister et. al. 2008). In general, methylated genes have intermediate levels of expression, such that hypo-methylated genes are both more-highly and less-highly expressed than hyper-methylated genes (Zemach et. al. 2010; Zhang et. al. 2006, Takuno and Gaut 2012). As expected, GE and genic methylation were correlated within tissues ($r = 0.287$; $p < 2.2e\text{-}16$; Table 1.2), but in a complex pattern

(Figure 1.3B). $prop_{CHG}$ and $prop_{CHH}$ were also correlated with GE but at lower levels (r = 0.046, $p = 1.21 \times 10^{-13}$ and r = 0.073, $p < 2.2 \times 10^{-18}$).

Lastly, we compared differential methylation to differential GE between tissues, focusing on either all of the 19,956 genes or just the 7,704 that were significantly differentially expressed. Differential GE and methylation were not correlated with $prop_{C}$, $prop_{CHH}$ or $prop_{CHG}$ (Table 1.3; Figure 1.3C) but were correlated between CG methylation and differential expression of the subset of 7,704 genes (Table 1.3). This significant correlation was negative, indicating that higher gene expression covaries with lower methylation levels. Note that the correlation, while significant, had a low absolute value (r = -0.0393; Table 1.3), suggesting that methylation differences explain at best a small proportion (3.9%) of the variance in GE between tissues. To sum: on a genome-wide scale, we uncovered moderate evidence that CG methylation and differential GE covary within genic regions.

Promoter methylation and GE: Differential methylation of promoter regions has been reported to correlate with GE during tomato ripening (Zhong et. al. 2013) and perhaps to tissue-specific GE of soybean genes (Song et. al. 2013). Accordingly we assessed relationships between promoter methylation and GE. For promoter regions there is a clear expectation of an inverse relationship between methylation levels and GE (Zhang et. al. 2006), such that higher expression correlates with lower levels of methylation.

We first assessed the pattern of DNA methylation within promoters and note that it varies as a function of both distance from the TSS and cytosine context. For example, CG and CHG methylation both reach a zenith ~750 bp from the TSS (Figure 1.5A and 1.5B), as documented previously (Cokus et. al. 2008; Lister et. al. 2008), but CHH methylation was

maximal ~500 bp from the TSS (Figure 1.5C). Within a tissue, promoters again exhibited the striking pattern of $prop_{CG}$ and $prop_{CHG}$ correlation, where the former is higher than the latter for 80% of observations (Figure 1.5D). The same relationships was evident between CG and CHH methylation (Figure 1.5F; Table 1.2) but not between CHG and CHH methylation (Figure 1.5E).

We expected a negative correlation between differential methylation and differential GE, and indeed the expected relationship was evident for both CG and CHG methylation (Table 1.3). With 1000 bp promoter regions, the correlation was as high as r = -0.0908 ($p = 4.99e10^{-14}$) for the subset of 7,704 differentially expressed genes (Figure 1.5G; Table 1.3). In contrast to CG and CHG methylation, $prop_{CHH}$ was significantly *positively* correlated with differential GE (Table 1.3), showing that higher CHH methylation relates to enhanced gene expression. Overall, for promoter regions we conclude that: *i*) CG and CHG methylation covary with differential GE in the expected direction, *ii*) that CG methylation explains up to ~9% of the variation in gene expression between tissues for differentially expressed genes, but *iii*) CHH methylation differs from the expected pattern.

TE methylation and GE: Because the methylation of TEs is known to suppress the expression of nearby genes (Lippman 2004; Hollister and Gaut 2013; Hollister et. al. 2011), we expected that differences in GE would correlate negatively with differential methylation of nearby TEs. That is, if a TE nearest to a gene is more highly methylated in floral bud, we predicted it should suppress GE in flowers, thus yielding a negative correlation in our analyses. We detected this predicted negative correlation but only in the CG context (Table 1.3). In contrast, correlations between differential GE and both $prop_{CHG}$ and $prop_{CHH}$ were

positive, with the $prop_{CHH}$ comparisons reaching statistical significance (Table 1.3). Across all contexts ($prop_{C}$), the relationship was also significantly positive, likely owing to the positive trends for $prop_{CHG}$ and $prop_{CHH}$ countervailing the trend for $prop_{CG}$. Finally, we also applied a linear model to disentangle the effects of methylation vs. the distance (in bases) of the TE from the gene (Table S1.4). In the linear model, the effect of methylation remained significant ($p < 10^{-3}$), but the effect of distance explained little and was not significant.

CONCLUSIONS

Although there is a widespread belief that methylation affects gene expression during development (Law et. al. 2010), relatively few studies have contrasted methylation and gene expression between tissues on a genomic scale. Moreover, BSseq data have rarely been replicated in these studies. Hence, our first goal was simple: to determine whether methylation between two tissues is, in fact, differentiated beyond the level expected from proper replication. For this comparison we chose two tissues that have been sampled commonly in other plant studies—leaves and floral buds. Overall, we were able to detect a significant signal of differentiation between tissue samples based on two methodological approaches (permutation tests and clustering analyses) and two measures of variation (DMSs and DMRs).

Nonetheless, a sobering observation was that the false positive rate (FPR) was extremely high for contrasts between single replicates. For DMS analyses, the lowest FPR in our analyses was 75%. In other words, had we based our inferences on single replicates, three-quarters of our inferences about the sites of "tissue-specific" methylation would have

been incorrect relative to inferences based on the larger, replicated dataset. The FPR for DMR analyses was similarly large, at least 80%. While there are ways to decrease the FPR statistically in theory, they may result in the cost of sensitivity and power. Such tradeoffs in the use of BSseq replicates are the topics of ongoing theoretical and algorithmic research, but thus far these render improvements only for data with less coverage those in this study (Sun et. al. 2014). Altogether, we conclude that reliance on single BSseq replicates may be misleading when the goal is to focus on specific DMRs or DMSs. For this reason, we recommend analyzes that summarize over a region–e.g., genes (Takuno and Gaut 2012) or promoters or TEs–as opposed to individual sites or individual DMRs. Moreover, because replication has been applied so rarely in plant studies, we hope that our description of within- and between-tissue replicates helps guide interpretation of the existing literature.

Although we detected significant methylation differentiation between tissues, our results were similar to previous studies in documenting that tissues are far more similar than different in their methylation patterns (Zemach et. al. 2010). For example, we detected ~37-fold more sites conserved between tissue samples than variable sites. Most of the observed differences occurred in the CHG context within TEs and promoters, but there were also slight biases in *total* methylation between leaf and floral bud. Overall, these observations add to the growing notion that methylation differences between plant tissues are slight, except for a few exceptional tissues, such as the endosperm and the pollen vegetative nucleus (Lauria et. al. 2004; Hsieh et. al. 2009; Ibarra et. al. 2012; Zemach et. al. 2010). Since neither of these tissues contributes to ensuing generations, these epigenetic changes may be of little evolutionary consequence, although it seems that the pollen

vegetative nucleus may play a role in generation-to-generation epigenetic reprogramming (Slotkin et. al. 2009).

We have shown that tissue-specific methylation differentiation is higher than variation among replicates, but do any of these methylation differences drive functional differentiation? To address this question, we generated RNAseq data for the same sets of replicates and examined the correlation between differential GE and differential methylation, many of which were significant. The most striking aspect of these results is that they vary by methylation context. In general, CG methylation correlates with GE as predicted: higher CG methylation in one tissue correlates with lower GE in that tissue. This relationship is true whether one examines genes, promoters or TEs (Table 1.3). In contrast, the results for CHH and CHG variation are more varied, with CHH methylation trending in the opposite direction than predicted for both promoters and TEs. These observations indicate that CG methylation is the primary component of variation to affect (or at least covary with) GE. This observation is consistent with the fact that genic expression in pines covaries with CG but not CHG methylation, even though pine genes are heavily methylated in both contexts (Takuno and Gaut 2012).

Another interesting aspect about methylation contexts is that they appear to be hierarchical, because typically neither CHH nor CHG variation exceeds CG methylation, regardless of the region under consideration (Figures 1.4A, 1.5D, 1.5E and 1.5F). These results suggest that CG methylation acts in some unknown way to limit methylation in the other contexts, at least in brachypodium. It remains to be seen whether this relationship holds for other species and additional tissues.

Overall, our study suggests that methylation patterns covary with tissue-specific expression, but also that differential CG methylation explains only a small proportion of tissue-specific variation in GE (i.e., between 1% and 9% of variation; Table 1.3). We note, however, that our study likely underestimates the magnitude of the effect, for at least two reasons. First, the predictive power will probably increase with the number of tissues sampled. An explicit goal of future studies should be to estimate the percentage of GE variation explained by DNA methylation based on a broader range of tissue-types; however, to do so will require better sampling–both in terms of tissues and replicates–than has been performed to date. Second, like most other papers in plant epigenetic research, our tissue samples undoubtedly included multiple cell types; indeed we suspect that the variation in cell types is the primary reason for high variation in DMSs and DMRs among biological replicates (Figure 1.1). A recent review has called to question the value of 'tissue' vs. 'cell' samples (Ji et. al. 2015). In the review, the authors argue that the signal of differentiation for highly specialized cells will be masked within tissue samples that contain multiple cell types. This may or not be true, as it depends critically on the as-yet-unknown pattern of cell differentiation and of course the cellular composition of tissue samples. Nonetheless, their point is well taken: it is possible that tissue, as opposed to cell-type, samples lead to underestimate of the overall contribution of epigenetic variation to gene expression.

MATERIALS & METHODS

BSseq Data and Mapping: The BSseq data were published previously (Takuno and Gaut 2013) and were available in the Short Read Archive (accession nos. SRX208151–

SRX208156). Briefly, three *B*. *distachyon* plants from the Bd21 line were grown under 20-h days to induce rapid flowering. Spikes and leaves were harvested at the beginning of anthesis. For each plant and tissue, ~two micrograms of genomic DNA was sonicated and purified using Qiagen DNeasy mini-elute columns (Qiagen). Sequencing libraries were constructed with the NEBNext DNA Sample Prep Reagent Set 1 (New England Biolabs, Ipswich, MA) but with methylated adapters in place of the genomic DNA adapters. Ligation products were purified with AMPure XP beads (Beckman, Brea, CA). DNA was bisulfite treated using the MethylCode Kit (Invitrogen, Carlsbad, CA) following the manufacturer's guidelines and then PCR amplified using Pfu Cx Turbo (Agilent, Santa Clara). Libraries were sequenced using the Illumina HiSeq 2000. The BSseq reads were mapped to the brachypodium reference genome (version 1.0) following (Takuno and Gaut 2013), which included filtering of low-quality reads and bases ($q < 20$) and mapping with BRAT software (Harris et. al. 2010). Mismatches for mapping were allowed only at potentially methylated sites.

mRNAseq data and analysis: RNAseq data were generated from the same tissue samples as BSseq (Takuno and Gaut 2013) and are publicly in the Short Read Archive (accession number SRP063465). RNAseq relied on total RNA isolation with the Qiagen RNeasy Kit, cDNA generation with the Ovation RNA-seq system v. 2 and library preparation with the Illumina TruSeq DNA Sample Prep. V2. The libraries were sequenced on the HiSeq2500 (100 cycle, single read) in the UCI High Throughput Genomics Facility in 2013. RNAseq reads were processed using Trimmomatic (v 0.30) to remove low quality reads (<20) and adapter sequence.

Analyses of RNAseq data was based on read mapping to the *B*. *distachyon* MIPs v.1.2 reference sequence, using TopHat (v1.49.0) (Trapnell et. al. 2012) with default parameters. In this analysis, reads were counted for each annotated gene, so long as that gene did not overlap with an annotated transposable element (see below). Reads were counted for each gene in each replicate, and then DESeq (v1.16.0) (Anders and Huber 2010) was employed to identify differential expression between tissues with a false discovery rate of $q < 0.01$. For the comparison of differential gene expression and differential methylation we used all genes that had the number of fragments per kilobase of transcript per million mapped reads (FPKM) > 0.02 in both tissues. The difference in gene expression was defined as [log2 (Flower_FPKM)/(Leaf_FPKM)], where Flower_FPKM and Leaf_FPKM were based on data from all three replicates.

Definitions of genomic features, DMSs, CMSs, and *prop* values: We used genome annotations to define genes, promoters and TEs. A gene was defined from the transcription start site (TSS) to the transcription stop site, including putative introns, using the MIPs (v1.2) annotation (Nussbaumer et. al. 2013). TEs were also based on the MIPs (v1.2) annotation. TEs that overlapped with genes were removed from analysis along with any genes that were contained in a TE. Gene annotations were also the basis for promoter annotation, which were defined as the 1.0 kb region upstream from the TSS.

To determine whether individual cytosines were methylated or unmethylated, we computed a binomial probability at a significance level of $p \leq 0.01$, following (Lister et. al. 2008). This probability required a rate of conversion error, which was calculated on contaminating chloroplast data and was ~1% (Takuno and Gaut 2013). The specific error rate for each tissue was found for each replicate and for each tissue (i.e., L1+L2+L3 and

F1+F2+F3; Table S1.1). Once a base was defined as methylated or unmethylated in each tissue, a base that was methylated in each tissue was deemed a conserved methylation site (CMS).

To identify Differentially Methylated Sites (DMSs), we applied a Fisher exact test (FET), which was based on a 2X2 table of the number of converted Cs to non-converted C's across the two samples (Lister et. al. 2008). A site was considered as differentially methylated between two samples–i.e., a DMS–when the FET yielded a p-value < 0.05.

DMRs were defined by the number of DMSs in a row that had a consistent direction of methylation bias (i.e., hypermethylation in leaf or flower), that were within 500 bp of each other and that were uninterrupted by a CMS or by a DMS in the opposite direction. We considered DMSs in all contexts (i.e., CG, CHG and CHH) to define DMRs. To assess significance, we calculated the length of DMRs (as defined by the number of unidirectional DMSs) expected to be found at random in the genome, given both the underlying distribution of cytosines in proper context and the numbers of DMSs and CMSs. To calculate the random expectation, we permuted DMSs and CMSs among genomic sites in their appropriate contexts, identified DMRs within permutated genomes, and ascertained DMR lengths. After permuting across the genone, we identified DMRs and noted the number of DMSs that constitute each DMR (Figure S1.1). DMRs that were of a length expected found at $p < 0.01$ in the permuted genome were considered 'significant' for analysis of observed data.

The final metric was the proportion of methylation or $prop_c$, which was used as a measure of methylation across a region. The *prop* value was determined by adding the total number of converted reads over the total number of reads for cytosines in a specific

context. The context could be CG ($prop_{CG}$), CHG ($prop_{CHG}$), CHH ($prop_{CHH}$) or all three contexts ($prop_{C}$).

Additional Statistical Analyses: To construct the trees in Figure 1.1, distance values were converted to Newick format and unrooted neighboring-joining trees were made using the ape and phyclust libraries in R (RDevelopment CoreTEAM 2010).

To determine whether DMRs were closer to differential expressed genes than other genes, we first labeled each gene as either differentially expressed or not. We then calculated both the observed distance from a differentially expressed gene to its closest DMR and its average across the genome. We then randomized the labels (differentially expressed or not) among genes within the genome and recalculated the average distance between a differentially expressed gene and its nearest DMR. The randomization was performed 1000 times to generate a distribution of the average distance from a DMR to a gene and to determine whether the observed average was extreme (Figure S1.4).

All correlations were based on cor.test in R, using the Spearman correlation.

FIGURES

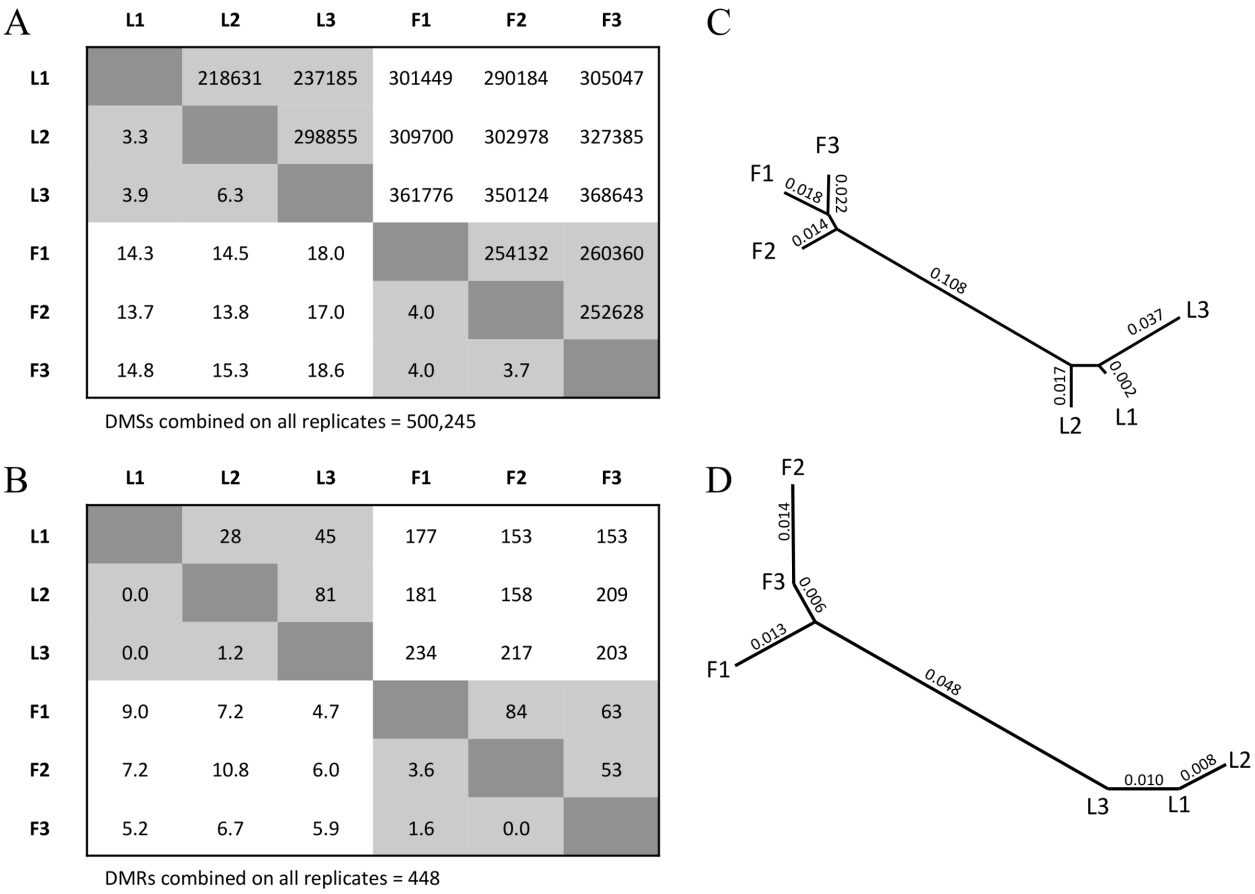

Figure 1.1: The inferred number of DMSs and DMRs between replicates. A) The upper matrix reports the number of DMSs between two BSseq replicates. The lower matrix reports the percentage of DMSs that map to the same location as the 500,245 DMSs inferred from the combined data sets. B) A neighbor-joining phylogeny representing the relationship among the six BSseq samples, based on distances defined by the lower matrix in A. C) The upper matrix reports the number of DMRs between two BSseq replicates. The lower matrix provides the percentage of DMRs that overlap with the 448 DMRs inferred from the combined data set. D) A neighbor-joining phylogeny representing the relationship among the six BSseq samples, based on distances defined by the lower matrix in C.

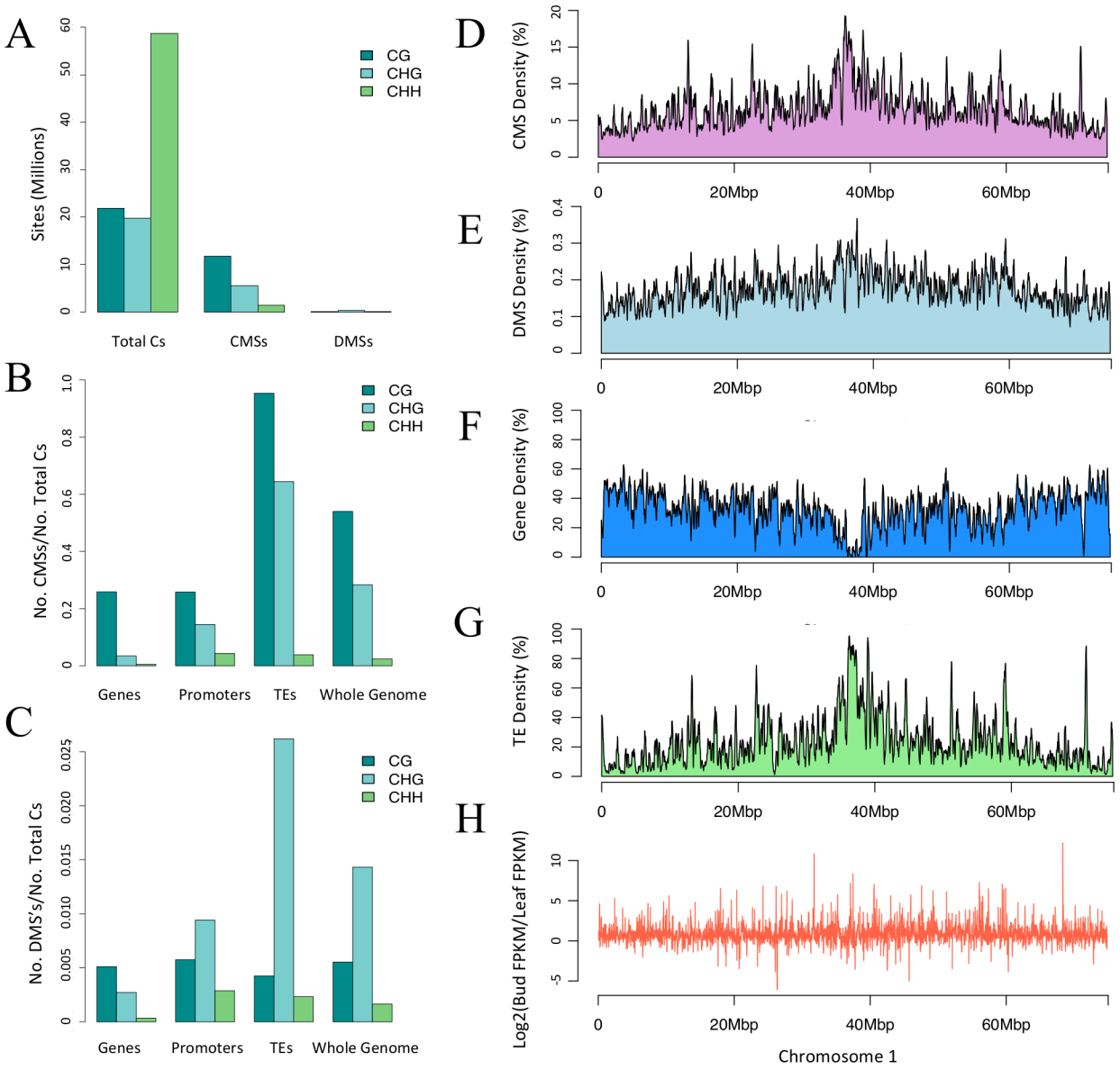

Figure 1.2: Context, direction, and regions of CMSs and DMSs. A) The number of sites in the correct context for methylation throughout the genome (Total Cs), along with the number of CMSs and DMSs in context. B) The proportion of CMSs relative to cytosines in the correct context for Genes, Promoters, TEs and the Whole Genome. C) The proportion of DMSs relative to cytosines in the correct context for Genes, Promoters, TEs and the Whole Genome. Note the difference in the scale of the y-axis between panels B and C. D to F) The graphs show the CMS, DMS, gene and TE density along chromosome 1. Density was

measured within a 50kb sliding window for smoothing. H) Differential gene expression plotted along the physical length of chromosome 1. The other chromosomes are represented in Figure S1.2.

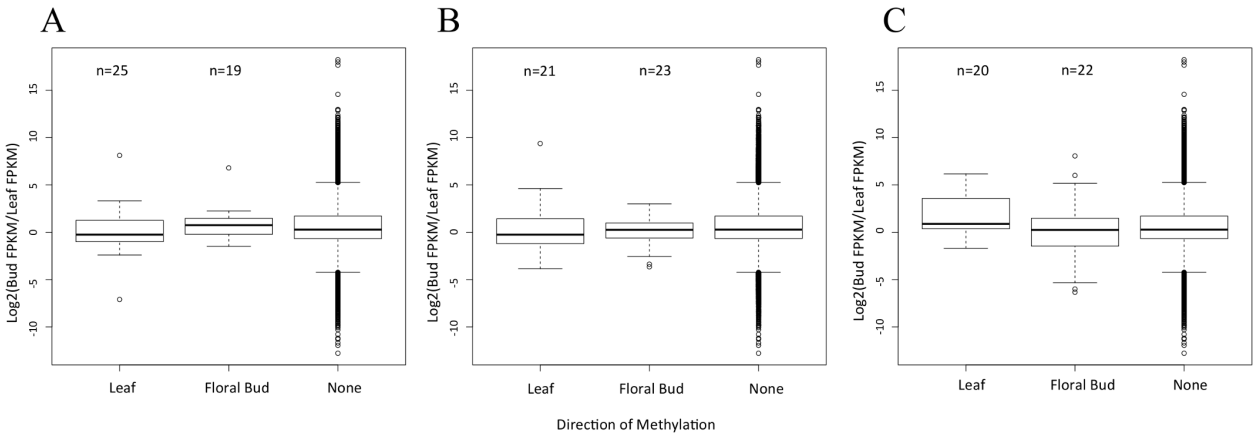

Figure 1.3: Gene expression with respect to DMRs and their direction. A) A graph of the distribution of gene expression when a DMR is located within a gene and hypermethylated in the Leaf or Floral Bud, or when there is no DMR in the gene (None). For the 25 genes hypermethylated in leaf, we predicted positive values on the y-axis, signaling higher expression in floral bud, but no bias was detected. For the 19 genes hypermethylated in floral bud, we predicted negative values on the y-axis, signaling higher expression in leaf, but again no bias was detected. B) The same graph of differential expression when the gene contains a DMR in its a promoter region. Again, there are no detectable biases in the direction of gene expression relative to genes that do not contain a DMR in their promoter region. C) A graph of differential gene expression when the TE nearest to a gene has a DMR that is hypermethylated in leaf, flower or no (None) DMR. For all graphs, the box plots represent the median, first, and third quartile. The whiskers represent the minimum and maximum. The numbers above the graph refer to sample size in each category.

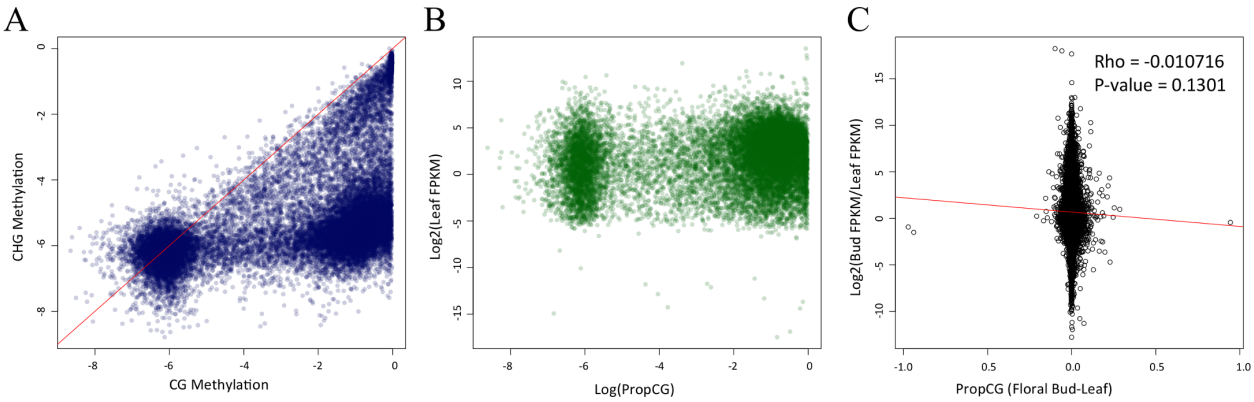

Figure 1.4: Methylation patterns within genes. A) The correlation between $prop_{CG}$ and $prop_{CHG}$ between genes for leaf tissue (r = 0.5977; p < 2.2e-16); floral bud tissue is not shown but the relationship is essentially identical. Methylation is plotted on a log scale. B) A comparison of $prop_{CG}$, on a log scale, and gene expression (FPKM) on a log2 scale within leaf (r = 0.2867; p <2.2e-16); again, floral bud tissue is not shown but essentially identical. C) A comparison of differential gene expression [log2fold (flower/leaf)] vs. the difference in $prop_{CG}$ between leaf and floral bud tissue.

