Extant variation in the maize pan-genome

A DISSERTATION
SUBMITTED TO THE FACULTY OF
UNIVERSITY OF MINNESOTA
BY

Alex B. Brohammer

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

Candice N. Hirsch

March 2019

Acknowledgements

The work herein would not be possible without the effort and support of many loved ones and friends. First and foremost, I would like to acknowledge my family who have given me continual love and encouragement. I would also like to thank my advisor, Dr. Candice Hirsch, for her guidance, patience, and dedication to providing me with opportunities to become a better scientist and a better person. Beyond my advisor, I also owe a great deal of gratitude to members of my graduate research committee, Dr. Nathan Springer, Dr. Robert Stupar, and Dr. Nevin Young. This work would simply not have been realized without their invaluable advice and feedback. Finally, this work was also made possible with the tremendous help of collaborators from the labs of Natalia de Leon and Shawn Kaeppler at the University of Wisconsin-Madison, Robin Buell at Michigan State University, Martha Yandeau-Nelson at Iowa State University, as well as Suzanne McGaugh, and Nathan Springer at the University of Minnesota, Twin-Cities.

I have been incredibly fortunate to receive funding and support from the DuPont Pioneer Bill Kuhn Memorial Fellowship, the MNDrive Scholarship Initiative, the Microbial and Plant Genomics Institute, the Maize Genetics Research Community, the Department of Agronomy and Plant Genetics, and others. I would like to express my sincere appreciation and gratitude for their support. Finally, I must also acknowledge the greater community of researchers and friends that I have had the privilege of interacting with during my time at the University of Minnesota, Twin-Cities. This network of friends and collaborators produced a positive environment that was incredibly fun to be a part of and led to many cherished memories that I will never forget.

Abstract

The publication of the B73 maize reference genome assembly in 2009 was a monumental achievement and marked an important milestone in the field of maize genetics. This resource has been pivotal to countless discoveries since its release. One of the most surprising of these discoveries, however has been the finding that many sequences are missing or significantly diverged from the reference genome. This realization has helped spur the generation of alternative maize reference genome assemblies including one for the elite inbred line, PH207. The first chapter in this work provides a detailed historical perspective of the study of structural variation in maize and presents a review of the current understanding of the maize pan-genome. The middle chapter consists of original research using the PH207 reference genome to understand the significance of differential fractionation to the prevalence of structural variation in maize. The third chapter explores the contribution of transposable elements to variation in the maize transcriptome. Together these sections highlight the importance of using multiple maize reference genomes to understand the extraordinary diversity in the maize genome and point towards the need for a nuanced and contextualized understanding of this sequence diversity.

Table of Contents

List of Tables

Chapter 1.

Chapter 2.

Chapter 3.

List of Figures

Chapter 1.

Chapter 2.

Chapter 3.

Chapter 1. The Maize Pan-Genome[1]

The pan-genome of a species is comprised of genes/sequences that are present in all individuals in the species (core genome) and genes/sequences that are present in only a subset of individuals within the species (dispensable genome). In maize, study of the pan-genome began in the 1940's through cytogenetic experiments and has seen an increase focus in research over the last decade largely driven by advances in genome sequencing technologies. It is estimated there are at least 1.5x as many genes in the pan-genome (greater than 60,000 genes) as there are in any individual's genome (~40,000 genes), with even more variation outside the gene space being observed. This variation has been associated with phenotypic variation and is hypothesized to be an important contributor to the high levels of heterosis often observed in maize hybrids. Due to the high level of variation and the existing genetic and genomic resources, maize has become a model species for plant pan-genomics studies. This chapter will review the mechanisms that can create genome content variation, tools that are available to study the pan-genome, the history of maize pan-genome research ranging from the early cytogenetic studies to today's genomics-based approaches, and the functional consequences of this variation.

---

[1] This work was submitted and accepted for publication to *Spinger* in February 2018, with full citation information provided below. This work was a collaboration between ABB, TJYK, and CNH. ABB, TJYK, and CNH conceptualized the manuscript. The manuscript was written by ABB except for section 2 and section 5.1, which were written by TJYK. All authors read and approved the manuscript.

Brohammer AB, Kono TJ Y, Hirsch CN. 2018. The Maize Pan-Genome. In: *The Maize Genome* Tuberosa R, et al., (eds). Springer.

Introduction

By definition, the pan-genome refers to the non-redundant set of sequences distributed throughout the population of a species. A pan-genome consists of two sets of sequences: those present in every individual in the population, the core genome, and those present in only a subset of individuals, the dispensable genome. The dispensable genome can be further partitioned based on a frequency spectrum. Genes present in low frequencies are part of the 'cloud' set, while those in intermediate and high frequencies are part of the 'shell' and 'soft core' sets, respectively (Koonin and Wolf 2008).

The concept of a pan-genome was introduced by the bacterial community to describe the extensive variation in genome content between species (Tettelin et al. 2005; Medini et al. 2005; Hogg et al. 2007; Tettelin et al. 2008). Technological advances and reduced sequencing technology costs have permitted the pan-genome concept to be extended beyond bacterial species to the plant and animal kingdoms (Li et al. 2010; Computational Pan-Genomics Consortium 2016). Within the plant kingdom, pan-genome analyses have been applied to a number of model and crop species such as *Arabidopsis thaliana* (Cao et al. 2011; 1001 Genomes Consortium 2016), *Brassica oleracea* (Golicz et al. 2016), *Glycine soja* (Li et al. 2014), maize (*Zea mays*; Hirsch et al. 2014), *Medicago truncatula* (Zhou et al. 2017), *Oryza sativa* (Yao et al. 2015), soybean (*Glycine max*; Anderson et al. 2014), and wheat (*Triticum aestivum*; Montenegro et al. 2017).

Depending on the number of genomes that need to be surveyed to capture the full suite of dispensable genes in a species, a pan-genome can be considered open or restricted. The former is common of bacterial species, where with each additional genome that is sequenced new genes are added to the species pan-genome (Tettelin et al.

2008). In contrast, restricted genomes like maize are typical of plant and animal species, where the majority of the pan-genome is captured in a relatively limited set of individuals. In maize, through a transcriptome-based analysis it was estimated that approximately 350 lines were needed to capture the suite of dispensable genes transcribed in the seedling (Hirsch et al. 2014).

Genome content variation in pan-genomes is often described in the context of gene copy number variation (CNV) and gene presence/absence variation (PAV). Copy number variation describes the situation in which additional copies of a particular gene exist in one individual compared to another, and PAV is simply the extreme form of CNV, where one individual possesses one or more copies and another has zero copies of the gene. Genome content variants can result from recombination-based mechanisms, replication-based mechanisms, or other molecular mechanisms and can be divided into two broad categories based on whether they lead to a balanced or unbalanced outcome. This chapter will expand on these mechanisms that generate genome content variation in plant pan-genomes, tools to measure genome content variation, historical and contemporary knowledge on the maize pan-genome, and the functional importance of this variation in driving phenotypic variation within the species.

Mechanisms that generate genome content variation

Transposable elements

Transposable elements (TEs) are genomic elements that have the ability to move in the genome either through a copy-and-paste or cut-and-paste mechanism. Transposable elements were first identified by Barbara McClintock through studying disruption of

pigments in maize kernels (McClintock 1950) and comprise approximately 85% of the maize genome (Schnable et al. 2009). In addition to having direct effects on protein coding sequence and transcript regulation (Tenaillon et al. 2010), TEs also provide multiple avenues for generation of genome content variation. Some classes of TEs "capture" and shuffle gene fragments or entire genes during transposition such as Pack-MULEs and *Helitrons*. Additionally, TEs are a form of dispersed homologous sequence throughout the genome, which can lead to ectopic recombination and the generation of novel gene sequences (Bennetzen and Wang 2014). Finally, the presence of TEs can stimulate meiotic recombination, presumably through the generation of transposase-induced double-strand-breaks (Yandeau-Nelson et al. 2005). Subsequent error-prone repair of these breaks then provides further opportunity for genome content variation.

Unequal recombination

Unequal recombination occurs when homologous chromosomes do not pair exactly during meiosis, and recombination results in gametes with differing DNA content. This is particularly prone to occur in regions of the genome that are already duplicated, because paired sequences may be locally homologous, but may not be globally homologous. Recombination between these improperly paired chromosomes then generates some gametes with more DNA than the progenitor cell, and some gametes with less DNA. Genes arranged in tandem duplicate arrays are common in maize (Messing et al. 2004; Schnable et al. 2009), and provide opportunities for genome content variation via unequal pairing and recombination of duplicated sequences. For example, the *A1-b* locus in maize is a naturally occurring tandem duplication of the

anthocyaninless1 (*a1*) gene that has been well characterized for unequal recombination (Yandeau-Nelson et al. 2006). In this case, unequal pairing of the duplicated genes occurred preferentially between homologous chromosomes but could also occur between sister chromatids. Unequal recombination rates at the duplicated locus were similar to equal recombination rates at non-duplicated *a1* loci, suggesting that unequal recombination is a common phenomenon at this locus.

Non-allelic homologues

Similarly to unequal recombination, segregation of single-copy homologues in non-allelic positions can also lead to changes in gene copy number in the genome (Emrich et al. 2007). Mating between two individuals carrying single-copy homologues in non-allelic positions will results in progeny that are hemizygous for each of the homologues. Independent assortment, or meiotic recombination if the homologues are linked, generates gametes that have variable copy number for the homologues. Inbred progeny produced from these gametes then have zero, one, or two copies of the non-allelic homologues, resulting in apparent *de novo* copy number variation. An example of this phenomenon in maize is two loci involved in elongation of fatty acid precursors for surface lipids, gl8a and gl8b. These two loci are unlinked paralogs with 96% nucleotide sequence identity in B73 that can form *de novo* copy number variation (Dietrich et al. 2005). On a genome-wide scale, several dozen genes were documented to be non-allelic homologues in a single recombinant inbred line population that showed apparent *de novo* copy number variation through segregation of the non-allelic homologues (Liu et al. 2012). This *de novo* copy number variation was hypothesized to contribute to the

phenotypic transgressive segregation observed in the population across a number of phenotypic traits.

Horizontal gene transfer

Horizontal gene transfer (HGT) refers to the asexual transfer of genes between organisms of divergent evolutionary lineages. Maintenance of a newly transferred gene as a segregating genome content variant depends on several events. First, the horizontally transferred gene must integrate into a cell that gives rise to gametes in order for it to be transmitted into subsequent generations. It must then not be lost due to genetic drift, and provide strong enough selective advantage to be maintained in a population. As such, it is hypothesized that horizontally transferred genes that persist as segregating variation within a population have a particularly high likelihood of contributing to phenotyping variation.

Horizontal gene transfer was first observed in bacteria (Freeman 1951), and is now known to be highly prevalent among bacterial species. In bacteria, HGT occurs through random uptake of extracellular DNA, incorporation of viral DNA into the host genome, or direct transfer of plasmids among individuals (Syvanen 2012). While rare in plants, HGT has been observed via viral DNA repeats in *Nicotiana tabacum* (Bejarano et al. 1996). Expressed transfer DNAs from *Agrobacterium rhizogenes* have also been observed in cultivated sweet potato (Kyndt et al. 2015). Plant-to-plant HGT has also been documented in parasitic species. For example, a nuclear gene in *Striga hermonthica*, a hemiparasitic plant that can cause devastating crop loss in species such as *Sorghum*

bicolor, has been found to have high similarity to genes from *S. bicolor*, suggesting HGT as an origin for this gene in *S. hermonthica* (Yoshida et al. 2010).

Genome duplication and fractionation

When a genome undergoes a whole genome duplication event, it generates four copies of each nuclear gene where there were previously just two. New mutations can then begin to cause the function of the duplicates to diverge. Under classical models, the net direction of molecular evolution will be toward the ancestral state of two functional copies of each gene. Three major paths to this outcome are that one duplicate evolves a new function (Ohno 1970), the copies are retained and each partially loses function (Force et al. 1999), or one of the copies completely loses function (Jacq et al. 1977). Following a whole genome duplication, the most common mechanism to restore the ancestral diploid function is through fractionation (Langham et al. 2004; Tang et al. 2008).

An ancient genome duplication event in the ancestor of maize resulted in two subgenomes in present day maize. Analysis of the B73 reference genome assembly showed that one subgenome has greater gene retention than the other, and these subgenomes were named “Maize1” and “Maize2”, respectively (Schnable et al. 2011). Presumably, the paralogues lost during fractionation are not completely consistent between individuals within the species and this variation in gene loss during fractionation generates genome content variation within the species. Many genes that show presence-absence variation within maize also show sequence similarity to genes in closely related grass species (Hansey et al. 2012; Hirsch et al. 2014). This suggests that these genes were

present before divergence of the maize lineage from other grass species and were differentially lost among maize individuals.

Contemporary tools to measure genome content variation

Reference based methods

Reference based methods used to measure genome content variation within species include oligonucleotide arrays and next generation sequencing (NGS) read mapping. Oligonucleotide arrays were the first reference-based method used for conducting genome-wide surveys of genome content variation within maize (Springer et al. 2009; Beló et al. 2010). A specific technique called array-based comparative genomic hybridization (aCGH) was particularly important to advancing our knowledge of PAV and CNV in maize. In this method two labeled DNA samples are hybridized to probe sequences designed to target regions throughout the genome, and signal intensity from each labeled sample indicates its relative copy number. A major limitation to aCGH, and arrays in general, is the inability to detect sequences absent from the reference genome since probes are often designed from a single reference individual. Related issues brought about by limitations of probe design from a single reference individual include biased CNV detection toward deletion discovery and a reduced ability to evaluate regions of high sequence diversity.

Unlike aCGH, NGS methods allow for the discovery of the full suite of structural variants within the species including sequences outside the reference genome (Young et al., 2016). There are three common NGS structural variant detection methods: read-depth, split-read, and read-pair. The read-depth method relies on sequence read depth

from mapping reads to a reference genome assembly as a proxy for copy number. Both the split-read and read-pair methods take advantage of imperfect mapping to identify genomic rearrangements and allow for the detection of all structural variant classes, including inversions and translocations. Paired-end and mate-pair sequence reads have an expected insert size between the two sets of reads. Deviation from these expected distances between the two reads can be used to identify structural variations. The read-pair method uses reads whose distance or orientation between mapped reads from the same fragment is discordant with the reference genome to detect structural variation. The split-read approach to structural variation detection uses information from paired-end sequence reads where one of the pairs maps accurately while the other pair maps only partially or fails to map entirely. The split-read approach can also be expanded to splitting an individual read and identifying reads in which only a portion of the read can accurately map to the reference genome as another method to identify structural variation.

Each method of NGS structural variation detection has its own set of biases (Alkan et al. 2011), and each have variable sensitivities. Many of the available structural variation callers were originally developed to work with human cancer data or model mammalian species and may provide unreliable results or require extensive knowledge and tuning of parameters to be properly used with plant genomes. Combining at least two of these structural variation detection methods into a hybrid structural variation caller (i.e. SURVIVOR; Jeffares et al. 2017) that reports consensus structural variations can overcome some of these issues. Additionally, some of these methods rely on imperfect read mapping, which can be prevalent when mapping short NGS reads to highly

repetitive plant genomes even in the case of reference genome reads mapping to the reference genome assembly. Increased read coverage and optimization of mate-pair library sizes can mitigate this challenge, however, long-read sequencing technologies offer the most promise for avoiding inconsistent structural variation detection in repetitive regions and for the detection of large structural variants.

Non-reference based methods

With reference-based variant detection there is an ascertainment bias that is caused by the reliance on a single reference genome assembly. One method for characterizing gene content variation beyond a single reference genome assembly is through direct comparison of multiple *de novo* genome assemblies. Schatz et al. demonstrated the power of this approach by generating *de novo* genome assemblies of *indica*, *aus*, and temperate *japonica* rice strains, where they identified several megabases of variable sequence between the three strains (Schatz et al. 2014). This approach has also been used in maize where approximately 2,700 novel genes were identified in a comparison of two *de novo* genome assemblies of elite inbred lines from opposite heterotic groups (Hirsch et al. 2016).

Direct comparison of whole genome *de novo* assemblies allows for detailed analysis of variation outside of a single reference genome, however a major disadvantage is the cost and computational effort required to bring these studies to fruition. This disadvantage is important for pan-genome studies because it often leads to a small number of genotypes being assayed and an underestimate of dispensable genome content within species. An alternative approach is to use the transcriptome as a proxy to evaluate

the gene space within a species pan-genome. This approach has the advantage of reducing both the amount of sequencing and computation required in pan-genome studies. In maize, the gene space is only ~97Mb of the genome, and as such, this approach was able to be used to study the maize pan-genome using over 500 accessions (Hirsch et al. 2014)

Recent improvements in assembly algorithms and the continued decline in sequencing costs are making multiple *de novo* genome assemblies within a species more practical (Schatz 2017; Wetterstrand 2018). An example of this shift towards generation of *de novo* genome assemblies for pan-genome analysis is the assembly and annotation of a panel of 54 *Brachypodium distachyon* accessions by Gordon and colleagues (Gordon et al. 2017). For seven years, only two reference genome assemblies for maize were available: the B73 reference genome, and Palomero Toluqueño, a popcorn landrace (Vielle-Calzada et al. 2009). In the span of just three years, nine additional genome assemblies were made publicly available (W22 - GenBank assembly accession GCA_001644905.2; F7 and Ep1 - (Unterseer et al. 2017); PH207 - (Hirsch et al. 2016); B73 - (Jiao et al. 2017); F2 – (Darracq et al. 2018); Mo17, B104, and CML247 (Maize Genetics and Genomics Database. 2017)).

New and emerging technologies that provide long-range information will help to further improve genome assembly and facilitate structural variant discovery. This information can come from special library preparation protocols for short read sequencing, long read sequencing, or large-scale optical maps. For example, 10x Genomics linked-reads are synthetic long reads that preserve single-molecule information through microfluidic encapsulation technologies. This technology is similar to Illumina

TruSeq Synthetic Long-reads (formerly Moleculo) but does not attempt to reconstruct each fragment. The Dovetail Chicago library preparation protocol relies on the Hi-C method of crosslinking DNA to capture long-range information and, like the 10x Genomics method, the processed reads can be read-out by a short-read sequencer such as an Illumina HiSeq. Third-generation single-molecule sequencing, which includes the technologies of Pacific Biosciences Inc. and Oxford Nanopore Technologies, sequence long DNA fragments to provide long-range linkage information. Finally, a separate method of preserving long-range information is through the construction of optical maps (i.e. OpGen and BioNano Genomics), which use restriction sites as "fingerprints" to resolve chimeric assemblies and identify large structural variations.

