EXAMINING THE CHROMATIN PROFILE AT TRANSPOSON-GENE BOUNDARIES AND DEVELOPING A GENETIC MAP OF THE B CENTROMERE IN MAIZE

by

NATHANAEL ANDREW ELLIS

(Under the Direction of R. KELLY DAWE)

ABSTRACT

The primary constriction site on a chromosome is called a centromere and is necessary for the faithful segregation of DNA during cell division. In maize, centromeres are primarily made up of the class I transposable elements (TEs), CRM2 and tandem repeat, CentC. These repetitive sequences interact with the centromere defining histone variant, CENH3. In this study, transposon display (TD) was carried out to amplify CRM2 junction sites, creating 40 unique markers that are specific to the B centromere. CRM2-TD markers were genetically mapped to the B centromere by assaying a series of lines with different centromere breakpoints, and the markers were joined to make a ~10kb pseudocontig or B minimal map. Chromatin immunoprecipitation for CENH3 associated DNA was carried out in B73 lines with and without B chromosome. Centromere specific reads were mapped to the B minimal map and B73 genome to identify CRM2-TD marker sequences associated with the active centromere and 31 markers were found that span the centromere cores, necessary for centromere formation. Lack of these markers were associated with ectopic neocentromere formation. The genetic map and minimal map of the B centromere will be essential for further analyzing of centromere deletions lines and formation of a physical map spanning the entirety of the B centromere. TEs play numerous important roles for genome evolution and centromere sequence is an example for class I elements. Non-autonomous derivatives of class II DNA transposons are called Miniature Inverted-Repeat Transposable Elements (MITE) and contribute to genetic diversity in maize. In the grasses, MITEs are abundant in the 3' and 5' regions of genes and associated with high gene expression. In this study, superfamilies of MITEs were analyzed by whether or not they act as boundary elements between genes and a group of class I TEs shown to spread heterochromatin into nearby low-copy regions. We found that when a MITE is present between a gene and spreading TE, gene expression levels are higher and DNA methylation levels lower than when a MITE is absent. Methylation levels drastically reduce over a subset of MITE superfamilies, and these MITEs have a unique chromatin profile and sequence content.

INDEX WORDS: *Zea mays*, centromere, MITE, transposable elements

EXAMINING THE CHROMATIN PROFILE AT TRANSPOSON-GENE BOUNDARIES AND DEVELOPING A GENETIC MAP OF THE B CENTROMERE IN MAIZE

by

NATHANAEL ANDREW ELLIS

BS, University of Missouri, 2008

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

DOCTOR OF PHILOSOPHY

ATHENS, GEORGIA

2015

EXAMINING THE CHROMATIN PROFILE AT TRANSPOSON-GENE BOUNDARIES AND DEVELOPING A GENETIC MAP OF THE B CENTROMERE IN MAIZE

by

NATHANAEL ANDREW ELLIS

Major Professor:	R. Kelly Dawe
Committee:	Jim Leebens-Mack
	Wolfgang Lukowitz
	Wayne Parrott
	Xiaoyu Zhang

Electronic Version Approved:

Suzanne Barbour
Dean of the Graduate School
The University of Georgia
August 2015

DEDICATION

To my loving and supportive family, Sarah and Beckett.

ACKNOWLEDGEMENTS

I would like to thank my advisor, Kelly Dawe, who has shown continuous support and who has been a constant positive influence throughout my graduate career. I would like to thank my committee members Jim Leebens-Mack, Wolfgang Lukowitz, Wayne Parrott, and Xiaoyu Zhang for their collaboration and guidance. Jonathan Gent provided unmeasurable guidance in the lab, with programing and experimental design. The Plant Biology office and greenhouse staff, Michael Boyd, Stephanie Chirello, Gretchen Bowen, Richard Hare, Beth Richardson, and Susan Watkins. Jason Strever and Caroline Jackson who helped tremendously with lab work. I want to thank all my friends and family for their support.

TABLE OF CONTENTS

Page

APPENDICES

LIST OF TABLES

LIST OF FIGURES

Page

CHAPTER 1

INTRODUCTION AND LITERATURE REVIEW

Introduction to chromatin

DNA is rarely found as naked DNA, stripped of all interacting proteins, and is generally wrapped around repeating protein structures known as nucleosomes. Nucleosomes are made up histones which can undergo post-translational changes and contribute to overall chromatin structure. Epigenomics is the study of these chromatin modifications, as well as DNA methylation at the whole genome level, usually using high-throughput sequencing technology. Chromosomes are broadly partitioned into different chromatin forms known as heterochromatin, euchromatin, and centromeric chromatin, which are differentiated based on visible staining intensity and assays of chromatin structure. Although many factors are involved, generally, DNA that is accessible for transcription is considered euchromatin, while inaccessible DNA is considered heterochromatin (Kouzarides 2007). However, recent studies depict chromatin as a spectrum rather than two mutually exclusive chromatin groups (Veiseth *et al.* 2011, Gent *et al.* 2014). Detailed chromatin structure studies of both genes and non-genic regions are essential for understanding the boundaries between heterochromatin and open genic regions.

In maize, gene rich euchromatic regions are found on the arms of chromosomes while heterochromatin is found in areas flanking the centromeres (Shi and Dawe 2006). Heterochromatin is made of transposable elements of all types and is mostly concentrated around centromeres; however, they are also widely dispersed among genes on chromosome arms. The combination of DNA methylation and specific histone modifications are associated with heterochromatin, and in the compact genome of Arabidopsis, are enriched in the pericentromeric region (Zhang *et al.* 2006, Lippman *et al.* 2004). In contrast, genes in maize are found in islands, often containing only one gene, surrounded by vast intergenic regions containing repetitive DNA (Liu *et al.* 2007). The complex genome of maize is more similar to other major crop species than Arabidopsis which has an unusually small and compact genome. The composition of the maize genome allows for a plethora of heterochromatin-euchromatin boundaries and is an excellent model to explore these interactions. A full understanding of chromatin structure genome-wide in maize, particularly at gene boundary regions, will be helpful in understanding how heterochromatin and genes are compartmentalized and maintained.

To understand different chromatin states, it is important to understand the major contributing factors. The overall structure of nucleosomes are fairly consistent, composed of a pair of tetramers each containing two heterodimers H2A - H2B and H3 - H4. The octamer is then wrapped 1.7 times by ~147 base pairs of DNA, and kept in place with the H1 protein by binding linker DNA (Luger *et al.* 1997). Nucleosome structure, composition and location within DNA are related to the modulation of chromatin structure. Histones H3 and H4 have an N-terminal tail emanating from the core of the nucleosome that undergo post-translational modification, and the combination of these modifications influences the structure of chromatin (Jenuwein and Allis 2001, Fuchs *et al.* 2006). Two of the most studied histone N-terminal tail modifications are acetylation and methylation, and each can be found at a number of lysine residues and in different combinations (Kouzarides 2007). Heterochromatin, which is associated with transposons and other repeats, is generally marked by H3K9me1, H3K9me2 and H3K27me, and hypo-acetylated (Pfluger and Wagner 2007). Euchromatin, which is generally associated with genes, often contains nucleosomes marked with acetylation and histone H3 di-methylation on lysine 4 (H3K4me2). Some histone modifications are found primarily in euchromatin but associated with inactive genes (Bernstein, Meissner, and Lander 2007, Li, Carey, and Workman 2007, Kouzarides 2007). In mammals and yeast, H3K9me3 and H3K36me3 are associated with suppressed genes when found in promoter regions, but associated with expression when found in gene bodies (Kouzarides 2007, Li, Carey, and Workman 2007). In plants, H3K27me3 is associated with repression at different developmental stages, specifically at a number of regulatory genes, such as transcription factors (Zhang *et al.* 2007). There are also many other euchromatin and heterochromatin-related modifications that are involved in the silencing of transposons and genes, and in some cases, the impact of histone modification depends on the sequence context and the extent to which it is found with other chromatin markers.

Another epigenetic system indirectly involved in the modification of chromatin is DNA methylation. An ancient silencing mechanism found in both prokaryotes and eukaryotic genomes involves adding a methyl group to the fifth position of the cytosine base, typically at CG-dinucleotide (Klose and Bird 2006). In animals it is predominantly found at CpG sites, but in plant is found in all sequence contexts, CG, CHG and CHH (CG, CHG, and CHH, H is C, T, or A) (Law and Jacobsen 2010). Both CG and CHG are symmetrical after replication and have dedicated methytransferases that maintain their methylation at hemimethylated sites (Law and Jacobsen 2010). DNA replication of CG and CHG sequence results in a symmetrical mark that is identifiable by methyltransferases for subsequent maintenance methylation. The CG methylation pathway in Arabidopsis is maintained by DNA METHYLTRANSFERASE 1 or MET1 which targets hemimethylated DNA (Goll and Bestor 2005). CG methylation is the most abundant form of methylation in the Arabidopsis genome with 56.6% of the CG motifs methylated (Takuno and Gaut 2012). CHG methylation maintenance involves another histone modification, H3K9me2.

The protein CHROMOMETHYLASE 3 (CMT3) carries out CHG methylation through a feedback loop involving the histone methyltransferase KRYPTONITE which is primarily responsible for H3K9 di-methylation (Jackson, J.P., *et al.* 2002). The interaction is described as a cross talk between histone and DNA methyltransferases. CHG methylation in Arabidopsis is the second most abundant DNA methylation type at 35% (Takuno and Gaut 2012). The majority of active genes are devoid of CHG methylation and H3K9me2 (Cokus *et al.*, 2008) and H3K9me2 is mainly found in silenced transposons (Du J., 2012).

In contrast to CG and CHG methylation, there are no proteins that maintain methylation in the CHH context and cytosines in this context must be persistently *de novo* methylated. In plants, 24-nt small interfering RNAs and long non-coding RNAs direct DNA methylation and are involved in gene silencing (Zhang, He, and Zhu 2014). RNA-directed DNA methylation (RdDM) is a specialized methylation pathway that involves siRNAs. The process includes 24-nt siRNA that directs de novo methylation to repetitive elements, transgenes and genes (Pikaard *et al.* 2008). Several essential components have been identified in the model plant Arabidopsis. This pathway involves RNA polymerase IV (Pol IV), which produces single stranded RNAs (ssRNA), and RNA-dependent RNA polymerase 2 (RDR2, also known as Mop1 in maize) which copies ssRNA into double stranded RNA (dsRNA) (Zhang, He, and Zhu 2014). The dsRNA is then diced by DICER-LIKE 3 or DLC3 into 24-nt siRNA, methylated by HUA ENHANCER or HEN1 and loaded into ARGONAUTE 4 or AGO4, to make what is known as the RdDM effector complex (He *et al.* 2009). Two other RNA polymerases, PolV and PolII are involved in producing noncoding-RNA which acts as a scaffold and provides a sequence specific target for siRNA, in cooperation with DOMAIN REARRANGED METHYLTRANSFERASE 2 or DRM2 to methylate in the CG, CHG, and CHH context (Liu *et al.* 2010).

In mammals, DNA methylation is predominantly found at CG sites and essential for development (Jones 2012). Specifically, DNA methylation plays a major role in silencing of sex chromosomes, such as X chromosome inactivation in humans, whole genome imprinting, as well as transposon silencing and regulation of genes (Li and Zhang 2014). Methylated promoter regions have been shown to silence genes, for example the binding of transcription factors at CpG sites are influenced by methylation (Chen *et al.* 2011; Gardiner-Garden and Frommer 1987). While DNA methylation was previously thought to be strictly for silencing of DNA, recent research is unraveling the associations of cytosine methylation and gene activity. Chen *et al.* 2011 found splice sites are enriched with methylation at non-CG sites compared to non-spliced sites (Chen *et al.* 2011). Also, DNA methylation in gene bodies positively correlate with gene expression (Varley *et al.* 2013; Salem *et al.* 2000). However, gene expression and gene body methylation are contextually dependent. For instance, in one study, genes with the highest methylation levels had moderate levels of gene expression (Maunakea *et al.* 2010). DNA methylation also plays a major role in intergenic regions to silence repetitive elements in mammalian cells (Maunakea *et al.* 2010). Overall, DNA methylation is abundant in the mammal genome and plays many important roles, both in genic and intergenic regions. The significance of methylation in CpG islands and in the CG and non-CG context within genic regions is still being unraveled.

In plants, the loss of DNA methylation results in developmental aberrations (Zhang and Jacobsen 2006). Different pathways exist for the maintenance of the CG and CHG, but CHH is persistently de novo methylated. Plants have developed *de novo* DNA methylation mechanisms that methylates at all cytosines, and is maintained by a number of pathways, including RNA-dependent DNA methylation (RdDM) as described in detail above (Law and Jacobsen 2010).

Suppressed repetitive sequences are enriched for CG and CHG methylation; however CHH methylation is also present at a lower frequency genome-wide (Feng *et al.* 2010). In *Arabidopsis*, DNA methylation occurs in promoters at a low frequency and is associated with tissue specific gene expression. Gene body methylation positively correlated with gene expression, and in some cases methylated genes were constitutively expressed (Zhang *et al.* 2006). CG methylation within the gene body is not the defining feature for gene silencing, and in general, methylated genes had moderate gene expression (Zilberman *et al.* 2007). In Arabidopsis about one-third of genes are associated with CG methylation in gene bodies and methylated genes tend to have moderate gene expression (Cokus *et al.* 2008). Pseudogenes and repeats tend to have high enrichment of CG and CHG methylation, while CHG and CHH are depleted in active genes (Cokus *et al.* 2008)

RNA has been shown to play important roles in many biological processes, such as defense against transposable element activation and viruses, gene regulation via microRNA, and gene regulation (Almeida and Allshire 2005). In maize, recent studies have described the association 24-nt RNA, as a part of the RdDM pathway, in promoter regions of genes and is associated with transposons (Gent *et al.* 2013). For the reason that the RdDM pathway is involved with *de novo* methylation in all sequence contexts, it would be presumed to be important for silencing, however CHH methylation, a defining feature of RdDM, is associated with highly expressed genes in maize (Gent *et al.* 2013). These RdDM-loci are also found to be depleted of H3K9me2 while having high levels of H3K27me2, making it more similar to euchromatin than heterochromatin (Gent *et al.* 2014). These data reveal a specialized chromatin structure intermediate to heterochromatin and euchromatin targeted by RdDM and enriched with repetitive elements as well as methylated by CHH more than any other region in the genome thus discovered and therefore called CHH islands.

The transposon landscape of maize

The vast intergenic repetitive regions of maize are mostly made up transposable elements, making up 85% of the genome (Schnable *et al.* 2009). Transposable elements are divided into two main classes called class I and class II. In maize, class I elements are abundant in deep heterochromatin regions near the centromere and intergenic regions between genes (Lamb, Meyer, *et al.* 2007, Schnable *et al.* 2009). Class I elements transpose via an RNA intermediate ("copy and paste"), and make up 75% of the maize genome (Schnable *et al.* 2009). Class II elements make up a smaller portion of the genome, transposing via a DNA intermediate ("cut and paste"), and are more abundant near genes. The class II elements were discovered and described as controlling elements by Barbara McClintock in response to her extensive studies on DNA transposons and their ability to affect endogenous genes (McClintock 1984). Transposon activity can increase the expression levels of genes if inserted nearby (Naito *et al.* 2009). However, if inserted into a gene, can cause insertional mutagenesis (Liu *et al.* 2009) or chromosomal rearrangements (Yu, Zhang, and Peterson 2011). Because of their inherently mutagenic features, both classes of transposons are targeted by silencing machinery that results in DNA methylation and/or histone modifications which result in their suppression (Lippman *et al.* 2004, Volpe *et al.* 2011).

Among eukaryotes, the successful accumulation of TEs by their transposition plays a large part in the differences of genome size (Kidwell 2002). Several mechanisms are involved in controlling transposable elements (Lisch 2009), which have the potential to impact nearby gene activity and large-scale chromatin structure (Hollister and Gaut 2009). For example, the RdDM pathway has the potential to result in both DNA and histone modification, changing the activity of both genes and transposons. The loss of these silencing mechanisms can cause sudden transcriptional reactivation (Kato *et al.* 2003; Johnson *et al.* 2007). Transposons can be affected by various regulatory mechanisms in response to developmental cues (embryonic stage) or environmental stresses (such as heat shock) (Gehring, Bubb, and Henikoff 2009). With current evidence, it is unclear to what extent and what frequency the mechanisms of TE regulation influence surrounding chromatin, or conversely how flanking chromatin influences TEs.

In *Arabidopsis thaliana*, methylated TEs were shown to be negatively correlated with gene expression (Hollister and Gaut 2009; Pereira, Enard, and Eyre-Walker 2009), and there is evidence that methylated TEs near genes are selected against (Hollister and Gaut 2009). However, the interaction of transposons and genes in maize is more complicated. Recent studies in maize have categorized class I elements based on their heterochromatin spreading into flanking low-copy regions (Eichten *et al.* 2012). Class I elements were grouped based on whether they showed evidence of spreading of H3K9me2 and 5mC, spreading of H3K9me2 only, or no evidence of spreading, called non-spreading. When either spreading group was found near genes, those genes had lower expression levels compared to genes with a transposon group of non-spreading (Eichten *et al.* 2012). In another study, it was found that DNA transposons (class II) have high levels of CHH methylation and are associated with highly expressed genes (Gent *et al.* 2013).

Centromeric structure and underlying chromatin and genetic factors

The most repeat-rich regions of all eukaryotes are the centromeres, which interact with kinetochore proteins to confer accurate chromosome segregation (Fukagawa and Earnshaw 2014). Centromeres are defined epigenetically by the H3 histone variant CENH3, first discovered in humans to be necessary for kinetochore formation, CENH3 is currently viewed as the foundation of all kinetochores (Howman *et al.* 2000; Earnshaw and Rothfield 1985; Earnshaw and Migeon 1985). Traditionally, the word "kinetochore" is defined as the proteinaceous complex that interacts with "centromere" DNA. All eukaryotic organisms require a kinetochore, and the underlying centromeric sequence varies among organisms, from the simplest point centromere, which contains one CENH3 over ~125 bp sequence in budding yeast and localization defined genetically, to the megabase sized centromeres in maize made of repetitive elements and activation defined epigenetically (Malik and Henikoff 2009; Dawe and Henikoff 2006).

Centromeric sequences in maize are generally composed of a tandem repeat and LTR retrotransposon families, called CentC and CRM respectively (Miller *et al.* 1998; Presting *et al.* 1998; Zhong 2002). Although there are very few single-copy regions in centromeres, centromeres 2 and 5 have been assembled entirely (Wolfgruber, Sharma, *et al.* 2009). The key to this success was to exploit unique junction sites formed when recently active transposons insert into older transposons (Wolfgruber, Sharma, *et al.* 2009; Shi *et al.* 2010). A modified AFLP protocol called transposon display (TD) was used to amplify sequences flanking CRM2 retrotransposons(Casa *et al.* 2000, Wolfgruber, Sharma, *et al.* 2009). CRM and other transposon junction sites were used to identify additional polymorphism to facilitiate assembly of existing BAC sequences (Wolfgruber, Sharma, *et al.* 2009, Luce *et al.* 2006). Once the centromere sequences were constructed, several questions about the centromere could be further tested, including assays of centromere diversity and age, mechanisms of change, and stability over lineages (Wolfgruber, Sharma, *et al.* 2009, Shi *et al.* 2010, Gent *et al.* 2015, Bilinski *et al.* 2014). For example, Wolfgruber *et al.* 2009 found that CRM2 elements target centromeres in a region-specific rather than a sequence-specific manner. Centromere locations appear to be fluid over evolutionary time frames (Wolfgruber, Molecular Biosciences and Bioengineering, *et al.* 2009). For instance all centromeres have blocks of CentC arrays (Jin *et al.* 2004, Wolfgruber, Sharma, *et al.* 2009), and recent studies have found that CentC arrays have been contracting since the domestication of maize (Bilinski *et al.* 2014). One hypothesis is that the reduction of CentC has occurred as a result of intra-strand recombination between joined elements (Wolfgruber, Sharma, *et al.* 2009). On smaller evolutionary time scales, however, the variation in centromere positioning between B73 inbred lines, separated by multiple generations, is small (Gent *et al.* 2015). Centromere movement is thought to result from major genetic changes rather than an inherent fluidity of centromere positioning (Gent *et al.* 2015).

The Centromeric Retroelements (CRM in maize) are members of the *Ty3-gypsy* family of retrotransposons and contain a polyprotein open reading frame flanked by tandem long terminal repeats (LTR) at the ends. Interestingly, the CR elements differ from other retroelements in the chromodomain of the integrase protein, which has been proposed to contribute to the apparent targeting to CENH3-rich regions (Gorinšek, Gubenšek, and Kordiš 2005). CentA, a non-autonomous CRM was the first of its kind found and discovered to be enriched in maize centromeres by use of fluorescent *in situ* hybridization (FISH) (Nagaki *et al.* 2004). The first two autonomous CRM elements in maize were CRM1 (Nagaki *et al.* 2004) and CRM2 (Zhong 2002) which were also found to be enriched in centromeric regions. Most recently, a third CentA-related non-autonomous CR element called CRM3 was described (Sharma and Presting 2008). Recently inserted CRM1 and CRM2 elements are associated with CENH3 regions, while older CRM elements tend to be outside CENH3 regions. When a kinetochore forms, the CENH3-bound and nonCENH3-bound DNA make up a higher-order chromosomal structure. The CRM subfamilies can be distinguished cytologically, such that CRM1 is primarily found in the interior cohesive region while CRM2 is found in the outer kinetochore-associated region (Wolfgruber, Sharma, *et al.* 2009). The available data describe a rapidly evolving centromeric sequence where new CR elements insert into active kinetochore regions, displacing genes and older elements to flanking regions.

CENH3 is necessary to initiate interacting kinetochore proteins which also interact with cellular structures, such as microtubules, to ensure the faithful transmission of the chromosomes. Better understanding of how centromeres function is essential for studying other mechansims involved, such as microtubule attachment or aspects of basic cellular division. Futhermore, understanding all features of centromeres and kinetochore formation will contibute to future biotechnology and engineering centromeres. Experiments in centromere function expand our knowledge of what determines centromere length and why centromere size may vary across taxa, but all the factors that contribute to, or limits, centromere length is still not clear. However, a recent study comparing grasses found 99% of the variation for CENH3 staining (estimate of centromeres size) can be explained as a function of genome size (Zhang and Dawe 2012). Understanding the factors involved in determining the optimal centromere size in a cell is essential for determing what fundamentally defines a centromere. For instance, positing that centromere size is not determined solely by chromosome size, we can predict what might occur if a chromosome enters a foreign cell with a larger genome. Maize has a smaller genome than oat, yet the two species can be crossed to create oat-maize addition lines (oat lines with an additional maize chromosome) (Wang *et al.* 2013). The centromere sizes of several separate maize chromosomes in an oat-maize addition lines were analyzed, and in nine cases, the maize centromere expanded, and in two of those the centromere moved to completely new locations

(Wang *et al.* 2013). The potential benefits of engineering a centromere and adjacent gene content are many (Houben and Schubert 2007), but it is not clear what properties of a centromere are necessary to maintain a fully functional centromere. A centromere that is too small may not segregate properly, and there are substantial technical limitations to engineering large centromeres. Although recent advances in genome engineering have resulted in transformation of a 1100kb repeat array in maize using biolistics, and the average centromere size in maize is around 1Mb (Zhang and Dawe 2012). Further, it is known that certain repetitive elements make up plant centromeres, but a kinetochore will not form over centromere-like repetitive regions spontaneously, even when containing specific elements that would theoretically promote activation (Zhang and Dawe 2012). In addition to agronomical benefits, experimenting with centromere length and deletion of important regions of the centromere could potentially reveal sequence features that are necessary for centromere function. Reducing centromere size is very difficult to study because doing so often results in reduced transmission rates, and when an organism is not viable, results are difficult to study.

Centromeres: neocentromere, inactivation and reactivation

Fundamental to understanding centromere evolution is the study of neocentromeres, which is the ectopic activation of a centromere at a previously inactive position, and occurs across taxa from fungi and mammals to plants (Burrack and Berman 2012). Neocentromere formation occurs when CENH3 localizes to a new, formally unassociated CENH3 sequence, either by expanding or forming at a completely new locus. In addition to forming at new locus, centromeres can become inactivated even though centromere sequence is still present (Earnshaw and Migeon 1985, Han, Lamb, and Birchler 2006). When the centromeric sequence is lost, neocentromeres have been found to form at ectopic sequences (Zhang *et al.* 2013). When a chromosome is fragmented and centromere DNA; CentC and CRM are lost, CENH3 protein produced by the oat integrates into neocentromere position of maize chromosome 3, where no former CentC or CRM2 is found (Topp *et al.* 2009). These instances would suggest that centromeric sequence is not sufficient for centromere localization, however in a recent study of the maize B chromosome, a previously inactivated centromere, reactivated roughly over the native centromeric region (Han *et al.* 2009). In maize, the abnormal chromosome 10 (Ab10) line can form neocentromeres at the heterochromatic knob positions (Kanizay *et al.* 2013), and although knobs in maize do not contain centromere sequence, they do contain tandem repeats called knob180 similar in length to CentC (Kanizay and Dawe 2009). This begs the question of what mechanisms define the centromere boundaries and why don't neocentromeres form all the time? Topp *et al.* 2009 proposed that centromere stability is established by a "critical mass" or threshold level of CENH3 enrichment, in which it can serve as a functional centromere. CENH3 likely targets centromere-like DNA, and once CENH3 is established, it acts as a self-perpetuating CENH3 loop, where the kinetochore is maintained unless disrupted externally. Centromeres then will expand to a necessary length appropriate for the organism's genome size (Zhang and Dawe 2012). Over evolutionary time, the establishment of CENH3 at a centromere like repetitive region or otherwise inactive chromatin, in which other repetitive elements target and in return become silenced (Wolfgruber, Sharma, *et al.* 2009). As transposons continue to insert, older transposons are pushed out and become part of the boundary defining heterochromatin surrounding the centromere. Chromatin not conducive for CENH3 localization provides a natural boundary for the centromere.

To aproach all unaswered questions of the centromere, a more pliable model is necessary, because often the disruption of the centromere or interacting proteins result in inviable organisms. The B chromosome discussed below serves as an excellent model to study centromere function because it can undergo dramatic centromere events while having limited consequences on plant health.

The maize B chromosome

The B cromosome in *Zea mays* has been an important genetic tool for studying the functional properties of centromeres. B chromosomes are not necessary for the survival of maize and are only detrimental in high copy numbers. They are maintained in maize populations by a preferential segregation mechanism through males (Jones and Rees 1982). If a B centromere is translocated to an autosome, the subsequent chromosome (such as TB-9Sb) can also be preferentially transmitted through males (Heitz 1928, Carlson 1970). B centromeres can undergo a fission event, in which the B centromere loses portions of the centromere domain, or may undergo fusion, in which two copies of the same centromere can join to form new centromere variants. These centromeric structural changes fall under the term “misdivision derivative” (Heitz 1928, Carlson and Phillips 1986). Carlson found a TB-9Sb line that underwent a chromosome fission event, resulting in a duplication of chromosome 9S. With the use of this line, Kaszás and Birchler 1996 went on to create several additional misdivision lines, undergoing fusion and fission events, resulting in different size centromeres (Kaszás and Birchler 1998, 1996).

The B chromosome is mostly made of heterochromatin, with an abundant B specific repeat named *Zea mays* B specific (ZmBs) (Alfenito and Birchler 1993). The sequence of the B repeat allowed Kaszàs and Birchler 1996 to use Southern blotting to interpret the repeat complexity and content in each misdivision derivative, and to correlate the results with the stability and transmission of the chromosomes (Kaszás and Birchler 1996). The first misdivision event of TB-9Sb produced a pseudoisochromosome, in which the chromosome arms appear to be duplicates of each other except that one arm contains a block of heterochromatin denoted as (+) and the other missing this domain and denoted with (-). When the pseudoisochromosome misdivides, one side of the centromere can be broken off and subsequently scored cytologically by the presence or absence of the heterochromatic region (Kaszás and Birchler 1996). By assaying ZmBs restriction fragment complexity and abundance, it was found that when the centromere is reduced in size, there is a decrease in the transmissibility of the B chromosome. Certain fragmented portions, for instance the 370-kb PmeI fragment, was associated with higher transmission rates of the B chromosome (Kaszás and Birchler 1998). Furthermore, a 55-kb sequence containing the ZmBs repeat is critical for meiotic function (Kaszàs and Birchler 1998, 1996). Although the ZmBs repeat is a useful marker for the B centromere, the repeat itself is not necessarily required for centromere function, and cytological studies have shown that CentC, CRM2, and ZmBs are interspersed along the length of the B centromere (Jin *et al.* 200; Lamb, Yu, *et al.* 2007).

High resolution mapping efforts have focused on using a technique called fiber-FISH where chromatin is stretched on a slide and hybridized with the major centromere repeats (Jin *et al.* 2005). The results revealed that the centromere domain of B chromosome is defined by a ~700-kb dense array of CentC, CRM and ZmBs repeats in a unique pattern (Jin *et al.* 2005). The core CentC domain appeared to co-localize with CENH3. Outside of the CENH3-enriched regions, all three repeats ZmBs, CentC and CRM2 are unassociated with centromere proteins and dispersed along the chromosome arm in a sparse pattern (Jin *et al.* 2005, Lamb, Yu, *et al.* 2007). The misdivision lines assayed involved simple linear deletions with breakpoints within the 700 kb core region, such that total centromere length was positively correlated with transmission levels (Jin *et al.* 2005).

The misdivision derivative lines have also been used to create chromosomes with two centromeres (dicentric chromosomes) (Han *et al.* 2006). Han *et al.* (2009) demonstrated that dicentric B chromosomes are unstable: when two different sized centromeres were pitted against each other on a dicentric chromosome, the smaller of the two was inactivated (Han, Lamb, and Birchler 2006). Such inactive B-centromeres can also be re-activated, and several newly-reactive centromere lines have been described (Han *et al.* 2009). Other more recent data have shown that neocentromeres occasionally form during the selection of misdivision derivatives, particularly if the core centromere is severely disrupted or deleted. In several cases ectopic centromeres formed at entirely new positions on the chromosome arm (Liu *et al.* 2015). These studies demonstrate that a B centromere can spontaneously inactivate, and then spontaneously re-form at roughly the same position as well as at sequence formally unassociated with centromere proteins. The large number of misdivision derivative should make it possible to study exactly which portions of the centromere must be missing for an ectopic centromere to form or what sequences are need to re-form a centromere at the native centromere site. However, such experiments will not be possible until a contiguous B centromere sequence is available as a scaffold for interpreting ChIP-seq data.

Purpose of study

Recent studies analyzing chromatin markers, such as DNA methylation and histone modifications, have focused on examples of gene-TE chromatin interaction. In maize, 39 class I families show spreading of DNA methylation and H3K9me2 into low-copy flanking regions of up to 3kb, resulting in genes being expressed at lower levels compared to non-spreading TEs

(Eichten *et al.* 2012). In addition, we showed that the upstream (1 kb) regions of most genes show unusually high levels of CHH methylation and siRNAs, and are often composed of MITEs (Gent *et al.* 2013). We propose to further investigate the maize genome and DNA methylation in relation to class II and class I elements, as well as test the effectiveness of MITE superfamilies as boundary elements between heterochromatin and genic regions. We also propose to develop the first molecular map of the B chromosome with use of transposon display markers, and create a discontinuous B centromere sequence representing the most robust sequence of the B centromere to date. Similar to efforts with centromere 2 and 5 centromere sequences in maize, we hope our data will contribute to a fully determined, contiguous sequence of the B centromere.