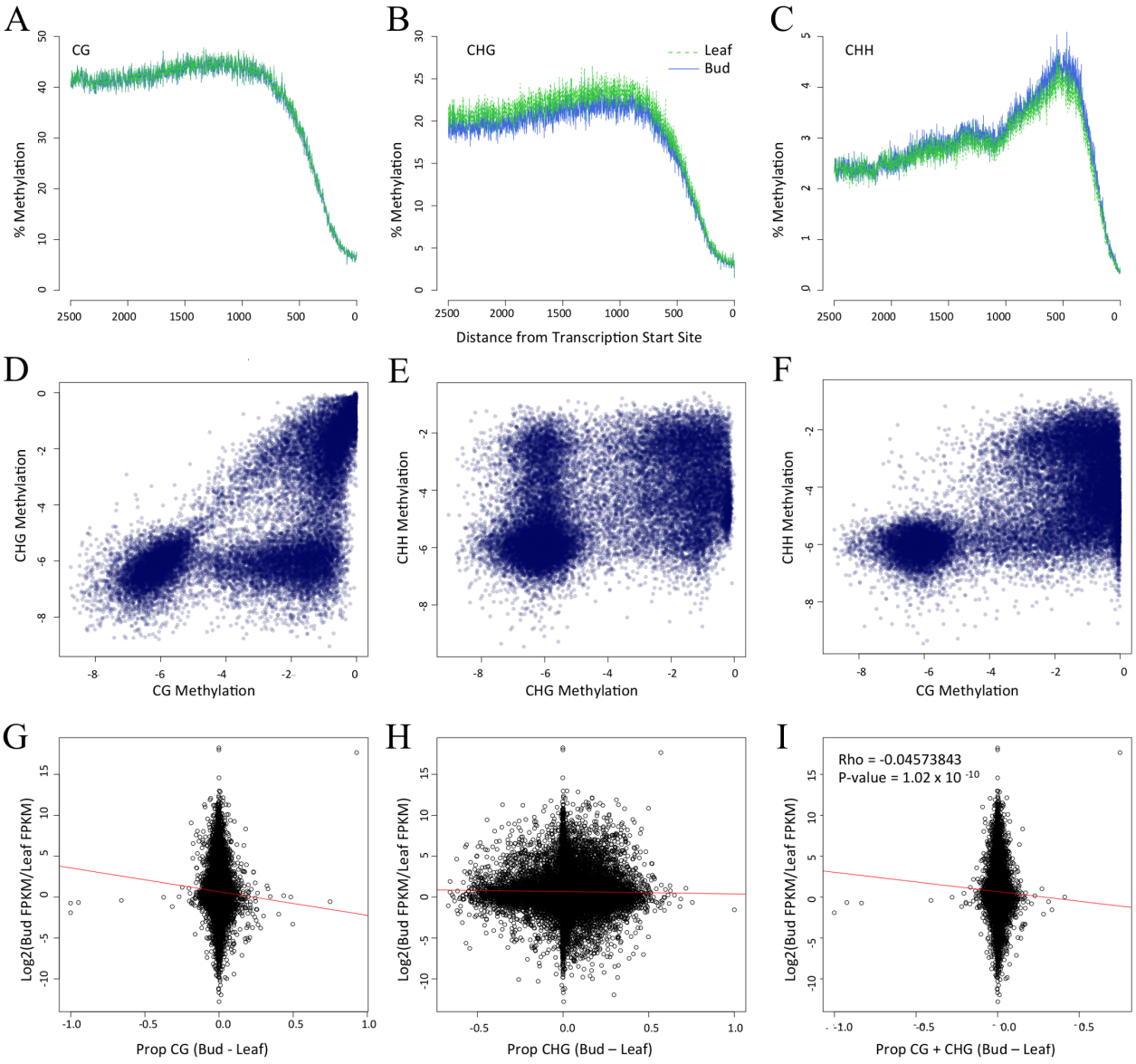

Figure 1.5: Methylation patterns within promoters and its relationship to gene expression. Graphs A,B and C present the level of CG, CHG and CHH methylation, respectively, in terms of distance from the Transcription Start Site. Graphs D, E and F compare methylation contexts, as measured by *prop* statistics in a log scale, within leaf tissue. Floral bud comparisons are not shown but are visually identical. Panels G, H and I compare differential gene expression [log2fold (FKPM_Flower/FKPM_Leaf)] vs. the

difference in prop between floral bud and leaf tissue. The correlation values for G and H are in Table 1.3.

TABLES

Table 1.1: The number of potential methylation sites, DMSs and CMSs in each of three sequence contexts (CG, CHG and CHH) throughout the entire *brachypodium* genome and also for three features separately (Genes, Promoters and TEs).

Region	CG	CHG	CHH	Total
		Potential methylation sites		
Genic	5,945,276	6,261,258	17,109,118	29,315,652
TE	7,066,299	5,524,931	17,168,841	29,760,071
Promoter	2,201,304	1,833,675	5,378,012	9,412,991
Whole Genome	21,814,767	19,722,162	58,692,551	100,229,480
		Differentially Methylated Sites (DMSs)		
Genic	30,396 (0.51)	17,059 (0.27)	6,009 (0.04)	53,464 (0.18)
TE	29,981 (0.42)	144,779 (2.62)	39,990 (0.23)	214,750 (0.72)
Promoter	12,658 (0.58)	17,276 (0.94)	15,500 (0.29)	45,434 (0.48)
Whole Genome	120,744 (0.55)	282,440 (1.43)	9,7061 (0.17)	500,245 (0.50)
		Conserved Methylated Sites (CMSs)		
Genic	1,537,550 (25.86)	216,844 (3.46)	88,467 (0.52)	1,842,861 (6.29)
TE	6,736,090 (95.33)	3,560,407 (64.44)	663,269 (3.86)	10,959,766 (36.83)
Promoter	568,236 (25.81)	265,069 (14.46)	232,488 (4.32)	1,065,793 (11.32)
Whole Genome	11,776,244 (52.98)	5,590,834 (28.35)	1,413,604 (2.41)	18,780,682 (18.74)

Numbers in parentheses represent the percentage of sites in context that are methylated. Those sites that are not DMSs or CMSs either lack evidence of methylation in both tissues or do not have a significant FET.

doi:10.1371/journal.pone.0150002.t001

Table 1.2: Spearman correlation coefficients between *prop* values within a tissue.

		Flower		Leaf	
		CHG	**CHH**	**CHG**	**CHH**
Genes	CG	0.5976	0.4393	0.5977	0.4416
	CHG	—	0.6089	—	0.6045
TEs	CG	0.3665	-0.1513	0.3558	-0.1743
	CHG	—	0.0639	—	0.06036
Promoters	CG	0.7970	0.5567	0.7495	0.5436
	CHG	—	0.7173	—	0.4293

The p-values of all coefficients are < 2.2 e-16 and significant after sequential Bonferroni correction.

doi:10.1371/journal.pone.0150002.t002

Table 1.3: Spearman correlations between the difference in *prop* values between tissues and the log2 fold change in gene expression.

		All Genes (19,956)		Differential Genes (7,704)[1]	
Region	**Context**	**Rho**	***p*-value[2]**	**Rho**	***p*-value[2]**
Gene	CG	-0.0107	0.1301	-0.0393	**0.0007**
	CHG	-0.0055	0.4358	-0.0064	0.5834
	CHH	0.0214	0.0025	0.0225	0.0529
	All	0.0045	0.5247	-0.0154	0.1857
Promoter (1kb)	CG	-0.0543	**1.696e-14**	-0.0908	**4.990e-15**
	CHG	-0.0357	**4.429e-07**	-0.0295	0.0112
	CHH	0.0767	**2.200e-16**	0.1183	**2.200e-16**
	All	0.0187	0.0083	0.0202	0.0821
TE	CG	-0.0438	**6.263e-10**	-0.0731	**3.018e-10**
	CHG	0.0205	0.0038	0.0262	0.0240
	CHH	0.0651	**2.200e-16**	0.0915	**3.152e-15**
	All	0.0454	**1.339e-10**	0.0659	**1.357e-08**

[1] Includes only the subset of genes that were significantly differentially expressed between tissues.

[2] **Bolded** values indicate correlations that remain significant at $p < 0.01$ after sequential bonferonni correction.

doi:10.1371/journal.pone.0150002.t003

SUPPORTING INFORMATION

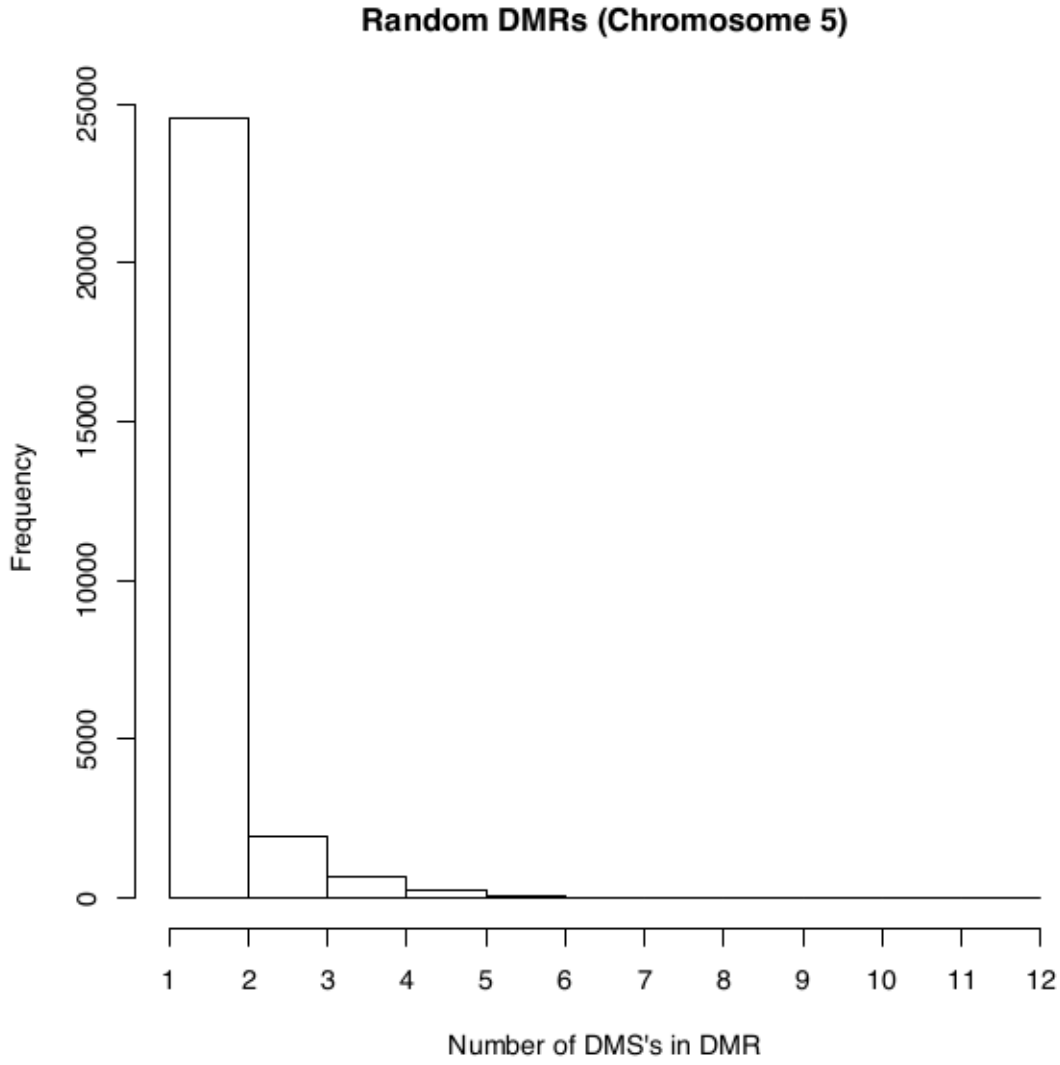

Figure S1.1: A histogram of the length of DMRs found after randomization of methylated cytosines within the *brachypodium* genome. Methylated cytosines were randomized in the proper context, and the number of DMRs in the same direction were counted. Within a randomized genome, a run of five or more methylated cytosines in length represented 1.3% of all potential runs; we defined a DMR to be ≥ 5 methylated cytosines in

the same direction, because this length represented a significant observation at the p ~0.01 threshold. See Materials and Methods for additional details.

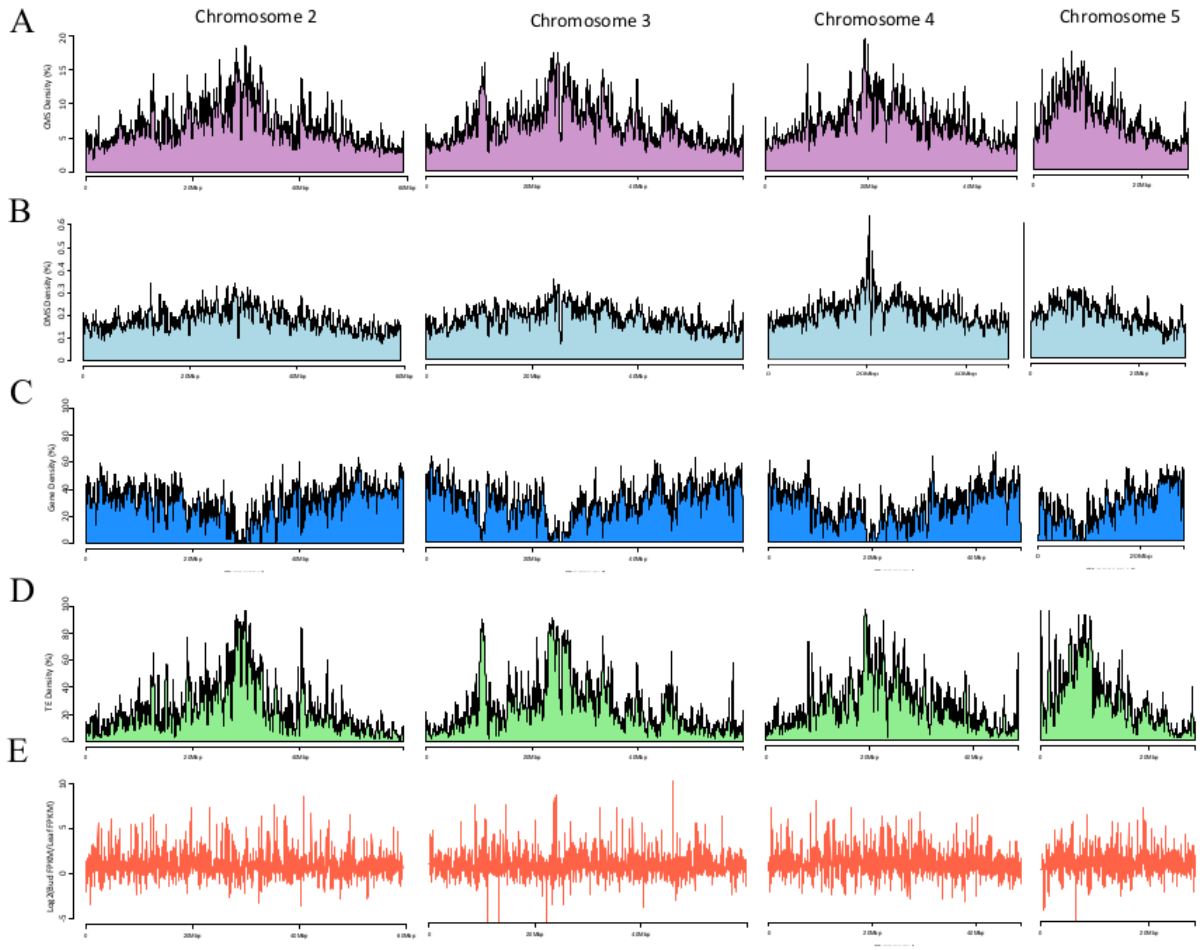

Figure S1.2: Plots of chromosomal densities of methylation features. Plots of chromosomal densities of A) CMSs, B) DMSs, C) genes, and D) TEs. Density was measured within a 50kb sliding window for smoothing. E) The graphs plot differential gene expression plotted along the physical length of chromosomes. This figure mimics Figure 1.2 of the main text, but includes the remaining four chromosomes.

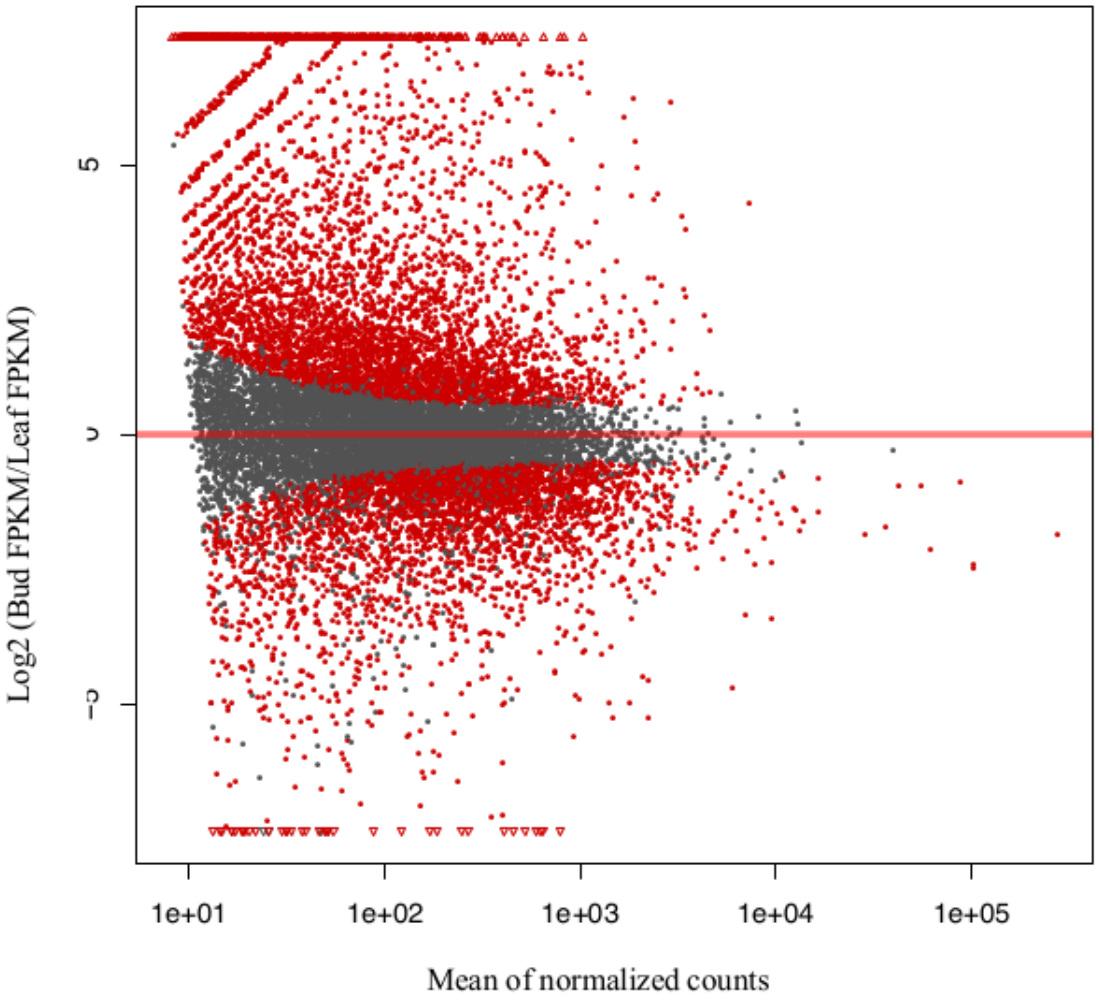

Figure S1.3: A volcano plot of the 26,072 genes tested for differential gene expression between leaf and floral tissue samples.

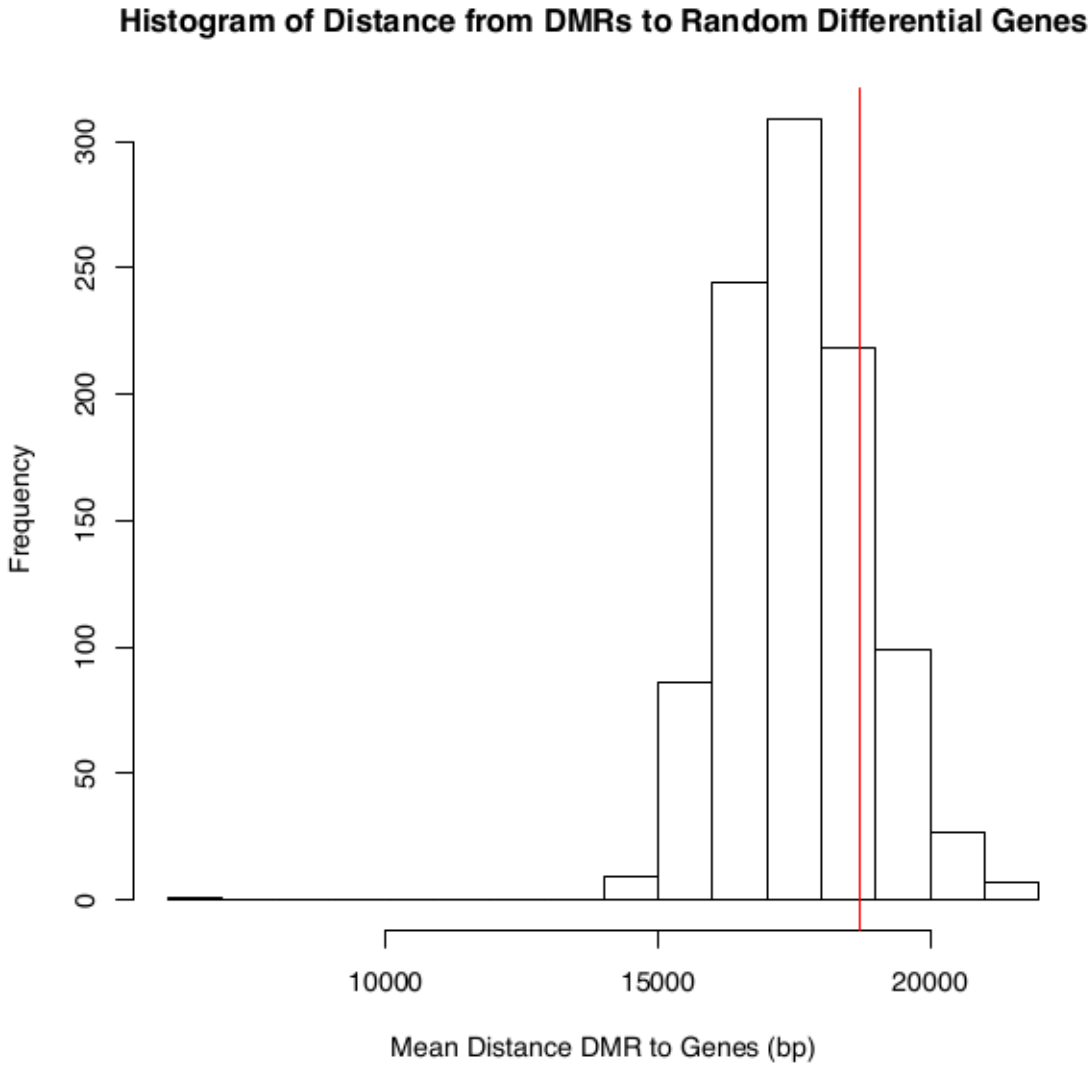

Figure S1.4: A histogram of the average distance between DMRs and genes. The histogram is based on 1000 randomizations (see Materials and Methods). The red line denotes the observed value.

Table S1.1: The estimate of the rate of conversion error. Estimates are based on analysis of chloroplast DNA in each replicate or on combined replicates.

Tissue/Replicate[1]	Error Rate
Leaf 1	0.0133180
Leaf 2	0.0106542
Leaf 3	0.0111591
Floral Bud 1	0.0086617
Floral Bud 2	0.0098665
Floral Bud 3	0.0094443
L1+L2+L3	0.0115213
F1+F2+F3	0.0092799

[1] Estimates for individual replicates are from (Takuno and Gaut 2013). Estimates for combined replicates are from the sum of reads from all leaf and floral bud replicates.

Table S1.2: A summary of RNAseq data. The table provides the number of reads after quality trimming, the number of reads that TopHat used to map for both left and right reads, and the maximum and minimum read lengths. The total number of transcripts is from based on output from cufflinks.

Replicate	No. In (Right)	No. Out (Right)	No. In (Left)	No. Out (Left)	Max Length	Min Length	Total
Leaf 1	18,160,475	18,143,542	18,160,475	18,135,783	100	36	146,973
Leaf 2	26,428,934	26,420,118	26,428,934	26,415,012	100	36	137,058
Leaf 3	17,868,229	17,848,933	17,868,229	17,848,410	100	36	142,192
Flower 1	14,734,746	14,723,047	14,734,746	14,720,678	100	36	163,786
Flower 2	24,718,807	24,700,755	24,718,807	24,690,319	100	36	177,207
Flower 3	12,599,231	12,580,503	12,599,231	12,578,360	100	36	160,370

Table S1.3: Go enrichment terms for differentially expressed genes between leaf and flower. Only the enrichments terms with a p-value < 0.01 are given.

GO Term		**P-value**
GO:004867	Serine-type endopeptidase inhibtor activity	0.0054
GO:0016706, GO:0010302	Oxidoreductase activity, acting on paired donors, with incorporation or reduction of molecular oxygen, 2-oxoglutarate as one donor, and incorporation of one atom each of oxygen into both donors	0.0010
GO:0016307	Phosphatidylinositol phosphate kinase activity	0.0072
GO:0046488, GO:0030384	Phosphatidylinositol metabolic process	0.0072
GO:0005874	Microtubule	0.0028
GO:0006184	GTP catabolic process	0.0006
GO:0007017	Microtubule-based process	0.0020
GO:0043234	Protein complex C	0.0007
GO:0051258	Protein polymerization	0.0002
GO:0044267	Cellular protein metabolic process	0.0087
GO:0042626	ATPase activity, coupled to transmembrane movement of substances	0.0056

GO:0006633, GO:0000037	Fatty acid biosynthetic process	0.0012
GO:0007165, GO:0023033	Signal transduction	0.0029
GO:0003777	Microtubule motor activity	1.418e-06
GO:0007018	Microtubule-based movement	1.418e-06
GO:0003924	GTPase activity	0.0086
GO:0016887, GO:0004002	ATPase activity	0.0034
GO:0004553, GO:0016800	Hydrolase activity, hydrolyzing O-glycosyl compounds	0.0027
GO:0005975	Carbohydrate metabolic process	0.0002
GO:0005840, GO:0033279	Ribosome	1.505e-14
GO:0055085	Transmembrane transport	0.0055
GO:0003735, GO:0003736, GO:0003738, GO:0003739, GO:0003740, GO:0003741, GO:0003742	Structural constituents of ribosome	8.320e-15
GO:0006412, GO:0006416, GO:0006453, GO:0043037	Translation	1.987e-12
GO:0016491	Oxidoreductase activity	0.0077
GO:0016021	Integral to membrane	0.0096
GO:0005622	Intracellular	2.428e-08

GO:0004672, GO:0050222	Protein kinase activity	0.0003
GO:0016020	Membrane	0.0012
GO:0006468	Protein phosphorylation	0.0003
GO:0055114	Oxidation-reduction process	0.0069
GO:0005515, GO:0045308	Protein binding	0.0005
GO:0005524	ATP binding	7.660e-11

Table S1.4: Results of the application of linear models. The models test for an effect of the TE distance to a gene and of TE methylation to differential gene expression.

