Comparative genomics of repetitive elements between maize inbred lines B73 and Mo17

by

PIERRE MIGEON

B.S., University of Dallas, 2013

A THESIS

submitted in partial fulfillment of the requirements for the degree

MASTER OF SCIENCE

Department of Plant Pathology and Interdepartmental Genetics Program
College of Agriculture

KANSAS STATE UNIVERSITY
Manhattan, Kansas

2017

Approved by:

Major Professor
Sanzhen Liu

Copyright

Abstract

The major component of complex genomes is repetitive elements, which remain recalcitrant to characterization. Using maize as a model system, we analyzed whole genome shotgun (WGS) sequences for the two maize inbred lines B73 and Mo17 using k-mer analysis to quantify the differences between the two genomes. Significant differences were identified in highly repetitive sequences, including centromere, 45S ribosomal DNA (rDNA), knob, and telomere repeats. Genotype specific 45S rDNA sequences were discovered. The B73 and Mo17 polymorphic k-mers were used to examine allele-specific expression of 45S rDNA in the hybrids. Although Mo17 contains higher copy number than B73, equivalent levels of overall 45S rDNA expression indicates that transcriptional or post-transcriptional regulation mechanisms operate for the 45S rDNA in the hybrids. Using WGS sequences of B73xMo17 doubled haploids, genomic locations showing differential repetitive contents were genetically mapped, revealing differences in organization of highly repetitive sequences between the two genomes. In an analysis of WGS sequences of HapMap2 lines, including maize wild progenitor, landraces, and improved lines, decreases and increases in abundance of additional sets of k-mers associated with centromere, 45S rDNA, knob, and retrotransposons were found among groups, revealing global evolutionary trends of genomic repeats during maize domestication and improvement.

Table of Contents

List of Figures v
List of Tables vi
Acknowledgements vii
Dedication vii
Chapter 1 - Topic Overview: Structural and Copy Number Variation and the Maize Genome 1
Chapter 2 - K-mer analysis of B73 and Mo17 Genomes 40
Chapter 3 - Conclusions and Perspectives 75
References 82
Appendix A - Supplemental Data 90

List of Figures

Figure 1.1 Glyphosate is a competitive inhibitor of ESPS……..20
Figure 1.2 Schematic of rDNA organization of 45S producing repeat element…..33
Figure 2.1 Comparison of k-mer spectra in B73 and Mo17……42
Figure 2.2 Comparison of high-copy k-mers between B73 and Mo17…….45
Figure 2.3 cnvQTL mapping of genomic locations contributing differential abundance of HAKmers……...49
Figure 2.4 B73 or Mo17 specific HAKmers…….53
Figure 2.5 Allelic expression of 45S rDNA in hybrids of B73 and Mo17……...56
Figure 2.6 Change of k-mer abundances in teosine, landrace, and improved maize.……59
Figure A.1 Genome-wide distribution of B73- and Mo17-gain rDNA k-mers……...93
Figure A.2 Distribution of differential abundance rDNA k-mers on 45S rDNA……..94
Figure A.3 Barplot of total abundance of B73- and Mo17-specific k-mers…..95
Figure A.4 Change patterns of k-mer abundance…..96

List of Tables

Table 2.1 k-mers from single-copy regions in B73 and Mo17……..……41
Table 2.2 Number of functionally classified k-mers in different clustering groups…….……58
Table A.1 Statistics of functional classes of HAKmers...……..……90
Table A.2 Sum of estimated copies of all k-mers in each functional class…..…….……91
Table A.3 Number of HAKmers showing various mapping peaks..……91
Table A.4 Number of each functional class of B73-gain HAKmers showing various mapping peaks……..……92
Table A.5 Number of each functional class of Mo17-gain HAKmers showing various mapping peaks……...……...92
Table A.6 K-mer abundance of three pairs of k-mers harboring a SNV at 26S rRNA.....……92

Dedication

βλέπομεν γὰρ ἄρτι δι' ἐσόπτρου ἐν αἰνίγματι, τότε δὲ πρόσωπον πρὸς πρόσωπον: ἄρτι γινώσκω ἐκ μέρους, τότε δὲ ἐπιγνώσομαι καθὼς καὶ ἐπεγνώσθην.
(translation by Reid Comstock)

Chapter 1 - Topic Overview: Structural and Copy Number Variation and the Maize Genome

In the present work, I describe comparative genomic analysis performed in satisfaction of the Masters of Science degree in Genetics at Kansas State University in the laboratory of Dr. Sanzhen Liu. For this work, we sought to analyze structural variation between the two genomes of inbred lines B73 and Mo17. Previous comparative work (Springer et al. 2009, Belo et al. 2010) using this model comparative system, while comprehensive, was limited by several features which we attempt here to address in our own analysis. Our work takes advantage of technical and analytic advances that facilitate resolution of repetitive components of the maize genome, not considered by previous studies. Past comparisons of these two lines were limited in resolution as they relied on comparative array hybridization, whereas we performed our analysis on the basis of a PE sequencing dataset that we generated for both samples. Our analysis using reference-free analysis methods permits complete removal of reference bias. Additionally, our analysis focuses on repetitive DNA sequences, which have not been analyzed in the past due to difficulty involved with accurately resolving these regions of the genome. Our analysis of copy number and structural variation in the maize genome sheds new light on this important aspect of genome biology and confirms past findings of pervasive, dramatic variation between the genomes.

Mechanistic basis for formation of Structural Variation and Related Complex Genome Dynamics

Numerous large scale chromosomal variants, abnormalities, and their mechanisms of formation are well documented and are common genomic features across the tree of life. Here I use genomic structural variation (SV) to indicate genomic alterations beyond SNVs (single-nucleotide variants i.e. single-nucleotide polymorphism or SNPs) and small scale INDELs/IDP (Insertions and deletions, Insertion Deletion Polymorphisms). Comprehensively, this includes differences of order, orientation (for instance in the case of inversions), location (translocations), copy number (CNV), or presence/absence (PAV) of genes and repetitive elements, or more broadly, segments of DNA sequence. Speaking generally, SV are either balanced, with no net gain or loss of DNA sequence between compared individuals (in the case of translocations and inversions for instance) or unbalanced, involving net gains and reciprocally net loss of DNA. Unequal crossing-over (in either mitotic or meiotic contexts) results in reciprocal gain and loss of a segment of DNA. For instance, in the meiotic context, unequal crossing over yields one gamete with a gain (duplication) and another with a loss (deletion). Such variation is sometimes referred to as a genomic imbalance. Similarly, genomic amplification might be used to refer to copy-number gain.

A significant contributor to the formation of SV and source of novel genome variation is the activity of mobile genetic elements. In addition to serving to generate structural variation within the native host genome, they themselves are frequently found to vary in copy and position within the genome between individuals of the same species. First discovered in maize by Barbara McClintock, transposable elements (TE) have long been appreciated as drivers of genome

evolution and structural rearrangements in the genome. Mobilization of transposable elements can have profound impact on the structure and function of a genome through several means. For instance, direct insertion of TE into a gene can abrogate its function, with obvious direct functional consequences. In maize, the *Mutator* (*Mu)* transposon system is known to be highly active and to show preference for insertion into or near protein coding regions of the genome (Lisch 2013, Dietrich et al. 2002). The *Mutator* family of transposable elements is diverse and primarily characterized by the presence of a conserved 220 bp terminal inverted repeat (TIR) - many of the elements are non-automonous and instead rely on the activity of autonomous *MuDR* elements for their own transposition. The majority of mutations in *Mutator* lines of maize are caused by non-autonomous elements, which far outnumber the autonomous elements. Some non-autonomous *Mu* elements contain portions of the host genome within their inverted repeats, allowing them to contribute to dynamic re-structuring of the host genome through their movement. Although *Mu* TEs have been most well studied in maize, *Mutator*-Like Elements (MULEs) are now known to be found in diverse plant species and thought to be a common element of angiosperm genomes (Jiang et al. 2004). MULEs that contain fragments of the host genome are referred to as Pack-MULES. A study of the rice genome found at least 1,380 Pack-MULEs within the reference genome sequence, and also found that in some cases Pack-MULEs capture regions resulted in the production of novel gene fusions not found elsewhere in the genome, further demonstrating the ability of transposable elements to contribute to genome evolution and innovation. Other distinct transposable element families known to have the capacity to capture and modify position or copy of genic sequences include the Helitron class of TEs, thought to replicate via a distinct rolling-circle replication mechanism in a manner similar to prokaryotic TEs (Lisch 2013). This class of TE is known to occur in several plants species (including maize) as well as a broad range of other

eukaryotes such as *Caenorbabditis elegans*. Another classic TE system in maize is the Ac/Dc (Activator/Dissociation) TE system, which was first described by McClintock. In addition to the type II DNA transposons just described, long terminal repeat (LTR) retrotransposons are also found in the maize genome. These are highly abundant genomic components, and can be the dominant element found in the genome, in part because their RNA-mediated integration results in amplification of their copy, unlike the type II DNA TEs which replicate by a copy-and-paste mechanism (Lisch 2013). Retrotransposons from the TY1/Copia-like and TY3/gypsy retrotransposons families are maintained at extremely high copy in the maize genome, >75% of which consists of these and other LTR retrotransposons (Schnable et al. 2009).

The mechanisms discussed thus far involve insertion into coding sequences such that genes are disrupted, or translocations of genes possibly impacting their copy or resulting in novel position effects. However, transposons can also impact the regulatory networks in a number of ways, for instance, modulating gene expression by insertion into gene promoter regions. Recently, a MITE (miniature inverted repeated transposable element) insertion into a promoter of a maize NAC gene was demonstrated to result in variation in drought tolerance (Mao et al. 2015). The insertion of this element resulted in novel methylation and other epigenetic marks, which was associated with reduced expression of a NAC gene, especially under stress conditions. Recently in rice, similar TE mediated dynamics were found to be important in the context of disease resistance. A protein that would normally inhibit a resistance gene's function was suppressed in tissue during the adult stage, while expression in pollen resulted in higher yield. In this case the authors found that this desirable tissue-specific expression was due to promoter localized MITEs, which experienced CHH methylation (methylation of cytosine followed two non-guanine nucleotides) in normal tissue, but

not in the pollen (Deng et al. 2017). These sorts of TE dynamics also have known roles during maize domestication. TE insertions roughly 60 kb upstream of the teosinte branched 1 (*tb1*) gene resulted in gain of function regulatory architecture for this gene which have played a role in important morphological changes during domestication, namely loss of branching as well contributions to ear morphology (Studer et al. 2011). These insertions result in higher expression of *tb1*, a transcription factor, in domesticated corn. It is thought that this higher expression increases apical dominance in maize and therefore contributes to the loss of branching. TE have long been known to be stress-responsive, and in the same way that temperate phages can become lytic under stress conditions and thus enhance their overall fitness by "escape" from the lysogenic state in the case that their host might perish from stress, many transposons are also known to be stress activated (Lisch 2013). Several retrotransposons are known to become transcriptionally activated in specific response to abiotic stress such as heat stress (Arabidopsis retrotransposon ONSEN) or salt/cold stress in the case of rice DNA transposon mPing (Makarevitch et al. 2015). Because TE can impact expression of proximal genes, novel TE insertions can lead to the production of transcripts that are stress-responsive whereas previously these transcripts were non-stress responsive. In the case that the gene confers special advantage under conditions of stress, improved stress resistance might evolve in this manner.

Diverse mechanisms can contribute to the formation of SV. Insults to DNA stability can cause multiple classes of DNA damage, which produce small scale nucleotide changes or give rise to larger aberrations that change the copy of affected sequence. Externally inflicted DNA damage includes primarily ionizing radiation as well as other environmental chemicals and toxins, for instance oxidizers or compounds that intercalate DNA molecules and therefore disrupt DNA

replication. In some cases toxins such as free-radicals may be generated internally in the normal course of metabolism. Numerous inevitable internal procedural imperfections, for instance in replication or recombination, are also known to be mutagenic. Mechanisms involved in repairing damaged DNA can be imperfect and also lead to the types of variants that are of interest to us. Double stranded breaks, the type of damage that is most likely to cause mutations larger than single nucleotides, can be repaired on the basis of homology via the homologous recombination pathway or by non-homologous end joining (NHEJ). Typically, the dominant repair pathway depends on cell type and lifecycle stage, and organism. The lower fidelity NHEJ repair pathway is much more mutagenic as it simply joins together broken ends of DNA, and often results in INDEL type mutations. Homologous recombination, however, can also be mutagenic when it results in incomplete crossing over. Incomplete crossing over yields reciprocal deletion and duplication, such that one homologous chromosome will undergo segmental copy gain while the other chromosome will undergo loss of this sequence. Numerous more complex cases also exist which may give rise to similar aberrations, for instance recombination involving looped chromosome arms. The second major context by which these types of aberrations might take place is during DNA replication. Simple Sequence Repeats (SSRs) and low-complexity regions of the genome can present challenges to DNA polymerases and incorrect replication at these loci is more frequent than at other regions of the genome, thought to be due to polymerase slippage. Stalled replication forks at loci containing such sequences can undergo homology based rescue, and the presence of multiple proximal homologous domains can cause sequences internal to the repeats to either be deleted or duplicated. This is sometimes referred to as microhomology mediated break induced repair, *MMBIR*, or fork stalling and template switching, FoSTeS (Zhang et al. 2009).

In addition to the mechanisms just discussed, which primarily cause genomic imbalance at single loci, presence of multiple copies of the same sequence at distinct loci can cause apparent *de-novo* copy number variation in the offspring relative to the parents due to segregation of these non-allelic regions. These sorts of dynamics may in some cases explain transgressive segregation, wherein some offspring of a cross will show a more extreme phenotype than either parent. The segregation of Non-allelic homologs (SNH) in copy number variation was explored in the work of Liu et al. 2012. This work used array comparative hybridization to examine copy number differences between Mo17, B73, and two inter-mated B73xMo17 RILs. Examining segments that were single copy in the B73 genome and which were not found to be significantly different between B73 and Mo17 genomes, the authors found that a proportion of these exhibited CNV in the offspring that was not found in the parents. There are several lines of evidence that might be used to distinguish SNH CNV from other mechanisms of formation. Firstly, these are likely to be loci subject to high levels of recurrent CNV, because we would expect them to occur in the progeny of any cross between individuals possessing non-allelic homologs. The authors analyzed ~300 IBM RILs (Recombinant inbred lines derived from intermated B73 and Mo17 lines) and found that this was in fact the case. One would also expect the gain and loss segments to be consistent with segments that were identified in the parents. Most obviously, the segments displaying SNH would also be expected to be non-allelic between the parents. The authors found strong support in their analysis for each of these lines of evidence for SNH. Further, they found evidence that these events involved likely functional protein coding genes. Extending this set to a group of 14 high confidence genes corresponding with de novo CNV regions in the two progeny assessed, the authors found two genes, loss of which were significantly associated with phenotypic variation in kernel diameter, row number, and tiller number.

Sequence context dictates dynamics of SV, and as a result genomic stability is not equally distributed throughout the genome. This has been documented in numerous plant studies (Springer et al. 2009, Belo et al. 2010) and is expected given varying intensities and types of selective pressure across the genome in the same way that levels of single nucleotide polymorphisms are known to vary considerably as a function of genomic landscape. As a result, highly conserved regions of low structural variation are observed as are CNV hotspots. Temporal dynamics of highly repetitive regions of the genome are expected to be entirely distinct from more highly conserved low copy regions of the genome, due to reduced selection as well as the propensity of repetitive sequences to be subject to phenomena such as unequal recombination. The genomic landscape of SV is further complicated by recombination dynamics, which might serve to redistribute SV. Low recombination regions such as certromeres and pericentromeric regions in many species are subject to higher levels of structural variation, and regions of reduced recombination are thought to permit repeat arrays to expand dramatically beyond what would be possible in recombination rich regions which in contrast are thought ultimately to reduce copy of repeat arrays (Hiatt et al. 2002). From the reverse perspective, SV and CNV also impact recombination, as recombination is expected to be reduced between homologous chromosomes over regions that are non-homologous. Heterochromatin is also thought to suppress recombination potentially due to the compressed, inaccessible state of the chromatin, and repetitive DNA often exists as heterochromatin due to being subject to RNA-directed DNA methylation. Recombination in the case of inverted DNA can result in di-centric and a-centric chromosomes and reduced gamete viability. As a result, there is less recombination at inverted sequences, and higher linkage disequilibrium in these regions. Genes found in these regions tend to form haplotype blocks due

to reduced recombination. In the case of haplotype specific fitness gains, it is clearly beneficial to the organism to prevent disruption of the haplotype by hosting it in a recombination depleted region of the genome such as within an inversion.

A classic example of a disease phenotype associated with expansion of repetitive sequence is Huntingtons disease, which is caused by abnormally high copy of the polyglutamine encoding CAG codon in the huntingtin gene. This repeat is usually found at 6-35 copies in healthy individuals, but in the disease state amplification of the repeat to several hundred copies is typical (Imarisio et al. 2008). Fragile X is another classic example of the impact of short tandem repeat expansion. In both of these cases, the repeat is a trinucleotide so its expansion does not cause frameshift in the disease gene nor does it cause changes in gene expression levels, but creates disease by disrupting normal protein function. The role of large structural and chromosomal abnormalities (i.e. resolvable by karyotypic analysis, several million basepairs and above) in human disease has been appreciated since the 1950s, for instance the role of aneuploidy in mental retardation. A region of chromosome 15 short arm (15q) has been implicated in both Prader-Willi Syndrome and Angelman Syndrome (the PWS/AWS region), which result from deletions in this region in either the paternal or maternal chromosomes, respectively (differences in sex-based genomic imprinting result in the distinctive phenotypes). This region is subject to complex architecture, with several known tandem duplications and large inverted repeats. Recurrent deletions in this region at characteristic genomic coordinates implicate the complex structure at the locus in the formation of these deletions. Not surprisingly, more recent work characterizing genome wide distributions of SV among populations have also observed more frequent CNV at loci containing segmental duplications.

Typically, the most direct way that SV impact an organism's biology is by modulation of gene expression, and the most intuitively obvious means by which this can occur is by changing net levels of cellular transcript abundance. Deletion of a sequence abrogates gene expression, and in this way discrete changes in expression levels are typical of presence/absence variable regions. Deletions spanning only a portion of a gene may destroy gene function while not leading to complete loss of transcript, perhaps resulting in pseudogenization of these genes. To my knowledge the evolutionary role of pseudogenization due to such deletions has not been explored, although it may be ripe for exploration. Deletions are thought to be under more negative selective pressure compared to other types of SV, especially relative to duplications, and evidence from population level studies supports this. On the other side of the spectrum, amplification of DNA segments tends to directly increase transcript levels for those genes found within the duplicated regions. Copy number gain can be local, as in the case of tandem duplications, or dispersed throughout the genome. Copy gain events increase gene dosage and thus potentially overall expression of genes found within these regions as well. Recent work on transgenic gene duplication has demonstrated that in the case of tandemly arrayed duplications, gene expression levels can increase as a non-linear function of copy number- in this case, gene expression for tandemly array gene duplications showed greater than the expected two-fold increases in expression. This over-activity was dependent on the tandem arrangement of the duplication and was not as pronounced in the case of dispersed gene duplications. The bar mutation in drosophila, resulting in reduced eye facet number and overall eye size, is caused by a segmental duplication. 4 copies of this duplication result in a generally severe phenotype, but the severity is exacerbated when three of the copies are tandemly arrayed on one chromosome (with the other homologous

chromosome possessing only a single copy of said segment) compared to the case of one duplication on each chromosome, despite the overall copy number remaining unchanged between the two cases. Again position effects are significant and these sorts of dynamics can become increasingly complex when considering non-haploid individuals, situations for which intermating can create a number of diverse permutations of the parental material. Combining diverse SV and CNV through inter-mating may also complement non-lethal but deleterious deficiencies and, thus, contribute to hybrid vigor (Springer et al. 2009). As with deletions, duplications can also incompletely span genic regions and thus potentially cause pseudogenization or the creation of novel gene fusions as a result.

In addition to impacting gene expression directly through modification of gene dosage, balanced changes in genome structure can also impact gene expression either by modifying functional cis-regulatory elements (such as enhancers or insulators) or by changing the location of a gene within the genome. Through either moving a gene further or closer to a regulatory element, or vice-versa, by moving a regulatory element relative to a gene, gene expression can be modified due to position effects. Unbalanced changes can also cause changes in gene expression, for instance through amplification or loss of activating or inhibitory elements. Several examples of enhancer hijacking, whereby rearrangements impact the position of a gene relative to an enhancer, are known to occur in cancer genomes, driving increased expression of proto-oncogenes. Noteworthy examples include the proto-oncogenes MYC, MYB and TAL1. Loss of an insulator sequence between a gene and a distal enhancer can also create a situation wherein gene expression increases for that gene, as in the case of insulator spanning IGF2 containing duplications (Beroukhim et al. 2017).

Aside from impacting gene expression, there are also a number of ways that SV might impact population dynamics. As discussed, haplotype blocks might be formed within an inversion in a way that might be beneficial to an organism. Additionally, large chromosomal aberrations may contribute to speciation. Such chromosomal differences might contribute to speciation by causing hybrid sterility as a result of aberrant meiosis (for instance due to improper pairing of homologous chromosomes).