References

Alfenito, Mark R, and James A Birchler. 1993. "Molecular characterization of a maize B chromosome centric sequence." *Genetics* 135 (2):589-597.

Almeida, Ricardo, and Robin C. Allshire. 2005. "RNA silencing and genome regulation." *Trends in Cell Biology* 15 (5):251-258. doi: 10.1016/j.tcb.2005.03.006.

Bernstein, Bradley E, Alexander Meissner, and Eric S Lander. 2007. "The mammalian epigenome." *Cell* 128 (4):669-681.

Bilinski, P., K. Distor, J. Gutierrez-Lopez, G. Mendoza Mendoza, J. Shi, R. K. Dawe, and J. Ross-Ibarra. 2014. doi: 10.1101/005058.

Burrack, Laura S, and Judith Berman. 2012. "Neocentromeres and epigenetically inherited features of centromeres." *Chromosome research* 20 (5):607-619.

Carlson, W. R. 1970. "Nondisjunction and isochromosome formation in the B chromosome of maize." *Chromosoma* 30:356-65. doi: 10.1007/BF00321067.

Carlson, Wayne R, and Ronald L Phillips. 1986. "The B chromosome of maize." *Critical reviews in plant sciences* 3 (3):201-226.

Casa, Alexandra M, Cory Brouwer, Alexander Nagel, Lianjiang Wang, Qiang Zhang, Stephen Kresovich, and Susan R Wessler. 2000. "The MITE family Heartbreaker (Hbr): molecular markers in maize." *Proceedings of the National Academy of Sciences* 97 (18):10083-10089.

Chen, Pao-Yang, Suhua Feng, JW Joo, Steve E Jacobsen, and Matteo Pellegrini. 2011. "A comparative analysis of DNA methylation across human embryonic stem cell lines." *Genome Biol* 12 (7):R62.

Cokus, Shawn J, Suhua Feng, Xiaoyu Zhang, Zugen Chen, Barry Merriman, Christian D Haudenschild, Sriharsa Pradhan, Stanley F Nelson, Matteo Pellegrini, and Steven E Jacobsen. 2008. "Shotgun bisulphite sequencing of the Arabidopsis genome reveals DNA methylation patterning." *Nature* 452 (7184):215-219.

Dawe, R Kelly, and Steven Henikoff. 2006. "Centromeres put epigenetics in the driver's seat." *Trends in biochemical sciences* 31 (12):662-669.

Earnshaw, W. C., and B. R. Migeon. 1985. "Three related centromere proteins are absent from the inactive centromere of a stable isodicentric chromosome." *Chromosoma* 92 (4):290-6.

Earnshaw, William C, and Naomi Rothfield. 1985. "Identification of a family of human centromere proteins using autoimmune sera from patients with scleroderma." *Chromosoma* 91 (3-4):313-321.

Eichten, Steven R, Nathanael A Ellis, Irina Makarevitch, Cheng-Ting Yeh, Jonathan I Gent, Lin Guo, Karen M McGinnis, Xiaoyu Zhang, Patrick S Schnable, and Matthew W Vaughn. 2012. "Spreading of heterochromatin is limited to specific families of maize retrotransposons." *PLoS genetics* 8 (12):e1003127.

Feng, Suhua, Shawn J. Cokus, Xiaoyu Zhang, Pao-Yang Chen, Magnolia Bostick, Mary G. Goll, Jonathan Hetzel, Jayati Jain, Steven H. Strauss, Marnie E. Halpern, Chinweike Ukomadu, Kirsten C. Sadler, Sriharsa Pradhan, Matteo Pellegrini, Steven E. Jacobsen, Suhua Feng, Shawn J. Cokus, Xiaoyu Zhang, Pao-Yang Chen, Magnolia Bostick, Mary G. Goll, Jonathan Hetzel, Jayati Jain, Steven H. Strauss, Marnie E. Halpern, Chinweike Ukomadu, Kirsten C. Sadler, Sriharsa Pradhan, Matteo Pellegrini, and Steven E. Jacobsen. 2010.

"Conservation and divergence of methylation patterning in plants and animals." *Proceedings of the National Academy of Sciences*. doi: 10.1073/pnas.1002720107.

Fuchs, Jörg, Dmitri Demidov, Andreas Houben, and Ingo Schubert. 2006. "Chromosomal histone modification patterns–from conservation to diversity." *Trends in plant science* 11 (4):199-208.

Fukagawa, Tatsuo, and William C Earnshaw. 2014. "The Centromere: Chromatin Foundation for the Kinetochore Machinery." *Developmental Cell* 30 (5):496-508. doi: 10.1016/j.devcel.2014.08.016.

Gardiner-Garden, M, and M Frommer. 1987. "CpG islands in vertebrate genomes." *Journal of molecular biology* 196 (2):261-282.

Gehring, Mary, Kerry L Bubb, and Steven Henikoff. 2009. "Extensive demethylation of repetitive elements during seed development underlies gene imprinting." *Science* 324 (5933):1447-1451.

Gent, Jonathan I, Nathanael A Ellis, Lin Guo, Alex E Harkess, Yingyin Yao, Xiaoyu Zhang, and R Kelly Dawe. 2013. "CHH islands: de novo DNA methylation in near-gene chromatin regulation in maize." *Genome research* 23 (4):628-637.

Gent, Jonathan I, Thelma F Madzima, Rechien Bader, Matthew R Kent, Xiaoyu Zhang, Maike Stam, Karen M McGinnis, and R Kelly Dawe. 2014. "Accessible DNA and relative depletion of H3K9me2 at maize loci undergoing RNA-directed DNA methylation." *The Plant Cell* 26 (12):4903-4917.

Gent, Jonathan I, Kai Wang, Jiming Jiang, and R Kelly Dawe. 2015. "Stable Patterns of CENH3 Occupancy Through Maize Lineages Containing Genetically Similar Centromeres." *Genetics*:genetics. 115.177360.

Goll, Mary Grace, and Timothy H Bestor. 2005. "Eukaryotic cytosine methyltransferases." *Annu. Rev. Biochem.* 74:481-514.

Gorinšek, Benjamin, Franc Gubenšek, and Dušan Kordiš. 2005. "Phylogenomic analysis of chromoviruses." *Cytogenetic and genome research* 110 (1-4):543-552.

Han, F., J. C. Lamb, and J. A. Birchler. 2006. "High frequency of centromere inactivation resulting in stable dicentric chromosomes of maize." *Proc Natl Acad Sci U S A* 103 (9):3238-43. doi: 10.1073/pnas.0509650103.

Han, Fangpu, Zhi Gao, James A. Birchler, Fangpu Han, Zhi Gao, and James A. Birchler. 2009. "Reactivation of an Inactive Centromere Reveals Epigenetic and Structural Components for Centromere Specification in Maize." *The Plant Cell Online* 21 (7):1929-1939. doi: 10.1105/tpc.109.066662.

He, Xin-Jian, Yi-Feng Hsu, Shihua Zhu, Andrzej T. Wierzbicki, Olga Pontes, Craig S. Pikaard, Hai-Liang Liu, Co-Shine Wang, Hailing Jin, and Jian-Kang Zhu. 2009. "An Effector of RNA-Directed DNA Methylation in Arabidopsis Is an ARGONAUTE 4- and RNA-Binding Protein." *Cell* 137 (3):498-508. doi: 10.1016/j.cell.2009.04.028.

Heitz, E. (1928). Das heterochromatin der moose. *I. Jahrb. Wiss. Botanik, 69*, 762-818.

Hollister, J. D., and B. S. Gaut. 2009. "Epigenetic silencing of transposable elements: a trade-off between reduced transposition and deleterious effects on neighboring gene expression." In *Genome Res*, 1419-28. United States.

Houben, Andreas, and Ingo Schubert. 2007. "Engineered plant minichromosomes: a resurrection of B chromosomes?" *The Plant Cell* 19 (8):2323-2327.

Howman, E. V., K. J. Fowler, A. J. Newson, S. Redward, A. C. MacDonald, P. Kalitsis, and K. H. Choo. 2000. "Early disruption of centromeric chromatin organization in centromere protein A (Cenpa) null mice." *Proc Natl Acad Sci U S A* 97 (3):1148-53.

Jenuwein, Thomas, and C David Allis. 2001. "Translating the histone code." *Science* 293 (5532):1074-1080.

Jin, W., J. C. Lamb, J. M. Vega, R. K. Dawe, J. A. Birchler, and J. Jiang. 2005. "Molecular and functional dissection of the maize B chromosome centromere." *Plant Cell* 17 (5):1412-23. doi: 10.1105/tpc.104.030643.

Jin, W., J. R. Melo, K. Nagaki, P. B. Talbert, S. Henikoff, R. K. Dawe, and J. Jiang. 2004. "Maize centromeres: organization and functional adaptation in the genetic background of oat." *Plant Cell* 16 (3):571-81. doi: 10.1105/tpc.018937.

Johnson, L. M., M. Bostick, X. Zhang, E. Kraft, I. Henderson, J. Callis, and S. E. Jacobsen. 2007. "The SRA methyl-cytosine-binding domain links DNA and histone methylation." In *Curr Biol*, 379-84. England.

Jones, Peter A. 2012. "Functions of DNA methylation: islands, start sites, gene bodies and beyond." *Nature Reviews Genetics* 13 (7):484-492. doi: 10.1038/nrg3230.

Jones, Robert Neil, and Hubert Rees. 1982. *B chromosomes*: Academic Press.

Kanizay, L., and R. K. Dawe. 2009. "Centromeres: long intergenic spaces with adaptive features." *Funct Integr Genomics* 9 (3):287-92. doi: 10.1007/s10142-009-0124-0.

Kanizay, Lisa B, Tanja Pyhäjärvi, Elizabeth G Lowry, Matthew B Hufford, Daniel G Peterson, Jeffrey Ross-Ibarra, and R Kelly Dawe. 2013. "Diversity and abundance of the abnormal chromosome 10 meiotic drive complex in Zea mays." *Heredity* 110 (6):570-577.

Kaszás, E., and J. A. Birchler. 1996. "Misdivision analysis of centromere structure in maize." *EMBO J* 15 (19):5246-55.

Kaszás, E., and J. A. Birchler. 1998. "Meiotic transmission rates correlate with physical features of rearranged centromeres in maize." *Genetics* 150 (4):1683-92.

Kato, M., A. Miura, J. Bender, S. E. Jacobsen, and T. Kakutani. 2003. "Role of CG and non-CG methylation in immobilization of transposons in Arabidopsis." In *Curr Biol*, 421-6. England.

Kidwell, M. G. 2002. "Transposable elements and the evolution of genome size in eukaryotes." *Genetica* 115 (1):49-63.

Klose, Robert J, and Adrian P Bird. 2006. "Genomic DNA methylation: the mark and its mediators." *Trends in biochemical sciences* 31 (2):89-97.

Kouzarides, T. 2007. "Chromatin modifications and their function." In *Cell*, 693-705. United States.

Lamb, J. C., J. M. Meyer, B. Corcoran, A. Kato, F. Han, and J. A. Birchler. 2007. "Distinct chromosomal distributions of highly repetitive sequences in maize." *Chromosome Res* 15 (1):33-49. doi: 10.1007/s10577-006-1102-1.

Lamb, J. C., W. Yu, F. Han, and J. A. Birchler. 2007. "Plant chromosomes from end to end: telomeres, heterochromatin and centromeres." *Curr Opin Plant Biol* 10 (2):116-22. doi: 10.1016/j.pbi.2007.01.008.

Law, Julie A., and Steven E. Jacobsen. 2010a. "Establishing, maintaining and modifying DNA methylation patterns in plants and animals." *Nat Rev Genet* 11 (3):204-220. doi: 10.1038/nrg2719.

Law, Julie A., and Steven E. Jacobsen. 2010b. "Establishing, maintaining and modifying DNA methylation patterns in plants and animals." *Nature Reviews Genetics* 11 (3):204-220. doi: 10.1038/nrg2719.

Li, Bing, Michael Carey, and Jerry L Workman. 2007. "The role of chromatin during transcription." *Cell* 128 (4):707-719.

Li, En, and Yi Zhang. 2014. "DNA methylation in mammals." *Cold Spring Harb Perspect Biol* 6:a019133.

Lippman, Z., A. V. Gendrel, M. Black, M. W. Vaughn, N. Dedhia, W. R. McCombie, K. Lavine, V. Mittal, B. May, K. D. Kasschau, J. C. Carrington, R. W. Doerge, V. Colot, and R. Martienssen. 2004. "Role of transposable elements in heterochromatin and epigenetic control." In *Nature*, 471-6. England.

Lisch, Damon. 2009. "Epigenetic Regulation of Transposable Elements in Plants." *http://dx.doi.org/10.1146/annurev.arplant.59.032607.092744*. doi: 10.1146/annurev.arplant.59.032607.092744.

Liu, Chunyan, Falong Lu, Xia Cui, and Xiaofeng Cao. 2010. "Histone Methylation in Higher Plants." *Annual Review of Plant Biology* 61 (1):395-420. doi: 10.1146/annurev.arplant.043008.091939.

Liu, R., C. Vitte, J. Ma, A. A. Mahama, T. Dhliwayo, M. Lee, and J. L. Bennetzen. 2007. "A GeneTrek analysis of the maize genome." *Proc Natl Acad Sci U S A* 104 (28):11844-9. doi: 10.1073/pnas.0704258104.

Liu, S., C. T. Yeh, T. Ji, K. Ying, H. Wu, H. M. Tang, Y. Fu, D. Nettleton, and P. S. Schnable. 2009. "Mu transposon insertion sites and meiotic recombination events co-localize with epigenetic marks for open chromatin across the maize genome." *PLoS Genet* 5 (11):e1000733. doi: 10.1371/journal.pgen.1000733.

Liu, Yalin, Handong Su, Junling Pang, Zhi Gao, Xiu-Jie Wang, James A Birchler, and Fangpu Han. 2015. "Sequential de novo centromere formation and inactivation on a chromosomal fragment in maize." *Proceedings of the National Academy of Sciences* 112 (11):E1263-E1271.

Luce, Amy C, Anupma Sharma, Oliver SB Mollere, Thomas K Wolfgruber, Kiyotaka Nagaki, Jiming Jiang, Gernot G Presting, and R Kelly Dawe. 2006. "Precise centromere mapping using a combination of repeat junction markers and chromatin immunoprecipitation–polymerase chain reaction." *Genetics* 174 (2):1057-1061.

Luger, Karolin, Armin W Mäder, Robin K Richmond, David F Sargent, and Timothy J Richmond. 1997. "Crystal structure of the nucleosome core particle at 2.8 Å resolution." *Nature* 389 (6648):251-260.

Malik, H. S., and S. Henikoff. 2009. "Major evolutionary transitions in centromere complexity." *Cell* 138 (6):1067-82. doi: 10.1016/j.cell.2009.08.036.

Maunakea, Alika K., Raman P. Nagarajan, Mikhail Bilenky, Tracy J. Ballinger, Cletus D’Souza, Shaun D. Fouse, Brett E. Johnson, Chibo Hong, Cydney Nielsen, Yongjun Zhao, Gustavo Turecki, Allen Delaney, Richard Varhol, Nina Thiessen, Ksenya Shchors, Vivi M. Heine, David H. Rowitch, Xiaoyun Xing, Chris Fiore, Maximiliaan Schillebeeckx, Steven J. M. Jones, David Haussler, Marco A. Marra, Martin Hirst, Ting Wang, and Joseph F. Costello. 2010. "Conserved role of intragenic DNA methylation in regulating alternative promoters." *Nature* 466 (7303):253-257. doi: 10.1038/nature09165.

McClintock, B. 1984. "The significance of responses of the genome to challenge." *Science* 226 (4676):792-801.

Miller, Joseph T, Fenggao Dong, Scott A Jackson, Junqi Song, and Jiming Jiang. 1998. "Retrotransposon-related DNA sequences in the centromeres of grass chromosomes." *Genetics* 150 (4):1615-1623.

Nagaki, Kiyotaka, Zhukuan Cheng, Shu Ouyang, Paul B. Talbert, Mary Kim, Kristine M. Jones, Steven Henikoff, C. Robin Buell, and Jiming Jiang. 2004. "Sequencing of a rice centromere uncovers active genes." *Nature Genetics* 36 (2):138-145. doi: 10.1038/ng1289.

Naito, Ken, Feng Zhang, Takuji Tsukiyama, Hiroki Saito, C. Nathan Hancock, Aaron O. Richardson, Yutaka Okumoto, Takatoshi Tanisaka, and Susan R. Wessler. 2009. "Unexpected consequences of a sudden and massive transposon amplification on rice gene expression." *Nature* 461 (7267):1130-1134. doi: 10.1038/nature08479.

Pereira, V., D. Enard, and A. Eyre-Walker. 2009. "The effect of transposable element insertions on gene expression evolution in rodents." *PLoS One* 4 (2):e4321. doi: 10.1371/journal.pone.0004321.

Pfluger, Jennifer, and Doris Wagner. 2007. "Histone modifications and dynamic regulation of genome accessibility in plants." *Current Opinion in Plant Biology* 10 (6):645-652. doi: 10.1016/j.pbi.2007.07.013.

Pikaard, Craig S, Jeremy R Haag, Thomas Ream, and Andrzej T Wierzbicki. 2008. "Roles of RNA polymerase IV in gene silencing." *Trends in plant science* 13 (7):390-397.

Presting, Gernot G, Ludmilla Malysheva, Jörg Fuchs, and Ingo Schubert. 1998. "ATY3/GYPSYretrotransposon-like sequence localizes to the centromeric regions of cereal chromosomes." *The Plant Journal* 16 (6):721-728.

Salem, Carol E, Isabel DC Markl, Christina M Bender, Felicidad A Gonzales, Peter A Jones, and Gangning Liang. 2000. "PAX6 methylation and ectopic expression in human tumor cells." *International journal of cancer* 87 (2):179-185.

Schnable, P. S., D. Ware, R. S. Fulton, J. C. Stein, F. Wei, S. Pasternak, T. K. Wolfgruber, L. Yang, Y. Yu, L. Zhang, S. Zhou, Q. Zhu, J. L. Bennetzen, R. K. Dawe, J. Jiang, N. Jiang, G. G. Presting, S. R. Wessler, S. Aluru, R. A. Martienssen, S. W. Clifton, W. R. McCombie, R. A. Wing, and R. K. Wilson. 2009. "The B73 maize genome: complexity, diversity, and dynamics." Science 326 (5956):1112-5. doi: 10.1126/science.1178534.Sharma, Anupma, and Gernot G Presting. 2008. "Centromeric retrotransposon lineages predate the maize/rice divergence and differ in abundance and activity." *Molecular Genetics and Genomics* 279 (2):133-147.

Shi, J., S. E. Wolf, J. M. Burke, G. G. Presting, J. Ross-Ibarra, and R. K. Dawe. 2010. "Widespread gene conversion in centromere cores." *PLoS Biol* 8 (3):e1000327. doi: 10.1371/journal.pbio.1000327.

Shi, Jinghua, and R Kelly Dawe. 2006. "Partitioning of the maize epigenome by the number of methyl groups on histone H3 lysines 9 and 27." *Genetics* 173 (3):1571-1583.

Varley, Katherine E, Jason Gertz, Kevin M Bowling, Stephanie L Parker, Timothy E Reddy, Florencia Pauli-Behn, Marie K Cross, Brian A Williams, John A Stamatoyannopoulos, and Gregory E Crawford. 2013. "Dynamic DNA methylation across diverse human cell lines and tissues." *Genome research* 23 (3):555-567.

Veiseth, Silje V, Mohummad A Rahman, Kyoko L Yap, Andreas Fischer, Wolfgang Egge-Jacobsen, Gunter Reuter, Ming-Ming Zhou, Reidunn B Aalen, and Tage Thorstensen. 2011. "The SUVR4 histone lysine methyltransferase binds ubiquitin and converts H3K9me1 to H3K9me3 on transposon chromatin in Arabidopsis." *PLoS Genet* 7 (3):e1001325.

Volpe, Tom, Robert A. Martienssen, Tom Volpe, and Robert A. Martienssen. 2011. "RNA Interference and Heterochromatin Assembly." *Cold Spring Harbor Perspectives in Biology* 3 (9). doi: 10.1101/cshperspect.a003731.

Wang, K., Y. Wu, W. Zhang, R. K. Dawe, and J. Jiang. 2013. "Maize centromeres expand and adopt a uniform size in the genetic background of oat." *Genome Research* 24 (1):107-116. doi: 10.1101/gr.160887.113.

Wolfgruber, T. K., A. Sharma, K. L. Schneider, P. S. Albert, D. H. Koo, J. Shi, Z. Gao, F. Han, H. Lee, R. Xu, J. Allison, J. A. Birchler, J. Jiang, R. K. Dawe, and G. G. Presting. 2009. "Maize centromere structure and evolution: sequence analysis of centromeres 2 and 5 reveals dynamic Loci shaped primarily by retrotransposons." *PLoS Genet* 5 (11):e1000743. doi: 10.1371/journal.pgen.1000743.

Wolfgruber, Thomas K *et al.* 2009. "Maize Centromere Structure and Evolution: Sequence Analysis of Centromeres 2 and 5 Reveals Dynamic Loci Shaped Primarily by Retrotransposons." *PLoS Genetics* 5 (11). doi: 10.1371/journal.pgen.1000743.

Yu, C., J. Zhang, and T. Peterson. 2011. "Genome rearrangements in maize induced by alternative transposition of reversed ac/ds termini." In *Genetics*, 59-67. United States.

Zhang, B., Z. Lv, J. Pang, Y. Liu, X. Guo, S. Fu, J. Li, Q. Dong, H. J. Wu, Z. Gao, X. J. Wang, and F. Han. 2013. "Formation of a functional maize centromere after loss of centromeric sequences and gain of ectopic sequences." *Plant Cell* 25 (6):1979-89. doi: 10.1105/tpc.113.110015.

Zhang, Han, and R. Kelly Dawe. 2012. "Total centromere size and genome size are strongly correlated in ten grass species." *Chromosome Research* 20 (4):403-412. doi: 10.1007/s10577-012-9284-1.

Zhang, Heng, Xinjian He, and Jian-Kang Zhu. 2014. "RNA-directed DNA methylation in plants." *RNA Biology* 10 (10):1593-1596. doi: 10.4161/rna.26312.

Zhang, X., and S. E. Jacobsen. 2006. "Genetic analyses of DNA methyltransferases in Arabidopsis thaliana." *Cold Spring Harb Symp Quant Biol* 71:439-47. doi: 10.1101/sqb.2006.71.047.

Zhang, X., J. Yazaki, A. Sundaresan, S. Cokus, S. W. Chan, H. Chen, I. R. Henderson, P. Shinn, M. Pellegrini, S. E. Jacobsen, and J. R. Ecker. 2006. "Genome-wide high-resolution mapping and functional analysis of DNA methylation in arabidopsis." In *Cell*, 1189-201. United States.

Zhang, Xiaoyu, Oliver Clarenz, Shawn Cokus, Yana V Bernatavichute, Matteo Pellegrini, Justin Goodrich, and Steven E Jacobsen. 2007. "Whole-genome analysis of histone H3 lysine 27 trimethylation in Arabidopsis." *PLoS Biol* 5 (5):e129.

Zhong, C. X. 2002. "Centromeric Retroelements and Satellites Interact with Maize Kinetochore Protein CENH3." *The Plant Cell Online* 14 (11):2825-2836. doi: 10.1105/tpc.006106.

Zilberman, Daniel, Mary Gehring, Robert K Tran, Tracy Ballinger, and Steven Henikoff. 2007. "Genome-wide analysis of Arabidopsis thaliana DNA methylation uncovers an interdependence between methylation and transcription." *Nature genetics* 39 (1):61-69.

CHAPTER 2

GENERATION OF A MAIZE B CENTROMERE MINIMAL MAP AND EVIDENCE FOR A NECESSARY CENTRAL CORE DOMAIN[1]

[1] Ellis, N. A., Douglas, R. N., Jackson, C. E., Birchler, J. A., Dawe R. K. To be submitted to G3.

Abstract

The maize B centromere has been exploited for gene mapping, gene dosage studies, building artificial minichromosomes, and for studying centromere epigenetics. However there are no sequence resources for this important model centromere. Here we used transposon display for the centromere-specific retroelement, CRM2, to identify a collection of 40 sequence tags that flank CRM2 insertion points on the B chromosome. These were confirmed to lie within the centromere by assaying deletion breakpoints from centromere misdivision derivatives (intra-centromere breakages caused by centromere malfunction). Markers were grouped together based on their association with other markers in the misdivision series, and assembled into a minimal map containing ~10 kb of sequence. To identify sequences that interact directly with centromere proteins, we carried out chromatin immunoprecipitation and used antibodies to Centromeric Histone H3, a defining feature of functional centromeric sequences. The ChIP-seq map was interpreted relative to the known transmission rates of centromere misdivision derivatives to identify a necessary centromere core domain spanning 31 markers. A line that is missing this region was recently shown to induce the formation of neocentromeres at ectopic sites. Our results support a growing body of evidence indicating that centromere stability relies on key structural motifs in a centromere core and that loss of the centromere core results in shifts in centromere function or location.

Introduction

Centromeres are structural features of the genome that serve as contact points between chromosomes and the spindles during cell division. The sequences within centromeres tend to be rapidly evolving repetitive elements that have evolved to stably recruit and maintain kinetochore components. In maize, the primary centromere sequences are a 156 bp tandem repeat, called CentC, and a family of retrotransposons called Centromeric Retroelements - Maize (CRM elements) (Ananiev, Phillips, and Rines 1998, Jiang *et al.* 2003). CentC occurs in continuous arrays extending for hundreds of kilobases, and CRM elements are clustered in nested arrangements along with less common retrotransposons (Jin *et al.* 2004, Wolfgruber *et al.* 2009). Genes and other single copy sequences are rare in centromeres. Nevertheless transposon junctions and restriction site polymorphisms made it possible to use standard BAC sequencing pipelines to assemble maize centromeres 2 and 5 in entirety (Wolfgruber *et al.* 2009). These sequences have been invaluable references for interpreting centromere structure and evolution (Gent *et al.* 2012, Wang *et al.* 2014, Bilinski *et al.* 2015, Gent *et al.* 2015). However, continued progress on assembling centromeres represents a major challenge, because while BAC assemblies can be powerful tools, they do not traverse all centromeres and are not available for most species.

Even where centromere sequences are known, the functional domains cannot be identified by sequence alone because centromere specification is heavily influenced by epigenetic factors (Allshire and Karpen 2008). In all species, the functional centromere domains are operationally defined by the presence of the centromere-specific histone H3 variant cenH3/CENP-A (CENH3 in plants), the basal-most centromere protein that serves to recruit overlying kinetochore proteins (Black and Bassett 2008). Chromatin immunoprecipitation with antibodies to CENH3 followed by high throughput sequencing (ChIP-seq) is considered the gold standard for defining centromere position. Interpreting ChIP-seq requires that there is a reference sequence that has sufficient polymorphism to unequivocally map short sequence reads. Although centromeres are generally composed of tandem repeats and transposons, the fact that they are interspersed with each other in maize (and other plants) provides for a remarkable degree of polymorphism (Luce *et al.* 2006, Shi *et al.* 2010). The maize CRM family contains many variants and over millions of years, have inserted into each other and into other repeats in random fashion (Sharma and Presting 2008). This diversity has been exploited using a technique called transposon display whereby primers to the long terminal repeat (LTR) of a common CRM element (CRM2) are used to amplify genomic DNA cleaved at a restriction site and ligated to an adapter primer (Shi *et al.* 2010). Transposon display yields marker bands that can be both genetically mapped and sequenced. If the marker sequence provides unique junction sites or polymorphisms, when combined with CENH3 ChIP- sequencing, the method can be further used to pinpoint functional centromere core domains.

The centromere of the maize B chromosome is an example of a centromere where sequence and marker information could be of great use. The B chromosome is a supernumerary chromosome that has the property of non-disjoining specifically at the second pollen mitosis so that it is preferentially transmitted (Roman 1948, Carlson 1978). Lines of maize with autosomal chromosomes containing knobs and B chromosomes were found to have increased instability, including breakage and loss of knobbed arms during the second mitotic division in pollen (Rhoades and Dempsey 1972, Rhoades, Dempsey, and Ghidoni 1967). The entire B chromosome is dispensable and has been exploited for multiple purposes including gene mapping, gene dosage analysis, genome engineering, and centromere studies (Nannas and Dawe 2015). The B chromosome contains a unique sequence called ZmBs (*Zea maize* B-specific) or simply the B-repeat (Alfenito and Birchler 1993). The B repeat is present in multiple locations on the B chromosome but is most abundant in and around the centromere as assayed by CENH3 staining (Lamb, Kato, and Birchler 2005, Jin *et al.* 2008). Fine-scale cytological analyses of stretched

DNA fibers (fiber-FISH) have revealed a 700 kb CentC and CRM-rich core that is interspersed with blocks of the B repeat (Jin *et al.* 2005).

Some isolates of the B chromosomes undergo frequent centromere misdivision and breakage of the centromere into two parts (Kaszas and Birchler 1996). Most centromere misdivision's result in smaller centromeres as assayed by the quantity of B repeat, however the size of the centromere can also increase as an outcome of misdivision (Kaszas and Birchler 1998). Transmission of the B chromosome is clearly reduced when centromeres become very small (Phelps-Durr and Birchler 2004). Five of the smaller misdivision-derived centromeres were assayed by fiber-FISH and shown to have deletions of the 700-kb B centromere core (Jin *et al.* 2005). Other screens have been used to identify chromosomes where the B centromere has become epigenetically inactivated (Han, Lamb, and Birchler 2006), and then from those lines chromosomes were identified that had regained centromere activity (Han, Gao, and Birchler 2009). Additional data demonstrate that centromere misdivision can involve the loss and reformation of centromeres at new locations (Liu *et al.* 2015). The molecular underpinning of these events is of great interest, however there are currently no mapping resources for the B centromere core.

Here we report the first molecular map of the maize B centromere and outline a method for assembling unique sequence from highly repetitive centromere regions. To this end we used CRM2 transposon display (TD) to identify 40 TD markers and mapped them within the centromere using misdivision derivatives. Markers were cloned, sequenced and assembled into a ~10 kb pseudocontig "minimal map", which was used to interpret CENH3 ChIP-seq data. We then compared the presence/absence of CENH3-enriched markers misdivision lines to identify a set of markers that are unique to the functional centromere core. We also generated seven PCR primer pairs that can be used to score the major domains of the centromere without employing CRM-TD. This important reference information can be used as a basis for interpreting the large collection of known B centromere variants and derivatives. In addition, the B centromere minimal map can serve as guide for a future full assembly of the B centromere.