Iterative mapping and assembly

A common approach to querying population-scale variation in plant pan-genomes is iterative mapping and assembly. An example of this approach was recently published by Yao et al., who analyzed 1,483 cultivated rice accessions to identify non-reference genome assembly sequences (Yao et al. 2015). In this strategy, all of the individuals are sequenced at low-coverage and then aligned to the reference genome. After filtering to remove contaminants and low-quality reads, the unmapped reads represent dispensable genome sequence. Yao et al., assembled the unmapped reads from *indica* and *japonica* separately so that the dispensable genome of each subspecies could be studied. After annotating protein-coding genes and transposable elements in each dispensable genome, they determined the genomic positions of ~80% of these features relative to the Nipponbare reference genome using linkage disequilibrium mapping. The iterative map

and assemble approach allows for a larger portion of the natural variation to be sampled at a relatively low cost compared to *de novo* assemblies. A limitation of the method is that the specific breakpoints of the PAV are often not clear.

History of maize genome content variation studies

Over the nine years that have passed since the publication of the B73 reference assembly (Schnable et al. 2009), the maize community has developed a nuanced understanding of genomic variation, in particular structural variation within the species. Maize genome content studies can be reviewed as a progression through four relatively distinct epochs: molecular and cytogenetic studies of large-scale chromosomal aberrations, Sanger sequencing applied to bacterial artificial chromosomes (BACs), whole genome scale studies using array technologies, and application of next-generation sequencing to study genome content variation across numerous genotypes. These eras represent a timeline that spans nearly 70 years, with a number seminal discoveries made during each era (Figure 1).

Molecular and Cytogenetic Era

The study of structural variation in maize can be traced back to early observations of genome-size variation among maize and its wild relatives. Extraordinary levels of variation for nuclear DNA content were observed between different maize inbreds and landraces ranging from 9.4 to 25.2 pg 4C content values (Laurie and Bennett 1985). Much of this variation in genome size was attributed to the presence of supernumerary B chromosomes (Ayonoadu and Rees 1971; Poggio et al. 1998), and variation in

heterochromatic knob content that makes up over 8% of the genome on average (Brown 1949; Kato 1976; McClintock et al. 1981; Peacock and Dennis 1981; Rayburn et al. 1985). Wide variation in the copy number of repeat sequences has also been widely observed in maize using molecular and cytogenetic approaches. These high-repeat sequences included ribosomal DNA (rDNA) repeats (Phillips et al. 1974; Buescher et al. 1984), centromere satellite DNA repeats (CentC) (Albert et al. 2010), telomere repeats (Burr et al. 1992), and dispersed repetitive sequences (Hake and Walbot 1980; Flavell 1986; Rivin et al. 1986; SanMiguel and Bennetzen 1998; Meyers et al. 2001). More recent surveys of the maize genome using modern cytogenetic and genomics techniques have confirmed these findings regarding dissimilarities in repetitive DNA content between maize lines (Kato et al. 2004; Liu et al. 2017).

Sanger Sequencing Era

The standardization of shotgun sequencing, improved protocols for BAC library construction, and development of bioinformatic algorithms gave rise to the next era in the study of maize genome content variation in the late 1990s and early 2000s. Comparisons of orthologous regions between related grasses using recombination maps generally revealed broad synteny (Whitkus et al. 1992; Ahn and Tanksley 1993), however, in some cases large-scale rearrangements were observed (Reviewed in Gale and Devos, 1998). Subsequently, sequencing based analyses of classical loci showed that smaller-scale rearrangements of orthologous sequence were much more common (Tikhonov et al. 1999). Soon thereafter, a landmark study discovered that the variation seen between orthologous regions could also be found between maize inbred lines. Using the inbred

lines McC and B73 to examine sequence variation at the *bz* locus, it was shown that four of the predicted genes in the McC haplotype were absent from B73 and many of the retroelements present were derived from independent insertion events (Fu and Dooner 2002). To determine if this result was due to a peculiarity between McC and B73, the region was evaluated across 10 separate inbred lines and four distinct structural variation haplotypes were found. In an accompanying commentary, it was hypothesized that the PAV between haplotypes was the result of differential fractionation between McC and B73 (Bennetzen and Ramakrishna 2002). The *z1C-1* locus was also evaluated using Sanger sequencing and significant variation in gene collinearity between the B73 and BSSS53 haplotypes was observed (Song and Messing 2003). A larger-scale comparison of 2.8 Mb of sequence between B73 and Mo17 revealed extensive stretches of nonhomology, in which more than one-third of the genes in the regions examined were variable in their presence (Brunner et al. 2005).

These studies raised numerous questions. What is the genetic mechanism that gives rise to these presence/absence variants? What proportion of the gene complement is dispensable? Do presence/absence variants encode functional proteins? The first of these questions was addressed in a follow-up study by Dooner and colleagues who found that the variability in genic content at the bz locus could be attributed to Helitron elements (Lai et al. 2005). This was further supported via a genome-wide comparison of the inbred lines, B73 and Mo17, in which it was estimated that only ~80% of genomic segments were shared between these two lines based on hybridization to probes designed from genic sequences (Morgante et al. 2005). In-depth characterization of nine of the nonshared sequences showed that all but one displayed the hallmarks of Helitron capture

(Morgante et al. 2005). At this time, prior to the completion of the B73 reference genome, it was hypothesized that any one line would contain only 85% of functional maize genes (Buckler et al. 2006).

Array-based comparative genomic hybridization era

The question of how many maize genes are affected by structural variation genome-wide was not addressed until the publication of the B73 reference genome (Schnable et al. 2009) and the subsequent development of an aCGH platform (Springer et al. 2009). A seminal paper from this era by Springer et al., showed that 180 high-confidence genes were present in B73 and absent in Mo17 (Springer et al. 2009). In addition to over 400 CNVs, a 2.6 Mb stretch of sequence harboring 31 genes was identified that was completely missing from 17 of 24 inbred lines that were subsequently evaluated. This pattern of CNVs being common in maize populations has been recapitulated in other studies. A comparison of 14 inbred lines showed approximately half of over 2,100 identified CNVs were at high allele frequency (Beló et al. 2010). In a further comparison of 19 diverse maize inbred lines and 14 teosinte accessions, 3,410 CNVs were detected, ~86% of which were shared between maize and teosinte (Swanson-Wagner et al. 2010). These studies marked an important advance in knowledge not only due to the genome-wide scale of the studies, but also because they showed that low-copy expressed genes can be PAVs and CNVs, not just repetitive elements and pseudogenes.

Second and third generation sequencing era

The growth of next-generation sequencing technologies is closely tied to the next era of maize genome content variation studies. The initial maize HapMap study utilized

sequencing-by-synthesis technology to inventory variation in the low copy portion of the genome across 27 diverse inbred lines and estimated that B73 contained only 70% of the low-copy maize sequence (Gore et al. 2009). The second-generation HapMap study also inventoried standing variation, but in an expanded collection of 103 inbred lines that included landraces and wild relatives (Chia et al. 2012). This study described the maize genome as being in 'flux' with high levels of read-depth variants (RDVs). This description was based on scanning the genome in 10-kb bins and finding that more than 90% of the tested bins displayed greater than twofold variation in read-depth across the individuals. Further, these RDVs were enriched for GWAS hits indicating their importance to phenotypic variation.

A number of subsequent studies have expanded beyond the reference genome assembly using iterative mapping and assembly approaches. In the first of this type, a set of six elite Chinese inbred lines were resequenced, and 570 novel gene sequences absent from the B73 assembly with an average coding sequence length of 527 bp were discovered (Lai et al. 2010). Of these 570 novel genes, 413 had high coverage from B73 resequencing reads while the remaining 157 did not, suggesting that the latter were true PAVs. Further analysis of the subset of the PAVs that did not have high resequencing coverage showed that many segregated in accordance with heterotic group and did not have paralogs elsewhere in the genome. A similar approach was taken using RNA-seq of 21 diverse inbred lines across heterotic groups that identified 1,321 novel transcripts outside of the reference genome assembly, in which ~11% were heterotic group specific (Hansey et al. 2012). Finally, in a study of 503 diverse inbred lines that again used an RNA-seq mapping and assembly approach, over 20,000 transcribed sequences were

identified that were not present in the B73 reference genome assembly, and it was determined that in this set of lines the closed maize pan-genome could be represented by ~350 lines (Hirsch et al. 2014). Subsequently, a novel method to convert GBS tags to pan-genome anchors across more than 14,000 genotypes found that B73 represents ~74% of the low-copy sequence present in maize (Lu et al. 2015). In this study, PAV SNPs were enriched for significant GWAS hits, but they were also negatively correlated with gene density and recombination frequency.

A new era in the study of maize genome content variation is emerging with the publication of multiple *de novo* genome assemblies and the availability of a new B73 reference genome assembly. The new B73 reference genome is a substantial improvement over the previous sanger-based assembly with a 52-fold increase in contig length. Comparisons of this B73 genome assembly with the optical maps of two other inbreds, Ki11 and W22, showed that only 32% and 39% of the optical maps could be mapped to B73, respectively. Moreover, a large proportion of the aligned region showed evidence for structural variation including 257 PAVs missing in Ki11 and W22 (Jiao et al. 2017). *De novo* assembly of iodent founder line PH207 allowed for a direct genome to genome comparison of gene content to B73 and reported 1,169 B73- and 1,545 PH207-specific genes in addition to extensive variation in gene family size (Hirsch et al. 2016). F2, an important inbred line in France was assembled and 88 Mb of sequence was reported as unique to F2 in a comparison to B73 (Darracq et al. 2018).

Functional importance of genome content variation

Gene and genome evolution

Genome content variation represents an important class of potentially functional genetic variation. Duplication or deletion of genomic regions may have strong impacts on phenotypic variation, presumably because they disrupt the stoichiometry of gene products in physiological contexts (Torres et al. 2008). This disruption, however, is not necessarily detrimental. In the short term, changes in genome content may confer resilience to sudden stresses (Yona et al. 2012). In longer terms, changes in genome content may provide the starting point for evolutionary novelty and species diversification (reviewed in Van de Peer et al. 2017). Considering a single genetic locus, individuals that contain a gene (or multiple copies of a gene) that is not in the genome of others in a population may be able to perform unique biochemical functions, which may then increase variation for fitness. This is a major mechanism underlying the rise and spread of resistance to certain biotic (Cook et al. 2012) or abiotic (Maron et al. 2013) stresses. Duplicated genes may also provide a starting point for the evolution of novel gene function, because one copy of the gene is potentially released from purifying selection, allowing it to diverge in function (Ohno 1970; Näsvall et al. 2012). Genome content variants outside of protein coding sequences may also have phenotypic effects, and thus contribute to fitness variation. For example, maize transposable elements have been shown to influence neighboring gene expression, resulting in alteration of plant morphology (Studer et al. 2011), and abiotic stress response (Makarevitch et al. 2015). However, maintenance of increased copy number or unique biochemical pathways come at a cost, and gene

duplicates are often purged in the absence of selective pressure to maintain them (Berglund et al. 2012).

Phenotypic association and cloned genes

The phenotypic importance of genome content variation (CNVs and PAVs) has been shown through a number of genome-wide studies. The second generation maize HapMap study (Chia et al. 2012), was particularly noteworthy as one of the first genome-wide studies to relate genome content variation to phenotypic variation in traits of agronomic importance. A subsequent association mapping experiment incorporated data from the HapMap studies to perform association mapping across 41 diverse phenotypes (Wallace et al. 2014). In both cases, the authors reported that while SNPs were most often associated with GWAS hits by virtue of their prevalence, CNVs ere the most highly enriched polymorphism class in GWAS hits relative to their genome-wide frequency. In another study that conducted GWAS for key developmental transitions including the juvenile-to-adult vegetative and the vegetative-to-reproductive transitions it was shown that novel gene associations were identified using transcript abundance and transcript PAV as markers relative to analyses that used only SNP markers (Hirsch et al. 2014). Presumably, some of the transcript PAV markers used in this study are based on genomic level PAV. A comparison of two maize *de novo* genome assemblies and the transcriptome profiles across six tissues from these genotypes revealed that approximately half of the transcript PAVs that were observed were the product of genome level PAV (Hirsch et al. 2016). Furthermore, a broad-scale study across more than 14,000 maize inbred lines found that phenotypic variation in four complex traits was

more associated with SNPs linked to PAVs than to SNPs not linked to PAVs (Lu et al. 2015). Finally, a diversity characterization of maize landraces found that the majority of SNPs associated with altitude adaptation overlapped regions of the genome with large-scale structural variation (Romero Navarro et al. 2017).

Despite the extensive levels of PAV and CNV detected across maize and the enrichment of structural variation in GWAS hits, there are relatively few examples of well characterized phenotypes in maize that result from a specific structural variant (Table 1). One of the first examples of a structural variant affecting a phenotype in maize was enhanced Aluminum tolerance resulting from copy number amplification of the MATE1 gene, a transporter from the multidrug and toxic compound extrusion family (Maron et al. 2013). The tunicate phenotype of pod corn (*Zea mays* var *tunicate*), is another example of a structural variant affecting a developmental phenotype (Wingen et al. 2012; Han et al. 2012). The characteristic phenotype of glume covered kernels in the *Tunicate1* (*Tu1)*, mutant is the result of ectopic expression of *Zmm19*, a MADS box transcription factor, in developing maize inflorescence. The ectopic expression of *Zmm19* is manifested through a ~1.8 Mb inversion associated with a *Mutator-like* transposon. A more extreme tunicate phenotype caused by duplication of two genes at the breakpoint of the rearrangement can also be seen. The *White Cap* (*Wc*), locus in maize is another example of structural variation brought about through transposon rearrangement (Tan et al. 2017). Variable repeats of a carotenoid-degrading enzyme, *Ccd1*, at this locus confers quantitative variation for grain color and is the basis for the white-endosperm phenotype. Another example of a structural variant associated with a mutant carotenoid phenotype is the Maize *white* seedling (*w3*) locus. This classical mutant phenotype was recently shown

to be caused by a complete gene deletion of a homogentisate solanesyl transferase (HST) gene (Hunter et al. 2018). Finally, at the *sugary-enhancer* (*Se1*) gene that is important for fresh market sweet corn, there is a recessive allele (*se1*) that is a 630 bp deletion, which eliminates the entire open reading frame of *Se1* and results in loss of normal *Se1* transcript and function. The recessive allele in combination with *sugary1* results in increased sugar content and high levels of water-soluble polysaccharide in the endosperm (Haro von Mogel et al. 2013).

While there are only a few examples of cloned genes in maize with natural PAV/CNV alleles, there are numerous other examples across the plant kingdom (Table 1). These cloned examples in other species have a range of phenotypic outcomes from biotic/abiotic stress tolerance to developmental impacts and production of novel secondary metabolites. The technological advances described earlier are decreasing the barriers to *de novo* genome assembly, which will facilitate CNV and PAV discovery and reduce the recalcitrant nature of studying the phenotypic outcomes of these genomic features. It is anticipated that as multiple reference genome assemblies become available for various plant species, including maize, the ability to identify and characterize functional structural variants will improve.

Heterosis

Since the discovery of interspecific gene content variation in maize there has been considerable interest in the potential role of variable genes in heterosis. Here we define heterosis in the breeding-context as the difference in performance of a hybrid relative to the performance of its better inbred parent, otherwise known as better parent heterosis.

Many non-mutually exclusive hypotheses have been put forward to explain the mechanism of heterosis in maize (reviewed Kaeppler 2012; Schnable and Springer 2013). The three classical quantitative genetics hypothesis include dominance, overdominance, and epistasis. The dominance hypothesis, which posits that heterosis results from the complementation of mildly deleterious alleles present in inbred parents, is most often invoked in the context genome content variation.

Based on early Sanger sequencing work it was hypothesized that maize genotypes with complementary dispensable gene subsets would produce hybrid offspring with a more complete suite of quantitative-effect dispensable genes (Fu and Dooner 2002). One of the reasons for invoking gene content variation in discussions of heterosis is that it is consistent with the breeding practice of crossing inbreds from complementary heterotic groups to form superior hybrids. Crosses between opposite pools (i.e. Stiff Stalk Synthetic x Non-Stiff Stalk Synthetic) would be expected to generate a more full gene complement compared to crosses that take place within heterotic group crosses (i.e. Stiff Stalk Synthetic x Stiff Stalk Synthetic). This model was supported by later work that demonstrated patterns in PAVs that reflect heterotic groups (Lai et al. 2010; Hansey et al. 2012). Lai et al. resequenced six elite Chinese breeding lines and found that many of the structural variants identified were private to a single heterotic group (Lai et al. 2010). A second study, based on RNA-seq of 21 diverse North American breeding lines, found 145 loci absent from B73 that also showed heterotic group patterning (Hansey et al. 2012). Further, in a comparison of two *de novo* assemblies from genotypes that have high specific combining ability, over 2,500 PAVs were identified as well as extreme expansion and contraction of gene families (Hirsch et al. 2016). While this association is

suggestive, clear evidence for a causal role of gene content variation in heterosis has yet to be realized.

Dosage balance

The concept of dosage balance has been formalized as the Gene Balance Hypothesis, which declares that balanced stoichiometry among members of multi-subunit complexes is critical for optimal function of the macromolecular complex (Birchler and Veitia 2007; Birchler and Veitia 2010; Birchler and Veitia 2012). In practical terms, this posits that gene products that function as part of a complex or interact closely within a certain biochemical framework will likely have an optimal ratio of subunits. Any change that modifies this ratio, such as alteration of gene copy number, will cause a deviation from the optimal balance. This can have important implications for gene expression regulation and, in the context of this chapter, on the evolutionary fate of CNVs. One line of evidence supporting this hypothesis comes from the study of genes retained in duplicate following the most recent polyploidization event in maize and other paleopolyploids. It has been shown that functional classes of genes that participate in macromolecular complexes such as transcription factors and signaling components, are more likely to be retained than other functional classes (Woodhouse et al. 2010). This bias also extends to non-polyploidy derived copy-number polymorphisms. Given that many CNVs segregate, inbreds may contain a more dramatic shift from optimal dosage when averaged across the genome compared to the hybrid state due to complementation. Under this model of heterosis, increased inbred performance is expected to lead to decline in the number CNVs observed across the genome (Kaeppler 2012).