Quantification of SV/CNV

Despite the pervasive character of the complex genomic components discussed, their characterization can prove a complex challenge and so a discussion of different methods of detection is warranted. There are several means of detecting structural or copy-number variation, which can be broadly categorized as either high or low throughput. Low throughput methods would primarily be used to investigate individual genes, for instance to validate copy number variation between different samples that was detected by other means or simply to investigate copy number of a gene of interest between different individuals. These low throughput methods include Southern blotting, Giemsa banding, quantitative PCR, and FISH or fiber-FISH. Southern blotting is an established method for detecting the presence of a gene or sequence within genomes. It involves size fractionation of restriction digested DNA sequences by gel electrophoresis, followed by transfer of DNA from a gel (typically agarose) onto a membrane, to which it is fixed, followed by probing with a probe specific to the sequence of interest. Visualization typically involves

detection using radioactive probes. When copy number differences result in different sizes of restriction fragments containing the sequence being probed for, then the Southern Blot will reflect this. Using Southern blotting for CNV studies is relatively uncommon, however. In this capacity it is best used to validate results. High sensitivity and accuracy possible with Southern blotting make it a gold standard in these respects. Zhang et al. 2015 represents a good example of successful use of Southern blotting to validate computational results of CNV. For large scale variation, giemsa staining allows for chromosome banding, which may be used to visualize chromosomes and can reveal heteromorphism. Chromosome banding techniques involve dyes that create differential staining patterns between heterochromatin and euchromatin in a way that creates distinct banding patterns for each chromosome. The distinct banding pattern allows identification of chromosomes as well as structural differences within these (Feuk et al. 2006). Other microscopy based methods are fluorescence in-situ hybridization (FISH) as well as fiber-FISH. Both methods involve hybridization of a fluorescent single-stranded nucleotide probe to whole chromosomes, in the case of FISH, or to long single strands of DNA (fiber-FISH) to visualize location and structure of sequences of interest. FISH and derivatives thereof, as well as chromosome banding techniques, all have the potential to be somewhat higher throughput compared to Southern blotting, but are limited in resolution by the resolution of microscope used. Finally, quantitative PCR (qPCR) may be used to assess DNA copy number. Doing so requires using PCR amplification of the sequence of interest from genomic DNA, using a known single copy gene as a reference (Heid et al. 1996, Ma et al. 2014).

High throughput methods have more recently become available to researchers and have significantly expanded our understanding of the spread and frequency of SV occurrence in

populations and the genome wide distribution of these events. The earliest method is array comparative genomic hybridization (aCGH) (Pollack et al. 1999). This method uses a microarray, consisting of a glass slide to which are fixed single stranded probes designed based on the genome sequence. Genomic DNA from a sample of interest is purified, labeled with a fluorophore, and hybridized to the array. Hybridization intensity demonstrates the level of homology between probe and sequence. Two distinct samples, hybridized to the array, can then be compared at a whole genome level on the basis of their respective hybridization intensities to the array. Because the method relies on hybridization to probes designed based on a reference sequence, there can be significant problems with bias towards the reference sequence, for instance, PAV occurring only in the alternative sample being considered will not be detected because they will not possess homology to the reference sequence. Array based methods are not capable of determining breakpoints with basepair resolution nor will they be able to detect balanced SV. Within their scope of inference, however, microarray based methods are considered to be reliable if these experiments are well designed. This requires that the researchers perform replicates and control measures such as the use of dye swapping as well as correct statistical analysis. aCGH also has medical applications in routine genotyping of tumor SV/CNV (Pinkel et al. 2005).

With the advent of affordable next generation sequencing methods, microarray based analysis has predominantly been replaced with direct sequencing methods. The primary challenge with regards to this type of data is correct analytic methods in order to allow assessment of CNV/SV. While genome assembly can often successfully reconstruct genome sequences, algorithms often struggle to accurately represent sequence copy and can be thwarted by the complexity inherent in copy variable and repetitive regions as well as by issues such as heterozygosity. Fragmented assemblies

might not encompass the entirety of the genes present in a genome, and this might be especially true in the case of genes belonging to paralogous gene families. Additionally, repetitive sequences are notoriously difficult to accurately assemble in a genome, and regions such as the centromeres are not typically resolved using assembly of short read methods. Rather than assembly, it is often most practical to quantify differences in copy on the basis of mapping reads of a sample of interest to a reference sequence. In this case there remains some level of reference bias, as obviously sequences completely absent from the reference will not be mapped, but is nonetheless of much higher positional and sequence resolution than array based methods and also resolves information regarding balanced variants as well as potentially allowing reconstruction of more complex variants. Typically the most straightforward means of quantifying copy number is through analysis of read depth of reference mapped reads. Duplications relative to the reference are expected to yield twice as many reads mapped to this region, increasing linearly as a function of copy. In the same way, deletions relative to the reference will manifest themselves as regions to which reads do not map. There are multiple ways that one might employ comparative read depth analysis, for instance performing within-sample normalization, or ratio based methods that compare the ratio of read depth of two samples mapped to a reference. Further sophistication such as segmentation, wherein genomic segments showing similar depth are grouped into bins for the purpose of analysis, might also be employed to detect CNV. Complementary to read depth analysis is split read and discordant mate pair analysis. In this case, reads with known insert size but mapping to much larger distances in the reference provide information about rearrangements that have occurred. A pair of reads with an insert size of 500 basepairs, but mapping to two different chromosomes, would provide evidence for the occurrence of a translocation. While discordant mapping represents SV with breakpoints that have occurred within insert between the reads, information about the SV can

also be gained from individual reads that span the breakpoint. A read that spans the breakpoint of an inversion, for example, will map to the reference partially in one orientation and partially in the other orientation, for example. Numerous software have been developed recently for this type of analysis which generally rely on analysis of previously generated alignment data, although in many cases high false positives have been observed and careful quality filtering for this data is necessary. For more discussion of the above methods, see Pirooznia et al. 2015.

More recently, technological innovations promise to make structural variation in the genome increasingly easy to detect and assess within samples. The primary advantage of these is increased raw data unit size. Third generation long-read sequencing, for example, makes it possible to capture break points of some CNV and SV in a single read, facilitating their reconstruction in the final assembly. These methods of SV detection might be especially useful in the case of tandem duplications or higher copy repetitive sequences, as typically assembly algorithms struggle to assemble repeats that are larger than read length. Additionally, these methods are also effective at identifying sequence that is completely unique to the sample, i.e. novel insertions relative to the reference, which are not detectable based on reference mapping techniques. Ultimately, however, repeat arrays larger than the read length afforded by these technologies exist and even the most sophisticated assemblies of the day remain incomplete. Therefore in addition to 3rd generation long read sequencing, additional non-PCR based methods for gaining information about long-range genome structure are emerging, such as high-throughput optical mapping (Bionano genomics) or other techniques such as those provided by Chicago dovetail and 10x genomics.

Case Studies: Known Biological Consequences and Phenotypic Implications

Genome wide studies of structural variation gained wide interest in the mid-2000s, especially with the publication of the two seminal papers of Sebat et al. and Iafrate et al. (both 2004), following shortly upon the completion of the human genome sequencing project. Previously, there was some appreciation for the role of chromosomal abnormality and structural and copy number variation in human disease, and some data suggested the genomic imbalance could play a role in cancer and other diseases. However, the levels of occurrence for these in healthy populations was very poorly understood. Both papers sample normal healthy individuals and use comparative array hybridization to detect large (several hundred kb) regions of genomic unbalance within this cohort, and report genome wide distribution for this variation. Computational results were validated using FISH in both cases. In some cases multiple CNV were found to occur near to each other, suggesting CNV hotspots. Variants were found in regions enriched for tandem duplications or other types of chromosomal rearrangements suggesting inherent genomic instability for these loci. CNV were shown to involve coding regions of the genome, with specific discussion of copy-variation in the amylase alpha 1a and 2a and neuropeptide-Y4 receptor with possible non-disease related phenotypic consequences as well as several genes relevant to neurological or other disease susceptibility. These works devote some of their scope to discussion of the frequencies of these variants and implications on selection is given. Since the publication of these works, much progress has been made towards understanding levels of natural variation in copy number within human populations as well as towards understanding the role that CNV has in several diseases. Significant roles for CNV in human diseases like cancer are now understood, as well as for conditions such as autism (Feuk et al. 2006).

Regions of the genome undergoing the most rapid evolution might be expected to be among the most highly variable loci among a population at the sequence level, for instance genes involved in immune function (NB-LRR resistance genes in plants and MHC complex genes in humans) that might be undergoing positive diversifying selection due to arms-race like dynamics with a pathogen. In a similar way, these regions might also show higher levels of SV and CNV variation for the same reason. These events might be important factors allowing rapid adaptive evolution to take place, and there are several clear cases that demonstrate the utility of these mechanisms for the establishment of resistance to diverse types of environmental challenges in plants. Here I will discuss several important examples of genomic amplification of a single or a handful of genes and their direct impact on phenotype and contribution to the survival of individuals that would typically not survive in an environment featuring the given stressors. While expansion of gene copy number might allow functional gene copies with reduced selective constraint and therefore allow for evolution of distinct and novel function, (as is typical of gene families under rapid evolution) over short evolutionary time periods, one might argue that changes in gene expression levels might more immediately tune existing functionality in way that requires less biological innovation. Organisms often have mechanisms for dealing with minor stresses, for instance removal of environmental toxins, and increased expression of these genes under conditions of extreme exposure provide a means for overcoming these more extreme conditions. Here I discuss cases of CNV involved in adaptation to a number of abiotic and biotic traits, but additional cases also exist in the literature of CNV impacting flowering time, plant size, and fruit morphology (Żmieńko et al. 2014).

One of the clearest examples of CNV playing a role in this type of adaptive evolution over an extremely short evolutionary time frame is seen in the case of glyphosate resistance in several weed species (Funke et al. 2006, Gaines et al. 2010, Jugulam et al. 2014). Glyphosate is an effective, broad spectrum pesticide that functions by blocking the shikimate pathway enzyme 5-enolpyruvylshikimate-3-phosphate synthase (EPSPS) enzyme. The enzyme catalyzes the conversion of Shikimate-3-phosphate to 5-Enolpyruvyl-shikimate-3-phosphate by addition of the enolpyruvate moity of phosphoenolpyruvate (PEP) to the 5-hydroxyl group of shikimate-3-phosphate in order to form an enolpyruvyl functional group (**Figure 1**). Glyphosate inhibits this step of the pathway by acting as a competitive inhibitor for PEP and a transition state analog. The shikimate acid pathway is necessary for the biosynthesis of the three aromatic amino acids phenylalanine, tyrosine, and tryptophan, as well as for that of folates. The pathway is present in plants and some microorganisms, whereas it is absent from animals (consequently phenylalanine and tryptophan are essential amino acids), and is essential for growth in these. Inhibition of the pathway is lethal in plants, explaining both the effectiveness of this pesticide as well as its broad spectrum of applicability. Glyphosate is typically used in conjunction with crop plants that have been engineered to be resistant to glyphosate, typically through transgenic introduction of the EPSPS gene found in *Agrobacterium tumefaciens*, which is not inhibited by glyphosate. In this way, fields of glyphosate resistant crops can be treated with glyphosate to effectively remove weeds while the crop of interest remains unaffected. Recently, there have been multiple reports of resistance to glyphosate developing in wild populations. Copy number variation of the native EPSPS gene has been implicated in glyphosate resistance in both *Kochia scoparia* (Jugulam et al. 2014) and *Amaranthus palmeri* (Gaines et al. 2010). In the case of *A. palmeri*, gene amplification is thought to involve the action of transposable elements, and the gene is found to be dispersed

throughout the genome in resistant individuals, whereas in *K. scoparia*, the gene is tandemly amplified. Irrespective of mechanism of amplification, in both cases amplification results in higher expression and protein levels of EPSPS. As a result, glyphosate is essentially titrated out, with EPSPS present at much higher levels than its inhibitor. Copy number among wild populations of *K. scoparia* has been demonstrated to be increasing over a period of less than a decade and corresponds linearly to tolerance of higher levels of glyphosate application.

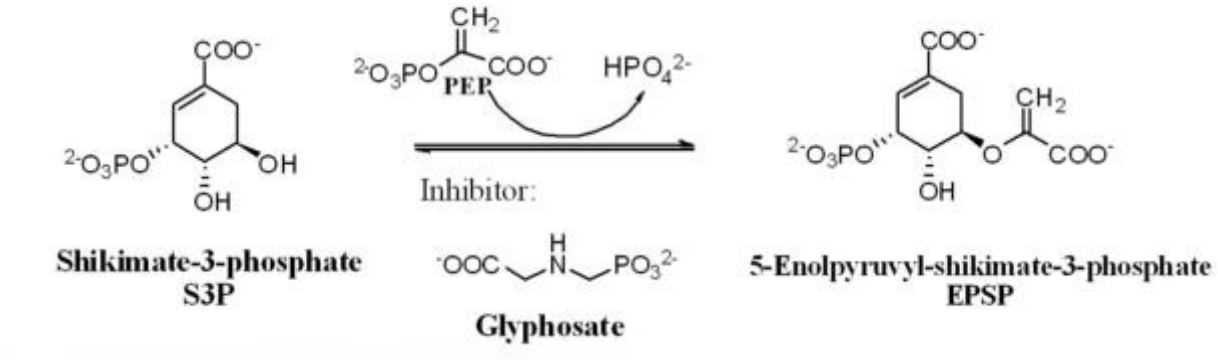

Figure 1.1: Glyphosate is a competitive inhibitor of ESPS (Funke et al. 2006)

The case of glyphosate resistance demonstrates how CNV can allow rapid adaptation to an environmental stressor, albeit one that is not agriculturally useful. However, there are several notable examples of CNV implicated in resistance to other types of abiotic stress which are of high agricultural interest. Along with pathogen related damage to crops, a major hurdle to food production world-wide remains other abiotic stresses that limit global acreage available to farmers from which to produce food. There are major calls underway for greater food production to feed the expanding global human population, with much of these efforts focused on either increasing the yield per acre possible, for instance through the use of disease resistant plants, or through increasing the total number of acres available for farming, for instance by increasing the amount of farmable land. Abiotic stresses of especial interest to farmers might include high heat, drought,

cold, and soil toxicity (for instance high salinity). These types of stressors might severely limit farming in impoverished regions of the world for which only limited agriculture is currently possible, but improved resistance to abiotic stress could also help increase agricultural output in more established parts of the world as well improving the variety of crops that can be grown in a region. Recently, copy number variation has been implicated in several important examples of plant resistance to abiotic stress in addition to the already discussed glyphosate resistance joins boron and aluminum tolerance in rice and maize respectively, submergence tolerance in rice, and cold tolerance (Żmieńko et al. 2014).

In maize, gene amplification and copy number gain has been implicated in soil aluminum tolerance. Aluminum toxicity is an important abiotic stressor that can significantly reduce the range of land that is available to farmers for the production of crop, in the USA as well as globally, and is known to be especially problematic at the equator in tropical ranges. Acidic soils make up 50% of the worlds potentially arable soil, and aluminum toxicity is the primary barrier to use of this land for agriculture. Soil acidity impacts solubilized aluminum, with aluminum in alkaline soils found complexed with other elements (in the form of aluminum oxides for example), whereas in acidic soils it exists as a non-complexed toxic free cation. Additionally, soil toxic aluminum content is dynamic, as different farming practices or acid rain might increase total soil acidity (Delhaize et al. 1995). Aluminum in acid soils significantly limits plant growth by slowing or preventing extension of roots at their apex, thereby stunting and preventing growth from an early stage. In plants, resistance to soil aluminum can be facilitated by release of organic acids into the soil, which serve to chelate toxic aluminum species, thereby preventing uptake by plant roots and subsequent toxicity. In maize, microarray analysis identified the MATE1 gene, a member of the

Multidrug and Toxin Extrusion family, as the most highly upregulated gene in Al resistant maize lines in comparison to non-resistant lines. Previous studies had implicated members of the MATE family of drug efflux pumps involved in the efflux of citrate in aluminum resistance in several plant species, such as barley, rice and Arabidopsis. Investigation into the MATE1 gene in maize revealed a role for CNV in aluminum resistance (Maron et al. 2013). CNV was demonstrated using both qPCR and FISH, and investigations using a RIL crossing population demonstrated high correlation of copy number (3 copies vs 1) with aluminum tolerance. Finally, the authors also found expression to be highly correlated with copy number as well, both under aluminum treated and non-treated conditions. The authors note that the CNV in this case is found predominantly in tropical lines where soil acidity is higher, suggesting that this variant is a specific adaptation for higher soil acidity found in these regions. While many SV are thought to predate domestication, several lines of evidence (perfect conservation of homology among the 3 copies of MATE1, lack of this variant in teosinte, and low frequency among maize inbreds) support relatively recent origin for this CNV, demonstrating the capacity for SV to generate rapid adaptive traits.

For adaptive traits, the majority of plant CNV studies have focused on abiotic stress. One well-known example of plant disease resistance is given by the work of Cook et al. 2012. Soybean (*Glycine max*) is the world's most widely used legume crop, with more than $35 billion in farm sale value per year in the United States alone. For soybean, the most economically damaging pathogen is Soybean Cyst Nematode (SCN). Resistance to this disease had been previously mapped to the *Rhg1* locus, a 67-kb region of the genome containing 11 putative protein coding genes. The resistant allele, *rhg1-b*, is now deployed in over 90% of all commercial varieties marketed as SCN resistant. Previous to this work, the resistance loci *Rhg1* had been identified as

a QTL on chromosome 18 of the soybean genome. Silencing experiments identified three genes overlapping with this QTL, silencing of which through RNA interference experiments resulted in reduced resistance to SCN. Concurrent analysis of the structure of the locus using fosmid clone sequencing revealed unique junctures not present in the reference genome, and the authors found that a region of around 30 kb (containing the genes found to contribute to resistance) was tandemly duplicated in resistant lines. Copy amplification of this region was also evidenced by read depth, which was found to be about ten-fold relative to surrounding regions as well as relative to homologous regions on other chromosomes. Follow up qRT-PCR experiments demonstrated higher expression levels for these genes. Additionally, the authors then used fiber-FISH to directly view the repeat arrangement for different haplotypes of the *Rhg1* locus, and found further evidence for multiple copies of the 30kb repeat in resistant lines, with between 3 and 10 copies of the 30 kb region.

Maize as a Model Organism

Zea mays ssp. mays, less formally known as maize, is a major global crop, grown for human consumption, biotechnology purposes such as ethanol and biofuels production, and as feed for livestock. It is among one of the most economically important crops, ranking with wheat and rice for global grain production and a market value of multibillions of dollar in annual revenue. While its status as a major global crop makes maize an important research system for applied agricultural and breeding purposes, maize has been an important model organism for research in basic biology since the early 20th century, although it was used in research by Gregor Mendel as early as 1869 and subsequently by Correns and de Vries in studies leading to the rediscovery and confirmation of Mendel's findings in the 20th century. Its status as contributing so extensively to both basic and applied research makes maize unique among model organisms. Numerous groundbreaking discoveries were first investigated in this system, famously mobile genetic elements by Barbara McClintock, as well as other phenomena such as that of paramutation, providing early insight into epigenetic biology, as well as early research into heterosis. The system was adapted as a model system for genetics in part because of the feasibility of performing genetic crosses, due to the fact that the male flower, found at the tassel at the top of the plant, is physically separated from the female. Thus researchers can perform controlled crosses, by covering tassels (typically using bags) and ears (in order to prevent undesired fertilization). Other cereals typically require laborious emasculation to allow controlled crosses and their crosses produce fewer seeds, compared to maize, which can produce abundant seed from a single cross. Finally, the large chromosomes and the presence of distinct chromosomal features such as heterochromatic knobs also make maize an ideal model for cytological analysis. Rollins A. Emerson of Cornell and advisor Edward M. East

of Harvard are considered to be the fathers of modern maize genetics, with Emerson mentoring maize giants such as George Beadle, Charles Burnham, Marcus Rhoades, and Barbara McClintock (Schnable et al. 2009, Coe 2001).

Among the cereal crops, maize is most closely related to sorghum (*Sorghum bicolor*), with the two sharing a common ancestor ~12 million years ago. Despite differences in overall genome size and ploidy, the cereal genomes share extensive collinearity, facilitating comparative evolutionary studies between the species. Many studies of dynamics of genome evolution have compared ortholog retention or loss between maize, sorghum, rice (*Oryza sativa*), *Brachypodium distachyon*, and *Setaria italica*. It is generally understood that major differences between these are due primarily to ploidy changes and subsequent repercussions of ploidy changes, and transposable element activity. Following the divergence of maize and sorghum lineages from their most recent common ancestor, maize underwent a whole genome duplication (WGD), and spent some of its evolutionary history (5-12 million years ago) as a tetraploid. Ancient genome duplications are not uncommon among other well studied organisms, for instance having been thought to have occurred in the models yeast and Arabidopsis. Following a period of tetraploidy, maize reverted to the diploid state. The process of diploidization involved fusion of chromosomes rather than loss of chromosomes, resulting in retention of both sets of original genomes. Chromosome level comparison to sorghum, which failed to undergo WGD, demonstrates two co-linear chromosomal regions within the maize genome for each sorghum chromosome. As a result of this process, maize has abundant gene paralogs, but many of these paralogs have been lost. The process of loss of paralogous genes in maize is known as fractionization, and has been demonstrated to favor one of the subgenomes over the other. As a result, the bulk of the extant maize genome consists of the

major subgenome while a smaller proportion consists of the remaining minor subgenome (1.26 and 0.75 Gb, respectively, Schnable et al. 2011).

Maize has also been an important model for understanding dynamics of domestication and evolution, and its domestication story has been extensively investigated historically (Chia et al. 2012, Matsuoka et al. 2002). Identification of the maize ancestor was subject to extensive decades long debate among the maize community, but extensive genetic and now sequencing based evidence indicates that maize was domesticated from the wild grass teosinte. The name teosinte refers to several species of the grass of the genus *Zea* that grow in Mexico and South America. The most similar to maize is *Z. mays ssp parviglumis*, from which modern maize was domesticated. Archeological, genetic, and sequencing evidence support the domestication of maize from *Z. mays ssp parviglumis* in the Balsas River basin of southwestern Mexico around 10,000 years ago (Doebley 2004). An initial domestication period was then followed by a period of improvement, during which increasingly desirable agronomic traits were selected. Extant maize landraces, having undergone domestication but not improvement, have also been used to shed light on genes selected for during maize domestication. Much of the debate over the identity of the maize progenitor was based on differences in morphology between teosinte and modern maize. While morphological differences are significant, more recent work has demonstrated that relatively few genetic changes were responsible for changes during domestication. Major differences between the two include branching patterns and ear morphology. While maize has just a single stalk, making it more amendable to agriculture, teosinte is branched. The most dramatic changes, however relate to ear morphology. Teosinte kernels are encased in a hard glume, making their consumption rather difficult. Maize kernels, on the other hand, have undergone softening. Maize

cobs feature far more kernels than do teosinte kernels, which are arrayed in near parallel to their axis, in contrast to maize kernels, which are arrayed perpendicularly to their axis. These changes were critical to changing maize from a wild grass to a crop. Another historically debated question concerning maize domestication was the number of distinct domestications resulting in domesticated maize. Many plant and animal species domesticated during the same period are known to have been domesticated multiple times, and the extensive variation evident in maize suggests the possibility of multiple domestication events. In Matsuoka et al. 2002, analysis of 99 SSR loci in maize and its ancestor shed light on the controversy surrounding the number of maize domestication events. Their phylogeny of diverse maize lines demonstrates a single ancestral branch and supports a single domestication as well as evidence for introgressions from *Zea mays ssp. Mexicana*, which, unlike *Zea mays ssp. parviglumis*, can be found growing as weed in maize fields in the highlands of Mexico and can easily hybridize with it. In the couple thousand years following this single domestication event, it is thought that maize spread from its center of origin in Southern Mexico over two main paths of dispersal. The first path follows northern and western Mexico, up through southwestern America and finally outwards towards east America and Canada. The Second path is thought to spread in the opposite direction, leading further south into Guatemala, the Caribbean Islands and into the Andes mountains. Post-Columbus, Europe also saw introduction of maize as well.