Results

Identification of B centromere sequences by CRM2-TD

When transposons insert into other repetitive sequences, unique junction sites are created (Luce *et al.* 2006, You *et al.* 2010). We targeted the regions flanking CRM2 using transposon display, a modified Amplified Fragment Length Polymorphism (AFLP) approach that takes advantage of restriction site polymorphism flanking the conserved LTR sequences of the transposon (Shi *et al.* 2010). CRM2 primers, labeled with radioactive or fluorochrome-labeled nucleotides, were paired with adapter primers that bind to BfaI cleavage sites. An additional PCR step using adapter primers with three selective bases was used to reduce the number of bands and improve the resolution on gels and genotyping software. To identify bands specific to the B chromosome, we compared the banding pattern from B73, the sequenced reference inbred (Schnable *et al.* 2009), to a B73 line containing several B chromosomes (called the B+ line, Figure 2.1).

Sequence of the TD markers

Each marker unique to the B+ line was extracted from a polyacrylamide gel, PCR amplified, and Sanger sequenced (Figure 2.1, Table 2.1). A total of 250 CRM2-TD markers were initially discovered. However, after sequencing the TD markers, we observed that the selective bases on adapter primers were only weakly selective and that many of the bands were duplicates. We ultimately chose a set of 61 markers that are unique to the (entire) B chromosome. This collection of markers was then analyzed in a line containing a “miniB” chromosome that contains only the centromere and flanking pericentromeric regions (Yu *et al.* 2007). Twenty-one markers proved to be absent on the miniB chromosome, thereby narrowing the number of potential centromeric markers to 40.

Nine of the 40 sequences flanking CRM2 contained either the B repeat or CentC sequence as expected for the B centromere core region (Jin *et al.* 2005). The other 31 markers contained sequences homologous to a variety of other transposable elements, including CRM1 (ten), Cinful-Zeon (seven), Xilon-Diguus (four), Sela (three), CRM3 (two), Doke (two), Huck (two), and one instance each of CRM4, Flip, Gyma, Puck, and Ji (Table 2.2). We also observed one instance of a repeating motif (TTTAGGG) that is observed in the B repeat (Alfenito and Birchler 1993).

Grouping markers within the B centromere using early misdivision derivatives

To make a genetic map of the B centromere, we analyzed a collection of lines with fragmented centromeres produced after centromere misdivision (Kaszas and Birchler 1998, 1996). All misdivision derivatives were originally derived from a translocation chromosome called TB-9Sb that involves the B chromosome and the short arm of chromosome 9. An unstable derivative of TB-9Sb called PI (for Pseudoisochromosome) was recovered by Carlson and Chou (1981). This chromosome contains two arms of chromosome 9 centered around a B centromere; however one arm contains a small knob that was derived from the B chromosome and the other arm does not, making it asymmetrical (Carlson and Chou 1981).

The misdivision derivatives derived from the PI chromosome are categorized as telocentric “Telo” or true isochromosomes “Iso” with identical arms (ring chromosomes were also recovered but are not studied here). In addition, the chromosomes are differentiated by the presence (+) or absence (-) of the small knob from the PI chromosome. We carried out CRM2-TD on the TB-9Sb progenitor and four early descendants: PI, Telo2-1(-), Iso3(-) and Telo2-2(-) (Figure 2.2). The observations are reported in Table 2.1. We found that five markers differentiate Tb-9Sb from PI, one marker differentiated PI and Telo2-1(-), one marker differentiated Telo2-1(-) and Iso3(-), and two markers differentiated Iso3(-) and Telo2-2, leaving 31 markers in the presumed core domain.

To confirm these patterns, simple PCR markers were developed to transposon junctions within seven key markers (Figure 2.4, Table 2.3). The presence or absence of the PCR bands confirmed the TD results in all cases. Using these PCR markers we also assayed two later-generation derivatives Telo4-11(-) and Telo3-3(-) (Kaszas and Birchler 1998), which were assayed in a prior fiber FISH study (Jin *et al.* 2005). The patterns of amplification indicated that Telo4-11(-) is as the same basic makeup as Telo2-2(-), and that Telo3-3 is missing all markers in the centromere core (Figure 2.4), consistent with the prior study (Jin *et al.* 2005). Markers with unique patterns were grouped together based on the patterns of centromere deletions (Figure 2.3, Table 2.1). The 31 markers of the presumed centromere core are likely to lie within or close to the 700 kb domain described by Jin et al (2005), but their relative arrangement and orientation within the core domain is not known.

Use of ChIP-seq to identify the centromere core on the B centromere minimal map

The screening method used to identify TD markers ensures that the combination of CRM2 insertion and BfaI restriction site are unique to the B chromosome. In addition, the internal sequences between the CRM2 and BfaI sites, while generally derived from transposable elements or other repeats, frequently have acquired polymorphism that makes them effectively unique over the length of a 150 bp Illumina sequence read. To build a map that could be used a reference sequence, we joined sequences from TD bands in the order shown in Table 1.1, with 101 N's separating each marker sequence (Figure 2.5, Suppl. Figure 2.1). The total amount of sequence in this pseudocontig, which we call the B centromere minimal map, is 10,100 bp.

Chromatin immunoprecipitation was carried out on TB-9Sb seedlings to identify sequences that interact with the centromeric histone variant CENH3. Nuclei from whole seedlings were isolated, MNase-digested and precipitated with maize CENH3 antibodies (Zhong *et al.* 2002). We also carried out ChIP on two control lines, the B73 reference inbred, and a line containing an inactive B centromere called 9-Bic-1 (Han, Lamb, and Birchler 2006). Both the TB-9Sb and 9-Bic-1 chromosomes had been backcrossed to the B73 inbred at least 5 times. Libraries were created from the precipitated DNA and sequenced using Illumina high-throughput sequencing. The reads were aligned to the B73 genome assembly (version 3) and the B centromere minimal map. Only unique hits were considered. The levels of CENH3 enrichment were calculated by comparing read depth over the centromere markers to the B73 whole-genome average.

Many of the B centromere markers showed significant enrichment for CENH3 in the TB9Sb sample by a first approximation (Figure 2.6). However, in some cases, the apparent enrichment for B centromere markers was also observed in B73 control sample that did not contain B chromosomes, indicating that the internal regions are not unique. In addition, the markers with the highest enrichment should be interpreted with caution, because although these regions are unique in comparison to the B73 genome, they may not be unique to the B chromosome (the B chromosome has long non-centromeric tracts of the B repeat, CentC and CRM (Lamb, Kato, and Birchler 2005)). At least five of the markers are clearly unique to the B centromere minimal map and showed uniform ChIP-seq coverage across the entire marker sequence including the CRM2 junction (TD9 10, 15, 18, and 21, Figure 2.6). These data confirm that the B centromere minimal map falls within the functional centromere core.

Discussion

Although the B chromosome is non-essential to maize, it has had a major role in many of the most important studies of plant centromeres (Birchler and Han 2009). Its dispensable nature has made the B centromere amenable to cytogenetic screens that would never be possible for a normal centromere, including the selection of deletion events through misdivision and the discovery of entirely inactive centromeres and their reactivated derivatives. These studies were facilitated by translocations between the B centromere and the short arm of chromosome 9 (such as TB-9Sb) that made it possible to identify centromere variants by simple visual screens. However the B chromosome has been left out of the recent surge in maize genomic studies. Sequences from small portions of the B chromosome have been published (Theuri *et al.* 2005, Cheng 2010) but there have been no systematic studies to identify sequences within the centromere core.

The most detailed study of the B centromere structure was a fiber-FISH analysis that identified a putative 700 kb centromere core based on a subset of centromere misdivision events. The putative centromere core contains five blocks of ZmBs-rich sequence that are flanked by regions of CRM-rich and CentC-rich sequence. The fiber-FISH map is a satisfying visual reference and has motivated multiple subsequent studies, however, it provides no sequence information for molecular analysis. Here we used transposon display to capture portions of the B centromere associated with CRM2 insertion points, sequenced, and mapped the markers relative to misdivision derivatives with various centromere sizes. The mapped and concatenated markers were then used to interpret CENH3 ChIP-seq results.

Two forms of data suggest that the B centromere minimal map traverses the functional centromere core. The first form of evidence is our ChIP-seq analyses, which demonstrate that markers of the minimal map interact with CENH3 and therefore identify the functional core (Figure 2.6). While a superficial analysis suggests that as many as half of the markers may interact with CENH3, for many of these we also observed apparent ChIP in the B73 sample and 9-Bic-1 samples, suggesting the sequences were not specific to the B chromosome. Minor differences in genetic background or differences with respect to the efficiencies of the ChIP experiments may be responsible for these observations. Nevertheless, using the most strict definition where no ChIP was observed for either of the B73 samples, no ChIP was observed for either of the 9-Bic-1 replicates, but ChIP reads covering the entire marker sequence were observed for all six TB-9Sb replicates, at least five markers (TD9 10, 15, 18, and 21) from the B centromere minimal map lie within the functional centromere core.

The second form of evidence that the B centromere minimal map identifies the functional centromere is a general concordance with the fiber-FISH data of Jin and coworkers (Jin *et al.* 2005). By definition, centromere misdivision derivatives must involve a breakage within the centromere core, as the mechanism requires that a single centromere erroneous divide into two parts such that the two parts separate to opposing poles (Yu and Dawe 2000). Fiber-FISH analysis of derivative Telo2-2(-) revealed a breakage that removed roughly one third of the 700 kb CentC-rich putative core domain (Jin *et al.* 2004). Similarly, we show here that the formation of Telo2-2(-) involved the removal of nine of the 40 TD markers, including TD9, which interacts with CENH3 (Figure 2.6). Jin and coworkers also studied Telo4-11(-), and showed that it had a nearly identical structure as Telo2-2(-). Using PCR markers developed from the TD markers, we confirmed that Telo4-4(-) and Telo2-2(-) have the same amplification pattern. The fact that at least five markers interact with CENH3, and that misdivision events involved breakages that separate markers of the core domain, demonstrate that our sequence identifies the functional maize B centromere.

Some early conclusions about the functional importance of the domains identified can also be made from our data. We can infer that the left-most markers are unlikely to be critical for centromere function, specifically the first seven markers that are absent in the PI derivative and Iso3(-) (Figure 2.6). Marker TD9 interacts with CENH3 and is presumably part of the B centromere core. Loss of this marker (and presumably other linked sequences) in Telo2-2(-) results in a reduction in transmission of the chromosome to 43% (from the 53% observed with PI). The ~400 kb section of the core domain that remains in Telo2-2(-) (Jin *et al.* 2004) presumably contains most of our remaining TD markers, and this region appears to be critical for centromere function. Derivatives with that show lower transmission than the 43% transmission are rare and tend to be unstable (Kaszas and Birchler 1998). The derivative Telo3-3(-) is the best studied of the small, unstable category of misdivision derivative (Jin *et al.* 2004). Importantly, we show here that Telo3-3(-) lacks the entire core domain as assayed by TD. Recent data corroborate this observation (Liu *et al.* 2015), showing that there are no detectable CRM elements in the Telo3-3(-) derivative. While a few small arrays of the B repeat are still present in Telo3-3(-) (Kaszas and Birchler 1998) and were originally thought to be responsible for the low transmission of this chromosome (Jin *et al.* 2004), Liu and coworkers have shown that the Telo3-3(-) chromosome has acquired a neocentromere in non-centromeric regions close to the original the B centromere (Liu *et al.* 2015). Taken together, the results indicate that the core domain identified by the B centromere minimal map is required for centromere function, and loss of this domain leads to centromere loss and the formation of new centromeres.

The availability of simple PCR markers (Table 2.3) and 10.1 kb of sequence data derived from the B centromere core (Suppl. Figure 1) will now make it possible to interpret other small unstable derivatives, and ultimately develop a full physical map of the B centromere for additional studies of centromere structure and function.

Materials and Methods

CRM-TD

To develop markers on the B chromosome, a modified amplified fragment length polymorphism (AFLP) approach was used to amplify centromere repeat using CRM2 LTR primers as described previously (Casa *et al.* 2000, Shi *et al.* 2010). Genomic DNA was digested with BfaI. For primary amplification step we used primer CRM2_R1 (5'-GAGGTGGTGTATCGGTTGCT) and BfaI + 0 (5'- GACGATGAGTCCTGAGTAG). For selective amplification we used P33- or FAM- labeled CRM2_R2 (5'-CTACAGCCTTCCAAAGACGC) and BfaI + 3 selective bases (5'-GACGATGAGTCCTGAGTAG + NNN). All 64 combinations of NNN were used as an attempt to identify all CRM2 elements in the B centromere. With this method, we compared a B73 line carrying the B chromosome (08109) to a B73 line (0878) without the B chromosome. We also analyzed lines containing TB9Sb, MiniB (#9), PI, Telo2-1(-), Iso3(-), Telo2-2(-) obtained from Jim Birchler.

Analysis of TD data

Many of the markers (all of those processed for sequencing) were analyzed on 6% polyacrylamide gels, blotted, and exposed to film (Shi *et al.* 2010). When using the FAM-labeled probe, samples mixed with a Geneflo 1000 size standard (ROX labeled) and genotyping at the

Georgia Genomics Facility. Genotyped results were analyzed using GeneMarker software (Holland and Parson 2011).

TD-marker sequencing

DNA from B-specific TD bands were extracted from polyacrylamide gels, amplified with pre-amplification primers and confirmed using 2.0% agarose gels. Only samples that showed a single band were further sequenced. PCR products were purified using a QIAGEN gel purification kit and were either directly sequenced or first cloned into a TOPO TA vector (Invitrogen, Carlsbad, CA) and Sanger sequenced. Two primers were used for the sequencing, a forward primer from the adapter sequence and a reverse primer for the ligated adapter, resulting in two separate sequences for each marker. Sequence alignment and quality was analyzed using Geneious software.

Sequence filtering for B specificity

After sequencing 250 TD bands, the number was further reduced based on number of duplicates. Blastn was used to identify and remove sequences that perfectly aligned to the maize B73 reference genome. Markers that were absent as perfect matches in the B73 reference were nevertheless annotated for their best match in NCBI to identify the repeats present.

Chromatin-immunoprecipitation and enrichment confirmation

A modified chromatin-immunoprecipitation (ChIP) protocol, developed by Zidian Xie of the Gernot Presting lab, was used on individual, 2 week old, maize seedlings (Li 2012). Briefly, the nuclei were isolated and chromatin digested with 12.5 U of MNase (Affymetrix, 70196Y) and enriched for mononucleosomes. For TB9Sb lines, enrichment for CENH3 associated chromatin used the antibody against CENH3 (Zhang *et al.* 2013). For 9-Bic-1 lines, the outer core of immature ears were digested and an antibody against the centromere specific histone

CENH3 was used to isolate centromere associated chromatin (Zhong 2002). CENH3-ChIP was performed on three TB9Sb lines and one 9-Bic-1 line, with two technical replicates for each sample. As controls for false positive enrichment, we utilized published B73 CENH3 ChIP and fragmented genomic sequences (Gent *et al.* 2014).

For enrichment confirmation of TB9Sb, qPCR primers for CRM/18s was used to ensure enrichment for each sample. Enrichment for all samples were calculated with Model-based Analysis of ChIP-Seq 2 (MACs 2) (Zhang *et al.* 2008, Feng *et al.* 2012). Enrichment was also calculated after CENH3 reads were mapped to fully sequenced centromeres 2 (Figure 2.6 A) and 5 (not shown).

Illumina library prep and sequencing

After ChIP, a library for Illumina sequencing was prepared by University of Missouri Core facility using a NEBNext Ultra Library Prep Kit (NEB, E7370L). All samples were sequenced on a single lane of 2x100bp on an Illumina HiSeq2000. The 9-Bic-1 sequence libraries were prepared without a kit but with TruSeq indexes and sequenced at Georgia Genomics Facility, with NextSeq500 High output flow, single-end 150-nucleotide length.

Sequence alignment and analysis protocol

Reads were processed with the following pipeline; quality trimmer: fastq_quality_trimmer -Q33 -t 20 -l 125; filtered by fastq_quality_filter -Q33 -q 20 -p 80; and the adapters removed by flexible adapter remover, /usr/local/far/latest/build/far -s -t -a --format fastq --trim-end right --adaptive-overlap yes --min-readlength 125.

Reads were then aligned to the maize genome with Burrows-Wheeler Aligner (BWA) (Li and Durbin 2009) with MEM alignment option at default except for seed alignment length (-k) increased to 40 bp. With this option, reads were first aligned with 40 bp of the read before checking the rest of the length. The output file of BWA mem was then filtered with Samtools view –bS –q 30 to filter out MAPQ scores lower than 30, and as a result aligned read had a 1% of being mapped incorrectly (Li *et al.* 2009). The Model-based Analysis of ChIP-Seq 2 (MACS2) (Feng *et al.* 2012) was used to calculate CENH3 enrichment relative to the genome average using the following parameters; macs2 callpeak -t -f BAM -g hs --outdir -B --SPMR --call-summits. Treat_pileup files were converted to bigwig files and displayed using Integrative Genome Viewer (Robinson *et al.* 2011).

References

Alfenito, M. R., and J. A. Birchler. 1993. "Molecular characterization of a maize B chromosome centric sequence." *Genetics* 135 (2):589-97.

Allshire, R. C., and G. H. Karpen. 2008. "Epigenetic regulation of centromeric chromatin: old dogs, new tricks?" *Nat Rev Genet* 9 (12):923-37. doi: 10.1038/nrg2466.

Ananiev, E.V., R.L. Phillips, and H.W. Rines. 1998c. "Chromosome-specific molecular organization of maize (*Zea mays* L.) centromeric regions." *Proc. Natl. Acad. Sci. USA* 95:13073-13078.

Bilinski, P., K. Distor, J. Gutierrez-Lopez, G. M. Mendoza, J. Shi, R. K. Dawe, and J. Ross-Ibarra. 2015. "Diversity and evolution of centromere repeats in the maize genome." *Chromosoma* 124 (1):57-65. doi: 10.1007/s00412-014-0483-8.

Birchler, J. A., and F. Han. 2009. "Maize centromeres: structure, function, epigenetics." *Annu Rev Genet* 43:287-303. doi: 10.1146/annurev-genet-102108-134834.

Black, B. E., and E. A. Bassett. 2008. "The histone variant CENP-A and centromere specification." *Curr Opin Cell Biol* 20 (1):91-100. doi: 10.1016/j.ceb.2007.11.007.

Carlson, W. R., and T. S. Chou. 1981. "B Chromosome Nondisjunction in Corn: Control by Factors near the Centromere." *Genetics* 97 (2):379-89.

Carlson, Wayne R. 1978. "The B chromosome of corn." *Annual review of genetics* 12 (1):5-5.

Casa, A. M., C. Brouwer, A. Nagel, L. Wang, Q. Zhang, S. Kresovich, and S. R. Wessler. 2000. "The MITE family heartbreaker (Hbr): molecular markers in maize." *Proc Natl Acad Sci U S A* 97 (18):10083-9.

Cheng, Y. M. 2010. "Evolution of the heterochromatic regions on maize B long arm based on the sequence structure of CL-repeat variants." *Chromosome Res* 18 (5):605-19. doi: 10.1007/s10577-010-9136-9.

Feng, J., T. Liu, B. Qin, Y. Zhang, and X. S. Liu. 2012. "Identifying ChIP-seq enrichment using MACS." *Nat Protoc* 7 (9):1728-40. doi: 10.1038/nprot.2012.101.

Gent, J. I., Y. Dong, J. Jiang, and R. K. Dawe. 2012. "Strong epigenetic similarity between maize centromeric and pericentromeric regions at the level of small RNAs, DNA methylation and H3 chromatin modifications." *Nucleic Acids Res* 40 (4):1550-60. doi: 10.1093/nar/gkr862.

Gent, J. I., T. F. Madzima, R. Bader, M. R. Kent, X. Zhang, M. Stam, K. M. McGinnis, and R. K. Dawe. 2014. "Accessible DNA and relative depletion of H3K9me2 at maize loci undergoing RNA-directed DNA methylation." *Plant Cell* 26 (12):4903-17. doi: 10.1105/tpc.114.130427.

Gent, J. I., K. Wang, J. Jiang, and R. K. Dawe. 2015. "Stable Patterns of CENH3 Occupancy Through Maize Lineages Containing Genetically Similar Centromeres." *Genetics*. doi: 10.1534/genetics.115.177360.

Han, F., Z. Gao, and J. A. Birchler. 2009. "Reactivation of an inactive centromere reveals epigenetic and structural components for centromere specification in maize." *Plant Cell* 21 (7):1929-39. doi: 10.1105/tpc.109.066662.

Han, F., J. C. Lamb, and J. A. Birchler. 2006. "High frequency of centromere inactivation resulting in stable dicentric chromosomes of maize." *Proc Natl Acad Sci U S A* 103 (9):3238-43. doi: 10.1073/pnas.0509650103.

Holland, Mitchell M, and Walther Parson. 2011. "GeneMarker® HID: A reliable software tool for the analysis of forensic STR data." *Journal of forensic sciences* 56 (1):29-35.

Jiang, J., J. A. Birchler, W. A. Parrott, and R. K. Dawe. 2003. "A molecular view of plant centromeres." *Trends Plant Sci* 8 (12):570-5.

Jin, W., J. C. Lamb, J. M. Vega, R. K. Dawe, J. A. Birchler, and J. Jiang. 2005. "Molecular and functional dissection of the maize B chromosome centromere." *Plant Cell* 17 (5):1412-23. doi: 10.1105/tpc.104.030643.

Jin, W., J. C. Lamb, W. Zhang, B. Kolano, J. A. Birchler, and J. Jiang. 2008. "Histone modifications associated with both A and B chromosomes of maize." *Chromosome Res* 16 (8):1203-14. doi: 10.1007/s10577-008-1269-8.

Jin, W., J. R. Melo, K. Nagaki, P. B. Talbert, S. Henikoff, R. K. Dawe, and J. Jiang. 2004. "Maize centromeres: organization and functional adaptation in the genetic background of oat." *Plant Cell* 16 (3):571-81. doi: 10.1105/tpc.018937.

Kaszas, E., and J. A. Birchler. 1996. "Misdivision analysis of centromere structure in maize." *EMBO J* 15 (19):5246-55.

Kaszas, E., and J. A. Birchler. 1998. "Meiotic transmission rates correlate with physical features of rearranged centromeres in maize." *Genetics* 150 (4):1683-92.

Lamb, J. C., A. Kato, and J. A. Birchler. 2005. "Sequences associated with A chromosome centromeres are present throughout the maize B chromosome." *Chromosoma* 113 (7):337-49. doi: 10.1007/s00412-004-0319-z.

Li, H., and R. Durbin. 2009. "Fast and accurate short read alignment with Burrows-Wheeler transform." *Bioinformatics* 25 (14):1754-60. doi: 10.1093/bioinformatics/btp324.

Li, Heng, Bob Handsaker, Alec Wysoker, Tim Fennell, Jue Ruan, Nils Homer, Gabor Marth, Goncalo Abecasis, and Richard Durbin. 2009. "The sequence alignment/map format and SAMtools." *Bioinformatics* 25 (16):2078-2079.

Li, X., Topp, C.N., Dawe, R.K. 2012. "Maize antibody procedures: Immunolocalization and chromatin immunoprecipitation." In *Plant Cytogenetics*, 271-286. New York: Springer.

Liu, Y., H. Su, J. Pang, Z. Gao, X. J. Wang, J. A. Birchler, and F. Han. 2015. "Sequential de novo centromere formation and inactivation on a chromosomal fragment in maize." *Proc Natl Acad Sci U S A* 112 (11):E1263-71. doi: 10.1073/pnas.1418248112.

Luce, A. C., A. Sharma, O. S. Mollere, T. K. Wolfgruber, K. Nagaki, J. Jiang, G. G. Presting, and R. K. Dawe. 2006. "Precise centromere mapping using a combination of repeat junction markers and chromatin immunoprecipitation-polymerase chain reaction." *Genetics* 174 (2):1057-61.

Nannas, N. J., and R. K. Dawe. 2015. "Genetic and genomic toolbox of Zea mays." *Genetics* 199 (3):655-69. doi: 10.1534/genetics.114.165183.

Phelps-Durr, T. L., and J. A. Birchler. 2004. "An asymptotic determination of minimum centromere size for the maize B chromosome." *Cytogenet Genome Res* 106 (2-4):309-13. doi: 10.1159/000079304.

Rhoades, MM, and Ellen Dempsey. 1972. "On the mechanism of chromatin loss induced by the B chromosome of maize." *Genetics* 71 (1):73-96.

Rhoades, MM, Ellen Dempsey, and Achille Ghidoni. 1967. "Chromosome elimination in maize induced by supernumerary B chromosomes." *Proceedings of the National Academy of Sciences of the United States of America* 57 (6):1626.

Robinson, James T, Helga Thorvaldsdóttir, Wendy Winckler, Mitchell Guttman, Eric S Lander, Gad Getz, and Jill P Mesirov. 2011. "Integrative genomics viewer." *Nature biotechnology* 29 (1):24-26.

Roman, Herschel. 1948. "Directed fertilization in maize." *Proceedings of the National Academy of Sciences of the United States of America* 34 (2):36.

Schnable, P. S., D. Ware, R. S. Fulton, J. C. Stein, F. Wei, S. Pasternak, C. Liang, J. Zhang, L. Fulton, T. A. Graves, P. Minx, A. D. Reily, L. Courtney, S. S. Kruchowski, C. Tomlinson, C. Strong, K. Delehaunty, C. Fronick, B. Courtney, S. M. Rock, E. Belter, F. Du, K. Kim, R. M. Abbott, M. Cotton, A. Levy, P. Marchetto, K. Ochoa, S. M. Jackson, B. Gillam, W. Chen, L. Yan, J. Higginbotham, M. Cardenas, J. Waligorski, E. Applebaum, L. Phelps, J. Falcone, K. Kanchi, T. Thane, A. Scimone, N. Thane, J. Henke, T. Wang, J. Ruppert, N. Shah, K. Rotter, J. Hodges, E. Ingenthron, M. Cordes, S. Kohlberg, J. Sgro, B. Delgado, K. Mead, A. Chinwalla, S. Leonard, K. Crouse, K. Collura, D. Kudrna, J. Currie, R. He, A. Angelova, S. Rajasekar, T. Mueller, R. Lomeli, G. Scara, A. Ko, K. Delaney, M. Wissotski, G. Lopez, D. Campos, M. Braidotti, E. Ashley, W. Golser, H. Kim, S. Lee, J. Lin, Z. Dujmic, W. Kim, J. Talag, A. Zuccolo, C. Fan, A. Sebastian, M. Kramer, L. Spiegel, L. Nascimento, T. Zutavern, B. Miller, C. Ambroise, S. Muller, W. Spooner, A. Narechania, L. Ren, S. Wei, S. Kumari, B. Faga, M. J. Levy, L. McMahan, P. Van Buren, M. W. Vaughn, K. Ying, C. T. Yeh, S. J. Emrich, Y. Jia, A. Kalyanaraman, A. P. Hsia, W. B. Barbazuk, R. S. Baucom, T. P. Brutnell, N. C. Carpita, C. Chaparro, J. M. Chia, J. M. Deragon, J. C. Estill, Y. Fu, J. A. Jeddeloh, Y. Han, H. Lee, P. Li, D. R. Lisch, S. Liu, Z. Liu, D. H. Nagel, M. C. McCann, P. SanMiguel, A. M. Myers, D. Nettleton, J. Nguyen, B. W. Penning, L. Ponnala, K. L.

Schneider, D. C. Schwartz, A. Sharma, C. Soderlund, N. M. Springer, Q. Sun, H. Wang, M. Waterman, R. Westerman, T. K. Wolfgruber, L. Yang, Y. Yu, L. Zhang, S. Zhou, Q. Zhu, J. L. Bennetzen, R. K. Dawe, J. Jiang, N. Jiang, G. G. Presting, S. R. Wessler, S. Aluru, R. A. Martienssen, S. W. Clifton, W. R. McCombie, R. A. Wing, and R. K. Wilson. 2009. "The B73 maize genome: complexity, diversity, and dynamics." *Science* 326 (5956):1112-5. doi: 10.1126/science.1178534.

Sharma, A., and G. G. Presting. 2008. "Centromeric retrotransposon lineages predate the maize/rice divergence and differ in abundance and activity." *Mol Genet Genomics* 279 (2):133-47.

Shi, J., S. E. Wolf, J. M. Burke, G. G. Presting, J. Ross-Ibarra, and R. K. Dawe. 2010. "Widespread gene conversion in centromere cores." *PLoS Biol* 8 (3):e1000327. doi: 10.1371/journal.pbio.1000327.

Theuri, J., T. Phelps-Durr, S. Mathews, and J. Birchler. 2005. "A comparative study of retrotransposons in the centromeric regions of A and B chromosomes of maize." *Cytogenet Genome Res* 110 (1-4):203-8. doi: 10.1159/000084953.

Wang, K., Y. Wu, W. Zhang, R. K. Dawe, and J. Jiang. 2014. "Maize centromeres expand and adopt a uniform size in the genetic background of oat." *Genome Res* 24 (1):107-16. doi: 10.1101/gr.160887.113.

Wolfgruber, T. K., A. Sharma, K. L. Schneider, P. S. Albert, D. H. Koo, J. H. Shi, Z. Gao, F. P. Han, H. Lee, R. H. Xu, J. Allison, J. A. Birchler, J. M. Jiang, R. K. Dawe, and G. G. Presting. 2009. "Maize Centromere Structure and Evolution: Sequence Analysis of Centromeres 2 and 5 Reveals Dynamic Loci Shaped Primarily by Retrotransposons." *Plos Genetics* 5 (11). doi: Artn E1000743

Doi 10.1371/Journal.Pgen.1000743.

You, F. M., H. Wanjugi, N. Huo, G. R. Lazo, M. C. Luo, O. D. Anderson, J. Dvorak, and Y. Q. Gu. 2010. "RJPrimers: unique transposable element insertion junction discovery and PCR primer design for marker development." *Nucleic Acids Res* 38 (Web Server issue):W313-20. doi: 10.1093/nar/gkq425.

Yu, H.-G., and R.K. Dawe. 2000. "Functional redundancy in the maize meiotic kinetochore." *J. Cell Biol.* 151:131-141.

Yu, W., F. Han, Z. Gao, J. M. Vega, and J. A. Birchler. 2007. "Construction and behavior of engineered minichromosomes in maize." *Proc Natl Acad Sci U S A* 104 (21):8924-9. doi: 10.1073/pnas.0700932104.

Zhang, B., Z. Lv, J. Pang, Y. Liu, X. Guo, S. Fu, J. Li, Q. Dong, H. J. Wu, Z. Gao, X. J. Wang, and F. Han. 2013. "Formation of a functional maize centromere after loss of centromeric sequences and gain of ectopic sequences." *Plant Cell* 25 (6):1979-89. doi: 10.1105/tpc.113.110015.

Zhang, Yong, Tao Liu, Clifford A Meyer, Jérôme Eeckhoute, David S Johnson, Bradley E Bernstein, Chad Nusbaum, Richard M Myers, Myles Brown, and Wei Li. 2008. "Model-based analysis of ChIP-Seq (MACS)." *Genome biology* 9 (9):R137.