Future bioinformatic challenges in the era of multiple genome assemblies

The number of sequenced and assembled plant genomes is growing at an exponential rate (Michael and Jackson 2013), and many species, including maize, have genome assemblies from multiple individuals within the species. This burst of activity is due to the realization that a single reference genome is not representative of the variation present in a species. The availability of additional reference genomes will greatly facilitate structural variation characterization and lead to a better understanding of the maize pan-genome. Before new genomic resources can be effectively used, however, current bioinformatic workflows need to be modified to accommodate multiple reference genomes. Some questions raised by the Computational Pan-Genomics Consortium (Computational Pan-Genomics Consortium 2016), include:

1. What is a reference genome? The genome of a selected individual, the consensus sequence from a population, or a maximal genome with all sequences detected?
2. How do we efficiently translate coordinates and compare genome features from one genome assembly to another genome assembly?
3. Should we abandon the concept of single, linear reference genome and move towards a graph-based approach?

The incorporation of alternative/novel loci is an important step towards more comprehensive representation of sequence diversity. One challenge associated with their adoption, is that read mapping software must be modified to support alternate loci. The

development of “alt-aware” algorithms is an area of extensive development. While these loci are useful for capturing variation at regions of interest, they do not attempt to fully represent variation at the pan-genome level. In order to best utilize the full suite of variation present in a population, research communities will need to move beyond the representation of reference genomes as linear strings. The idea of adopting a graph-based genome has been advanced by the Computational Pan-Genomics Consortium, which has advocated for a paradigm shift in how we think of reference genomes. Graph based structures are already commonly used in assembly software in the form of de Bruijn graphs, which are directed graph structures in which nodes represent *kmers* (unique strings of length *k*) and edges represent an overlap of *k-1* bases between two nodes. Similarly, a basic graph structure might encode shared sequences as nodes in a graph and novel sequences as edges.

Moving from a reference genome being a linear representation of single genotype to a graph based data structure that represents an amalgam of haplotypes will require new a consensus data structure, new coordinate systems, and the modification of genome browsers and other tools. However, as additional genome assemblies become available and our knowledge of the size and complexity of species pan-genomes continues to grow the difficulty in these challenges will be far outweighed by the benefit to biological understanding and utilization of diversity in plant species.

Table 1. Examples of copy number variants (CNVs) and presence/absence variants (PAVs) with known phenotypic outcomes.

Species	Variant Type	Trait	Reference
Barley	CNV	Boron toxicity tolerance	(Sutton et al. 2007)
Barley	CNV	Freezing tolerance	(Knox et al. 2010)
Barley	CNV	Flowering time	(Nitcher et al. 2013)
Cucumber	CNV	Reproductive morphology	(Zhang et al. 2015)
Maize	CNV	Tunicate phenotype	(Wingen et al. 2012; Han et al. 2012)
Maize	CNV	Aluminum tolerance	(Maron et al. 2013)
Maize	CNV	Grain color	(Tan et al. 2017)
Maize	PAV	Carotenoid synthesis	(Hunter et al., 2018)
Opium poppy	PAV	Noscapine synthesis	(Winzer et al. 2012)
Palmer amaranth	CNV	Glyphosate resistance	(Gaines et al. 2010)
Rice	PAV	Phosphorus uptake	(Schatz et al. 2014)
Rice	PAV	Submergence tolerance	(Schatz et al. 2014)
Soybean	CNV	SCN resistance	(Cook et al. 2012)
Tomato	CNV	Fruit size	(Xiao et al. 2008)
Wheat	CNV	Photoperiod response	(Díaz et al. 2012)
Wheat	CNV	Dwarfing	(Li et al. 2012)
Wheat	CNV	Freezing tolerance	(Zhu et al. 2014)
Wheat	CNV	Winter hardiness	(Würschum et al. 2017)

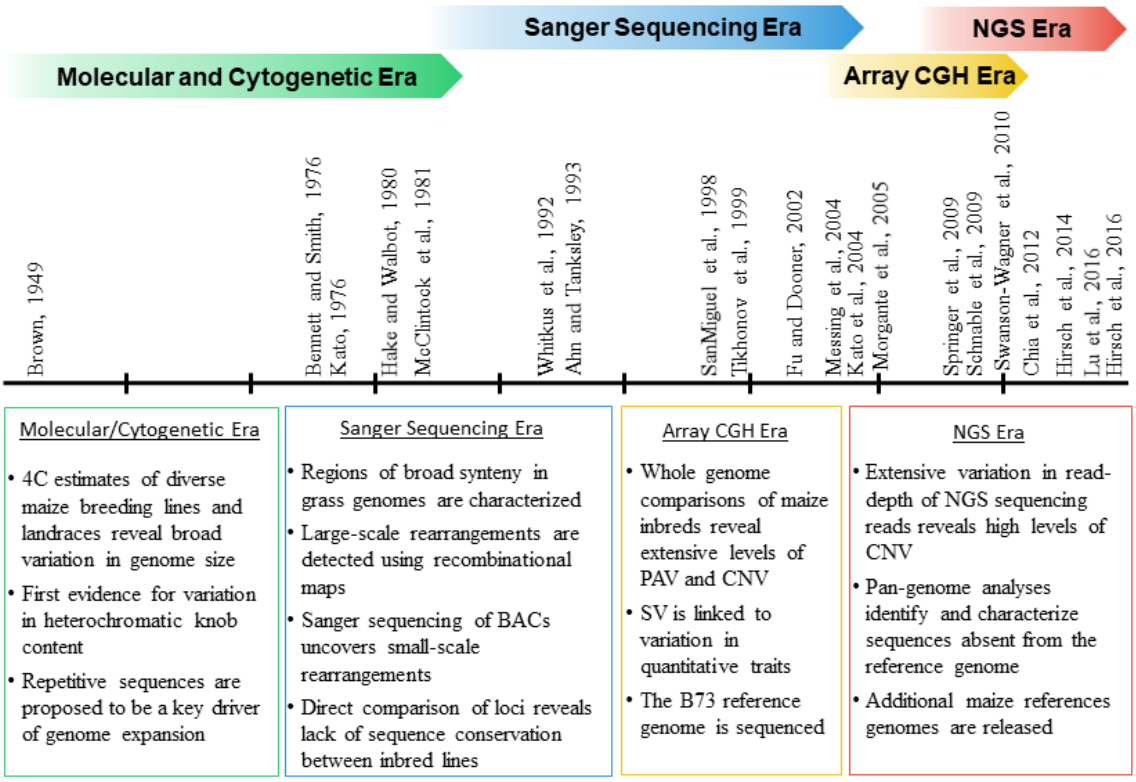


Figure 1. Timeline of seminal studies leading to our current understanding of the maize pan-genome and functional consequences of genome content variation within maize. BAC – bacterial artificial chromosome; PAV – presence-absence variation; CNV – copy number variation; SV – structural variation; NGS – next-generation sequencing

Chapter 2. Limited role of differential fractionation in genome content variation and function in maize (Zea mays L.) inbred lines[2]

Maize is a diverse paleotetraploid species with considerable presence/absence variation and copy number variation. One mechanism through which presence/absence variation can arise is differential fractionation. Fractionation refers to the loss of duplicate gene pairs from one of the maize subgenomes during diploidization. Differential fractionation refers to nonshared gene loss events between individuals following a whole genome duplication event. We investigated the prevalence of presence/absence variation resulting from differential fractionation in the syntenic portion of the genome using two whole genome *de novo* assemblies of the inbred lines B73 and PH207. Between these two genomes, syntenic genes were highly conserved with less than 1% of syntenic genes being subject to differential fractionation. The few variably fractionated syntenic genes that were identified are unlikely to contribute to functional phenotypic variation, as there is a significant depletion of these genes in annotated gene sets. In further comparisons of 60 diverse inbred lines, non-syntenic genes were six times more likely to be variable compared to syntenic genes, suggesting that comparisons among additional genome

[2] This work was published in *The Plant Journal* in January 2018, with full citation information provided below. This analysis was a collaborative effort, with several authors contributing to the final manuscript: the experiment was conceptualized by ABB, TJYK, NMS, SEM, and CNH. Data analysis and programming was performed by ABB except for the tandem duplicate analysis which was performed by TJYK. The manuscript was written by ABB, TJYK, SEM, and CNH. All authors read and approved the manuscript.

Brohammer AB, Kono TJY, Springer NM, McGaugh SE, Hirsch CN. 2018. The limited role of differential fractionation in genome content variation and function maize (Zea mays L.) inbred lines. *Plant Journal*. 93:131-141.

assemblies are not likely to result in the discovery of large-scale presence/absence variation among syntenic genes.

Introduction

Whole-genome duplication events are prevalent throughout the lineage of many plant species. The diversification of seed plants and angiosperms occurred shortly after two whole-genome duplication events (Jiao *et al.*, 2011). Whole-genome duplication events are a major driving force for angiosperm diversification and may provide a conduit for domestication to occur (Tank *et al.*, 2015; Salman-Minkov *et al.*, 2016). These events have profound impacts on genome structure, transcriptional regulation, biochemical functions, and ultimately phenotypes (Reviewed in PS. Soltis and DE. Soltis, 2012).

Maize (*Zea mays*) has a long history of genetic analyses that have led to the current understanding of its polyploid history, including the most recent allopolyploid event. The ancient progenitors of maize split from one another around 12 million years ago, closely following the divergence of maize and sorghum (Swigonová *et al.*, 2004). The subsequent hybridization of the two maize progenitors created duplicate copies of genes genome-wide (homeologs) (Gaut and Doebley, 1997). Synteny analysis revealed that ~60% of maize genes are co-orthologous to a location in the ancestral state of brachypodium, rice, or sorghum (Schnable *et al.*, 2012). Each homoeologous gene is derived from one of the maize progenitors and as such there exists two maize subgenomes, maize1 and maize2, that together make up the modern paleotetetraploid maize genome.

Following this most recent whole-genome duplication event, maize underwent chromosomal breakage and fusion events that returned the 2n = 40 allotetraploid state to the 2n = 20 diploid state. This process of diploidization led to extensive fractionation, the loss of a gene from a homoeologous gene pair, in the maize genome (Langham *et al.*, 2004; Woodhouse *et al.*, 2010). This phenomenon of genome fractionation is distinct from other forms of DNA removal in that it is associated with the loss of genic sequence, rather than repetitive sequence, despite occurrence via the same intrachromosomal deletion mechanism (Woodhouse *et al.*, 2010). Homeolog loss has been shown to be biased such that homeologs from the maize2 subgenome are 2.3 times more likely to be lost than the homeologs from the maize1 subgenome (Schnable *et al.*, 2011). Maize also exhibits unbalanced homeolog expression bias, with the maize1 gene copy often being more highly expressed when both homeologs are retained (Schnable *et al.*, 2011). Furthermore, syntenic genes and, in particular, genes from the maize1 subgenome are more likely to contribute to maize phenotypic variation (Renny-Byfield *et al.*, 2017; Schnable and Freeling, 2011), and are subject to higher levels of purifying selection (Pophaly and Tellier, 2015).

Analyses within maize have shown that copy number variation (CNV) and presence/absence variation (PAV) are common throughout the genome (Springer *et al.*, 2009; Swanson-Wagner *et al.*, 2010; Lai *et al.*, 2010; Hansey *et al.*, 2012; Hirsch *et al.*, 2014). There is an emerging body of evidence that indicates this genome content variation has important consequences for the extensive phenotypic variation present in maize (Chia *et al.*, 2012; Maron *et al.*, 2013; Hirsch *et al.*, 2014; Lu *et al.*, 2015). Copy number variation and PAV has been shown to underlie variation for important agronomic

traits such as aluminum tolerance (Maron *et al.*, 2013), starch metabolism (Haro von Mogel *et al.*, 2013), flowering time (Nitcher *et al.*, 2013), biochemical networks (Winzer *et al.*, 2012), and disease resistance (Cook *et al.*, 2012). This form of variation is dispersed throughout the genome and is enriched among loci identified in genome wide association (GWAS) studies (Chia *et al.*, 2012; Wallace *et al.*, 2014; Lu *et al.*, 2015).

Fractionation is thought to be an ongoing process (Woodhouse *et al.*, 2010, Schnable et al., 2011), and as such can lead to continual differential fractionation between individuals within the species. Differential fractionation is one mechanism that can create PAV within individuals of a species and is defined by multiple cycles of post-tetraploidy gene loss that results in differences in syntenic gene content among individuals within a species. Maize is uniquely positioned to assess differential fractionation and differential loss of both syntenic gene copies, hereafter referred jointly as differential fractionation, due to the recent whole genome duplication event and the availability of two whole genome *de novo* assemblies with structural annotations (B73 and PH207; Jiao *et al.*, 2017; Hirsch *et al.*, 2016). In this study, we sought to evaluate the prevalence of differential fractionation among maize inbred lines and to assess the functional significance of this variation with regards to expression variation and phenotypic variation.

Results

Macro-level synteny between B73 and PH207 is nearly identical

SynMap (Lyons *et al.*, 2008) was used to identify blocks of syntenic genes between the maize genome (both B73 and PH207) and the sorghum and rice genome

(Figure S1). On a macro scale, there were no substantial differences in syntenic block composition or subgenome assignment between the two maize genomes (Figure 1 and Figure S1). The maize1 subgenome collectively encompassed 55% (1.16 Gb) of the B73 genome and 54% (1.13 Gb) of the PH207 genome, while the maize2 subgenome encompassed 32% (0.66 Gb) of the B73 genome and 30% (0.63 Gb) of the PH207 genome with the remaining 13% (0.28 Gb) and 16% (0.34 Gb) consisting of non-syntenic regions.

Our syntenic annotation for the B73 v4 reference genome was largely consistent with reports based on current and past versions of the B73 genome assembly (Schnable *et al.*, 2009; Jiao *et al.*, 2017). In total, the raw SynMap analysis identified 21,568 genes in B73 and 20,446 genes in PH207 with syntenic orthologs in sorghum plus an additional 408 genes in B73 and 421 genes in PH207 that could only be identified through comparisons of rice orthologs.

Curation of syntenic assignments

The number of syntenic genes identified through SynMap could be underestimated due to limitations of the syntenic identification software, assembly errors, or incomplete gene annotation in one or both assemblies. Significant effort was undertaken to validate and identify missed syntenic assignments including a series of BLAST alignments, alignment of resequencing data, and curating assembly gaps (see Experimental Procedures).

There were many regions in the B73 and/or PH207 genome with homology to syntenic loci that were not annotated as gene models that may reflect putative,

unannotated genes. Inconsistent annotation of these genes in B73 or PH207, could lead to false positive identification of differentially fractionated genes. All putative differentially fractionated genes were aligned to the expected syntenic position in the opposite genotype (e.g. the B73 gene putatively lost in PH207 was aligned to the expected syntenic location in PH207) and mapping coordinates of each gene that could be aligned in place of an annotated gene model were included in downstream analysis. The final list of syntenic assignments included 354 loci in B73 and 1,148 loci in PH207 that likely represent *bona fide* gene models that were not previously annotated. In B73, 49.9% of the loci, and in PH207 55.6% of the loci had RNA-seq read coverage from at least one of five sampled tissues, providing evidence that these loci largely represent missed gene annotations.

Another major source of false negative assignments resulted from fused gene models (i.e. two separate gene models in one genotype correspond to a single "fused" gene model in another), which were identified based on significant mapping of a single gene model in one genotype to two or more adjacent genes in the opposite genotype. In total, 442 instances of fused gene models in B73 and 314 instances in PH207 were identified and removed from downstream analysis.

To correct for false positive differentially fractionated genes in downstream analyses due to gaps in either of the maize genome assemblies, the putative differentially fractionated genes were aligned to annotated genes present on scaffolds or contigs and significant hits were incorporated into the working list of syntenic assignments. Putative differentially fractionated genes within gaps in the assembly were also identified and this information was included in the list of syntenic gene assignments (Table S1). This list

contains 11 B73 and 219 PH207 putative gene models that could not be identified due to assembly gaps.

After the extensive curation of incorrect assignments and recovery of missing assignments the final set of maize syntenic orthologs identified in B73 included 24,514 genes and 24,454 PH207 genes (Table 1). These assignments confirm previous observations of biased fractionation between the maize1 and mazie2 subgenomes in B73 (Schnable *et al.*, 2011), and extend this observation to a second maize genome. We identified 9,255 and 9,239 maize1 singleton genes (maize2 copy fractionated) in B73 and PH207, respectively, compared to 3,777 (B73) and 3,789 (PH207) maize2 singleton genes (maize1 copy fractionated) which supports the finding that genes from the maize2 subgenome are approximately 2.5 times more likely to fractionate than genes from the opposite subgenome (Table 1).

Differential fractionation is not a primary driver of gene content variation

Fractionation is an ongoing process within genomes (Woodhouse *et al.*, 2010), and as such can lead to differential fractionation between individuals within a species. After validating the working list of syntenic assignments and recovering missed assignments, we sought to characterize the prevalence of differential fractionation between B73 and PH207. In total, we identified 112 genes that were putatively fractionated only in B73 and 172 that were putatively fractionated only in PH207 (Figure 2A; Class II-IV). Figure 2A shows an example of a differentially fractionated gene. Of the differentially fractionated genes, 49 (B73) and 93 (PH207) were lost from the maize1 subgenome, while 63 (B73) and 79 (PH207) were lost from the maize2 subgenome.

While there are 12 possible differential fractionation scenarios, only a subset of all possible fractionation scenarios was observed. We characterized the types and relative frequency of fractionation scenarios observed for all putative differentially fractionated genes (Figure 2B). Differential fractionation scenarios have different expectations for frequency based on the level of functional redundancy and the number of unique events that are required to derive the observed state. We hypothesized that segregation for a gene loss would occur most often if another copy was present in the other subgenome, as this redundancy would be less likely to lead to a negative fitness impact. The observed frequencies were consistent with this hypothesis. The most common differential fractionation scenario observed was the presence of a singleton in one genotype and retention of both subgenome copies in the other genotype (Figure 2B; Class II). The next most frequent differential fractionation scenario was when a singleton was retained only in one genotype and the copy from the other subgenome was fractionated in both genotypes (Figure 2B; Class III). The final scenario of differential fractionation required multiple independent loss events between the two genomes, and as expected was the least frequently observed scenario (Figure 2B; Class IV).