	Coefficient Methylation	P-value[1]	Coefficient Distance	P-value
*Prop*C	-0.7233***	**0.000118**	3.895e-06	0.398970
DMSs	-0.0193	0.102	4.040e-06	0.382

[1] bolded values denote significance after sequential Bonferroni correction.

Chapter 2: Modeling interactions between transposable elements and the plant epigenetic response: a surprising reliance on element retention

Angiosperm genomes vary more than 1000-fold in size, and this variation correlates strongly with transposable element (TE) content. For plant species with small genomes, like *Arabidopsis thaliana* or *Brachypodium distachyon*, DNA derived from TEs constitute 20-30% of the genome (Vogel et. al 2010; Kaul et. al. 2010). Species with larger genomes have commensurately larger proportions of TE-derived DNA. For example, TE-derived DNA represents >85% of the barley (*Hordeum vulgare*) and maize (*Zea mays* ssp. *mays*) genomes (Wicker et al., 2005, Schnable et al., 2009). When one considers that the average size of a diploid angiosperm genome is similar to that of barley genome, at 6400Mb, then it is clear that most extant plant DNA is derived from TEs (Tenaillon et al., 2010).

Despite the obvious evolutionary success of TEs, their proliferation is checked by their plant host. The two entities engage in a continuous arms-race, where TEs seek to proliferate and hosts attempt to control them (Lisch and Slotkin, 2011). In fact, most (if not all) TEs are epigenetically silenced under normal conditions (Lisch, 2009). The plant host exerts this control by suppressing TE activity both before and after transcription. Post-transcriptional modification relies chiefly on *RNAi* that recognizes and degrades TE mRNA produced by RNA polymerase II (*Pol II*). Degradation requires associated factors like *RNA-polymerase 6* (RDR6), which converts single-stranded to double-stranded RNA (dsRNA); the Dicer-like proteins *DCL2* and *DCL4* that cleave dsRNAs to produce 21 and 22 nucleotide (nt) small interfering siRNAs; and the *Argonaute1* (*AGO1*) protein that guides

siRNAs to mRNAs for cleavage (Fultz et al., 2015). Presumably, 21-22nt siRNAs can prime multiple cycles of mRNA cleavage, but they may have another important function, which is to initiate transcriptional silencing (Nuthikattu et al., 2013, McCue et al., 2015). Hence, 21-22nt siRNAs can be seen as dual-purpose, because they are involved in post-transcriptional silencing and also because they initiate DNA methylation (Cuerda-Gil and Slotkin, 2016).

Transcriptional silencing is achieved through epigenetic modifications like DNA methylation, histone modifications and shifts in nucleosome positioning (Bernatavichute et al., 2008, Chodavarapu et al., 2010). The first of these, DNA methylation, relies on the RNA-directed DNA methylation (RdDM) pathway. RdDM begins when the plant- specific *Pol IV* transcribes a TE. The resulting single-stranded RNA is processed into 24nt siRNAs by *RDR2* and *DCL3*, two homologs that are distinct from those employed in *RNAi.* Ultimately, the 24nt siRNAs guide protein complexes to homologous DNA sequences that are then targeted for cytosine methylation. Once DNA methylation is established, at least two mechanisms act to maintain it. The first is a positive feedback loop: *Pol IV* and *Pol V*, the RNA polymerases involved in RdDM, preferentially act on methylated DNA (Law et al., 2013, Johnson et al., 2014), thereby reinforcing silencing (Panda and Slotkin, 2013, Bousios and Gaut, 2016). The second is the maintenance of symmetric CG and CHG (where H = A, C, or T) methylation during DNA replication and cell division (Law and Jacobsen, 2010). In theory, then, once a TE is targeted for DNA methylation, the host genome employs feedbacks to ensure that the TE reaches and maintains a quiescent state.

Numerous molecular studies have dissected the *RNAi* and RdDM pathways (reviewed in (Law and Jacobsen, 2010, Fultz et al., 2015, Matzke et al., 2015)), but several important questions remain about systems0level interactions between TEs and their plant

hosts. One major question is why the host relies on two mechanisms – i.e., *RNAi* and RdDM – to silence TEs. Presumably both pathways are capable of silencing; they are thus overlapping and potentially redundant. Moreover, both pathways are energetically costly, because they require the production of myriad polymerases, methylases and small RNAs (Bousios and Gaut, 2016). Why are two pathways maintained despite an energetic cost? One working hypothesis is that the two pathways act synergistically, but this hypothesis has yet to be explored.

A second major question concerns 24nt siRNAs. As mentioned above, 24nt siRNAs are predominantly produced by the RdDM pathway, which preferentially acts on TEs that have already been targeted for silencing. An important feature of these 24nt siRNAs is that they can act in *trans* to guide the methylation of TEs that have similar sequence characteristics to the original TE template (Slotkin et al., 2005, Teixeira et al., 2009, Ito et al., 2011, Ye et al., 2012, Fultz et al., 2015). Under this process, 24nt siRNAs may constitute a kind of 'immune memory' that act as a buffer against the possibility of TE activity (Fultz et al., 2015). If true, this implies that the strength of the host epigenetic response is proportional to the number of similar TEs in the genome that have already been silenced. Yet, no studies have explored the potential co0dependence between TE copy numbers and the strength of the host response.

Our final systems-level question concerns a separate process that occurs in cells associated with (but not part of) the germline. In cells such as the pollen vegetative nucleus (Slotkin et al., 2009), some TEs are actively demethylated, expressed, and utilized to produce 21-22nt siRNAs. These siRNAs are then transported to the germline, where they presumably contribute to stable TE silencing across generations (Slotkin et al., 2009, Ibarra

et al., 2012, Martínez et al., 2016, Martinez and Köhler, 2017). But what is the systems-level benefit of this additional step in the host response, given that there are already at least two overlapping pathways dedicated to silencing TEs and also that symmetric DNA methylation is typically inherited faithfully?

Here, we address these questions by building a model of host:TE interactions based on ordinary differential equations (ODEs). ODE models have been used widely to study biological phenomena that range from population growth (Malthus, 1798), to predator0prey interactions (Volterra, 1926), to the dynamics of viral infection and reproduction (Perelson, 2002). ODE models have also studied the interactions between TEs and the host response, but without a focus on plants and with few details of host response mechanisms (Abrusán and Krambeck, 2006). Our model includes proxies for RNAi, RdDM and additional factors like TE propagation and TE deletion. We study properties of the model but also estimate reasonable biological parameters by fitting the model to biological data, specifically from the study of the accumulation of the *Evade* element in an *A. thaliana* inbred line (Mari-Ordonez et al., 2013). Given these parameter estimates, we explore dynamics of the model and address systems0level questions about host:TE interactions. We focus on three sets of questions: *i*) Are both pre0 and post-transcriptional silencing necessary to control TEs? If not, what advantage is gained by having two mechanisms? *ii*) Given that methylated TEs may be an important source of immune memory, does TE deletion affect the dynamics of the host response? And, finally, *iii*) What is the added benefit of a third mechanism for generating 21-22nt siRNAs in the germline?

RESULTS

A model of TE propagation and silencing: Our model assumes that a TE begins as single copy and expresses mRNA at rate *v* (Figure 2.1; Table 2.1). Among the produced mRNA, a proportion *p* is transposed into new genomic copies of the TE per host generation. Another proportion, ε, of the TE mRNA is processed into 21-22nt siRNAs. Note that $p + \varepsilon \leq 1.0$ under our model. We assume that the 21-22nt siRNAs degrade at rate δ and initiate TE silencing at rate *i*. Initiation encompasses both post-transcriptional silencing (*RNAi*) and the onset of methylation, following previous models (Nuthikattu et al., 2013, McCue et al., 2015). Finally, 24nt siRNAs reinforce methylation at rate *r*, representing RdDM. In our model, 24nt siRNAs are *trans*-acting and thus may affect numerous TE insertions, including active elements. Overall, active TEs (aTEs) may become silenced TEs (sTEs) through 21-22nt siRNAs, 24nt siRNAs, or by a combination of both (Figure 2.1).

The model includes two additional parameters. The first is TE deletion from the genome, which occurs at rate *d* for both aTEs and sTEs. The second is the loss of silencing from TEs over time (*e.g*. through the loss of methylation), which can lead to reactivation of TEs at rate *u*. Methylation loss has been shown to occur in mutation accumulation lines (Schmitz et al., 2011), and rateu is included to reflect this biological process. When $u = 0$, maintenance of silencing is perfect, but silencing is not maintained when $u = 1$. The model is represented diagrammatically in Figure 2.1 and consists of three differential equations:

$$\frac{d(aTE)}{dt} = (v * p - d - i * siRNA - r * sTE) * aTE \; + \; u * sTE$$

$$\frac{d(sTE)}{dt} = (i * siRNA + r * sTE) * aTE - (d + u) * sTE$$

$$\frac{ds(siRNA)}{dt} = \varepsilon * v * aTE - \delta * siRNA$$

The first equation describes the change in the number of aTEs over time; the second describes the change in the number of sTEs over time, and the third monitors numbers of 21-22nt siRNAs over time. While these three equations represent our basic model, Figure 2.1 includes a dashed arrow representing a fourth process, the epigenetic remodeling of TEs in the germline. This process will be incorporated after we first explore the dynamics of the basic model.

Model Equilibria: Once a TE has invaded a host it has three possible fates: it may fail to successfully invade and be lost completely; it may establish itself and reach an equilibrium number of copies over time; or it may expand in copy number unabated. An advantage of ODE models is that we can analytically solve the equilibrium points to understand TE invasion behavior and parameter dependence. We identified two equilibrium points in our system. The first is when there are no Tes and, hence, no 21-22nt and 24nt siRNAs in the host. That is, the equilibrium points for the active copies (aTE_{eq}), silenced copies (sTE_{eq}) and siRNA ($siRNA_{eq}$) are equal to zero. Stability around this point provides information as to whether a TE will successfully invade the genome or be lost. We investigated stability (Methods; see Equation 3) and found that it does not rely on any of the parameters associated with epigenetic processes 0 i.e., I, ε or r. Instead, stability relies only on the parameters for TE expression, propagation, and deletion (v, p, and d). Although Equation 3 is complex, the Jacobian matrix (see Methods) suggests the intuitive notion that invasion proceeds when expression and propagation ($v*p$) outcompetes deletion (d).

Once a TE has established its presence in the host, it may increase in number until the second equilibrium point (see Methods). The equilibrium points for aTEs and sTEs are given by:

$$\frac{d(aTE)}{dt} = (v * p - d - i * siRNA - r * sTE) * aTE \; + \; u * sTE$$

$$\frac{d(sTE)}{dt} = (i * siRNA + r * sTE) * aTE - (d + u) * sTE$$

$$\frac{ds(siRNA)}{dt} = \varepsilon * v * aTE - \delta * siRNA$$

The first equation describes the change in the number of aTEs over time; the second describes the change in the number of sTEs over time, and the third monitors numbers of 21-22nt siRNAs over time. While these three equations represent our basic model, Figure 2.1 includes a dashed arrow representing a fourth process, the epigenetic remodeling of TEs in the germline. This process will be incorporated after we first explore the dynamics of the basic model.

Model Equilibria: Once a TE has invaded a host it has three possible fates: it may fail to successfully invade and be lost completely; it may establish itself and reach an equilibrium number of copies over time; or it may expand in copy number unabated. An advantage of ODE models is that we can analytically solve the equilibrium points to understand TE invasion behavior and parameter dependence. We identified two equilibrium points in our system. The first is when there are no TEs and, hence, no 21-22nt and 24nt siRNAs in the host. That is, the equilibrium points for the active copies (aTE_{eq}), silenced copies (sTE_{eq}) and siRNA ($siRNA_{eq}$) are equal to zero. Stability around this point provides information as to whether a TE will successfully invade the genome or be lost. We investigated stability (Methods; see Equation 3) and found it does not rely on any of the parameters associated with epigenetic processes - i.e., *i*, *e* or *r*. Instead, stability relies only on the parameters for TE expression, propagation, and deletion (*v, p,* and *d*). Although Equation 3 is complex, the Jacobian matrix (see Methods) suggests the intuitive notion that

invasion proceeds when expression and propagation ($v*p$) outcompetes deletion (d), leading to a possible invasion.

Once a TE has established its presence in the host, it may increase in number until the second equilibrium point (see Methods). The equilibrium points are given by:

$$aTE_{eq} = \frac{1}{\frac{r}{d} - \frac{i*\varepsilon}{(\frac{d}{v} - p)}} \qquad \textbf{[Equation 1]}$$

$$sTE_{eq} = \frac{\left(\frac{v*p}{d}\right) - 1}{\frac{r}{d} - \frac{i*\varepsilon}{(\frac{d}{v} - p)}} \qquad \textbf{[Equation 2]}$$

with $siRNA_{eq}$ given by Equation 4 (see Methods). These two equations illustrate that aTEs and sTEs have similar parameter dependencies. However, equilibrium values of sTEs depend more explicitly on v and p in the numerator than does the equilibrium values of aTEs. This is an interesting observation because v and p are properties of aTEs; it drives home the point that equilibria copy number of sTEs relies intricately on the properties of their active counterparts. The denominator of the two equations clearly indicates that increasing r tends to decrease both aTE_{eq} and sTE_{eq}. Finally, the equations also hint at a complex relationship between equilibrium copy numbers and d, because the latter appears twice in the denominator (and once in the nominator for sTE_{eq}). As dincreases, these appearances have opposite effects on equilibrium values.

To explore these and other parameter dependencies further, we first fit the model to biological data and then perturbed parameter values separately to assess their effects on TE copy numbers.

Fitting the model to biological data: It can be difficult to identify biologically reasonable parameter values for ODE models. To address this concern, we fitted our model to experimental data from the study of Mari-Ordonez et al. (2013) (Mari-Ordonez et al., 2013), who characterized the expression and transposition of a single-copy of the *Evade* retroelement that had become unmethylated in *A. thaliana met1*-mutant epigenetic recombinant inbred lines. By following two lines to generations 14 and 15, they showed that *Evade* was highly expressed until generation 11 and 7, respectively, after which expression plummeted precipitously, presumably due to host silencing. The number of *Evade* copies increased rapidly while its expression was high, to a maximum of ~40 copies after 11 and 7 generations.

To fit our model to their data, we extracted information about *Evade* copy numbers and relative expression (see Methods). We focused on one inbred line (*met1*) from their study, because this was the only line for which data were sampled for consecutive generations: in total seven generations (from 8 to 14) since the reactivation of the single *Evade* element. We fitted the model to the *Evade* data with a Monte Carlo approach that concurrently considered the total TE copy number (i.e., the combined total of aTEs and sTEs) and expression. Our set of fitted parameter values are reported in Table 2.1. These parameter values produced a good fit to the copy number data, and a curve of similar shape to the observed relative expression data over time (Figure 2.2A). [Note that our measure of expression is only a proxy for expression measured experimentally; see Methods.] We recognize that we have fitted a complex model to relatively simple data and that our fitted parameters may represent one of many potential reasonably fitting parameter sets. They nonetheless provide a biologically-plausible foundation for examining model behavior.

Model behavior under fitted parameters: Given the fitted parameters, we explored host:TE dynamics over 500 generations, monitoring numbers of aTEs, sTEs and total TE copy number (=aTEs + sTEs) (Figure 2.2B). With these parameter values, the model produces oscillations of all three entities for ~200 generations until it reaches an equilibrium. The oscillations of aTEs and sTEs are somewhat out of phase with one another. We interpret these results as reflecting feedbacks in the epigenetic system. When a TE first invades a host, the combination of expression (fitted value v = 1.63; Table 2.1) and propagation (p = 0.340) create an initial burst in TE copy number. If TEs were able to grow unabated, there would be an exponential increase at a rate of 0.554 (=v*p) TEs per generation. However, some transcripts are processed into 21-22nt siRNAs (ε = 0.051)that silence TEs at rate i = 0.062. These 21-22nt siRNAs degrade quickly for each host generation (δ=0.999), and therefore any new 21-22nt siRNAs are not residual, but must be made from active TEs. However, once initiation of methylation has begun as part of i, reinforcement quickly takes hold at rate r = 0.025. Eventually, the number of sTEs increases and the number of aTEs decreases, so that total expression begins to decline. As TEs become silenced, they have two fates under our model: they can be deleted from the genome or become active again due to loss of silencing (Figure 2.1). Since loss of silencing was very low (u = 4 x 10^{-6}) in the fitted parameter set, the main fate of sTEs is to be deleted (d = 0.16). As these quiescent TEs are lost, so is the source of reinforcing 24nt siRNAs. When reinforcement becomes unreliable, the host loses epigenetic control, the subset of remaining aTEs propagate, and the phased cycle begins again. These cycles dissipate in amplitude until equilibria are reached at ~20 total TE copies, with more sTEs (~14) than

aTEs (~6) (Figure 2.2B). It is important to note that the equilibrium is not necessarily static; it can be reached when equal numbers of TEs are created vs. deleted.

These phased interactions occur with the fitted parameters but also occur regularly with other parameter combinations. They are not, however, a necessary outcome of the model (see examples below).

Examining initiation (*i*) and reinforcement (*r*): We have shown that the model can have complex, oscillating dynamics based on parameters inferred from biological data. These parameters can be modified independently to explore the importance of various processes. In this section, we assess the effect of perturbing the system by varying either initiation (*i*) or reinforcement (*r*), or both, while holding the remaining parameters to the values estimated from the *Evade* data. We first set $i = 0$, and the result was both intuitive and trivial. With $i = 0$ silencing never begins. Hence, the number of aTEs trended upward at an exponential rate, with no resulting sTEs (Figure 2.3A).

The effect of setting r to zero was less straightforward, because i was > 0 and hence silencing was initiated. Without reinforcement, copy numbers no longer oscillated, but instead burst and rapidly reached a maximum for both aTEs and sTEs. These copy numbers remained flat, implying a steady state in which silencing was initiated by 21-22 siRNAs and there was sufficient transposition to counteract TE deletion. Under these parameter values, the steady state of sTEs was higher than that of aTEs (Figure 2.3A), as with the equilibria reached with fitted parameters (Figure 2.2B).

If initiation by 21-22nt siRNAs is sufficient to reach a steady state and to control TEs, then what is the advantage of reinforcement by 24nt siRNAs? To address this question, we varied *i* and *r* across their parameter ranges and assessed total copy numbers (=aTEs +

sTEs). To help characterize effects, we focused on two descriptive statistics, TE_{max} and TEf_{inal} (see Figure 2.2B). TE_{max} is the highest total TE copy number achieved under a set of model parameters, and TEf_{inal} is the total copy number after 5,000 generations, a point by which total aTE and sTE copy numbers have typically reached a steady0state. When varying *i* and *r*, we found that their relationship was non-linear (Figure 2.3B). Briefly, if *r* ≥0.5 then any change in *i* had little effect on TE_{max} and TE_{final}, so long as there was at least some initiation. In contrast, when r was low (e.g., *r*≤ 0.1), the value of i had notable effects on both TE_{max} and TE_{final}. For example, when *r* = 0.001, TE_{max} varied over 2 orders of magnitude as a function of *i*. Similarly, TE_{final} differed ~330fold when *i* ranged from 0.001 to 0.99 (Figure 2.3B). This relationship implies that reinforcement can counter TE propagation efficiently, even when initiation of silencing is weak. This observation held true when also adjusting for TE expression (*v*) and deletion (*d*) (Figure S2.1).

The effects of TE deletion (*d*): Theoretically, high TE deletion rates should be advantageous for the plant host, because they limit opportunities for transposition and consequent deleterious mutations. However, high amounts of TE deletion could have consequences for immune memory, because quiescent TEs may be a major source of trans-acting 24nt siRNAs (Teixeira et al., 2009, Ito et al., 2011, Fultz et al., 2015). Hence, high deletion rates may adversely affect the epigenetic response. To illustrate the effect of deletion on TE copy numbers, we varied the deletion parameter *d* from 0.001 to 0.99 (Figure 2.4), while holding the remaining parameters to their fitted values (Table 2.1).

The model produced four noteworthy results. First, when TE deletion was very low (*d* =0.001), aTEs burst quickly to high copy number (~30). After peaking at a total copy number of ~80, all TEs were silenced and the population of sTEs declined slowly over time,

reflecting the low rate of deletion (Figure 2.4). Throughout this process, there were no aTEs after the initial burst. Second, when d increased ($0.01 \leq d < 0.5$), the system generated oscillations in the number of aTE and sTEs. The amplitude, frequency, and equilibrium values (TE_{final}) vary with d. Note that the running average of sTEs exceeded that of aTEs for these parameter values (Figure 2.4). Third, when TE deletion was at intermediate levels ($d = 0.5$), aTEs reached a steady-state, but there were very few sTEs. Finally, when the rate of TE deletion was very high ($d = 0.99$), all TEs were removed from the genome. Overall, we interpret these results to convey a somewhat counterintuitive idea: if the goal is to have few aTEs, then it is beneficial either to have dramatically high rates of TE deletion (e.g., $d = 0.99$) or to have such low (e.g., 0.01-0.1) deletion rates to preserve a reservoir of sTEs that contribute to reinforcement of silencing. This supports our observation, based on equilibrium equations (Equations 1 and 2), that deletion plays a complex role in determining aTE_{eq} and sTE_{eq}.

Additional parameters: We also varied values of expression (v), propagation (p), and loss of silencing (u), while the remainder of the parameters were held at their fitted values. The parameter v was arbitrarily ranged between 0 and 5. The chief effect of this range was on the amplitude and periodicity of TE oscillations. Higher expression levels led to more dramatic copy number oscillations (Figure S2.2). Importantly, at low parameter values (e.g., $v \leq 0.5$) TEs either did not invade the genome or were maintained at very low copy numbers (<5 total TEs) over the long term. Varying p produced results similar to varying v (Figure S2.3). Increasing p did, however, tend to lead to higher TE_{max} and average copy numbers relative to the parameter values we explored for v (Figure S2.2). This was presumably because there is a trade off with v; as it increases, so does the production of 21-

22nt siRNAs, which then potentially affect *RNAi*. Propagation (p), on the other hand, contributes only to the proliferation of more TEs. Note that low levels of propagation ($p < 0.25$) resulted in no invasion. Hence, TEs cannot invade if expression or propagation are low.

Our model also assumes a process of silencing loss (u), for which the most likely example is methylation loss. Methylation loss is known to be low based on empirical data because symmetric methylation is typically maintained faithfully through cell division (Becker et al., 2011). Indeed, our fitted parameter estimate was $u=4 \times 10^{-6}$, suggesting that a very low amount of sTEs become aTEs due to, for example, leaky maintenance of symmetrical methylation. Overall, we found that varying the u parameter had little effect on model behavior at parameter values < 0.01 (Figure S2.4). This implies that variation in spontaneous demethylation rates are likely to have few effects on the dynamics of host:TE interactions unless u varies by several orders of magnitudes from our fitted estimate.

TE reactivation dampens TE oscillations: Finally, we incorporated an interesting biological observation – i.e., the fact that TEs are activated in some reproductive tissues, ostensibly to ensure the transmission of a complement of siRNAs to egg and sperm (Slotkin et al., 2009, Ibarra et al., 2012, Martínez et al., 2016, Martinez and Köhler, 2017). TEs are known, for example, to be demethylated and reactivated in the pollen vegetative nucleus, which accompanies the sperm cell, but does not contribute DNA to the fertilized zygote. The reactivated TEs are sources of 21-22nt sRNAs that are transported to the sperm and presumably target silencing of TEs in the zygote (Slotkin et al., 2009). The net effect of this process is to increase the numbers of 21022nt siRNAs in germline cells; these 21-22nt sRNA originate not only from aTEs, but also from sTEs (see below).

We added this mechanism to our model with an equation that increases the number of 21-22nt siRNAs in the system at a level proportional to the number of sTEs that were demethylated in the companion cells. That is,

$$\frac{dsiRNA}{dt} = \varepsilon * v * (TE + mTE) - \delta * sRNA$$

This equation is represented by the dotted arrow in Figure 2.1. We evaluated the effects of this additional process on the system with fitted parameter values. The effects were consistent: it decreased TEmax, TEfinal and the periodicity of copy number oscillations (Figure 2.5). Thus, this additional process yields notable decrements in TE copy numbers.

DISCUSSION

In this study, we have devised an ODE model to examine the systems dynamics of TE propagation within the context of the epigenetic response of a plant host (Figure 2.1). Although there are clear limitations to our approach, the model has produced at least four fundamental insights. The first is the prediction of oscillating copy numbers typified by a burst of TE activity, followed by silencing, deletion and then reactivity. Despite these oscillations, the system often reached equilibrium copy numbers (Figure 2.2). Second, our model emphasizes the importance of reinforcement by RdDM-like processes, because it buffers potential upstream inefficiencies in the initialization of silencing (Figure 2.3). Third, we show that these outcomes are linked to the rate of TE deletion. Somewhat non-intuitively, the model predicts that either low or very high levels of deletion lead to more

efficient control of the number of aTEs (Figure 2.4). Finally, we show that de-methylation within germline cells reinforces host defenses by dampening TE bursts and lowering steady-state copy numbers (Figure 2.5). Below, we first discuss the caveats of our ODE model before placing our insights into the context of plant genome structure and evolution.

Caveats: Every model has limitations, and ours is no exception. One important consideration is that our biological knowledge of the host response is incomplete. For example, the details of the initiation of methylation are not yet clear, because there are at least two competing (but likely non-exclusive) hypotheses as to how the host transitions from *RNAi* to the RdDM response (Mari-Ordonez et al., 2013, Nuthikattu et al., 2013, McCue et al., 2015). Furthermore, some aspects of the host response have not been included in our model, such as recent discoveries that 18-22nt tRNA fragments (Martinez et al., 2017, Schorn et al., 2017) and some miRNAs (Creasey et al., 2014) may interfere with TE replication and propagation. However, these additional host mechanisms fit relatively easily in our model, because they would likely affect conversion (ε) and initiation (i) (Figure 2.1). In this sense, our model already implicitly accounts for some exciting new findings, but other new insights may require model modifications.

Another limitation is that we have studied the invasion of only one TE family. In reality, plant genomes harbor a multitude of TE types that may interact with each other and also vary with respect to the host response. For example, some but not all TE families in *A. thaliana* are recognized by endogenous miRNAs (Creasey et al., 2014), and short, non-autonomous DNA elements are methylated less efficiently than longer, autonomous elements (Hollister and Gaut, 2009), perhaps in part due to biases in genomic location (Zemach et al., 2013). Finally, we have used only one dataset to fit the model, which

followed the invasion of the *Evade* TE for a short period of few host generations (Mari0Ordonez et al., 2013). The reliance on *Evade* reflects the fact that very few studies have monitored the copy number and expression of TEs within a plant genome over time, particularly beginning from recent invasion or reactivation. We recognize the limitations of the empirical data, but they nonetheless allow a glimpse into model behavior under relevant parameter values.