Zhong, C. X. 2002. "Centromeric Retroelements and Satellites Interact with Maize Kinetochore Protein CENH3." *The Plant Cell Online* 14 (11):2825-2836. doi: 10.1105/tpc.006106.

Zhong, C. X., J. B. Marshall, C. Topp, R. Mroczek, A. Kato, K. Nagaki, J. A. Birchler, J. Jiang, and R. K. Dawe. 2002. "Centromeric retroelements and satellites interact with maize kinetochore protein CENH3." *Plant Cell* 14 (11):2825-36.

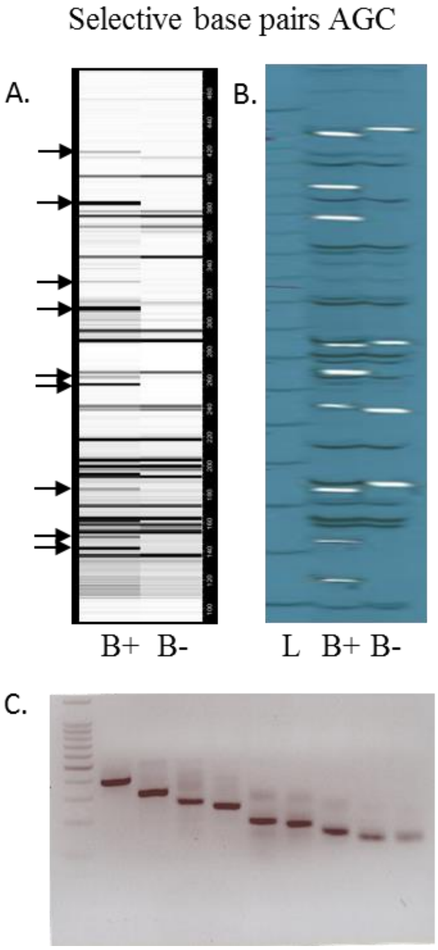


Figure 2.1. Both radio- and fluorescence-labeled transposon display of CRM2. These data were acquired using the selective bases AGC on the adapter primer. Lanes are labeled with B+ (carrying multiple copies of the B chromosome), B- (control line without B chromosomes), and L (ladder). **A**. Fluorescence-labeled TD. Digital data were converted to a pseudogel format using GeneMarker software. Arrows indicate bands that were chosen sequencing. **B**. The same adapter primers used in A were p33-labeled for TD, and the results separated by PAGE and exposed to film. The film was placed back onto the gel, and a blade used to cut through the film and gel; the gel slice was then used for re-amplification PCR. **C**. PCR re-amplification for sequencing of the extracted bands.

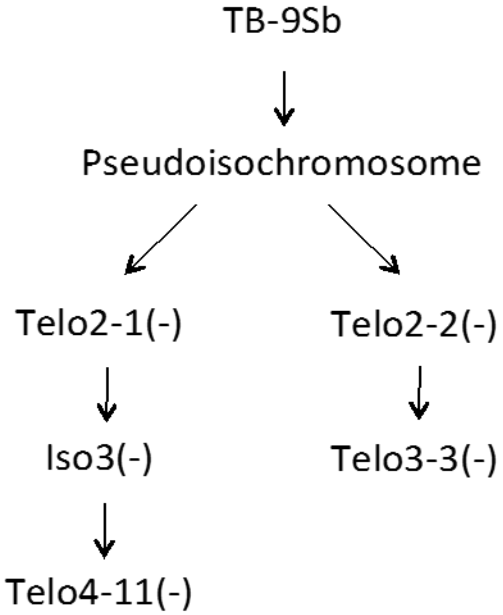


Figure 2.2. Pedigree of the misdivision lines used here. All derivatives were derived from TB-9Sb, which gave rise to the Pseudoisochromosome (PI). Two second-generation derivatives (Telo2-1(-) and Telo2-2(-)) were derived from PI. We also studied two third generation derivatives (Iso3(-) and Telo3-3(-)), and one fourth generation derivative (Telo4-11(-)).

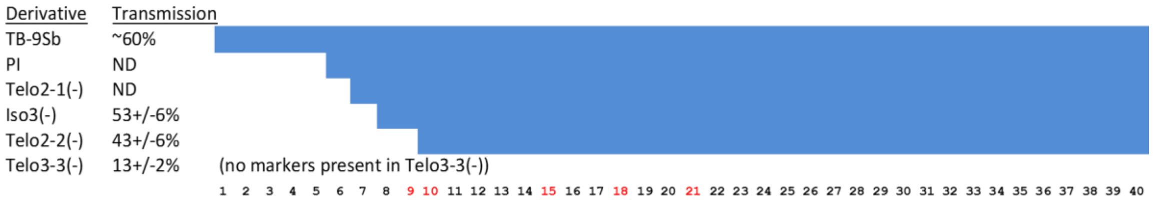
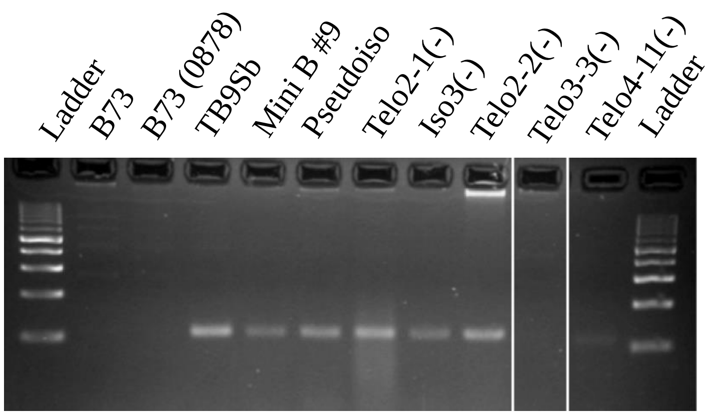


Figure 2.3. Transmission frequencies and relative sizes of B centromere misdivision derivatives. The derivative and its transmission frequency are shown to the left (Kaszas and Birchler 1998). No transmission data (ND) are available for Telo2-1(-) and Iso3(-). Markers in the B centromere minimal map are shown below; those with strong CENH3 ChIP are highlighted in red.



Primer pair	TD number	B73	TB9Sb	PI	Telo2-1(-)	Iso3(-)	Telo2-2(-)	Telo3-3(-)	Telo4-11(-)
CRM2-ATA-CCA-342	TD7	-	+	+	+	-	-	-	-
CRM2-TCC-310	TD8	-	+	+	+	+	-	-	-
CRM2-AGC-GGC-GAC-382	TD17	-	+	+	+	+	+	-	+
CRM2-214	TD10	-	+	+	+	+	+	-	+
CRM2-CGG-326	TD15	-	+	+	+	+	+	-	+
CRM2-ACC-ATG-CTG-351	TD21	-	+	+	+	+	+	-	+
CRM2-AGG-328	TD40	-	+	+	+	+	+	-	+

Figure 2.4. PCR confirmation of TD marker patterns. Upper image shows the amplification of CRM2-AGG-328 (TD40) in the lines studied here. All lanes were from the same gel (regions where lanes were removed are indicated by white lines). Lower table shows the amplification of the seven simple PCR markers and confirmation of the TD patterns. Telo3-3(-) and Telo4-11(-) were not assayed by TD and only assayed using simple PCR markers.

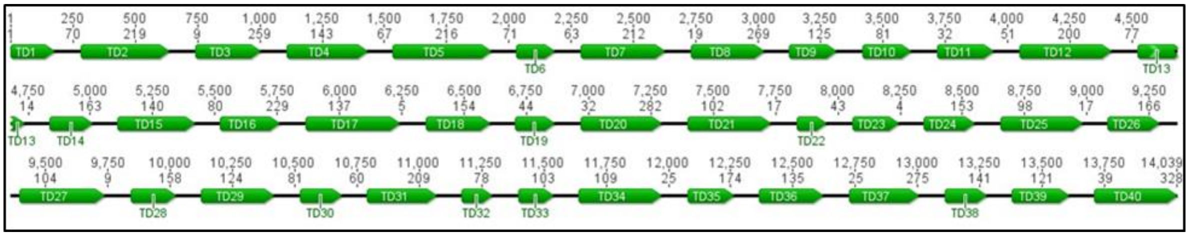


Figure 2.5. Diagram of the B centromere minimal map. The minimal map pseudocontig contains 40 CRM2-TD markers each separated by 101 N base pairs (black lines).

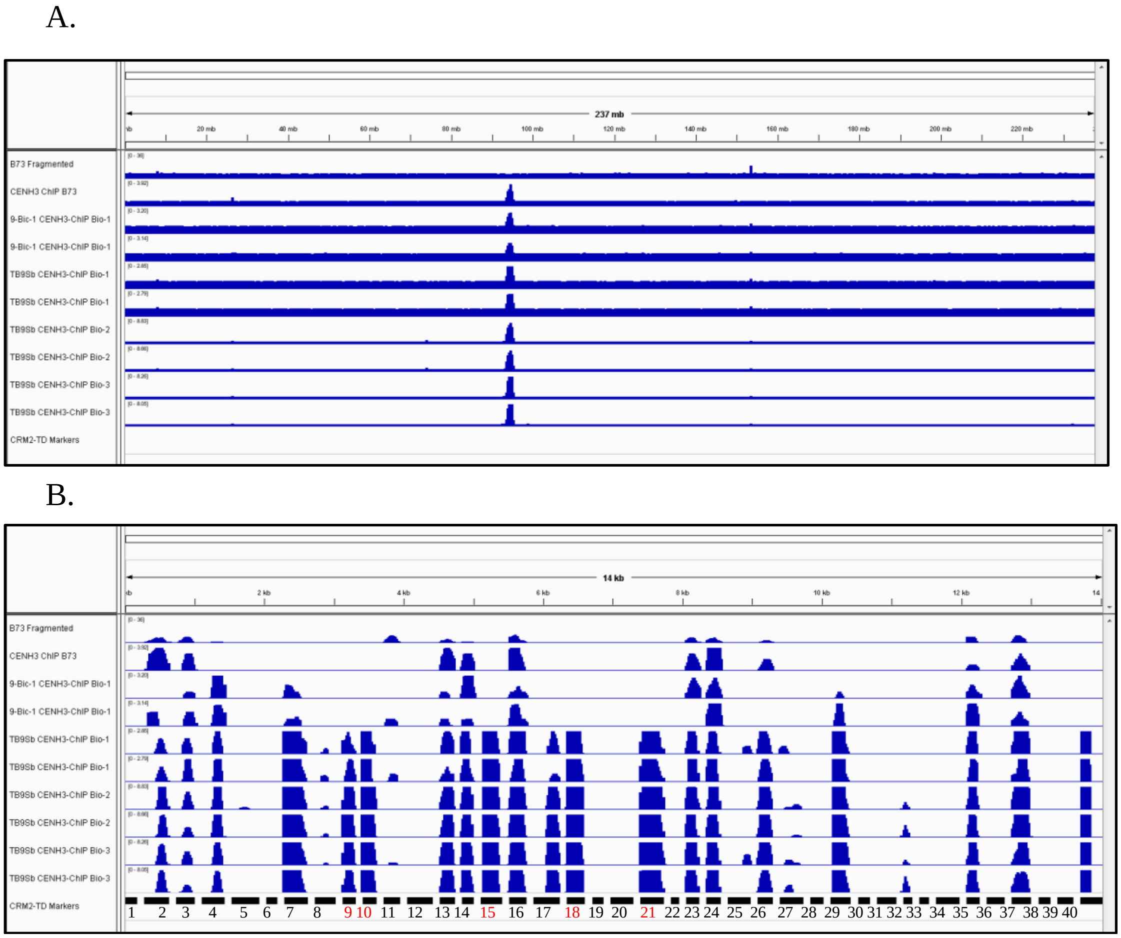


Figure 2.1. CENH3 ChIP-seq reads aligned to the B73 genome and B centromere minimal map. Graphs are obtained from IGV (Robinson *et al.* 2012). In both panels, the first bar shows an alignment of fragmented B73 DNA as a negative control. The second row shows the results of ChIP-seq using B73 tissue, which does not contain B chromosomes. The third and fourth rows show the results from two technical replicates from a plant carrying 9-Bic-1 (a B chromosome variant with an inactive centromere). The last six rows are TB9Sb lines, including three biological replicates with two technical replicates each A. Alignment to B73 chromosome 2 showing the centromere position as a single sharp peak. B. Alignment to the B centromere minimal map with each marker shown and numbered below. Enrichment values varied dramatically among markers. In order to visibly display all of the data, the full height of many of the peaks are not shown. Five markers that show uniform enrichment over the entire marker sequence specifically in the TB-9Sb lines are highlighted in red.

Table 2.1. CRM2-TD markers mapped to misdivision derivatives. The marker names include all the selective base pairs that amplified the sequence. TD indicates short pseudonym for each marker. The B+ and B- lines were paired samples obtained from James Birchler. The presence or absence of the band is noted by a "+" or "-" in the body of the table. As noted*, TD7, TD8, TD10, TD15, TD21, and TD40 were also converted to simple PCR markers.

CRM2 TD Markers	TD	B73	B+	B-	TB9Sb	PI	Telo2-1(-)	Iso3(-)	Telo2-2(-)
CRM2-ACC-CCA-AGA-ATA-196	TD1	-	+	-	+	-	-	-	-
CRM2-CCA-375	TD2	-	+	-	+	-	-	-	-
CRM2-CCC-280	TD3	-	+	-	+	-	-	-	-
CRM2-CCC-TCC-342	TD4	-	+	-	+	-	-	-	-
CRM2-GGA-GAC-405	TD5	-	+	-	+	-	-	-	-
CRM2-TAT-GAT-CTT-173	TD6	-	+	-	+	+	-	-	-
CRM2-ATA-CCA-342*	TD7	-	+	-	+	+	+	-	-
CRM2-TCC-310*	TD8	-	+	-	+	+	+	+	-
CRM2-TCT-209	TD9	-	+	-	+	+	+	+	-
CRM2-AGA-214*	TD10	-	+	-	+	+	+	+	+
CRM2-CTA-CTG-GTA-247	TD11	-	+	-	+	+	+	+	+
CRM2-CTG-389	TD12	-	+	-	+	+	+	+	+
CRM2-GAT-228	TD13	-	+	-	+	+	+	+	+
CRM2-GTC-188	TD14	-	+	-	+	+	+	+	+
CRM2-CGG-326*	TD15	-	+	-	+	+	+	+	+
CRM2-AGC-GGC-257	TD16	-	+	-	+	+	+	+	+
CRM2-AGC-GGC-GAC-382*	TD17	-	+	-	+	+	+	+	+
CRM2-CTT-256	TD18	-	+	-	+	+	+	+	+
CRM2-TAT-161	TD19	-	+	-	+	+	+	+	+
CRM2-TCC-345	TD20	-	+	-	+	+	+	+	+
CRM2-ACC-ATG-CTG-351*	TD21	-	+	-	+	+	+	+	+
CRM2-AGA-138	TD22	-	+	-	+	+	+	+	+
CRM2-AGA-AGG-GGA-TCG-207	TD23	-	+	-	+	+	+	+	+
CRM2-ATC-TCA-220	TD24	-	+	-	+	+	+	+	+
CRM2-CAG-335	TD25	-	+	-	+	+	+	+	+
CRM2-CAT-AGT-227	TD26	-	+	-	+	+	+	+	+
CRM2-CGA-361	TD27	-	+	-	+	+	+	+	+
CRM2-CTC-199	TD28	-	+	-	+	+	+	+	+
CRM2-GCC-TTC-CCC-TAC-CCT-309	TD29	-	+	-	+	+	+	+	+
CRM2-GGG-186	TD30	-	+	-	+	+	+	+	+
CRM2-GGG-298	TD31	-	+	-	+	+	+	+	+
CRM2-TAC-140	TD32	-	+	-	+	+	+	+	+
CRM2-TAT-CAT-159	TD33	-	+	-	+	+	+	+	+
CRM2-TCC-350	TD34	-	+	-	+	+	+	+	+
CRM2-AAT-GAT-208	TD35	-	+	-	+	+	+	+	+
CRM2-AGC-CAG-CCG-275	TD36	-	+	-	+	+	+	+	+
CRM2-ATG-298	TD37	-	+	-	+	+	+	+	+
CRM2-GAC-AGC-AGT-CGT-ATT-AAT-185	TD38	-	+	-	+	+	+	+	+
CRM2-GTC-247	TD39	-	+	-	+	+	+	+	+
CRM2-AGG-328*	TD40	-	+	-	+	+	+	+	+

*Sequenced marker also converted to simple PCR marker

Table 2.1. CENH3 fold enrichment and sequence homology of the TD markers. CENH3 enrichment shows raw values calculated as the read depth in the TB-9Sb lines compared to the B73 control line. Identity of the CRM2-flanking sequence was interpreted using Blastn. All of the markers contain sequences homologous to known retroelements. Some markers also show homology to known centromere repeats.

Marker	CENH3 Enrichment	Retroelement homology	Other Repeat homology
TD1		Huck (RLG)	
TD2		CRM1 (RLX)	
TD3		CRM1 (RLX)	
TD3		CRM4 (RLX)	
TD4		Huck (RLG)	
TD5	0.9	Xilon-Diguus (RLG)	
TD6	16.8	CRM1 (RLX)	
TD7	62.0	CRM2 (RLX)	B repeat
TD8	3.8	Sela (RLX)	B repeat
TD9	15.1	CRM2 (RLX)	Centc
TD10	31.9	TTTAGGG	B repeat
TD11	7.3	Ji (RLC)	
TD12		CRM1 (RLX)	
TD13		Xilon-Diguus (RLG)	
TD14		CRM3 (RLX)	
TD15	16.1	CRM2 (RLX)	B repeat
TD16		CRM1 (RLX)	
TD16		CRM1 (RLX)	
TD16		CRM1 (RLX)	
TD17	13.4	CRM1 (RLX)	
TD18	40.9	CRM2 (RLX)	CentC
TD19	2.8	CRM2 (RLX)	
TD20	2.4	Flip (RLG)	
TD21	12.1	CRM2 (RLX)	CentC
TD23		CRM1 (RLX)	
TD24		Xilon-Diguus (RLG)	
TD25		Doke (RLG)	
TD26	0.1	Puck (RLG)	
TD27	1.5	Doke (RLG)	
TD28	9.5	Xilon-Diguus (RLG)	
TD29	53.0	Sela (RLX)	
TD29		Sela (RLX)	B repeat
TD30	7.1	CRM1 (RLX)	
TD31	0.1	Gyma (RLG)	
TD35	2.1	CRM3 (RLX)	
TD36		Cinful-Zeon (RLG)	
TD36		Cinful-Zeon (RLG)	
TD36		Cinful-Zeon (RLG)	
TD37		Cinful-Zeon (RLG)	
TD38		Cinful-Zeon (RLG)	
TD38		Cinful-Zeon (RLG)	
TD39	1.8	Cinful-Zeon (RLG)	
TD40	5.2	CRM2 (RLX)	CentC

Table 2.3. Primers used to confirm TD results.

TD marker	TD Name	Primer Name	Sequence
CRM2-ATA-CCA-342	TD7	CRM2-ATA-CCA-342-6-F	AAACGCTATAGGACAGGCCC
		CRM2-ATA-CCA-342-6-R	TCTTTGGAGGCTGTAGTCGG
CRM2-TCC-310	TD8	CRM2-TCC310-5-F	CATAAACCCTAAAGCCCAAACC
		CRM2-TCC310-5-R	TCTTTGGAAGGCTGTAGTCGG
CRM2-214	TD10	CRM2-AGA-214-5-F	CGGGTGCACATCAACTAACC
		CRM2-AGA-214-5-R	GAGTTTGGGTTTTTGGATTTATGG
CRM2-CGG-326	TD15	CRM2-CGG-326-4-F	GGGTGCACATCAAGAACCAT
		CRM2-CGG-326-4-R	CGAAAACACCCCAAAGATGA
CRM2-AGC-GGC-GAC-382	TD17	CRM2-AGC-GGC-GAC-382-1-F	CCAACGGGTGCACATCAC
		CRM2-AGC-GGC-GAC-382-1-R	CCCCCTGCTGTTGTTAACCT
CRM2-ACC-ATG-CTG-351	TD21	CRM2-ACC-ATG-CTG-351-6_F1	CTAGTCGATTCGGCATGTTCGTTGCG
		CRM2-ACC-ATG-CTG-351-6_R2	GGTGCACATCATTTCGCGCAATTCAG
CRM2-AGG-328	TD40	CRM2-AGG-328-4_F3	CGGTAACGTACGGCAACG
		CRM2-AGG-328-4_R2	CATCAAGAACCATTTCTACGTTTATCG

CHAPTER 3

DO MINIATURE INVERTED-REPEAT TRANSPOSABLE ELEMENTS (MITES) ACT AS BOUNDARY ELEMENTS FOR GENES?

Abstract

The maize genome is robust and primarily composed of heavily methylated heterochromatic transposable elements (Schnable *et al.*, 2009). There are two main types of transposons, retrotransposons (class I) and DNA transposons (class II). Both types are targeted by various mechanisms to ensure their silenced state. Alternately, the majority of maize genes occur in gene islands that are flanked by heterochromatic transposons. These gene islands often maintain a functional state with a chromatin make up to allow gene expression. Recent studies categorized retrotransposons based on their propensity to spread heterochromatin into flanking genic regions and found that in general genes nearby have lower expression levels (Eichten *et al.*, 2012). Unlike retrotransposons, class II MITE transposons are associated with genes categorized as highly expressed in the grasses (Gent *et al.*, 2013; Han, Qin, & Wessler, 2013), though the MITEs themselves are maintained in a silenced state. MITEs have a propensity to target flanking regions of genes which often result in change in gene expression (Naito *et al.*, 2006). It is not yet understood why MITEs are associated with highly expressed genes. In this experiment spreading heterochromatin TEs were used to test whether MITEs act as boundary elements for genes, and genes were high in expression if a MITE present, and had lower methylation levels. Further, MITEs that show severe boundary effect also have high A/T% content and low CG% motifs in their sequence.

Introduction

Transposable elements (TE) make up a large fraction of many plant and animal genomes (Copenhaver *et al.*, 2011; Schnable *et al.*, 2009). About 45% of the human genome is made up of transposable elements and in plants, such as maize, the proportion can be as high as 85% (Cordaux & Batzer, 2009; Schnable *et al.*, 2009). Active transposons have the potential to result in major genomic changes that affect genome size, genome rearrangements and cause gene mutation (Kumar & Bennetzen, 1999). Multiple mechanisms have evolved to silence and reduce the spread of transposons, including DNA methylation and histone modifications that are associated with heterochromatic chromatin regions (Lisch, 2009; Richards, 2006; Weil & Martienssen, 2008). Genes are also marked by histone modifications, such as those to ensure constitutive gene expression or to direct expression levels spatially or temporally (Cazzonelli *et al.*, 2009; Ha, Ng, Li, & Chen, 2011). Here we investigated the murky interface between inactivated transposons targeted by silencing mechanisms and nearby genes often targeted by machinery promoting expression, and found that small class II transposons called Miniature Inverted-Repeat Transposable Elements or MITEs have special properties that may allow them to act as boundary elements that separate inactive heterochromatic regions from active genes.

Traditionally, as a means to categorize the genome as either heterochromatic (inaccessible chromatin) or euchromatic (accessible chromatin), carmine acidic acid was used to identify densely stained chromatin defined as heterochromatin (Heitz, 1928). Euchromatin was defined as chromatin that underwent decondensation during the cell cycle. In maize and many other large-genome grasses, heterochromatin is found in vast transposon-rich areas flanking centromeres, whereas euchromatin is found on gene-rich chromosome arms (Shi & Dawe, 2006). And while DNA methylation is often associated with silencing, recent studies have found that methylation plays many roles in gene regulation, such as hyper- or hypomethylation at promoter regions, gene bodies, and start sites (Jones, 2012). Even in the euchromatic regions of maize, genes are generally found in islands, surrounded on both sides by large expanses of transposon-derived repetitive DNA (R. Liu *et al.*, 2007). This arrangement creates heterochromatin-euchromatin transition zones at nearly every gene. These transition zones give the heterochromatin the potential to significantly impact gene expression (Gent *et al.*, 2013; Gent *et al.*, 2014; Veiseth *et al.*, 2011).

In mammals, DNA methylation is primarily at CG sites and essential for proper gene expression (Li, Bestor, & Jaenisch, 1992) whereas in plants, DNA methylation is present in all sequence contexts and functions to suppress transposons and to regulate endogenous genes (X. Zhang & Jacobsen, 2006). Plants have developed a *de novo* DNA methylation mechanism which occurs at all cytosines (CG, CHG, and CHH, H is any nucleotide but G), and is controlled by a number of pathways, including RNA-dependent DNA methylation (RdDM) (Law & Jacobsen, 2010). DNA methylation in the CG and CHG form are maintained by a group of methyltransferases that target hemi-methylated DNA after replication (Goll & Bestor, 2005). The distribution of CHG methylation, for example, correlates with the methylation of histone H3 at lysine 9 (H3K9me). Specialized pathways exist for the maintenance of methylation at CG and CHG sites, but CHH methylation occurs primarily through the RdDM pathway and is *de novo* methylated at each cell cycle (Feng *et al.*, 2010). Over one third of the genes in *Arabidopsis* maintain expression while being methylated in their transcribed regions (X. Zhang & Jacobsen, 2006), and moderately expressed genes are more likely to be methylated than highly expressed genes. In general methylation is biased away from the ends of genes, being found primarily in the gene body (Zilberman, Gehring, Tran, Ballinger, & Henikoff, 2007). DNA methylation and gene expression is not fully understood and more research is needed to fully understand the relationship (Jones, 2012).

Transposable elements are divided into two main classes. Class I, which transpose via an RNA intermediate (“copy and paste”), and make up 75% of the maize genome. The mechanisms that control and inactivate class I transposable elements involve histone modifications and heavy DNA methylation that have strong negative effects on both transposon and gene expression (Hollister & Gaut, 2009; Johnson *et al.*, 2007; Kato, Miura, Bender, Jacobsen, & Kakutani, 2003). In maize, retrotransposons are more likely to be found in deep intergenic heterochromatic regions between genes, or pericentromeric regions. In *Arabidopsis thaliana*, methylated class I elements near genes are negatively correlated with gene expression (Hollister & Gaut, 2009; Pereira, Enard, & Eyre-Walker, 2009), and there is evidence that methylated TEs near genes are selected against (Hollister & Gaut, 2009). Recent work indicates that heterochromatic marks, such as cytosine methylation (5-mC) or H3K9me2, marking class I transposons spread into flanking genic regions, and are linked to genes with lower expression (Eichten *et al.*, 2012).

The other category of TE, class II elements, have more than one way to transpose, but traditionally are defined by a DNA intermediate (“cut and paste”) method, and are more abundant near genes. They were discovered and described as controlling elements by Barbara McClintock due to their ability to affect endogenous genes (McClintock, 1984). DNA transposons can occur as non-autonomous element, in which the DNA transposon has lost the ability to transpose itself. One category of non-autonomous DNA transposons is known as miniature inverted-repeat transposable elements (MITEs). MITEs are generally categorized into five different families based on their shared structural features, such as their TIR sequence and their target site duplication (TSD): *CACTA, Mutator, PIF/Harbinger, hAT,* and *TC1/mariner*

(Feschotte & Pritham, 2007; Yuan & Wessler, 2011). MITEs may sometimes insert directly into coding regions to cause mutant phenotypes (S. Liu *et al.*, 2009), however they are more frequently found at the edges of genes or within introns (Oki *et al.*, 2008).

Analyses of MITEs across the cereals have revealed differences and similarities between the superfamilies (Han *et al.*, 2013). The *Tc1/mariner* and *PIF/Harbinger* superfamilies tend to have positive effects on nearby gene expression. However, *CACTA* elements, which are generally found furthest from genes, tend to negatively affect gene expression when close to a gene (Han *et al.*, 2013). MITEs are methylated like class I elements, but particularly when they lie at the edges of genes, have exceptionally high levels of CHH methylation. These domains have been called “CHH islands” (Gent *et al.*, 2013). On average, CHH islands are associated with increased levels of gene expression. The observations that DNA methylation of class I elements tend to have suppressive effects when in close proximity to genes, whereas the unusual CHH methylation associated with MITEs is correlated with increased gene expression, appear to be in conflict. This probably reflects the complex interactions that occur when two forms of chromatin meet or overlap (Xiaoyu Zhang, 2008). Here we investigate chromatin states of genic boundaries, focusing on changes in DNA methylation, gene expression and histone modifications as it relates to MITE content.

Results

Overall distribution of DNA methylation and transposons in the maize genome

To understand the general trends of methylation in maize, we calculated genome distributions of DNA methylation along meta-chromosome plots with average relative methylation calculated across the length of all 10 maize chromosomes. While CG and CHG methylation followed the trend of transposon enrichment (Figure 3.1 A, left y-axis), CHH methylation was enriched on the arms of chromosomes, similar to genes (Figure 3.1 A, right y-axis). To test whether internal methylation of transposons varied by their distance from the centromere, we calculated relative methylation for class I and class II transposons along the length of all 10 chromosomes in Figure 3.1B and1C. The centromere in Figure 3.1A is depicted by a black dot in the middle of the x-axis and represented by a 0 in Figure3.1B and 3.1C. Class I internal methylation is relatively consistent for CG methylation and CHG methylation, regardless of location on chromosome, while class II decreases towards the ends of the chromosomes (Figure 3.1B). Class I CHH methylation increased on the chromosome arms (Figure 3.1C), but not much higher than the genome average of 5% (not shown). Class II elements generally have higher CHH methylation than class I, and this CHH methylation substantially increases on the chromosome arms (Figure 3.1C).

DNA methylation differences among transposon types

Retrotransposons can be divided into four superfamilies; copia (RLC), gypsy (RLG), LINEs (RIL), and unknown (RLX). The methylation levels for CG, CHG and CHH class I superfamilies are at 80%, 65% and 4% respectively, and there is quite a bit of variation within each superfamily (Figure 3.2A). In comparison, class II superfamilies have on average higher CG, CHG, and CHH methylation (Figure 3.2B). In general when comparing class II superfamilies, the lower the CG methylation for each superfamily, the higher its CHH methylation (Figure 3.2B). For instance, *CACTA* has the highest CG methylation at around 90% and CHH methylation below 10%, while *Tc1-Mariner* has 79% CG methylation and 20% CHH methylation (Figure 3.2B).

These data correlated with the average distance a MITE is found from a gene. *CACTA* are found furthest from genes in comparison to all other superfamilies (Figure A, B) while *Tc1-*

Mariner or Stowaway tend to be closest (Figure 3.3A, B). However, all MITE superfamilies are highly enriched on the arms of chromosomes (Figure 3.3A).