Differentially fractionated genes between B73 and PH207 are more likely to exhibit PAV among diverse inbred lines than other syntenic genes

The direct comparison of syntenic gene content between B73 and PH207 revealed little PAV among syntenic genes. However, high levels of PAV may still be observed in non-syntenic genes. To further determine the PAV frequency of the differentially fractioned genes identified above within the species and to extend our analysis to the non-

syntenic gene set, we resequenced 60 diverse inbred lines. All genes with coverage of less than 20% of the gene model length from 12x-65x depth resequencing data were considered significantly deleted or lost. Using this criterion for PAV, 10.1% of B73 syntenic genes and 11.3% of PH207 syntenic genes were classified as PAV among the 60 lines, while 62.5% and 58.2% of B73 and PH207 non-syntenic genes were classified as PAV (Figure 3A). On average, syntenic genes with PAV were absent across 9.1% of the inbred lines, while the subset of non-syntenic genes with PAV were absent across nearly twice as many of the inbred lines (17.2%).

There is a clear distinction between the PAV frequency distribution of syntenic and non-syntenic genes across diverse maize inbred lines (Figure 3B). The frequency distribution for differentially fractionated syntenic genes roughly follows that of non-syntenic genes, which shows substantially higher absence frequency across diverse maize lines than syntenic non-differentially fractionated genes (Figure 3B). The deviation that is seen in the plot for the distribution of syntenic differentially fractionated genes is a result of the small number of genes in this subset (112 in B73 and 172 in PH207). Additionally, the PH207 genome likely has more TEs annotated as genes than the B73 genome, causing this line to have some deviation from the distribution observed for non-syntenic genes. On average, differentially fractionated syntenic genes were absent across 16.2% of the inbred lines compared to less than 1% for non-differentially fractionated syntenic genes.

Differentially fractionated genes with a non-allelic homolog likely represent misassemblies rather than biological observations

Differentially fractionated genes that are lost from a syntenic position in the genome may have their function buffered by additional copies of the gene present in non-syntenic locations in the genome. Previous literature suggested that the maize genome contains many homologs present in non-allelic positions throughout the genome and numerous near-identical paralogs (Liu *et al.*, 2012; Emrich *et al.*, 2007). Potential buffering for differentially fractionated genes was analyzed by examining coverage over non-syntenic gene models in resequencing data. We found that several of these genes (11/111 in B73; 43/161 in PH207), could be uniquely mapped back to a single gene model elsewhere in the genome (Table S2). Thirty-seven of the 43 PH207 genes and 2 of the 11 B73 genes that mapped to a non-syntenic position were on a chromosome that did not contain either of the collinear ancestral blocks. The remaining genes mapped to the expected chromosome but outside the syntenic block.

To determine if these non-allelic homologs were shared with any other genotypes, the draft assemblies of maize inbred lines W22 (GenBank assembly accession GCA_001644905.2), CML247 (Maize Genetics and Genomics Database, 2017), F7 (Unterseer *et al.*, 2017) and Ep1 (Unterseer *et al.*, 2017) were analyzed. All of the cases in which PH207 contained a non-allelic homolog on a chromosome without a collinear block were private to PH207 and retained at the B73 position in the other genotypes. Similarly, the two B73 genes with non-allelic homologs on unexpected chromosomes were only observed in B73 and not in any of the other assemblies including the previous version of the B73 assembly (Figure 4). Although we did not rule out that these are

biological, evidence from multiple genomes suggests that these are predominantly the product of missassembly. These cases of potential buffering were excluded from further analysis of functional properties of differentially fractionated genes.

Differentially fractionated genes are underrepresented among genes of functional significance

The function of genome content variants (CNV and PAV) is generally not well understood and the specific contribution of differential fractionation to phenotypic variation has not been extensively studied (Renny-Byfield *et al.*, 2017). To evaluate the functional consequences of differential fractionation to phenotypic variation, we tested whether differentially fractionated genes were enriched or depleted compared to non-differentially fractionated genes in various annotated gene sets that would indicate importance to phenotypic variation.

The first of these annotated gene sets was the maize classical gene set, which consists of 424 annotated B73 gene models that have been extensively cited in the literature and have previously been shown to be enriched for the presence of syntenic genes (Schnable and Freeling, 2011). An expanded set of 4,461 named genes manually curated by the Maize Genetics and Genomics Database (MaizeGDB.org) was combined with the classical gene list. Of the 4,649 genes in the combined list of non-redundant curated genes, 3,989 were in our list of syntenic genes. There was a significant under-representation of differentially fractionated genes amongst these curated genes (Chi-square with Yates correction, one-tailed p-value 0.0274) with only 16 differentially fractionated genes overlapping with this gene set (Table 2).

Next, a set of highly interconnected 'hub-genes' from mRNA and protein based regulatory networks (Walley *et al.*, 2016) were tested for enrichment for fractionated genes. Among the 4,280 hub-genes overlapping syntenic genes, only 11 were differentially fractionated, which was a highly significant under-representation of hub-genes among the set of differentially fractionated genes (Chi-square with Yates correction, one-tailed p-value 0.0005; Table 2).

Finally, we tested a list of curated maize NAM-GWAS hits across 41 agronomically relevant traits (Wallace *et al.*, 2014). As with the previous gene sets, there were few differentially fractionated genes overlapping GWAS hits, and less than 1% of GWAS hits were differentially fractionated. However, the proportion of differentially fractionated GWAS hits was not significantly different from the proportion of differentially fractionated genes among non-GWAS hits (Chi-square with Yates correction, one-tailed p-value 0.3735; Table 2).

Differential fractionation is also limited at the transcriptome level

The term fractionation is most often invoked in terms of structural gene loss, however fractionation can also be considered at the transcriptome level. We hypothesize that transcriptional fractionation occurs at a higher rate than genome fractionation because gene inactivation can occur through mechanisms other than sequence deletion. To test the rate of transcriptional fractionation, we only considered genes for which both the maize1 and maize2 copies were retained in both B73 and PH207, and at least one homolog was expressed for a total of 4,498 maize1 homolog and 4,436 maize2 homolog sets. We detected 174 cases of "on/off" expression with the B73 gene active and 157

cases in which the PH207 gene was active across five distinct tissues. The rate of false positive detection of "off" genes based on these five tissues was evaluated in B73 using the maize gene atlas, a resource consisting of RNA-seq based expression values for 60 tissues across B73 (Stelpflug et al., 2016). Only 13 of the 174 B73 genes were confirmed as being not expressed across the larger set of tissues. Contrary to our hypothesis, the rate of transcriptional fractionation was as rare as the rate of genome fractionation (~0.001% of genes with transcriptional fractionation and ~0.006% of genes with genome fractionation). These results emphasize the high degree to which these genes are conserved at both the genome and transcriptome levels.

Discussion

Capturing genome diversity in the context of a species pan-genome has received much interest in maize and other plant research communities. These research efforts can be substantially improved by the availability of multiple reference genome assemblies within a species. Recent and ongoing efforts to assemble and annotate additional maize genomes enables more accurate detection of PAV through direct genome comparisons. A deeper understanding of the role of dispensable maize genes and the mechanisms through which gene content variation is created is critical to broader pursuits such as synthetic biology. Here, we take advantage of the two annotated whole-genome maize assemblies that are currently available to study the impact of differential fractionation of genome content variation in maize and the functional consequences of this variation.

Using the B73 and PH207 assemblies, we found the vast majority of post-tetraploidy loss events were shared across inbred lines. Only 112 and 172 fractionation

events were specific to B73 and PH207, respectively. This number, which is likely an overestimate, indicates that differential fractionation has played a limited role in generating the extensive PAV that has been documented in maize (Springer *et al.*, 2009; Swanson-Wagner *et al.*, 2010; Lai *et al.*, 2010; Hansey *et al.*, 2012; Hirsch *et al.*, 2014). B73 and PH207 are inbred lines that have undergone substantial selection and improvement for North American agricultural environments. Additional variation in syntenic gene content may be present in landraces and other diverse sets of germplasm that may have been subject to weaker or divergent selective pressures. Given that syntenic genes are highly enriched among functionally important genes (Schnable and Freeling, 2011), much of the variation in syntenic gene content, if it exists, is likely deleterious variation in the context of modern agricultural systems.

The few syntenic genes that are variable across maize breeding lines are an exception and are unlikely to underlie major phenotypes. Only a small proportion of the variable genes we identified overlap with hub-genes, community curated genes, or GWAS hits. Except for the latter, these overlaps indicate a significant under-representation of differentially fractionated genes among functional gene sets. The failure to meet the significance threshold for the enrichment test of GWAS hits may be a result of differentially fractionated genes being more likely to confer minor quantitative effects compared to genes with major qualitative importance. These findings could also be the result of noise associated with assigning a significant non-genic GWAS hit to the nearest gene or linkage disequilibrium between a causal allele and a neighboring allele with an association of higher significance.

Transcriptional loss can also result in phenotypic outcomes that are equally impactful as those generated through sequence loss. There are several examples of altered transcriptional patterns brought about through transposon insertions that are associated with discernible phenotypic effects (i.e. *Vgt1,* Salvi *et al.*, 2007 and *Tb1,* Studer *et al.*, 2011), and GWAS using transcript PAV as a marker identified significant associations where there was not allelic variation (Hirsch et al., 2014). Due to the numerous mechanisms involved with terminating transcription of a gene (i.e. promoter disruption, methylation changes, etc.) versus sequence deletion, we hypothesized that this variation would be more common. However, we observed transcriptional fractionation at a rate at least as low as the rate observed for genome fractionation of syntenic genes. While there are some examples in the literature and we identified a small number (<20 after accounting for false positives) of additional transcriptional fractionation events, these events are not likely to be driving substantial phenotypic variation within the species. This lack of transcriptional fractionation is consistent with previous results showing an underrepresentation of syntenic genes among differentially expressed genes and genes with non-additive gene action (Baldauf et al., 2016).

Our goal was to assess the contribution of differential fractionation to maize genome content variation and to better understand how this variation relates to functional outcomes. Our results suggest that differential fractionation of syntenic genes plays a minor role in the high levels of PAV in the maize genome and much of the existing PAV among syntenic genes is not likely to have major functional significance. Syntenic and non-syntenic gene sets have different evolutionary constraints and it is becoming increasingly clear that PAV distributions follow different frequencies among these

classes of genes. While PAV among syntenic genes may be an "evolutionary dead-end" due to purifying selection, the same form of variation persists among non-syntenic genes and may contribute to quantitative variation for traits of agronomic importance.

Experimental Procedures

Syntenic gene identification

To identify maize genes in syntenic blocks relative to the ancestral state, we ran the SynMap pipeline for both the B73 v4 (Yinping Jiao *et al.*, 2017), and PH207 v1 (Hirsch *et al.*, 2016), maize assemblies against the sorghum v3.1 (http://phytozome.jgi.doe.gov/) and rice v7 (Ouyang *et al.*, 2007) genomes as the ancestral anchor states. All assemblies were downloaded from Phytozome 12.0.2 (https://phytozome.jgi.doe.gov), except for the B73 v4 assembly, which was downloaded from Gramene release-33 (http://www.gramene.org). SynMap was run using Quota Align to merge syntenic blocks with a coverage depth ratio of 1:2 and the tandem duplication distance was set to 15. All other SynMap parameters were set to default. Homologous genes between B73 and PH207 were identified by assignment to the same ancestral orthologs based first on assignment to the same sorghum syntenic ortholog and then the same rice syntenic ortholog for any maize genes that did not have a defined ancestral state from the comparison with sorghum.

Curation of tandem duplicate genes

Clusters of genes identified as tandem duplicates were filtered to a single representative copy prior to assignment to syntenic regions in the maize genomes (Figure S2). For each group of putative tandem duplicate genes, the amino acid sequences from

the longest transcript of each gene were aligned using clustal-omega (Sievers *et al.*, 2011), back-translated to nucleotides using the annotated CDS of each gene, and pairwise similarity between the genes was estimated with the compute program from the analysis package available in the libsequence evolutionary genetic analysis library (Thornton, 2003). Tandem duplicate genes that had less than 75% sequence identity or more than 50% gapped sites were considered misassigned tandem duplicates.

For correctly assigned tandem duplicates, the furthest upstream gene of each cluster was chosen to represent the syntenic relationship. For the misassigned cases, syntenic relationships were inferred based on synonymous divergence from a putative ancestral gene. For each group of false tandem duplicate genes, clustal-omega (Sievers *et al.*, 2011) was used to align the amino acid sequences of the longest transcripts and the amino acid sequence of the ancestral gene, as reported by SynMap (Lyons *et al.*, 2008). Alignments were back-translated to nucleotide sequences, and synonymous divergence between each maize gene and the ancestral gene was estimated with the yn00 program in PAML (Yang, 2007). The maize gene that showed the lowest synonymous divergence from the ancestral gene was chosen to represent the syntenic relationship.

Orthofinder (Emms and Kelly, 2015) was used to identify additional ancestral orthologs within the misassigned tandem duplicates. Representative amino acid sequences from 13 grass species, excluding maize, were used as input for Orthofinder (*Aegilops tauschii*, ASM34733v1; *Brachypodium distachyon*, 3.1; *Hordeum vulgare*, ASM32608v1; *Leersia perrieri*, Lperr_V1.4; *Oropetium thomaeum* 1.0; *Oryza sativa*, IRGSP-1.0; *Panicum hallii*, 2.0; *Panicum virgatum*, 1.1; *Phyllostachys edulis*, 1.0; *Setaria italica*, 2.2; *Sorghum bicolor* 3.1; *Triticum aestivum*, TGACv1; *Triticum Urartu*,

ASM34745v1). Amino acid sequences from B73 and PH207 were treated as originating from separate species, to allow for genotype-specific orthology inference. Orthologous relationships between maize and sorghum, and maize and rice were used to identify additional syntenic ancestral orthologs for non-representative false tandem duplicates.

Validation and recovery of syntenic assignments

The subgenome identity of each chromosome was determined using a previously described method (Schnable et al., 2011). The percent of the genome in syntenic blocks was calculated by determining the order of each maize gene on its respective chromosome, scanning for consecutive runs of genes that were within 20 genes of one another, and appending the distance between each consecutive gene. If two genes were separated by more than 20 genes within a chromosome, a new block was formed.

To remove false positive syntenic assignments, all possible pairwise BLASTN (Altschul *et al.*, 1990) alignments of CDS sequences between homeologs within B73 and PH207 and between homologs across B73 and PH207 were made. An alignment threshold of 75% identity over at least 50% of the sequence length was used to remove any false or highly diverged assignments that were present in the raw SynMap output. Some genes may fail to meet the pairwise alignment thresholds due to inconsistencies in annotation. Genes that did not meet the threshold were realigned to the genome using BLASTN with the requirement that the gene map within 1-Mb of the original gene coordinates using the same coverage and identity criteria. Genes that did not meet the alignment criteria to the genome were filtered from the working file of syntenic assignments and replaced with NA (Table S1).

To remove false negative syntenic assignments that were not assigned in the original SynMap output, BLASTN was used to find significant alignments in collinear regions. A significant hit (E-value < 1e-30 and at least 75% identity over at least 50% of the sequence length), was classified as collinear by scanning for the nearest upstream and downstream syntenic genes on the expected chromosome, extracting the coordinates for these genes, and requiring that the hit be located within the window between those two genes. A buffer of 50-kb was added on both sides of the window to allow for local rearrangements that are biological or brought about by misassembly. The mapping coordinates of the genes that aligned to the expected syntenic positions were used as input to the intersect tool implemented in the BEDTools suite v2.25.0 (Quinlan and Hall, 2010) to determine if the mapping position corresponded to an annotated gene model. If no intersect with an annotated gene model was present the coordinates of the alignment were filled in to the syntenic assignments.

In some cases, an ancestral gene in either sorghum or rice or a maize gene in either B73 or PH207 was duplicated in our working list of syntenic assignments. In a subset of these cases a gene was ambiguously assigned to different gene models due to gene models that physically overlapped one another in the genome. To remove any incorrect assignments due to this case, the CDS sequence of all maize genes associated with the duplicated gene were extracted and aligned to the ancestral sequences associated with the duplicated gene using TBLASTX (Altschul *et al.*, 1990). The true orthologous gene was chosen based on the highest alignment score. If a gene aligned significantly to two or more adjacent genes due to inconsistent annotations across genomes, the maize genes from both B73 and PH207 were excluded from subsequent analyses.

Curation of differentially fractioned genes

Differentially fractionated genes were defined as those present in one maize genome (i.e. B73 or PH207) and absent in the other based on the synteny assignments described above (Table S1). The list of differentially fractionated genes was curated to remove false positives by first aligning all syntenic genes present only in one genome to the scaffold and contig sequences in the opposite genome using BLASTN. Genes that mapped significantly (E-value < 1e-30) to these locations were filtered from the working list of differentially fractionated genes.

Resequencing reads from both B73 and PH207 were then used to determine if reads from the genotype containing the fractionated gene could be aligned to the retained gene in the genome that had the retained copy. Contaminants in reads were identified with FastQC version 0.11.5 (Babraham Bioinformatics, https://www.bioinformatics.babraham.ac.uk/projects/fastqc/). Reads were cleaned of adapter contamination with Cutadapt version 1.13 (Martin, 2011). The sequences that were targeted for removal were the Illumina universal adapter, the index-specific adapter for each library, and any contaminating sequences identified by FastQC. Reads were then trimmed of low quality bases with sickle version 1.33 (Joshi and Fass, 2011), with a minimum length of 20bp, and a minimum mean quality score of 20. When one of the read pairs failed quality control, its mate was written into a single-end read file that was aligned separately. Cleaned single-end and paired-end reads were mapped to the B73 and PH207 reference genomes using bowtie2 version 2.3.0 (Langmead and Salzberg, 2012). The seed length was set to 12 bp, to adjust the sensitivity of the mapping to account for

the average nucleotide diversity of maize. BAM alignments were cleaned of unmapped reads and sorted with SAMtools version 1.4 (Li *et al.*, 2009). Duplicate reads were removed, and read groups were added with Picard version 2.9.2 (http://www.github.com/broadinstitute/picard). Coverage was then determined by calculating coverage of exon sequence from the longest transcript using BEDTools coverage v2.25.0 (Quinlan and Hall, 2010). Features with coverage over less than 20% of the CDS sequence were interpreted to represent high confidence gene losses.