Invasion and Oscillations: How long does it take to silence TEs *in vivo*? Our understanding of the duration and intensity of TE amplification bursts remains limited (Bousios and Gaut, 2016). In order to be silenced, a TE must first invade. Based on our model and analyses of the stability of the first equilibrium point (where aTEs=sTEs=siRNA=0), invasion depends on expression (v), propagation (p), and deletion (d) but not on downstream properties of the host response, such as conversion of TE transcripts to 21-22nt siRNAs (ε), initiation (i) and reinforcement (r). Put simply, $v*p$ needs to outpace d for a TE to successfully invade the host. We also investigated invasion by modifying v and p from the fitted parameter values (Table 2.1); invasion did not occur when expression or propagation were low ($v < 0.5$, Figure S2.2; $p < 0.25$ Figure S2.3).

Assuming a TE invades successfully, it has the potential to increase rapidly in copy number. Under many parameter values explored in this work, the maximum duration of a TE burst lasts for only a few dozen generations before they are temporarily silenced and decrease in copy number (Figures 2.2, 2.4, S2.2, S2.3, S2.4). These results likely reflect our reliance on data from a study in which silencing occurred rapidly (Mari-Ordonez et al., 2013), but there is other experimental evidence that host defenses react quickly to silence active TEs within a few host generations (Ito et al., 2011; Teixeira et al., 2009). It is

interesting to note that these experimental studies contradict numerous genome-wide analyses, which suggest that TE families experience massive bursts lasting thousands or even millions of years (Piegu et al., 2006, Schnable et al., 2009, Bousios et al., 2012, Daron et al., 2014). One likely explanation for this incongruence may be the difficulty of resolving the occurrence of multiple rounds of episodic bursts within the expanded timeframes reported by the genome-wide studies. Limited resolution may be due to technical issues related to *in silico* TE identification, accurate age estimation, and perhaps even heterogeneous rates of TE sequence loss and decay across the genome (Tian et al., 2009). No matter the cause, the apparent gaps between experimental and genome-wide studies deserves further thought and consideration. Longer-term experimental studies that monitor TE copy numbers over time and under different stress conditions would certainly be welcome contributions to our empirical understanding of host:TE interactions.

Equilibria: Another question is whether TEs reach long-term equilibria within a genome. In our model, the oscillations often reduce in intensity over time to reach a steady state (Figures 2.2, 2.4, S2.2, S2.3, S2.4). In this equilibrium, sTEs are found in higher numbers than aTEs whenever $(v^*p)/d > 2$ (Equations 1 and 2).

Our ODE-based approach regularly predicts two phases of host:TE dynamics: one shaped by frequent changes in TE numbers, and another characterized by an equilibrium. TE evolution has been modeled extensively with population genetic approaches (Charlesworth and Charlesworth, 1983, Charlesworth et al., 1994, Brookfield, 2005, Le Rouzic and Deceliere, 2005), too, and the basic models predict that TEs reach steady-state copy numbers after the first TE invasion through a transposition-selection equilibrium. In other words, they do not predict oscillations prior to an equilibrium. In contrast, some

studies have expanded their models to include TE sequence evolution or competition between TEs, and these often predict oscillations in TE copy numbers (Le Rouzic and Capy, 2006, Le Rouzic et al., 2007a, b). For example, Le Rouzic et al. (2007) investigated host-parasite interactions between autonomous TEs and their non-autonomous counterparts, and they found oscillations in copy numbers between both entities. Notably, the oscillations continued indefinitely; an equilibrium was rarely reached unless there were very low mutation rates and few adaptive TE insertions. Le Rouzic et al. (2007), and also Brookfield (2005), have argued that equilibriums are reached under conditions that are probably unrealistic for *in vivo* TEs. This is because the parameters that affect TE dynamics such as selection, transposition and deletion are likely to change at faster rates than the time required to reach an equilibrium. Our model does not include autonomous and non0autonomous TEs, nor does it allow perturbations in subsequent generations. Yet, the focus on active and silenced TEs may mimic some characteristics of host0parasite relationships and may contribute to our observed oscillating dynamics. We must caution, however, that our model is not explicitly evolutionary, because it does not consider fitness or population variation.

The Importance of Overlapping Mechanisms: Why do plants maintain two overlapping and energetically costly pathways (*RNAi* and RdDM) to silence TEs? Here it encompasses post-transcriptional silencing and the initiation of methylation, and rrepresents RdDM (Figure 2.1). Our results show that only a small amount of *i* is needed to begin silencing of an unrecognized TE, but r is necessary to counter propagation efficiently. For example, the host maintains TE copy numbers at low levels even when iis inefficient (e.g., $i = 0.001$), so long as r reinforces silencing by a value of $r \geq 0.1$ (Figure 2.3B). *RNAi* is

clearly not as efficient at limiting TE copy numbers when there is no RdDM, yet it is essential for silencing TEs (Figure 2.3A). Hence, to the extent the model is correct, it implies that plants must have *RNAi* to start the process of silencing, but RdDM vastly enhances host control over TEs. The inclusion of another, apparently overlapping mechanism – i.e., the active demethylation of TEs in cells that contribute siRNAs to germline cells – further enhances host silencing (Figure 2.5).

Our data are consistent with the argument that 24nt siRNAs are important for buffering TE activity, even though they seem unnecessary because most methylation is maintained independently of RdDM in heterochromatic regions (Zemach et al., 2013). In fact, it was recently shown that these heterochromatic regions also produce 24nt siRNAs, albeit to a smaller extent (Li et al., 2015). These findings are consistent with the idea that 24nt siRNAs may act as immune memory (Fultz et al., 2015), based on evidence that they may play a key role in suppressing reactivated TEs (Teixeira et al., 2009, Ito et al., 2011, Fultz et al., 2015).

The curious case of TE deletion: If 24nt siRNAs act as a source of immune memory, then the retention of silenced TEs may be a benefit to the host, because they may be the template for 24nt siRNA production. This relationship is implied by our analyses of the deletion (*d*) parameter under the *Evade* model (Figure 2.4). If the goal is simply to rid the genome of TEs, the most efficient method is to have a very high *d* (> 0.5) that removes all aTEs and sTEs. However, deletions are mediated by ectopic recombination and illegitimate recombination (Devos et al., 2002) that may introduce a substantial fitness cost due to the potential for catastrophic mutations (Langley et al., 1988). Assuming that high ectopic recombination carries an unacceptable fitness cost, our model suggests that the

next best solution to limit the number of aTEs is to have very low rates of TE deletion ($d \leq 0.01$).

Hence, our argument is that the retention of silenced TE benefits the host by boosting immune memory. In theory, this immune memory provides a defense against the invasion of new TEs that have sequence homology to existing genomic TEs and also against TEs that have escaped silencing and need to be re-silenced. Two interesting features of acquired immune memory are that it is energetically expensive but also maintained under frequent cycles of reinfection (Best and Hoyle, 2013). Under the parameter values explored with our model, the system usually reaches a steady state in which the copy number of aTEs is greater than zero. To the extent that these dynamics reflect reality, a non-zero equilibrium of aTEs defines a system in which reinfection is not merely frequent but constant. This observation may explain one feature of the selective pressure to maintain RdDM-like mechanisms, even though it seems as if most TEs within plant genomes are effectively silenced. There is also a conjecture that 'zombie' TEs are maintained in the genome in order to produce siRNAs that boost immune memory and can trigger the trans-silencing of active relatives (Lisch, 2009). Indirect *in silico* evidence for the existence of zombie TEs has been recently produced in maize (Bousios et al., 2016).

Finally, if low rates of TE deletion are somehow beneficial to the host response, this process could drive genome size increases over evolutionary time, because each new TE infection or TE reactivation adds copies that are silenced, retained and not quickly deleted. This model offers a partial explanation for the high TE contents and genome sizes of plant genomes. We also note that this is unlikely to be a run-away process, because there is

evidence for selection on genome size (Diez et al., 2013; Bilinski et al., 2017), especially when genome size gets too large (Knight et al., 2005).

Future Directions: This is the first study to explicitly incorporate features of the plant host response into a quantitative model of host:TE dynamics.We view this model as a foundation for further extensions that will continue to elucidate important features of host:TE interactions. One promising avenue will be to extend our model to include populations, genetic drift and fitness (Szitenberg et al., 2016), perhaps with a potential for rare beneficial effects (Le Rouzic et al., 2007a). Such an approach is likely to yield more realistic understandings of the evolution of host:TE interactions than are available at present. It will also be illustrative to model multiple TE families, including autonomous and non-autonomous elements, and the possibility for siRNA cross-homologies between them. Finally, an important future goal will be to mimic reality by introducing perturbations into the model. One potential example is polyploidy, which is thought to lead to epigenetic re0patterning (Matzke et al., 1999), but for which the causes remain a mystery.

METHODS

Equilibria and Stability: To find equilibria, we solved for aTE, sTE, and siRNA when all equations were equal to zero. For all analyses of equilibria and stability we assumed u=0 and δ=1 for simplicity, but also because it is biologically reasonable to assume that maintenance of the silenced state is strong (u=0), based on the conservation of symmetric methylation, and that siRNAs degrade rapidly (δ=1). The first equilibrium point was $aTE_{eq} = sTE_{eq} = siRNA_{eq} = 0$. To derive the stability of this equilbrium, we calculated the Jacobian matrix for the ODEs, which provided:

$$\boldsymbol{J_{TE}}(0,0,0) = \begin{bmatrix} v*p-d & 0 & 0 \\ 0 & -d & 0 \\ \varepsilon*v & 0 & 1 \end{bmatrix}$$

The eigenvalues yielded:

$$det(\boldsymbol{J}_{TE}(0,0,0) - \lambda * \boldsymbol{I}) = \lambda^3 + v*p*\lambda^3 + (v*p*d - v*p + d^2)*\lambda - d*v*p - d^2$$

[**Equation 3**]

where λ is the eigenvalue. The equation clearly communicates that stability depends on a complex relationship among v,p, and d but only these parameters. The second equilibrium point is shown in equations 1 and 2 for aTE_{eq} and sTE_{eq}; the corresponding equation for $siRNA_{eq}$ is:

$$siRNA_{eq} = \frac{v}{\frac{r}{\varepsilon*d} - \frac{i}{(\frac{d}{v}-p)}} \qquad \text{[Equation 4]}$$

We also examined the Jacobian matrix and eigenvalues to study stability for this equilibrium point. However, the stability equation was complex and yielded no simple and general trends for relationships between stability and individual parameters.

Fitted parameters: We obtained the data from Mari-Ordonez et al. (2013) by loading their Figure 2.3a onto WebPlotDigitizer (http://arohatgi.info/WebPlotDigitizer/). To estimate model parameters that fit the empirical data, we used the sum of least squares method, based on the following formula:

$$sqEr = \sum (E_{CN} - O_{CN})^2 + w * \sum (E_{Exp} - O_{Exp})^2$$

In this formula, E_{CN} and O_{CN} are the expected and observed copy number respectively. The expected copy number was defined as the sum of aTEs and sTEs obtained from the model. E_{Exp} and O_{Exp} are the expected and observed values, respectively, for relative expression.

The expected relative expression for generation n was obtained from the model by taking the total expression in generation 8, which is equal to v multiplied by the aTE copy number at generation 8, and comparing that to the total expression at generation n, which is equal to *v* * the number of aTEs in generation n. Note, however, that our measure of relative expression may not correspond perfectly to that from Mari-Ordonez et al (2013), because the empirical data on relative expression actually compares two genes (*Evade* and *ACT2*) within each generation and also because qRT-PCR can be inaccurate, especially when it is used as a ratio (of *ACT2* vs. *Evade* expression). In the *sqER* equation, we assigned w a weight of 40 to reflect the magnitude of difference in the empirical data, because copy number reached ~40 and relative expression plateaued at ~1 (Figure 2.2A).

We used a Monte Carlo approach to estimate fitted parameters. In this approach, all seven parameters were initialized with randomly0drawn values from a uniform distribution between 0 and 1, except for v, which was ranged between 0 and 20. We also imposed the constraint that $p + \varepsilon \leq 1.0$. Given initial parameters, the square error (*sqEr*) was calculated as above. A single parameter was then altered, with a step size between 00.1 and 0.1 for all parameters (except v where step size was between 01.0 and 1.0). The *sqEr* was calculated would be calculated and the iteration would only move forward if $sqEr_n > sqEr_{n+1}$; otherwise a new step size would be calculated. All the parameters (in the following order: *v, d, p, i, r,* ε, and δ) were iterated through 100 times with 50 steps for each

parameter, until the final fitted parameters were found with the smallest *sqEr* for each run. The initialization and iteration of all parameters was performed >10,000 times; the lowest sqEr across all 10,000 runs was used to define the fitted parameters.

Running the ODE model: The ODE model was run using odeint from the scipy.integrate package and python (v2.6.6). Figure 2.1A was made with draw.oi, all other figures were made with R (v. 3.3.2). The heatmaps were made with heatmap2, from the gplots library.

FIGURES

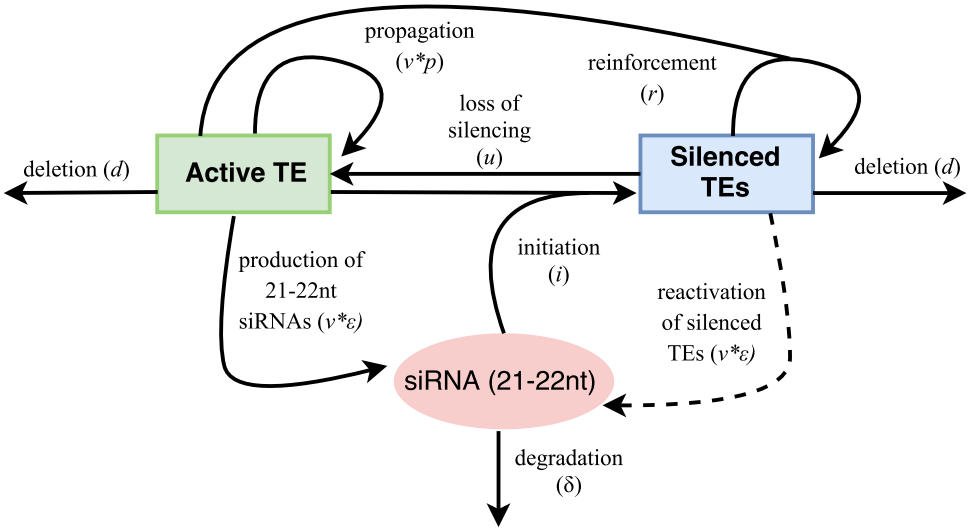

Figure 2.1: A schematic of the model, with details provided in the text. The dashed arrow represents a step specific to cells that contribute to germline material.

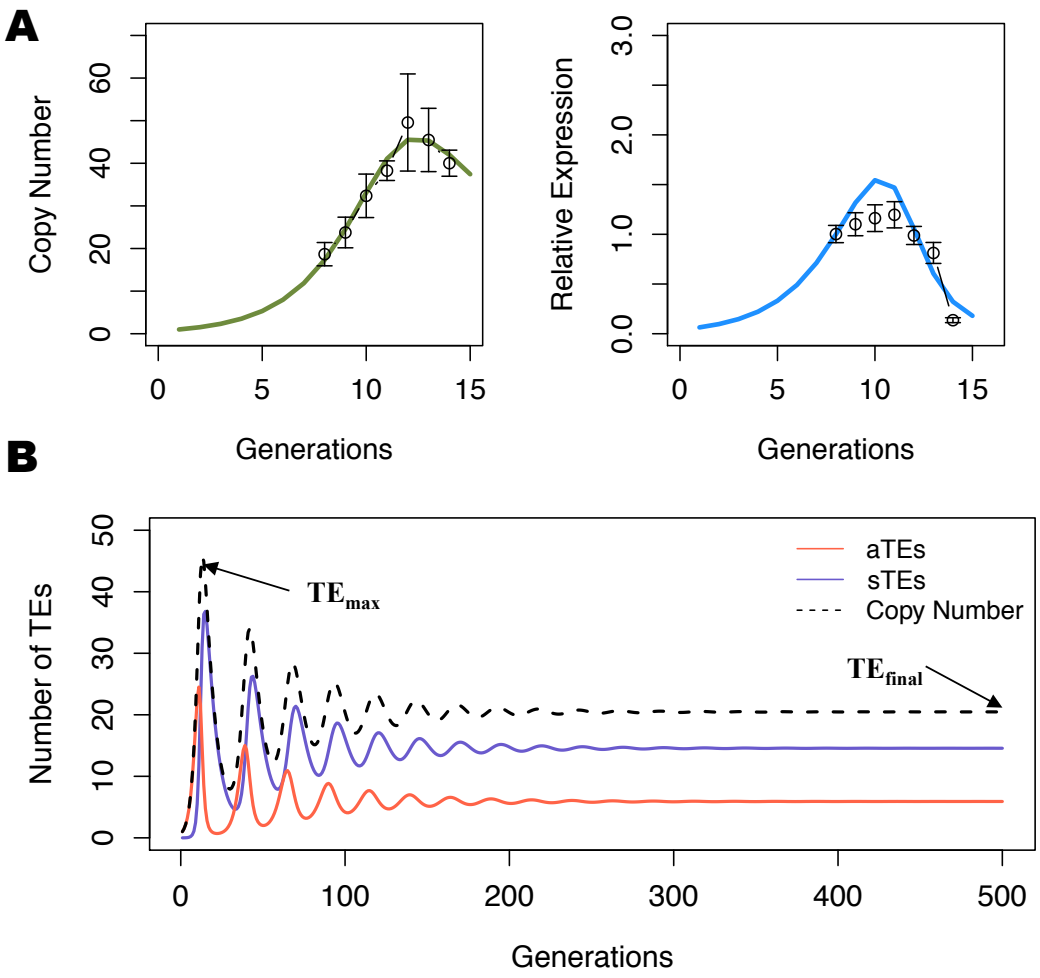

Figure 2.2: A) Model fit to the *Evade* data for total copy number (left) and relative expression (right). The empirical data from the *Evade* study are represented by circles; the whiskers indicate standard deviation. The model results based on the fitted parameters (Table 2.1) are represented by the solid line. B) Long-term behavior of the model, based on the fitted parameters to the *Evade* data. Arrows show TE_{max} and TE_{final}, which are defined in the text. Copy number refers to the summation of aTEs and sTEs.

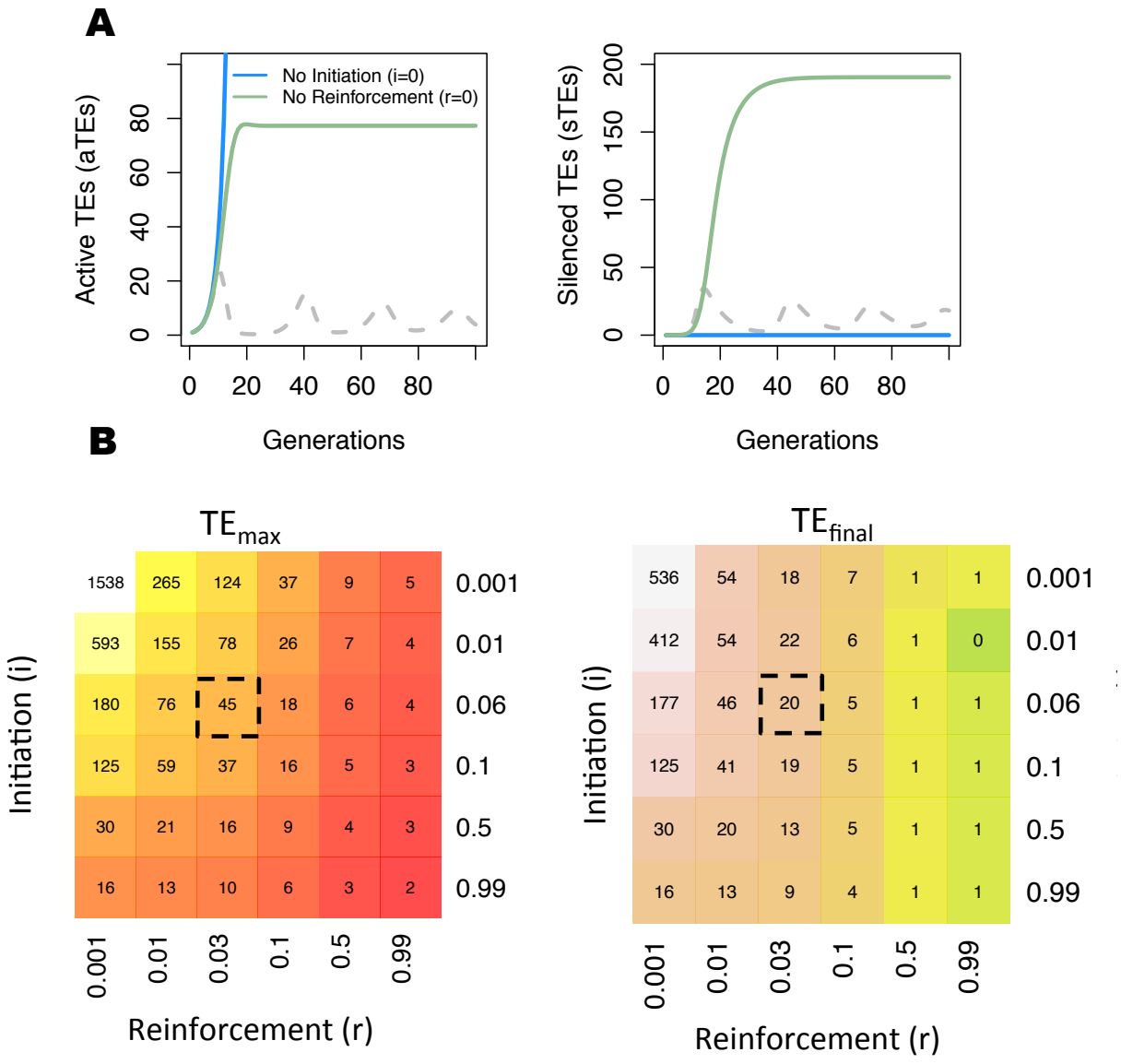

Figure 2.3: Model behavior with the fitted values for all parameters but initiation (*i*) and reinforcement (*r*). A) Graphs illustrate the effect of setting initiation and reinforcement parameters to zero for active TEs (left) and methylated TEs (right). In both graphs, the gray dashed lines represent the number of TEs based on the fitted model parameters to the *Evade* data (see also Figure 2.2B). B) Heat maps showing the TE_{max} (left) and TE_{final} (right) for the total copy number (= aTEs + sTEs) based on varied values of initiation (y-axis) and

reinforcement (x-axis), with copy number displayed in each cell. The dashed cell in each heat map represents the fitted values (Table 2.1).

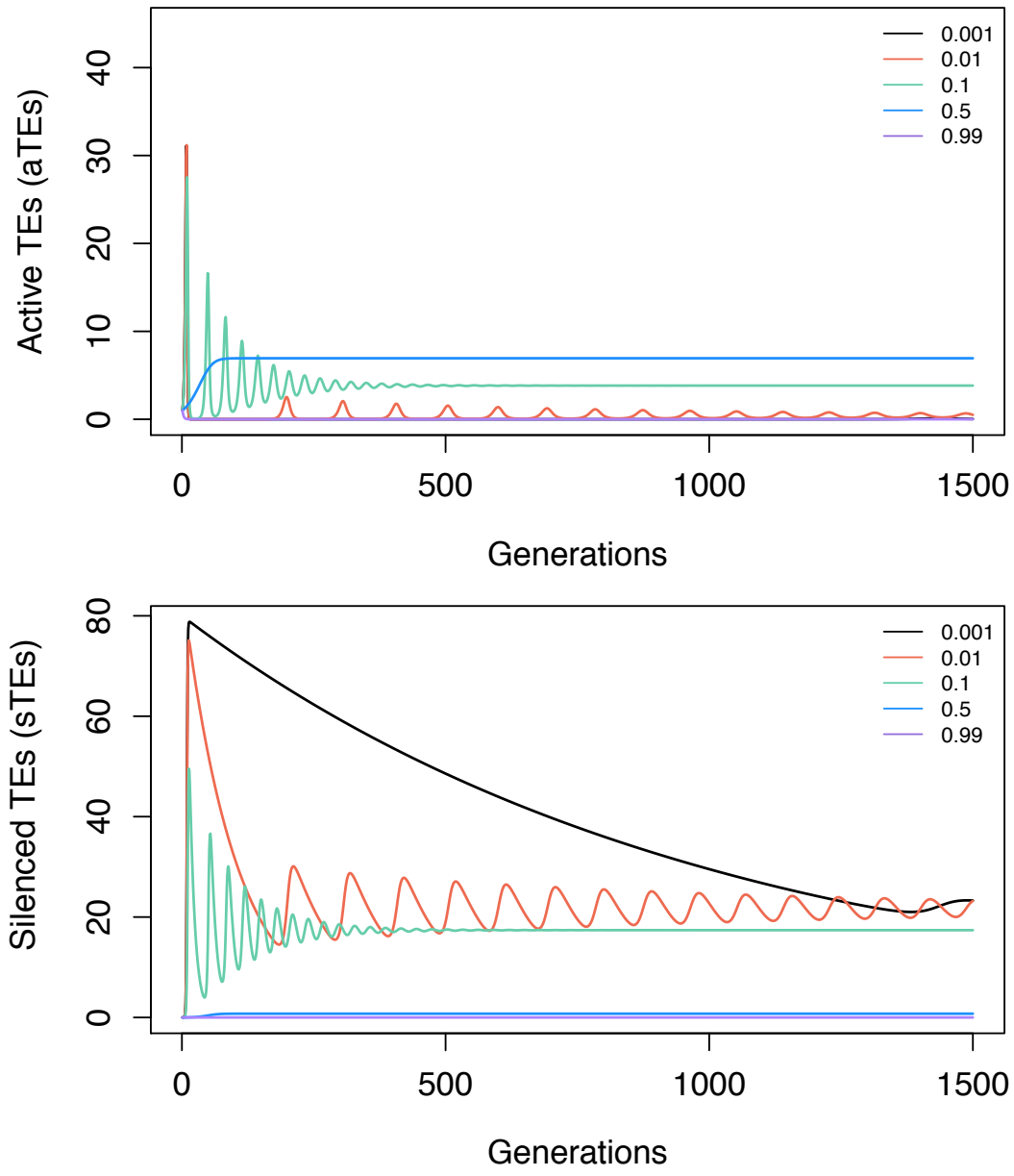

Figure 2.4: Model behavior with the fitted values for all parameters but TE deletion (*d*), which is varied from 0.001 to 0.99.

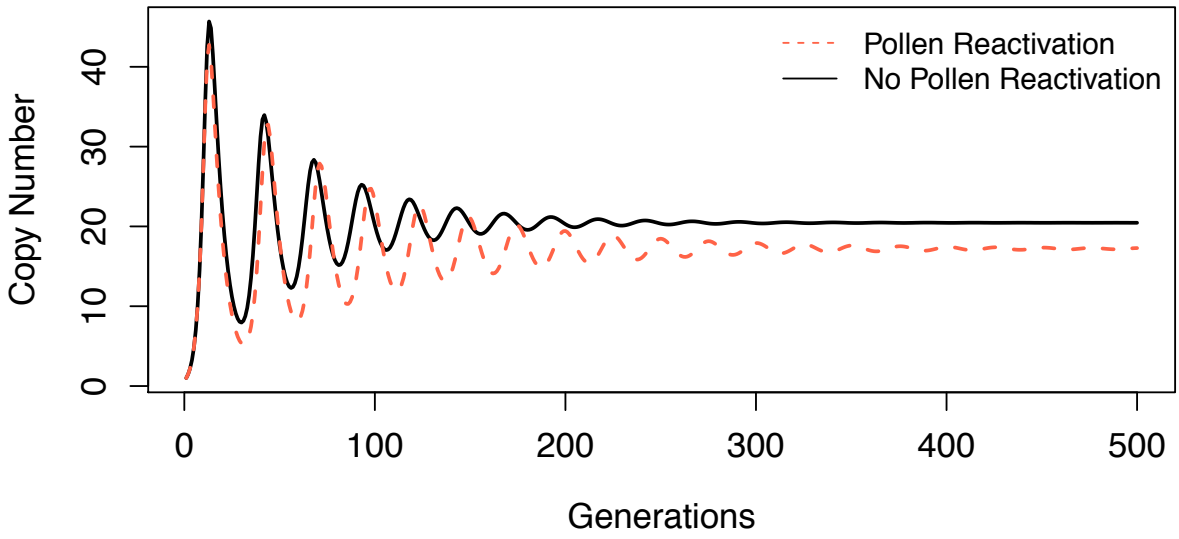

Figure 2.5: TE reactivation in pollen. The black line is based on the model with fitted parameters (no pollen reactivation); the dashed line is using the same parameters but including additional feedback for pollen guard cells (pollen reactivation). Both lines indicate total copy numbers (=aTEs + sTEs). The additional mechanism in pollen guard cells is denoted by the dashed arrow in Figure 2.1.