More recently, extensive sequencing resources have also made maize a powerful system for studying genomics and an ideal model for understanding complex and dynamic eukaryotic genomes. The first genome sequence was completed in 2009 (Schnable et al. 2009), with extensive improvements to the sequence made since then (the long promised reference version 4 having been

completed using long-read single molecule sequencing technologies, both Pacific Biosciences SMART sequencing and Bionano Genomics optical mapping). Additionally, resequencing of hundreds of diverse maize lines representing geographically distinct populations as well as ancestral and improved lines has been performed, resulting in rich information to draw from for this research. The extreme phenotypic variation seen in maize is now also appreciated at the nucleotide level and the extent of single nucleotide variants (SNV), indels and SV in the genome can be quantified from this data. The temperate inbred lines B73 (which served as the basis of the reference sequence) and Mo17 have frequently used as models for comparative analysis at targeted loci as well as at the whole genome level. SNV are found between these two lines approximately every 80 bp and IDP (insertion / deletion polymorphism) every 300 bp (Springer et al. 2009). On average, for any two randomly selected maize lines, SNV polymorphisms will be observed at a rate higher than that of between humans and chimpanzees. The highly complex and variable genome therefore make maize an ideal system for studying CNV and SV at a genome wide scale.

One of the most notable early examples of a genome wide survey of SNV/CNV in maize was that of Springer et al. 2009. In this work, the authors perform comparison of the maize elite inbred lines B73 and Mo17 in order to evaluate CNV, which they define as sequences present in both lines at different copy, and presence absence variation (PAV), which they define as being present in one genome but completely absent in the other, for this comparative system. The authors developed a high density (2.1 million feature) oligo microarray for this analysis using B73 BAC sequences. The authors found extensive genome wide variation between the lines, and also identified regions for which there was little or no variation between the two lines. Variation was not equally distributed. The authors found that there was typically low levels of variation found at the

centromeres, but also identified a roughly 19Mb region on chromosome 8 and a 17Mb region on chromosome 1 for which no variation was detected. Seven low diversity regions greater than 10Mb were identified using a sliding window approach. The authors identified a hypervariable region on chromosome 6, close to the Nucleolus Organizing Region (NOR) of maize. The authors identified a B73 specific sequence which was deleted in Mo17 of roughly 2.6 Mb, and confirmed its absence in Mo17 using amplification of 32 PCR primer pairs spanning the length of the segment. They also tested for presence/absence of this segment in 22 other inbred lines and found that 16 of the lines tested possessed the segment while the other 6 did not. The authors performed some analysis of the 31 predicted genes contained within this PAV and provide evidence that many of these genes are functional. Subsequent research in maize using array hybridization methods using an expanded sample of individuals show extensive SV within the maize population impacting regions enriched for lack of putative orthologs in other species (Swanson-Wagner et al. 2010). Additionally, these found that SV were observed in multiple lines, suggesting moderate frequencies of these variants within modern maize lines. Additionally, many variants were also conserved between modern and ancestral teosinte lines, suggesting an ancient origin for these.

Characteristic Maize Repetitive Genomic Elements and Functional Implications

With the advent of whole genome sequencing in the late 2000s and the subsequent fall in sequencing prices, in addition to the new appreciation for the extent of genome copy and structural

variation came also a new understanding of the pervasive extent of non-coding, repetitive DNA within many eukaryotic genomes and especially the maize genome. With the completion of the maize genome sequence in 2009, transposable elements, already understood to be significant features of the genomic landscape, were shown to make up as much of 85% of the *Zea mays* genome (Schnable et al. 2009). Differential TE activity in maize lineages during domestication helps to explain some of the extreme variation seen between modern maize lines, some of which differ in genome content by as much as 50% (Lu et al. 2015). Beyond repetitive TE content, maize is also known to harbor high levels of other diverse classes of repetitive DNA content. While the maize genome is known to be highly polymorphic, reduced selective pressure at repetitive loci and the propensity for unequal crossing seen within repeat arrays results in especially high levels of variation in sequence and copy number for these repetitive sequences. Some cytogenetics, genetics, and a few genomics studies have documented variation for many of high repetitive sequences among maize lines as well during evolution (Schnable et al. 2009, Wolfgruber et al. 2009, Bilinski et al. 2015, Schneider et al. 2016). Several major classes of repetitive DNA previously identified in maize are ribosomal DNA (rDNA), knob repeats, centromere satellite C DNAs (CentC), telomere repeats, various retrotransposon families, including centromeric retrotransposons, and knob repeats. While some of the repeats, such as the knob repeats, have no known functionality for the organism, others may play crucial structural roles, for instance telomeres serving as the ends of chromosomes, or centromere repeats ensuring proper chromosome segregation during cell division.

Centromeres are primarily made up with tandem satellite repeated CentC and interspersed centromeric retrotransposons of maize (CRM), both of which exhibit varying abundance across

taxa. Cytological evidence indicates that CRM elements, as the name implies, are largely located at centromeres (Lamb et al. 2006). Recently, studies using next-generation sequencing (NGS) data discovered that the abundance of CentC repeats is reduced in domesticated maize, while the contents of CRM are increased in domesticated maize, in comparison with the wild progenitor teosinte (Bilinski et al. 2015, Schneider et al. 2016). Maize centromeres are especially large, typically several million base pairs in length (~2-10 Mb) although the functional centromeres may be smaller than the total repetitive centromere sequence. Centromeres are hallmarks of eukaryotic biology, serving to delineate the position of the functional centromere and location of kinetochore assembly (required for chromosome segregation during meiotic and mitotic cell division). Despite this, little to no sequence homology is seen in between different species, suggesting a sequence independent mechanism for centromere function. How function might be conserved despite lack of sequence conservation has intrigued scientists, leading them to dub the phenomenon the "centromere paradox." Epigenetic factors are therefore thought to be critically important to proper positioning of the centromeres, which are functionally delimited by the presence of CENH3, a centromere-specific histone, which in maize has shown to localize to both the centC satellite repeats, as well as to the CRM elements (Zhong et al. 2002, Dawe 2003).

Telomeres are the natural ends of eukaryotic chromosomes. Telomere repeats typically consist of 5 to 8 nucleotide highly conserved motifs, which function to recruit the proteins of the nucleoprotein complex and protect chromosomes from instability. In most plants, the conserved motif is TTTAGGG (McKnight et al. 2004, Yu et al. 2006). Sub-telomeres are DNA sequences immediately adjacent to the telomere repeats. Hybridization, using telomere-specific probes, revealed that telomere lengths vary within a range of more than 25-fold among 22 surveyed maize

inbred lines. Genetic mapping analysis mapped additional *in trans* elements that control telomere length (Burr et al. 1992). Maize sub-telomeres consist of highly repetitive tandem sequences (Li et al. 2009). Here, telomere will be used as a general term for both telomere and sub-telomere repeats.

rDNA repeats are functional coding repetitive elements of genomes common to all organisms, although their copy is known to be especially high in the maize genome. rDNA is transcribed into rRNA, which serve a catalytic role in ribosomal translation of mRNA into protein. While 4 total RNA units are found in ribosomes, there are two major classes of rDNA repeats, either the 45S encoding rDNA locus or the 5S encoding locus. The former is transcribed as a single transcript (45S transcript), which is further processed into 18S, 5.8S and 26S mature rRNAs (**Figure 2**). These are subsequently assembled with the 5S rRNA into the ribosomal subunits within the sub-nuclear domain known as the nucleolus (Layat et al. 2012). Each repeat class is tandemly arrayed in distinct genomic regions. 5S rDNA loci are physically located at the distal of the long arm of chromosome 2 (Li et al. 2001), while 45S rDNA tandem arrays are clustered at the nucleolus organizer region (NOR) located at the short arm of chromosome 6 in maize (Phillips et al. 1974). The tandem arrays making up the NOR are the site of nucleolus formation, allowing direct localized processing of the 45S rRNA transcripts and assembly of ribosomal large and small subunits previous to nuclear export of these. Demand for high concentrations of ribosomes within the cell is reflected by the high copy number of the ribosomal RNA (rRNA) 45S operon and 5S unit. The copy number of 45S rDNA repeats is highly variable between different maize lines, possibly due to unequal crossover within large tandem repeats (Buescher et al. 1984). 5S rDNA

loci, in contrast, appear relatively stable (Rivin et al. 1986). Epigenetic regulation of the rDNA locus is dynamic and tightly controlled in many organisms.

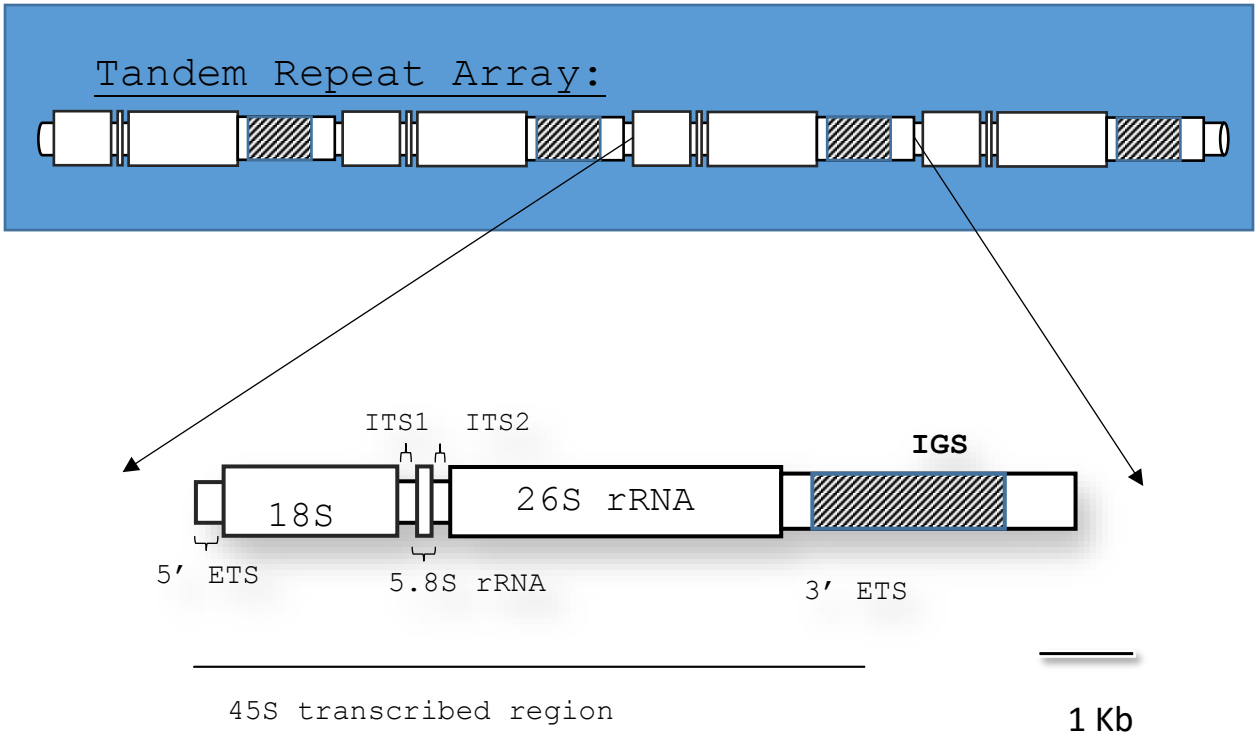

Figure 1.2: Schematic of rDNA organization of 45S producing repeat element
Shown is the External Transcribed Regions (5' and 3' ETS, RNA pol 1 promoter), Internal Transcribed Region (ITS) 1 and 2, and the IGS (Intergenic Spacer), as well as the 18S, 5.8S and 26S portions of the unit. Box with grey bars within the IGS represents complex repetitive region characteristic of the IGS.

Cytogenetic characterization of metaphase chromosomes has demonstrated the presence of large heterochromatic regions across numerous eukaryotic genomes and maize is no exception in this regard. Large conspicuous heterochromatic structures known as knobs were first observed in maize by Barbara Mcclintock. These structures are highly variable in terms of location within the genome

(having been observed at at least 34 different cytological positions within maize and teosinte) and in number, but most maize lines posses between 4 and 8 knob regions per haploid genome, typically at mid-chromosome arm positions (Ghaffari et al. 2013). Considering both extremes of the spectrum, "knobless" lines have been isolated, while a line from the Mescalero Apache Tribe known for high knob content possesses 13 knob regions based on cytogenetic evidence. In addition, our results (**chapter 2**) suggest that additional smaller clusters of knob repeats exist which are too small to detect through microscopy. Typically, however, knobs are multi-megabase structures, often visable through much of the cell cycle. Because of this, they have been successfully used as cytological markers.

Knobs consist of tandem repeat arrays of either a 180 bp repeat or the more recently described (Ananiev et al. 1998a) 350 bp "TR-1" repeat. Ananiev and coworkers observed TR-1 repeats in oposite orientation to each other, causing them to speculate that knob repeats might have the capacity to gain genomic mobility and transpose themselves to other regions of the genome. Precedent for this type of phenomenom comes from "Mega-transposons" in drosophila known to behave in this fashion, and this would help to explain the wide spread over which knob regions have been observed within the maize genome. Some levels of homology are found between the TR-1 and the 180 bp repeat (two subdomains of 31 and 12 basepairs between the repeat subunits retain around 70% homology each) and it is therefore thought that perhaps the knob repeats evolved from a common ancestor. In some cases, knobs are known to consist entirely of one repeat or the other, or to be a mixture of the two repeat types. In the latter case, the repeats are still clustered together tandemly, maintaining separate subdomains within the knob. Some observations suggest that exclusively TR-1 containing knobs tend to be much less frequent than 180bp repeat

exclusive knobs, which were observed at around the same frequency as mixed knobs (Hiatt et al. 2002).

While the persistance of some forms of repetitive DNA in the genome can be understood in the context of its function (telomere, centromere and rDNA repeats, as already discussed), the persistance of knob repeats within the genome can be less intuitively understood given lack of known function for these repeats. It is thought that the presistance of these repeats is in fact not due to their importance to the host, but rather due to their ability to cause segregation distortion as selfish genetic elements within the miotic drive system. Functionally, it is known that this process involves the capacity for knobs to function as centromeres in the presence of abnormal chromosome 10 (Ab10), which is an additional sequence on the end of chromsome 10 containing genes required for centromere activity at the knobs (termed "neocentromeres") and consequential segregation distortion. Specifically, within this system, neocentromeric knobs are pulled towards the spindle poles ahead of native centromeres during anaphase 1 and 2. As a result of rapid pole-ward movement, knob containing chromosomes are much more likely to be found in the upper and lower spores of a linear tetrad, and therefore to be found in the sole surviving megaspore in female flowers. It is thought that this system has allowed the knob elements to be maintained within the maize genome and that of its wild relatives (Hiatt et al. 2002). In addition, larger knobs have greater neocentromeric capacity, resulting in selection for increased knob size and helping to explain their remarkable size.

Novel K-mer Approach to Facilitate Unbiased Genome Analysis

NGS has provided in depth sequence data. However, accurate assessment of genome structure and dynamics of repetitive sequence evolution using large NGS datasets remains challenging due to the difficulty of unambiguous genome mapping and of accurately reconstructing repetitive sequences with high-copy number. Additionally, analysis relying on mapping reads to a reference assembly is subject to ascertainment bias. Analysis independent of a reference genome sequence could reduce biases of genome comparisons. In our study we quantify and characterize genome dissimilarity through the comparison of k-mer abundances directly determined from near-raw sequencing data.

Kmers are short substrings of larger fragments of DNA sequence, typically short read sequencing data, of fixed length K. For example, in the case of a 100 basepair sequencing read, one could generate kmers from this read of size K. Starting from the beginning of the read and moving towards its end moving 1 basepair at a time, it is possible to generate multiple unique kmers in this way. At most, there are N – k + 1 unique kmers that can be generated, so in the case of the 100 bp read, we would be able to generate 100 – 25 + 1 = 76 possible unique 25-mers, as long as there are no repeats within the read. Repeats within the read would result in a reduction of total unique k-mers from within that read, while the repetitiveness of the sequence would be reflected by the higher frequency of k-mers originating from the repetitive sequence. The total number of distinct kmers possible for k-mer length k is n^K, with n being the total number of possible letters at each site (in the case of nucleotide sequence, 4). Therefore for a nucleotide sequence of 25 bps in length, there are 4^{25} or $\sim 1*10^{15}$ possible kmers.

K-mer decomposition of sequencing data and analysis of said k-mers has been widely applied in many genomic analyses and bioinformatics tools, such as genome assemblies, genome characterization, and metagenomic analysis (Compeau et al. 2011, Williams et al. 2013, and Guo et al. 2015). It is often advantageous to reduce large complex data sets in such a way in order to reduce the overall complexity of the data as well as the computational burden associated with analysis. De bruijn graph assembly algorithms are a classic example of k-mers in assembly. These algorithms begin by generating k-mers from raw sequencing data, and then removing redundant k-mers in order to reduce the overall computational burden. Overlaps between k-mers are then used to construct a de bruijn graph, consisting of connections between k-mers based on overlap between these. Unambiguous connections then allow extension of sequence, and assembly proceeds on this basis. K-mers are also utilized heavily by different alignment algorithms. The Basic Local Alignment Search Tool (BLAST) for example, generates a k-mer based hash table for the genome which is used as the database for the alignment during the first stage of the alignment algorithm (Altschul et al. 1990). During this stage, a seed from the query is matched within the hash table, which contains indexed k-mers. Storing the genome in a database in this way reduces the number of initial comparisons which must be made dramatically, improving the speed of the search such that it is feasible to perform otherwise intractable analysis given computational and time limitations.

More recently, some novel applications of k-mer analysis have been proposed, notably for direct comparison of mutant and wildtype individuals for the purpose of identification of mutations between these (Nordstrom et al. 2013). In the past, forward genetic screens relied on intensive

mapping strategies followed by complementation studies in order to identify mutations responsible for a phenotype of interest. Advances in sequencing technologies have facilitated more rapid means for mapping mutations, for instance methods such as bulk segregant analysis seq. However, these methods are still hindered by the requirement for a quality reference genome sequence. By using kmer decomposition of near-raw, unassembled sequencing data (quality and adaptor trimming or other error correction methods are still necessary), it is possible to identify causative mutations more efficiently and rapidly, and bypass potential errors associated with assembly and alignment. Alignment-free methods are especially useful in the case of non-model organisms for which quality reference sequence is not available. Between sample comparisons through re-sequencing efforts rely on high levels of sequence homology and genic synteny between the reference and subject of interest and so are typically confounded by comparisons within species that show high levels of sequence divergence. In order to overcome these hurdles to mapping mutations using sequencing data, Nordstrom et al. proposed a k-mer analysis based comparison for direct comparison of mutant and wild-type individuals using sequencing data generated for each. Their method, described as NIKS (Needle in the K-stack), relies on analysis of kmer frequencies to identify mutations. Our study is an extension of the analysis proposed in their work. However, our analysis involves more radically divergent samples, and the aim of our analysis was comparative rather than functional. Given that the maize genome primarily consists of repetitive DNA, that these portions of the genome are the most difficult to resolve using a traditional assembly based methods, and that the k-mer method is well suited for this type of analysis, we chose to focus on the highly abundant portions of the maize genome, represented by the highly abundant k-mers derived from the sequencing data. We use maize lines B73 and Mo17 to this end.

The inbred lines B73 and Mo17 represent two of the most appreciated models for understanding maize genome diversity with respect to small-scale polymorphisms (Barbazuk 2007, Liu et al. 2010, Fu et al. 2006) and large-scale structural variation (Springer et al. 2009, Belo et al. 2010). In addition, mapping populations of inter-mated B73xMo17 recombinant inbred lines and double haploids have been generated to facilitate genetic analyses (Liu et al. 2012, Liu et al. 2015). Numerous comparative genomics studies of other maize cultivars and wild ancestors have examined the origin of maize as well as events of adaptation and artificial selection (Chia et al. 2012, da Fonseca et al. 2015, Hufford et al. 2012, Jiao et al. 2012, Jin et al. 2016, Swanson-Wagner et al. 2010, van Heerwaarden et al. 2011). However, the studies are limited to comparisons of non-repetitive and low-repetitive sequences. Using B73 and Mo17 whole genome shotgun (WGS) sequencing data, we quantified the level of the difference between the two genomes at both non-repetitive and highly repetitive genome sequences. Genomic locations influencing variation in copy number at highly repetitive sequences were genetically mapped using WGS sequencing data of 280 intermated B73 and Mo17 double haploid (Liu et al. 2010). Furthermore, highly variable k-mers in diverse lines using *Zea mays* HapMap2 WGS data (Chia et al. 2012, Hufford et al. 2012) were identified, revealing significant changes on highly repetitive sequences during maize domestication and improvement.