A subset of class I transposons cause spreading of DNA methylation into flanking sequences

Retrotransposons in the maize genome were divided into three categories, those with evidence of spreading of both DNA and histone methylation ("5mc/H3K9me2 group") those that showed spreading of H3K9me2 but not DNA methylation ("H3K9me2 group"), and those that showed no evidence of spreading ("Non spreading")(Eichten *et al.*, 2012). The two spreading groups were only a fraction of the total number of retrotransposon families, but make up over 50% of the maize genome. In general, the closer a spreading transposons is found to a gene, the lower that gene's expression (Eichten *et al.*, 2012). In comparison to the genome average of superfamilies in Figure 3.2A, both the 5mc/H3K9me2 and H3K9me2 groups had higher internal CG and CHG methylation at around 90% and 80% respectively. The non-spreading group on the other hand had slightly lower CG methylation than the spreading groups at 88%. The 5mC/H3K9me2 group had significantly higher CG and CHG methylation than non-spreading (Figure 3.2C). The H3K9me2 only group did not have significantly higher CG methylation than non-spreading but did show significantly higher CHG methylation than non-spreading group. Interestingly, the non-spreading group had internal CHH methylation levels significantly higher than other two spreading groups at around 9% (Figure 3.2 C).

Next we looked at the 1kb upstream and downstream flanking regions of each defined spreading or non-spreading group. The relative methylation of CG, CHG and CHH was calculated across a 1kb region. The flanking of transposons was only included if another transposable element was not present and called "low-copy" (see methods). The methylation of flanking sequences was lowest for the non-spreading group at 58% CG and 34% CHG methylation. CHH methylation was low at around 5% in all groups.

Distribution of the transposons in maize genome

We also looked at the genomic distribution of the three categories of transposons, anticipating that spreading transposons may be preferentially distributed in pericentromeric areas where there are fewer genes, and their propensity to affect the methylation of flanking sequences would have less of an impact. Transposons were mapped by their physical position in the genomes, defined by Repeatmasker, across a meta-chromosome plot that averages all 10 chromosomes (Smit, Hubley, & Green, 1996). We found that the 5mc/H3K9 group was enriched in the pericentromere overall (Figure 3.4 A). In contrast, the H3K9 only and non-spreading transposons are primarily found on chromosome arms (Figure 3.4 A). Both spreading groups were on average further from genes, when comparing average base pair distance from nearest transcription start site, compared to non-spreading (Figure 3.4 B). All three groups were further from genes in comparison to MITEs except for non-spreading which were closer to genes than *CACTA* elements (Figure 3.4 B; Figure 3.3A). These data suggest that the transposons which cause spreading of DNA methylation are adapted to regions with few genes.

Similar to the previous meta-chromosomal plots, we compared the three spreading type transposon group's internal methylation across all 10 chromosomes. Similar to all class I elements in Figure 3.1A, CG methylation was consistent across the chromosome (Figure 3.5). The Non-spreading group's CHG methylation slightly reduced on the arms of chromosomes (Figure 3.5). Similar to MITEs, the internal CHH methylation of non-spreading transposons increased on the arms of chromosomes (Figure 3.5). There are some similarities between the non-spreading group and the MITEs, such as distribution, and similar changes in methylation across the chromosome.

Effects of spreading transposons on gene expression

When either spreading group is near a gene, its expression level is on average lower than compared to the non-spreading group or when an LTR is absent upstream (Eichten *et al.*, 2012). A similar experiment was carried out with bisulfite and mRNA sequencing data, illustrated in Figure 3.6A. For our experiments, the combining of 5mc/H3K9 and H3K9 spreading groups were put into one group called "spreading" as well as included the non-spreading group. We then divided up all genes based on whether they had a spreading group or non-spreading group upstream 1kb. An additional control was included, a group of genes called "without spreading" which includes all other genes with a non-spreading or spreading TE absent upstream. Genes with a spreading group had significantly lower expression levels compared to non-spreading or without spreading groups (Figure 3.6B).

If the expression levels of genes are dependent on the nearby transposon's methylation status, then we would expect genes near spreading TEs to have higher methylation levels than genes near non-spreading TEs. The methylation of the first 10% for each of the defined groups of genes were measured for cytosine methylation (Figure 3.6C). The groups of genes with a spreading group upstream had significantly higher levels CG and CHG methylation compared to non-spreading and without spreading.

CHH islands and positive effect on gene expression

Genes are mostly found surrounded by vast seas of transposable elements. In between genes are sequences high in both CG and CHG DNA methylation, which drastically decline near genes. Methylation in the CHH context occurs primarily in CHH islands upstream and downstream of genes (Gent *et al.*, 2013). Consistent with the fact that RNA-dependent de novo methylation involves the production of small RNAs, the CHH islands are also islands of small RNA production. We found when splitting all genes into four categories based on upstream 1kb CHH methylation, there is a positive correlation with gene expression or RPKM, which is defined as the number of reads mapped to a gene and normalized by gene size and library size (Figure 3.7 A).

When studying the internal methylation of all transposon super-families, MITEs were found to have the highest CHH methylation. In addition, MITEs are most frequently found in the 1kb upstream and downstream region of genes (Han *et al.*, 2013). Analysis of the 1kb upstream regions revealed a clear positive correlation between MITE number and CHH methylation (Figure 3.7 B).

The distribution of MITEs near genes, their high internal CHH methylation, enrichment for mapped siRNA, and their correlation with highly expressed genes make them of prime interest. Why do MITEs, found enriched near genes and targeted by silencing mechanisms, associate with the group of genes with highest expression? A likely scenario is that they target highly expressed. We hypothesize that once DNA transposons insert into gene rich regions, MITEs act as boundary element or spacers that function to separate spreading class I elements from genes.

Gene expression levels positively correlated with upstream 1kb CHH methylation levels (Figure 3.7 A). We divided genes based on upstream 1kb CHH methylation levels to investigate gene methylation as it relates to upstream DNA methylation, in Figure 3.8 A. The highest CHH methylation levels upstream 1kb are quartile 1 while the genes with the lowest CHH methylation are quartile 4 (Figure 3.8 A). Previous studies found that the end of active genes are often devoid of methylation, and the first 10% of a gene is an accurate indicator of gene expression (Zilberman, Gehring, Tran, Ballinger, & Henikoff, 2006). We found that methylation in the first 10% of a gene is strongly correlated with gene expression and upstream CHH methylation. For example genes with the highest expression CHH methylation had the lowest CG and CHG methylation in the first 10% of the gene (Figure 3.8 A).

MITEs mark chromatin boundaries

If MITEs act as boundary elements, we would expect them to block, dampen, or at least be correlated with less spreading of DNA methylation into nearby genes. We devised three genomic experiments or comparisons to test if MITEs reduce methylation compared to different random controls. The first experiment compares MITE superfamilies (Figure 3.9), the second tests if the distance created by a MITE alone accounts for methylation change (Figure 3.10), and third experiment tests if there are regions upstream of genes that show reduction in methylation and if the difference found is similar to MITEs (Figure 3.11). Finally, we used published *mop1* mutant sequenced read to test whether the RdDM pathway is essential to the reduction of methylation over MITEs (Figure 3.12).

In the first experiment, we divided genes based on whether there is a MITE present or absent in the 3 kilobases upstream of the start site (always "both spreading" upstream of MITE). As expected from Figure 3.2 B, when a MITE was present cytosine was methylated at higher levels than when there was no MITE (Figure 3.8 B.). However, we also found the methylation level more sharply decrease upstream of the gene and within the gene body compared to when a MITE is not present (Figure 3.8 B). Whether CHH methylation was higher upstream of a gene or not, the gene body had low CHH methylation (Figure 3.8 B). CG and CHG methylation was lower across the entire length of the gene when a MITE was present upstream (Figure 3.8 B).

To compare different MITE superfamilies we analyzed the proportional change in CG and CHG methylation over MITEs in comparison to random controls (Figure 3.9 A, B). The MITEs with the largest change in methylation are *Mutator*, with a 27% CG and 24% CHG reduction over the length of a *Mutator* and 26% CG and 23% CHG reduction over *Tourist* (Figure 3.9 B). *CACTA* showed the lowest reduction in methylation compared to all other MITEs (Figure 3.9 B).

Boundary effect contributed to MITE size?

To further test the boundary effects of MITEs, we focused on the most abundant MITEs in the maize genome which also showed the largest reduction in Figure 3.9B, *Tourist* and *Mutator*. For the test in Figure 3.9, we used a 200bp length for the “random” sample. For the next experiment, instead of a static random sample size, we made a sample size based on the same length and total number of *Tourist* and *Mutator* elements (Figure 3.10A). For instance, if there were three *Tourist* elements with sizes 200bp, 300bp, and 150bp, then there would be three random samples with the same lengths, except their position would be randomized over a 1kb regions upstream of a gene. This was to test whether the total length of a MITE was the contributor to reduction of methylation. The reduction over the two MITE superfamilies are compared to the “Random” sample, and the proportional change in methylation over each was calculated. *Tourist* and *Mutator* both showed a reduction of CG and CHG methylation of 50 percent, while random reduced at around 20 percent (Figure 3.10B). These data would indicate that the length of a MITE alone does not contribute to the total reduction of methylation over a MITE. However, MITEs could target the region upstream of genes where the greatest reduction of methylation naturally occurs.

Boundary effect compared to sliding winding

Next we tested whether there is a region upstream of gene in which the reduction of CG and CHG methylation is greatest and compared it to all MITEs superfamilies in the same group of genes. We developed a similar experiment to Figure 3.9, except we calculated change in CG and CHG methylation over 200bp windows across 1kb upstream of genes (with spreading TE upstream) in an attempt to identify the greatest change in methylation (Figure 3.11A). For instance, for the position 200-400 bp upstream of a gene, we subtracted 100bp flank methylation downstream from 100bp methylation upstream of 400bp. We checked the methylation change over a sliding window for the highest expressed gene, or quartile 4 (genes were divided into four groups based on expression levels and the highest quartile was used only). The largest reduction in CG methylation over a 200 base pair region was at 800-1000bp upstream position at ~15%. The highest reduction of CHG methylation is over 600-800, upstream of genes, at ~22.5%. We averaged the change in methylation over all MITE superfamilies, but each MITE had to be within 1kb upstream of gene, and found ~48% reduction in methylation for CG and 53% reduction for CHG over a MITE. There did not seem to be particular upstream region of genes that naturally reduces in methylation at the same level of MITEs.

Boundary effects in mop1 mutant

MITEs are enriched for CHH methylation and siRNA indicating that they are targeted by the RdDM pathway. The RNA-dependent RNA polymerase mutant, mop1 has been shown to be necessary for siRNA production and mutants show reduced CHH methylation (Alleman *et al.*, 2006; Gent *et al.*, 2014). To test whether siRNA or CHH methylation is necessary for the change in methylation measured over MITEs, we analyzed sequence *mop1* mutant and control lines (Gent *et al.*, 2014). Using the same position comparison in Figure 3.11, we compared *mop1* mutant to the control sequence reads. For *mop1*, (which is *mop1* mutant backcrossed to B73) MITEs had 56% and 59% reduction of methylation CG and CHG, compared to the B73 sample of 52% CG and 58% CHG methylation reduction (Figure 3.12). We found very little difference between *mop1* mutant versus the control sample, indicating that siRNA and CHH methylation are not necessary for reduction of methylation over a MITE. MITEs in the control and *mop1* mutant had a greater reduction in methylation than any other upstream sliding window samples (Figure 3.12). These data indicate that the production and targeting of these site, downstream of RDR2 is not necessary for the change in methylation.

MITES have unusual sequence features and chromatin profiles

Sequence content or presence of a particular motif in a MITE could affect the capacity of methylation to spread across their length. We investigated the frequency of CG motifs (which are subject to CG methylation) and overall A/T content among spreading (Both class I spreading groups) elements, non-spreading (class I) elements 1kb flanks, and MITEs (class II) of the major categories (Figure 3.13 A, B). Although there is considerable variation among transposons within these families, it is noteworthy that *Tourist* and *Stowaway*, which are the two MITE families with highest CHH methylation (Figure 3.2B) had the most severe barrier effects, also have fewest CG motifs and highest A/T content (Figure 3.13A, B). *Mutator*, which is abundant in the genome, close to genes, has relatively high CHH methylation, and severe barrier effect also had low CG motif% and high A/T% (Figure 3.13B). *CACTA*, which had the least severe barrier effect found furthest from genes on average, has a negative effect on gene expression, has lowest CHH methylation as well as lowest A/T richness. Conversely spreading transposons showed a high level of CG motifs in their flanking regions relative to other repetitive elements, non-spreading transposon's flanking regions, random samples, and genes (Figure 3.13A). These results generally point to the possibility that sequence content, specifically fewer CG-motifs and higher A/T content could contribute to reduction of methylation over MITEs.

Finally, we investigated the chromatin make up of MITEs when downstream of spreading TEs, anticipating a unique pattern of markers. Prior data indicated that methylation at both histone 3 lysine 27 di-methylation (H3K27me2) and lysine 9 di-methylation (H3K9me2) are associated primarily with class I retrotransposons and “deep heterochromatin” (Gent *et al.* 2014). However, CHH islands tend to have relatively low H3K9me2 methylation and relatively accessible chromatin as measured by micrococcal nuclease sensitivity (Gent *et al.* 2014). Given that MITEs are a major component of CHH islands, its stands to reason that they might also show relatively low enrichment for H3K9me2. An analysis of the major MITE families revealed that each group has a similar profile to RdDM loci defined by (Gent *et al.*, 2014). As can be seen in Figure 3.14, non-RdDM loci have similar levels of H3K27me2 and H3K9me2, whereas RdDM loci and all MITE families have significantly reduced H3K9me2 (Figure 3.14). MITEs have unique chromatin structure and sequence content which could contribute to their role as a boundary element.

Discussion

Recent studies found several families of retrotransposons that show signs of spreading 5mC and H3K9me2 into low-copy regions (Eichten *et al.* 2012). With the use of bisulfite sequence, the gold standard for methylation analyses, we found that the flanking regions of both spreading groups had a higher methylation level in the 1kb flanking regions when compared to non-spreading groups. When combining both groups and analyzing genes within 3kb of these “spreading” TEs, their expression levels were low compared to the genome average and had relatively high levels of CG and CHG methylation in the first 10% of the gene. These general observations allowed the opportunity to test whether MITEs act as boundary elements between heterochromatin and genes.

This study provides several promising associations for a subset of MITE superfamilies to act as boundary elements for genes. Previous studies show evidence that MITEs are targeted by RdDM pathway and have chromatin structure unlike other repetitive elements in the genome (Figure 13; (Gent *et al.*, 2013). While MITEs are targeted by siRNA and have high CHH% methylation (Gent *et al.*, 2013), when comparing *mop1* mutants to a control, MITEs still showed severe methylation reduction. These results indicate that the RdDM pathway is not associated with the reduction of methylation over MITEs.

Genes with a spreading TE upstream had low expression levels and higher levels of gene methylation. When a MITE was found between a spreading TE and the gene, gene body methylation was lower and expression levels higher. Further, by comparing MITEs to genomic positions similar in length, MITEs show significantly reduced methylation over their own sequence in comparison. In addition, MITE sequences are A/T-rich with few CG motifs, unfavorable for cytosine methylation. MITEs are abundant between genes and heterochromatin regions, allowing them to create space between deep heterochromatin regions and euchromatic genes. Since methylation over a MITE reduces more than genomic space of the same length, MITEs not only act as a spacer, but as a boundary element.

The differences found between MITE superfamilies with respect to methylation, distribution, and sequence content indicate that some MITEs are more effective boundary elements. For class II elements, higher CHH methylation generally correlates with lower internal CG and CHG methylation. For example, *CACTA* had the lowest CHH methylation and highest CG and CHG methylation, while *Tourist* and *Stowaway* had one of the highest CHH methylation and lowest CG and CHG methylation. *Tourist* had one of the highest change in methylation of 25% CG and 28% CHG reduction in methylation compared to 18% reduction of both CG and CHG methylation for *CACTA*. *CACTA* has low A/T content, low CHH methylation, high percent of CG-motifs and showed more reduction of methylation than the random sample. However, the *CACTA* superfamilies contain the longest length of the MITEs (Han *et al.*, 2013) and we did not compare it to a random sample of similar length positions. Overall, the results for *CACTA* exemplifies what would be expected of a moderately effective barrier. *Tourist*, a highly abundant MITE, second smallest (length) superfamily, with high CHH methylation, low CG-motif content and A/T-rich, has 30% more reduction in methylation compared to positions of a similar length. Furthermore, *Tourist* and *Stowaway* showed the lowest H3K9me2 enrichment at 0.5 and 0.6 respectively, a level closer to the exons of genes than to non-RdDM loci. *CACTA* elements may be poor barriers as a result of being far from genes, while *Tourist* elements are closer to genes (Figure 3.3). One possibility is that near-gene *CACTA* have been selected against because they negative effects on genes, while *Tourist* elements are still enriched in genic regions because they have a general positive effect on gene expression.

Transposable elements are often described as inherently mutagenic. Paradoxically, transposon enriched regions are referred to as “junk” DNA. Retrotransposable elements are primarily found in the pericentromeric regions, often inserted into LTRs of transposons, likely allowing for their high copy numbers. The general consensus from recent studies are that DNA transposons play a significant role in the plant genomes and contribute to genetic diversity (Bennetzen, 2000). In addition to the broad implications of genetic diversity, studies have found DNA transposons targeting 5’ and 3’ of genes in a short-term burst of activity while avoiding exons resulted in either no change or increased gene expression, (Naito *et al.*, 2006). DNA transposons insertion into the 5’ and 3’ region of genes has been hypothesized to contribute to MARs (Matrix attachment region), forming a higher order chromatin structure and acting as a boundary for genes (Bennetzen, 2000). Not many studies have further investigated this hypothesis. One possible reason for MITEs to target the more active regions is that they have an open chromatin structure, allowing for easy transposition. In fact, this would corroborate with the results of increased number of MITEs correlating positively with higher expressed genes. Also, some class II elements have been found to target specific sequences, such as TAA preference for *Tourist* (Wessler, Bureau, & White, 1995). Recent *mPing,* of the *PIF/harbinger* superfamily, insertional analyses in rice and yeast found *mPing* to have a preference for 9 bp A/T rich sequence (Hancock, Zhang, & Wessler, 2010). In general, DNA containing CG bonds are considered more stable than AT bonds, and having an AT-rich upstream region of a gene may contribute to more open chromatin. In cereal grasses, all families except AT-poor MITE *CACTA* are associated with genes at expression levels higher than average (Han *et al.,* 2013). These data would suggest MITEs target other MITEs which are inherently A/T-rich, and as a result increase boundary length between genes and heterochromatin. Evolutionarily, this would be a niche for MITEs to proliferate without harming the organism by high gene mutation rates and contributing positively to the genomic environment by separating harmful heterochromatic regions and transcriptionally active genes.

Materials and Methods

DNA extraction and Bisulfite-treated DNA

DNA was extracted from the outer tissues of B73 ears whose silks had emerged but had not been fertilized. Sodium bisulfite-treated Illumina sequencing libraries were prepared using a method similar to that of (Lister *et al.*, 2009). Alignment to the genome (AGPv2) and identification of methylated cytosines was performed using BS Seeker (Chen, Cokus, & Pellegrini, 2010). A total of 198,333,982 single-end reads with unique alignments specifically on the ten chromosomes were obtained, with an average genome-matching read length of 72.8 bases (Eichten *et al.*, 2012).

The level of methylation in CG, CHG and CHH contexts and the total proportion of DNA methylation was calculated for non-repeat masked sequences (as annotated within ZmB73_5a_MTEC_repeats) located within 1 kb of each retrotransposon family. Percent methylation is defined as the number of methylated Cs per total number of Cs for a region. BEDTools (Quinlan & Hall, 2010) was used to identify low-copy sequences flanking retrotransposons. Superfamilies of MITEs were done using, Mite hunter (Han & Wessler, 2010) and to identify MITE positions in the genome, Repeatmask and .out file as defined positions in the genome. Class I superfamilies were obtained from the set of MTEC repeats (version 5a; http://ftp.maizesequence.org/current/repeats/).

mRNA and expression analysis

mRNA reads of 100bp in length and divided into separate file to ensure highest quality reads were used as previously described (Gent *et al.*, 2013).

Genome-wide alignment, methylation calculations

Genome alignment of mRNA sequence reads was done with version 3 genome using Bowtie, v2 (Langmead, Trapnell, Pop, & Salzberg, 2009). We used specific parameters for mapping reads and methylation calculations with custom python scripts (Gent *et al.*, 2013). Coordinates used in this study, such as whole genes and exons, were downloaded from

http://ftp.maizesequence.org/current/filtered-set/.

H3K9me2, H3K27me2, mop1 data

Mop1 mutant, mop1 control, H3K9me2 and H3K27me2 ChIP reads were obtained (Gent *et al.*, 2014) and contained single-end, 151 base pair length. All reads were aligned to the RefGen_v3 maize genome and analyzed with the use of BEDtools (Quinlan & Hall, 2010) and the use of custom python3 scripts to calculate RPKM at specific positions in the genome, and to form meta-chromosome and meta-gene analyses. A line that was *mop1-1* mutant in a B73 genetic background and a sibling which was homozygous wild-type was used for a control. Publically available bisulfite treated reads sequence were processed in similar manner to passed study (Gent *et al.*, 2014).

Significance tests

To calculate p-value for comparisons of methylation levels between groups of transposons and expression levels between gene groups we performed two-tailed Student's t-tests.

Analyses and python programs

We divided the length of genes, genomes, and upstream regions in specific bin sizes and calculated the relative cytosine methylation for all sequence context. To analyze instances of specific genomic positions or sequence of interest (SOI), custom python programs were created. Bisulfite treated reads were aligned using BS Seeker software (Chen *et al.*, 2010) as previously carried out (Gent *et al.*, 2013). BEDtools was used for several analyses to find positions with overlapping MITE, spreading TEs, etc. (Quinlan & Hall, 2010). Meat-gene analyses were compared to recently publically available similar program at https://github.com/tjparnell/biotoolbox to ensure accuracy.

To make meta-chromosome plots, a custom python script called “Methylation Distribution Meta-Chromosome” (Appendix A) was created, and calculates the relative methylation (methylated Cs/possible Cs), total methylation (total number of methylated Cs), the number SOI (in this case transposons), the total length of each SOI bisulfite sequence reads. Genomic positions or sequence of interest need to be in the .gff format and for this experiment, we looked at internal transposon and flanking 1kb methylation. Once the main counts are done the average for each chromosome is calculated for each bin (user defined) and averaged across all 10 chromosome of maize.

To make meta-gene plots, another program called “Methylation Distribution Meta-Gene” (Appendix B), in which methylation levels can be measured and summarized for SOI, such as a genic regions or whole transposons. The input file must be a BS Seeker output file which have been extracted overlapping read (program not shown). For gene methylation analysis for example, all genes were divided into groups and genes in each group produced a meta-gene. Calculations for relative methylation and total methylation in each sequence context was applied to each gene and averaged for each defined bin. These examples of python scripts were created in collaboration with Jonathan Gent.

References

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

Bennetzen, J. L. (2000). Transposable element contributions to plant gene and genome evolution. *Plant molecular biology, 42*(1), 251-269.

Cazzonelli, C. I., Cuttriss, A. J., Cossetto, S. B., Pye, W., Crisp, P., Whelan, J., . . . Pogson, B. J. (2009). Regulation of carotenoid composition and shoot branching in Arabidopsis by a chromatin modifying histone methyltransferase, SDG8. *The Plant Cell, 21*(1), 39-53.

Chen, P.-Y., Cokus, S. J., & Pellegrini, M. (2010). BS Seeker: precise mapping for bisulfite sequencing. *BMC bioinformatics, 11*(1), 203.

Copenhaver, G. P., de Koning, A. P. J., Gu, W., Castoe, T. A., Batzer, M. A., & Pollock, D. D. (2011). Repetitive Elements May Comprise Over Two-Thirds of the Human Genome. *PLoS Genetics, 7*(12), e1002384. doi: 10.1371/journal.pgen.1002384

Cordaux, R., & Batzer, M. A. (2009). The impact of retrotransposons on human genome evolution. *Nature Reviews Genetics, 10*(10), 691-703.

Eichten, S. R., Ellis, N. A., Makarevitch, I., Yeh, C.-T., Gent, J. I., Guo, L., . . . Vaughn, M. W. (2012). Spreading of heterochromatin is limited to specific families of maize retrotransposons. *PLoS Genetics, 8*(12), e1003127.

Feng, S., Cokus, S. J., Zhang, X., Chen, P.-Y., Bostick, M., Goll, M. G., . . . Jacobsen, S. E. (2010). Conservation and divergence of methylation patterning in plants and animals. *Proceedings of the National Academy of Sciences*. doi: 10.1073/pnas.1002720107

Feschotte, C., & Pritham, E. J. (2007). DNA transposons and the evolution of eukaryotic genomes. *Annual review of genetics, 41*, 331.

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

Gent, J. I. *et al.* (2014). Accessible DNA and relative depletion of H3K9me2 at maize loci undergoing RNA-directed DNA methylation. *The Plant Cell, 26*(12), 4903-4917.

Ha, M., Ng, D. W. K., Li, W. H., & Chen, Z. J. (2011). Coordinated histone modifications are associated with gene expression variation within and between species. *Genome Research, 21*(4), 590-598. doi: 10.1101/gr.116467.110

Han, Y., Qin, S., & Wessler, S. R. (2013). Comparison of class 2 transposable elements at superfamily resolution reveals conserved and distinct features in cereal grass genomes. *BMC genomics, 14*(1), 71.

Han, Y., & Wessler, S. R. (2010). MITE-Hunter: a program for discovering miniature inverted-repeat transposable elements from genomic sequences. *Nucleic Acids Research,* gkq862.

Hancock, C. N., Zhang, F., & Wessler, S. R. (2010). Transposition of the Tourist-MITE mPing in yeast: an assay that retains key features of catalysis by the class 2 PIF/Harbinger superfamily. *Mobile DNA, 1*(1), 5. doi: 10.1186/1759-8753-1-5

Heitz, E. (1928). Das heterochromatin der moose. *I. Jahrb. Wiss. Botanik, 69,* 762-818.

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

Johnson, L. M., Bostick, M., Zhang, X., Kraft, E., Henderson, I., Callis, J., & Jacobsen, S. E. (2007). The SRA methyl-cytosine-binding domain links DNA and histone methylation *Curr Biol* (Vol. 17, pp. 379-384). England.

Jones, P. A. (2012). Functions of DNA methylation: islands, start sites, gene bodies and beyond. *Nature Reviews Genetics, 13*(7), 484-492. doi: 10.1038/nrg3230

Kato, M., Miura, A., Bender, J., Jacobsen, S. E., & Kakutani, T. (2003). Role of CG and non-CG methylation in immobilization of transposons in Arabidopsis *Curr Biol* (Vol. 13, pp. 421-426). England.

Kumar, A., & Bennetzen, J. L. (1999). Plant retrotransposons. *Annual review of genetics, 33*(1), 479-532.

Langmead, B., Trapnell, C., Pop, M., & Salzberg, S. L. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome biol, 10*(3), R25.

Law, J. A., & Jacobsen, S. E. (2010). Establishing, maintaining and modifying DNA methylation patterns in plants and animals. *Nat Rev Genet, 11*(3), 204-220. doi: 10.1038/nrg2719

Li, E., Bestor, T. H., & Jaenisch, R. (1992). Targeted mutation of the DNA methyltransferase gene results in embryonic lethality *Cell* (Vol. 69, pp. 915-926). United States.

Lisch, D. (2009). Epigenetic Regulation of Transposable Elements in Plants. *Annual Review of Plant Biology, 60*(1), 43-66. doi: doi:10.1146/annurev.arplant.59.032607.092744

Lister, R., Pelizzola, M., Dowen, R. H., Hawkins, R. D., Hon, G., Tonti-Filippini, J., . . . Ngo, Q.-M. (2009). Human DNA methylomes at base resolution show widespread epigenomic differences. *Nature, 462*(7271), 315-322.

Liu, R., Vitte, C., Ma, J., Mahama, A. A., Dhliwayo, T., Lee, M., & Bennetzen, J. L. (2007). A GeneTrek analysis of the maize genome. *Proc Natl Acad Sci U S A, 104*(28), 11844-11849. doi: 10.1073/pnas.0704258104

Liu, S., Yeh, C. T., Ji, T., Ying, K., Wu, H., Tang, H. M., . . . Schnable, P. S. (2009). Mu transposon insertion sites and meiotic recombination events co-localize with epigenetic marks for open chromatin across the maize genome. *PLoS Genet, 5*(11), e1000733. doi: 10.1371/journal.pgen.1000733

McClintock, B. (1984). The significance of responses of the genome to challenge. *Science, 226*(4676), 792-801.

Naito, K., Cho, E., Yang, G., Campbell, M. A., Yano, K., Okumoto, Y., . . . Wessler, S. R. (2006). Dramatic amplification of a rice transposable element during recent domestication. *Proceedings of the National Academy of Sciences, 103*(47), 17620-17625.

Oki, N., Yano, K., Okumoto, Y., Tsukiyama, T., Teraishi, M., & Tanisaka, T. (2008). A genome-wide view of miniature inverted-repeat transposable elements (MITEs) in rice, Oryza sativa ssp. japonica. *Genes & genetic systems, 83*(4), 321-329.

Pereira, V., Enard, D., & Eyre-Walker, A. (2009). The effect of transposable element insertions on gene expression evolution in rodents. *PLoS ONE, 4*(2), e4321. doi: 10.1371/journal.pone.0004321

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

Richards, E. J. (2006). Inherited epigenetic variation—revisiting soft inheritance. *Nature Reviews Genetics, 7*(5), 395-401.

Schnable, P. S., *et al.* (2009). The B73 Maize Genome: Complexity, Diversity, and Dynamics. *Science, 326*(5956), 1112-1115. doi: 10.1126/science.1178534

Smit, A. F., Hubley, R., & Green, P. (1996). RepeatMasker Open-3.0.

Veiseth, S. V., Rahman, M. A., Yap, K. L., Fischer, A., Egge-Jacobsen, W., Reuter, G., . . . Thorstensen, T. (2011). The SUVR4 histone lysine methyltransferase binds ubiquitin and converts H3K9me1 to H3K9me3 on transposon chromatin in Arabidopsis. *PLoS Genet, 7*(3), e1001325.

Weil, C., & Martienssen, R. (2008). Epigenetic interactions between transposons and genes: lessons from plants. *Current opinion in genetics & development, 18*(2), 188-192.

Wessler, S. R., Bureau, T. E., & White, S. E. (1995). LTR-retrotransposons and MITEs: important players in the evolution of plant genomes. *Current opinion in genetics & development, 5*(6), 814-821.

Yuan, Y.-W., & Wessler, S. R. (2011). The catalytic domain of all eukaryotic cut-and-paste transposase superfamilies. *Proceedings of the National Academy of Sciences, 108*(19), 7884-7889.