False positive differentially fractionated genes could also arise from gaps in either of the genome assemblies. To correct these false positives, the coordinates of expected flanking syntenic genes were extracted. If 40% or more of the sequence space between the flanking syntenic genes was N's, the fractionated gene was replaced with the coordinates for the flanking sequences and a flag indicating a gap (GAP:) was recorded in the synteny assignment file (Table S1). To identify smaller gap sequences, 5-kb of sequence on both sides of differentially fractionated genes was extracted and aligned to the homologous region from the opposite genotype using LAST (Kiełbasa *et al.*, 2011). The number of N's between the aligned sequences was calculated as described previously and alignments with greater than 40% gaps were flagged as a gap (Table S1).

Resequencing data from 60 diverse inbred lines (Table S3 and Table S4) was used to assess the frequency of gene deletions. These sequence data were produced by the US Department of Energy Joint Genome Institute http://www.jgi.doe.gov/ in collaboration with the user community. Reads were processed, aligned, and exon coverage was calculated as described above for the B73 and PH207 resequencing reads.

Identification of non-allelic homologs

Fractionated genes that had read coverage over greater than 20% of the exon sequence during the Curation of Differentially Fractionated Genes (see above) can result from non-allelic homologs present in non-syntenic locations in the genome. To find the prevalence of non-allelic homologs, reads from the fractionated genome that mapped uniquely to the retained gene in the opposite genome (MAPQ score > 20) were extracted from the alignment file using Sambamba v0.6.6 (Tarasov *et al.*, 2015), converted to fastq format with BEDtools bamtofastq v2.25.0 (Quinlan and Hall, 2010), and remapped back to the genotype of origin using Bowtie2 version 2.2.4 (Langmead and Salzberg, 2012) with default parameters. To determine whether the uniquely mapping reads corresponded to gene models, BEDtools intersect (v2.25.0; Quinlan and Hall, 2010) was used and coverage of each gene model was calculated using the method described above. Gene models with greater than 20% of the exon sequence covered were considered non-allelic homologs.

Functional significance of differentially fractionated genes

Enrichment of differentially fractionated genes in lists of functional genes was used to assess the functional significance of differentially fractionated genes. The list of genes that were tested included the maize classical gene set and curated gene set (http://maizegdb.org/gene_center; accessed June 7, 2017; Schnable and Freeling, 2011), network hub-genes (Walley *et al.*, 2016), and curated NAM-GWAS hits from Panzea (http://cbsusrv04.tc.cornell.edu/users/panzea/filegateway.aspx?category=GWASResults; accessed June 7, 2017; Wallace *et al.*, 2014). These gene lists were generated from

previous versions of the B73 genome assembly and gene annotation. Gene models were converted to the B73 v4 gene models using a conversion list obtained from Gramene (ftp://ftp.gramene.org/pub/gramene/CURRENT_RELEASE/data/gff3/zea_mays/gene_id_mapping_v3_to_v4/maize.v3TOv4.geneIDhistory.txt). Enrichment tests were conducted using Fisher's exact test with Yates' correction in R (R Core Team, 2014).

Transcriptional variation analysis

RNAseq reads for B73 and PH207 from blade, cortical parenchyma, germinating kernel, root, and stele tissues were downloaded from the NCBI SRA accession number PRJNA258455. Three replicates were available for B73 and two replicates were available for PH207. Adapters were trimmed from the RNAseq reads using Cutadapt version 1.8.1 (Martin, 2011) with the quality cutoff option set to 20 and the minimum length option set to 50. Cleaned reads from B73 were aligned to the B73 genome assembly and the PH207 reads were aligned to the PH207 genome assembly using Bowtie2 version 2.2.4 (Langmead and Salzberg, 2012) and TopHat2 version 2.0.13 (Kim *et al.*, 2013). Mapping parameters were set to defaults except for minimum intron size and maximum intron size, which were set to 5 bp and 60,000 bp, respectively. Transcript abundance values for longest CDS feature of each gene were calculated with HTSeq (version 0.7.2; (Anders *et al.*, 2015) using the union mode and the non-strand-specific option. Gene models between the two genomes were linked using the syntenic assignments described above (Table S1). Since the abundance values were calculated using different CDS models between the two genomes that were of variable lengths, the counts were normalized by respective CDS lengths and corrected for library size differences (Table S5). For the

binary expression analysis, a gene was considered transcriptionally inactive if it had three or fewer normalized counts averaged across replicates in all tissues. Any gene that had a count of eight reads (a log2 fold change of 1.5 compared to the cutoff for considering the gene off) in at least one tissue was considered transcriptionally active.

Table 1. Summary of retained duplicate, singleton, and total maize1 and maize2 genes in the B73 and PH207 genomes based on comparison to the ancestral state from sorghum and rice.

Gene Classification	B73	PH207
Retained Duplicates	5,741	5,713
Maize1 Singletons	9,255	9,239
Maize2 Singletons	3,777	3,789
Total maize1 genes	14,996	14,952
Total maize2 genes	9,518	9,502

Table 2. Overlap of differentially fractionated genes and functional maize gene lists.

Dataset	Reference	Overlapping Syntenic Genes	Overlapping Differentially Fractionated Genes	Chi-Square Test P-value
Curated Genes	(Schnable and Freeling, 2011; MaizeGDB.org)	3,989	16	0.0274
Hub Genes	(Walley et al., 2016)	4,280	11	0.0005
GWAS Hits	(Wallace et al., 2014)	6,537	39	0.3735

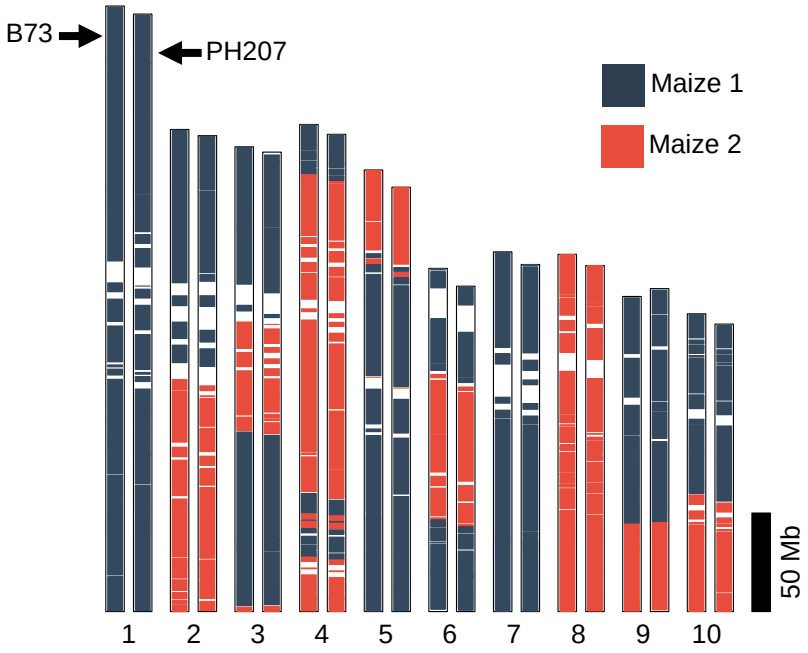


Figure 1. Maize subgenome syntenic blocks in the B73 and PH207 genomes. The 10 maize chromosomes are represented with B73 on the left and PH207 on the right. Syntenic blocks for the maize1 and maizd2 subgenomes were determined based on comparison with the ancestral state from sorghum and rice.

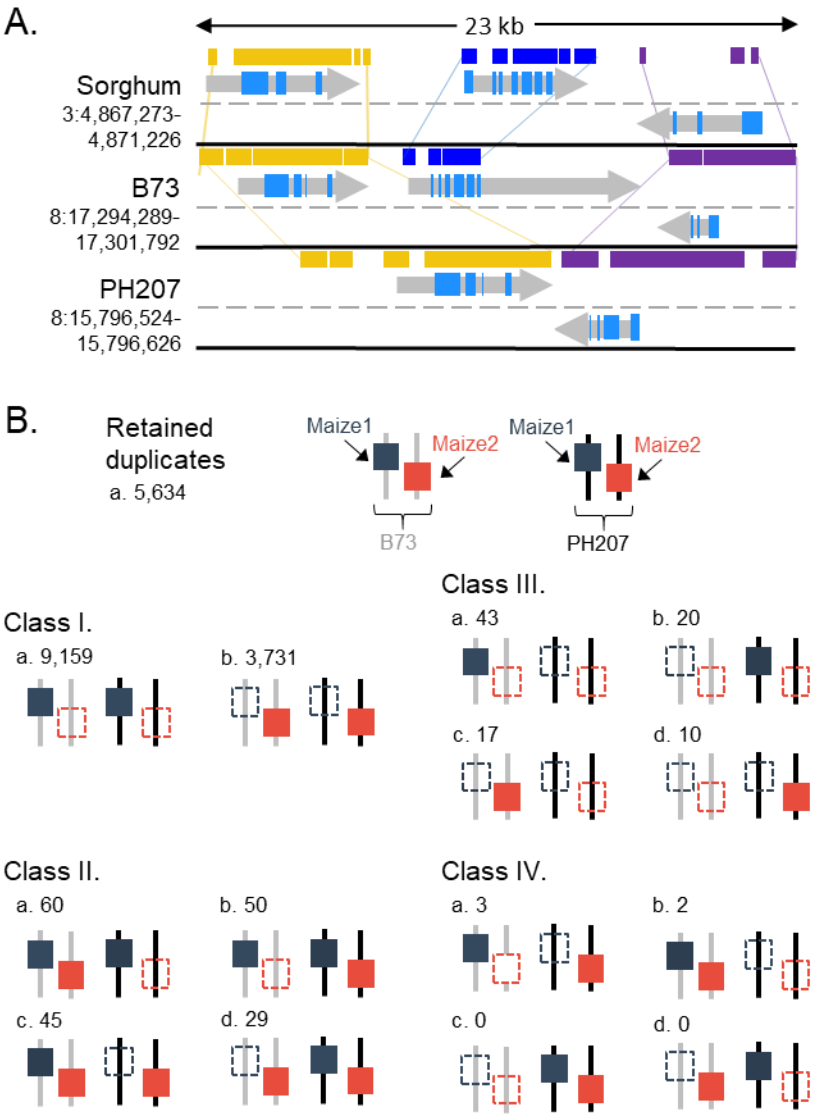


Figure 2. Differential gene fractionation scenarios. A) Example of a differentially fractionated gene. Shown are three pairs of orthologous genes in B73 and sorghum. PH207 contains only the flanking orthologous genes but is missing the homologous B73 gene, Zm00001d008692, and the sorghum gene, Sobic.003G053700. Colored blocks represent high-scoring pairs (HPSs) from alignment between each orthologous gene. B) Differential fractionation events are grouped according to the fractionation outcome observed between B73 (grey, left) and PH207 (black, right).

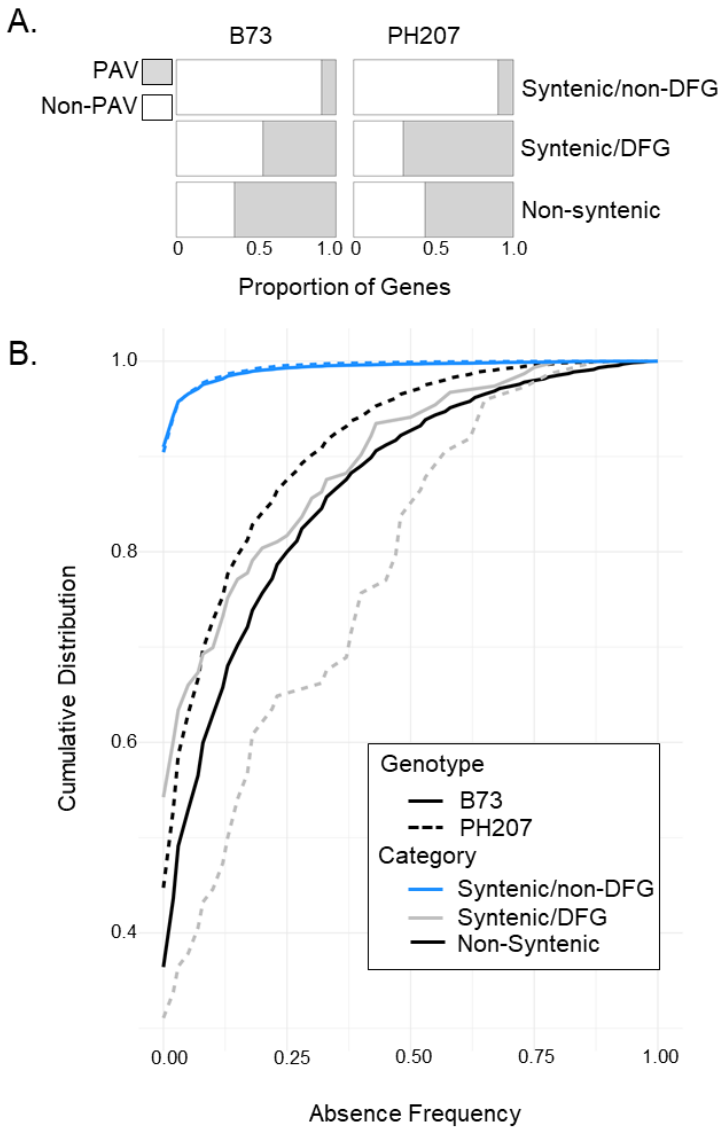


Figure 3. Presence/absence variation (PAV) frequency distribution of differentially fractionated genes (DFGs). A) Proportion of genes that show PAV among 60 diverse lines. B) Cumulative distribution of PAV frequencies across a panel of 60 diverse inbred lines. A PAV frequency of zero indicates that a gene is not variable across any of the resequenced inbreds, while a PAV frequency of one indicates that a gene is private to B73 and/or PH207, and not contained in any of the other maize lines. Criteria for absence or substantial deletion was resequencing coverage across less than 20% of representative CDS sequence. Differentially fractionated genes are defined based on comparison of the genomes of B73 and PH207 to the ancestral states based on sorghum and rice. The distinction of B73 and PH207 in both plots is based on which genome the resequencing reads were mapped to and the subset of differentially fractionated genes that are private to either genome.

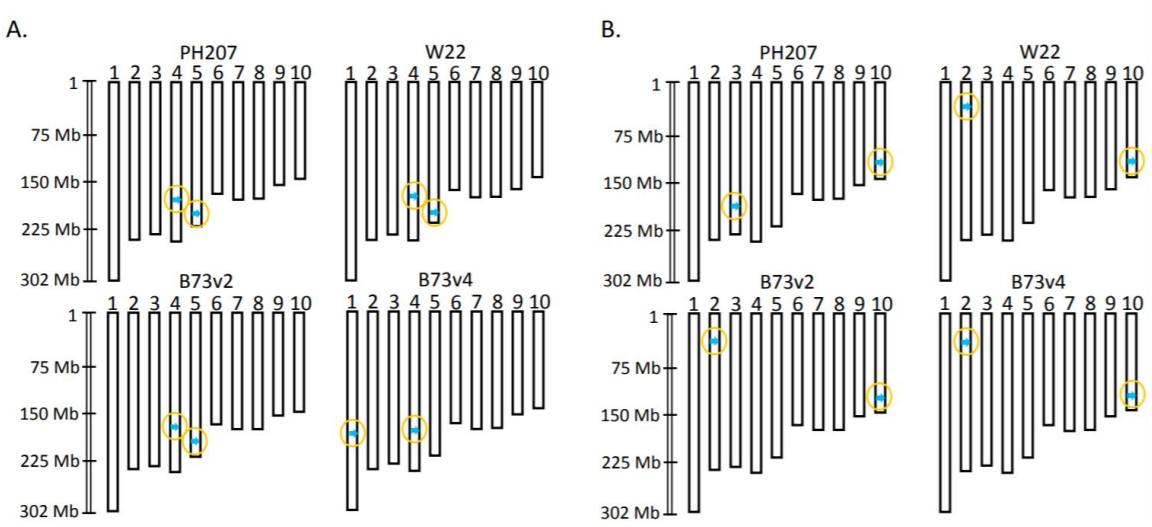


Figure 4. Examples of potential buffering loci from non-allelic homologs. A) The B73 gene Zm00001d031670 is present on chromosome 1 in the B74v4 assembly (blue arrow), and present on chromosome 5 in all other assemblies (PH207, W22, CML247, F7, Ep1). The homeolog, Zm00001d052213, is present on chromosome 4 in all assemblies. B) The PH207 gene Zm00008a013424 is present on chromosome 3 in the PH207 assembly but present on chromosome 2 in all other assemblies. The homeolog of this gene, Zm00008a038016, is present on chromosome 10 in all assemblies.

Chapter 3. The influence of TEs to variation in maize (Zea mays L.) gene expression.[3]

Genome-wide experiments often focus on low-copy and genic portions of the genome, however in a species like maize this means ignoring more than 70% of the genome that is comprised of transposable elements (TEs). The relative proportion among TE families that make up this sequence space are largely consistent between genotypes, yet the presence of individual insertions is remarkably variable. While most of these insertions are unlikely to lead to phenotypic differences, multiple studies have indicated that TEs play an important role in shaping genome evolution and transcriptional regulation. In this study, we build on these findings by exploring the potential influence of TEs to gene expression variation using direct genome-wide comparisons for the maize inbred lines B73 and W22. Using an RNA-seq dataset representative of the maize transcriptome throughout development we show that there are higher numbers of differentially expressed genes associated with nonshared TEs between the genomes. Despite this observation, we did not find a consistent pattern to support the hypothesis that nonshared TEs between these individuals are responsible for driving differential expression. This result was consistent across an expanded panel of 20 inbred lines that showed high rates of TE insertional polymorphisms but lacked a clear panel association

[3] This work was a collaboration between ABB, PJM, SNA, NMS, SEM, and CNH. ABB, and CNH conceptualized the manuscript. The manuscript was written by ABB. SNA provided the data on B73 and W22 TEs. PEM provided alignment files from the diversity panel. All other data was analyzed by ABB.