TABLES

Table 2.1: Summary of parameters and their fitted estimates

Parameter*	**Description**	**Fitted Estimate**
v	Amount of Pol II mRNA expressed by active TEs	1.63
p	Proportion of mRNA that contributes to transposition	0.340
e	Proportion of mRNA that contributes to 21nt siRNA production	0.0510
i	The rate at which 21-22 nt siRNA initiate methylation	0.0619
r	The rate at which 24 nt siRNA reinforce methylation	0.0250
l	The rate of TE removal per generation	0.161
u	The rate of methylation loss per generation	0.000004
δ	The rate of degradation of 21-22nt siRNAs per generation	0.999

SUPPORTING INFORMATION

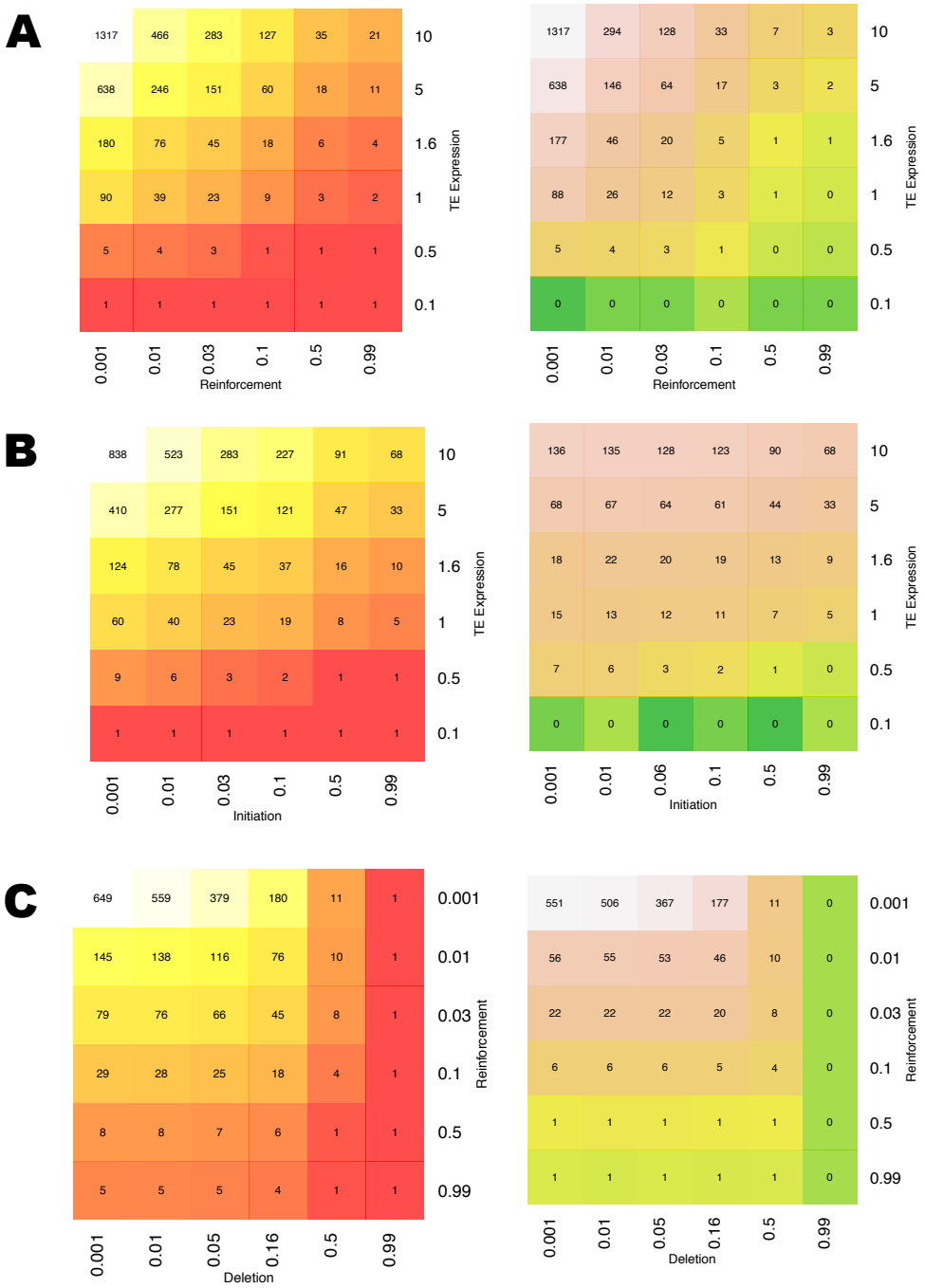

Figure S2.1: Heatmaps of TE_{max} (left) and TE_{final} (right) that vary TE expression (v), initiation (i), reinforcement (r) and deletion (d) while holding all other parameters to fitted values. Actual numbers of TE_{max} and TE_{final} are located in heatmap cells. A. Varies

reinforcement from 0.001 to 0.99 and TE expression from 0.1 to 10. B. Varies initiation from 0.001 and 0.99 and TE expression from 0.1 to 10. C. Varies reinforcement from 0.001 to 0.99 and deletion from 0.001 to 0.99.

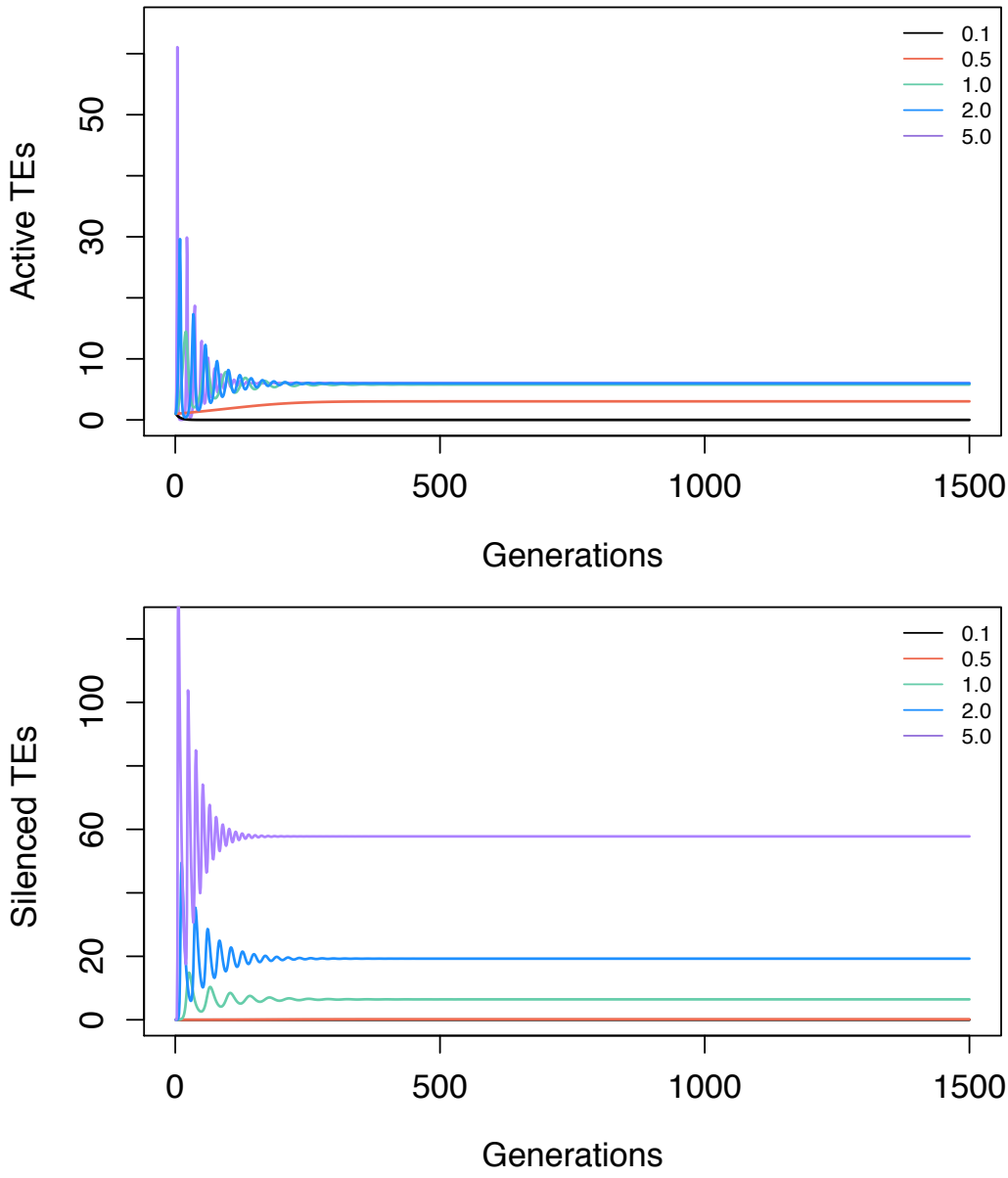

Figure S2.2: Model behavior with the fitted values for all parameters but TE expression (*v*), which is varied from 0.1 to 5.0 in these examples.

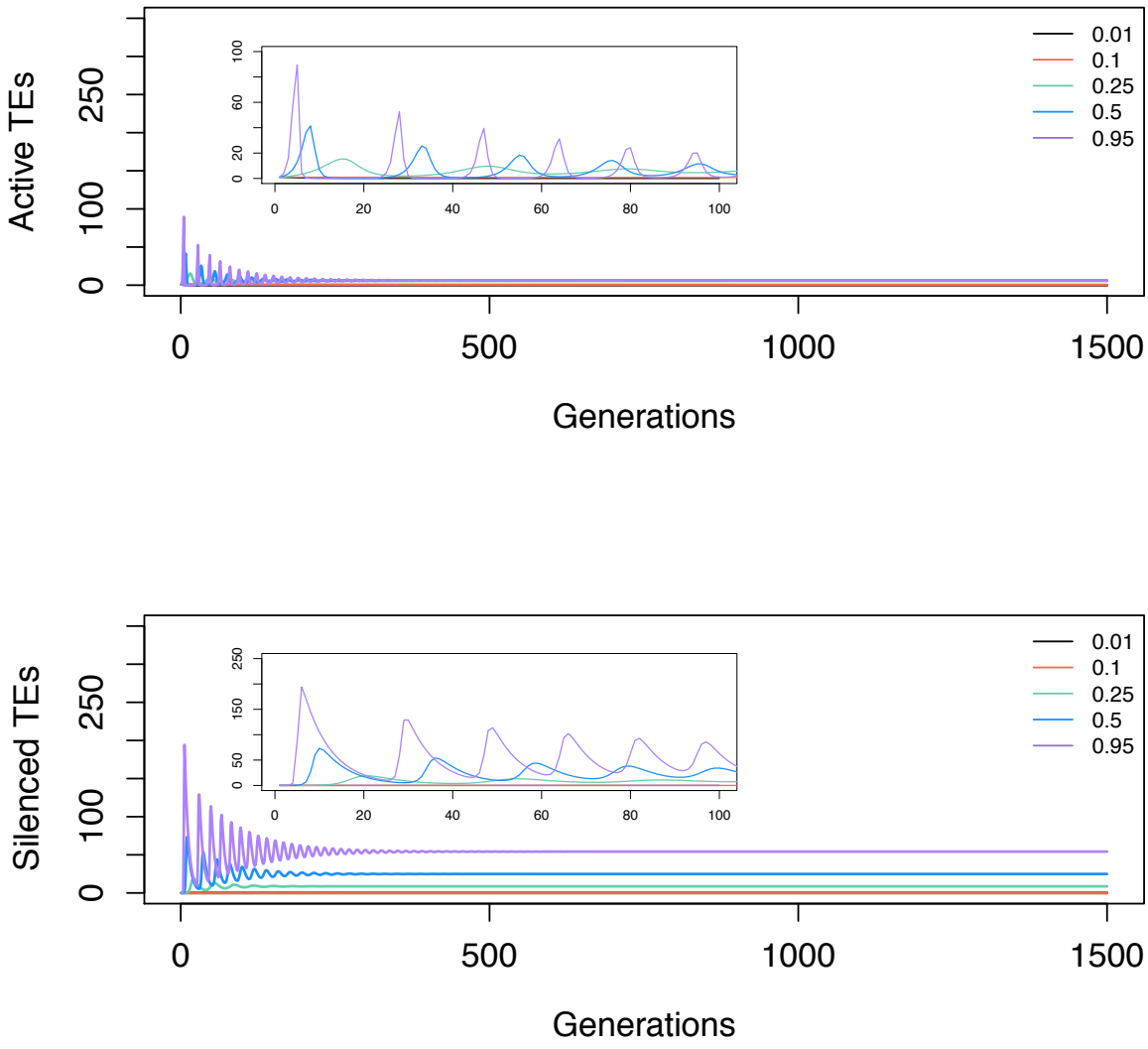

Figure S2.3: Model behavior with the fitted values for all parameters but TE propagation (p), which is varied from 0.001 to 0.99. The insets illustrate the effect over a shorter time-frame of 100 generations.

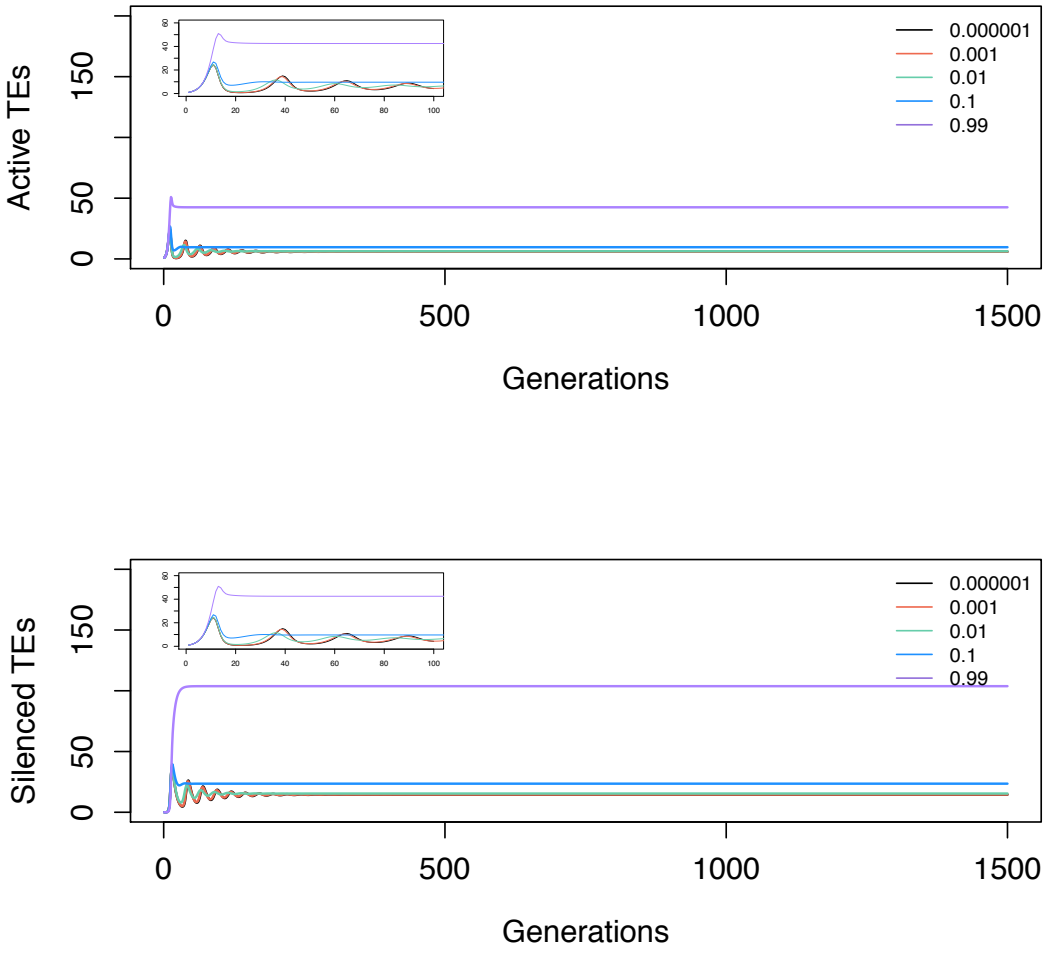

Figure S2.4: Model behavior with the fitted values for all parameters but methylation loss (u), which is varied from 0.000001 to 0.99. The insets illustrate the effect over a shorter time-frame of 100 generations.

Chapter 3: The genomic effects of selfing in maize

Charles Darwin was interested in the prevalence of outcrossing and the many mechanisms that prevent self-fertilization across plant species. He was the first to show that self-fertilization can lead to detrimental fitness effects (Darwin 1876). After his seminal work, scientists have shown that inbreeding lowers fitness across many plant species (reviewed in Charlesworth and Willis 2009). In fact, fitness can become so low in some experimental systems that inbred lines go extinct; that is, they have no reproductive fitness. Interestingly, however, when inbred lines are intercrossed they often produce hybrids that are not only of higher fitness than the parents, they can also exceed the performance of best parent (Zhang et. al 2008). This phenomenon, known as 'heterosis,' is frequently used to improve important food crops such as maize (Zea mays ssp, mays) and Oryza sativa (rice). It is important to understand both inbreeding and heterosis, because they are valuable tools for crop breeding.

In this paper we focus on the genomic effects of inbreeding, including expected increases in homozygosity (Charlesworth and Willis 2009). One explanation for reduced fitness involves the uncovering of recessive detrimental alleles. After inbreeding, more deleterious recessive alleles may become homozygous, hence increasing the genetic load and decreasing fitness. One potential way a genome can combat this increased genetic load is to remove or 'purge' alleles from the population. For these alleles to be purged they need to be of large effect or semilethal when in a homozygous state (Lande and Schemske 1985; Charlesworth et al. 1990; Hedrick 1994; Schultz and Willis 1995). If purging does occur, then the relative fitness of inbred individuals need not decrease (Crnokrak and Barrett

2002). For this reason, purging is a likely a method for a species to combat the harmful effects of inbreeding.

Another predicted, and perhaps related, effect of inbreeding is a reduction in genome size (GS), because selfing species tend to have smaller genomes than outcrossers (Price 1976; Govindaraju and Cullis 1991). A likely reason for GS differences between selfing and outcrossing plant species is disparities in the number and extent of repetitive regions, because repetitive regions constitute the vast majority of most plant genomes, including >85% of the genomes in maize (Tenaillon et. al. 2010) [In contrast, gene content constitutes a smaller, more highly conserved component of plant genomes.] Consistent with this view, GS and transposable element (TE) content is strongly and positively correlated across angiosperm species. Differences in GS and TE content occur not only between species, but within species as well. For example, two maize genomes have been shown to differ by 22% in size, with 90% of that difference due to repetitive elements (Vielle-Calzada et. al. 2009). This observation, along with previous comparisons among maize individuals suggests that GS can change rapidly (Diez et. al. 2013).

If it is the case that repetitive elements are lost in genomes after selfing, what is the cause for this loss of TEs? As TEs are a source of deleterious mutations, we expect after generations of selfing they will be purged from the genome (Wright et. al. 2013). In addition, we predict that TEs near genes will especially be prone to loss, because they have deleterious effects on gene expression (Hollister and Gaut 2009). If this loss of TEs is large enough, it could cause the declines in GS that are observed across selfing species. Finally, if genomic content is being lost, what is the mechanism that removes repetitive elements? One mechanism that has been investigated in the literature is ectopic recombination (along

with unequal recombination), which can remove TE insertions from the genome and lesson the TE load (Kalendar et. al. 2000; Vitte and Bennetzen 2006). It is not clear, however, if these mechanisms act rapidly enough to cause effects over the timespan of generations. An alternative mechanism in the context of selfing is segregation that systematically biases against the retention of heterozygous TE insertions.

In this experiment we ask two questions: what happens to genomes over the first few generations after a transition to selfing, and how might these changes reflect purging? As of yet, no experiments have been designed to explicitly address these questions. In contrast, there have been many studies that contrast outcrossing to selfing plants in phenology, population structure, genomic diversity and evolutionary fate (Wright et al. 2013; Takebayashi and Morrell 2001). Yet these effects probably accrue after not during, the transition to selfing. A smaller number of studies have found evidence for purging by comparing inbreeding depression between naturally inbreeding and naturally outcrossing species (Weller et. al. 2005; Crnokrak and Barrett 2002) but without assessing the genomic effects of inbreeding. Here we take an experimental approach to examine the immediate effects of selfing on genome content and to characterize potential mechanisms of purging.

To explore these questions, we employ an experimental evolution approach. The experiment begins with a single heterozygous parent from each of 11 landraces of maize. The parent is selfed (S1), three S1 individuals are sampled, and the fourth is selfed again (S2) (Figure 3.1). This process proceeds until S6. Here we monitor shifts in GS and genome content between the S1 and the S6 generations, representing the onset of selfing and the end of the experiment (Table 3.1). For individuals grown from S1 and S6 seed, we have

used cell flow cytometry to estimate GS and also applied whole genome sequencing to study changes in genomic content.

Using this experimental setup we address three questions about selfing in plant genomes. First, since other studies have shown that there is a smaller GS for inbreeding species, we investigate whether a change in GS can occur after only six generations of selfing. Second, we use whole-genome sequencing to explore whether specific genomic components have changed through six generations. Our primary prediction is that TEs are lost through the process of selfing. Finally, we speculate the question of potential mechanisms for TE loss in the context of our data.

MATERIAL AND METHODS

Plant materials. We performed our experiment on 11 maize landraces (Table 3.1) that were selfed by John Doebley (U. Wisconsin) and maintained through single-seed descent for six generations. Landraces were chosen as the parental material because they are typically highly heterozygous. As noted above, four sibling seeds were retained from each generation. One seed was grown and selfed, and the remaining three seeds were saved for sampling. All of the sibling seeds were grown in the UC Irvine greenhouses after germination on petri dishes. Leaf and tassel tissue were collected from all samples. Based on this design, we expected to sample materials from 198 plants, but only 73 grew, due to germination failures.

Genome size estimation with cell flow cytometry. Plant material was collected from a single leaf harvest with the same level of maturity for each plant. To estimate GS,

leaf samples from S1 and S6 from each landrace were sent to Plant Cytometry Services (Schijndel. Netherlands). Flow cytometry used 4'6-diamindino-2-phenylindole (DAPI) staining and Ilex crenata 'Fastigiata' (2C = 2.2pg) as an internal standard. Flow cytometry is well suited for detection of small variation of total DNA among samples (Dolezel and Bartos 2005). In order to limit technical error, three technical replicates were performed for each plant, and the reference maize inbred line B73 was included in every batch. To assess whether GS had changed as a consequence of selfing, we performed a t-test between the S1 and S6 generations, including both sibling and technical replicates (Table 3.1).

Whole-genome Sequencing. We selected six landraces and 32 individuals for whole-genome sequencing, focusing on the S1 and S6 generations. DNA was extracted from frozen leaf tissue using the QIAGEN DNeasy Plant Mini kit. DNA was multiplexed into libraries with Illumina TruSeq PCR Free kit. The libraries were sequenced on the HiSeq2500 (100 bp read length, paired-end, 2 lanes) in the UCI High Throughput Genomics Facility in 2015 (landraces MR01, MR08, MR18, and MR19) and on the HiSeq3000 (150 bp read length, paired-end, 1 lane) in the UC Davis DNA Technologies Core in 2016 (landraces MR09 and MR22). Sequenced reads were processed by Trimmomatic (v0.35) to remove barcodes and low quality reads (<20). Reads were also required to have a minimum read length of 36.

Mapping and normalizing counts. Processed reads were mapped onto maize genome AGP version 4 (AGPv4; Jiao et. al. 2017) using BWA-MEM (v0.7.12) with the–k 9 and –T 25 options. To prevent double counts of a feature, only one of the paired reads was used and only the primary alignment was kept for each read (samtools v1.3). Individuals

were sequenced to an average coverage of ~4X, but they ranged from ~2X to 10X (Table S3.1)

Altogether, we counted read counts for four genomic components: genes, chromosomal knobs, rDNA and TEs. The annotation features for protein coding genes and TE gff files for AGPv4 were obtained from the Gramene database on 1/5/17. To identify regions containing knob and rDNA sequences, fasta files of both features were mapped to the genome using blat (v36). The regions mapping to either feature were then added to gff files (blattogff v3) for read count analyses. To count reads, all features were merged (bedtools merge v.2.25.0) to avoid double counting. Bedtools coverage was used to count reads that overlapped at least 90% with each feature.

We used BUSCO genes to normalize between libraries, on the expectation that these highly conserved genes represent an invariant component of the genome. To identify a conserved set of BUSCO genes, we ran BUSCO (v3) on the maize B73 genome. From the resulting set of 1,309 single-copy BUSCO genes, we eliminated any that appeared to be multi-copy or that overlapped with TE annotations in AGPv4. Any gene, knob, or rDNA annotation that overlapped with a TE was also removed.

We identified TEs from the AGPv4 gff file and employed their TE family designations for additional analyses. For some analyses, we investigated TEs that were near genes; these 'genic TEs' were defined as falling within the gene or within 5kb on either side of the gene. Normalized counts were counts divided by the total number of BUSCO genes for that genome. To verify that our use of BUSCO genes was accurate, we simulated datasets with BUSCO normalizations based on Chromosome 10 (see Results below).

Relationship between GS and genomic components. Linear regression and ANOVA analyses were performed to assess relationships between GS and genomic components. Normalized counts for genomic components and the response variable cell flow cytometry for GS data were applied in a linear model (GS ~ genes + TEs + knobs + rDNA) with no transformations to find associations. We also used ANOVA to test for significant differences between generations (g), landraces (l), and interactions (g x l) for each of the four components (Normalized Counts ~ g + l + g:l). This same ANOVA was applied to genic and nongenic TEs and TE families. All statistics were performed in R (v.3.34).

RESULTS

Selfing can lead to rapid declines in GS. We first measured GS for each line, using triplicated flow cytometry measurements on S1 and S6 plants from all 11 lines (Table 3.1). We note that some of the lines did not have three S6 individuals; for example only a single S6 individual germinated for landrace Araguito (MR01) and for other lines we employed S7 individuals (Table 3.1).

Across all eleven landraces, three lines had a significant decrease in GS ($p<0.05$) between S1 and S6, with up to 20% estimated loss in GS (Figure 3.2; Table 3.1). For the remaining 8 lines, we did not detect difference in GS based on flow cytometry. Interestingly, none of the lines exhibited larger GS after selfing, suggesting that the probability of GS loss is significantly higher than that of GS gain. To make this inference, we measured the probability of loss as 3 lines with observed GS loss out of 11 total lines (=0.273). If the probability of GS gain were equivalent, then zero observations of GS increases would be

improbable even with this relatively small (n=11) number of observations (binomial, $p < 0.03$). Thus, in our experiment, selfing is biased toward GS reductions over GS increases.