Chapter 2 - K-mer analysis of B73 and Mo17 Genomes

K-mer analysis of genome dissimilarity between two maize inbred lines

B73 and Mo17 are two maize elite inbred lines that are widely used in maize genetic and genomic research. The two genomes have been extensively compared in both small and genome-wide scales (Barbazuk et al. 2007, Liu et al. 2010, Fu et al. 2006, Springer et al. 2009, Belo et al. 2010). However, previous studies largely relied on a reference genome, which produces systemic biases. To perform genome comparison with an unbiased k-mer method that is independent of the reference genome, two HiSeq2500 lanes of Illumina data, using PCR-free prepared DNA libraries, were generated for each of the two maize inbred lines B73 and Mo17, resulting in 450.9 and 445.3 millions of pairs of 2x125 paired-end reads, respectively. More than 99% reads were retained after the adaptor and quality trimming. The genome coverage of sequencing data (~46x) for each genome enable the employment of error correction of sequencing reads. We use abundance to represent counts of k-mer from sequencing data and use copy number to represent sequence copies in a genome. The corrected reads were subjected to 25-nt k-mer counting, resulting in approximately 749.7 and 738.7 millions of non-redundant k-mers for B73 and Mo17, respectively. The similar shapes of the distributions of k-mer abundances (Fig. 2.1A) and the curves of cumulative contribution of k-mers with different abundances to the genomes (Fig. 2.1B) indicate that B73 and Mo17 exhibit overall similar levels of genome complexities. The B73 and Mo17 abundance peaks are presumably located at in single-copy k-mers (http://www.broadinstitute.org/software/allpaths-lg/blog/wp-content/uploads/2014/05/KmerSpectrumPrimer.pdf), which occur only once in a genome (Fig. 2.1A). The merged B73 and Mo17 k-mer abundances form a curve with two peaks in k-mer abundances (Fig. 2.1A). The lower abundance peak underneath the original uncombined peaks

consists of k-mers specific to either B73 or Mo17, while the second higher frequency peak represents the common k-mers of the two genomes. This novel approach was employed to visualize the difference of non-repetitive genomic sequences between the two genomes. K-mer comparison indicates that only 60.9% of single-copy k-mers are shared between the two maize cultivars, leaving a remaining 39.1% of the single-copy k-mers specific to each genome (Table 2.1). Based on the k-mer distribution, the B73 genome size was estimated to be 2.38 Gb and consisted of 24.9% single-copy k-mers, while 2.48 Gb with 23.7% single-copy k-mers for Mo17. The B73 genome size estimation agrees with that of 2.3 Gb estimated from the B73 genome sequencing project (Schnable et al. 2009). The slightly larger estimated genome size of Mo17 versus B73 but the smaller proportion of single-copy sequences in Mo17 implies that distinct contributions of repetitive sequences to two genomes, which indeed can be observed on the curves of cumulative k-mer contribution to the genome at high abundant k-mers that are representatives of highly repetitive sequences (Fig. 2.1B).

Table 2.1: k-mers from single-copy regions in B73 and Mo17[*]

Category	Number of single-copy k-mers	% single-copy k-mers in either B73 or Mo17[#]
B73 & Mo17 common	285,759,048	-
B73 specific	183,644,569	39.1%
Mo17 specific	183,441,358	39.1%

[*] k-mers with counts between 20 and 50 are considered to be single-copy k-mers

[#] percentage of genotype specific k-mers in all single-copy k-mers in either B73 or Mo17

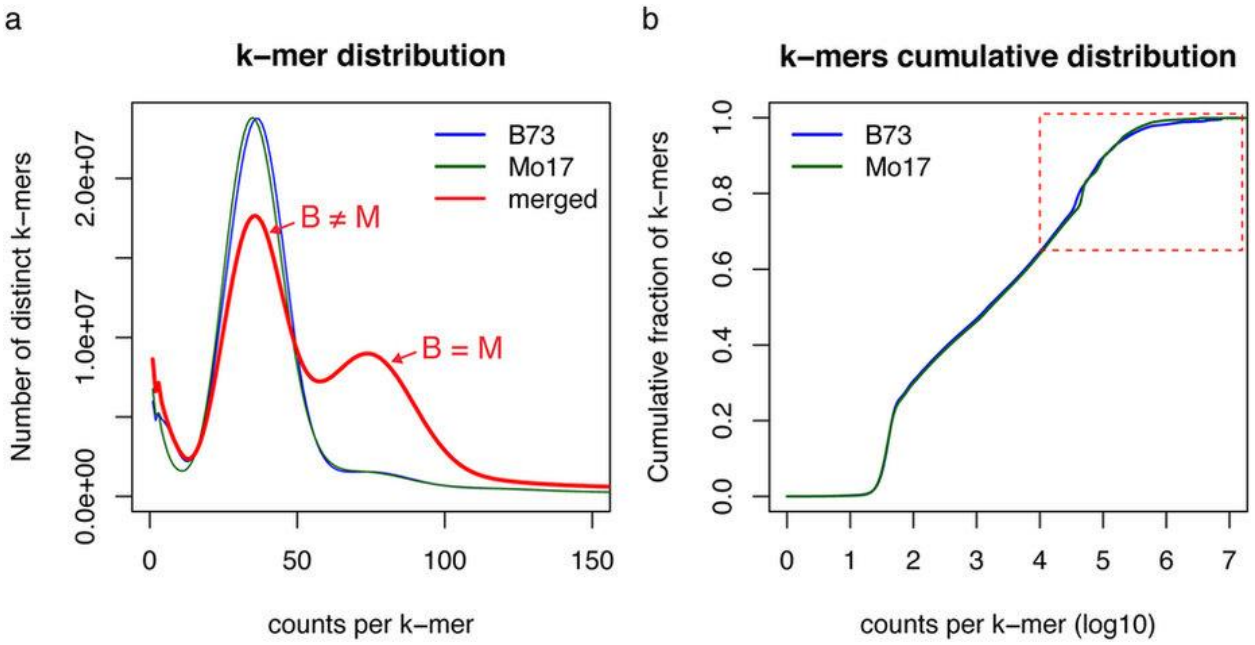

Figure 2.1: Comparison of k-mer spectra in B73 and Mo17.

(a) Distributions of k-mers at different abundance in B73, Mo17, and merged B73 and Mo17. Merged k-mer counts are the total counts from both B73 and Mo17. Only the range of 1–150 on the x-axis was plotted to show the distribution of low-copy k-mers. (b) Cumulative fraction of k-mers of different abundance in each genome. The dashed-line box highlights high abundance k-mers.

Divergence in copy number exhibited on highly repetitive DNA sequences

Owing to the implication of the distinct constitution of high-copy genomic sequences between B73 and Mo17, highly abundant k-mers (HAKmers, N= 802,668) in either B73 or Mo17 or both were examined. The majority of HAKmers exhibit similar abundance in the two genomes but some are highly different (Fig. 2.2A). Functional annotation through a BLASTN of HAKmers to a *Zea mays*

repeat database results in 552,371 annotated HAKmers each of which has at least one hit with the minimum e-value of 0.1. The best hit of each HAKmer was referred to as the k-mer's functional class. The major classes include retrotransposon, knob, rDNA, CentC, telomere, and a variety of DNA transposon members (Table S1).

χ^2 statistical tests with a multiple test correction using the cutoff of 5% false discovery rate (FDR) were performed to identify HAKmers showing differential abundance between B73 and Mo17. A minimum of two-fold change in k-mer abundance was also required. As a result, 11,413 and 2,633 differential abundance HAKmers respectively showing higher abundance in B73 and Mo17 were identified, and, hereafter, referred to as B73-gain and Mo17-gain HAKmers. Four major functional annotation classes, knob, 45S rDNA, CentC, and telomere, were found in these differential abundance HAKmers (Fig. 2.2B). Although retrotransposon derived k-mers (retrotransposon k-mers hereafter and a similar expression was applied to other classes of k-mers, e.g., 45S rDNA k-mers to represent k-mers derived from 45S rDNA) represent the largest class of HAKmers, relatively few of these differ significantly in abundance (Table S1). Many knob k-mers were identified and all belong to B73-gain k-mers, indicating more knob sequences in the B73 genome. This is consistent with the previous cytological observation that B73, but not Mo17, contains knobs at the long arms of chromosomes 5 and 7 (He et al. 2014, Kato et al. 2004). Despite the changes in the knob content detected, no differential abundance HAKmers were found to be TR-1 repeats. A similar finding was made for B73-gain telomere k-mers although the number is much smaller (Fig. 2.2B). Moreover, a number of k-mers derived from 45S rDNA and CentC show gains in either B73 or Mo17. More 45S rDNA k-mers and less CentC k-mers showing higher abundance were identified in B73 versus Mo17. Genomic locations of these differential abundance HAKmers

on the B73 genome were mapped through aligning k-mers to the B73 reference genome (B73Refv3) (Fig. 2.2C). From the result, knob k-mers are clustered on multiple chromosomes (e.g., long arm of chromosomes 1, 4, 5, 7 and a distal short arm region at chromosome 9), CentC k-mers are largely located at or around centromeres, and telomere k-mers are identified at the ends of chromosomes 1, 2, 4, 8 and 10. 45S rDNAs k-mers are predominantly clustered at the short arm on chromosome 6, presumably the NOR. Note that such distributions based on the reference genome rely on the quality of assemblies, and the assembly quality of different regions might vary. The genome distribution plot also shows that 45S rDNA k-mers are pervasive in other genome regions in addition to the NOR (Fig. 2.2C, Fig. A.1).

To understand copy numbers of different classes of highly repetitive sequences in two genomes, the total count of all the k-mers of each class was determined and normalized, which represents the relative level of repetitiveness of each class. As a result, compared to B73, approximately 55% and 22% reduction were respectively observed on knob and telomere repeats, while 71%, 34%, 25% increased on CentC, 45S rDNA and 5S rDNA, respectively, in Mo17 (Table S2). We also used abundances of the k-mers (N=3,533) from the 45S rDNA regions conserved among multiple plant species to estimate the copy number of 45S rDNA (Methods). The copy numbers of 45S rDNA in B73 and Mo17 were estimated to be around 3,658 and 5,063, respectively. Our estimation is in the range of a previous estimation of placing rRNA gene number from 2,500 to12,500 in 16 maize lines (Buescher et al. 1984). Collectively, we discovered several major classes of repetitive sequences showing differential copy number between B73 and Mo17, suggesting that two genomes experience pronounced divergence with respect to copy number of highly repetitive sequences. Because these repetitive sequences are largely clustered and tandemly arrayed, high

levels of copy number variation at these loci are likely caused by insertions or deletions of large genomic segments due to aberrant crossing over or replication errors.

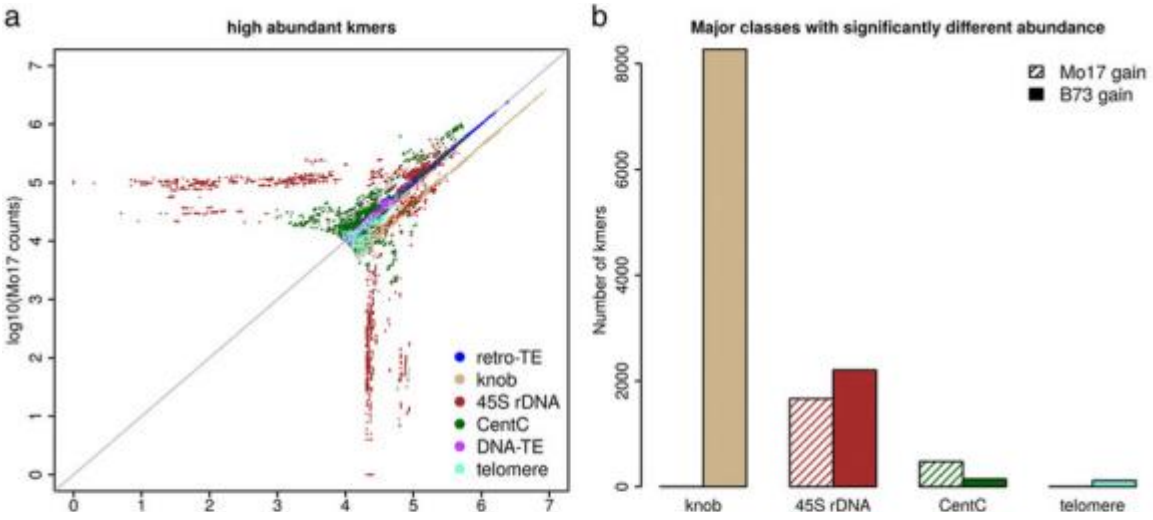

Figure 2.2: Comparison of high-copy k-mers between B73 and Mo17.

(a) A scatter plot of counts of high abundance k-mers from error corrected WGS reads. K-mers were annotated by BLASTN to the maize repeat database. (b) Four major repeat classes containing k-mers that exhibit statistically significant differential counts and at least two-fold changes between B73 and Mo17, were shown. Two types, Mo17-gain and B73-gain, respectively represent more counts in Mo17 and more counts in B73. (c) Genome-wide view of the distribution on the B73Ref3 reference genome of k-mers with differential B73 and Mo17 counts. All perfect hits each of which is end-to-end and 100% matching to the reference genome were used for determining the number of hits per bin (100 kb). The number of hits in each bin with at least 10 hits was plotted versus bin physical locations on the B73Ref3. Different functional groups were color- and shape-coded.

Genetic mapping of genomic locations showing differential copy number of repetitive sequences between B73 and Mo17

Differential abundance of HAKmers from B73 and Mo17 results from distinct copy numbers of genomic repetitive sequences from which k-mers have originated. The segregation of such genomic sequences in a segregating population (e.g., recombinant double haploids) derived from B73 and Mo17 results in different copy number among the offspring. To map genomic locations showing the differentiation of copy number between B73 and Mo17, low-coverage WGS sequencing of 280 individuals from inter-mated B73xMo17 double haploids (IBM DHs) (Liu et al. 2015) was analyzed. First, the abundance of each of differential abundance HAKmers from each DH line was determined and normalized (Methods). K-mer abundance resembles a quantitative trait value, and the genomic elements contributing their genomic copy number variation can be genetically mapped using a quantitative trait locus (QTL) mapping approach (referred to as copy number variation QTL, cnvQTL, hereafter). Using a high-density genetic map developed with the same WGS data set from these 280 DH lines (Liu et al. 2015), the normalized counts of a k-mer were input as phenotypic values for a genetic mapping analysis using the R package rqtl. In total, 11,413 and 2,633 of B73- and Mo17-gain HAKmers were analyzed, respectively. To determine the cutoff of log10 likelihood ratio (LOD) of cnvQTL, each of 1,000 randomly selected HAKmers was subjected to a permutation test to determine the LOD cutoff. All of these LOD cutoffs with the 5% type I error are in between 3 and 4. Therefore the minimum LOD of 4 was used to declare mapping cnvQTL peaks (Table S3). Only 0.3% B73-gain and 3% Mo17-gain HAKmers could not be mapped using this approach. The majority of HAKmers, 74.5% B73-gain and 83.5% Mo17-gain, were mapped to single major genomic locations, and the rest were mapped to 2-4 genomic locations.

Functional annotation analysis of these mapped HAKmers revealed distinct mapping locations for different sources of k-mers (Fig. 2.3). For B73-gain HAKmers, knob and 45S rDNA are two major sources (Table S4). Knobs k-mers were mapped to the long arms on chromosomes 1, 5, and 7, of which the regions on chromosomes 5 and 7 were reported to have differential knobs between B73 and Mo17 (He et al. 2014, Kato et al. 2004). All 2,205 45S rDNA k-mers were mapped to around 13.5 Mb on chromosome 6 to which 11 retrotransposon k-mers were also mapped. This mapping region is located at a short arm region on chromosome 6, which exhibits a presence-and-absence variation (PAV) that was identified in previous comparative studies (Springer et al. 2009, Belo et al. 2010). Substantial copy gains of some type of 45S rDNA and some retrotransposons in B73 at this region indicate the long PAV segment harbors rich repetitive sequences. The differential abundance 45S rDNA k-mers are largely located at the intergenic spacer (IGS) between 18S and 26S of 45S rDNA and a small proportion are located at internal transcribed spacer (ITS) and 26S rRNA gene (Fig. A.2). On the same chromosome, CentC k-mers were mapped to 62.8 Mb, suggesting the two genomes contain distinct centromere compositions on chromosome 6. Moreover, telomere k-mers were mapped to the ends of short arms of chromosomes 1, 2, 4, and 5. The further analysis shows that B73 contains more copies of telomere repeats than Mo17 at chromosomes 2, 4, 5, but less copies at chromosome 1.

45S rDNA and CentC are two major sources for Mo17-gain HAKmers (Table S5). Interestingly, similar to B73-gain 45S rDNA HAKmers, Mo17-gain counterparts were mapped to around 13.6 Mb on chromosome 6, although a long DNA segment on the B73 reference genome around that region is absent in Mo17. This indicates that B73 and Mo17 likely contain different versions of

45S rDNA at the NOR. Furthermore, four 5S rDNA k-mers (N=4) showing higher abundance in Mo17 were mapped to around 222.5 Mb on chromosome 2, consistent with a previous FISH result in which 5S rDNA was mapped to the distal of chromosome 2 (Li et al. 2001). Significantly, Mo17-gain CentC k-mers were mapped to multiple chromosomes. The centromeric regions at chromosomes 2, 4, 7, 8, and 9 contribute to varying abundance of CentC k-mers. The same k-mers can be mapped to the centromeres on multiple chromosomes, suggesting multiple centromeres co-evolved to change CentC abundance.

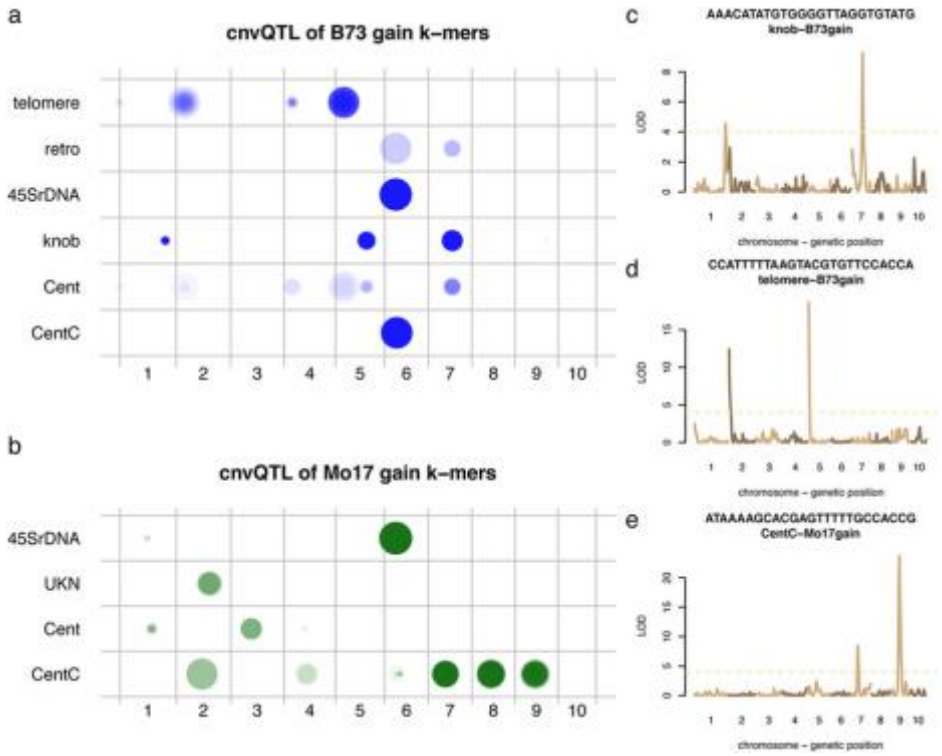

Figure 2.3: cnvQTL mapping of genomic locations contributing differential abundance of HAKmers.

WGS of 280 IBM DHs was used to determine abundance of differential abundance HAKmers. A QTL approach was employed to map genomic locations influencing k-mer abundance in DH lines. (a,b) The mapping results of B73-gain HAKmers (a) and Mo17-gain HAKmers (b) were plotted for each annotated class. A mapping location of each k-mer is designated by a dot. Transparent factor (0.02) was used for a dot from each k-mer. The sizes of dots represent the logarithm 10 scaled LOD values from QTL analyses. retro, Cent, and UKN represent retrotransposon, centromere elements, and unknown elements, respectively. (c,d,e) Three examples of the QTL results of knob B73-gain (c), telomere B73-gain (d), and CentC Mo17-gain HAKmers (e), are shown.

Different evolutionary origins of 45S rDNAs of B73 and Mo17, likely expanded, and spread to regions other than the NOR after domestication

From differential abundance HAKmers, an extreme type of k-mer was surprisingly observed in which the k-mer was highly abundant in B73 or Mo17 but absent or very low in the other, which are referred to as genotype-specific HAKmers (Fig. 2.4A). In total, 162 B73-specific HAKmers and 103 Mo17-specific HAKmers were obtained. These genotype-specific HAKmers were verified by using independent B73 and Mo17 WGS sequencing data (Chia et al. 2012) without error correction. Additionally, all of the B73-specific HAKmers can be perfectly aligned to the B73 reference genome, while only 3/103 Mo17 specific HAKmers were perfectly aligned to single locations at the NOR region. This result confirms, at least, that Mo17 specific HAKmers are highly abundant in Mo17 but hardly identified in the B73 genome. Interestingly, all of these genotype-specific HAKmers are annotated to the class of 45S rDNA. K-mer analysis using IBM DH lines WGS sequencing data indicates that each DH line predominated by either B73- or Mo17-specific k-mers (Fig. A.3). Genetic mapping analysis of both B73- and Mo17-specific HAKmers through cnvQTL shows that the NOR where 45S rDNA repeats are clustered is largely responsible for the segregation of B73- and Mo17-specific HAKmers, further suggesting that distinct types of high-copy 45S rDNAs are included at the B73 and Mo17 NORs (Fig. 2.4B). A detailed analysis found that all these genotype-specific k-mers were mapped to the IGS of the 45S rDNA unit.