Zhang, X. (2008). The Epigenetic Landscape of Plants. *Science, 320*(5875), 489-492. doi: 10.1126/science.1153996

Zhang, X., & Jacobsen, S. E. (2006). Genetic analyses of DNA methyltransferases in Arabidopsis thaliana. *Cold Spring Harb Symp Quant Biol, 71*, 439-447. doi: 10.1101/sqb.2006.71.047

Zilberman, D., Gehring, M., Tran, R. K., Ballinger, T., & Henikoff, S. (2006). Genome-wide analysis of Arabidopsis thaliana DNA methylation uncovers an interdependence between methylation and transcription. *Nature Genetics, 39*(1), 61-69. doi: 10.1038/ng1929

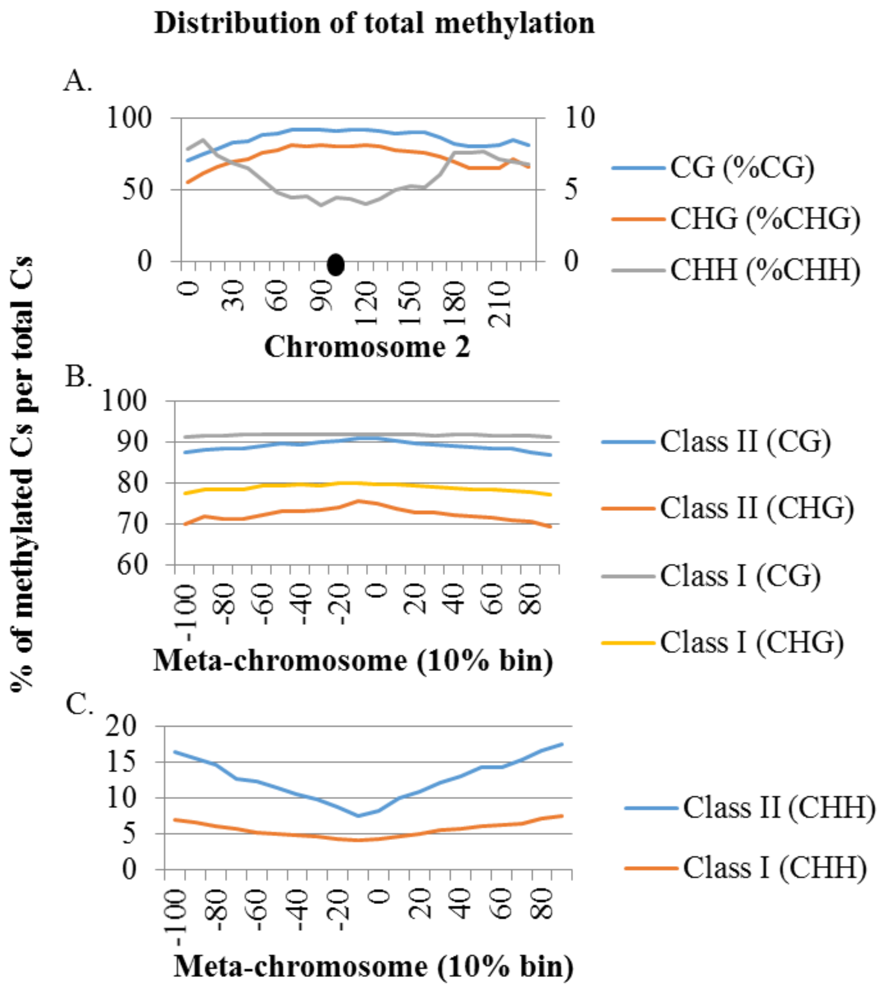
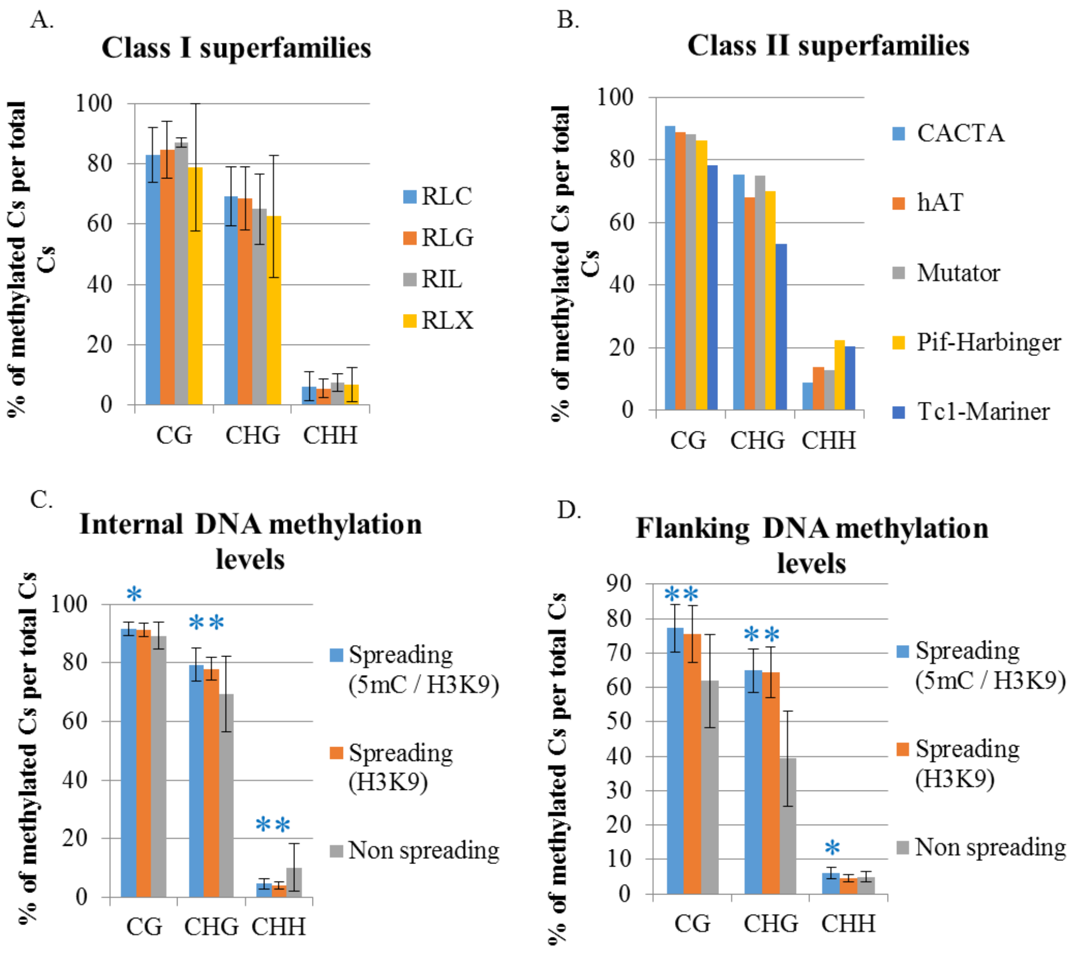


Figure 3.2. Genome-wide DNA methylation and internal class I and class II internal methylation distribution. A. Genome-wide distribution for each DNA methylation type across chromosome 2 (x-axis is in MB), similar trend for all chromosomes (black dot represents centromere. B. Distribution of internal DNA methylation (CG and CHG) for class I and class II element as relative enrichment across a meta-chromosome plot. The arms are divided into 10% bins and the 0 represents the centromere. C. CHH percent methylation for class I and class II element relative to distribution along a meta-chromosome plot.



Error bars: standard deviation

*=Bar significantly different than non-spreading

Figure 3.3. Internal and flanking DNA methylation levels for different groups of transposons. The error bar represents the standard deviation. A. Internal methylation averages in class I superfamilies. B. Internal DNA methylation levels for class II superfamilies. C. Internal methylation average for the spreading and non-spreading groups. D. Flanking DNA methylation frequency in the low-copy flanking regions for each spreading group. (low-copy = non-repetitive)

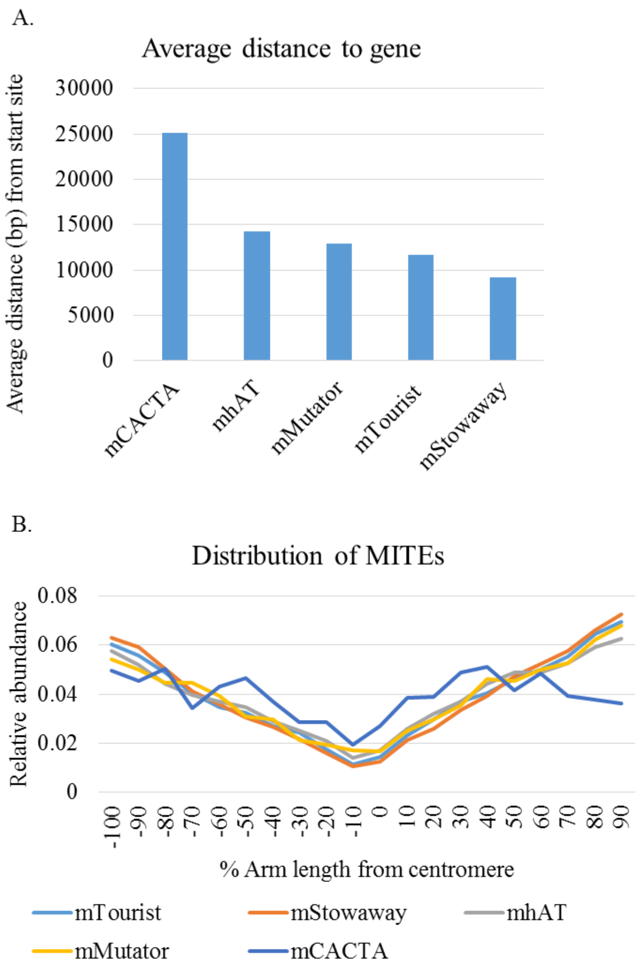


Figure 3.3. Genome distribution averages for class II superfamily MITE derivatives. A. Average distance of MITE superfamily to nearest start site of a gene. B. Distribution of MITEs across meta-chromosome plot. Arms are divided into 10% and 0 represents the centromere. Each arm of a chromosome is divided into 10% bins and the average number of each MITE superfamily is calculated and averaged across all 10 chromosomes. (m = MITE form)

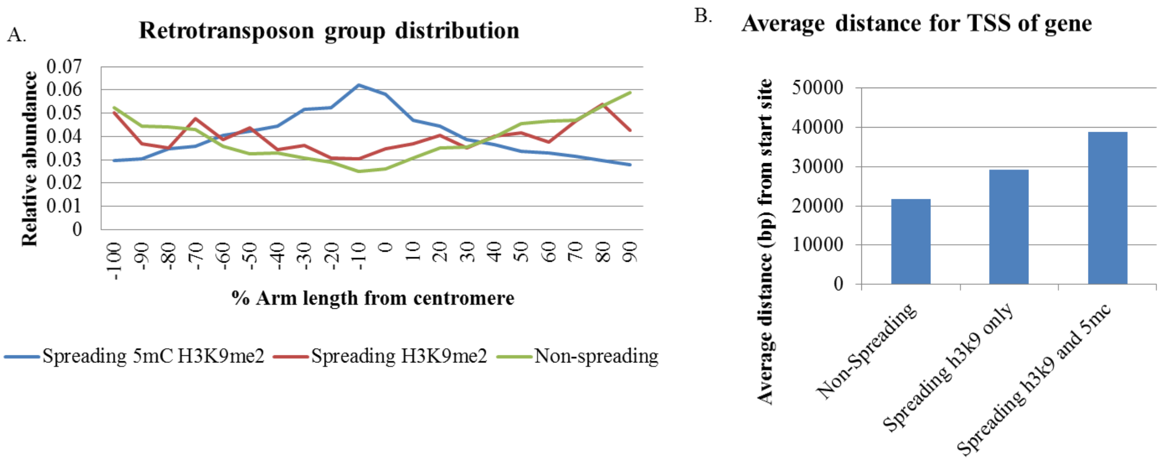


Figure 3.4. Genome distribution of spreading groups and non-spreading group. A. Relative abundance of each group along the length of a meta-chromosome plot. Each arm of a chromosome is divided into 10% bins and the average number of each TE is calculated and averaged across all 10 chromosomes. B. Average base pair distance of each group to the nearest transcription start site of a gene.

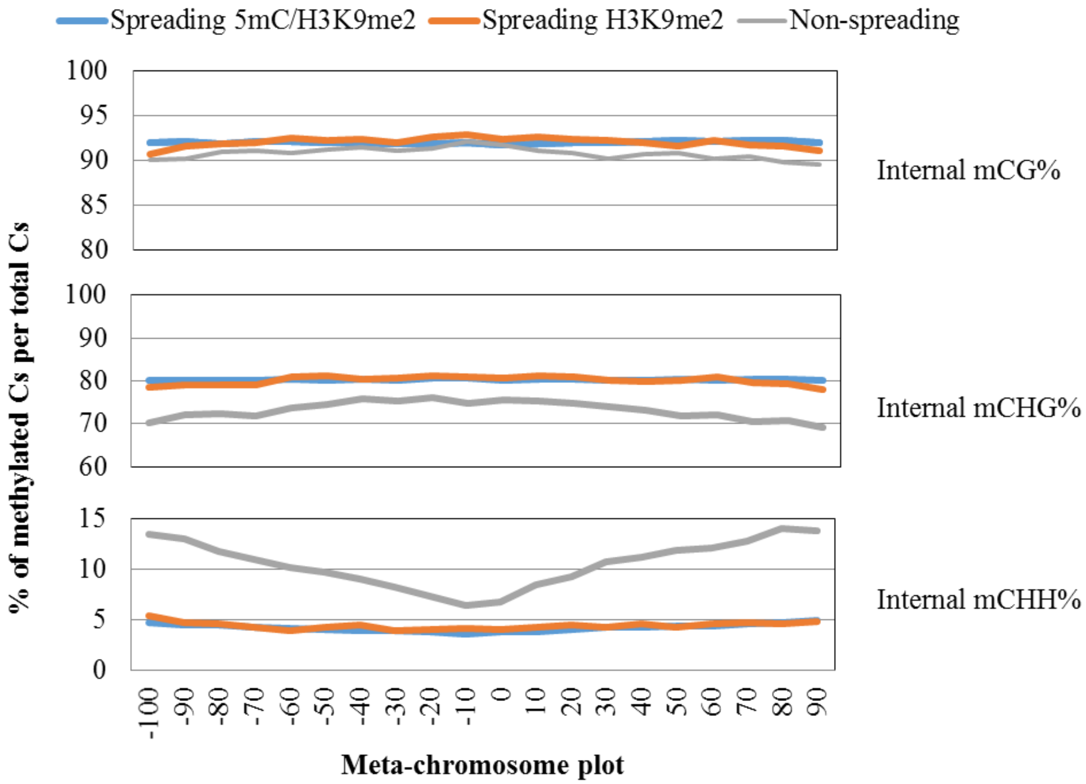
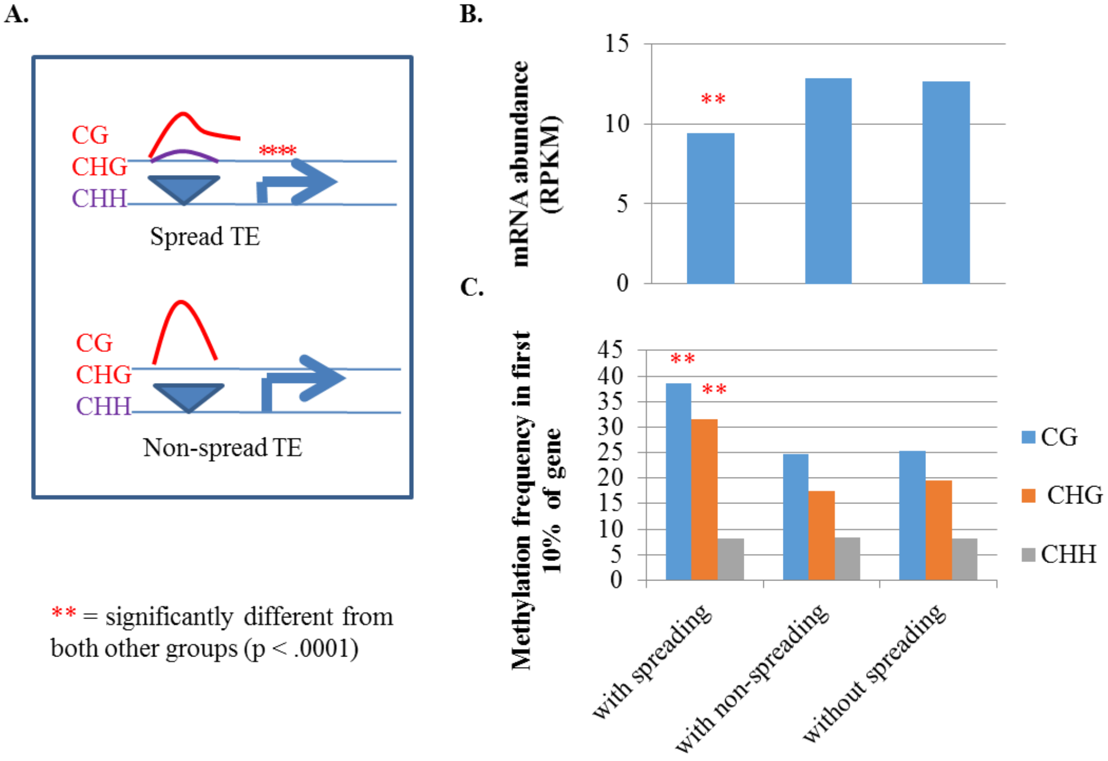


Figure 3.5. Internal methylation of spreading groups relative to distribution along meta-chromosome plots for mCG (m= methylated CG), mCHG, and mCHH. Each graph is divided by methylation type and 0 represent the centromere.



** = significantly different from both other groups (p < .0001)

Figure 3.6. Comparison of gene expression and gene methylation based on upstream 1kb content. Groups of genes are divided based on type of class I overlapping upstream 1kb or no TE present. A. Diagram representing hypothetical scenario of CG and CHG heterochromatin spread into nearby gene compared to a non-spreading TE. B. Comparison of expression levels calculated by reads per kilobase per million mapped reads (RPKM) for each group. Two asterisks indicates that the value is significantly different than both “with non-spreading” and “without spreading”. C. Comparison of DNA methylation calculated for the first 10% of genes (including exons and introns) for each group.

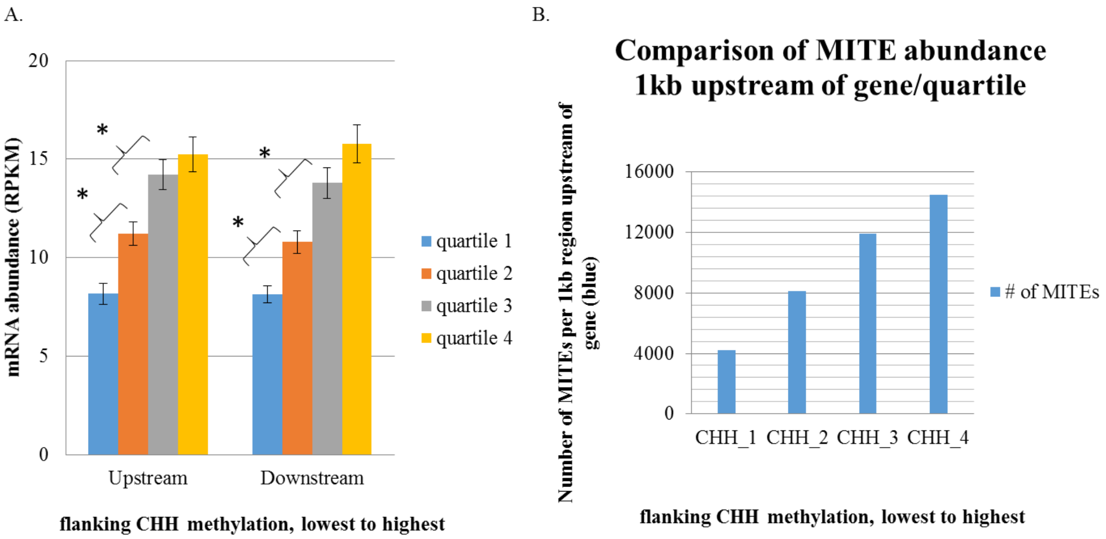


Figure 3.7. Comparison of gene groups divided by upstream 1kb CHH% average. A. Quartiles of genes based on upstream 1kb CHH methylation levels and near gene expression level average. Asterisk indicates significant difference. B. Quartiles of genes based on upstream 1kb CHH methylation levels and total number of MITEs in upstream region for each quartile. CHH_1 is quartile one or lowest CHH methylation levels while CHH_4 has the highest CHH levels.

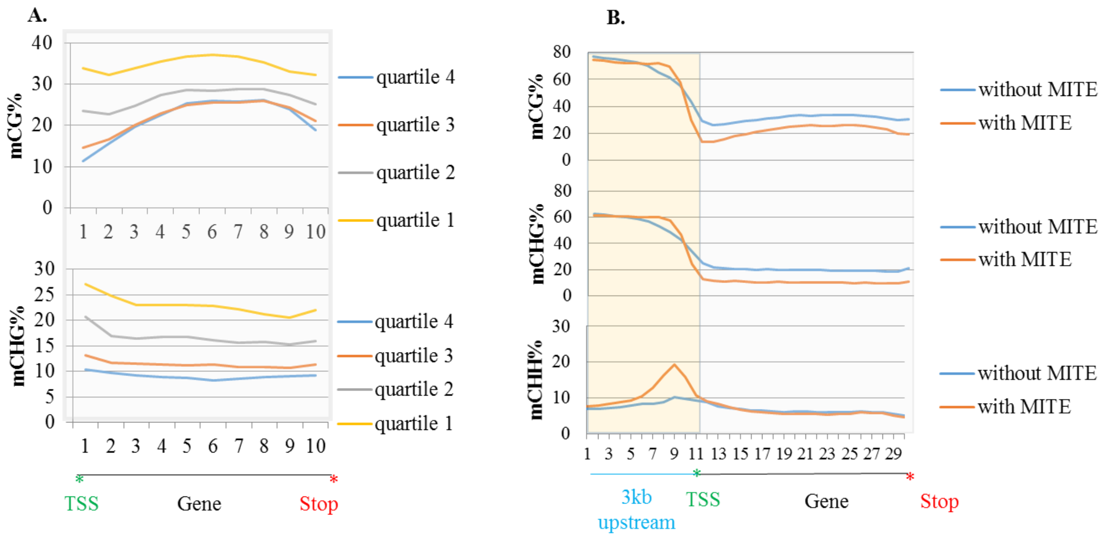


Figure 3.8. Comparison of gene body methylation to upstream 1kb CHH and gene body methylation with or without MITE. Genes were averaged and plotted into unit domains starting at the transcription start site (TSS) and ending at the stop codon. Only the exon portions of a gene were calculated for cytosine methylation. A. All genes in the maize genome were divided up into four equal sized quartiles based on 1kb upstream CHH methylation levels and plotted in 10% unit domains. Quartile 1 has the lowest upstream mCHH while quartile 4 has the highest mCHH methylation. B. All genes were divided into two groups, genes “with MITE” or at least one MITE present in their upstream 1kb region and “without MITE” which have none. The upstream 3kb promoter regions were averaged and plotted as 300bp per unit domain while the genes are in 5% bins.

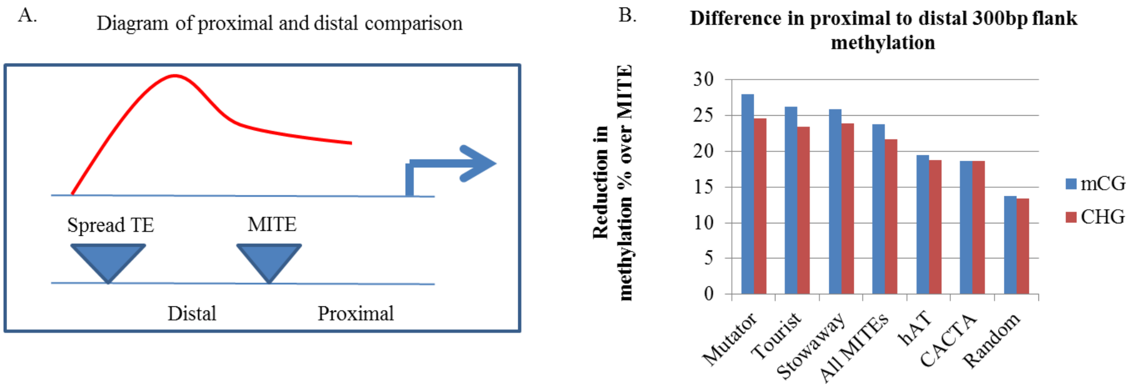


Figure 3.9. Methylation level change between the distal and the proximal flanks of a MITEs. A. Diagram illustrating the change in methylation over a MITE when upstream of a gene. Methylation change is calculated by the distal methylation percent minus the proximal methylation percent. B. Bar graph comparing the methylation change between the distal 300bp flank and the proximal (to gene) flank, and the difference is represented for mCG and mCHG methylation. The average length of a MITE was calculated to be 179bp. Random is defined as a 200bp region equidistant to a spreading TE and nearest gene, then the change in flank methylation was calculated.

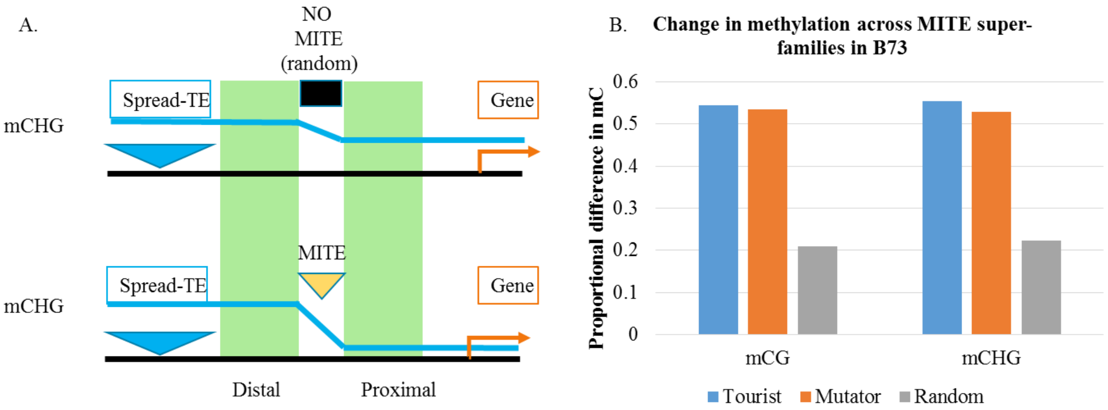


Figure 3.10. Comparison of DNA methylation change over MITE positions of *Tourist* and *Mutator* to Random, non-MITE positions of the same size. A. Diagram of comparison of a MITE upstream to a random non-mite position of equal size and number of positions. The green bars represent the flanking region of the mite or random position to calculate change in methylation. In all cases a spread group of both (H3K9me2 and 5mC/H3K9me2) were within 1kb. B. comparison of the *Tourist* and *Mutator* to random sampling. Proportional difference CG and CHG methylation was calculated.

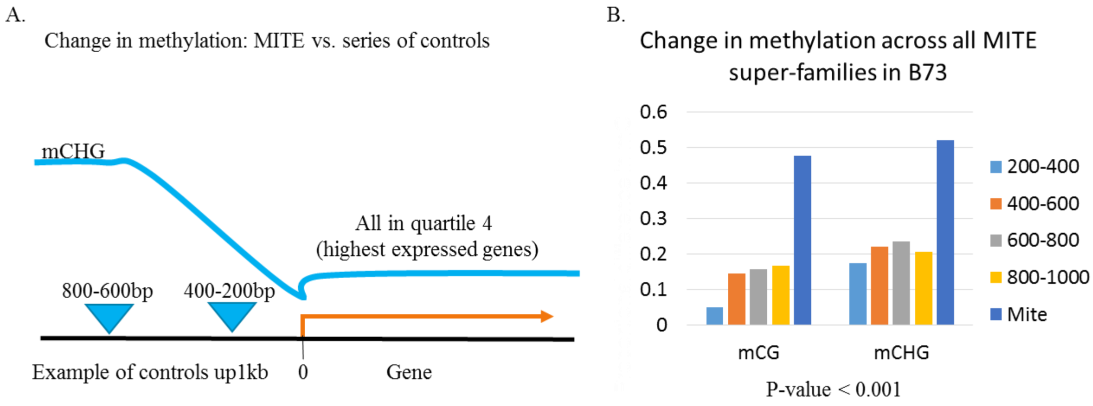


Figure 3.11. Comparison of methylation change for all MITEs sliding window samples. A. Diagram illustrating the comparison of CHG DNA methylation over sliding window controls, and upstream with at least one upstream 3kb spreading TE. B. Proportional difference in C methylation between sliding window positions compared to the average for MITEs. All MITEs were included in the analysis, and the flanking of MITEs and sliding windows 100 bp flanks were calculated.

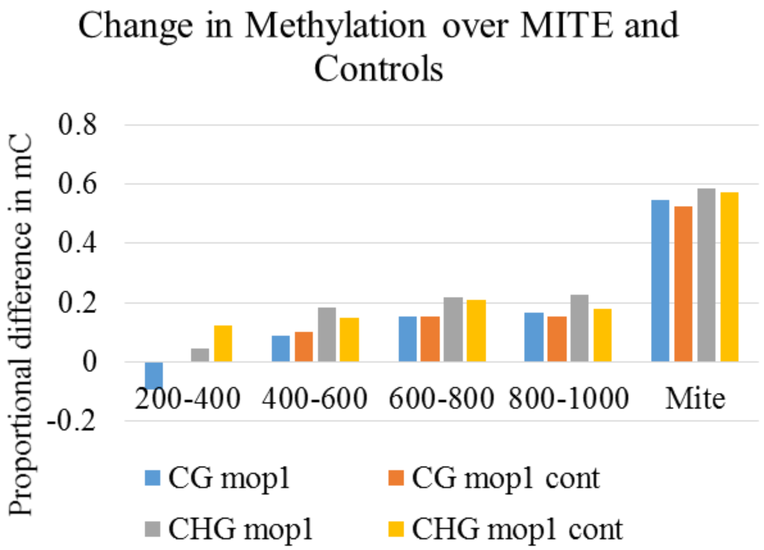


Figure 3.12. Change in methylation over all MITEs in *mop1* mutant. Proportional difference in methylation of 200bp sliding windows upstream of genes with spreading TE upstream 3kb. All MITEs were included in the analysis, and the flanking of MITEs and sliding windows 100 bp flanks were calculated.

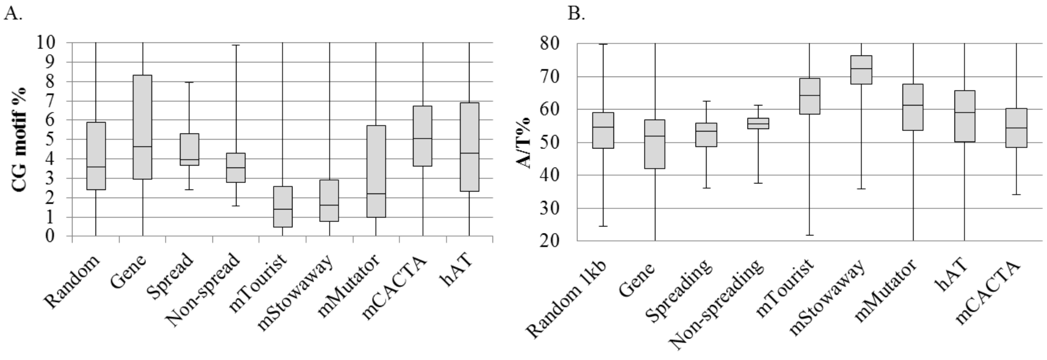


Figure 3.13. Box-and-whisker plots comparing CG motif percent and A/T percent at various genomic positions. The internal sequence of MITEs were calculated, mTourist (m = MITE form), mStowaway, mMutator, mCACTA, and mhAT. Full length genes were calculated, including exons and introns. The flanking regions of Spread (5mc/H3K9 and H3K9 only) and Non-spread were calculated. A. CG motif percent and B. A/T percent in each defined positions. Random is defined as 1kb position randomly distributed around the maize genome.