Brohammer AB, Monnahan PJ, Anderson SN, Springer NM, McGaugh SE, Hirsch CN. 2018. The influence of TEs to variation in maize (Zea mays L.) gene expression. *In preparation.*

to DE. We also show that allele bias is not a prominent feature between B73 and W22 genes with proximal nonshared TEs. Together these data suggest that a nuanced understanding of the role of TEs in altering transcriptional regulation is required and these effects are less likely to be observed in the absence of specific environmental conditions.

Introduction

Since the time Barbara McClintock first discovered transposable elements (TEs) in maize (McClintock, 1948; McClintock, 1950), there has been interest in identifying the influence of TEs in shaping genome evolution. The presence of transposable elements has been linked to plant phenotypes (*Vgt1,* Salvi *et al.*, 2007; *Knotted1,* Greene *et al.*, 1994; *waxy,* Wesler and Varagona, 1985) and helped refine our understanding of the role of TEs in genomes as something closer to the 'controlling-elements' paradigm that McClintock first envisioned than to the framework that emphasizes TEs as ineffectual 'junk'. This view is supported by the numerous potential routes for TEs to generate genomic novelty including influencing rearrangement through ectopic recombination, bringing about changes in chromatin, or providing novel regulatory sequence. The ability to easily visualize kernel pigment aberrations was key to McClintock's discovery of TEs in maize, but numerous other characteristics have also made maize central to the study of TEs.

One of the characteristics that makes maize an ideal species in which to study TEs is the high content of TEs in the genome compared to genic, and intergenic sequence. Structural annotation of the B73v4 genome assembly indicates that 1,352 Mb out of the

2,114 Mb of assembly space or about 64% of the genome is comprised of intact TEs (Jiao *et al.,* 2017). The large proportion of TEs in maize is primarily due to a large expansion of LTR elements, which represent class I TEs that transpose through an RNA intermediate and comprise about 60% of the maize genome. Class II TEs, which transpose through a DNA rather than an RNA intermediate and often contain terminal inverted repeats (TIRs), are much less abundant than class I elements and constitute less than one percent of the genome partially due to their average length of 700 bp. Both classes of TEs can be divided into orders and even further subdivided into superfamily and family-based classifications (Wicker *et al.,* 2007), which highlights the degree to which individual elements within each class can differ in repeat structure and key characteristics like insertion-site preference (Bureau and Wessler, 1992; Bureau *et al.*, 1996; SanMiguel *et al.*, 1996; Baucom *et al.*, 2009; Wei *et al.*, 2009; Tenaillon *et al.*, 2011). *Helitrons*, for example, are a type of class II element but lack the canonical TIR sequences and transpose through a rolling-circle replication mechanism rather than the traditional "cut-and-paste" mechanism. These elements comprise 3.6% of the genome and are notable for their ability to capture gene fragments.

Given that most of the maize genome consists of TEs, it is not surprising that many elements have inserted near or within genes. Based on the B73 assembly, more than 70% of genes have a TE within 5 kb of the gene (Jiao *et al.,* 2017). Apart from TE insertions into coding sequence that are often highly deleterious, the maize genome has developed ways to insulate genes from the effects of TEs near and within genic sequence (Gent *et al.*, 2013; Li *et al.,* 2015). As such, the majority of intergenic or intronic insertions are unlikely to have an impact on the function of nearby genes. However,

numerous reports have shown that in certain conditions TEs can influence gene expression levels or patterns (Reviewed in Hirsch and Springer, 2017). This influence can occur at both the transcriptional and post-transcriptional levels and be created through several mechanisms including modifications of mRNA processing (Varagona *et al.*, 1992), influencing chromatin state (Hollister and Gaut, 2009; Eichten *et al.*, 2011), providing *cis*- or *trans*- regulatory features (Barkan and Martienssen, 1991; Sundaram *et al.*, 2014; Makarevitch *et al.*, 2015), or insertional mutagenesis (B. Greene *et al.*, 1994; Schneeberger *et al.*, 1995; Muehlbauer *et al.*, 1999).

The number of transcripts that TEs contribute to the global transcriptome is a small proportion compared to the contribution of genes, yet there is a notable signal of TE expression across diverse spatiotemporal conditions. One recent estimate attributes approximately 5% of the transcriptome to TEs and reports that this expression is dynamic with many TE families displaying tissue and developmental specific expression patterns (Anderson, *et al.*, 2018). In the case of the maize shoot apical meristem, nearly 10% of all transcripts are derived from retrotransposons (Ohtsu *et al.*, 2007). Many TEs have their own regulatory modules that lead to their activation under certain environmental conditions, especially those associated with abiotic or biotic stress (Naito, *et al.*, 2009; Mhiri *et al.*, 1997; Ivashuta *et al.*, 2002; Ito *et al.*, 2011; Grandbastien *et al.*, 2005; Buchmann *et al.*, 2009; De Felice *et al.*, 2009). In rice, for example, class II type *mPing* elements tend to be up-regulated in response to cold and salt stress and are associated with the coordinated response of several genes to these conditions (Naito, Zhang, Tsukiyama, Saito, C Nathan Hancock, *et al.*, 2009; Yasuda *et al.*, 2013).

In addition to their own expression, TEs can also influence gene regulation by providing novel *cis*- regulatory variation. It is known that many promoter sequences contain motifs that are associated with TEs (White *et al.*, 1994), with one study in the human literature finding that 25% of identified promoter sequences contained evidence of TE associated sequence (Jordan *et al.*, 2003). A comprehensive mapping of transcriptional enhancer candidates in maize found that ~30% of tissue specific enhancers overlapped with TEs and identified three LTR Gypsy families that were enriched for putative enhancers (Oka *et al.*, 2017). Also in maize, Makarevitch and colleagues reported that several TE families are associated with stress-responsive expression of proximate genes likely as a result of the TEs acting as local enhancers (Makarevitch *et al.*, 2015). These observations fit with the idea that TE regulatory sequences can act as cryptic promoters for nearby genes and even shape the gene regulatory landscape after exaptation (Feschotte, 2008).

Here, we combine dense resequencing data across a set of diverse maize inbred lines with an RNA-seq dataset representing the maize transcriptome across nine tissues throughout development to better understand the impacts of variable TE content on differential gene expression that is observed between maize inbred lines under standard growth conditions. These data indicate that genes with nearby TEs that were nonshared between haplotypes were more likely to exhibit variable expression levels than genes with only shared insertions. Despite this observation, our results do not indicate a clear pattern for the association of gene proximal TE insertions to expression variability in the absence of conditions of stress.

Results

Gene proximal TEs have a high rate of presence/absence variation among maize inbred lines

The goal of this study was to determine the impact that TEs have on expression of nearby genes. There are many different metrics that can be used to define the regulatory region for a gene, and regulatory regions can extend quite far from genes (Salvi *et al.*, 2007; Castelletti *et al.*, 2014), particularly in species such as maize where the distance between genes is substantially further than in smaller genomes such as Arabidopsis (Keller and Feuillet, 2000). For this study, we used the distribution of TE density to set the boundary to define the criteria for considering a TE as gene proximal. This was accomplished by calculating the density of TEs in 1 kb windows surrounding the midpoint of a gene (Supplemental Figure 1A). For windows that were within 5 kb of a gene the density of TEs was quite low and a dramatic increase in TE density was observed in windows of increasing distance from genes. Thus, for this study we define gene proximal TEs as those that are within 5 kb of gene coding sequence.

To assess the impact of TEs on gene expression variation we focused on genes that have proximal TEs that are polymorphic between maize inbred lines. The presence or absence of a TE insertion was called using criteria defined by Anderson *et al.* (In preparation). In brief, this method consisted of identifying stretches of collinear genes between pairwise comparisons of genomes (Supplemental Figure 1B) and using these genes as anchor points in which to define windows where flanking sequences of TEs are expected to align (see Methods). The TE annotation was filtered to include only TIR and LTR elements in the “DT” and “RL” orders, respectively. The list of TEs considered was

further filtered to include only those elements that could be confidently called shared or nonshared. A summary of the filtered and unfiltered TE elements is included in Table 1 but all downstream analyses only consider this filtered set of TEs.

The set of genes with at least one proximal TE before filtering included ~86% (33,882) of all annotated B73 gene models (see Supplemental Table 1 for summary). Of all the genes with a proximal filtered TE 54% (15,024/27,550) of those genes contained at least one nonshared proximal TE, while 10,130 contained a single nonshared TE and 3,772 contained a single TE that was nonshared. Beta versions of structural TE annotations for W22 and PH207 were used to assess the consistency of the results obtained using B73. For each genome comparison and for each distance category there were consistent numbers of shared and nonshared TEs with a range of ~38% percent of TEs within genes being nonshared compared to ~46% of the TEs 5-kb downstream of a gene being nonshared (Supplemental Table 1 and Supplemental Figure 1C).

Genes with nonshared proximal TE insertions are associated with elevated rates of differential expression

To determine the effect of nonshared TEs on proximal gene expression an RNA-seq dataset consisting of nine spatiotemporally diverse tissues was used that included primary root 6 days-after-planting (DAP; R), shoot and coleoptile 6 DAP (SC), internode at vegetative 11 (V11; I), middle of the 10^{th} leaf (V11; L10), leaf 30 DAP (L), immature ear at V18 (IE), anthers at reproductive 1 (R1; A), and endosperm 16 DAP (E). These tissues were chosen to broadly capture the dynamic maize transcriptome based on the maize B73 gene atlas (Stelpflug *et al.*, 2016) and were sampled contemporaneously from

B73 and W22. Differential gene expression (DE; $p_{adj} < 0.1$ and minimum 1-fold log_2 change) was determined by mapping processed reads to the B73 AGPv4 assembly. On average, 4,149 of 39,498 total B73 gene models were DE in each tissue, with a minimum of 2,686 DE genes in shoot and coleoptile and a max of 6,044 DE genes in immature ear (Supplemental Figure 2).

In seeking to determine the impact of TEs on gene expression variation we first looked at the comparison of genes that do or do not have a proximal TE. The proportion of genes with a proximal TE that were DE was nearly identical to the proportion of genes without proximal TEs (Figure 1A) in separate calculations including all genes and only expressed genes, suggesting that the presence or absence of proximal TEs alone does not affect DE (Chi-square test with Yates correction p ~0.7). However, when considering the association of DE with the presence or absence of proximal nonshared TEs the proportion of genes that have at least one nonshared TE and are DE is consistently elevated compared to the proportion of genes that do not have a proximal nonshared TE and are DE (Figure 1B; Chi-square test with Yates correction $p < 0.01$ in all tissues). Across tissues, genes with at least one nonshared TE were slightly enriched for DE with 12% of these genes being DE compared to 7% of the genes with only shared proximal TEs between B73 and W22. This trend is consistent across tissues but most divergent in 10^{th} leaf (nearly 5% difference) and holds for genes with multiple proximal TEs or only a single proximal TE (Supplemental Figure 3).

There did not appear to be a significant trend associated with the different categories of proximal TEs based on relative proximity to the gene (i.e. within gene, 1-kb upstream, 1-kb downstream, etc.), with every category having nearly identical

proportions of DE genes (Supplemental Figure 4). The lowest and highest proportions were observed for genes with no proximal TE and genes that contained a TE within 5-kb downstream although the proportion were nearly identical, 10.2% and 10.7%, respectively. There was also minimal variation after isolating genes with only a single proximal TE that was nonshared and recalculating the relative enrichment/de-enrichment of the proximal categories (Supplemental Table 2) suggesting that background sequence variation may be responsible for the elevated rate of DE observed.

Differentially expressed genes with nearby nonshared TEs are not associated with allele bias

If nonshared TEs play a role in differential gene expression, it is hypothesized that the effect will result in allelic expression bias of the nearby gene. With this reasoning, expression levels in B73 and W22 were compared to determine whether the haplotype with the TE insertion (B73 allele) resulted in lowered or elevated expression compared to the haplotype without the insertion (W22 allele). For both DE and non-DE genes that contained nonshared TEs, expression levels for B73 and W22 were compared and are plotted in Figure 1C. The number of genes with a single TE that is nonshared and are DE that favor the B73 allele versus the W22 allele is not statistically significant (Chi-square test with Yates correction p-value 0.89). This result is consistent across both TE classes with 53% and 51% of DE genes with a single nonshared TIR and LTR favoring the B73 over the W22 allele, respectively.

This potential bias was also examined by identifying genes in which there was consistent bias towards one allele in every tissue in which the gene was expressed. For

genes with only a single proximal TE in which the insertion was nonshared, 56% were expressed but not consistent in the allele with higher expression, 17% were not expressed, 15% consistently favored the haplotype with the TE insertion (B73 allele), and 11% consistently favored the haplotype without the TE insertion (W22 allele). Despite the slight increase towards the B73 allele, these proportions were nearly identical to the proportions calculated using genes that had a single proximal TE that was shared (Supplemental Table 3)

TE effects across a maize diversity panel

To further study the association of nonshared TEs with gene expression variation we examined a larger set of genotypes that represent a subset of lines from the Wisconsin Diversity Panel (WiDiv; Hansey *et al.*, 2011), a set of maize inbred lines that reliably mature in an upper-Midwest growing environment. Using resequencing reads, each B73 TE was scored for presence or absence across 51 of these inbred lines. A comparison of the TE calls obtained from resequencing reads to the TE calls obtained from the method used to compare whole genome assemblies for B73 and W22 resulted over 95% concordance. Based on the resequencing reads about 37% of all the tested insertions were found to be present in at least 95% of the lines (Figure 2A). Still, a significant proportion of the TEs that are annotated in the B73 reference genome assembly were not detected in a majority of the resequenced individuals. The median number lines in which the TE is absent in the diversity panel was 36 for nonshared DE genes and 26 for shared DE genes.

For 21 of the 51 lines (including B73) RNA-seq data was available for seedling tissue and was used to explore the relationship between TE presence and proximal gene

expression on a broader set of germplasm. The expression level of each gene in the resequenced WiDiv lines was compared to the expression level in B73. Genes that had a $\log_2$ fold change of 1.5 or greater compared to the B73 expression level were considered DE and any insertion that was not positively identified in B73 was excluded. The total list of genes was then filtered to include only genes with a single proximate TE to reduce any possible noise in the dataset. After these filters were applied, 10,371 genes remained for downstream analyses.

To investigate the relationship between DE and the presence of a proximal TE, the rate of DE was examined in relation to the frequency in which a TE was present in the population. Figure 2B-C. show that higher levels of DE were observed for TEs that were specific to only a few lines compared to TEs that were present in the majority of lines with about twice as many genes being DE from one terminus to one another (14.3% DE compared to 7.7% DE). These results were consistent after filtering the list of TEs to only contain the filtered list of class I and class II elements used in the B73 versus W22 analyses. After separating the list of 1,279 TIRs from the 5,693 LTRs there appeared to be slightly higher rates of DE among TIRs that were variable at moderate to high levels compared to LTRs (Supplemental Figure 6). The average proportion of DE genes across all TE frequencies was nearly identical however with 9.8% and 9.6% percent of genes of DE for TIR and LTR elements, respectively.

Few TE families are enriched for characteristics indicative of an influence with gene expression

A confounding factor with the finding that nonshared proximal TEs are enriched among DE genes is background sequence divergence. The perceived effect of a TE

insertion is difficult to isolate from the effect of other polymorphisms in the same haplotype. For this reason, potential influences on gene expression were also studied at the TE family level. This analysis focused on identifying TE families that exhibited multiple characteristics that might suggest an ability to influence *cis*- regulatory variation including consistency in allele direction, rate of presence/absence variability, and rate of proximal insertions. For each category, the proportion of insertions for each TE family was compared against the proportion of insertions among the broader TE superfamily classification. The set of TE families with values greater than or less than that observed across the cognate TE superfamily for a response were identified and plotted (Figure 3A-C).

1. Proximal insertions

First, each TE family was characterized by the proportion of total insertions that were gene proximal versus non-gene proximal with the reasoning that TE families with the ability to introduce novel *cis*- regulatory variation may exhibit a preference for insertion near coding sequence (Figure 3A). The opposite, TE insertions near coding sequence that are rare, may also be significant thus families with a depletion of gene proximal elements were also considered. The proportion of TE superfamilies with proximal insertions ranges from 64% to less than 20%. The range for TE families was even broader with one superfamily having only a single proximal TE out of 61 insertions (RLG00077) compared to another family (DTA00310) in which every insertion is gene proximal after filtering for at least 40 and 20 insertions, respectively. Considering enrichment/de-enrichment in proximal TEs, the most enriched family LTR is

"RLX00094", for which there is 64% difference between the family percentage and the superfamily percentage, while the most de-enriched family, "DTT10864", has a 46% difference in the opposite direction.

2. Nonshared insertions

A TE family that alters *cis*- regulatory variation of nearby genes may be likely to lead a fitness impact. In the case of an average positive fitness effect it might be expected that insertions be on average more conserved compared to other TE insertions or the opposite if the effect was deleterious. Thus, the dataset was filtered for TE families that were had large differences for the number of shared insertions compared to the distribution of values from the superfamily classification (Figure 3B). The most elevated family had an elevated proportion of nonshared TEs of 32%, while the most depleted had a difference of 38% from its superfamily value. By comparison the median difference of elevated and depleted families was 12% and 9%, respectively.

3. Allele bias

A TE family that influences proximal gene expression is expected to result in a consistent pattern of proximal gene expression being elevated or lowered compared to a haplotype without the presence of the insertion as described above. The TE families shown in Figure 3C were identified after filtering the dataset for TE families that had at least 10 nonshared insertions and displayed a consistent expression pattern in favor of one allele over another. Only nonshared insertions were considered in the calculation. The TE family with the highest proportion of insertions associated with a consistent expression pattern is the TIR family, "C00119". Eleven of the 32 insertions of this family

favored the haplotype with the TE insertion (B73 allele) everywhere where the proximal gene was expressed compared to 32/112 from its respective superfamily, a proportional difference of 34%. Of the top twenty families that were enriched for one allele over another, 11 were consistently elevated towards the haplotype with the TE insertion (B73 allele) and nine towards the haplotype without the TE insertion (W22 allele).