To understand the origin of these genotype-specific k-mers, maize HapMap2 WGS sequencing data, which includes lines from teosinte, landrace, and improved maize (Chia et al. 2012, Hufford 2012), were subjected to k-mer analyses. The count of each of B73- and Mo17-specific k-mers was determined for each HapMap2 line. To account for the variation of k-mer abundance owing

to non-genetic factors, such as sequencing depth and organelle DNA contamination, a novel normalization approach was developed of which normalization factors were determined by using the total counts of a set of conserved single-copy k-mers across HapMap2 lines. Briefly single-copy k-mers were first obtained from both B73 and Mo17 and the correlation of counts of each k-mer with the library sizes of all the HapMap2 lines determined. Based on the assumption that a conserved single-copy k-mer exhibits a high correlation with the sequencing library size, the top 5% k-mers (N=49,955) with highest correlation efficiencies were used to calculate the normalization factors. A principal component analysis (PCA) was performed using normalized abundances of genotype-specific HAKmers (N=265) of HapMap2 lines. At a result, the first two components (PC1 and PC2) explain 72.4% variation in normalized abundance (Fig. 2.4C). From the PCA plot, three distinct branches were formed and teosinte lines were centralized at the intersection. Mo17 is located on the distal position of one branch but B73 is not located at any of the branches. The PCA analysis implies that not all the HapMap2 lines exhibit either of two extremely divergent patterns possessed in B73 and Mo17.

To understand the abundance of these genotype-specific HAKmers in each HapMap2 line, the total normalized counts of all the B73- and Mo17-specific HAKmers were separately determined. Total counts of the B73- and Mo17-specific HAKmers vary dramatically among the HapMap2 lines (Fig. 2.4D). It is notable that all teosinte lines exhibit relatively low abundance, while many but not all maize lines show high abundance in total counts. This result indicates that these particular types of 45S rDNA repeats likely experienced appreciable expansion after domestication or shrinkage in teosinte and some maize lines. Evidence was also found that B73-specific k-mers are largely, but not only, located at the NOR. Indeed, the B73-specific k-mers can be identified at

many locations on all the chromosomes in the B73 genome (Fig. 2.4A). Presumably, the scattered distribution of these k-mers across all the chromosomes is the consequence of the 45S rDNA spreading from the NOR. Moreover, all teosinte lines and the majority of maize lines contain only either B73- or Mo17-specific HAKmers, while a few landrace and improved lines consist of both. Our cnvQTL mapping result indicated that both B73- and Mo17-specific HAKmers are predominantly located at the NOR. The observed mixture of two rDNA types in some maize inbred lines are likely the consequence of heterozygous residues or recombination at the NORs, although meiotic recombination is substantially suppressed at the NOR (Bauer et al. 2013). It is also notable that the proportion of lines with B73-specific types of 45S rDNAs in the improvement levels is increased from teosinte to landrace, and from landrace to improved lines (Fig. 2.4D), possibly due to positive selection on the NOR or nearby regions. Previous studies also suggested that this region was under selection during either domestication (Hufford et al. 2012) or maize improvement (Jin et al. 2016).

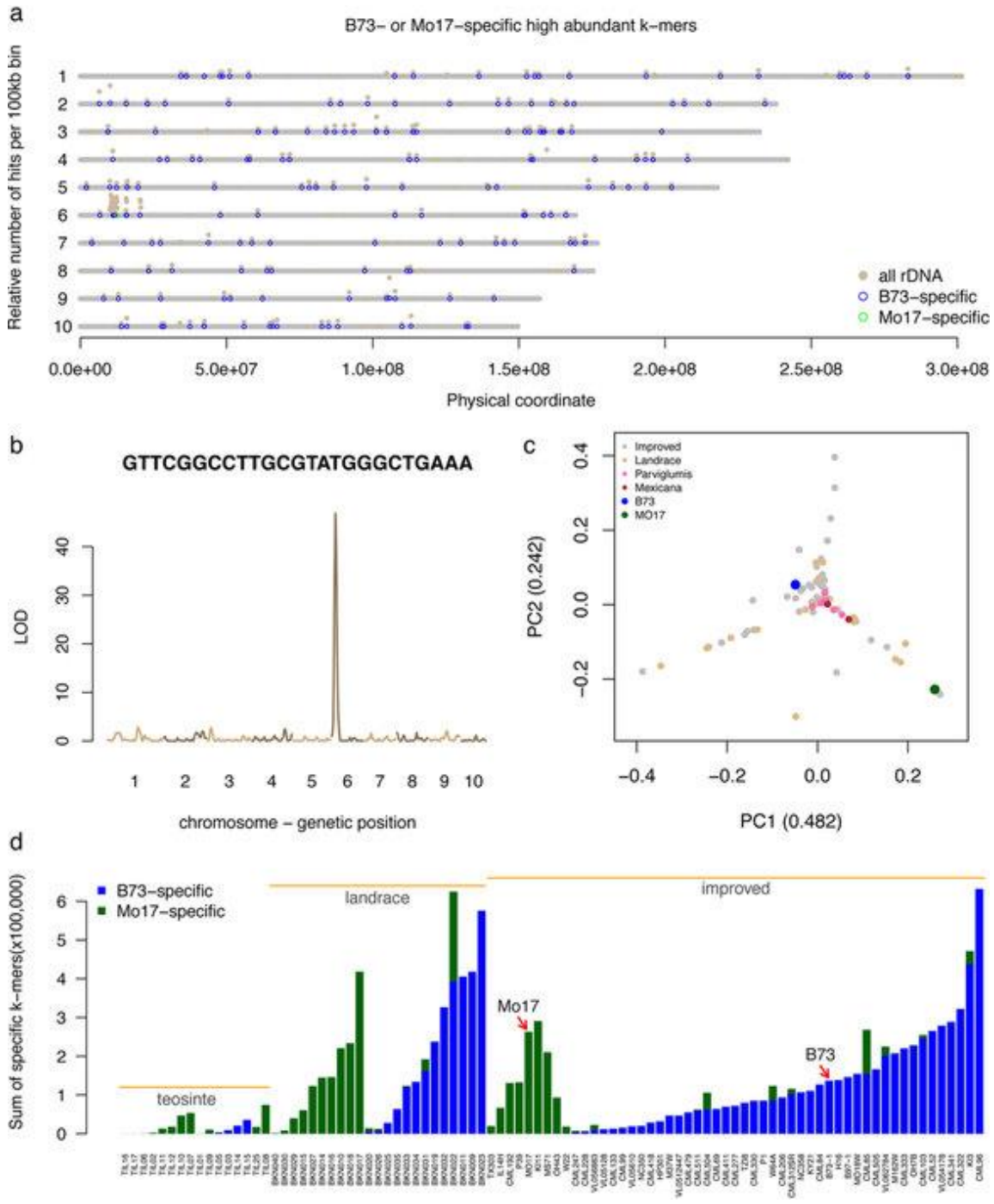

Figure 2.4: B73 or Mo17 specific HAKmers.

Genome-wide distributions of all rDNA-related k-mers, B73- and Mo17-specific k-mers that can be perfectly aligned to the B73Ref3 reference genome. Alignment counts (e.g., 100 was used to represent alignments of 10 k-mers and 10 hits per k-mer) per bin (100 kb) were plotted versus bin

physical locations at the B73Ref3. (b) Counts of a Mo17-specific k-mer in IBM DH lines were treated as a trait and the genomic loci (or locus) contributing the levels of counts in DH lines were mapped. (c) Principal component analysis (PCA) was performed using normalized counts of each B73- or Mo17-specific k-mer in multiple teosinte, landrace, and improved maize lines. Numbers in parentheses are percentages of variation in normalized counts explained by principal component (PC) 1 and 2. (d) Sum of normalized counts of all B73-specific k-mers (blue) or Mo17-specific k-mers (green) in different lines from HapMap2 WGS sequencing data without error correction. Bars were sorted in the subspecies order, teosinte (first Parviglumis then Mexicana), landrace, and improved maize lines. Within each subspecies, bars were sorted by total counts of B73-specific k-mers first and then by total counts of Mo17-specific k-mers.

Allelic expression of 45S rDNA in hybrids of B73 and Mo17

The differences of 45S rDNA sequences in B73 and Mo17 enables the investigation of the expression of two types of 45S rDNA in the hybrid of B73 and Mo17. Messenger RNA (mRNA) is typically selected and enriched in final sequencing libraries in the regular RNA-Seq (mRNA sequencing) procedure. However, it is almost impossible to completely remove all rRNA, which allows the study of the expression of rRNA using mRNA sequencing data. Two sets of RNA-Seq data were used. One is the transcriptomic data of young maize primary roots in the B73, Mo17 and the reciprocal hybrids (Paschold et al. 2014). The other is transcriptomic data of whole kernels at 0, 3, and 5 days after pollination (DAP) and endosperms at 7, 10, and 15 DAP from reciprocal hybrids of B73 and Mo17 (Xin et al. 2013). From both data sets, many sequences were aligned to

45S rDNA, proving that rRNA sequences remained in mRNA sequencing data. The B73- and Mo17-specific 45S rDNA k-mers can be used to trace the genotype-specific expression of 45S rDNA if their k-mer abundance could be reliably measured in RNA-Seq. However, all these genotype-specific k-mers are located at the IGS. The IGS is either not transcribed or accumulated at a level as high as the rRNA genes (5.8S, 18S, and 26S), and IGS expression therefore cannot be reliably detected. Fortunately, a single-nucleotide variant (SNV), A/T, was discovered on the 26S rRNA gene and three pairs of k-mers harboring this SNV were identified in both genomic sequencing and RNA-Seq data (Table S6). 72% and 28% B73 rDNAs carry A and T, respectively, while almost 100% of Mo17 rDNAs carry T. A-carrying rDNAs nearly completely dominated rRNA expression in primary roots of B73 (Fig. 2.5A), suggesting that not all rDNAs, as previously reported (McStay et al. 2006), are transcribed. In Mo17, T-carrying rDNA is the only type of expressed rDNA. In the reciprocal hybrids, both types were almost identically expressed in primary roots, although in both the reciprocal hybrids the A and T types of rDNAs are unequal in abundance in their genomes (Fig. 2.5A).

Using the time-course transcriptomic sequencing data of whole kernels and endosperms, the allelic expression levels of the SNV (A/T) on the 26S rRNA gene in the reciprocal hybrids of B73 and Mo17 were also examined (Fig. 2.5B). As a result, detected rRNA almost entirely belong to the maternal type in whole kernels at 0 DAP. Paternal rRNA accumulation levels were gradually increased in whole kernels from 0 to 5 DAP. In endosperms at 7, 10, and 15 DAP, the ratios of maternal to paternal rRNA expression are not far from 2:1 that is the actual copy number ratio of maternal to paternal genomes, indicating that both maternal and paternal rRNA copies are expressed at equal rates in early endosperms.

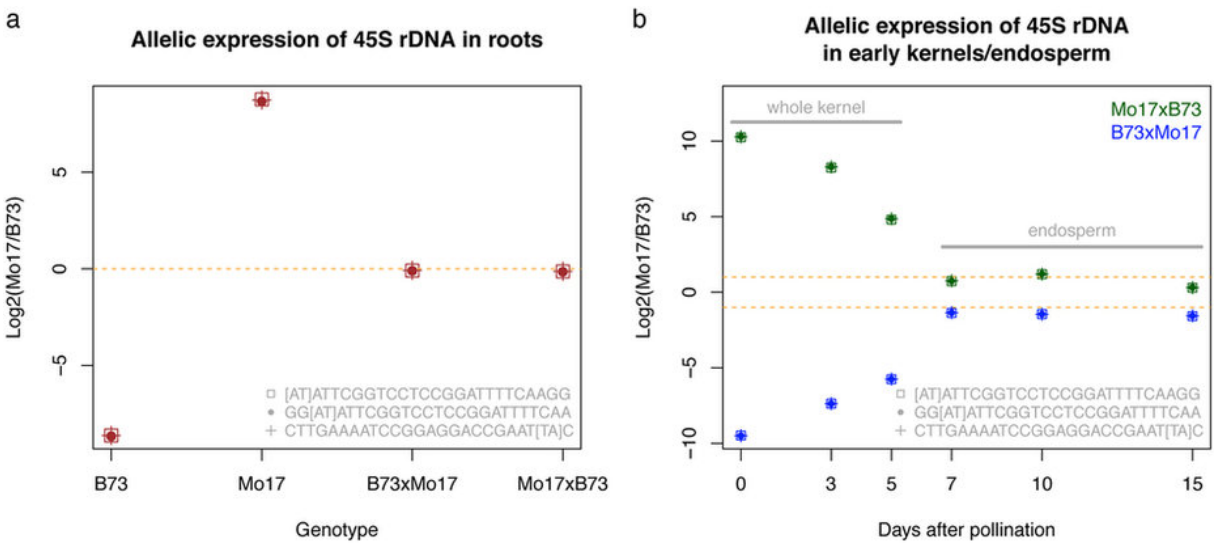

Figure 2.5: Allelic expression of 45S rDNA in hybrids of B73 and Mo17.

(a) A single-nucleotide variant was identified at the 26S rRNA gene of the 45S rDNA unit. Three pairs of k-mers harboring this single-base variant are listed in the figure. Two bases within square brackets represent the allele type highly enriched in B73 and Mo17 respectively (B73 k-mer and Mo17 k-mer). The log2 of the ratio of expression abundance of each Mo17 k-mer to that of its paired B73 k-mer was plotted for four genotypes, B73, Mo17 and the reciprocal hybrids. The expression data were from maize primary root RNA-Seq. Expression abundance is the average of four biological replicates. (b) The log2 of the ratios of expression abundance of each Mo17 k-mer to that of its paired B73 k-mers were determined for samples (whole kernel or endosperm) from different developmental stages, and plotted versus the days after pollination. The expression data are from maize RNA-Seq of B73 and Mo17 reciprocal hybrids. The reciprocal hybrids were plotted in either blue (B73 as the female parent) or green (Mo17 as the female parent).

Marked changes of multiple types of highly repetitive genomic sequences during domestication and maize improvement

The finding that B73 and Mo17 exhibit substantial variation at high-copy genomic sequences inspired an investigation of such variation among the HapMap2 lines. B73 and Mo17 are included in the HapMap2 lines but in this analysis we wanted to identify k-mers highly variable across the whole HapMap2 set, rather than the genotype-specific high abundance k-mers defined using these two inbred lines. Using the HapMap2 WGS sequencing data, k-mers showing high abundance (>1,000 counts per k-mer) in at least five HapMap2 lines but low abundance (<10 counts per k-mer) in at least five other lines were extracted, resulting in 8,462 highly variable k-mers. To examine the change of these k-mers among three evolutionary groups, teosinte, landrace, and improved, an ANOVA test was performed for each k-mer and a Bonferroni correction was conducted to account for multiple testing. As a result, 2,016 k-mers exhibit significantly differential abundance among three groups at the 5% type I error. Functional annotation through a BLASTN of k-mers to the repeat database results in 1,090 annotated k-mers (Methods). The k-mers exhibiting significantly differential abundance among evolutional groups were annotated to the functional classes of 45S rDNA, CentC, retrotransposon (copie and gypsy), and knob. The low rate (only ~54%) of k-mers that are annotated using the repeat database is because a relatively high proportion of k-mers are derived from organelle genomes, which likely reflects the diversity of organelle genomes. To focus on highly repetitive sequences from nuclear genomes, only the functionally classified k-mers were subjected to a clustering analysis using the software MCLUST (Fraley et al. 1999), resulting in 12 clusters (Fig. A.4). Nine major clusters were further manually grouped into two groups (Table 2.2, Fig. 2.6A, B). In detail, k-mer abundance of the group 1 was significantly decreased during maize domestication and/or improvement. K-mers from this group

are largely annotated as CentC (example in Fig. 2.6C) and 45S rDNA, as well as a small number of k-mers from knob, DNA transposons, and retrotransposons (Table 2.2). K-mer abundance of the group 2 was substantially increased during maize domestication and/or improvement. K-mers from this group are annotated as retrotransposon members (CRM and unclassified retrotransposon) (example in Fig. 2.6D) and 45S rDNA. The observation of 45S rDNA in both groups 1 and 2 suggests that some types of 45S rDNA sequences experienced substantial expansion while others experienced substantial shrinkage during maize domestication and improvement.

Table 2.2: Number of functionally classified k-mers in different clustering groups

Class	Decrease during domestication	Increase during domestication
CentC	212	0
CRM*	0	121
Knob	12	0
45S rDNA	266	81
DNA transposon	9	0
Retrotransposn$	54	138

* k-mers were annotated unknown centromere retrotransposons

$ unclassified retrotransposon

Abundance of k-mers that were generated from the conserved regions of 45S and 5S rDNA across multiple plant species was estimated for each HapMap2 line. The median of abundances of all the 45S rDNA k-mers from a HapMap2 line and the counterpart of 5S rDNA k-mers were used to represent the genomic copy number level of 45S and 5S rDNA of the line, respectively. Most

landrace maize lines exhibit lower copy number than teosinte, while maize improved lines shows much higher diversity in term of 45S rDNA copy number (Fig. 2.6E). This observation suggests there were a possible shrinkage or a strong selection on the NOR region during domestication, and a re-expansion of 45S rDNAs during improvement. No association with evolutionary groups was observed for copy number of 5S rDNAs. Additionally, the correlation of copy number of 45S and 5S rDNAs among HapMap2 lines is weak ($R^2 = 0.059$), suggesting that dosage balance in genomic copy number between 45S and 5S rDNAs, which was observed in human and mouse genomes (Gibbons et al. 2015), was not required in *Zea* genomes.

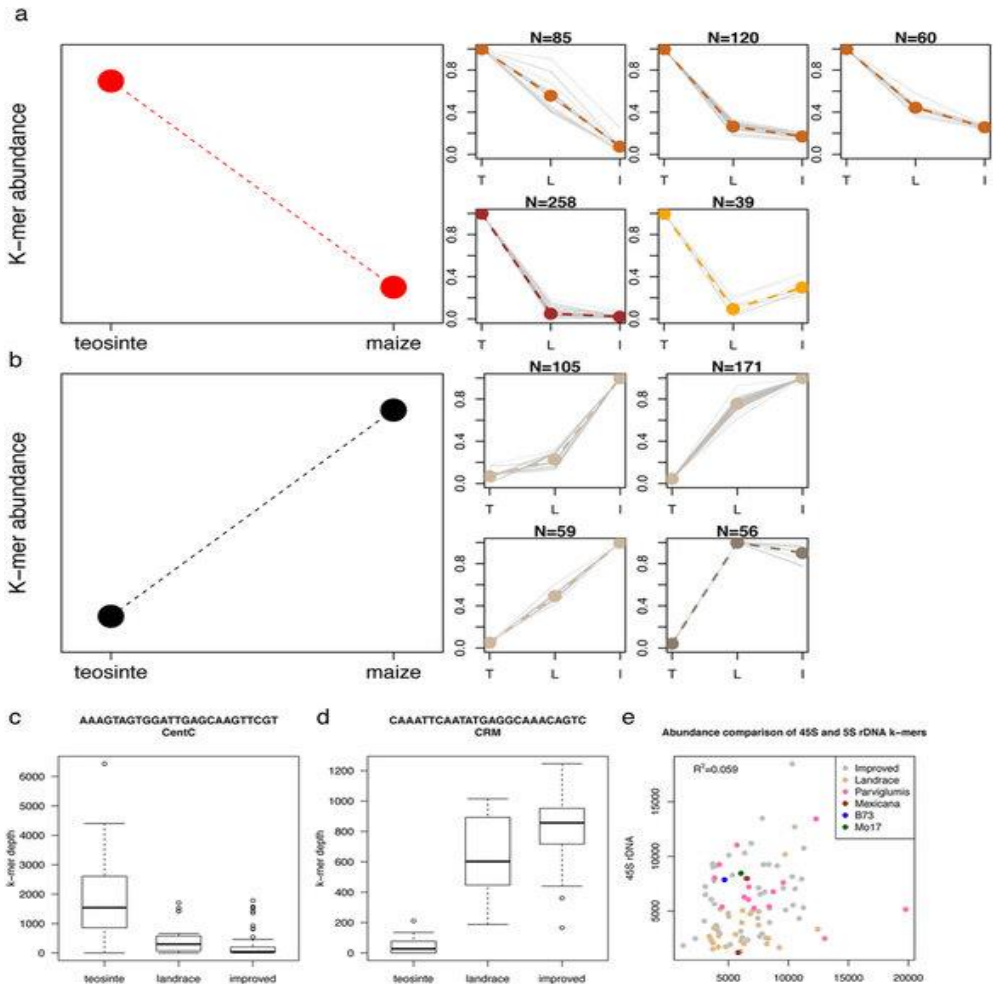

Figure 2.6: Change of k-mer abundances in teosine, landrace, and improved maize. (**a**,**b**). K-mers with significantly differential abundance in teosine, landrace, and improved maize were clustered. Nine major clusters were further manually divided into two groups. K-mers in group 1 (**a**) exhibit markedly higher abundance in teosinte relative to maize, while k-mers in group 2 (**b**) exhibit the opposite. Smaller plots provide the details of the clusters in each group. Each grey line in the smaller plots represents a k-mer. Colored lines are average values from all the k-mers in each cluster. Clusters with a similar pattern were highlighted by the same color. T, L, I on the x-axes represent teosinte, landrace, and improved lines, respectively. (**c**,**d**) Boxplots of three representative k-mers that are separately derived from CentC (**c**) and CRM (**d**). (**e**) The median abundance of 45S rDNA k-mers generated from the conserved 45S rDNA sequence in each HapMap2 line was plotted versus the median abundance of 5S rDNA k-mers generated from the conserved 5S rDNA sequence of the same HapMap2 line. Each dot represents a line, which is color-coded by genotype groups.

Discussion

This study employs a novel k-mer analysis strategy for comparative genomics. Reference-independent quantification of NGS data allows precise and unbiased comparison of the genomic constitutions, particularly highly repetitive sequences that are generally overlooked from regular analyses. Our results offer insightful information about copy number abundance, genomic locations, and evolution of highly repetitive sequences among maize genomes, and provide an

unbiased genome comparative method for mining existing and incoming deluge of NGS data to gain biological insights.