Counting genomic components and normalizing across libraries. The flow cytometry results suggest that GS can vary both dramatically, with up to 20% loss of the GS, and rapidly, in just six generations. These observations lead to two questions about the observed GS declines: What causes some genomes to decline and not others? And which genomic components are lost during this decline? To answer these questions we performed whole-genome sequencing on six lines: three with decreased GS and three with constant GS. For each line that was sequenced, between one and three individuals were sequenced for S1 and S6 to provide a replicated measure (Table S3.1).

We mapped reads to the reference genome and separated the genome into four different components: genes, TEs, knobs, and rDNA. We investigated these repetitive elements because TEs (Tenaillon et. al. 2011), heterochromatic knobs (McClintock 1978), and rDNA (Havlova et. al. 2016) have been found to account for within-species variation in GS. We counted the total number of reads for each of the components and normalized by counts from 761 BUSCO genes.

To compare counts among individuals, it is important to assess the accuracy of our normalization approach. We tested BUSCO normalization via simulations of TE loss and gain. For the simulations, we used the smallest chromosome 10 for computational efficiency. We randomly removed either 10% or 20% of TEs from the chromosome, duplicated 10% of TEs, or did not change the chromosome. Each treatment was repeated five times with different random TEs removed or gained. The short-read simulator wgsim was used to simulate datasets with ~2x and 10X coverage, mimicking the potential for

different coverages among our libraries. For each simulation, reads were mapped to chromosome 10, counted across annotation features (non-BUSCO genes, TEs, knobs and rDNA) and then normalized by dividing by the total counts for BUSCO genes on chromosome 10. Based on these simulations, we were able to recover the expected decrease in genomic components (Figure S3.1), but it did not recapitulate genome gain in TEs as accurately. It is likely that the inability to estimate TE gains is a feature of our simulations, because we duplicated TEs as exact, tandem copies of chromosomal TEs, which would lead to systematic undercounting of the duplicated TEs. Nonetheless, our simulations indicate that our normalization approach is sufficient to compare TE loss among datasets with different coverages and different degrees of TE loss with the assumption that BUSCO genes are constant between landraces.

Genomic components and GS. Given normalized counts for all individuals and for each of the four genomic components, we investigated the genomic components associated with GS. To do this we used a linear regression model that compared cell flow cytometry (as the measure of GS) to the normalized counts for each of the genomic components (Figure 3.3). After applying the regression to all 32 individuals, we found no discernible relationship between GS and gene counts. However, GS was associated significantly ($p<0.01$) with TEs, with a high correlation coefficient ($r=0.95$, $p<0.001$). For each increase in a normalized TE count there was a predicted increase in 0.004 pg/1C or 400 kbp in GS, when genes, knobs and rDNA were held constant. In the linear model rDNA also associated significantly with GS ($p=0.018$) when holding all other features constant and had a significantly high correlation (0.71, $p<0.001$). To our surprise given previous research (Chia et. al. 2012), we did not find knobs to be significant with GS when holding TEs and

rDNA constant. This may be due, in part, to confounding effects between knobs and TEs and the correlation between them (cor=0.84, p<0.001) (Figure S3.2; see Discussion). When both are placed in the linear model TEs have a stronger linear association with GS than knobs. However, knobs also have a strong correlation with GS (0.79, p<0.001); we therefore suspect that they contribute to GS shifts in our experiment (see below).

Shifts in genomic components over generations. Over all 32 individuals, we have shown that three genomic components correlate with GS, but what about the specific lines that exhibited a GS reduction from S1 to S6 (Figure 3.2)? We plotted the normalized counts for each of the four genomic components, separating the three lines [Araguito (MR01), Costeno (MR08) and Reventador (MR18)] with reductions in GS from those with no observable shift in GS [Cravo Riograndense (MR09), Santo Domingo (MR19) and Tuxpeno (MR22)]. For those with a GS reduction, they exhibited losses of counts within TEs, rDNA and knobs from generation S1 to S6. On average there was a significant ~7% (p=0.017) and ~8% (p=0.012) loss of TEs and knobs, respectively, between S1 and S6 for landraces that decreased GS through selfing. In contrast, the counts in the other landraces remained similar from S1 to S6 for all four genomic components (Figure 3.4), although Santo Domingo shows an intriguing hints of TE increase (2%, not significant).

We applied an ANOVA to the counts of each of the four genomic components to partition the variance across generations (g), landraces (l), and interaction (g x l). There were no significant components of variation for genes (Table 3.2). In contrast, TEs, Knobs and rDNA varied significantly across landraces (p < 10-9). Only TEs and knobs were significant across generations (p ≤ 0.012) and for g x l interactions (p ≤ 0.011). The results

corroborate our previous inference that TEs and knobs are the two components most responsible for genome changes from S1 to S6.

Which TEs are being lost? Since TEs are the largest component of the maize genome and associated with GS, we suspect that they are the primary driving force behind GS loss. We further hypothesize that the loss is attributed to TE insertions that are deleterious. To investigate this possibility, we analyzed TEs that are in and around genes – i.e, that overlap with or are within a 5kb window of genes (genic TEs). We focused on genic TEs because they have been shown to affect the expression of nearby genes and thus represent a component of genetic load (Hollister et. al. 2009;Wright et. al 2003). We found that landraces that had a GS decrease began with ~3.5% ($p=0.002$) higher amounts of genic TEs in S1 than landraces with no change in GS (Figure 3.5). Interestingly, the landrace with the highest total amount of TE counts (Santo Domingo; Figure 3.4) had a relatively low number of genic TE counts in S1. The difference in genic TE content disappeared by S6; that is, for S6 there was no a significant difference ($p=0.680$) in genic TE counts between landraces with and without a GS decrease. Hence, selfing appears to decrease GS from landraces that have high burdens of genic TEs. We note, however, that the number of genic TEs is not sufficient to fully explain GS decreases, because they represent only 11% of the total TE content. Nonetheless, these analyses suggest that the number of genic TEs could be important for determining whether there is GS loss due to selfing.

In addition to looking at regions where TEs have been lost, we also explored whether specific TE types were purged. We put TE families into four broad categories: terminal repeat retrotransposons (LTRs), helitrons, SINEs and LINEs. We predicted that the trend of TE loss is most likely due to loss of long terminal repeat retrotransposons (LTRs),

both because LTRs represent the largest component of the maize genome (SanMiguel et. al. 1996; Schnable et. al. 2009) and because terminal repeats are vulnerable to ectopic recombination (Devos et. al. 2002). Indeed, ~7% of LTRs were lost from S1 to S6 in landraces that saw a GS reduction ($p=0.011$; Figure S3.3). However, loss was not specific to LTRs, because Helitrons and SINEs have a similar pattern of loss in landraces that had a shift in GS, with 4% and 1% lost on average, respectively, from S1 to S6. Interestingly, LINEs, which have relatively low counts, are strongly reduced from S1 to S6 for Santa Domingo (MR19), which did not have an overall significant change in GS, suggesting that there were ongoing shifts in genome content even in those lines that did not exhibit a GS shift as measured by flow cytometry. Finally, we applied an ANOVA to the TE types among all six lines for S1 and S6. The ANOVA indicated that all four TE types differ significantly among landraces ($p < 10\text{-}6$), that three of the four types – LTRs, helitrons and SINEs) – differed between generations ($p < 0.02$), and that two of these (LTRs and helitrons) exhibited g x l interactions ($p < 0.02$) (Table S3.3).

DISCUSSION

In this study we document that some genomes decrease rapidly in size after the onset of selfing. Three of the eleven inbred lines exhibited a GS loss, as measured by flow cytometry, after only six generations of selfing. Remarkably, none of the inbred lines exhibited an increase in GS, suggesting that selfing introduces a biological bias toward smaller genomes. It is also possible that some aspect of the growing conditions drove GS reductions, because it has been shown (for example) that GS can decrease with selection for rapid flowering in maize (Rayburn et. al. 1994). Nonetheless, our results are

concordant with macroanalyses that identify a correlation between GS and selfing across 176 seed plants (Govindaraju and Cullis 1991). We nonetheless find it surprising to see such a drastic reduction in GS after only six generations, because previous studies have examined much longer timescales, such as that between selfers and outcrossing subspecies (Albach and Greilbuber 2004). Previous studies have shown that GS can vary significantly within and between subspecies of Z. mays (Diez et. al. 2013), but we have now shown the rapidity at which this can occur in the context of selfing. Our results are also consistent with the fact that the GS of maize inbred lines is smaller than some open-pollinated maize varieties and the wild relatives of maize (Laurie and Bennett, 1985). Overall our results and previous results suggest that average GSs have decreased through the processes of domestication and subsequent crop improvement, perhaps because of inbreeding and strong selection against lower fit individuals (Rayburn et al., 1994).

Given GS variation, we investigated the genomic components that contribute to size variation – that is, features such as genes, TEs, rDNAs and heterochromatic knobs. One somewhat trivial possibility is that chromosomes with large knobs were somehow disfavored during selfing, such that GS decreases are driven solely by knob content. While we do find a significant association between knobs and GS (Figures 3.3 and 3.4), the association with TEs is stronger and accounts for a larger proportion of GS change, because knobs constitute only 8% of the maize genome (Ananiev et al., 1998) while TEs compose >85% (Schnable et. al. 2009).

Our results complement evidence that maize genomes vary substantially in TE content (Wang and Dooner, 2006; Morgante et. al. 2007). In more recent studies, the measurement of knobs and TEs has been enhanced by the use of high-throughput

sequencing (Tenaillon et al., 2011). For example, one study found genomic differences among that maize inbred lines are driven by differences in knob repeats and that the overall TE content of these lines correlates negatively with GS (Chia et al., 2012). In contrast, we find that both TEs and knobs correlate positively with GS. The differences between our studies and past studies may be from differences in methods. Past studies have mapped to exemplar TEs; presumably TE content can be underestimated, depending on the exemplar set. Our study has used whole genome mapping to specific TEs, which likely results in a higher proportion of counts mapping to TEs, especially with the improved long-read assembled reference genome (Jiao et. al. 2017).

That is not to say that our method is perfect. Even with improvement of the maize genome assembly, repeat regions such as knob repeats are difficult to assemble, annotate and map, and so they may be underrepresented. However, our reliance on AGPv4 should minimize this to the extent possible. It is also possible that we are underestimating the number of knobs because we remove any that overlap with TEs (~47%) causing a bias towards TEs. That said, we still find a small change in the knob counts is associated with changes in the GS. Finally, we have not included centromeric or telomeric repeats in our study, and these could also contribute to shifts in GS. Overall, our four components account for >75% of mapped reads, on average.

In the transition from outcrossing to selfing, we believe that a decrease in GS can be attributed to two different factors. The first is pressure for a smaller genome, which results in the loss of repetitive regions. It has long been suggested that large genomes are disadvantageous (Knight et al., 2005) over short time scales (Smarda et al., 2007) and hence there could be selection against large GS (Smarda et al., 2010). There is also some

evidence of GS limitations in maize (Poggio et al., 1998). If this were a strong force, however, we would expect GS losses to occur in all landraces that transition to sefling, but this is not true. In fact, one of our landraces (Santo Domingo) with a large GS remains constant and may even increase in the number of TEs and knobs (Figure 3.5). Thus, we do not have strong evidence that there are limitations GS per se in the context of selfing.
A second factor could be loss of genic TEs that are deleterious and thus purged from the genome after selfing. We find some intriguing evidence for this, because all three landraces that had a GS decrease contained higher counts of genic TEs in S1 than landraces whose GS remained constant (Figure 3.4). At the cessation of selfing, nearly all landraces had a similar number of TEs in genic regions at S6. We propose that GS reduction is a complex combination of both GS constraints and purging of deleterious TEs, especially near genes.

We know very little about the efficiency of DNA removal across TE families and genomic components or about the possible causes of GS loss. One mechanism that could remove TEs is ectopic or unequal recombination. Ectopic recombination causes double-strand DNA breaks in the terminal repeat sequences of an intact LTR retrotransposons, and solo LTRs result from errors in homologous-recombination-mediated repair (Mani and Chinnaiyan 2010). Since the maize genome is composed mainly of LTRs, we should therefore expect to see solo LTR sequences as a consequence of ectopic recombination (Fehry et. al. 2015). The maize genome contains many solo-LTRs, but we did not find any evidence for increased mapping to these solo-LTRs in S6 (Figure S3.4). That said, we consider this to be a preliminary result, because a more fitting metric will be to contrast the ratio of counts between LTRs and internal TE regions across generations. This approach might be particularly insightful in the context of recombination rates, because unequal

recombination events and TEs are closely linked and where genomic regions have high recombination they tend to have a high density of TEs (Tian et. al. 2009, Kent et. al 2017). These regions may have a more difficult time purging TEs during the transition to selfing. Therefore, the genomes with more TEs in high recombination regions may be subject to higher levels of ectopic recombination and thus loss in genome size.

Another mechanism that could lead to smaller GS is biased segregation against heterozygous TE insertions. Outcrossing species may have many TE insertions as heterozygotes, but it is possible that segregation is biased against large TE insertions, particularly near genes. Over generations of selection this phenomenon might eventually lead to smaller genome sizes and TE decreases in euchromatic regions. This mechanism does not depend on the GS per se, but on the location of TE insertions. That said, it is still somewhat mysterious how selection against individual TE insertions could lead to smaller genome sizes, because six generations is likely not enough to uncouple linkage among TE insertions. Perhaps segregation tends to favor shorter chromosomes, and only parents with substantial differences in TE insertions ultimately exhibit GS decreases. If this is the case, then we predict that the three lines with a GS reduction originated from more heterozygous parental chromosomes and specifically chromosomes that are more heterozygous for TE insertions. Further work is being implemented on SNP heterozygosity of these lines to explore this hypothesis.

We have shown that the transition from outcrossing to selfing can shift GS and genomic content swiftly. As GS is associated with heterchromatic knobs and TEs, we find these are most likely the driving force for a reduction in GS. Where these repeat sequences are located and the total number may predict whether this any loss in GS at all. TEs have been

found to repeatedly make significant contributions to genome evolution. With better sequencing technology we can further uncover the effects of inbreeding and its relationship with GS and TE content. There is still much to learn about the majority of components of higher plant genomes and processes that shape them.

FIGURES

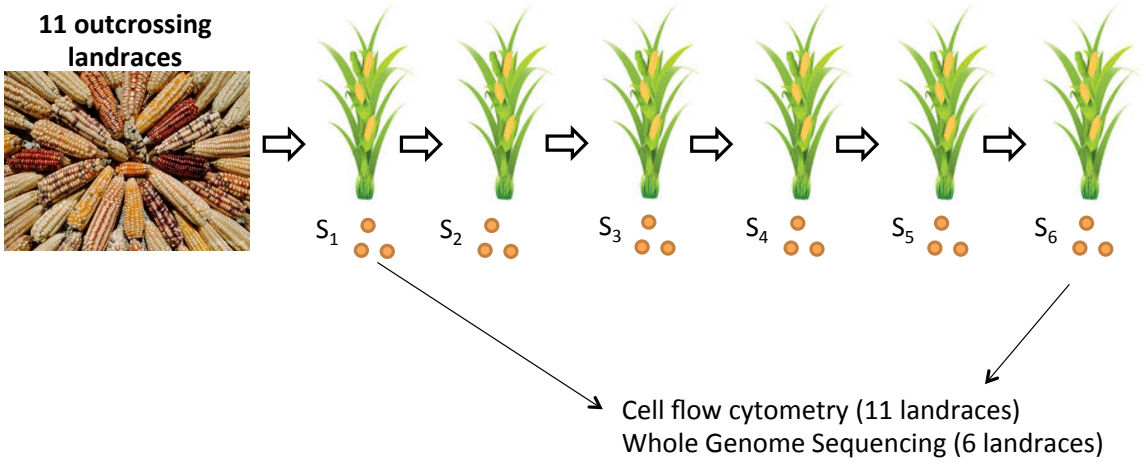

Figure 3.1: Schematic of experimental setup. A single individual from each of 11 outcrossing landraces were selfed with single seed descent. For each generation, three sibling seeds were retained. Generations S1 and S6 were measured with cell flow cytometry and whole genome sequencing.

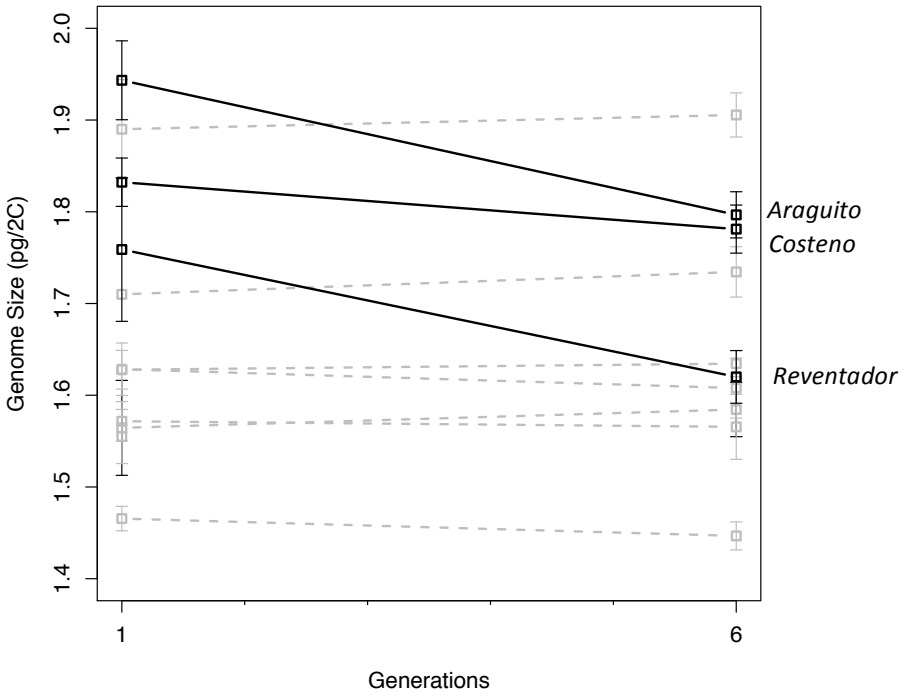

Figure 3.2: Cell flow cytometry measurements for all 11 landraces for generation S1 and S6. Error bars show both sibling seed replicates and technical triplicates. Black solid lines show landraces with a significant decrease in GS and gray dashed lines show landraces with no change in GS.

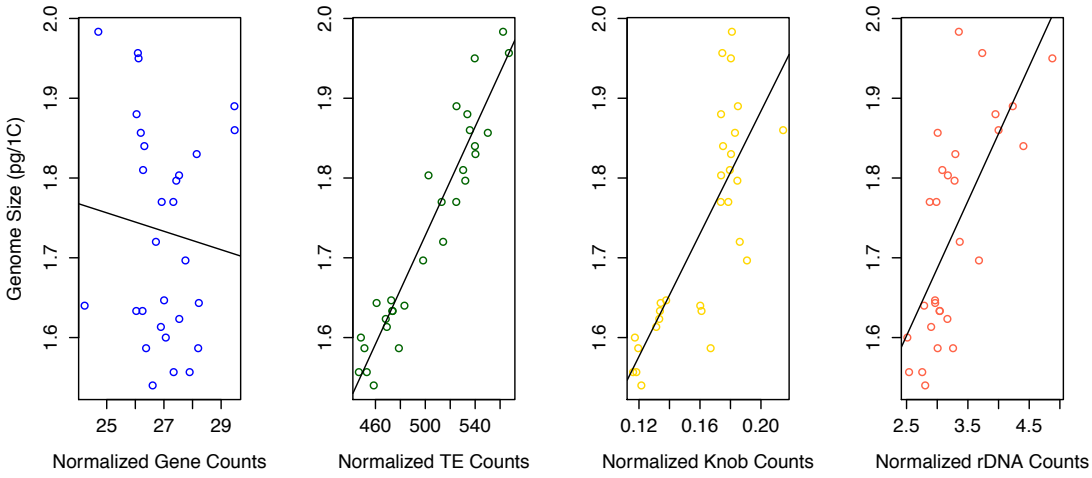

Figure 3.3: Linear relationships between GS (pg/1C) and normalized counts of genes, TEs, knobs, and rDNA. Each dot represents a single individual, and the individuals constitute the entire sample of S1 and S6 individuals across 11 landraces. The GS of each individual is the average among technical replicates. Correlations for all graphs can be found in Figure S3.2.

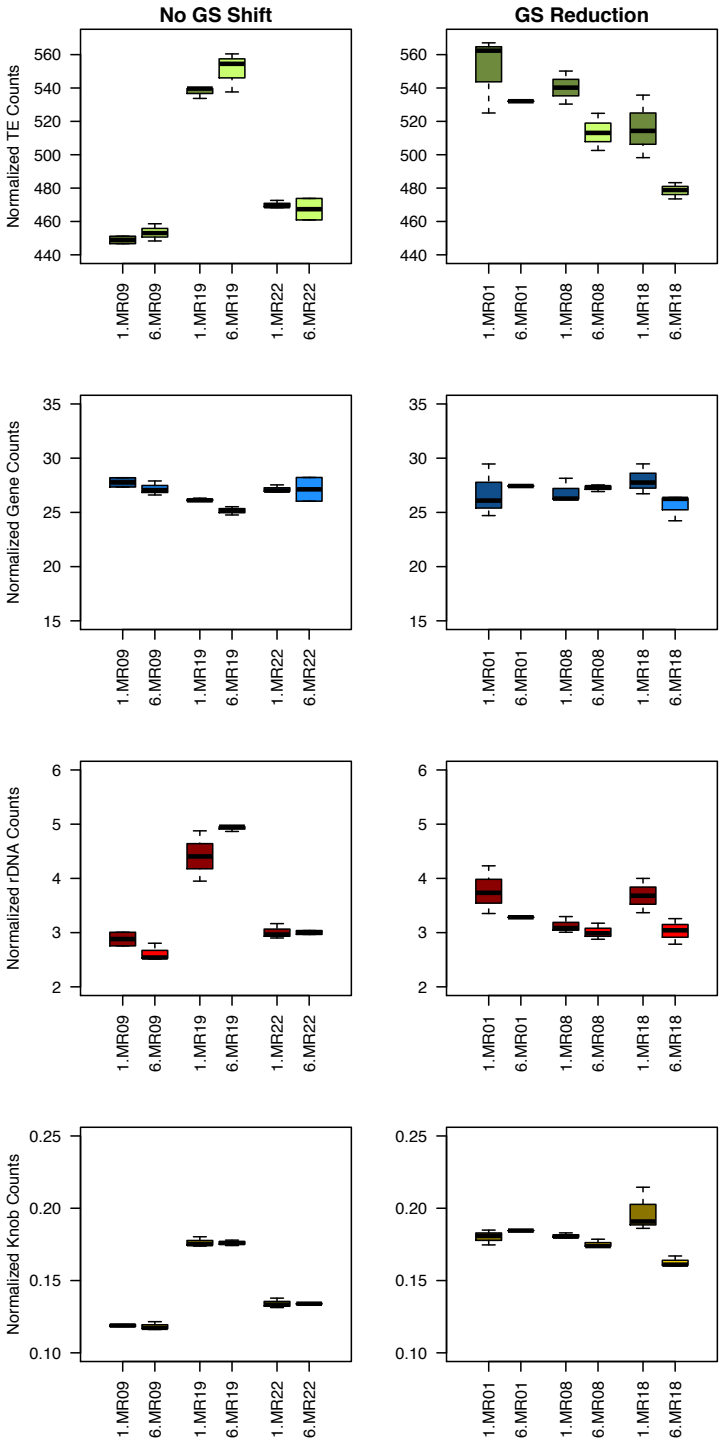

Figure 3.4: Differences in genomic components between landraces and generations. The plots on the left show results for landraces without a significant GS change, as measured by

flow cytometry (Table 3.1). The plots on the right illustrate results based on landraces that had a reduction in GS (Table 3.1). The plots with blue boxplots represent genes, green boxplots represent TEs, yellow boxplots represent knobs and red boxplots represent rDNA. The generation and landrace are indicated in the *x*-axis of each plot – e.g., "1.MRO1" represents the S1 generation of landrace *Araguito*. See Table 3.1 for the names of landraces.

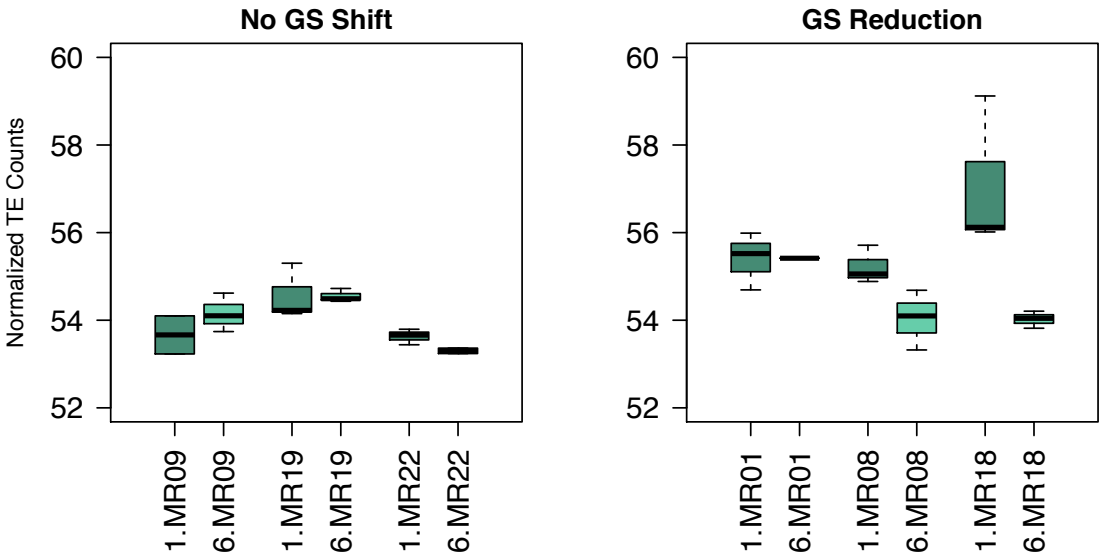

Figure 3.5: Differences in genic TEs between landraces and generations. Left is landraces that saw a reduction in GS, right is landraces that saw no change in GS.

TABLES

Table 3.1: Shows all eleven landraces and the number of sibling replicates available from generations 1 and 6. The t-test examines the null hypothesis of equal GS between S1 and S6.

*We also sampled two individuals from S5 for MR01, and the t-test contrast was also significant for S1 vs. S5 and S6; $p = 1.87 \times 10^{-6}$)

**MR13, MR19 did not have any viable S6 samples, but S7 was used for cell flow cytometry measurements, in addition

**For MR23, S2 was used in place of S1.

Landrace	**Sibling Replicates (S1,S6)**	**Genome Reduction? (t.test p-value)**	**Whole-genome sequencing?**
MR01 (*Araguito*)*	3,1	Yes (2.95×10^{-4})	Yes
MR05	3,3	No (0.333)	No
MR08 (*Costeno*)	3,3	Yes (7.91×10^{-4})	Yes
MR09 (*Cravo Rioganense*)	2,3	No (0.684)	Yes
MR11	2,3	No (0.223)	No
MR13	2,1**	No (0.759)	No
MR14	3,1	No (0.149)	No
MR18 (*Reventador*)	3,3	Yes (5.24×10^{-4})	Yes
MR19(*Santo Domingo*)	3,3**	No (0.447)	Yes
MR22 (*Tuxpeno*)	3,2	No (0.444)	Yes
MR23	3***,3	No (0.638)	No

Table 3.2: Results of the linear model that includes GS (pg/1C) and four genomic components. * represent significance.