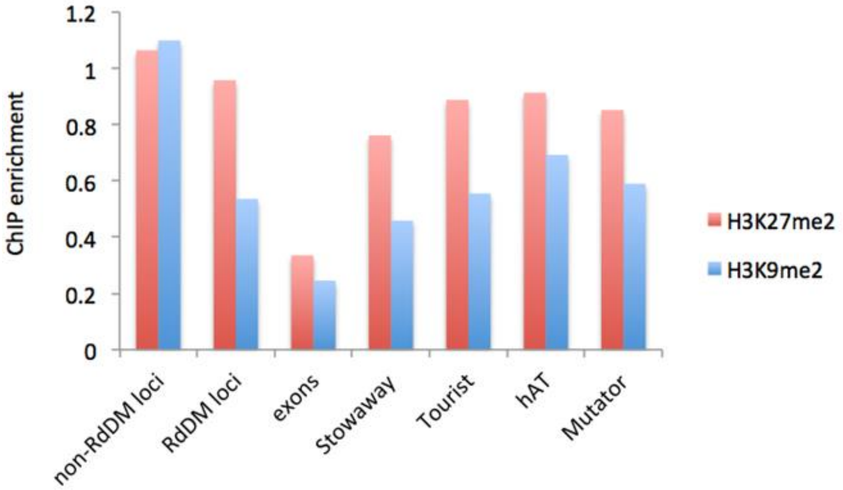


Figure 3.14. Comparison of H3K9me2 and H3K27me2 enrichment of RdDM loci (defined in Gent *et al.* 2015) exons and MITEs. The internal regions of exons and MITEs were calculated for ChIP enrichment for antibodies against H3K27me2 and H3K9me2. RdDM loci are locations targeted by the RdDM pathway and are enriched for CHH methylation, while non-RdDM loci do not show these features.

CHAPTER 4

CONCLUSION

The maize B centromere is essential for gene mapping, gene dosage studies, building artificial chromosomes and for studying centromere epigenetics. In our study we developed the first sequence resource for this important model centromere by identifying and mapping 40 CRM2-TD markers to the centromere core. The availability of sequence and PCR markers derived from the B centromere will be essential for interpreting other misdivision derivatives, and additional studies on centromere activation and inactivation. Plans are underway to use the TD markers and 10kb sequence as a scaffold for creating a BAC-based assembly and complete sequence of this centromere, which we hope will be used to further develop the B centromere as a model centromere for plant biology.

Genes in maize are primarily found in islands, surrounded by transposable elements targeted by silencing mechanisms. Although genes are flanked by heterochromatin, they maintain an open chromatin state that allows for transcription. MITEs are also targeted by silencing mechanisms, such as RdDM, but tend to be near genes with relatively high expression levels. We found correlations with the number of MITEs and reduced gene methylation. MITEs also had a chromatin profile that could not be categorized as deep heterochromatin, as they have H3K9me2 levels closer to exons than to heterochromatic regions, yet have high levels of CHH methylation. MITEs maintained their boundary effect in *mop1* mutants even with the loss of siRNA at MITE sites. These data suggest MITEs play an important role as an intermediate sequence between gene and other transposons. The unique sequence context of MITEs, such as low CG-motif content and high A/T levels, and change in methylation over a MITE suggests they act as boundary elements between heterochromatic regions and euchromatic regions.

Transposable elements are more likely to be found in high copy number in the genome if they have either no disadvantage (cause no harm) or because their insertions are beneficial. While there are examples of seemingly benign transposons providing benefit, such by providing promoters for other genes, the fact that a transposon flourishes does not necessarily imply that it positively contributes to the genome. By defining the distribution of transposons, in relation to age and chromatin structure, it is possible to interpret their impact on the genome and understand how they have been successful. CRM2 elements are one example of a successful transposon family, as they have been shown to target CENH3-rich regions where there are few genes, and the negative consequences are minimal. The abundance of CRM2 is a mutually beneficial relationship in which CRM2 can increase in number and CENH3 has more regions for binding and centromere size may fluctuate. MITEs are a second example, as they target upstream and downstream of genes, where the negative consequences are minimal. The abundance of MITEs is mutually beneficial, as the interface between deep heterochromatin and genes exists in a chromatin state that is not entirely inactive, and allows MITEs to maintain a relative high transposition rate. Genes and the genome appear to benefit from the presence of MITEs because they have features that dampen the negative consequences of retroelements that cause the spreading of inactive chromatin marks, thereby isolating and insulating genes from the sea of flanking deep heterochromatin that is typical of large genome species.

Appendix A

Sequence of the B centromere minimal map.

TAGACACGTTAGCACACTGCTACACCCCCCATTGTACACCTGGACCCTCT
CCTTACGCCTATAAAAGCAATGATGTGCACCCGTTGGGTGATTGCCCGAT
CTTTCGATGAGAGGGTGTGGAATAACTCGATTGGGGGAGGAGACGACGTT
CACGGCCCGACTACAGCCTTCCAAAGACGCNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNTAGCCATGCTATGATTGCC
ACTAACCATGCAGAAAATGAGGAAGTCCACGTTGATCCCATCGATGCCGA
TAGGTATGAGAGTCTTGTTGTGCAGCATGTTCTCAGCACACAGGTTGGCC
AGGCCGAAAAAAATCAGCGACACACTCTATTCCATACCAAGGGCGTTGTG
CACGAACGGTCGATTCGCATCATCATCGATAGTGGCAGCTGCAACAATTT
GGCAAGTACAGCTTTGGTACAGAAATTATCTGATGTGCACCCGTTGGGTG
ATTGCCCGATCTTTCGATGAGAGGGTGTGGAATAACTCGATTGGGGAAGG
AGACGACGTTCACGGCCCGACTACAGCCTTCCAAAGACGCNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNTAGCCCGCA
TGACCCGTTTCAACGCATCCCAAGTTTGTGGCATGTTATTAGGATTCTTC
TTACCATGTTCTATCCACCAAACAGAAGCAAATTCAGTAAACTCACTAGT
AGCAGCTCTAACCCGCGCATTCTCAGGAAATTCATGGCATGCAAACTGAT
GTGCACCCGTTGGGTGATTGCCCGATCTTTCGATGAGAGGGTGTGGAATA
ACTCGATTGGGGGAGGAGACGACGTTCACGGCCCGACTACAGCCTTCCAA
AGACGCNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNTAGTCCAGATGGGTCGACGCGAGGCCGAGCGCGAAAGGGGGAA
GTGAGGTGGCCGGAGACCGGCGTGAGAGAGGTGGAAATCCCGCGGCCTTC
GTGTTCGTCCCGCGCCCAGGTCGGGTGCGCTTGCAGTAGGGGGTTACAAG
CGTCCACGCGGGAGAGGGAGCGAGCGGCCTCACGCGAGCGCCTGTCTCGT
CCTCGTCCCCGCGCGGCCAACCCTGATGTGCACCCGTTGGGTGATTGCCC
GATCTTTCGATGAGAGGGTGTGGAATAACTCGATTGGGGGAGGAGACGAC
GTTCACGGCCCGACTACAGCCTTCCAAAGACGCNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNTAGTGCTCGCTCTGCC
AAGGGGGTGTAGAGGTCCATTCGATTTGGTTTAGTTAGTCACCCACCTTG
GGGAGGTGTACTGCATTTGTACGACTGGCGAAACCTAACGAGCAGCTATG
CACTAGGGGAGTCTTTGTAAAGGCTACGTAATGTATCCCTGGCCATTCAC
CTCGATAGTGAAGATCGAGTCTATACAACCCCGGCTGGAGAGGGATCATG
ACTCGTGGGTAAAGTGTGCAACCTCTACAAAGTGTTAGAAACTGGTATAT
CAGCAGAGCTCACGGTTAATGATGTGCACCCATTGGGTAAATGCCCGATC
TTTCGATGAGAGGGTGTGGAACAACTCGATCGGTGGAGGAGACAACGTTC
ATGGCCCGACTACAGCCTTCCAAAGACGCNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNTAGCATCGCATTCAAGCTCA
AACACTTTATTAAAATCAGGCAATTGCTGATGTGCACCCGTTGGGTGATT

```
GCCCGATCTTTCGATGAGAGGGTGTGGAATAACTCGATTGGGGGAGGAGA
CGACGTTCACGGCCCGACTACAGCCTTCCAAAGACGCNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNCTGAGTAGCCAA
CATAAACGCTATAGGACAGGCCCAAACAAGTATTGTTTACAGTCTGGCAA
TGAGCAAATTGTCATGTAATTTCCTTTATTATATGTTTTGTAAAAAATAT
CATTAGTCCCATACTTGTTTTTTGAGTGGCCACAAACTTTCATTGATGTC
CATAACCAAGAAACATTTGAAATAGCACTAAATATCTTTTAAGATAACAA
GCCATAACCAGTATTGTTTATGATGTGCACCCGTTGGGTGATTGCCCGAT
CTTTCGATGAGAGGGTGTGGAATAACTCGATTGGGGAAGGAGACGACGTT
CACGGCCCGACTACAGCCTTCCAAAGACGCNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNTAGTCCTGAACCATAACCA
TAAACCCTAAAGCCCAAACCATAGACCCTAAACCCTTTGTACTTAAGGCT
AACGCCTATGGATTTTTTACCTCCTGACCATTATAAATGTGTACAGATAA
GAAAATGTTTCCCTAAGTCAAAAACATAAGCCATAACCCCAAACACATTA
GCACCAGTCAGTAATGATGTGCACCCGTTGGGTGATTGCCCGATCTTTCG
ATGAGAGGGTGTGGAATAACTCGATTGGGGGAGGAGACGACGTTCACGGC
CCGACTACAGCCTTCCAAAGACGCNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNTAGTCTAGGTCCAAAACTCATGTTT
GGGGTGGTTTCGCGCAATTTCGTTGCCGCACGTCACCCATTCCGAAAACG
GGTATCAGTGATGTGCACCCGTTGGGTGATTGCCCGATCTTTCGATGAGA
GGGTGTGGAATAACTCGATTGGGGGAGGAGACGACGTTCACGGCCCGACT
ACAGCCTTCCAAAGACGCNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNTAGAGATTTTGAGTTTGGGTTTTTGGATTTA
TGGTTTAGGTTTATGGTTCAGGGATTAGGGTTTATGGTTTAGTGTTTAGG
GGTTAGTTGATGTGCACCCGTTGGGTGATTGCCCGATCTTTCGATGAGAG
GGTGTGGAATAACTCGATTGGGGGAGGAGACGACGTTCACGGACCGACTA
CAGCCTTCCAAAGACGCNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNTAGCTACGGGTGCATAGGTTTCACCGAAATCC
AAACCTTCGACTTGGGAGTATCCCTTGGCCACAAGTCGGGCTTTGTTCCT
TGTCACCACACCATGCTCATGTTGCTTGTTGCGGAACACTGATGTGCACC
CGTTGGGTGATTGCCCGATCTTTCGATGAGAGGGTGTGGAATAACTCGAT
TGGGGGAGGAGACGACGTTCACGGCCCGACTACAGCCTTCCAAAGACGCN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
TAGCTGGAACACACAGCTTGTTAGCGCGAAACAGGAACCCATCCTGTATG
TGAAATTTGCCCCATGGTATCCCATTAATACAATGGCCGAAAGCATCTTT
AAAATCAGCATCGTCAACATATTGATCCTTCACAGTGTCCAAACCAAAGA
TTTTAAAATCTAACTGTGACAGCATGGTATAGCGACGAGACAAAGCATCA
GCAATAACATTGTCCTTCCCGTTCTTGTGTTTAATAATGTAAGGAAAGGA
CTCAATGAATTCGTGATGTGCACCCGTTGGGTGATTGCCCGATCTTTCGA
TGAGAGGGTGTGGAATAACTCGATTGGGGGAGGAGACGACGTTCACGGCC
```

CGACTACAGCCTTCCAAAGACGCNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNTAGGATTGTCCTCACCGATAGGTGAA
ACAATCACGCCTCCTGTGCTGCCAGATGGACGGCGGGGTAATACTTCAG
GTCGAGATCGTCAGCTATCCCACCGATGATGTGCACCCGTTGGGTGATTG
CCCGATCTTTCGATGAGAGGGTGTGGAATAACTCGATTGGGGGAGGAGAC
GACGTTCACGGCCCGACTACAGCCTTCCAAAGACGCNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNTAGGTCCAGCACA
GTACCACCGTTGAACGATTATTCAGCTTGCTTTTGTAACTAACGTGGTAT
GATGTGCACCCGTTGGGTGATTGCCCGATCTTTCGATGAGAGGGTGTGGA
ATAACTCGATTGGGGGAGGAGACGACGTTCACGGACCGACTACAGCCTTC
CAAAGACGCNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNTAGCGGATTCGGCATGTTCATTGCGAAAAACAAAGAAATG
GTTTTGGTGGCAAAACTCGTGCTTTGTGTGCACCCCGATACCCGTTTTCC
GAATCGGTAACGTACGGCAACGAAATTGCGCGAAAACACCCCAAAGATGA
GTTTTGGACCTAAACTGGTGGATTCAGCATGTACGTTGCGATAAACGTAG
AAATGGTTCTTGATGTGCACCCGTTGGGTGATTGCCCGATCTTTCGATGA
GAGGGTGTGGAATAACTCGATTGGGGGAGGAGACGACGTTCACGGCCCGA
CTACAGCCTTCCAAAGACGCNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNTAGRGCTGCTAAAACTGCACAACAAGTCA
AATCTGATGAAGCTGAGGTGCCCGATCTTTCGGCGAGTAGAGATAATTCC
GATTTGGCGGAAGATGACCCTTGCGATCCGACTACGACGAGCAAGCCCGA
GGTGATGTGCACCCGTTGGGTGATTGCCCGATCTTTCGATGAGAGGGTGT
GGAATAACTCGATTGGGGGAGGAGACGACGTTCACGGCCCGACTACAGCC
TTCCAAAGACGCNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNGATGAGTCCTGAGTAGGGCTGCTAAAACTGCACAACA
AGTCAAATCGACATCAGCCGCACCTATTAAATCTGAAATCAAGTTGCACT
CTCCTGTTTTACTTGCTACACGTGCTGATTTTGATGATCTCCATGAAGCT
CATATGCCCTGTTATGCACTTGTATGCTCGCGCATGCTTGTTCCGCTTGA
TGATGCACCGTCTTTGGATATACCCCCTGCTGTTGTTAACCTTTTGCAGG
AGTATGCTGATGTTTATCCTACGGACTTACCACCGTGATGTGCACCCGTT
GGGTGATTGCCCGATCTTTCGATGAGAGGGTGTGGAATAACTCGATTGGG
GGAGGAGACGACGTTCACGGCCCGACTACAGCCTTCCAAAGACGCNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNTAGC
TTAGGTCCAAAAATCATGTTTGGGGTGTTTTCGCGCAATTTCGTTGCCGC
ACGTTACCGATTCCGAAAACGGGTATCGGGGTGCACACAAAGCACGAGTT
TTTGCCACCGGAACCATTTCTTCCTTTTTCGAAAAGAACATGCCCAATGA
TGTGCACCCGTTGGGTGATTGCCCGATCTTTCGATGAGAGGGTGTGGAAT
AACTCGATTGGGGGAGGARACAACTTTCACGGCCCGACTACAGCCTTCCA
AAAACNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN

```
NNNNNNTAGTATGTTAGTATCATAATAGTATGAAGATGTGCACCCGTTGG
GTGATTGCCCGATCTTTCGATGAGAGGGTGGGAATAACTCGATTGGGGGA
GAGGACGATGTTCACGGCCCGACTACGGCCTTCCAAAGACGCACCGACTA
CAGCCTTCCAAAGACGNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNTAGTCCTCAGGGCACTATTCACTCGTGTGACT
GACAGCATGAAAACCCCATAGGATCTACCCTAATCTAACCCGCGAGCCCT
GGGTTCTTGTAACAGTACTTGCGGCAAAGATTACGGTGGCGAGACTCACC
GATGGCGAAATTGGTCCGGAGAAGTGGTCGAATGATGTGGGGGAAGTCGC
GGCGGTCATGTCGGTGTGCAGGTCATCACCGGGGATTGATGTGCACCCGT
TAGGTGATTGCCCCATCTTTCGATGAGAGGGTGTGGAATAACTCGATTGG
GGGAGGAGACGACGTTCACGGCCCGACTACAGCCTTCCAAAGACGCANNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNTA
GACCTTTGTGTGCACCTCGATACCCGTTTTCGGAATCGGTAACGTGCGGC
AACGAAATTGCGCGAAACCACCCCATACATGAGTTTTGGACCTAAACTAG
TTGATTCGGCATGTTCGTTGCGAAAAACGTAGAAATGGTTCCGGTGGCAA
AAACTCGTGCTTTGTATGCACCCTGACACCCGTTATCGGAATGGGTGACG
TGCGACAACTGAATTGCGCGAAATGATGTGCACCCGTTGGTTGATTGCCC
GATCTTTCGATGAGAGGGTGTGGAATAACCAGATTGGGGGAGGAGACGAC
GTTCACGGCCCGACTACAGCCTTCCAAAGACGCNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNTAGAGATATAGTGATG
TGCACCCGTTGGGTGATTGCCCGATCTTTCGATGAGAGGGTGTGGAATAA
CTCGATGGGGGGGGAGACCAACATTCTCGGGCCCACTCCACCCTCCCCCA
GACACGCNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNTAGAGAAGGTGCTGCTGGTGTAGATGTGCTCGTGTTGGTGTT
GATGTCCGCATCATCCTCCCCTTCTTGAACTGAAGTTGATGTGCACCCGT
TGGGTGATTGCCCGATCTTTCGATGAGAGGGTGTGGAATAACTCGATTGG
GGGAGGAGACGACGTTCACGGCCCGACTACAGCCTTCCAAAGACCGNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNTAG
ATCATGGATTGGGTAATTCTTCTCATGAACCTTCAGCCGTCGGGACGAGT
AAGCCACAACTCTTCCCTCTTGCATCAACACACATCCCAAATGATGTGCA
CCCGTTGGGTGATTGCCCGATCTTTCGATGAGAGGGTGTGGAATAACTCG
ATTGGGGGAGGAGACGACGTTCACGGCCCGACTACAGCCTTCCAAAGACG
CNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNTAGCAGACTTCGTCGTCGATTGGACCGGGCCAATAACACAGCCGGACC
CGTCCGCAGAGAAGGTTTGGACAATCCACTGCGACGGCGCATGGTGCCAT
GCGGGGGCAGGCGCTGCCGCAGTCATCACCTCACCCGCCAGGGTCAAACA
CAGATACGCGGCACGCTTAAGCTTCGCTCTGGAATCCGACAGATGCACAA
ACAACATAGTAGAATACGGAGCTTGATGTGCACCCGTTGGGTGATTGCCC
GATCTTTCGATGAGAGGGTGTGGAATAACTCGATNGGGGGAGGAGACGAC
GTTCACGGCCCGACTACAGCCTTCCAAAGACGCNNNNNNNNNNNNNNNNN
```

NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNTAGCATGAGTCGGGCG
AGATCTTTAGTGGCGTCTCGGGGCGTCAGTGGGGGAATCCTTCTTTAAAA
AGGGGTTCATCCCTTGAGTAGCAGCCATGCCTTGCTTCTTGATGTGCACC
CGTTGGGTGATTGCCCGATCTTTCGATGAGAGGGTGTGGAATAACTCGAT
TGGGGGAGGAGACGACGTTCACGGCCCGACTACAGCCTTCCAAAGNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNTAGC
GACGTCTTCGGCACCAGATGACTTAGTCGAATTGGTCCCTCGGAGGGCAA
ATGTTGGGGCGAAGGCGAAGACGCTACCCTTCGCTCGAAGTCTTCGCCAA
TCTCCCTGCACCAACGGAGGCGAAACGACCAACGGGTTCCACCCTTCGTC
CACTGCGTTGCAAGATGAAGGCTTACAATGAGGTCGCCCCGTCCCTCGCC
CTCGTCCATCCCGGAGGCCCACGGGGAATTCTGATGTGCACCCGTTGGGT
GATTGCCCGATCTTTCGATGAGAGGGTGTGGAATAACTCGATTGGTGGAG
GAGACGACGTTCACGGCCCGACTACAGCCTTCCAAAGACGCNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNTAGCTCAT
GAACCACACTTTCCCCACCTTCCGGCATTTGATAGACAGAGCAATCATGA
CTGAAAGGAAACGTCAGGTGATGTGCACCCGTTGGGTGATTGCCCGATCT
TTCGATGAGAGGGTGTGGAATAACTCGATTGGGGGAGGAGACGACGTTCA
CGGCCCGACTACAGCCTTCCAAAGANNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNTAGTTCTGAACCATAACCATAAAC
CCTAAAGCCCAAACCATAGACCCTAAACCCTTTGTACTTAAGGCTAACGC
CTATGGATTTTTTACCTCCTGACCATAATAAATGTCTACAGATAAGAAAA
TGTTTCCCTAACTCAAAAACCTAAGCCATAACCCCAAACACTTTAGCACC
AGTCACTAATGATGTGCACCCGTTGGGTGATTGCCCGATCTTTCGATGAG
AGGGTGTGGAATAACTCGATTGGGGGAGGAGACGACGTTCACGGCCCGAC
TACAGCCTTCCAAAGACGCNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNTAGGGGCTGCGCCTCCTTTGTTTTCGAGTT
CTGTTTTGATGTGCACCCGTTGGGTGATACTGATGTGCACCCGTTGGGTG
ATTGCCCGATCTTTCGATGAGAGGGTGTGGAATAACTCGATTGGGGGAGG
AGACGACGTTCACGGCCCGACTACAGCCTTCCAAAGACGCNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNTAGGGGATA
TTCCTCTTTGACGAATAGGGGTGGACGAGTTTCAGAGAAGTCCAGATAGG
TGGTCTCACGGGGACCTTGTGGATAACCTCTTCCCGCCTGAGGCTGGAAG
TGATCTGACCCGCCATTTCCCTAAGGAAGCGGGTATTATCGTCGGTGGCA
TTCATCAAGGCGGCTGATGTGCACCCGTTGGGTGATTGCCCGATCTTTCG
ATGAGAGGGTGTGGAATAACTCGATTGGGAGGAGACGACGTTCACGGCCC
GACTACAGCCTCCAAAGACGCNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNTAGTACAACGTTAGTGATGTGCACCCGT
TGGGTGATTGCCCGATCTTTCGATGAGAGGGTGTGGAATAACTCGATTGG
GGGAGGAGACGACGTTCACGGCCCGACTACAGCCTTCCAAAGACGCNNNN

```
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNTAG
TATAGTAAGAGATGATGATATGGAGTATTCTGATGTGCACCCGTTGGGTG
ATTGCCCGATCTTTCGATGAGAGGGTGTGGAATAACTCGATTGGGGGAGG
AGACGACGTTCACGGCCCGACTACAGCCTTCCAAAGACGCNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNTAGTCCTTT
GTGTGCACCTCGATATGCGTTTTAGGAATGGGTAATCTGGGGCAATGAAA
ATGCGCGAAACCACCCCATACATGATGTTTGGACCTAAAGTAGTTGATTC
GGCAGCTTAGTTGCGAAAAACGTAGAAATGGTTGCGGTGGCAAAAACTCG
TGAAAATAGGCACCCTGACACCCGTTATCGGAATGGGTGACGTGCGACAA
ATCACCAGTGAGAAATGATGTGCACCCGTTGGGTGATTGCCCGATCTTTC
GATGAGAGGGTGTGGAATAACTCGATTGGGGGAGGAGACGACGTTCACGG
CCCGACTACAGCCTTCCAAAGACGCNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNTAGAATTTGAGAAGTCAACAAGGC
AATAAAAGCTGCTCCACTTGATGGGATTTTGTAACTGCAGCAAGTGCAAG
TTGAAGATGATGTGCACCCGTTGGGTGATTGCCCGATCTTTCGATGAGAG
GGTGTGGAATAACTCGATTGGGGGAGGAGACGACGTTCACGGCCCGACTA
CAGCCTTCCAAAGANNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNTAGCAGCCTCTTGACTTCCAAACACAGCAATAGGA
CCTTGGTCCGATGGTTTCTTCATGCACAGATATACTGGATGCAATAGGAA
ACCCTTTGATAACACATGATATGTCCATGGCGGCATTATGGGGGTAATCC
TTGAGTTGAGGGTCTGATGTGCACCCGTTGGGTGATTGCCCGATCTTTCG
ATGAGAGGGTGTGGAATAACTCGATTGGGGGAGGAGACGACGTTCACGGC
CCGACTACAGCCTTCCAAAGACGCNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNTAGATGAGAAGCATCATCAAAATCA
AAATTACCATGATGGATATCATCATACCATTCAACATCGTTGTACGAATT
GTCTTTATGAAGCTCTCTATGTTGAGGCCTTCAGTCTTGGTCATCTTGGG
TGAGATGACACATTTCCTCGGCGGCCTCATCAATCTTCCTCTGAAGATCT
GATGTGCACCCGTTGGGTGATTGCCCGATCTTTCGATGAGAGGGTGTGGA
ATAACTCGATTGGGGGAGGAGACGACGTTCACGGCCCGACTACAGCCTCC
AAAGACGCNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNTAGAGTTGAGCGTGAGGGAGACGAAGCCGAAGGAGCGGAGA
CTGTTCCTATTTCTGAAGTGATGTGCACCCGTTGGGTGATTGCCCGATCT
TTCGATGAGAGGGTGTGGAATAACTCGATTGGGGGAGGAGACGACGTTCA
CGGCCCGACTACAGCCTTCCAAAGACGCNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNTAGGTCAGCAGAGAGAAATTT
ACCATTCTTCGAAGTGTTGAAGTCAGTGGAAGTTTTTCAATGGGGACCAG
TTCAGTAGAAAGCTTTCGACGAGCTGAAGCAGTATCTGATAGACCTAACA
TGATGTGCACCCGTTGGGTGATTGCCCGATCTTTCGATGAGAGGGTGTGG
AATAACTCGATTGGGGGAGGAGACGACGTTCACGGCCCGACTACAGCCTT
```

CCCAAGACGCNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNGACGATGANTCTGAGTAGAGGATTCGGCATGTTCATTGC
GAAAAACAAAGAAATGGTTCTGGTGGCAAAAACTCGTGCTTTGTGTGCAC
CCCGATACCCGTTTTCCGAATCGGTAACGTACGGCAACGAAGATGCGGGA
WAACACCCCRRAGAAGARTNYTGGACCWARASTGAYGGATTCAGCATGTA
CGNTTGNCRATAAACGTAGAAATGGTTCTTGATGTGCACCCGTTGGGTGA
TTGCCCGATCTTTCGATGAGAGGGTGTGGAATAACTCGATTGGGGGAGGA
GACGACGTTCACGGCCCGACTACAGCCTTCCAAAGACGC