Unbiased k-mer analysis

K-mers represent all the possible subsequences of length k from a sequencing read. For genome assembly using short NGS reads, k-mers are typically generated from sequencing data to construct *de Bruijn* graphs (Compeau et al. 2011). In addition to genome assemblies, K-mer analysis has been applied to many other genomic analyses, including but not limited to characterization of repeat content and heterozygosity (Williams et al. 2013), estimation of genome size (Guo et al. 2015), evaluation of metagenomic dissimilarity (Dubinkina et al. 2016), and identification of causal genetic variants conferring phenotypic traits (Nordstrom et al. 2013). Any size of k-mers can be used for k-mer analysis. Using smaller sized k-mers, sequencing data are condensed to less total k-mers, and a smaller number of k-mers are derived from single-copy regions, resulting in higher degree of information loss. Increasing size of k-mers increases both the total k-mer number and the number of single-copy k-mers, which is compromised by increased computation cost. Additionally, higher size of k-mers is more vulnerable to sequencing errors contained in sequencing reads. The impact of sequencing errors could be alleviated by error correction of sequences. The choice of k-mer length of 25 nt is an optimal size for human-sized genomes which was used in ALLPATHS-LG for analyzing k-mer abundance spectrum (Butler et al. 2008).

K-mer based methods are independent of read mapping that typically relies on a reference sequence, which allows the establishment of a fair comparison between genomes. For WGS data from either the same or different species, k-mer analysis can be directly applied to quantify the level of dissimilarity between individuals as long as WGS data are comparable. Low-coverage WGS data are sufficient to deliver reliable counts for k-mers derived from highly repetitive sequences. The critical issue is to develop a reliable normalization approach to account for non-genomic variation in data due to different sequence depths, varying levels of organelle DNAs, or contaminations from other species, particularly from microbes. In this study, we used total counts of a great number of single-copy k-mers that are conserved in the examined individuals to determine normalization factors. This normalization method is expected to well account for non-genomic variation. With high-coverage WGS data from multiple individuals, any types of genomic polymorphisms at either low or highly repetitive genomic regions would be unbiasedly represented by abundance of corresponding k-mers. In particular, copy number variation can be well captured by analyzing k-mer abundances. With that respect, one of potential applications of k-mer analysis is to perform genome-wide association with abundances of k-mers, which could retrieve some associated genetic elements that are unable to be detected using reference-based approaches. Collectively, the k-mer based approach alleviates ascertainment biases introduced by reference-based methods, and should provide the complement to many existing genome analyses.

HAKmer copy number variation QTL mapping

Using low-coverage WGS sequencing data of the IBM DH lines, a cnvQTL genetic mapping strategy was developed to map the genomic regions determining variation of k-mer abundance among DHs. As a result, the vast majority of differential abundance HAKmers between B73 and Mo17 were confidently mapped. The success of mapping differential abundance HAKmers from a variety of sources, including 45S rDNA, CentC, knobs, and telomeres, proved the effectiveness of the cnvQTL mapping. The fact that k-mers from rDNAs, CentC, telomeres, and knobs were all mapped to the expected regions where they are physically located suggests that no recognizable *trans* elements control the segregation of these repetitive sequence copies. The lack of *trans* elements makes sense because these repetitive sequences, although they evolve rapidly, are steadily maintained in each of two maize inbred lines.

We obtained a high-resolution map identifying coordinates contributing to differences in abundance of k-mers for many types of repetitive sequences in B73 and Mo17. These mapped genomic regions accurately mark the locations of clusters of repetitive sequences and corroborate many previous findings, as well as provide additional insight into the differentiation between B73 and Mo17. For example, both B73- and Mo17-gain 45S rDNA k-mers were mapped to around 13.5 Mb (B73Ref3) on chromosome 6 where a large PAV on the order of a megabase between B73 (presence) and Mo17 (absence) has been found. The result that Mo17-gain 45S rDNA k-mers were mapped at this PAV region, presumably located on the NOR, indicates that Mo17 has distinct 45S rDNA sequences to replace the missing version of 45S rDNA at the Mo17 NOR. Moreover, some Mo17-gain 45S rDNA k-mers were mapped to 210.5 Mb at the long arm of chromosome 1 that was not discovered previously, suggesting Mo17 contains a 45S rDNA cluster with significantly elevated copy number of 45S DNA at that region. Using a set of k-mers

from the 45S rDNA specific sequences that are conserved among maize, rice, and barley, we estimated that the copy number of 45S rDNA in B73 and Mo17 is 3,658 and 5,063, respectively. The Mo17-gain of 45S rDNA at chromosome 1, at least partially, explains higher copy number of 45S rDNA in Mo17 relative to B73.

Our cnvQTL mapping data genetically confirm the differential abundance of knob contents between B73 and Mo17. In addition to the long arms on chromosomes 5 and 7 that were reported previously (He et al. 2014, Kato et al. 2004), a distal region (293.5 Mb) at the long arm of chromosome 1 shows higher abundance of knob repeats in B73. The reduction of knob repeats on chromosomes 1, 5, 7 primarily accounts for the 55% loss of knob repeats in Mo17. What is more, detailed differentiation in CentC and telomere sequences were revealed. The increase of CentC repeats in multiple chromosomes in Mo17 indicates a possible common driving force involved in these parallel directional changes in copy number in a genome.

45S rRNA expression in hybrids

Nucleolar dominance is a phenomenon specifically observed in hybrids in which the NOR of one parent are dominant over the other of which rRNA is silenced. rRNA silencing involves epigenetic modifications of chromatin (McStay 2006). To examine nucleolar dominance in hybrids, allelic expression of rRNA needs to be precisely quantified. We have showed that rRNA is well represented even in mRNA sequencing data where rRNA was selected against. The divergence of 45S rDNA sequences between B73 and Mo17 provides the possibility for

examining rRNA allelic expression in their hybrids. However, most polymorphisms of 45S rDNA are located at IGS and ITS whose expression is hardly detected using the examined mRNA sequencing data. Fortunately, we identified the k-mers harboring a SNV polymorphic site on the 26S rRNA gene. The paired polymorphic k-mers are respectively, and nearly exclusively, expressed in one of B73 and Mo17 inbred lines, which sets an ideal marker to measure the expression of two types of 45S rDNA in the hybrid of B73 and Mo17. The k-mer abundance analysis indicates that Mo17 contains higher copy number of 45S rDNA than B73. Using transcriptomic sequencing data of primary roots, we observed the expression levels of rRNAs derived from two parents were equalized in both reciprocal hybrids, suggesting no nucleolar dominance occurs in the primary roots of the hybrid of B73 and Mo17 and also implying that an unknown mechanism exists to regulate dosage compensation.

Using transcriptomic sequencing data of early whole kernels and endosperms, we observed that the maternal rRNA expression is almost completely dominant in the whole kernels at 0 DAP, followed by the gradual increase of paternal rRNA expression from 0 to 5 DAP. It is not clear that inequality of maternal and paternal rRNA expression in early whole kernels is merely due to the distinct proportions of maternal and paternal genomes or its combination with the transcriptional suppression of paternal rRNA. Further examination through precise quantification of both rRNA and rDNA could address this question. Maize endosperm is a triploid, containing 2n of the material genome and 1n of the paternal genome. In early endosperms at 7, 10, and 15 DAP, the maternal rRNA expression is around twice as high as the paternal rRNA expression, indicating both maternal and paternal rRNA function, and, therefore, no nucleolar dominance was observed at the tissues examined.

Implications for maize evolution

Maize was domesticated from a wild species teosinte (*Zea mays* ssp. *parviglumis*) approximately 9-10 thousands years ago (Piperno et al. 2009, van Heerwaarden et al. 2011). Genetic evidence supports a single domestication and the post-domestication introgression from other wild relatives including *Zea mays* ssp. *Mexicana* (Hufford et al. 2012, Matsuoka et al. 2002, van Heerwaarden et al. 2011). The two distinct versions of 45S rDNA repeats traced by B73- and Mo17-specific k-mers at the NOR can be identified in different teosinte lines, indicating maize NORs originated from multiple ancient sources. The lower abundance of B73- and Mo17-specific k-mers in all examined teosinte but higher abundance in most landraces and improved maize lines suggests an expansion of certain types of rDNA repeats after domestication. Our observation that identical genotype-specific sequences are spread throughout the entire genome also raises interesting questions about the evolutionary past and origin of these sequences in relation to the NOR. Given evidence for a single domestication event and our observation of the local expansion of genotype-specific 45S rDNA sequences during maize domestication and improvement, flow of rDNA repeats away from the NOR following domestication is a more likely hypothesis. While the translocating mechanism can be either RNA- or DNA-mediated, our observation that spread regions consist of tandem arrays of intact 45S rDNA repeats suggests that this translocating mechanism is likely DNA-mediated. Spreading phenomena were observed for knob repeats and centromere retrotransposon members in both our results (Fig. 2.2) and previous studies (Ananiev et al. 1999b, Ghaffari et al. 2013, Lamb et al. 2007, Wolfgruber et al.

2009). Spreading sequences might serve as seeds that could eventually form new clusters of repetitive sequences, such as nascent knobs or NORs.

To further characterize flux of repetitive DNA during evolution, we identified k-mer sequences showing strikingly differential abundance among three groups, teosinte, landraces, and improved lines. Nearly all of these differential abundance k-mers displayed distinct patterns of either increase or decrease in abundance from teosinte to maize. rDNA k-mers make up the largest class of differential abundance k-mers. While 83 45S rDNA k-mers showed increasing abundance during this evolutionary time-frame, 266 showed marked loss. Additional analysis of relative copy numbers of 45S rDNA of HapMap2 lines also showed shrinkages and expansions of 45S rDNA repeats from teosinte to maize lines. In contrast, all differential abundance CentC k-mers were observed to decrease in abundance, strongly suggesting the shrinkage of CentC during domestication. The reverse trend is seen for CRM k-mers, which are dramatically elevated during domestication. This result replicates similar findings discussed in two recent centromere publications (Bilinski et al. 2015, Schneider et al. 2016). In addition, other retrotransposon members vary greatly among historical groups. Increasing evidence shows that transposons play important roles in adaptation and evolution (Lisch et al. 2013, Studer et al. 2011). The dramatic change in copy number of transposon elements during maize domestication could affect transcription and gene function by disrupting genes via direct integration in functional genic regions, providing new regulatory elements, and spreading epigenetic status to nearby genes (Lisch 2009, Makarevitch et al. 2015). In summary, our k-mer analyses offers a single-base resolution to trace dynamics of *Zea mays* genomes which has been appreciated

through cytogenetics, molecular, genetics, and genomics studies, providing valuable insights into the contents and organization of highly repetitive sequences in maize.

Materials and Methods:

Plant materials and extraction of nucleus genomic DNA

Two sources of B73 (PI 550473) were used, including seeds from Patrick Schnable laboratory and North Central Regional Plant Introduction Station (NCRPIS). All Mo17 (PI 558532) seeds were originated from NCRPIS. Seeds of two genotypes were geminated and grown in growth chamber at 28 ℃, with a photoperiod of 14:10 h (light:dark). 15~20 grams of fresh leaves of seedlings at 2–3 leaf-stage were harvested, frozen in liquid nitrogen, and homogenized with liquid nitrogen to fine powder. The nuclei were isolated using a protocol modified from Zhang's approach (Zhang et al 2012), followed by using the Qiagen DNeasy Plant Mini Kit protocol to extract nucleus DNA.

WGS sequencing of B73 and Mo17

Genomic DNAs from nuclei were used for PCR-free library preparation. Two replicates of each of B73 and Mo17 were whole genome shotgun sequenced with one sample per lane in HiSeq2000. 2 × 125 bp paired-end data were generated. Sequencing was conducted at BGI Genomics Co., Ltd., Shenzhen, China.

Error correction and genome size estimation

B73 and Mo17 whole genome sequences were trimmed to remove adaptor contaminations and low quality sequences with Trimmomatic version 3.2 (Bolger et al. 2014). The clean data were subjected to error correction using the error correction module (ErrorCorrectReads.pl) in ALLPATHS-LG49with the parameters of "PHRED_ENCODING = 33 PLOIDY = 1". Genome size was estimated during the procedure of error correction.

K-mer counting

Corrected sequences were subjected to k-mer counting using the count function in JELLYFISH (Marcais et al. 2011) with the k-mer size of 25 nt.

Estimation of genomic copy number of 45S rDNAs in B73 and Mo17

Quantification of rDNA copy number was performed using k-mers generated from the conserved regions of 45S rDNA among maize, rice, and barley. K-mers were aligned to the Zea mays repeat database (TIGR_Zea_Repeats.v3.0) to exclude any k-mers aligning to non-45S rDNA repeats, and to the B73Ref3 mitochondrial and plastid sequences to exclude k-mers that are not exclusively nuclear. Abundance of the 45S rDNA k-mers was evaluated for each B73 and Mo17. Abundances of these conserved k-mers in B73 and Mo17 were normalized by division by the respective estimated abundances for single-copy k-mers in order to estimate the number of 45S rDNA repeats in each genome. The median value of all conserved k-mers was the estimation of the rDNA copy number.

Identification of HAKmers with significant differential abundance between B73 and Mo17

High-abundance k-mers (HAKmers) in B73 or Mo17 were extracted, each of which is required to have at least 20,000 of total of B73 and Mo17 counts. A $\chi 2$ statistical test for each HAKmer was performed to test the null hypothesis of no relationship between k-mer counts and the genotypes (B73 and Mo17). P-values of all HAKmers were corrected to account for multiple tests (Benjamini et al. 1995). The differential abundance of HAKmers were declared if adjusted p-values are smaller than 5% and fold change in k-mer abundance between B73 and Mo17 is not less than 2.

Functional annotation of HAKmers

The Zea mays repeat database (TIGR_Zea_Repeats.v3.0) was downloaded from the plant repeat database that is currently maintained by Michigan State University (plantrepeats.plantbiology.msu.edu). BLASTN was performed with the word size of 12 to identify hits in the Zea mays repeat database for each HAKmer. The top hit with the e-value cutoff of 0.1 was referred to as the functional annotation.

K-mer mapping to the B73 reference genome

K-mer mapping to the B73 reference genome (B73Ref3) was conducted by using Bowtie (version 1.1.2) to identify all possible perfect hits.

Genetic mapping of HAKmers via cnvQTL

Resequencing data of 280 DH lines of the IBM Syn10 population used to build an ultra-high density genetic map (Liu et al. 2015) were trimmed with Trimmomatic version 3.2 (Bolger et al. 2014). Remaining clean reads were subjected to k-mer counting with JELLYFISH (Marcais et al.

2011). The k-mer size is 25 nt. The abundance of each HAKmer with differential abundance in B73 and Mo17 was determined in each DH line. The total counts (C) of a million of randomly selected B73 and Mo17 common single-copy k-mers in each DH line were determined. The normalization factor for the ith line was calculated by using the formula $NC_i/\sum_{i=1}^{n} C_i$, where N is the total number of IBM DH lines. The designation single-copy was determined by k-mer abundance from whole genome sequencing data for both B73 and Mo17 and confirmed by alignments to the B73ref3. Normalized abundance of a HAKmer was treated as a quantitative trait. For each HAKmer, a genetic mapping resembling a QTL detection implemented in an R package rqtl (Broman et al. 2003) was performed to identify genomic locations contributing the HAKmer abundance.

Identification of B73- and Mo17-specific HAKmers

To identify extremely unbalanced HAKmers that show extremely low abundance in one of two datasets from B73 and Mo17, the maximum number of 10 was used as the cutoff. Note that the minimum total abundance from B73 and Mo17 is 20,000 for HAKmers. If a HAKmer exhibits extremely low abundance (<= 10) in one genotype, it must be high (>19,990) in the other genotype. An extremely unbalanced HAKmer of which only one genotype, B73 or Mo17, showing high abundance is called B73 or Mo17 specific HAKmers.

HapMap2 data and k-mer analysis

Resequencing data of Zea mays HapMap2 lines (Chia et al. 2012, Hufford et al 2012) were downloaded and trimmed with Trimmomatic version 3.2 (Bolger et al. 2014), followed by 25 nt k-mer analysis using JELLYFISH (Marcais et al. 2011). To make comparable k-mer abundances

in different lines, a novel normalization method was developed. In this method, a set of "conserved single-copy k-mers" across HapMap2 lines was identified, which are single-copy in almost all lines. For each of these k-mers, k-mer abundances of HapMap2 lines should show a high correlation with their sequencing library sizes. In detail, the k-mer abundance of each HapMap2 line was determined for each of one million of B73 and Mo17 common single-copy k-mers that we identified. For each k-mer, a correlation of k-mer abundances of HapMap2 lines with their library sizes was calculated. The top 5% k-mers with the highest correlations (N = 49,955) were selected, which are deemed as "conserved single-copy k-mers". The total counts (C) of conserved single-copy k-mers per Hapmap2 line were determined. The normalization factor for the ith line was calculated by using the formula $NC_i/\sum_{i=1}^{n} C_i$, where N is the total number of HapMap2 lines.

PCA of k-mer abundance of B73 and Mo17 specific k-mers in HapMap2

PCA was performed using normalized k-mer abundances of B73 and Mo17 specific k-mers in HapMap2. The R function of prcomp was used for the PCA.

Allelic expression of rRNA in hybrids of B73 and Mo17

The RNA-Seq data of young maize primary roots in the B73, Mo17 and their reciprocal hybrids (Paschold et al. 2014) and the time-course sequencing RNA-Seq data of whole kernels at 0, 3, and 5 DAP and endosperms at 7, 10, and 15 DAP from reciprocal hybrids of B73 and Mo17 (Xin et al. 2013) were downloaded. Sequencing reads were subjected to quality, adaptor trimming, and k-mer counting with the size of 25 nt. The expression abundance of 45S rDNA k-mers harboring a polymorphic site was used to assess allelic expression.

Identification of highly variable k-mers in Zea mays

Abundances of k-mers were determined in each HapMap2 line. K-mer abundances were normalized using normalization factors calculated from a "conserved single-copy k-mers". Highly variable k-mers were extracted using the hard-filtering criteria that require >1,000 counts per k-mer per line in at least five HapMap2 lines but <10 counts in at least five other lines.

Identification of highly variable k-mers with significant differential abundance among evolutionary groups

Normalized counts of each k-mer for all HapMap2 lines were subjected to an ANOVA test. The genotype variable has three levels: teosinte, landrace, and improved. The null hypothesis is that k-mer abundances are independent of the genotype evolutionary groups. Then the Bonferroni approach was applied for multiple test correction at the 5% type I error.

MCLUST to classify highly variable k-mers showing significantly differential abundance among evolutionary groups

K-mers exhibiting significant differential abundance among three genotype groups were subjected to a clustering analysis using MCLUST (Fraley et al. 1999). For each k-mer, each count was scaled by being divided by the maximum count value of this k-mer. Scaled counts of k-mers were then used for the clustering using the parameters of "G = 1:12, modelNames = 'EEE'".

Estimation of relative genomic copy number of rDNAs in HapMap2 lines

The 45S rDNA k-mers used to estimate 45S rDNA copy number in B73 and Mo17 were used to estimate relative copy number level of each HapMap2 line. In each line, the median abundance value of k-mers represents the 45S rDNA copy number. The same method was used to determine 5S rDNA copy number level. The 5S rDNA k-mers were derived from the 5S rDNA sequence that is conserved among maize, rice, and wheat and were not aligned to B73 organelle genomes and other repetitive sequences.

Data access

B73 and Mo17 Illumina sequencing data have been deposited at Sequence Read Archive (SRA accession number: SRP082260).

Chapter 3 - Conclusions and Perspectives

The turgid and complex nature of the *Zea mays* genome requires equally sophisticated means of reducing the complexity to a level at which information can be accurately resolved, coupled with analysis that allows meaning to be extracted and condensed to its simplest and most concise form. The present work represents the application of a novel technique, k-mer analysis, employed towards resolving the complexity of the maize genome and shedding light on previously opaque genomic elements.

Large Genomes are complex not only in that they contain larger numbers of genes, gene families, higher order regulatory components and actual physical structure, but also because they are composed primarily of large numbers of repetitive DNA and guest mobile genetic elements. It is not so much the complexity of these elements that makes them intractable, but the complexity of the computational challenges involved in resolving these elements from data generated by short read technologies. Very large complex elements can often have profound impacts on the biology of an organism, yet methods of discovery are lacking with current technologies. These elements often fail to resolve in genome assemblies, resulting in fragmented assemblies, stitched together by gaps of N-base nucleotides representing the dark matter of the genome. Such ghost elements are not only themselves concealed, but their presence fragments the assembly, confusing the order and orientation by which contigs are placed within scaffolds. In this thesis, I describe the application of a novel method of analysis designed to confront these issues directly. The k-mer approach which our team has developed allows quantification of 'dark matter' repetitive elements in the genome in a way that is highly accurate and quantitative with base-pair resolution.

We were able to reach several biologically significant insights through our methods, especially in terms of broad trends within the maize genome and as well as in the context of maize domestication. In comparing the two genomes, the most extreme differences were found at the level of highly abundant k-mers corresponding to highly repetitive DNA. Within these classes, the most striking differences were within the rDNA repeats, from which k-mers determined to be unique to one line or the other were derived. While the trend observed is quite notable, the significance of the trend is not necessarily immediately apparent. Further future studies might be necessary to help to understand the mechanisms underlying this trend, as well as to help understand why the trend is seen so dramatically for rDNA repeats, while it is not as evident for the remaining repeat types. One possible explanation for this observation might be that the repeats are subject to high levels of concerted evolution. Concerted evolution is a phenomenon which results in greater homology between sequences not directly related through descent within a species than that which is found between sequences which share a recent common ancestor. That is, in a genic context, two genes in species A sharing homology, but more distantly related than the same genes between species A and B, are more homologous to each other than they are to their proper homologs between the two species. This phenomenon may in some cases be a result of gene conversion, causing the sequence at one locus to be replaced by the sequence at a non-allelic locus to be replaced by it. This process might also occur in the course of unequal crossing over, such that one variant expands in number within the repeat array while the other shrinks until the original sequence has been replaced by the new sequence. Given the tandem arrangement of the rDNA repeats in the maize genome, this mechanism is more likely than that of gene conversion to be the active mechanism driving homogenization of the repeat sequence,

and resulting in dramatic differences between the rDNA k-mers for each respective inbred line. Concerted evolution has been observed in the past acting on the IGS sequence of 45S rDNA repeats of *Xenopus* rDNAs, for instance (Nei et al. 2005).