	Estimate	**Std. Error**	**t value**	**p-value**
Genes	0.007	0.007	-1.15	0.269
TE	0.004	$3.97x10^{-4}$	9.18	$2.53x10^{-9}$***
Knobs	-0.962	0.515	-1.87	0.074
rDNA	0.042	0.017	2.55	0.018*

Table 3.3: P-values for ANOVA for each of the genomic features whether differences in variation are explained by landrace, generation or an interaction between landrace and generation. * represent signifigance

	Landrace	**Generation**	**Landrace:Generation**
Genes	0.182	0.151	0.323
TEs	$2.75x10^{-11}$***	0.012*	0.011*
Knobs	$2.69x10^{-14}$***	$1.32x10^{-3}$***	$1.56x10^{-4}$***
rDNA	$9.03x10^{-10}$***	0.151	0.020*

SUPPORTING INFORMATION

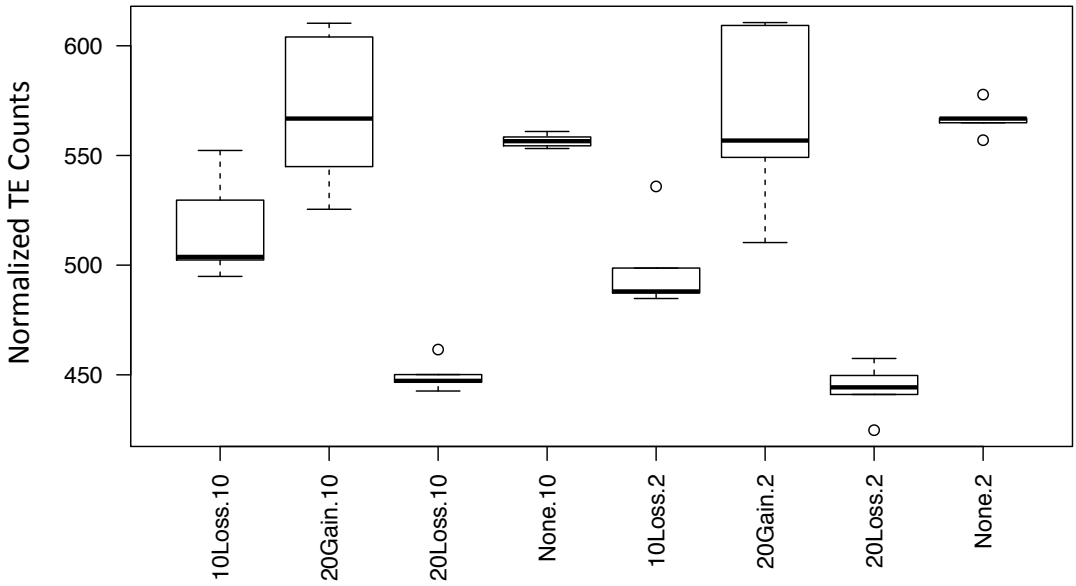

Figure S3.1: Simulation results for BUSCO method on chromosome 10, where either 10% or 20% of TEs were lost (10Loss, 20Loss), 10% of TEs were gained (10Gain) or there was no change to the genome (None). The simulations were run with either 10X or 2X coverage. The simulation parameters are indicated in the x-axis labels; ; for example, "10Loss.2" simulated the loss of 10% of TEs with genome coverage of 2x. The boxplots represent results from 5 simulations. The black bar within the boxplot is the median, and the whiskers represent the 25th and 75th percentile.

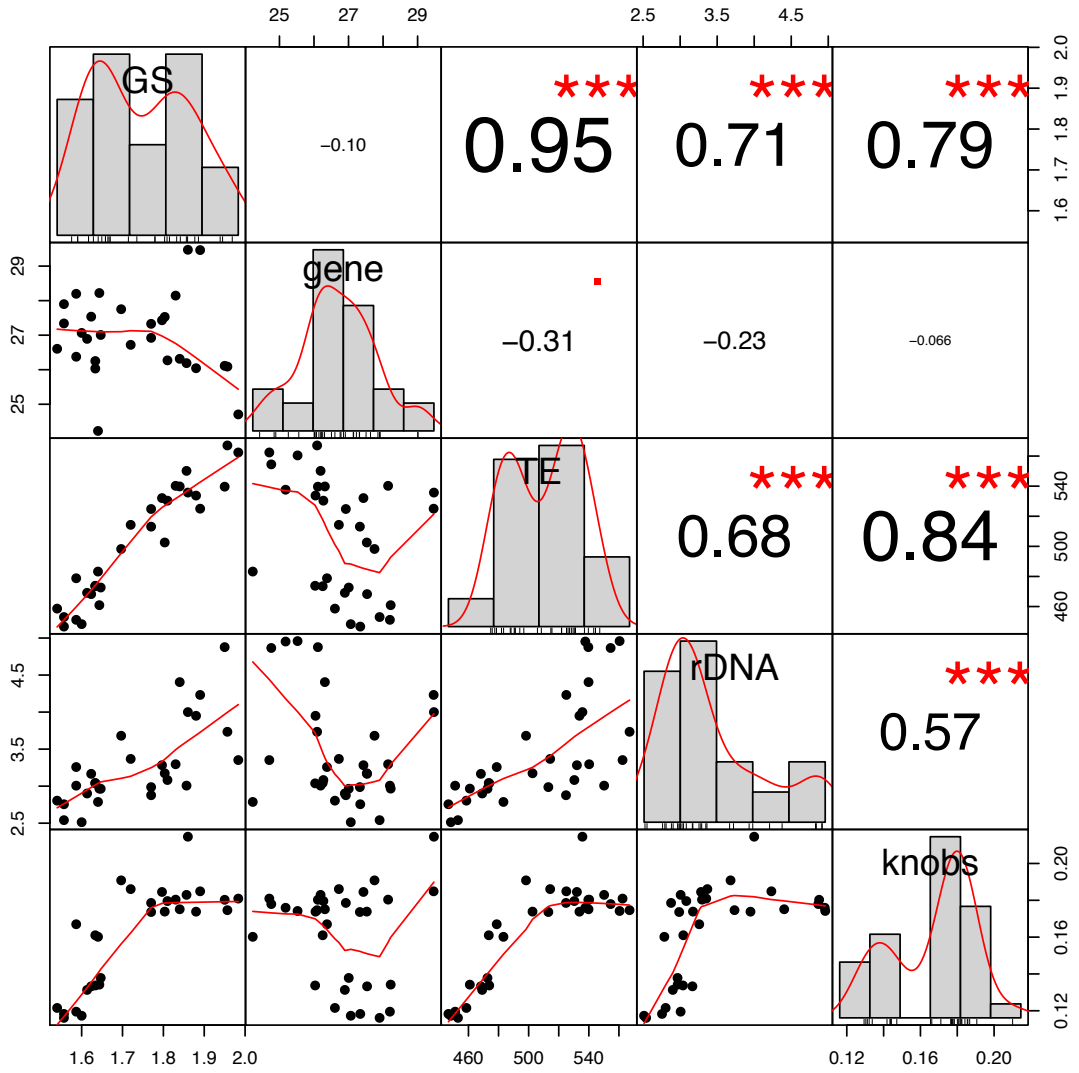

Figure S3.2: Correlations between genomic components. The diagonal line has histograms of GS (pg/1C) and genomic components (normalized counts). Top-right shows correlations between GS and genomic components with stars representing significance. Bottom-left shows plots along with best fit lines.

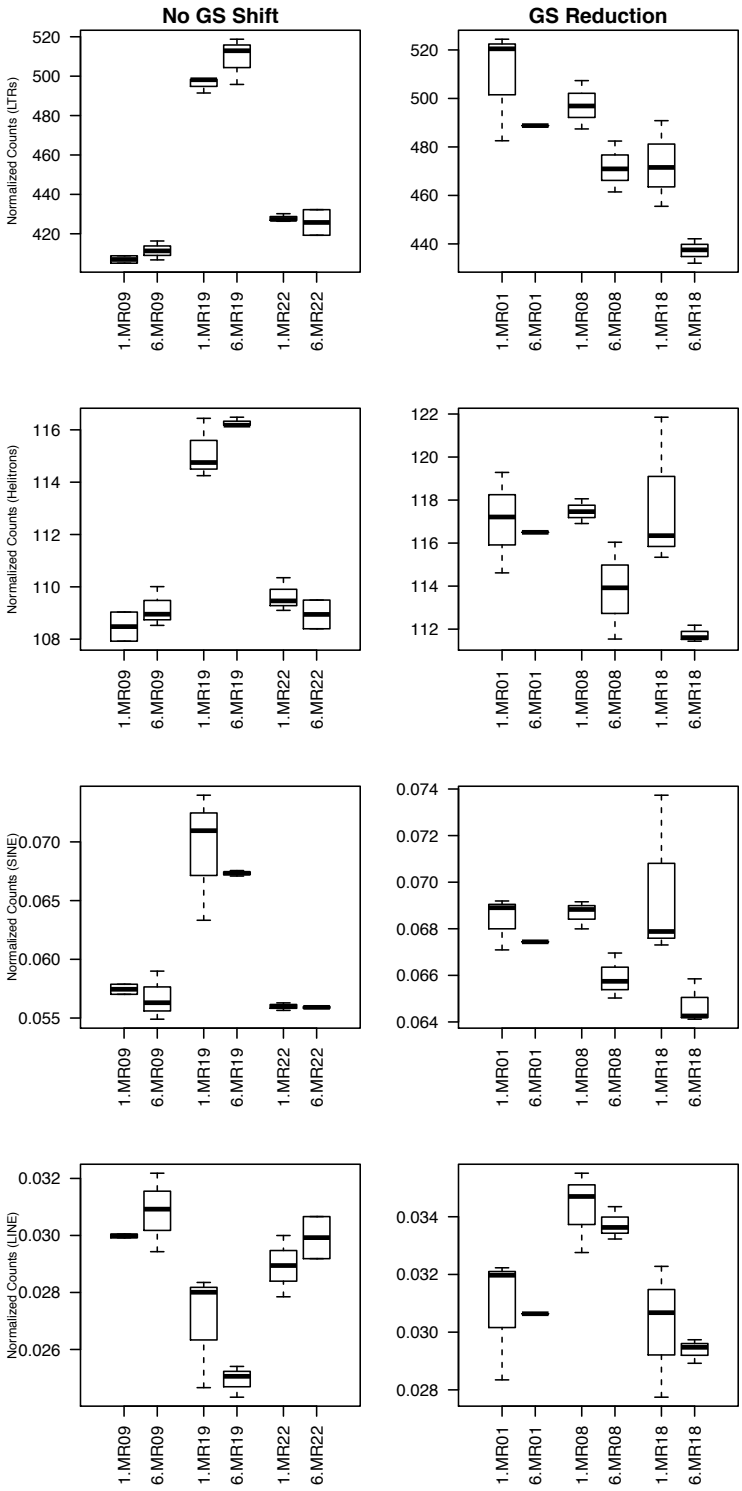

Figure S3.3: Differences in TE families between generations and landraces

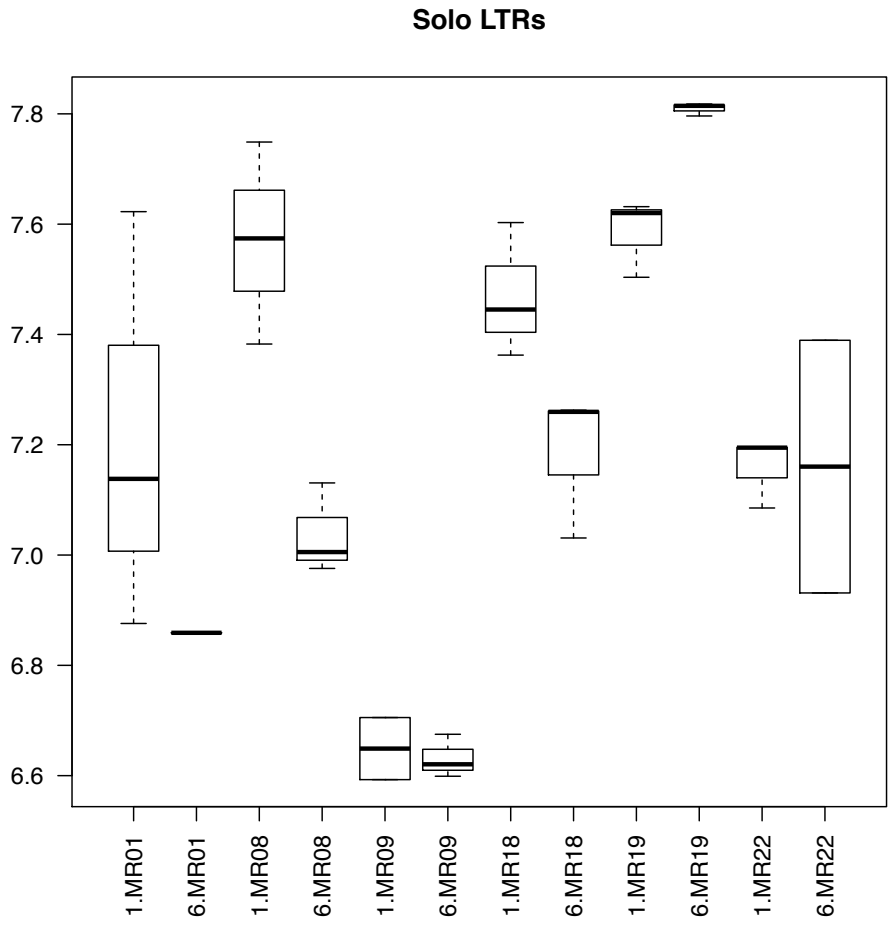

Figure S3.4: Differences in solo-LTRs between landraces and generations.

Table S3.1: Shows number of total reads (read1 and read2) and the number of mapped reads after filtering for quality and trimming

Plant	Landrace	Generation	Total Raw Reads	Mapping & Filtering
6	MR01	1	27,983,116	26,668,715
50	MR08	1	26,986,772	25,573,929
54	MR08	6	24,736,425	19,722,223
129	MR01	6	25,711,281	27,246,849
130	MR08	1	26,012,183	24,360,061
134	MR08	6	26,376,723	24,897,807
158	MR01	1	23,942,518	22,549,068
183	MR08	1	26,003,832	24,820,978
187	MR08	6	26,614,842	25,539,393
24	MR19	1	30,255,041	28,314,178
29	MR19	6	30,396,790	28,288,266
60	MR18	1	22,990,097	21,426,200
65	MR18	6	23,031,606	21,489,571
83	MR18	1	26,956,437	25,230,033
88	MR18	6	21,093,782	19,673,250
94	MR19	1	21,750,767	20,279,482
99	MR19	6	27,445,943	25,288,443
141	MR18	1	26,114,974	23,878,265
146	MR18	6	25,478,194	23,681,434
204	MR19	1	26,701,082	24,735,502
209	MR19	1	25,592,077	23,853,751
135	MR22	1	33,047,402	31,847,965
140	MR22	6	36,347,768	35,025,312

164	MR22	1	35,933,487	34,665,308
169	MR22	6	34,623,472	33,276,187
203	MR22	6	34,157,920	23,978,643
1	MR09	1	35,316,631	33,981,512
5	MR09	6	35,407,139	34,075,274
31	MR22	1	31,346,952	30,133,410
108	MR09	1	33,526,492	32,336,580
112	MR09	6	36,830,659	35,544,758

Table S3.2: ANOVA p-values for genic TEs

	Landrace	**Generation**	**Landrace:Generation**
Genic TEs	0.001**	0.005**	0.006**

Table S3.3: ANOVA p-values for TE families between generations, landraces and the interaction between generations and landrace

	Landrace	**Generation**	**Landrace:Generation**
LTRs	$2.09x10^{-11}$***	0.015*	0.013*
Helitrons	$1.07x10^{-7}$***	0.007**	0.009**
SINEs	$3.72x10^{-9}$***	0.017*	0.557
LINEs	$4.37x10^{-13}$***	0.427	0.524

REFERENCES

Abrusán, G. and Krambeck, H.J., 2006. Competition may determine the diversity of transposable elements. *Theoretical population biology*, *70*(3), pp.364-375.

Albach, D.C. and Greilhuber, J., 2004. Genome size variation and evolution in Veronica. *Annals of Botany*, *94*(6), pp.897-911.

Ananiev, E.V., Phillips, R.L. and Rines, H.W., 1998. Chromosome-specific molecular organization of maize (Zea mays L.) centromeric regions. *Proceedings of the National Academy of Sciences*, *95*(22), pp.13073-13078.

Anders, S. and Huber, W., 2010. Differential expression analysis for sequence count data. *Genome biology*, *11*(10), p.R106.

Becker, C., Hagmann, J., Müller, J., Koenig, D., Stegle, O., Borgwardt, K. and Weigel, D., 2011. Spontaneous epigenetic variation in the Arabidopsis thaliana methylome. *Nature*, *480*(7376), pp.245-249.

Bernatavichute, Y.V., Zhang, X., Cokus, S., Pellegrini, M. and Jacobsen, S.E., 2008. Genome-wide association of histone H3 lysine nine methylation with CHG DNA methylation in Arabidopsis thaliana. *PloS one*, *3*(9), p.e3156.

Best, A. and Hoyle, A., 2013. The evolution of costly acquired immune memory. *Ecology and evolution*, *3*(7), pp.2223-2232.

Biémont, C., 1992. Population genetics of transposable DNA elements. *Genetica*, *86*(1-3), pp.67-84.

Bilinski, P., Albert, P.S., Berg, J.J., Birchler, J., Grote, M., Lorant, A., Quezada, J., Swarts, K., Yang, J. and Ross-Ibarra, J., 2017. Parallel Altitudinal Clines Reveal Adaptive Evolution Of Genome Size In Zea mays. *bioRxiv*, p.134528.

Bird, A., 1997. Does DNA methylation control transposition of selfish elements in the germline?. *Trends in Genetics*, *13*(12), pp.469-470.

Bousios, A. and Gaut, B.S., 2016. Mechanistic and evolutionary questions about epigenetic conflicts between transposable elements and their plant hosts. *Current opinion in plant biology*, *30*, pp.123-133.

Bousios, A., Diez, C.M., Takuno, S., Bystry, V., Darzentas, N. and Gaut, B.S., 2016. A role for palindromic structures in the cis-region of maize Sirevirus LTRs in transposable element evolution and host epigenetic response. *Genome*

research, *26*(2), pp.226-237.

Bousios, A., Kourmpetis, Y.A., Pavlidis, P., Minga, E., Tsaftaris, A. and Darzentas, N., 2012. The turbulent life of Sirevirus retrotransposons and the evolution of the maize genome: more than ten thousand elements tell the story. *The Plant Journal*, *69*(3), pp.475-488.

Bradley, D., Carpenter, R., Sommer, H., Hartley, N. and Coen, E., 1993. Complementary floral homeotic phenotypes result from opposite orientations of a transposon at the plena locus of Antirrhinum. *Cell*, *72*(1), pp.85-95.

Brookfield, J.F., 2005. The ecology of the genome—mobile DNA elements and their hosts. *Nature Reviews Genetics*, *6*(2), pp.128-136.

Candaele, J., Demuynck, K., Mosoti, D., Beemster, G.T., Inzé, D. and Nelissen, H., 2014. Differential methylation during maize leaf growth targets developmentally regulated genes. *Plant physiology*, *164*(3), pp.1350-1364.

Charlesworth, B. and Charlesworth, D., 1983. The population dynamics of transposable elements. *Genetics Research*, *42*(1), pp.1-27.

Charlesworth, B. and Langley, C.H., 1989. The population genetics of Drosophila transposable elements. *Annual review of genetics*, *23*(1), pp.251-287.

Charlesworth, B., Sniegowski, P. and Stephan, W., 1994. The evolutionary dynamics of repetitive DNA in eukaryotes. *Nature*, *371*(6494), pp.215-220.

Charlesworth, D. and Willis, J.H., 2009. The genetics of inbreeding depression. *Nature Reviews Genetics*, *10*(11), pp.783-796.

Charlesworth, D., Morgan, M.T. and Charlesworth, B., 1990. Inbreeding depression, genetic load, and the evolution of outcrossing rates in a multilocus system with no linkage. *Evolution*, *44*(6), pp.1469-1489.

Chia, J.M., Song, C., Bradbury, P.J., Costich, D., De Leon, N., Doebley, J., Elshire, R.J., Gaut, B., Geller, L., Glaubitz, J.C. and Gore, M., 2012. Maize HapMap2 identifies extant variation from a genome in flux. *Nature genetics*, *44*(7), pp.803-807.

Chodavarapu, R.K., Feng, S., Bernatavichute, Y.V., Chen, P.Y., Stroud, H., Yu, Y., Hetzel, J.A., Kuo, F., Kim, J., Cokus, S.J. and Casero, D., 2010. Relationship between nucleosome positioning and DNA methylation. *Nature*, *466*(7304), pp.388 392.

Choi, Y., Gehring, M., Johnson, L., Hannon, M., Harada, J.J., Goldberg, R.B., Jacobsen, S.E. and Fischer, R.L., 2002. DEMETER, a DNA glycosylase domain protein, is

required for endosperm gene imprinting and seed viability in Arabidopsis. *Cell*, *110*(1), pp.33-42.

Cokus, S.J., Feng, S., Zhang, X., Chen, Z., Merriman, B., Haudenschild, C.D., Pradhan, S., Nelson, S.F., Pellegrini, M. and Jacobsen, S.E., 2008. Shotgun bisulphite sequencing of the Arabidopsis genome reveals DNA methylation patterning. *Nature*, *452*(7184), pp.215-219.

Coleman-Derr, D. and Zilberman, D., 2012. Deposition of histone variant H2A. Z within gene bodies regulates responsive genes. *PLoS genetics*, *8*(10), p.e1002988.

Creasey, K.M., Zhai, J., Borges, F., Van Ex, F., Regulski, M., Meyers, B.C. and Martienssen, R.A., 2014. miRNAs trigger widespread epigenetically activated siRNAs from transposons in Arabidopsis. *Nature*, *508*(7496), pp.411-415.

Crnokrak, P. and Barrett, S.C., 2002. Perspective: purging the genetic load: a review of the experimental evidence. *Evolution*, *56*(12), pp.2347-2358.

Cuerda-Gil, D. and Slotkin, R.K., 2016. Non-canonical RNA-directed DNA methylation. *Nature plants*, *2*, p.16163.

Daron, J., Glover, N., Pingault, L., Theil, S., Jamilloux, V., Paux, E., Barbe, V., Mangenot, S., Alberti, A., Wincker, P. and Quesneville, H., 2014. Organization and evolution of transposable elements along the bread wheat chromosome 3B. *Genome biology*, *15*(12), p.546.

Darwin, C., 1876. *The effects of cross and self fertilisation in the vegetable kingdom*. J. Murray.

Devos, K.M., Brown, J.K. and Bennetzen, J.L., 2002. Genome size reduction through illegitimate recombination counteracts genome expansion in Arabidopsis. *Genome research*, *12*(7), pp.1075-1079.

Díez, C.M., Gaut, B.S., Meca, E., Scheinvar, E., Montes-Hernandez, S., Eguiarte, L.E. and Tenaillon, M.I., 2013. Genome size variation in wild and cultivated maize along altitudinal gradients. *New Phytologist*, *199*(1), pp.264-276.

Diez, C.M., Roessler, K. and Gaut, B.S., 2014. Epigenetics and plant genome evolution. *Current opinion in plant biology*, *18*, pp.1-8.

Doležel, J. and Bartoš, J.A.N., 2005. Plant DNA flow cytometry and estimation of nuclear genome size. *Annals of Botany*, *95*(1), pp.99-110.

Dowen, R.H., Pelizzola, M., Schmitz, R.J., Lister, R., Dowen, J.M., Nery, J.R., Dixon, J.E. and Ecker, J.R., 2012. Widespread dynamic DNA methylation in response to

biotic stress. *Proceedings of the National Academy of Sciences*, *109*(32), pp.E2183-E2191.

Fedoroff, N.V., 1989. About maize transposable elements and development. *Cell*, *56*(2), pp.181-191.

Feng, S., Cokus, S.J., Zhang, X., Chen, P.Y., Bostick, M., Goll, M.G., Hetzel, J., Jain, J., Strauss, S.H., Halpern, M.E. and Ukomadu, C., 2010. Conservation and divergence of methylation patterning in plants and animals. *Proceedings of the National Academy of Sciences*, *107*(19), pp.8689-8694.

Flavell, A.J., Pearce, S.R. and Kumar, A., 1994. Plant transposable elements and the genome. *Current opinion in genetics & development*, *4*(6), pp.838-844.

Frahry, M.B., Sun, C., Chong, R.A. and Mueller, R.L., 2015. Low levels of LTR retrotransposon deletion by ectopic recombination in the gigantic genomes of salamanders. *Journal of molecular evolution*, *80*(2), pp.120-129.

Franklin, G.C., Adam, G.I.R. and Ohlsson, R., 1996. Genomic imprinting and mammalian development. *Placenta*, *17*(1), pp.3-14.

Fultz, D., Choudury, S.G. and Slotkin, R.K., 2015. Silencing of active transposable elements in plants. *Current opinion in plant biology*, *27*, pp.67-76.

Gehring, M., Bubb, K.L. and Henikoff, S., 2009. Extensive demethylation of repetitive elements during seed development underlies gene imprinting. *Science*, *324*(5933), pp.1447-1451.

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

Govindaraju, D.R. and Cullis, C.A., 1991. Modulation of genome size in plants: the influence of breeding systems and neighbourhood size. *Evolutionary Trends in Plants (United Kingdom)*.

Greaves, I.K., Groszmann, M., Ying, H., Taylor, J.M., Peacock, W.J. and Dennis, E.S., 2012. Trans chromosomal methylation in Arabidopsis hybrids. *Proceedings of the National Academy of Sciences*, *109*(9), pp.3570-3575.

Gregory, T.R. and Hebert, P.D., 1999. The modulation of DNA content: proximate causes and ultimate consequences. *Genome Research*, *9*(4), pp.317-324.

Harris, E.Y., Ponts, N., Levchuk, A., Roch, K.L. and Lonardi, S., 2009. BRAT: bisulfite treated reads analysis tool. *bioinformatics*, *26*(4), pp.572-573.

Havlová, K., Dvořáčková, M., Peiro, R., Abia, D., Mozgová, I., Vansáčová, L., Gutierrez, C. and Fajkus, J., 2016. Variation of 45S rDNA intergenic spacers in Arabidopsis thaliana. *Plant molecular biology*, *92*(4-5), pp.457-471.

Hawkins, R.D., Hon, G.C., Lee, L.K., Ngo, Q., Lister, R., Pelizzola, M., Edsall, L.E., Kuan, S., Luu, Y., Klugman, S. and Antosiewicz-Bourget, J., 2010. Distinct epigenomic landscapes of pluripotent and lineage-committed human cells. *Cell stem cell*, *6*(5), pp.479-491.

Hedrick, P.W., 1994. Purging inbreeding depression and the probability of extinction: full-sib mating. *Heredity*, *73*(4), pp.363-372.

Hollister, J.D. and Gaut, B.S., 2009. Epigenetic silencing of transposable elements: a trade-off between reduced transposition and deleterious effects on neighboring gene expression. *Genome research*, *19*(8), pp.1419-1428.

Hollister, J.D., Smith, L.M., Guo, Y.L., Ott, F., Weigel, D. and Gaut, B.S., 2011. Transposable elements and small RNAs contribute to gene expression divergence between Arabidopsis thaliana and Arabidopsis lyrata. *Proceedings of the National Academy of Sciences*, *108*(6), pp.2322-2327.

Hsieh, T.F., Ibarra, C.A., Silva, P., Zemach, A., Eshed-Williams, L., Fischer, R.L. and Zilberman, D., 2009. Genome-wide demethylation of Arabidopsis endosperm. *Science*, *324*(5933), pp.1451-1454.