Appendix B

Python script: “Methylation Distribution Meta-chromosome”:
Python 2.7 script to make meta-features, such as a meta-gene plots.

```
#This program maps repeats to whole chromosome with methylation % within bins
#It counts CG,CHG,CHH methylation of TEs or sequence of interest (SOI) position within a bin
AND flanking regions of the TE
#ONLY of SOI larger than 101 base pairs in length
####################################################################################
############################
#USER INPUT
####********************#######
#NEEDED IMPROVEMENTS#
#Can't currently count reads which are bigger than the sequence of interest
#Main problem is when counting reads that fall near right side and only counting
#portion with SOI. The problem is that I take TE[i][2] - read length which if the read
#length is longer than the length of the TE, will count reads that are actually outside the SOI

#For each chromosome add brackets in Chrm_list and the length of the chromosome
#e.g. [[1,1233343], [2,12343234],[3,98657888]]
Chrm_list = [[Chrnum,length],[X]]

#this determines a bin_size
#if you want single basepair resultion then it
#If looking looking a single TE, then bin size can't be larger than the TE length
#Replace X with bin number
bin_size = X
#USER INPUT
#FILE INPUT of position with methylation info.
methyl_file = open('Bisulfite treated reads')
#This automatically opens the correct file, if the name is in the same format
#Needs to be list of C-some and positions - ex: 2,203000,302000

#INPUT FILE is output file from
#Repeatmasker converter v2.2 BLAST output converter
TE_list = open (RepeatMasker output file)
#READ SIZE
#This section is if you want to look at a single TE, and can look at flanking region's methylation
left_flank = 0
right_flank = 0
####################################################################################
#############################
###Create list with chromosome, start_postition and end_position within script
###Still have this to calculation percentage of TEs in a bin
TE = []
```

```
for line in TE_list:
    TE_input = [0,0,0,0]
    cols = line.split(',')
    '''cols3 = int(cols[3])'''

    #This will ignore all TEs that were found to have both flanks near other TEs
        ###### need to add a fourth if you want to add orientation

    TE_input[0] = int(cols[0])
    TE_input[1] = int(cols[1])
    TE_input[2] = int(cols[2])
    '''TE_input[3] = int(cols[3])'''
    TE.append(TE_input)
#Make list for each chromosome
TE_Chrm_list = [[] for i in range(len(Chrm_list))]
#Places position in list based on chromosome number
for i in range(len(TE)):
    chromosome_number = TE[i][0]
    TE_Chrm_list[chromosome_number - 1].append(TE[i])

#This will split up the TE list into ten percent of whatever length the chromosome_length is
TE_per = [[[] for number in range(10)] for number_of_chromosomes in range(len(Chrm_list))]

for x in range(len(TE_Chrm_list)):
    for i in range(len(TE_Chrm_list[x])):
        chromosome_number = TE_Chrm_list[x][i][0]

        if TE_Chrm_list[x][i][1] <= (Chrm_list[chromosome_number - 1][1] * 0.10):
            TE_per[chromosome_number-1][0].append(TE_Chrm_list[x][i])
        elif (Chrm_list[chromosome_number-1][1] * 0.1) < TE_Chrm_list[x][i][1] <=
(Chrm_list[chromosome_number-1][1] * 0.2):
            TE_per[chromosome_number-1][1].append(TE_Chrm_list[x][i])
        elif (Chrm_list[chromosome_number-1][1] * 0.2) < TE_Chrm_list[x][i][1] <=
(Chrm_list[chromosome_number-1][1] * 0.3):
            TE_per[chromosome_number-1][2].append(TE_Chrm_list[x][i])
        elif (Chrm_list[chromosome_number-1][1] * 0.3) < TE_Chrm_list[x][i][1] <=
(Chrm_list[chromosome_number-1][1] * 0.4):
            TE_per[chromosome_number-1][3].append(TE_Chrm_list[x][i])
        elif (Chrm_list[chromosome_number-1][1] * 0.4) < TE_Chrm_list[x][i][1] <=
(Chrm_list[chromosome_number-1][1] * 0.5):
            TE_per[chromosome_number-1][4].append(TE_Chrm_list[x][i])
        elif (Chrm_list[chromosome_number-1][1] * 0.5) < TE_Chrm_list[x][i][1] <=
(Chrm_list[chromosome_number-1][1] * 0.6):
            TE_per[chromosome_number-1][5].append(TE_Chrm_list[x][i])
        elif (Chrm_list[chromosome_number-1][1] * 0.6) < TE_Chrm_list[x][i][1] <=
(Chrm_list[chromosome_number-1][1] * 0.7):
```

```
            TE_per[chromosome_number-1][6].append(TE_Chrm_list[x][i])
        elif (Chrm_list[chromosome_number-1][1] * 0.7) < TE_Chrm_list[x][i][1] <=
(Chrm_list[chromosome_number-1][1] * 0.8):
            TE_per[chromosome_number-1][7].append(TE_Chrm_list[x][i])
        elif (Chrm_list[chromosome_number-1][1] * 0.8) < TE_Chrm_list[x][i][1] <=
(Chrm_list[chromosome_number-1][1] * 0.9):
            TE_per[chromosome_number-1][8].append(TE_Chrm_list[x][i])
        elif (Chrm_list[chromosome_number-1][1] * 0.9) < TE_Chrm_list[x][i][1] <=
(Chrm_list[chromosome_number-1][1] * 1.0):
            TE_per[chromosome_number-1][9].append(TE_Chrm_list[x][i])

counter = [[],[],[],[],[],[],[],[],[],[]]
Left_counter = [[],[],[],[],[],[],[],[],[],[]]
Right_counter = [[],[],[],[],[],[],[],[],[],[]]
def counter_bins(Counter_type, bin_size):
    for z in range(len(Chrm_list)):
        chromosome_number = Chrm_list[z][0]
        Chrm_bin_calc = ((Chrm_list[z][1])//(bin_size)+1)
        y = Chrm_list[z][0]
        for i in range(Chrm_bin_calc):
            Counter_type[y-
1].append([0,0,0,0,0,0,0,0.0000,0.0000,0.0000,0.0000,0.0000,0.0000,0,0.0000,0])

counter_bins(counter,bin_size)
counter_bins(Left_counter,bin_size)
counter_bins(Right_counter,bin_size)

######################################################################################
#####################################
#Counter, TE the variable that changes based on the position of the read
#!!!!!!!!!!!!should add read cut off paramaters for the flanking regions
######################################################################################
#####################################
def counter_TE(counter,Left_counter, Right_counter, TE, meth):
    i = 0
    while i < len(TE):
        #Main SOI
            #Read position checker
            #This section is to compare Bisulfite read coordinates to the coordinates of input, there
are three sections:
            #the fist checks for positions within the BLAST start and end(-49) positions,
            #the second checks for read within 49 basepairs to the end (needs at least one base pair
outside sequence),
            #the third part checks for reads that are 49 base pairs before the beginning sequence
        #Ignores reads larger than the length of the TE
        if TE[i][2] - TE[i][1] > read_length:
```

```
    #Start within TE
    if (TE[i][1]) <= pos < (TE[i][2] - (read_length - 2)):  ##To see if read falls within the
position of each TE.
                                                          #If the read does not fall within the first TE positions,
        k = (pos) #starting point of read                 #then it will go on to check if found at other
positions

        j =(k)//bin_size #converts position to bin
        counter[j][6] += read_length#Any read/positions that meets the above criteria is
counted.
        for x in meth:
            if x == 'x':
                counter[j][0] += 1
            elif x == 'y':
                counter[j][1] += 1
            elif x == 'z':
                counter[j][2] += 1
            elif x == 'X':
                counter[j][3] += 1
            elif x == 'Y':
                counter[j][4] += 1
            elif x == 'Z':
                counter[j][5] += 1
        i = len(TE)
    #Right side within TE and counting read if goes into flank
    #First right side within TE
    elif (TE[i][2] - (read_length - 2)) <= pos <= (TE[i][2]):
        if strand == 1:
            meth = meth[::-1]
        truncated_read = ((TE[i][2] - pos) +1) #To include reads pos that equal TE[i][2]
                                              #must add 1 to each position
        meth = meth[:truncated_read]  #read until this position
        k = (pos)
        j =(k)//bin_size #converts position to bin
        counter[j][6] += truncated_read
        #Count for SOI portion of read
        for x in meth:

            if x == 'x':
                counter[j][0] += 1
            elif x == 'y':
                counter[j][1] += 1
            elif x == 'z':
                counter[j][2] += 1
            elif x == 'X':
                counter[j][3] += 1
            elif x == 'Y':
                counter[j][4] += 1
            elif x == 'Z':
                counter[j][5] += 1
        #Count for flanking portion of read
        col6_right = meth[truncated_read:]
        Right_counter[j][6] += read_length - truncated_read
        for x in col6_right:  #this goes through the read information and counts the number
nonmethylated and methylated CG, CHG, and CHH
            if x == 'x':
                Right_counter[j][0] += 1  #For example this counts the number of lower case xs
and feeds them to the counter. i = given TE (defined at the beginning of "Read position checker
section",
                                    #and [0] indicates the first position of the counter list
            elif x == 'y':
                Right_counter[j][1] += 1
            elif x == 'z':
                Right_counter[j][2] += 1

            elif x == 'X':
                Right_counter[j][3] += 1

            elif x == 'Y':
                Right_counter[j][4] += 1

            elif x == 'Z':
                Right_counter[j][5] += 1


        i = len(TE) #This will break the loop

    #To get reads that start before the TE start position, spilts flank and SOI

    #First, cut off read belonging to SOI and count for within
        #****Need to make sure if SOI is shorter than read, part of sequence may be part of
Right FLANK
    elif (TE[i][1] - (read_length - 1)) <= pos < (TE[i][1]):#This is to check to see if read falls
within the position of each TE. If the read does not fall within the first TE positions,

        if strand == 1:
            meth = meth[::-1]

        k = (pos)

        truncated_read = (TE[i][1] - pos)
        #Need something here to make sure SOI of read is counted
        meth = meth[truncated_read:]  ###If read is larger than TE, then can add
"(truncated_read+(TE[i][2] - TE[i][1]))" to other side of : to ensure it is cut of at end of read
        j =(k)//bin_size #converts position to bin
        counter[j][6] += read_length - truncated_read  #This counts only the part of the read
that is used


        for x in meth:

          if x == 'x':
            counter[j][0] += 1

          elif x == 'y':
            counter[j][1] += 1

          elif x == 'z':
            counter[j][2] += 1

          elif x == 'X':
            counter[j][3] += 1


          elif x == 'Y':
            counter[j][4] += 1


          elif x == 'Z':
            counter[j][5] += 1


      #Second is for counting part of read in flanking region
        meth_flank = meth[:truncated_read]
        Left_counter[j][6] += truncated_read  #This counts only the part of the read that is
used in flank
                                #If at least one base pair is within SOI, then we will only
counted the position out side the SOI

        #Column 6 is the column with the read information
        for x in meth_flank:  #this goes through the read information and counts the number
nonmethylated and methylated CG, CHG, and CHH
          if x == 'x':
            Left_counter[j][0] += 1  #For example this counts the number of lower case xs
and feeds them to the counter. i = given TE (defined at the beginning of "Read position checker
section",
                              #and [0] indicates the first position of the counter list
              elif x == 'y':
                 Left_counter[j][1] += 1
              elif x == 'z':
                 Left_counter[j][2] += 1

              elif x == 'X':
                 Left_counter[j][3] += 1

              elif x == 'Y':
                 Left_counter[j][4] += 1

              elif x == 'Z':
                 Left_counter[j][5] += 1

           i = len(TE)
            ####Need to make it so if flanks not used then this part of the program is not used
           ####May be better to use non flanking counters and regions if not using the flank
option
        #LEFT FLANKING REGION OF Main SOI

        elif (TE[i][1] - left_flank) <= pos < (TE[i][1] - (read_length - 1)): #First part, check
position of read between flank start to the end
                                                    #of the start of the soi MINUS the length of the
read size minus one bp, this is
                                                    #so that we do NOT count reads that could
possibly have one bp in the
                                                      #main sequence yet (notice less than sign).
This will be done in the next counter

           k = (pos)
           j =(k)//bin_size
           Left_counter[j][6] += read_length

          #Column 6 is the column with the read information
           for x in meth:  #this goes through the read information and counts the number
nonmethylated and methylated CG, CHG, and CHH
              if x == 'x':
                 Left_counter[j][0] += 1  #For example this counts the number of lower case xs
and feeds them to the counter. i = given TE (defined at the beginning of "Read position checker
section",
                              #and [0] indicates the first position of the counter list
              elif x == 'y':
                 Left_counter[j][1] += 1
              elif x == 'z':
                 Left_counter[j][2] += 1
```

```
                elif x == 'X':
                    Left_counter[j][3] += 1

                elif x == 'Y':
                    Left_counter[j][4] += 1

                elif x == 'Z':
                    Left_counter[j][5] += 1

            i = len(TE) #This will break the loop


        #Reads left side of left flank
            #Counts read data inside flanking region with a read starting postion Left of the LEFT flanking regions.

        elif (TE[i][1] - (left_flank + (read_length - 1))) <= pos < (TE[i][1] - (left_flank)):



            if strand == 1:
                meth = meth[::-1]

            k = (pos)

            truncated_read = ((TE[i][1] - left_flank) - pos)

            meth = meth[truncated_read:] #This will cut off the read so that only the part within the flank will be counted
            j =(k)//bin_size #converts position to bin
            Left_counter[j][6] += read_length - truncated_read  #This counts only the part of the read that is used
           #Column 6 is the column with the retruncated_readad information
            for x in meth:  #this goes through the read information and counts the number nonmethylated and methylated CG, CHG, and CHH
                if x == 'x':
                    Left_counter[j][0] += 1  #For example this counts the number of lower case xs and feeds them to the counter. i = given TE (defined at the beginning of "Read position checker section",
                                        #and [0] indicates the first position of the counter list
                elif x == 'y':
                    Left_counter[j][1] += 1
                elif x == 'z':
                    Left_counter[j][2] += 1
                    elif x == 'X':
                        Left_counter[j][3] += 1

                    elif x == 'Y':
                        Left_counter[j][4] += 1

                    elif x == 'Z':
                        Left_counter[j][5] += 1
                i = len(TE) #This will break the loop


            #Right Flanking Region

            elif (TE[i][2]) < pos < ((TE[i][2]) + (right_flank) - (read_length - 2)):   #Counting reads
in righ flank unless

                k = (pos)
                j =(k)//bin_size
                Right_counter[j][6] += read_length

                #Column 6 is the column with the read information
                for x in meth:  #this goes through the read information and counts the number
nonmethylated and methylated CG, CHG, and CHH
                    if x == 'x':
                        Right_counter[j][0] += 1  #For example this counts the number of lower case xs
and feeds them to the counter. i = given TE (defined at the beginning of "Read position checker
section",
                                            #and [0] indicates the first position of the counter list
                    elif x == 'y':
                        Right_counter[j][1] += 1
                    elif x == 'z':
                        Right_counter[j][2] += 1

                    elif x == 'X':
                        Right_counter[j][3] += 1

                    elif x == 'Y':
                        Right_counter[j][4] += 1

                    elif x == 'Z':
                        Right_counter[j][5] += 1

                i = len(TE) #This will break the loop
            elif ((TE[i][2]) + (right_flank) - (read_length - 2)) <= pos <= ((TE[i][2]) + (right_flank)):
#This will be for cutting off reads that overlap with flank on right side of TE


                if strand == 1:
                    meth = meth[::-1]
                truncated_read = (((TE[i][2] + right_flank) - pos) +1) #To include reads pos that equal
TE[i][2]

                meth = meth[:truncated_read]  #read until this position

                k = (pos)

                j =(k)//bin_size #converts position to bin
                Right_counter[j][6] += truncated_read
                for x in meth:  #this goes through the read information and counts the number
nonmethylated and methylated CG, CHG, and CHH
                    if x == 'x':
                        Right_counter[j][0] += 1  #For example this counts the number of lower case xs
and feeds them to the counter. i = given TE (defined at the beginning of "Read position checker
section",
                                            #and [0] indicates the first position of the counter list
                    elif x == 'y':
                        Right_counter[j][1] += 1
                    elif x == 'z':
                        Right_counter[j][2] += 1

                    elif x == 'X':
                        Right_counter[j][3] += 1

                    elif x == 'Y':
                        Right_counter[j][4] += 1

                    elif x == 'Z':
                        Right_counter[j][5] += 1

                i = len(TE) #This will break the loop

            else:
                i +=1

        else:
            i +=1
```

```
################################################################
#BRING IN BS READS
################################################################
for line in methyl_file: #Bring in file with BS reads
    cols = line.split('\t')

    #Chromosome Conversion for the read information
    #Extract chromosome number and check to see if read on chromosome 1, if so, next step
    chrom_col = cols[3]
    BS_chrom_num = chrom_col[2:4]
    if BS_chrom_num == '02':
        chrom_num = 1
    elif BS_chrom_num == '05':
        chrom_num = 4
    elif BS_chrom_num == '03':
        chrom_num = 2
    elif BS_chrom_num == '04':
        chrom_num = 3
    elif BS_chrom_num == '06':
        chrom_num = 5
    elif BS_chrom_num == '08':
        chrom_num = 7
    elif BS_chrom_num == '09':
        chrom_num = 8
    elif BS_chrom_num == '07':
        chrom_num = 6
    elif BS_chrom_num == '10':
        chrom_num = 9
    elif BS_chrom_num == '01':
        chrom_num = 10
    pos = int(chrom_col[5:]) #this variable is the starting position number of the read

    col6 = cols[6].strip()
    read_length = len(col6)

    if cols[2][0] == '+':
        strand = 0
    else:
        strand = 1

    ###CHECK which TE section it will check through
    if pos <= (Chrm_list[chrom_num - 1][1] * 0.10):
        counter_TE(counter[chrom_num - 1],Left_counter[chrom_num -
1],Right_counter[chrom_num - 1], TE_per[chrom_num -1][0],col6)
    elif (Chrm_list[chrom_num - 1][1] * 0.1) < pos <= (Chrm_list[chrom_num - 1][1] * 0.20):
        counter_TE(counter[chrom_num - 1],Left_counter[chrom_num -
1],Right_counter[chrom_num - 1], TE_per[chrom_num -1][1],col6)
    elif (Chrm_list[chrom_num - 1][1] * 0.2) < pos <= (Chrm_list[chrom_num - 1][1] * 0.30):
        counter_TE(counter[chrom_num - 1],Left_counter[chrom_num -
1],Right_counter[chrom_num - 1], TE_per[chrom_num -1][2],col6)
    elif (Chrm_list[chrom_num - 1][1] * 0.3) < pos <= (Chrm_list[chrom_num - 1][1] * 0.40):
        counter_TE(counter[chrom_num - 1],Left_counter[chrom_num -
1],Right_counter[chrom_num - 1], TE_per[chrom_num -1][3],col6)
    elif (Chrm_list[chrom_num - 1][1] * 0.4) < pos <= (Chrm_list[chrom_num - 1][1] * 0.50):
        counter_TE(counter[chrom_num - 1],Left_counter[chrom_num -
1],Right_counter[chrom_num - 1], TE_per[chrom_num -1][4],col6)
    elif (Chrm_list[chrom_num - 1][1] * 0.5) < pos <= (Chrm_list[chrom_num - 1][1] * 0.60):
        counter_TE(counter[chrom_num - 1],Left_counter[chrom_num -
1],Right_counter[chrom_num - 1], TE_per[chrom_num -1][5],col6)
    elif (Chrm_list[chrom_num - 1][1] * 0.6) < pos <= (Chrm_list[chrom_num - 1][1] * 0.70):
        counter_TE(counter[chrom_num - 1],Left_counter[chrom_num -
1],Right_counter[chrom_num - 1], TE_per[chrom_num -1][6],col6)
    elif (Chrm_list[chrom_num - 1][1] * 0.7) < pos <= (Chrm_list[chrom_num - 1][1] * 0.80):
        counter_TE(counter[chrom_num - 1],Left_counter[chrom_num -
1],Right_counter[chrom_num - 1], TE_per[chrom_num -1][7],col6)
    elif (Chrm_list[chrom_num - 1][1] * 0.8) < pos <= (Chrm_list[chrom_num - 1][1] * 0.90):
        counter_TE(counter[chrom_num - 1],Left_counter[chrom_num -
1],Right_counter[chrom_num - 1], TE_per[chrom_num -1][8],col6)
    elif (Chrm_list[chrom_num - 1][1] * 0.9) < pos <= (Chrm_list[chrom_num - 1][1] * 1.0):
        counter_TE(counter[chrom_num - 1],Left_counter[chrom_num -
1],Right_counter[chrom_num - 1], TE_per[chrom_num -1][9],col6)
 ################
#TE COUNTER
#Goes through TE list and counts how many repeats in each bin and length
#################
for i in range(len(TE)):
    y= TE[i][0] - 1
    x = (TE[i][1]) //bin_size
    counter[y][x][13] += (TE[i][2] - TE[i][1]) # TE length, to later calculate total percent length in
bin
    counter[y][x][15] += 1

##########################################################################
#Calculation def
def calc(TE, counter):
    for z in range(len(Chrm_list)):

        #X calculations
        for i in range(len(counter[z])):
            #####********#####Can delete each for i range(len.. and calculate all methyl calculations for each postion
            try:
                counter[z][i][7] = ((float(counter[z][i][3])/float(counter[z][i][3] + counter[z][i][0])) * float(100))

            except ZeroDivisionError:
                counter[z][i][7] = ' '

            try:
                counter[z][i][8] = ((float(counter[z][i][3])/float(counter[z][i][6])) * float(100))

            except ZeroDivisionError:
                counter[z][i][8] = ' '

            try:
                counter[z][i][9] = ((float(counter[z][i][4])/float(counter[z][i][4] + counter[z][i][1])) * float(100))

            except ZeroDivisionError:
                counter[z][i][9] = ' '
            try:
                counter[z][i][10] = ((float(counter[z][i][4])/float(counter[z][i][6])) * float(100))

            except ZeroDivisionError:
                counter[z][i][10] = ' '

            try:
                counter[z][i][11] = ((float(counter[z][i][5])/float(counter[z][i][5] + counter[z][i][2])) * float(100))

            except ZeroDivisionError:
                counter[z][i][11] = ' '

            try:
                counter[z][i][12] = ((float(counter[z][i][5])/float((counter[z][i][6])) * float(100)))

            except ZeroDivisionError:
                counter[z][i][12] = ' '
################################################################################
##CALCULATIONS
################################################################################
#Main Counter
calc(TE, counter)
for z in range(len(Chrm_list)):
    for i in range(len(counter[z])):
        try:
            counter[z][i][14] = str((float(counter[z][i][13])/float(bin_size)) * float(100))
        except ZeroDivisionError:
            counter[z][i][14] = ' '
#LEFT FLANK
calc(TE, Left_counter)
#RIGHT FLANK
calc(TE, Right_counter)
#Enumerate is like an iterator, it will go through the reads and
                #and allows you to differentiate between the position of a character within the
sequence and the character
                #z is the position in the string and a is the character

'''for i in range(len(counter)):
    if Left_counter[i][7] or Right_counter[i][7] != ' ':
        Right_counter[i][13] = ((float(Left_counter[i][7]) + float(Right_counter[i][7]))/float(2))
    if Left_counter[i][9] or Right_counter[i][9] != ' ':
        Right_counter[i][14] = ((float(Left_counter[i][9]) + float(Right_counter[i][9]))/float(2))
    if Left_counter[i][11] or Right_counter[i][11] != ' ':
        Right_counter[i][15] = ((float(Left_counter[i][11]) + float(Right_counter[i][11]))/float(2))
'''
#OUTPUT FILE
#***********OUTPUT PRINT FILE FOR EACH CHROMOSOME FOR JONATAHN'S
PROGRAM

for i in range(len(Chrm_list)):
    out_file = open
('/rcc_home/krdlab/gent/Output/TE_methylation//Tourist_repeats_methyl_chrom_flank_' +
str(left_flank) + '_JUNK_' + str(bin_size) + '_chrm_' + str(i + 1) + '.txt.', 'w')
    out_file.write('Bin_size_' + str(bin_size) + 'left_flank_' + str(left_flank) + 'right_flank_' +
str(right_flank)+ 'chromosome_number_' + str(i + 1))
    out_file.write('\t' + 'BIN\tx count\ty count\tz count\tX count\tY count\tZ count\tread total\tCG
(%CG)\tCG (%N)\tCHG (%CHG)\tCHG (%N)\tCHH (%CHH)\tCHH (%N)\tTotal nucleotide
TE\tTE%\tTE total\tLeft_Flank CG (CG%)\tLeft_Flank CHG (CHG%)\tLeft_Flank CHH
(CHH%)\tRight_Flank CG (CG%)\tRight_Flank CHG (CHG%)\tRight_Flank CHH
(CHH%)\tAverage Flank CG (CG%)\tAverage Flank CHG (CHG%)\tAverage Flank CHH
(CHH%)\n')

    for z,a in enumerate(counter[i]):
        out_file.write('\t' + str(z)
                        + '\t' + str(counter[i][z][0])
                        + '\t' + str(counter[i][z][1])
                        + '\t' + str(counter[i][z][2])
                        + '\t' + str(counter[i][z][3])
                        + '\t' + str(counter[i][z][4])
                        + '\t' + str(counter[i][z][5])
```

```
                    + '\t' + str(counter[i][z][6])
                    + '\t' + str(counter[i][z][7])
                    + '\t' + str(counter[i][z][8])
                    + '\t' + str(counter[i][z][9])
                    + '\t' + str(counter[i][z][10])
                    + '\t' + str(counter[i][z][11])
                    + '\t' + str(counter[i][z][12])
                    + '\t' + str(counter[i][z][13])
                    + '\t' + str(counter[i][z][14])
                    + '\t' + str(counter[i][z][15])
                    + '\t' + str(Left_counter[i][z][7])
                    + '\t' + str(Left_counter[i][z][9])
                    + '\t' + str(Left_counter[i][z][11])
                    + '\t' + str(Right_counter[i][z][7])
                    + '\t' + str(Right_counter[i][z][9])
                    + '\t' + str(Right_counter[i][z][11])
                    + '\n')
out_file.close()
TE_list.close()
methyl_file.close()
```

Appendix C

Pythonscript: “Methylation Distribution Meta-gene”:
Python2.7 script of the program which calculate methylation levels over genes

```
#FeatureMethylBin - This program takes the output file of BS_Extractor and creates a metafeature graph while considering the features orientation
#Cannot skipp any BS file lines
###Reads must go through extractor AND BE WITHIN A FEATURE...otherwise an error occurs..Reads will be too large to fit within feature length

#USER INPUT
import sys
FILENAME = sys.argv[1]
#INITIALIZE VARIABLES
#BIN_size - there are some limits to what bin number can be used...if remainder, must add 1 to master_counts array
#Can put bin_size parameters here to automatically change master_count array size
bin_size = int(sys.argv[2])

###This is temporarily included to keep track of program speed
from datetime import datetime
startTime = datetime.now()
from itertools import repeat

#I/O
in_file = open('/panfs/pstor.storage/grphomes/krdlab/nellis/Output/' + str(FILENAME) + '.extract', 'r') #open input file
GFF_file = open('/panfs/pstor.storage/grphomes/krdlab/nellis/Jobs/' + str(FILENAME), 'r')
out_file = open('/panfs/pstor.storage/grphomes/krdlab/nellis/Output/' + str(FILENAME) + '_binsize_' + str(bin_size) + '_MethBin.txt', 'w') #open output file

#This takes the info from the first line
with open(gff_path, 'r') as f:
    first_line = f.readline()
    GFF_cols = first_line.split('\t')
    GFF_orient = (GFF_cols[6])
    GFF_start = int(GFF_cols[3])
    GFF_end = int(GFF_cols[4])
    feature_size = (GFF_end - GFF_start) + 1
    GFF_chr = int(GFF_cols[0])
    bin_size_element = feature_size/(100/bin_size)

##############################################################################################################
#Have to multiply by a whole number, so if bin size is not evenly divided then will need
    #to adjust this statement to make sure there is a bin present for the remainder
```

```
methyl_counts = [[0 for c in range(7)] for pos in range(feature_size)] #This keeps track of
methylation info for individual nucleotides
master_counts = [[0,0,0,0,0,0] for i in range(int(100/bin_size))] #mC % is then stored in the
master_counts array for each Bin
##############################################################################
######################
#keeps track of line number in GFF file, starts at line 0 in GFF file
k = 0
GFF_file.seek(0)
lines = GFF_file.readlines()
File_length = (len(lines))
##############################################################################
#################################
#DEFINED FUNCTIONS: Each time a BS read no longer belongs to the current GFF feature,
the data stored will be processed and stored. Then the information of the next line in the GFF file
will be used
##############################################################################
#################################
####Methylation counting mechanism####
def count(methyl_counts, GFF_orient):
    if GFF_orient == '-':
        for e in range(feature_size):
            #calculate methylation frequency per position
            #get methylation freqs and counts for current bin
            for c in range(3):
                m = methyl_counts[e][c]
                M = methyl_counts[e][c + 3]
                #write frequency of each methylation to correct column of e output file, both in terms
methylation relative to C's and relative to total nucleotides.
                if (M + m) > 0: #avoid division by zero

                    x = (int(100/bin_size) - (int(e//bin_size_element)) ) - 1

                    master_counts[x][c] += (M/(M+m)*100)
                    master_counts[x][c + 3] += 1
    else:
        for e in range(feature_size):
            #calculate methylation frequency per position
            #get methylation freqs and counts for current bin
            for c in range(3):
                m = methyl_counts[e][c]
                M = methyl_counts[e][c + 3]
                #write frequency of each methylation to correct column of e output file, both in terms
methylation relative to C's and relative to total nucleotides.
```

```
                if (M + m) > 0: #avoid division by zero

                    x = int(e//bin_size_element)
                    if x == int(100/bin_size):
                        x = (int(100/bin_size) - 1)

                    master_counts[x][c] += (M/(M+m)*100)
                    master_counts[x][c + 3] += 1

#################################
#BRING IN THE BS READS!!!!
#################################

for line in in_file: #Bring in BS file
    cols = line.split('\t')
    col3 = cols[3] #column with chromosome number and start position
    BS_chr = (int(col3[2:4])) #fix chromosome numbers in BS file

    orient = cols[2][0]
    pos_col = cols[3] #element and position
    pos1 = int(pos_col[5:]) #left edge of position on chromosome (BS seeker position is zero based, from left edge regardless of orientation)
    cols6 = cols[6].strip() #location of C's, methylated and unmethylated within sequence
    read_length = len(cols6)
    meth = cols6

    for o, line2 in enumerate(lines): #Enumerate GFF file and make o = line number

        if GFF_chr == BS_chr:

            if GFF_end >= pos1 >= GFF_start - read_length:
                pos = int(pos1 +1 - GFF_start) #(BS seeker position is zero based)

                if (len(meth) + pos1) <= GFF_end: #If GFF features overlap, it cannot count it if the read is too long (which means that read belongs to the next GFF feature)

                    if orient == '-':   #Check Orientation to count in correct direction
                        meth = meth[::-1]

                    for i in range(len(meth)):
                        ##Tally methylation information in read
                        methyl_counts[pos + i][6] += 1
                        if meth[i] == 'x':
                            methyl_counts[pos + i][0] += 1
                        elif meth[i] == 'y':
                            methyl_counts[pos + i][1] += 1
                        elif meth[i] == 'z':
                            methyl_counts[pos + i][2] += 1
                        elif meth[i] == 'X':
                            methyl_counts[pos + i][3] += 1
                        elif meth[i] == 'Y':
                            methyl_counts[pos + i][4] += 1
                        elif meth[i] == 'Z':
                            methyl_counts[pos + i][5] += 1
                else:
                    #If read to large
                    #BS read does not meet criterion, so calculate current stored data for GFF feature
and reset for next line
                    count(methyl_counts, GFF_orient)
                    if k > 39950:
                        GFF_chr = int(GFF_cols2[0])
                    if k == File_length:
                        break
                    GFF_cols2 = lines[k+1].split('\t')
                    GFF_orient = (GFF_cols2[6])
                    GFF_end = int(GFF_cols2[4])
                    GFF_start = int(GFF_cols2[3])
                    feature_size = GFF_end - GFF_start + 1
                    GFF_chr = int(GFF_cols2[0])
                    bin_size_element = feature_size/(100/bin_size)
                    methyl_counts = [[0,0,0,0,0,0,0] for i in repeat(None, feature_size)]
                    k += 1
            else:
                #If read is not within GFF positions
                #BS read does not meet criterion, so cacluate current stored data for GFF feature and
reset for next line
                count(methyl_counts, GFF_orient)
                if k > 39950:
                    GFF_chr = int(GFF_cols2[0])

                if k == File_length:
                    break
                GFF_cols2 = lines[k+1].split('\t')
                GFF_orient = (GFF_cols2[6])
                GFF_end = int(GFF_cols2[4])
                GFF_start = int(GFF_cols2[3])
                feature_size = GFF_end - GFF_start + 1
                GFF_chr = int(GFF_cols2[0])
                bin_size_element = feature_size/(100/bin_size)
                methyl_counts = [[0,0,0,0,0,0,0] for i in repeat(None, feature_size)]
                k += 1
        else:
            #If read is not on same chromosome
            #BS read does not meet criterion, so cacluate current stored data for GFF feature and reset for next line
            count(methyl_counts, GFF_orient)
            if k == File_length:
                break
            if k > 39950:
                GFF_chr = int(GFF_cols2[0])

            GFF_cols2 = lines[k+1].split('\t')
            GFF_orient = (GFF_cols2[6])
            GFF_end = int(GFF_cols2[4])
            GFF_start = int(GFF_cols2[3])
            feature_size = GFF_end - GFF_start + 1
            GFF_chr = int(GFF_cols2[0])
            bin_size_element = feature_size/(100/bin_size)
            methyl_counts = [[0,0,0,0,0,0,0] for i in repeat(None, feature_size)]
            k += 1


#This section is to add the data calculated for the last line in the GFF file
count(methyl_counts, GFF_orient)

##########################
#PRINTING AND CALCULATIONS
##########################
#CALCULATE PERCENTAGE OF METHYLATION FOR EACH BIN
#write headlines
out_file.write('mCG (%C)\tmCHG (%C)\tmCHH (%C)\n')
for e in range(len(master_counts)):
    try:
        mCG = str((master_counts[e][0]/master_counts[e][3]))
    except ZeroDivisionError:
        mCG = '0'

    try:
        mCHG = str((master_counts[e][1]/master_counts[e][4]))
    except ZeroDivisionError:
        mCHG = '0'
    try:
        mCHH = str((master_counts[e][2]/master_counts[e][5]))
    except ZeroDivisionError:
        mCHH = '0'

    out_file.write((mCG) + '\t' + (mCHG) + '\t' + (mCHH) + '\n')
```

```
#close files
in_file.close()
out_file.close()
print(datetime.now()-startTime)
```