The question of rDNA spreading or translocation from the NOR to distinct loci within the maize genome also merits further investigation, and also illustrates that despite advances allowed in quantification of repeat copy number afforded by our study, further analysis can still be hindered by poor representation in the reference sequence. Based on alignment of our B73 unique k-mer sequences to the B73 reference 3 genome assembly, it seems that there are multiple loci containing the rDNA repeats above and beyond the classic locus at the NOR. This finding relies heavily on the assembly, and requires further validation. We also observed other classes of repeats dispersed throughout the reference genome, as have other authors (Ghaffari et al. 2013). Whether this phenomenon is due to incorrect assembly, or whether it is due to some inherent property of repetitive sequences within the context of the maize genome, remains to be seen. Previous non-sequencing based research (Phillips et al. 1973) investigated total rDNA copy in genomic DNA in normal and monosomic maize lines. Their results confirmed the location of the primary rDNA encoding locus belonging to chromosome 6, and they also found no reduction in total rDNA content in lines monosomic for chromosome 8. However, they still saw an overall reduction in total rDNA content in lines monosomic for chromosome 10, which supports our hypothesis that rDNA can be found on maize chromosomes other than 6. Microscopic evidence also supports the occurrence of knob repeats at diverse loci within the maize genome, and it is possible that there is some inherent property of extended long arrays of repetitive DNA that makes it susceptible to translocation. Assuming equal probability of double stranded breaks

genome wide, sequences which make up a greater proportion of the genome are more likely to experience double stranded breaks than other regions of the genome. Therefore, the more extended the length of a repetitive sequence, the more likely that two double-stranded breaks might occur within said repetitive sequence and potentially result in translocation of the repetitive element outside of its normal genomic range. Translocated regions could then go on to serve as seeds for repetitive regions which might expand or contract stochastically in the course of evolutionary time. Maize knobs are known to be highly variable in position within the maize genome, so there is certainly precedent for active positional dynamics of repetitive DNA in the genome.

Along with allowing comparative and quantitative analysis of maize genomic dark matter, another important advance allowed by our methods was mapping of variation between the compared genomes. Our use of maize inbreds B73 and Mo17 facilitated this approach, as there is sequencing data available for a crossing population using these as the parents. QTL mapping traditionally utilizes SNP markers to associate differences quantitative levels of traits with a specific genotype. Our innovation in this case was using k-mer abundance as the quantitative trait. In this case, we choose the line-unique k-mers for the analysis. While this approach does not necessarily allow for mapping of repetitive sequence shared by both lines, it is nonetheless useful for mapping loci responsible for large differences in variation. Accurate placement of large tracts of repetitive elements in genome assemblies remains challenging. Our cnvQTL approach might be useful to help resolve this problem in the future. For instance, the maize reference is limited in both placement and representation of the knob regions (Ghaffari et al. 2013). Given that these sequences consist of repeated sequence several million base pairs in

length, this is not surprising. The same has also been true of other repetitive elements discussed in our study, for instance the centromere repeats, which some re-sequencing efforts have attempted to capture more accurately through the use of long-read sequencing technology.

In the past, repetitive and dark matter DNA has been difficult to study using either molecular biology techniques or sequencing based approaches, and as a result, the biological relevance of these sequences has often eluded researchers. However, as these genomic elements become more accessible to researchers, both as a result of more advanced sequencing methods as well as analytic approaches, it will become easier to ask questions and test hypotheses regarding these. Our k-mer approach is one such method that researchers might employ. Often, to understand the significance of a biological component, researchers will consider different contexts in which the component exists, and by better understanding these components, the researchers can come to a more complete understanding of the importance and function of these components. For instance, utilizing the hapmap2 dataset allowed to investigate k-mers which were highly variable during domestication, allowing us to examine repeat content in an evolutionary context. An alternative approach might be to focus on repetitive content that is highly conserved within an evolutionary timeframe. It is difficult to demonstrate selective forces driving trends in repetitive DNA. For instance, while we observed changes that occurred during evolution, it is difficult to determine whether these changes were random or driven by selection. It is unclear what sorts of selective advantages repetitive elements might have. Conservation of repetitive sequence might be a better indicator of selection than overall trends, however. This conservation might be at the sequence level, at the level of copy number, or might be positional. For instance, one could speculate that some selective force is acting to maintain the NOR at its location on chromosome 6, in contrast

to the knob repeat arrays which seem to be highly mobile within the maize genome. Finally, especially of interest might be the way that these regions interact with more conventionally understood elements. Repetitive elements might be interesting in the respect that they might influence recombination, or contribute to rapid evolution in gene islands found within them, for instance. Or, speaking more hypothetically, they might serve unknown structural roles in a similar manner to that of centromeres, forming functional protein/DNA complexes with possible roles within the nucleus such as regulating chromatin structure or formation of nuclear micro-domains. Epigenetic dynamics of these elements might also demonstrate themselves to be fruitful to investigators, as these regions are subject to varying types of epigenetic regulation. Centromeres and rDNA, for example, are known to be subject to dynamic epigenetic regulation (Layat et al. 2012, Zhong et al. 2002).

In conclusion, we develop the application of a novel method for analysis of sequencing data that lends itself well to comparative analysis of repetitive DNA elements that are not otherwise accessible to researchers. We found marked differences in highly repetitive DNA, demonstrating the capacity for the technique to illuminate genomic dark matter. We demonstrated the applicability of this technique to mapping repetitive regions of the genome, and we generate an overview of the genetic changes that occur during domestication. In addition, we perform analysis of RNA sequencing datasets to explore the question of nucleolar dominance within Mo17/B73 hybrids. K-mer analysis yielded intriguing insights and the analysis can be easily extended to other systems or experimental questions. Additionally, the wealth of existing publicly available sequencing datasets can be potentially re-analyzed using these methods, yielding new insights without requiring the generation of new data, and in that sense our methods

are especially thrifty. This is even true in terms of computational resources, as k-mer generation and counting using Jellyfish is computationally very efficient. Even within our own datasets, further data mining and extensions of the analysis can be made, for instance of k-mers for which annotation was not feasible.

References

Altschul S, Gish W, Miller W, Myers E, Lipman D. Basic local alignment search tool. Journal of Molecular Biology. 1990;215(3):403–410. doi:10.1016/S0022-2836(05)80360-2.

Ananiev EV, Phillips RL, Rines HW. A knob-associated tandem repeat in maize capable of forming fold-back DNA segments: are chromosome knobs megatransposons? PNAS. 1998;95(18):10785-90.

Ananiev EV, Phillips RL, Rines HW. Complex structure of knob DNA on maize chromosome 9. Retrotransposon invasion into heterochromatin. Genetics. 1998;149(4):2025-37.

Barbazuk WB, Emrich SJ, Chen HD, Li L, Schnable PS. SNP discovery via 454 transcriptome sequencing. The Plant Journal. 2007;51(5):910-8. doi: 10.1111/j.1365-313X.2007.03193.x.

Bauer E, Falque M, Walter H, Bauland C, Camisan C, Campo L, et al. Intraspecific variation of recombination rate in maize. Genome Biology. 2013;14(9):R103. doi: 10.1186/gb-2013-14-9-r103.

Belo A, Beatty MK, Hondred D, Fengler KA, Li B, Rafalski A. Allelic genome structural variations in maize detected by array comparative genome hybridization. Theoretical and Applied Genetics. 2010;120(2):355-67. doi: 10.1007/s00122-009-1128-9.

Benjamini Y, Hochberg Y. Controlling the False Discovery Rate - a Practical and Powerful Approach to Multiple Testing. Journal of the royal statistical society. Series B (Methodological). 1995. 57(1):289-300.

Beroukhim Rl, Zhang X, Meyerson M. Copy number alterations unmasked as enhancer hijackers. Nature Genetics. 2017;49: 5-6. doi:10.1038/ng.3754.

Bilinski P, Distor K, Gutierrez-Lopez J, Mendoza GM, Shi J, Dawe RK, et al. Diversity and evolution of centromere repeats in the maize genome. Chromosoma. 2015;124(1):57-65. doi: 10.1007/s00412-014-0483-8.

Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics. 2014;30(15):2114-20. doi: 10.1093/bioinformatics/btu170.

Broman KW, Wu H, Sen S, Churchill GA. R/qtl: QTL mapping in experimental crosses. Bioinformatics. 2003;19(7):889-90.

Buescher PJ, Phillips RL, Brambl R. Ribosomal RNA contents of maize genotypes with different ribosomal RNA gene numbers. Biochemical Genetics. 1984;22(9-10):923-30.

Burr B, Burr FA, Matz EC, Romero-Severson J. Pinning down loose ends: mapping telomeres and factors affecting their length. The Plant Cell. 1992;4(8):953-60. doi: 10.1105/tpc.4.8.953.

Butler J, MacCallum I, Kleber M, Shlyakhter IA, Belmonte MK, Lander ES, et al. ALLPATHS: de novo assembly of whole-genome shotgun microreads. Genome Research. 2008;18(5):810-20. doi: 10.1101/gr.7337908.

Chia JM, Song C, Bradbury PJ, Costich D, de Leon N, Doebley J, et al. Maize HapMap2 identifies extant variation from a genome in flux. Nature Genetics. 2012;44(7):803-7. doi: 10.1038/ng.2313.

Coe, EHJ. The origins of maize genetics. Nature Reviews Genetics. 2001; 2:898–905. doi: 10.1038/35098524.

Compeau PEC, Pevzner PA, Tesler G. How to apply de Bruijn graphs to genome assembly. Nature Biotechnology. 2011;29(11):987-91. doi: 10.1038/nbt.2023.

Cook DE, Lee TG, Guo X, Melito S, Wang K, Bayless AM, et al. Copy number variation of multiple genes at Rhg1 mediates nematode resistance in soybean. Science. 2012;338:1206–1209. doi: 10.1126/science.1228746.

da Fonseca RR, Smith BD, Wales N, Cappellini E, Skoglund P, Fumagalli M, et al. The origin and evolution of maize in the Southwestern United States. Nature Plants. 2015;1:14003. doi: 10.1038/nplants.2014.3.

Dawe RK. RNA Interference, Transposons, and the Centromere. The Plant Cell. 2003;15(2):297-301. doi:10.1105/tpc.150230.

Delhaize E, Ryan PR. Aluminum Toxicity and Tolerance in Plants. Plant Physiology. 1995; 107(2):315-321.

Deng Y, Zhai K, Xie Z, Yang D, Zhu X, Liu J, et al. Epigenetic regulation of antagonistic receptors confers rice blast resistance with yield balance. Science. 2017;355(6328): 962-965. doi: 10.1126/science.aai8898.

Dietrich CR, Cui F, Packila ML, Li J, Ashlock DA, Nikolau BJ, et al. Maize Mu transposons are targeted to the 5' untranslated region of the gl8 gene and sequences flanking Mu target-site duplications exhibit nonrandom nucleotide composition throughout the genome. Genetics. 2002;160:697-716.

Doebley J. The genetics of maize evolution. Annual Reviews in Genetics. 2004;38:37–59. doi: 10.1146/annurev.genet.38.072902.092425.

Dubinkina VB, Ischenko DS, Ulyantsev VI, Tyakht AV, Alexeev DG. Assessment of k-mer spectrum applicability for metagenomic dissimilarity analysis. BMC Bioinformatics. 2016;17:38. doi: 10.1186/s12859-015-0875-7.

Feuk L, Carson AR, Scherer SW. Structural variation in the human genome. Nature Reviews Genetics. 2006;7: 85–97. doi:10.1038/nrg1767.

Fraley C, Raftery AE. MCLUST: Software for model-based cluster analysis. J Classif. 1999;16(2):297-306. doi: DOI 10.1007/s003579900058.

Fu Y, Wen TJ, Ronin YI, Chen HD, Guo L, Mester DI, et al. Genetic dissection of intermated recombinant inbred lines using a new genetic map of maize. Genetics. 2006;174(3):1671-83. doi: 10.1534/genetics.106.060376.

Funke T, Huijong H, Healy-Fried ML, Fischer M, Schönbruun E. Molecular basis for the herbicide resistance of Roundup Ready crops. PNAS. 2006;103:13010-13015. doi: 10.1073/pnas.0603638103.

Gaines TA, Zhang W, Wang D, Bukun B, Chrisholm ST, Shaner DL, et al. Gene amplification confers glyphosate resistance in amaranthus palmeri. PNAS. 2010;107:1029–1034.

Ghaffari R, Cannon EK, Kanizay LB, Lawrence CJ, Dawe RK. Maize chromosomal knobs are located in gene-dense areas and suppress local recombination. Chromosoma. 2013;122(1-2):67-75. doi: 10.1007/s00412-012-0391-8.

Gibbons JG, Branco AT, Godinho SA, Yu S, Lemos B. Concerted copy number variation balances ribosomal DNA dosage in human and mouse genomes. PNAS. 2015;112(8):2485-90. doi: 10.1073/pnas.1416878112.

Guo LT, Wang SL, Wu QJ, Zhou XG, Xie W, Zhang YJ. Flow cytometry and K-mer analysis estimates of the genome sizes of Bemisia tabaci B and Q (Hemiptera: Aleyrodidae). Frontiers in Physiology. 2015;6:144. doi: 10.3389/fphys.2015.00144.

He S, Yan S, Wang P, Zhu W, Wang X, Shen Y, et al. Comparative analysis of genome-wide chromosomal histone modification patterns in maize cultivars and their wild relatives. PLoS One. 2014;9(5):e97364. doi: 10.1371/journal.pone.0097364.

Heid CA, Stevens J, Livak KJ, Williams PM. Real time quantitative PCR. Genome Research. 1996;6:986–994.

Hiatt EN, Kentner EK, Dawe RK. Independently regulated neocentromere activity of two classes of tandem repeat arrays. The Plant Cell. 2002;14:407–420. doi: 10.1105/tpc.010373.

Hufford MB, Xu X, van Heerwaarden J, Pyhajarvi T, Chia JM, Cartwright RA, et al. Comparative population genomics of maize domestication and improvement. Nature Genetics. 2012;44(7):808-11. doi: 10.1038/ng.2309.

Iafrate AJ, Feuk L, Rivera MN, Listewnik ML, Donahoe PK, Qi Y, et al. Detection of large-scale variation in the human genome. Nature Genetics. 2004;36:949–951. doi:10.1038/ng1416.

Imarisio S, Carmichael J, Korolchuk V, Chen CW, Saiki S, Rose C, et al. Huntington's disease: from pathology and genetics to potential therapies. Biochemical Journal. 2008;412(2):191-209. doi: 10.1042/BJ20071619.

Jiang N, Bao Z, Zhang X, Eddy SR, Wessler SR. Pack-MULE transposable elements mediate gene evolution in plants. Nature. 2004;431:569-573. doi:10.1038/nature02953.

Jiao Y, Zhao H, Ren L, Song W, Zeng B, Guo J, et al. Genome-wide genetic changes during modern breeding of maize. Nature Genetics. 2012;44(7):812-5. doi: 10.1038/ng.2312.

Jin ML, Liu HJ, He C, Fu JJ, Xiao YJ, Wang YB, et al. Maize pan-transcriptome provides novel insights into genome complexity and quantitative trait variation. Scientific Reports. 2016; 6:18936. doi: ARTN 1893610.1038/srep18936.

Jugulam M, Niehues K, Godar AS, Koo DH, Danilova T, Friebe B et al. Tandem Amplification of a Chromosomal Segment Harboring 5-Enolpyruvylshikimate-3-Phosphate Synthase Locus Confers Glyphosate Resistance in Kochia scoparia. Plant Physiology. 2014;166(3):1200-1207. doi:10.1104/pp.114.242826.

Kato A, Lamb JC, Birchler JA. Chromosome painting using repetitive DNA sequences as probes for somatic chromosome identification in maize. PNAS. 2004;101(37):13554-9. doi: 10.1073/pnas.0403659101.

Lamb JC, Birchler JA. Retroelement genome painting: cytological visualization of retroelement expansions in the genera Zea and Tripsacum. Genetics. 2006;173(2):1007-21. doi: 10.1534/genetics.105.053165.

Lamb JC, Meyer JM, Corcoran B, Kato A, Han F, Birchler JA. Distinct chromosomal distributions of highly repetitive sequences in maize. Chromosome Research. 2007;15(1):33-49. doi: 10.1007/s10577-006-1102-1.

Layat E, Saez-Vasquez J, Tourmente S. Regulation of Pol I-transcribed 45S rDNA and Pol III-transcribed 5S rDNA in Arabidopsis. Plant Cell Physiology. 2012;53(2):267-76. doi: 10.1093/pcp/pcr177.

Li J, Yang F, Zhu J, He S, Li L. Characterization of a tandemly repeated subtelomeric sequence with inverted telomere repeats in maize. Genome. 2009;52(3):286-93. doi: 10.1139/G09-005.

Li L, Arumuganathan K. Physical mapping of 45S and 5S rDNA on maize metaphase and sorted chromosomes by FISH. Hereditas. 2001;134(2):141-5.

Lisch D. Epigenetic regulation of transposable elements in plants. Annual Review of Plant Biology. 2009;60:43-66. doi: 10.1146/annurev.arplant.59.032607.092744.

Lisch D. How important are transposons for plant evolution? Nature Reviews Genetics. 2013;14(1):49-61. doi: 10.1038/nrg3374.

Liu H, Niu Y, Gonzalez-Portilla PJ, Zhou H, Wang L, Zuo T, et al. An ultra-high-density map as a community resource for discerning the genetic basis of quantitative traits in maize. BMC Genomics. 2015;16:1078. doi: 10.1186/s12864-015-2242-5.

Liu S, Chen HD, Makarevitch I, Shirmer R, Emrich SJ, Dietrich CR, et al. High-throughput genetic mapping of mutants via quantitative single nucleotide polymorphism typing. Genetics. 2010;184(1):19-26. doi: 10.1534/genetics.109.107557.

Liu S, Ying K, Yeh CT, Yang J, Swanson-Wagner R, Wu W, et al. Changes in genome content generated via segregation of non-allelic homologs. The Plant Journal. 2012;72(3):390-9. doi: 10.1111/j.1365-313X.2012.05087.x.

Lu F, Romay MC, Glaubitz JC, Bradbury PJ, Elshire RJ, Wang T, et al. High-resolution genetic mapping of maize pan-genome sequence anchors. Nature Communications. 2015;6:6914. doi: 10.1038/ncomms7914.

Lu F, Romay MC, Glaubitz JC, Bradbury PJ, Elshire RJ, Wang T, et al. High-resolution genetic mapping of maize pan-genome sequence anchors. Nature Communications. 2015;6:6914. doi:10.1038/ncomms7914.

Makarevitch I, Waters AJ, West PT, Stitzer M, Hirsch CN et al. Transposable elements contribute to activation of maize genes in response to abiotic stress. PLoS Genetics. 2015;11(10): e1005566. doi: 10.1371/journal.pgen.1005566.

Ma L, Chung WK. Quantitative Analysis of Copy Number Variants Based on Real-Time LightCycler PCR. Current Protocols in Human Genetics. 2014;80:Unit 7.21. doi:10.1002/0471142905.hg0721s80.

Makarevitch I, Waters AJ, West PT, Stitzer M, Hirsch CN, Ross-Ibarra J, et al. Transposable elements contribute to activation of maize genes in response to abiotic stress. PLoS Genetics. 2015;11(1):e1004915. doi: 10.1371/journal.pgen.1004915.

Mao H, Wang H, Liu S, Li Z, Yang X, Yan J, et al. A transposable element in a NAC gene is associated with drought tolerance in maize seedlings. Nature Communications. 2015;6:8326. doi:10.1038/ncomms9326.

Marcais G, Kingsford C. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. Bioinformatics. 2011;27(6):764-70. doi: 10.1093/bioinformatics/btr011.

Maron LG, Guimaraes CT, Kirst M, Albert PS, Birchler JA, Bradbury PJ, et al. Aluminum tolerance in maize is associated with higher MATE1 gene copy number. PNAS. 2013;110: 5241–5246. doi: 10.1073/pnas.1220766110.

Matsuoka Y, Vigouroux Y, Goodman MM, Sanchez GJ, Buckler E, Doebley J. A single domestication for maize shown by multilocus microsatellite genotyping. PNAS. 2002;99(9):6080-4. doi: 10.1073/pnas.052125199.

McKnight TD, Shippen DE. Plant telomere biology. The Plant Cell. 2004;16(4):794-803. doi: 10.1105/tpc.160470.

McStay B. Nucleolar dominance: a model for rRNA gene silencing. Genes and Development. 2006;20(10):1207-14. doi: 10.1101/gad.1436906.

Nei M, Rooney AP. Concerted and Birth-and-Death Evolution of Multigene Families. Annual review of genetics. 2005;39:121-152. doi:10.1146/annurev.genet.39.073003.112240.

Nordstrom KJ, Albani MC, James GV, Gutjahr C, Hartwig B, Turck F, et al. Mutation identification by direct comparison of whole-genome sequencing data from mutant and wild-type individuals using k-mers. Nature Biotechnology. 2013;31(4):325-30. doi: 10.1038/nbt.2515.

Paschold A, Larson NB, Marcon C, Schnable JC, Yeh CT, Lanz C, et al. Nonsyntenic genes drive highly dynamic complementation of gene expression in maize hybrids. The Plant Cell. 2014;26(10):3939-48. doi: 10.1105/tpc.114.130948.

Phillips RL, Weber DF, Kleese RA, Wang SS. The Nucleolus Organizer Region of Maize (ZEA MAYS L.): Tests for Ribosomal Gene Compensation or Magnification. Genetics. 1974;77(2):285-97.