Hu, T.T., Pattyn, P., Bakker, E.G., Cao, J., Cheng, J.F., Clark, R.M., Fahlgren, N., Fawcett, J.A., Grimwood, J., Gundlach, H. and Haberer, G., 2011. The Arabidopsis lyrata genome sequence and the basis of rapid genome size change. *Nature genetics*, *43*(5), pp.476-481.

Ibarra, C.A., Feng, X., Schoft, V.K., Hsieh, T.F., Uzawa, R., Rodrigues, J.A., Zemach, A., Chumak, N., Machlicova, A., Nishimura, T. and Rojas, D., 2012. Active DNA demethylation in plant companion cells reinforces transposon methylation in gametes. *Science*, *337*(6100), pp.1360-1364.

Ito, H., Gaubert, H., Bucher, E., Mirouze, M., Vaillant, I. and Paszkowski, J., 2011. An siRNA pathway prevents transgenerational retrotransposition in plants subjected to stress. *Nature*, *472*(7341), pp.115-119.

Ji, L., Neumann, D.A. and Schmitz, R.J., 2015. Crop epigenomics: identifying, unlocking, and harnessing cryptic variation in crop genomes. *Molecular plant*, *8*(6), pp.860-870.

Jiao, Y., Peluso, P., Shi, J., Liang, T., Stitzer, M.C., Wang, B., Campbell, M.S., Stein, J.C., Wei, X., Chin, C.S. and Guill, K., 2017. Improved maize reference genome with single-molecule technologies. *Nature*.

Johnson, L.M., Du, J., Hale, C.J., Bischof, S., Feng, S., Chodavarapu, R.K., Zhong, X., Marson, G., Pellegrini, M., Segal, D.J. and Patel, D.J., 2014. SRA-and SET domain-containing proteins link RNA polymerase V occupancy to DNA methylation. *Nature*, *507*(7490), pp.124-128.

Kalendar, R., Tanskanen, J., Immonen, S., Nevo, E. and Schulman, A.H., 2000. Genome evolution of wild barley (Hordeum spontaneum) by BARE-1 retrotransposon dynamics in response to sharp microclimatic divergence. *Proceedings of the National Academy of Sciences*, *97*(12), pp.6603-6607.

Kaul, S., Koo, H.L., Jenkins, J., Rizzo, M., Rooney, T., Tallon, L.J., Feldblyum, T., Nierman, W., Benito, M.I., Lin, X. and Town, C.D., 2000. Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. *nature*, *408*(6814), pp.796-815.

Kellogg, E.A. and Bennetzen, J.L., 2004. The evolution of nuclear genome structure in seed plants. *American Journal of Botany*, *91*(10), pp.1709-1725.

Kent, T.V., Uzunović, J. and Wright, S.I., 2017. Coevolution between transposable elements and recombination. *Phil. Trans. R. Soc. B*, *372*(1736), p.20160458.

Knight, C.A., Molinari, N.A. and Petrov, D.A., 2005. The large genome constraint hypothesis: evolution, ecology and phenotype. *Annals of Botany*, *95*(1), pp.177-190.

Lande, R. and Schemske, D.W., 1985. The evolution of self-fertilization and inbreeding depression in plants. I. Genetic models. *Evolution*, *39*(1), pp.24-40.

Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W. and Funke, R., 2001. Initial sequencing and analysis of the human genome. *Nature*, *409*(6822), pp.860-921.

Langley, C.H., Montgomery, E., Hudson, R., Kaplan, N. and Charlesworth, B., 1988. On the role of unequal exchange in the containment of transposable element copy number. *Genetics Research*, *52*(3), pp.223-235.

Lauria, M., Rupe, M., Guo, M., Kranz, E., Pirona, R., Viotti, A. and Lund, G., 2004. Extensive maternal DNA hypomethylation in the endosperm of Zea mays. *The Plant Cell*, *16*(2), pp.510-522.

Laurie, D.A. and Bennett, M.D., 1985. Nuclear DNA content in the genera Zea and Sorghum. Intergeneric, interspecific and intraspecific variation. *Heredity*, *55*(3), pp.307-313.

Law, J.A. and Jacobsen, S.E., 2010. Establishing, maintaining and modifying DNA methylation patterns in plants and animals. *Nature Reviews Genetics*, *11*(3),

pp.204-220.

Law, J.A., Du, J., Hale, C.J., Feng, S., Krajewski, K., Palanca, A.M.S., Strahl, B.D., Patel, D.J.and Jacobsen, S.E., 2013. Polymerase IV occupancy at RNA-directed DNA methylation sites requires SHH1. *Nature*, *498*(7454), pp.385-389.

Le Rouzic, A. and Capy, P., 2006. Population genetics models of competition between transposable element subfamilies. *Genetics*, *174*(2), pp.785-793.

Le Rouzic, A. and Deceliere, G., 2005. Models of the population genetics of transposable elements. *Genetics Research*, *85*(3), pp.171-181.

Le Rouzic, A., Boutin, T.S. and Capy, P., 2007. Long-term evolution of transposable elements. *Proceedings of the National Academy of Sciences*, *104*(49), pp.19375-19380.

Le Rouzic, A., Dupas, S. and Capy, P., 2007. Genome ecosystem and transposable elements species. *Gene*, *390*(1), pp.214-220.

Li, S., Vandivier, L.E., Tu, B., Gao, L., Won, S.Y., Li, S., Zheng, B., Gregory, B.D. and Chen, X., 2015. Detection of Pol IV/RDR2-dependent transcripts at the genomic scale in Arabidopsis reveals features and regulation of siRNA biogenesis. *Genome research*, *25*(2), pp.235-245.

Lippman, Z., Gendrel, A.V., Black, M., Vaughn, M.W., Dedhia, N., McCombie, W.R., Lavine, K., Mittal, V., May, B., Kasschau, K.D. and Carrington, J.C., 2004. Role of transposable elements in heterochromatin and epigenetic control. *nature*, *430*(6998), pp.471-476.

Lisch, D. and Slotkin, R.K., 2011. Strategies for silencing and escape: the ancient struggle between transposable elements and their hosts. *Int Rev Cell Mol Biol*, *292*, pp.119-152.

Lisch, D., 2009. Epigenetic regulation of transposable elements in plants. *Annual review of plant biology*, *60*, pp.43-66.

Lister, C., Jackson, D. and Martin, C., 1993. Transposon-induced inversion in Antirrhinum modifies nivea gene expression to give a novel flower color pattern under the control of cycloidearadialis. *The Plant Cell*, *5*(11), pp.1541 1553.

Lister, R., O'Malley, R.C., Tonti-Filippini, J., Gregory, B.D., Berry, C.C., Millar, A.H. and Ecker, J.R., 2008. Highly integrated single-base resolution maps of the epigenome in Arabidopsis. *Cell*, *133*(3), pp.523-536.

Lister, R., Pelizzola, M., Dowen, R.H., Hawkins, R.D., Hon, G., Tonti-Filippini, J., Nery,

J.R., Lee, L., Ye, Z., Ngo, Q.M. and Edsall, L., 2009. Human DNA methylomes at base resolution show widespread epigenomic differences. *nature*, *462*(7271), pp.315-322.

Lorincz, M.C., Dickerson, D.R., Schmitt, M. and Groudine, M., 2004. Intragenic DNA methylation alters chromatin structure and elongation efficiency in mammalian cells. *Nature structural & molecular biology*, *11*(11), pp.1068 1075.

Luco, R.F., Pan, Q., Tominaga, K., Blencowe, B.J., Pereira-Smith, O.M. and Misteli, T., 2010. Regulation of alternative splicing by histone modifications. *Science*, *327*(5968), pp.996-1000.

Ma, L., Hatlen, A., Kelly, L.J., Becher, H., Wang, W., Kovarik, A., Leitch, I.J. and Leitch, A.R., 2015. Angiosperms are unique among land plant lineages in the occurrence of key genes in the RNA-directed DNA methylation (RdDM) pathway. *Genome biology and evolution*, *7*(9), pp.2648-2662.

Malthus, T.R., 1798. *An Essay on the Principle of Population, as it Affects the Future Imporvement of Society, with Remarks on the Speculations of Mr. Godwin, M. Condorcet, and Other Writers*. The Lawbook Exchange, Ltd..

Mani, R.S. and Chinnaiyan, A.M., 2010. Triggers for genomic rearrangements: insights into genomic, cellular and environmental influences. *Nature Reviews Genetics*, *11*(12), pp.819-829.

Marí-Ordóñez, A., Marchais, A., Etcheverry, M., Martin, A., Colot, V. and Voinnet, O., 2013. Reconstructing de novo silencing of an active plant retrotransposon. *Nature genetics*, *45*(9), pp.1029-1039.

Martinez, G. and Köhler, C., 2017. Role of small RNAs in epigenetic reprogramming during plant sexual reproduction. *Current opinion in plant biology*, *36*, pp.22 28.

Martinez, G., Choudury, S.G. and Slotkin, R.K., 2017. tRNA-derived small RNAs target transposable element transcripts. *Nucleic Acids Research*, *45*(9), pp.5142 5152.

Martínez, G., Panda, K., Köhler, C. and Slotkin, R.K., 2016. Silencing in sperm cells is directed by RNA movement from the surrounding nurse cell. *Nature plants*, *2*, p.16030.

Matzke, M.A., Kanno, T. and Matzke, A.J., 2015. RNA-directed DNA methylation: the evolution of a complex epigenetic pathway in flowering plants. *Annual review of plant biology*, *66*, pp.243-267.

Matzke, M.A., Scheid, O.M. and Matzke, A.J.M., 1999. Rapid structural and epigenetic changes in polyploid and aneuploid genomes. *Bioessays*, *21*(9), pp.761-767.

McClintock, B., 1950. The origin and behavior of mutable loci in maize. *Proceedings of the National Academy of Sciences*, *36*(6), pp.344-355.

McClintock, B., 1978. Mechanisms that rapidly reorganize the genome.

McCue, A.D., Panda, K., Nuthikattu, S., Choudury, S.G., Thomas, E.N. and Slotkin, R.K., 2014. ARGONAUTE 6 bridges transposable element mRNA-derived siRNAs to the establishment of DNA methylation. *The EMBO journal*, p.e201489499.

McDonald, J.F., 1993. Evolution and consequences of transposable elements. *Current opinion in genetics & development*, *3*(6), pp.855-864.

Michael, T.P., 2014. Plant genome size variation: bloating and purging DNA. *Briefings in functional genomics*, *13*(4), pp.308-317.

Morgante, M., De Paoli, E. and Radovic, S., 2007. Transposable elements and the plant pan-genomes. *Current opinion in plant biology*, *10*(2), pp.149-155.

Nussbaumer, T., Martis, M.M., Roessner, S.K., Pfeifer, M., Bader, K.C., Sharma, S., Gundlach, H. and Spannagl, M., 2012. MIPS PlantsDB: a database framework for comparative plant genome research. *Nucleic acids research*, *41*(D1), pp.D1144-D1151.

Nuthikattu, S., McCue, A.D., Panda, K., Fultz, D., DeFraia, C., Thomas, E.N. and Slotkin, R.K., 2013. The initiation of epigenetic silencing of active transposable elements is triggered by RDR6 and 21-22 nucleotide small interfering RNAs. *Plant physiology*, *162*(1), pp.116-131.

Ong-Abdullah, M., Ordway, J.M., Jiang, N., Ooi, S.E., Kok, S.Y., Sarpan, N., Azimi, N., Hashim, A.T., Ishak, Z., Rosli, S.K. and Malike, F.A., 2015. Loss of Karma transposon methylation underlies the mantled somaclonal variant of oil palm. *Nature*, *525*(7570), pp.533-537.

Ortiz, D.F. and Strommer, J.N., 1990. The Mu1 maize transposable element induces tissue-specific aberrant splicing and polyadenylation in two Adh1 mutants. *Molecular and cellular biology*, *10*(5), pp.2090-2095.

Panda, K. and Slotkin, R.K., 2013. Proposed mechanism for the initiation of transposable element silencing by the RDR6-directed DNA methylation pathway. *Plant signaling & behavior*, *8*(8), pp.116-31.

Perelson, A.S., 2002. Modelling viral and immune system dynamics. *Nature Reviews Immunology*, *2*(1), pp.28-36.

Piegu, B., Guyot, R., Picault, N., Roulin, A., Saniyal, A., Kim, H., Collura, K., Brar, D.S., Jackson, S., Wing, R.A. and Panaud, O., 2006. Doubling genome size without polyploidization: dynamics of retrotransposition-driven genomic expansions in Oryza australiensis, a wild relative of rice. *Genome research*, *16*(10), pp.1262-1269.

Poggio, L., Rosato, M., Chiavarino, A.M. and Naranjo, C.A., 1998. Genome size and environmental correlations in maize (Zea mays ssp. mays, Poaceae). *Annals of Botany*, *82*(suppl_1), pp.107-115.

Price, H.J., 1976. Evolution of DNA content in higher plants. *The Botanical Review*, *42*(1), pp.27-52.

Ravindran, S., 2012. Barbara McClintock and the discovery of jumping genes. *Proceedings of the National Academy of Sciences*, *109*(50), pp.20198 20199.

Rayburn, A.L., Dudley, J.W. and Biradar, D.P., 1994. Selection for early flowering results in simultaneous selection for reduced nuclear DNA content in maize. *Plant Breeding*, *112*(4), pp.318-322.

RDevelopment CORE TEAM, 2010. *R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria*. ISBN 3 900051–07–0, URL: http://www. R-project. org.

Regulski, M., Lu, Z., Kendall, J., Donoghue, M.T., Reinders, J., Llaca, V., Deschamps, S., Smith, A., Levy, D., McCombie, W.R. and Tingey, S., 2013. The maize methylome influences mRNA splice sites and reveals widespread paramutation-like switches guided by small RNA. *Genome research*, *23*(10), pp.1651-1662.

Reik, W., 2007. Stability and flexibility of epigenetic gene regulation in mammalian development. *Nature*, *447*(7143), pp.425-432.

Richards, E.J., 1997. DNA methylation and plant development. *Trends in Genetics*, *13*(8), pp.319-323.

Robbins, T.P., Carpenter, R. and Coen, E.S., 1989. A chromosome rearrangement suggests that donor and recipient sites are associated during Tam3 transposition in Antirrhinum majus. *The EMBO journal*, *8*(1), p.5.

SanMiguel, P., Tikhonov, A., Jin, Y.K., Motchoulskaia, N., Zakharov, D., Melake-Berhan, A., Springer, P.S., Edwards, K.J., Lee, M., Avramova, Z. and Bennetzen, J.L., 1996. Nested retrotransposons in the intergenic regions of the maize genome. *Science*, *274*(5288), pp.765-768.

Schmitz, R.J., He, Y., Valdés-López, O., Khan, S.M., Joshi, T., Urich, M.A., Nery, J.R.,

Diers, B., Xu, D., Stacey, G. and Ecker, J.R., 2013. Epigenome-wide inheritance of cytosine methylation variants in a recombinant inbred population. *Genome research*, *23*(10), pp.1663-1674.

Schmitz, R.J., Schultz, M.D., Lewsey, M.G., O'Malley, R.C., Urich, M.A., Libiger, O., Schork, N.J. and Ecker, J.R., 2011. Transgenerational epigenetic instability is a source of novel methylation variants. *Science*, *334*(6054), pp.369-373.

Schmitz, R.J., Schultz, M.D., Urich, M.A., Nery, J.R., Pelizzola, M., Libiger, O., Alix, A., McCosh, R.B., Chen, H., Schork, N.J. and Ecker, J.R., 2013. Patterns of population epigenomic diversity. *Nature*, *495*(7440), pp.193-198.

Schnable, P.S., Ware, D., Fulton, R.S., Stein, J.C., Wei, F., Pasternak, S., Liang, C., Zhang, J., Fulton, L., Graves, T.A. and Minx, P., 2009. The B73 maize genome: complexity, diversity, and dynamics. *science*, *326*(5956), pp.1112-1115.

Schorn, A.J., Gutbrod, M.J., LeBlanc, C. and Martienssen, R., 2017. LTR Retrotransposon control by tRNA-derived small RNAs. *Cell*, *170*(1), pp.61-71.

Schultz, S.T. and Willis, J.H., 1995. Individual variation in inbreeding depression: the roles of inbreeding history and mutation. *Genetics*, *141*(3), pp.1209-1223.

Schwarz-Sommer, Z., Leclercq, L., Göbel, E. and Saedler, H., 1987. Cin4, an insert altering the structure of the A1 gene in Zea mays, exhibits properties of nonviral retrotransposons. *The EMBO journal*, *6*(13), p.3873.

Seymour, D.K., Koenig, D., Hagmann, J., Becker, C. and Weigel, D., 2014. Evolution of DNA methylation patterns in the Brassicaceae is driven by differences in genome organization. *PLoS genetics*, *10*(11), p.e1004785.

Shukla, S., Kavak, E., Gregory, M., Imashimizu, M., Shutinoski, B., Kashlev, M., Oberdoerffer, P., Sandberg, R. and Oberdoerffer, S., 2011. CTCF-promoted RNA polymerase II pausing links DNA methylation to splicing. *Nature*, *479*(7371), pp.74-79.

Slotkin, R.K. and Martienssen, R., 2007. Transposable elements and the epigenetic regulation of the genome. *Nature Reviews Genetics*, *8*(4), pp.272-285.

Slotkin, R.K., Freeling, M. and Lisch, D., 2005. Heritable transposon silencing initiated by a naturally occurring transposon inverted duplication. *Nature genetics*, *37*(6), pp.641-644.

Slotkin, R.K., Vaughn, M., Borges, F., Tanurdžić, M., Becker, J.D., Feijó, J.A. and Martienssen, R.A., 2009. Epigenetic reprogramming and small RNA silencing of transposable elements in pollen. *Cell*, *136*(3), pp.461-472.

Slotkin, R.K., Vaughn, M., Borges, F., Tanurdžić, M., Becker, J.D., Feijó, J.A. and Martienssen, R.A., 2009. Epigenetic reprogramming and small RNA silencing of transposable elements in pollen. *Cell*, *136*(3), pp.461-472.

Smallwood, S.A., Lee, H.J., Angermueller, C., Krueger, F., Saadeh, H., Peat, J., Andrews, S.R., Stegle, O., Reik, W. and Kelsey, G., 2014. Single-cell genome-wide bisulfite sequencing for assessing epigenetic heterogeneity. *Nature methods*, *11*(8), pp.817-820.

Šmarda, P. and Bureš, P., 2010. Understanding intraspecific variation in genome size in plants. *Preslia*, *82*(1), pp.41-61.

Šmarda, P., Bureš, P., Horová, L., Foggi, B. and Rossi, G., 2007. Genome size and GC content evolution of Festuca: ancestral expansion and subsequent reduction. *Annals of botany*, *101*(3), pp.421-433.

Song, Q.X., Lu, X., Li, Q.T., Chen, H., Hu, X.Y., Ma, B., Zhang, W.K., Chen, S.Y. and Zhang, J.S., 2013. Genome-wide analysis of DNA methylation in soybean. *Molecular plant*, *6*(6), pp.1961-1974.

Soppe, W.J., Jacobsen, S.E., Alonso-Blanco, C., Jackson, J.P., Kakutani, T., Koornneef, M. and Peeters, A.J., 2000. The late flowering phenotype of fwa mutants is caused by gain-of-function epigenetic alleles of a homeodomain gene. *Molecular cell*, *6*(4), pp.791-802.

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

Sun, D., Xi, Y., Rodriguez, B., Park, H.J., Tong, P., Meong, M., Goodell, M.A. and Li, W., 2014. MOABS: model based analysis of bisulfite sequencing data. *Genome biology*, *15*(2), p.R38.

Szitenberg, A., Cha, S., Opperman, C.H., Bird, D.M., Blaxter, M.L. and Lunt, D.H., 2016. Genetic drift, not life history or RNAi, determine long-term evolution of transposable elements. *Genome biology and evolution*, *8*(9), pp.2964-2978.

Takebayashi, N. and Morrell, P.L., 2001. Is self-fertilization an evolutionary dead end? Revisiting an old hypothesis with genetic theories and a macroevolutionary approach. *American Journal of Botany*, *88*(7), pp.1143 1150.

Takuno, S. and Gaut, B.S., 2011. Body-methylated genes in Arabidopsis thaliana are functionally important and evolve slowly. *Molecular Biology and Evolution*, *29*(1), pp.219-227.

Takuno, S. and Gaut, B.S., 2013. Gene body methylation is conserved between plant orthologs and is of evolutionary consequence. *Proceedings of the National Academy of Sciences*, *110*(5), pp.1797-1802.

Takuno, S., Ran, J.H. and Gaut, B.S., 2016. Evolutionary patterns of genic DNA methylation vary across land plants. *Nature plants*, *2*, p.15222.

Teixeira, F.K., Heredia, F., Sarazin, A., Roudier, F., Boccara, M., Ciaudo, C., Cruaud, C., Poulain, J., Berdasco, M., Fraga, M.F. and Voinnet, O., 2009. A role for RNAi in the selective correction of DNA methylation defects. *Science*, *323*(5921), pp.1600-1604.

Tenaillon, M.I., Hollister, J.D. and Gaut, B.S., 2010. A triptych of the evolution of plant transposable elements. *Trends in plant science*, *15*(8), pp.471-478.

Tenaillon, M.I., Hufford, M.B., Gaut, B.S. and Ross-Ibarra, J., 2011. Genome size and transposable element content as determined by high-throughput sequencing in maize and Zea luxurians. *Genome biology and evolution*, *3*, pp.219-229.

Tian, Z., Rizzon, C., Du, J., Zhu, L., Bennetzen, J.L., Jackson, S.A., Gaut, B.S. and Ma, J., 2009. Do genetic recombination and gene density shape the pattern of DNA elimination in rice long terminal repeat retrotransposons?. *Genome research*, *19*(12), pp.2221-2230.

Trapnell, C., Roberts, A., Goff, L., Pertea, G., Kim, D., Kelley, D.R., Pimentel, H., Salzberg, S.L., Rinn, J.L. and Pachter, L., 2012. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nature protocols*, *7*(3), pp.562-578.

Vielle-Calzada, J.P., de la Vega, O.M., Hernández-Guzmán, G., Ibarra-Laclette, E., Alvarez-Mejía, C., Vega-Arreguín, J.C., Jiménez-Moraila, B., Fernández-Cortés, A., Corona-Armenta, G., Herrera-Estrella, L. and Herrera-Estrella, A., 2009. The Palomero genome suggests metal effects on domestication. *Science*, *326*(5956), pp.1078-1078.

Vining, K.J., Pomraning, K.R., Wilhelm, L.J., Priest, H.D., Pellegrini, M., Mockler, T.C., Freitag, M. and Strauss, S.H., 2012. Dynamic DNA cytosine methylation in the Populus trichocarpa genome: tissue-level variation and relationship to gene expression. *Bmc Genomics*, *13*(1), p.27.

Vitte, C. and Bennetzen, J.L., 2006. Analysis of retrotransposon structural diversity uncovers properties and propensities in angiosperm genome evolution. *Proceedings of the National Academy of Sciences*, *103*(47), pp.17638-17643.

Vogel, J.P., Garvin, D.F., Mockler, T.C., Schmutz, J., Rokhsar, D., Bevan, M.W., Barry, K.,

Lucas, S., Harmon-Smith, M., Lail, K. and Tice, H., 2010. Genome sequencing and analysis of the model grass Brachypodium distachyon. *Nature*, *463*(7282), pp.763-768.

Volterra, V., 1926. Fluctuations in the abundance of a species considered mathematically. *Nature*, *118*(2972), pp.558-560.

Wang, Q. and Dooner, H.K., 2006. Remarkable variation in maize genome structure inferred from haplotype diversity at the bz locus. *Proceedings of the National Academy of Sciences*, *103*(47), pp.17644-17649.

Weller, S.G., Sakai, A.K., Thai, D.A., Tom, J. and Rankin, A.E., 2005. Inbreeding depression and heterosis in populations of Schiedea viscosa, a highly selfing species. *Journal of evolutionary biology*, *18*(6), pp.1434-1444.

White, S.E., Habera, L.F. and Wessler, S.R., 1994. Retrotransposons in the flanking regions of normal plant genes: a role for copia-like elements in the evolution of gene structure and expression. *Proceedings of the National Academy of Sciences*, *91*(25), pp.11792-11796.

Wicker, T., Zimmermann, W., Perovic, D., Paterson, A.H., Ganal, M., Graner, A. and Stein, N., 2005. A detailed look at 7 million years of genome evolution in a 439 kb contiguous sequence at the barley Hv-eIF4E locus: recombination, rearrangements and repeats. *The Plant Journal*, *41*(2), pp.184-194.

Wright, S.I., Agrawal, N. and Bureau, T.E., 2003. Effects of recombination rate and gene density on transposable element distributions in Arabidopsis thaliana. *Genome Research*, *13*(8), pp.1897-1903.

Wright, S.I., Kalisz, S. and Slotte, T., 2013, June. Evolutionary consequences of self fertilization in plants. In *Proc. R. Soc. B* (Vol. 280, No. 1760, p. 20130133). The Royal Society.

Ye, R., Wang, W., Iki, T., Liu, C., Wu, Y., Ishikawa, M., Zhou, X. and Qi, Y., 2012. Cytoplasmic assembly and selective nuclear import of Arabidopsis Argonaute4/siRNA complexes. *Molecular cell*, *46*(6), pp.859-870.

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

Zemach, A., Kim, M.Y., Silva, P., Rodrigues, J.A., Dotson, B., Brooks, M.D. and Zilberman, D., 2010. Local DNA hypomethylation activates genes in rice endosperm. *Proceedings of the National Academy of Sciences*, *107*(43), pp.18729-18734.

Zemach, A., McDaniel, I.E., Silva, P. and Zilberman, D., 2010. Genome-wide evolutionary analysis of eukaryotic DNA methylation. *Science*, *328*(5980), pp.916-919.

Zhang, H.Y., He, H., Chen, L.B., Li, L., Liang, M.Z., Wang, X.F., Liu, X.G., He, G.M., Chen, R.S., Ma, L.G. and Deng, X.W., 2008. A genome-wide transcription analysis reveals a close correlation of promoter INDEL polymorphism and heterotic gene expression in rice hybrids. *Molecular Plant*, *1*(5), pp.720-731.

Zhang, M., Xu, C., von Wettstein, D. and Liu, B., 2011. Tissue-specific differences in cytosine methylation and their association with differential gene expression in sorghum. *Plant physiology*, *156*(4), pp.1955-1966.

Zhang, X., Yazaki, J., Sundaresan, A., Cokus, S., Chan, S.W.L., Chen, H., Henderson, I.R., Shinn, P., Pellegrini, M., Jacobsen, S.E. and Ecker, J.R., 2006. Genome-wide high-resolution mapping and functional analysis of DNA methylation in Arabidopsis. *Cell*, *126*(6), pp.1189-1201.

Zhong, S., Fei, Z., Chen, Y.R., Zheng, Y., Huang, M., Vrebalov, J., McQuinn, R., Gapper, N., Liu, B., Xiang, J. and Shao, Y., 2013. Single-base resolution methylomes of tomato fruit development reveal epigenome modifications associated with ripening. *Nature biotechnology*, *31*(2), pp.154-159.

Zilberman, D., Gehring, M., Tran, R.K., Ballinger, T. and Henikoff, S., 2007. Genome wide analysis of Arabidopsis thaliana DNA methylation uncovers an interdependence between methylation and transcription. *Nature genetics*, *39*(1), pp.61-69.

Ziller, M.J., Hansen, K.D., Meissner, A. and Aryee, M.J., 2015. Coverage recommendations for methylation analysis by whole-genome bisulfite sequencing. *Nature methods*, *12*(3), pp.230-232.