After labelling each TE family according to characteristics described above, we looked to see if any of the families identified as having extreme values in one category were identified in other categories as well (Figure 4A) There appeared to be little overlap between the TE families identified across each of the three categories. Given the randomness associated with mutation, this result is not entirely unexpected. Using the twenty most divergent categories to label TE insertions we also re-analyzed the association of individual elements with DE compared to insertions from families that did not meet the extreme criteria (Figure 4B). There did not appear to be any difference in the proportion of DE genes associated with these families providing little evidence for the existence of families enriched for altering expression variation under standard growth conditions. This result is in contrast to previous work that has shown enrichment of TE families altering expression of nearby genes (Makarevitch *et al.*, 2015).

Discussion

Given that most of the maize genome is made up of repetitive sequence and TEs have their own regulatory activities, it is intriguing to consider TEs as a genome-wide source of regulatory elements. There are numerous studies that provide support for a regulatory role of TEs using multiple layers of biological data. For example, it has long

been known some TEs become activated in response to stress (Casacuberta and Gonzalez, 2013) and RNA-seq based studies of gene expression have shown that this TE activation can also lead to altered expression of nearby genes in conditions of plant stress (Makarevitch *et al.*, 2015). There is also a growing body of evidence for the role of TEs in providing novel *cis*- regulatory variation from studying methylation patterns chromatin accessibility and DNA binding (Hirsch and Springer, 2017). Genome-wide discovery of transcriptional enhancer candidates has also pointed to the importance of TE sequence (Oka *et al.*, 2017). Most importantly studies have linked individual insertions to important phenotypes such as *tb1* (Studer *et al.*, 2011), *Vgt1* (Castelletti *et al.*, 2014), and *ZmCCT* (Yang *et al.*, 2013).

A concomitant research interest has centered around understanding structural variation and dispensable gene content among maize haplotypes. This is especially true in the case of TEs, which display exceptional allelic diversity as shown by early genomics projects that compared such haplotypes. One such study sequenced 151 kb of common sequence surrounding the *bronze1* (*bz1*) locus in a comparison of maize inbred lines B73 and McC and found that only 30% of the sequences were shared in the region. (Fu and Dooner, 2002). Additional studies have documented novel TE content at the *a1-sh2* interval (Yao *et al.*, 2002), the *z1C* locus (Song and Messing, 2003), and several regions in a comparison of inbred lines B73 and Mo17 (Brunner *et al.*, 2005). A more recent comparison examined variation in TE content across the entire genomes of five maize inbred lines for which genome assemblies were available and estimated that over 500 Mb of TE sequence was nonshared in at least one of the pair-wise comparisons (Anderson, *et al.*, In preparation). With regards to B73 versus W22, this study found that over 10 Mb of

TE sequence is nonshared between the two genomes. We hypothesized that some of these nonshared TE sequences could lead to novel regulatory variation across the haplotypes.

The primary motivation for this study was to test the hypothesis that nonshared TE sequences affect proximal gene expression in standard growth conditions and thus contribute novel regulatory variation. Analysis of nonshared TEs between B73 and W22 failed to support this hypothesis and did not find strong evidence for the association between nonshared TEs and variation in gene expression levels. In one-way this result is not unexpected, given that deleterious insertions that disrupt expression of key genes or regulatory sequences are likely to be quickly purged or silenced especially among elite breeding lines that are subject to intense selection pressure (Slotkin and Martienssen, 2007). In a genome that is majority comprised of TEs, it is critical that the host genome develops mechanisms that insulate functional DNA sequence from the effects of TE insertions to maintain stability (Bennetzen, 2000). Despite these mechanisms to tolerate the deleterious effects of TE insertions, it is clear that TEs have a major role in shaping genome evolution and providing a substrate for new *cis*- regulatory features in the form of motifs and binding sites (Springer *et al.*, 2016). There appears a clear discrepancy in this appreciation for the role of TEs as a larger force in the evolution of a novel regulatory sequence and the lack of clear evidence presented for the ability of nonshared TEs to broadly influence expression under standard growth conditions. How do we reconcile this understanding with the results presented herein?

One explanation is that the influence of TEs on gene expression is constrained to certain environmental conditions not captured in this experiment. Many of the previous experiments that have demonstrated an association between TE families and expression

changes have observed these effects after subjecting plants to various stresses or environmental changes (Naito, *et al.*, 2009; Yasuda *et al.*, 2013; Ito *et al.*, 2016). The majority of the TEs that have been shown to influence gene expression in response to stress are themselves upregulated during these conditions, thus the proper environmental conditions may not have been met in the experimental setup. Another reason for the lack of a positive response may be related to focusing on a single pairwise comparison. The TEs capable of influencing expression may be present in the population at high frequencies and significant variation may simply not be captured by a pairwise comparison of two genotypes. A third explanation is that the current study focused on intact and annotated elements. Experiments that have associated regulatory elements with TEs highlight specific sequence motifs left by decayed TEs (Zhao et al., 2018). These more ancient TE derived sequences are also more likely to be fixed in the genome and would not be captured through a method focused on nonshared elements.

In this study, we presented data that shows nonshared TEs between two maize genomes do not significantly contribute to transcriptional variation in the form of differential expression. It is important to note that these results do not rule out the potential for TEs to play important roles in contributing to regulatory variation. Rather, this work suggests that a nuanced understanding of the role of TE in transcriptional regulation is required. Future experiments might take advantage of the new genomic resources available to revisit previous experiments that have focused on stress conditions or incorporate data on chromatin to develop a more complete picture of the role of TEs in transcriptional regulation.

Methods

Plant material sampling and sequencing

Tissues sampled for this experiment included primary root 6 days after planting, shoot and coleoptile 6 day after planting, internode, middle of the 10th leaf, leaf at 30 days after planting, immature ear at the V11 stage, anthers at the R1 stage, endosperm 16 days after pollination, and embryo 16 days after pollination. Sampling of each tissue was done as previously reported (Sekhon et al., 2011; Stelpflug et al., 2016). For every sample three plants were pooled per biological replicate. Seedling samples were harvested from greenhouse grown plants. Greenhouse conditions were 27°C/24°C day/night and 16 h light/8 h dark, and seeds imbibed in water for 24 hours were planted in Metro-Mix 300 (Sun Gro Horticulture) with no additional fertilizer. The remaining seven tissues were harvested from field grown plants grown at the Minnesota Agricultural Experiment Station, St. Paul, MN, in Summer 2015 under standard agricultural management practices.

RNA from two biological replicates for each of the nine tissues from B73 and W22 were extracted using the Qiagen RNeasy Plant Mini Kit (Qiagen, Hilden, Germany). Sequence libraries were prepared by the University of Minnesota BioMedical Genomics Center following the standard TruSeq library preparation protocol (Illumina, San Diego, CA). Samples were sequenced on a single run of a HiSeq 2000 as 50-nucleotide single-end reads on an Illumina HiSeq 200 (Illumina, San Diego, CA) at the University of Minnesota BioMedical Genomics Center.

RNAseq quality control and processing

RNA-seq quality was first inspected by evaluating the per base sequence quality of reads and other quality control metrics available in the FastQC tool (version 0.11.5; Babraham Bioinformatics, https://www.bioinformatics.babraham.ac.uk/projects/fastqc/). RNA-seq reads were pre-processed using CutAdapt (1.16, Martin, 2011) requiring a minimum length of 25 bp and low-quality ends were trimmed using the '--quality-cutoff' option with a value of 20. Reads were aligned to the B73 AGPv4 reference assembly using STAR (2.5.3a, Dobin *et al.*, 2013). The genome indices for the STAR alignments were created by providing a GTF annotation file and setting the '—sjdbOverhang' option to the read length -1 (50-1=49). The number of reads mapping to each CDS feature were calculated using featureCounts (Liao *et al.*, 2014) and grouped by gene model. To test the biological integrity of the samples, a principle component analysis was run using the abundance counts for each feature (Supplemental Figure 5). This analysis showed distinct clusters of tissues with both reps from a genotype together indicating the data was of high quality. Differential expression analysis was conducted using DESeq2 (1.18.1, Love *et al.*, 2014). Samples were normalized using the median-of-ratios approach provided within DESeq2, FDR < 0.05 and minimum 1 $\log_2$ fold change. Default settings were used for all remaining options.

Identification of nonshared TEs

Nonshared TEs between the *de novo* genome assemblies were identified using the methods described in full in Anderson et al. (In preparation). A working version of the TE annotation prior to publication by Anderson et al., was made available for this study. This working version of the TE annotation has been made available on the GitHub

repository associated with this study (https://github.umn.edu/broha006/te-expression; B73v4_structural_filtered_Feb92017.gff3). In brief, to determine whether a TE was shared or nonshared between B73 and W22 sequences flanking the 5’ and 3’ ends of each TE in the query genome were mapped to the comparand genome using BWA MEM algorithm (version 0.7.17) with default parameters. Only reads with mapping score above 30 were retained and flanking sequences were required to map in a syntenic location defined by anchor coordinates in each genome. Anchor coordinates were defined by homologous genes in collinear arrangement and shared between genomes (Supplemental Figure 1B). Homologous genes were identified as previously described (Brohammer et al., 2018). Any TE flank sequences that failed to map were further inspected in a pairwise sequence alignment of the sequence between anchor features using LASTZ (version 1.03.02) with parameters "--gfextend", "--chain", "--ambiguous=n", "--gap=450,30", "--gapped", "--markend", "--hspthresh=3500", "--gappedthresh=3500", "--filter=identity:80". The insertions that were classified as having only partial homology, usually a result of a significant deletion, were removed from downstream analyses. The remaining set of TE insertions were primarily LTR and TIR elements with high confidence shared or nonshared calls between genomes.

Classification of TEs as gene proximal TEs

The distance of TEs from genes that was considered proximal was chosen based on visualizing the density of TE insertions in 1-kb windows away from the transcription start site of genes (Supplemental Figure 1A). There was a sharp decrease in TE density starting near 5-kb away genic sequence and this observation was used to defined the

boundary of gene proximal TEs for this study. Due to inconsistencies in the inclusion of 5' and 3' UTR elements across the B73 and W22 annotations, a standard buffer distance of 450 base pairs was applied to each gene model extending from the boundaries of the CDS sequence. The distance of 450 base pairs was chosen based on the average length of UTR features in the B73 v4 annotation. TEs were initially classified into six distance categories based on their proximity to the gene. These categories included encompassing (gene is entirely within TE coordinates), within gene (TE is entirely with gene coordinates), 1-kb upstream or downstream, and 5-kb upstream or downstream. The majority of the TEs that made up the 'encompassing' category were Helitron elements. The nature of the replication mechanism used by these elements makes their annotation more difficult leading to less confidence in their annotation this category, as such this category was not considered in downstream analyses. The downstream analyses thus focused on LTR and TIR elements in the "RL" and "DT" order of TEs.

Genomic and transcriptomic analysis of a diverse panel of lines

RNA-seq data for lines in the Wisconsin Diversity Set were obtained from the Sequence Read Archive using the accession number 'PRJNA189400' (Hirsch et al., 2014). RNA-seq quality was examined using the FastQC tool (version 0.11.5; Babraham Bioinformatics, https://www.bioinformatics.babraham.ac.uk/projects/fastqc/)). RNA-seq reads were pre-processed using CutAdapt (1.16, Martin, 2011) requiring a minimum length of 20 bp and low-quality ends were trimmed using the '--quality-cutoff' to remove reads with PHRED score below 20. Reads were aligned to the B73 AGPv4 reference assembly using STAR (2.5.3a, Dobin *et al.*, 2013). The number of reads mapping to each

CDS feature were calculated using featureCounts (Liao *et al.*, 2014) and aggregating to gene level counts. Read counts were normalized using DESeq2 (1.18.1, Love *et al.*, 2014). Normalized reads were averaged across reps for each genotyping and compared to the count obtained from mapping the B73 reads. Any samples with a 1.5-fold change compared to the B73 expression level were considered DE.

Moderate depth (~20x), short-read resequencing data on a subset of these lines was generated using Illumina Hiseq from seedling tissue. Reads were mapped to the B73 AGPv4 reference genome using Bowtie2 (version 2.2.4) using default options and mapped reads were filtered to retain the uniquely mapped portion (MAPQ 20). The processed reads were filtered to retain only those reads that mapped to the flanking ends of annotated B73 TE sequences. TE insertions based on the B73 structural TE annotation that had at least three mapped reads flanking at least 1-bp of both TE flanks were called shared and others were considered nonshared. After scoring the presence of TEs, results were compiled in a matrix and the results were filtered to retain only those TEs in which the resequencing reads from B73 itself were called shared.

Analysis scripts are available through GitHub at the following link: https://github.umn.edu/broha006/te-expression

Table 1. Number of genes associated with proximal TEs and a breakdown of B73 proximal TE insertions by distance to proximal gene. The unfiltered category includes all elements in the original annotation except for TE elements that entirely encompass genes, which are not included. The filtered category includes only elements in the "RL" and "DT" order. The TE distance categories within the TE summary include redundant elements if they are proximal to multiple genes prior to the filtering criteria, thus a TE may be counted more than once if it is proximal to more than gene, however a TE is only represented by the most proximal category for any one gene (i.e. if counted as 1-kb upstream it will not be included in the tally for 5-kb upstream).

Gene count summary		
Genes with no proximal TEs	5,616	
Genes with singe proximal TE	12,652	
Genes with multiple proximal TEs	21,230	
Total genes	39,498	
TE count summary	**Unfiltered**	**Filtered**
Within gene TE elements	4,930	3,876
1-kb upstream TE elements	9,178	5,897
1-kb downstream TE elements	8,233	5,202
5-kb upstream TE elements	20,143	15,802
5-kb downstream TE elements	18,624	14,508
Non-redundant proximal TE elements	48,174	36,269
Non-redundant non-proximal TE elements	125,686	120,289
Total TE elements	171,427	156,558

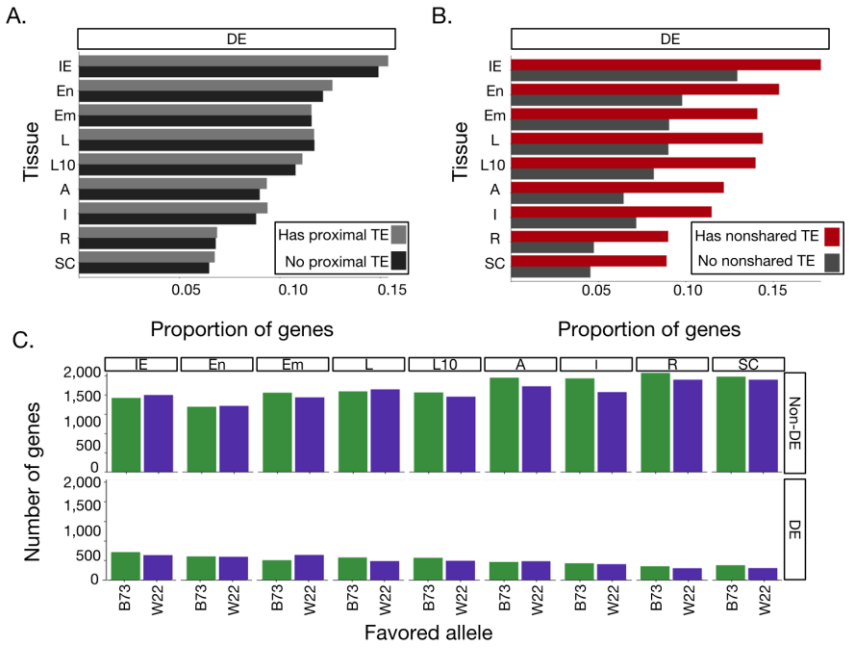


Figure 1. Association of differential expression (DE) with proximal TE insertions and nonshared proximal TE insertions and relationship between proportion of DE genes with allele bias. A.) Relationship between proportion of DE genes and presence of proximal TE insertions. B.) Relationship between proportion of DE genes and the presence of nonshared proximal TE insertion. C.) Proportion of DE and non-DE genes that consistently favor the haplotype with a shared TE insertion (B73) versus a haplotype without a nonshared TE insertion (W22).

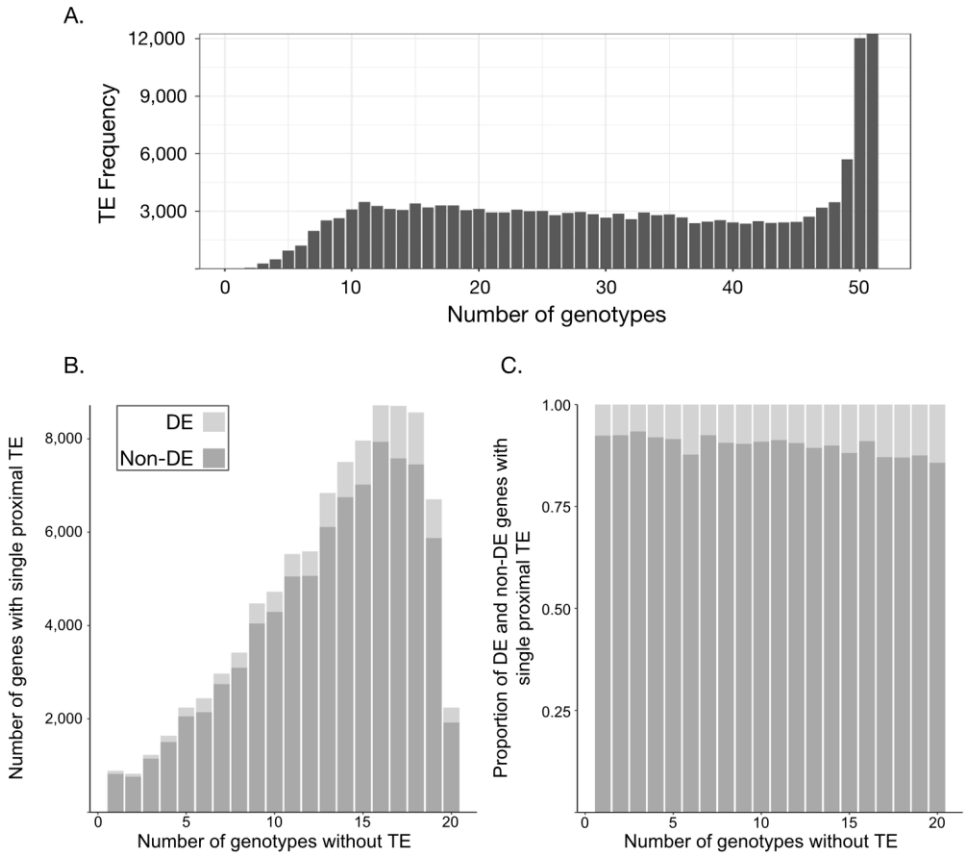


Figure 2. TE flank frequency and relationship with DE status across maize diversity panel. A.) Histogram of the number of maize inbred lines from the Wisconsin Diversity Panel that share a B73 TE insertion as measured by mapping resequencing reads to the B73 reference genome. TE frequency, shown as the y-axis indicates the frequency in which a TE is shared across the number of genotypes in the corresponding x-axis. B.) The relationship between the frequency of TE presence and DE status using RNA-seq data from 20 diverse inbred lines. The height of the bars reflects the number of observations for both DE and non-DE genes in seedling tissue. C.) The proportion of DE and non-DE genes for each x-axis category in panel B.