Pinkel D, Donna A. Array comparative genomic hybridization and its applications in cancer. Nature Genetics. 2005;37:211-S17. doi:10.1038/ng1569.

Piperno DR, Ranere AJ, Holst I, Iriarte J, Dickau R. Starch grain and phytolith evidence for early ninth millennium B.P. maize from the Central Balsas River Valley, Mexico. PNAS. 2009;106(13):5019-24. doi: 10.1073/pnas.0812525106.

Pirooznia M, Goes FS, Zandi PP. Whole-genome CNV analysis: advances in computational approaches. Frontiers in Genetics. 2015;6:138. doi: 10.3389/fgene.2015.00138.

Pollack JR, Perou CM, Alizadeh AA, Eisen MB, Pergamenschikov A, Williams CF, et al. Genome-wide analysis of DNA copy-number changes using cDNA microarrays. Nature Genetics. 1999;23(1):41−46. doi:10.1038/12640.

Rivin CJ, Cullis CA, Walbot V. Evaluating quantitative variation in the genome of Zea mays. Genetics. 1986;113(4):1009-19.

Schnable JC, Springer NM, Freeling M. Differentiation of the maize subgenomes by genome dominance and both ancient and ongoing gene loss. PNAS. 2011; 108(10):4069-74. doi: 10.1073/pnas.1101368108.

Schnable PS, Ware D, Fulton RS, Stein JC, Wei F, et al. The B73 maize genome: complexity, diversity, and dynamics. Science. 2009;326(5956):1112-5. doi: 10.1126/science.1178534.

Schneider KL, Xie Z, Wolfgruber TK, Presting GG. Inbreeding drives maize centromere evolution. PNAS. 2016;113(8):E987-96. doi: 10.1073/pnas.1522008113.

Sebat J, Lakshmi B, Troge J, Alexander J, Young J, Lundin P, et al. Large-scale copy number polymorphism in the human genome. Science. 2004;305:525–528. doi: 10.1126/science.1098918.

Springer NM, Ying K, Fu Y, Ji T, Yeh CT, Jia Y, et al. Maize inbreds exhibit high levels of copy number variation (CNV) and presence/absence variation (PAV) in genome content. PLoS Genetics. 2009;5(11):e1000734. doi: 10.1371/journal.pgen.1000734.

Strable J, Scanlon MJ. Maize (Zea mays): A Model Organism for Basic and Applied Research in Plant Biology. Cold Spring Harbor Protocols. 2009; (10):pdb.emo132. doi:10.1101/pdb.emo132.

Studer A, Zhao Q, Ross-Ibarra J, Doebley J. Identification of a functional transposon insertion in the maize domestication gene tb1. Nature Genetics. 2011;43(11):1160-3. doi: 10.1038/ng.942.

Swanson-Wagner RA, Eichten SR, Kumari S, Tiffin P, Stein JC, Ware D, et al. Pervasive gene content variation and copy number variation in maize and its undomesticated progenitor. Genome Res. 2010;20(12):1689-99. doi: 10.1101/gr.109165.110.

van Heerwaarden J, Doebley J, Briggs WH, Glaubitz JC, Goodman MM, de Jesus Sanchez Gonzalez J, et al. Genetic signals of origin, spread, and introgression in a large sample of maize landraces. PNAS. 2011;108(3):1088-92. doi: 10.1073/pnas.1013011108.

Williams D, Trimble WL, Shilts M, Meyer F, Ochman H. Rapid quantification of sequence repeats to resolve the size, structure and contents of bacterial genomes. BMC Genomics. 2013;14:537. doi: 10.1186/1471-2164-14-537.

Wolfgruber TK, Sharma A, Schneider KL, Albert PS, Koo DH, Shi J, et al. Maize centromere structure and evolution: sequence analysis of centromeres 2 and 5 reveals dynamic Loci shaped primarily by retrotransposons. PLoS Genet. 2009;5(11):e1000743. doi: 10.1371/journal.pgen.1000743.

Xin M, Yang R, Li G, Chen H, Laurie J, Ma C, et al. Dynamic expression of imprinted genes associates with maternally controlled nutrient allocation during maize endosperm development. The Plant Cell. 2013;25(9):3212-27. doi: 10.1105/tpc.113.115592.

Yu W, Lamb JC, Han F, Birchler JA. Telomere-mediated chromosomal truncation in maize. PNAS. 2006;103(46):17331-6. doi: 10.1073/pnas.0605750103.

Zhang F, Khajavi M, Connolly AM, Towne CF, Batish SD, Lupski JR. The DNA replication FoSTeS/MMBIR mechanism can generate genomic, genic and exonic complex rearrangements in humans. Nature Genetics. 2009;41:849-853. doi:10.1038/ng.399.

Zhang MP, Zhang Y, Scheuring CF, Wu CC, Dong JJ, Zhang HB. Preparation of megabase-sized DNA from a variety of organisms using the nuclei method for advanced genomics research. Nature Protocols. 2012;7(3):467-78. doi: 10.1038/nprot.2011.455.

Zhang Z, Mao L, Chen H, Bu F, Li G, et al. Genome-Wide Mapping of Structural Variations Reveals a Copy Number Variant That Determines Reproductive Morphology in Cucumber. The Plant Cell. 2015;27(6):1595-604. doi: 10.1105/tpc.114.135848.

Zhong CX, Marshall JB, Topp C, Mroczek R, Kato A, et al. Centromeric retroelements and satellites interact with maize kinetochore protein CENH3. The Plant Cell. 2002;14(11):2825-2836. doi:10.1105/tpc.006106.

Żmieńko A, Samelak A, Kozłowski P, Figlerowicz M. Copy number polymorphism in plant genomes. Theoretical and Applied Genetics. 2014;127(1):1-18. doi:10.1007/s00122-013-2177-7.

Appendix A - Supplemental Data

Table A.1: Statistics of functional classes of HAKmers

Code	Class	Number of all non-redundant k-mers	Number of differential abundance k-mers	Differential/all (%)
TERTOOT	Unclassified Retrotransposons	470,660	27	0.006
TERT002	Ty3-gypsy	21,824	0	0
OTKN000	Knob	10,371	8,269	79.732
CMCMOOT	Unclassified Centromere Sequences	9,837	81	0.823
OTOT000	Unclassified	9,330	41	0.439
RGRR000	45S rDNA	9,173	3,867	42.156
TERT001	Ty1-copia	7,657	0	0
TETN002	CACTA, En/Spm	4,469	0	0
TETNOOT	Unclassified Transposons	3,112	0	0
CMCM002	Centromeric satellite repeats	1,838	621	33.787
RGRR005	5S rDNA	1,397	4	0.286
TRTM000	Telomere	656	117	17.835
TEMT059	mPIF	567	0	0
CMCM001	Centromere-specific Retrotransposons	450	0	0
TETN003	Mutator (MULE)	311	0	0
TETN001	Ac/Ds	227	0	0
TEMT055	Heart breaker	135	0	0
TEMTOOT	Unclassified MITEs	134	0	0
TERT003	LINE	127	0	0
TEMT056	Frequent Flyer	33	0	0
TETN004	Mariner (MLE)	27	0	0
TEMT002	Stowaway	14	0	0
TEMT057	Heart Healer	12	0	0
TEMT006	Castaway	7	0	0
TEMT001	Tourist	3	0	0

Table A.2: Sum of estimated copies of all k-mers in each functional class

Class	Sum of copies of all k-mers per class*		Mo17/B73
	B73	Mo17	
knob	1,979,523,103	896,095,038	0.453
45S rDNA	807,833,346	1,080,176,065	1.337
Ty3-gypsy	802,093,819	813,043,361	1.014
Unclassified	301,005,739	310,232,395	1.031
Unclassified Centromere Sequences	289,605,185	276,527,056	0.955
Ty1-copia	239,439,384	245,138,170	1.024
5S rDNA	96,397,625	120,000,151	1.245
Centromeric satellite repeats (CentC)	90,457,272	154,423,006	1.707
CACTA, En/Spm	67,240,372	66,214,078	0.985
Unclassified Transposons	65,768,661	65,767,164	1.0
mPIF	14,768,671	15,618,841	1.058
Telomere	13,554,843	10,501,889	0.775
Centromere-specific Retrotransposons	9,128,860	8,158,594	0.894
Mutator (MULE)	7,728,574	8,316,731	1.076
Ac/Ds	7,479,272	8,370,891	1.119
LINE	2,541,319	2,837,120	1.116
Heart breaker	2,070,377	2,162,051	1.044
Unclassified MITEs	1,912,841	1,904,293	0.996
Frequent Flyer	566,569	538,890	0.951
Mariner (MLE)	458,570	465,303	1.015
Stowaway	436,074	614,332	1.409
Heart Healer	411,751	515,739	1.253
Castaway	326,505	319,793	0.979
Tourist	64,804	61,979	0.956

* k-mer counts were corrected by 36 and 35 which represent the k-mer abundance of single-copy k-mers in B73 and Mo17, respectively.

Table A.3: Number of HAKmers showing various mapping peaks*

Number of peaks	0	1	2	3	4
No. B73-gain HAKmers (%)	35 (0.3)	8,499 (74.5)	2,840 (24.9)	28 (0.2)	11 (0.1)
No. Mo17-gain HAKmers (%)	78 (3.0)	2,198 (83.5)	290 (11.0)	66 (2.5)	1 (0.04)

* The maximum one peak was considered for each chromosome

Table A.4: Number of each functional class of B73-gain HAKmers showing various mapping peaks

Repeat category		Number of mapping peaks				
Code	Class	0	1	2	3	4
CMCM002	Centromeric satellite repeats	0	134	10	0	0
CMCMOOT	Unclassified centromere sequences	0	19	24	0	3
OTKN000	Knob	32	5660	2569	8	0
RGRR000	45S rDNA	0	2205	0	0	0
TERTOOT	Unclassified retrotransposons	0	27	0	0	0
TRTM000	Telomere	0	40	58	12	5

Table A.5: Number of each functional class of Mo17-gain HAKmers showing various mapping peaks

Repeat category		Number of mapping peaks				
Code	Class	0	1	2	3	4
CMCM002	Centromeric satellite repeats	1	267	165	43	1
CMCMOOT	Unclassified Centromere Sequences	0	0	28	7	0
OTOT000	Unclassified	0	40	1	0	0
RGRR000	45S rDNA	0	1645	17	0	0
RGRR005	5S rDNA	0	4	0	0	0
TRTM000	Telomere	0	0	2	0	0

Table A.6: K-mer abundance of three pairs of k-mers harboring a SNV at 26S rRNA

Pair	K-mers*	Abundance (% of total of a pair) in B73	Abundance (% of total of a pair) in Mo17	Type
1	GG**A**ATTCGGTCCTCCGGATTTTCAA	82,924 (72%)	40 (0%)	B73gain
1	GG**T**ATTCGGTCCTCCGGATTTTCAA	32,602 (28%)	152,178 (100%)	Mo17gain
2	CTTGAAAATCCGGAGGACCGAAT**T**C	82,900 (72%)	40 (0%)	B73gain
2	CTTGAAAATCCGGAGGACCGAAT**A**C	32,628 (28%)	152,071 (100%)	Mo17gain
3	**A**ATTCGGTCCTCCGGATTTTCAAGG	83,074 (72%)	40 (0%)	B73gain
3	CCTTGAAAATCCGGAGGACCGAAT**A**	32,750 (28%)	152,247 (100%)	Mo17gain

* the polymorphic site was highlighted in red and the k-mers of the third pair were reversely complemented.

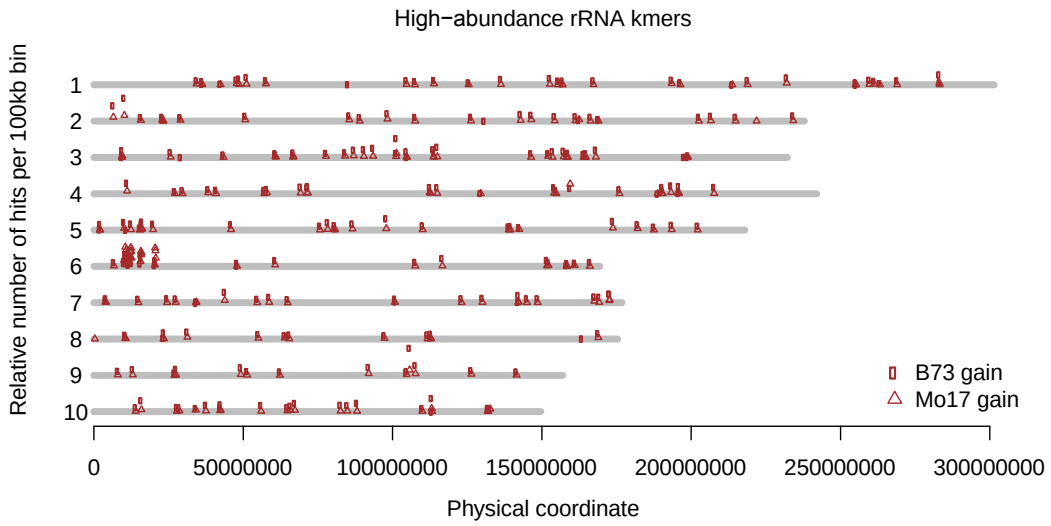

Figure A.1: Genome-wide distribution of B73- and Mo17-gain rDNA k-mers.

B73- and Mo17-gain rDNA k-mers that can be perfectly aligned to the reference genome (B73Ref3). Alignment numbers per bin (100 kb) were plotted versus bin physical locations at the B73Ref3. The 10 minimum alignment hits per bin were required for each circle/triangle points.

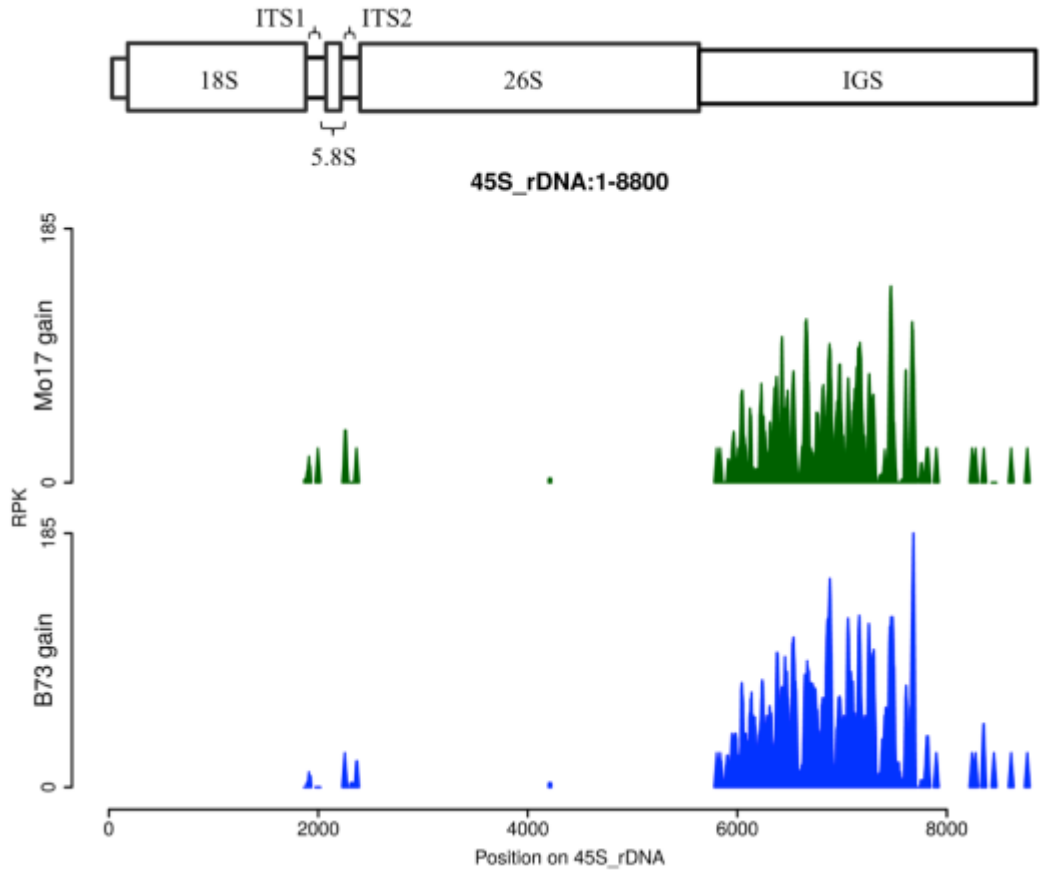

Figure A.2: Distribution of differential abundance rDNA k-mers on 45S rDNA.
Differential abundance rDNA k-mers, including B73-gain (blue) and Mo17-gain (green), were aligned to the 45S rDNA sequence. The count per 1,000 k-mer alignments (RPK) at each position was plotted versus the position on the 45S rDNA. On the top of the figure, the model structure of 45S rDNA was depicted. Three genes, 18S rRNA, 5.8S rRNA, and 26S rRNA, are included in a 45S rDNA unit. ITS and IGS designate internal transcribed spacer and intergenic spacer, respectively.

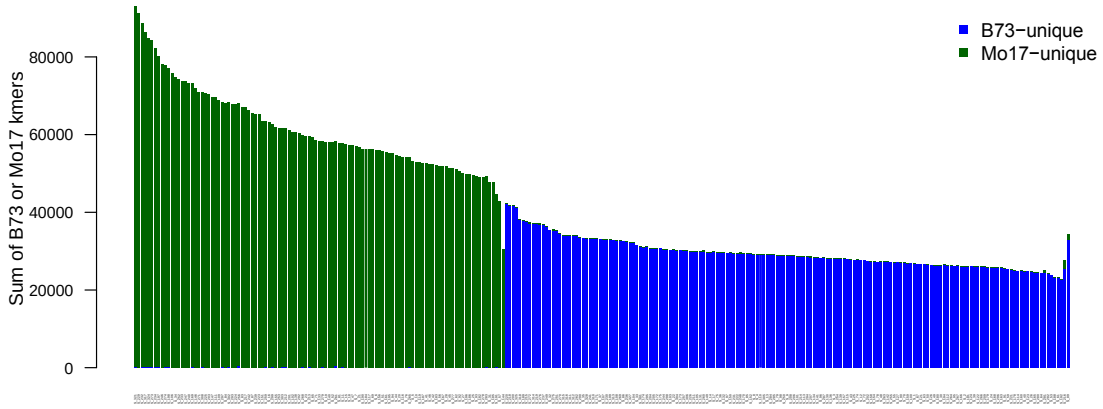

Figure A.3: Barplot of total abundance of B73- and Mo17-specific k-mers.

The total abundance of B73- and Mo17-specific k-mers was determined, normalized, and plotted for each IBM DH line. Bars were colored coded by which genotype of unique k-mers is predominant in that DH line.

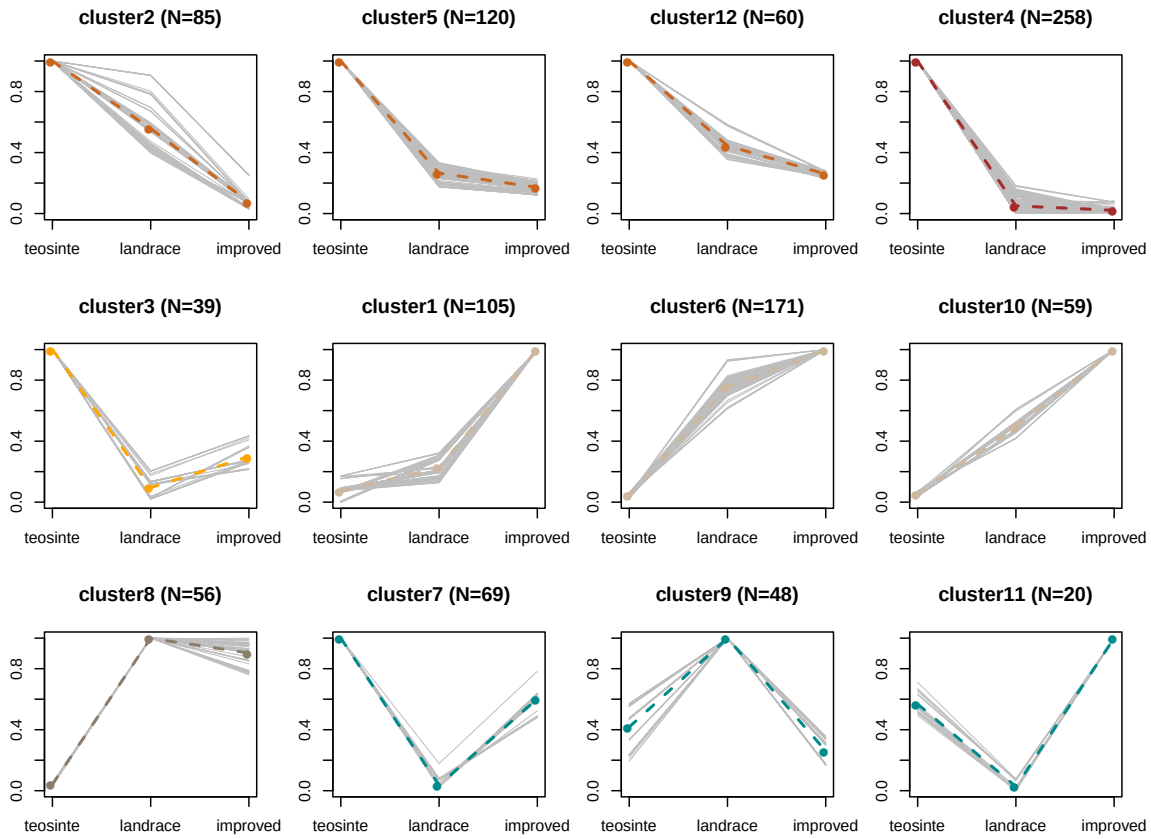

Figure A.4: Change patterns of k-mer abundance

K-mers with significantly differential abundance in teosine, landrace, and improved maize were clustered, resulting in 12 clusters. Each grey line in the figures represents a k-mer. Colored lines are average values from all the k-mers in each cluster. Clusters with a similar pattern were highlighted by the same color.