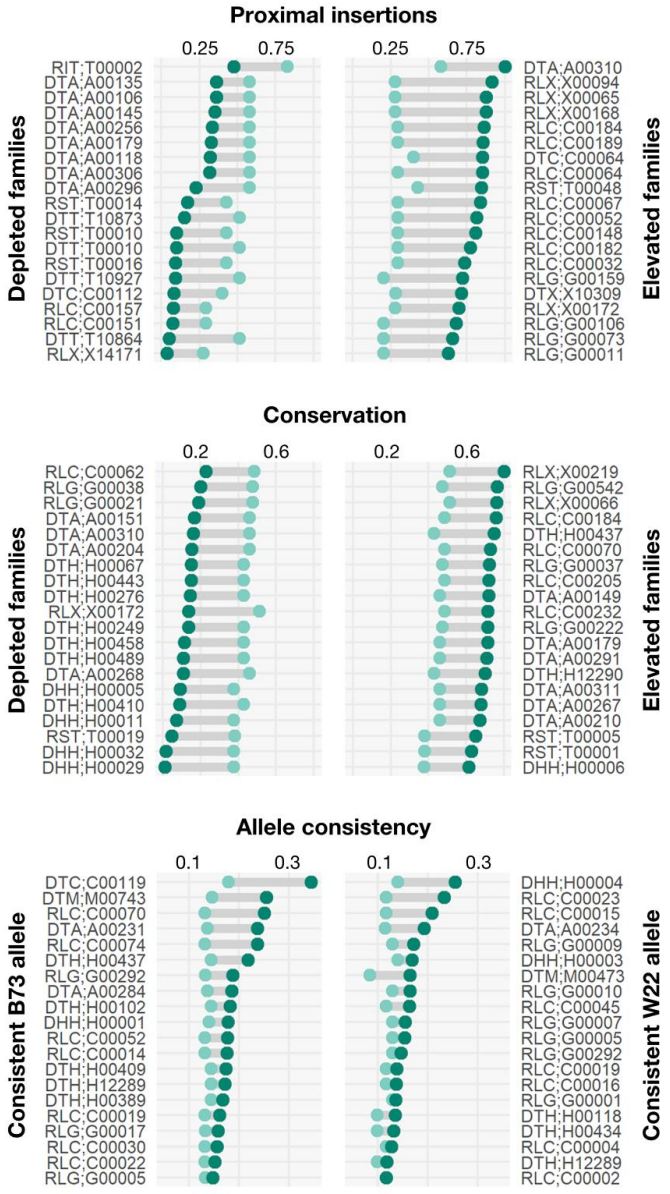


Figure 3. TE family deviations from superfamily proportions. Parts A-C. depict TE families that deviate most from their respective superfamilies across three categories. Part A. shows families that have enriched or depleted numbers of gene proximal insertions compared to each TEs respective superfamily. Part B. shows families that have enriched or depleted numbers of shared TEs compared to the superfamily amount. Part C. considers TE families that consistently favored the allele with the TE insertion, B73, or the opposite allele without the insertion, W22. In all three parts, the light green circles indicate the superfamily value, while the dark green circles represent the family value. The line connecting the two points indicates the difference between these values.

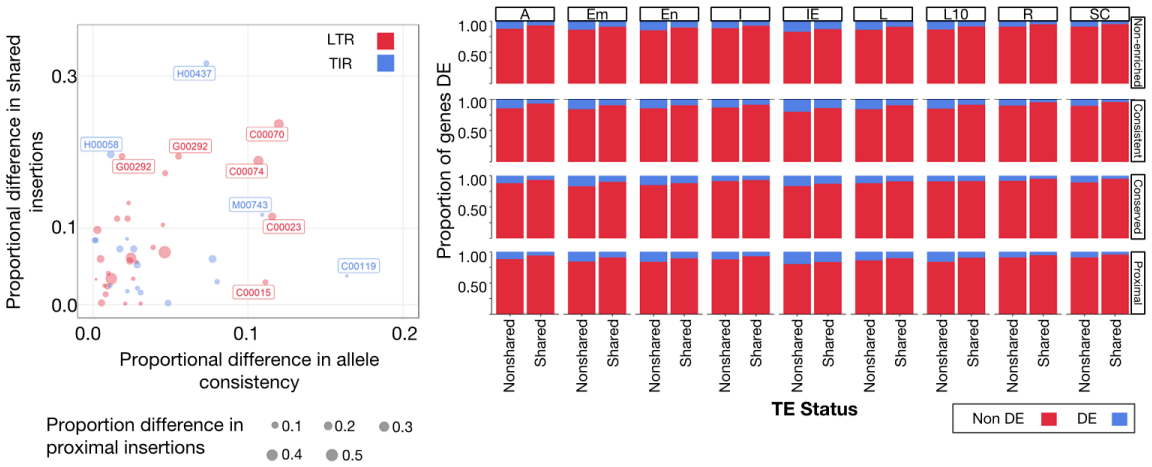


Figure 4. TE proportional difference summary and association of nonshared TEs with DE. A.) Relationship between each of the categories used for the deviation calculations presented in Figure 4. B.) Relationship between expression pattern and genes with at least one nonshared TE between B73 and W22. Genes with consistent expression were defined as favoring one allele (B73 or W22) in every tissue in which the gene was expressed. Genes labeled as 'inconsistent' were expressed in at least one tissue but did not consistently favor one allele.

Bibliography

1001 Genomes Consortium (2016) 1,135 Genomes Reveal the Global Pattern of Polymorphism in Arabidopsis thaliana. *Cell.* 166, 481–491.

Ahn S, Tanksley SD (1993) Comparative linkage maps of the rice and maize genomes. *Proc. Natl. Acad. Sci. USA.* 90, 7980–7984.

Albert PS, Gao Z, Danilova TV, Birchler JA (2010) Diversity of chromosomal karyotypes in maize and its relatives. *Cytogenet Genome Res.* 129, 6–16.

Alkan C, Coe BP, Eichler EE (2011) Genome structural variation discovery and genotyping. *Nature.* 12, 363–376.

Altschul, S., Gish W., Miller W., Myers, E.W. and Lipman, D.J. (1990). Basic Local Alignment Search Tool. *J Mol Biol*., 215, 403–410.

Altschul, S.F., Madden, T.L., Schäffer, A.A., et al. (1997). Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucl Acid S.*, 25, 3389–3402.

Anders, S., Pyl, P.T. and Huber, W. (2015) HTSeq—a Python framework to work with high-throughput sequencing data. *Bioinformatics*. 31, 166–169.

Anderson JE, Kantar MB, Kono TY, et al (2014) A roadmap for functional structural variants in the soybean genome. *G3-Genes Genom Genet.* 4, 1307–1318.

Anderson, S.N., Stitzer, M.C., Noshay, J., Brohammer, A.B., Zhou, P., Hirsch, C.N., Ross-Ibarra, J., Hirsch, C.D. and Springer, N.M. (2018) Transposable element contribution to the dynamic maize genome and transcriptome. In Preparation.

Ayonoadu UW, Rees H (1971) effects of B chromosomes on the nuclear phenotype in root meristems of maize. *Heredity.* 27, 365–383.

Baldauf, J. A., Marcon, C., Paschold, A. & Hochholdinger, F. (2016) Nonsyntenic genes drive tissue-specific dynamics of differential, nonadditive, and allelic expression patterns in maize hybrids. *Plant Physiol.* 171, 1144–1155.

Barkan, A. and Martienssen, R.A. (1991) Inactivation of maize transposon Mu suppresses a mutant phenotype by activating an outward-reading promoter near the end of Mu1. *Proc. Natl. Acad. Sci.* 88, 3502–3506.

Baucom, R.S., Estill, J.C., Chaparro, C., Upshaw, N., Jogi, A., Deragon, J.-M., Westerman, R.P., SanMiguel, P.J. and Bennetzen, J.L. (2009) Exceptional diversity, non-random distribution, and rapid evolution of retroelements in the B73 maize genome. *PLoS Genet*, 5, e1000732.

Bejarano ER, Khashoggi A, Witty M, Lichtenstein C (1996) Integration of multiple repeats of geminiviral DNA into the nuclear genome of tobacco during evolution. *Proc. Natl. Acad. Sci. USA.* 93, 759–764.

Beló A, Beatty MK, Hondred D, et al (2010) Allelic genome structural variations in maize detected by array comparative genome hybridization. *Theor Appl Genet.* 120, 355–367.

Bennetzen, J.L. (2000) Transposable element contributions to plant gene and genome evolution. *Plant Mol. Biol.* 42, 251–269.

Bennetzen JL, Ramakrishna W (2002) Exceptional haplotype variation in maize. *Proc. Natl. Acad. Sci. USA*. 99, 9093–9095.

Bennetzen JL, Wang H (2014) The contributions of transposable elements to the structure, function, and evolution of plant genomes. *Annu Rev Plant Biol.* 65,505–530.

Berglund J, Nevalainen EM, Molin A-M, et al (2012) Novel origins of copy number variation in the dog genome. *Genome Biol.* 13, R73.

Birchler JA, Veitia RA (2007) The gene balance hypothesis: from classical genetics to modern genomics. *Plant Cell.* 19:395–402.

Birchler JA, Veitia RA (2010) The gene balance hypothesis: implications for gene regulation, quantitative traits and evolution. *New Phytol.* 186, 54–62.

Birchler JA, Veitia RA (2012) Gene balance hypothesis: connecting issues of dosage sensitivity across biological disciplines. *Proc. Natl. Acad. Sci. USA.* 109, 14746–14753.

Brohammer AB, Kono TJY, Springer NM, McGaugh SE and Hirsch CN. (2018) The limited role of differential fractionation in genome content variation and function in maize (Zea mays L.) inbred lines. *Plant Journal* 93, 131–141

Brown WL (1949) Numbers and Distribution of Chromosome Knobs in United States Maize. *Genetics.* 34, 524–536.

Brunner S, Fengler K, Morgante M, et al (2005) Evolution of DNA sequence nonhomologies among maize inbreds. *Plant Cell.* 17, 343–360.

Buchmann, R.C., Asad, S., Wolf, J.N., Mohannath, G. and Bisaro, D.M. (2009) Geminivirus AL2 and L2 Proteins Suppress Transcriptional Gene Silencing and Cause Genome-Wide Reductions in Cytosine Methylation. *J. Virol.* 83, 5005–5013.

Buckler ES, Gaut BS, McMullen MD (2006) Molecular and functional diversity of maize. *Curr Opin Plant Biol.* 9, 172–176.

Buescher PJ, Phillips RL, Brambl R (1984) Ribosomal RNA contents of maize genotypes with different ribosomal RNA gene numbers. *Biochem Genet.* 22, 923–930.

Bureau, T.E., Ronald, P.C. and Wessler, S.R. (1996) A computer-based systematic survey reveals the predominance of small inverted-repeat elements in wild-type rice genes. *Proc. Natl. Acad. Sci*. 93, 8524–8529.

Burr B, Burr FA, Matz EC, Romero-Severson J (1992) Pinning down loose ends: mapping telomeres and factors affecting their length. *Plant Cell.* 4, 953–960.

Cao J, Schneeberger K, Ossowski S, et al (2011) Whole-genome sequencing of multiple Arabidopsis thaliana populations. *Nat Genet.* 43, 956–963.

Casacuberta E, Gonzalex J. The impact of transposable elements in environmental adaptation. *Mol Ecol.* 22, 1503–1517

Castelletti, S., Tuberosa, R., Pindo, M. and Salvi, S. (2014) A MITE Transposon Insertion Is Associated with Differential Methylation at the Maize Flowering Time QTL Vgt1. *G3*. 4, 805–812

Chia, J-M, Song, C., Bradbury, P.J., et al. (2012) Maize HapMap2 identifies extant variation from a genome in flux. *Nat Genet*. 44, 803–807.

Chuck, G., Cigan, A.M., Saeteurn, K. and Hake, S. (2007) The heterochronic maize mutant Corngrass1 results from overexpression of a tandem microRNA. *Nat Genet*. 39, 544–549.

Chuck, G.S., Tobias, C., Sun, L., et al. (2011) Overexpression of the maize Corngrass1 microRNA prevents flowering, improves digestibility, and increases starch content of switchgrass. *Proc. Natl. Acad. Sci. USA.* 108, 17550–17555.

Computational Pan-Genomics Consortium (2016) Computational pan-genomics: status, promises and challenges. *Brief Bioinform*. 19, 118–135.

Cook, D.E., Lee, T.G., Guo, X., et al. (2012) Copy number variation of multiple genes at Rhg1 mediates nematode resistance in soybean. *Science*. 338, 1206–1209.

Darracq, A, Vitte, C, Nicolas, S, et al. (2018). Sequence analysis of European maize inbred line F2 provides new insights into molecular and chromosomal characteristics of presence/absence variants. *BMC genomics.* 19, 119.

Dietrich CR, Perera MADN, D Yandeau-Nelson M, et al (2005) Characterization of two GL8 paralogs reveals that the 3-ketoacyl reductase component of fatty acid elongase is essential for maize (Zea mays L.) development. *Plant J.* 42, 844–861.

Díaz A, Zikhali M, Turner AS, et al (2012) Copy number variation affecting the Photoperiod-B1 and Vernalization-A1 genes is associated with altered flowering time in wheat (Triticum aestivum). *PLoS ONE.* 7, e33234.

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

Eichten, S.R., Swanson-Wagner, R.A., Schnable, J.C., et al. (2011) Heritable epigenetic variation among maize inbreds. *PLoS Genet*, 7, e1002372.

Emms, D.M. and Kelly, S. (2015) OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. *Genome Biol.* 16, 157.

Emrich, S.J., Li, L., Wen, T-J, et al. (2007) Nearly identical paralogs: implications for maize (Zea mays L.) genome evolution. *Genetics*, 175, 429–439.

Felice, B. De, Wilson, R., Argenziano, C., Kafantaris, I. and Conicella, C. (2009) A transcriptionally active copia -like retroelement in Citrus limon. *Cell. Mol. Biol. Lett.* 14, 289–304.

Feschotte, C. (2008) Transposable elements and the evolution of regulatory networks. Nat. Rev. Genet. 9, 397–405.

Flavell RB (1986) Repetitive DNA and chromosome evolution in plants. *Philos Trans R Soc London.* 312, 227-242.

Force A, Lynch M, Pickett FB, et al (1999) Preservation of duplicate genes by complementary, degenerative mutations. *Genetics.* 151, 1531–1545.

Freeman VJ (1951) Studies on the virulence of bacteriophage-infected strains of Corynebacterium diphtheriae. *J Bacteriol.* 61, 675–688.

Fu H, Dooner HK (2002) Intraspecific violation of genetic colinearity and its implications in maize. *Proc. Natl. Acad. Sci. USA.* 99, 9573–9578.

Gaines TA, Zhang W, Wang D, et al (2010) Gene amplification confers glyphosate resistance in Amaranthus palmeri. *Proc. Natl. Acad. Sci. USA.* 107, 1029–1034.

Gale MD, Devos KM (1998) Comparative genetics in the grasses. *Proc. Natl. Acad. Sci. USA.* 95, 1971–1974.

Gaut, B.S., Doebley, J.F. (1997). DNA sequence evidence for the segmental allotetraploid origin of maize. *Proc. Natl. Acad. Sci. USA.* 94, 6809–6814.

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

Golicz AA, Bayer PE, Barker GC, et al (2016) The pangenome of an agronomically important crop plant Brassica oleracea. *Nat Commun.* 7, 13390.

Gore MA, Chia J-M, Elshire RJ, et al (2009) A first-generation haplotype map of maize. *Science.* 326, 1115–1117.

Gordon, S.P., Contreras-Moreira, B., Woods, D.P., Des Marais, D.L., Burgess, D., Shu, S., Stritt, C., Roulin, A.C., Schackwitz, W., Tyler, L. and Martin, J., (2017) Extensive gene content variation in the Brachypodium distachyon pan-genome correlates with population structure. *Nature Commun.* 8, 2184.

Grandbastien, M.A., Audeon, C., Bonnivard, E., et al. (2005) Stress activation and genomic impact of Tnt1 retrotransposons in Solanaceae. *Cytogenet. Genome Res.* 110, 229–2

Greene, B., Walko, R. and Hake, S. (1994) Mutator insertions in an intron of the maize knotted1 gene result in dominant suppressible mutations. *Genetics.* 138, 1275–1285.

Hake S, Walbot V (1980) The genome of Zea mays, its organization and homology to related grasses. *Chromosoma.* 79, 251–270

Han J-J, Jackson D, Martienssen R (2012) Pod corn is caused by rearrangement at the Tunicate1 locus. *Plant Cell.* 24, 2733–2744.

Hansey, C.N., Johnson, J.M., Sekhon, R.S., Kaeppler, S.M. and Leon, N. de (2011) Genetic diversity of a maize association population with restricted phenology. Crop Sci. 51, 704–715.

Hansey, C.N., Vaillancourt, B., Sekhon, R.S., de Leon, N., Kaeppler, S.M. and Buell, C.R. (2012) Maize (Zea mays L.) Genome Diversity as Revealed by RNA-Sequencing A. Moustafa, ed. *PLoS ONE*. 7, e33071.

Haro von Mogel K, Hirsch CN, De Vries B, et al (2013) The mapping, genetic analysis, and phenotypic characterization of sugary enhancer1 (se1). In Maize Genetics Conference Abstracts vol. 55 p T16

Hirsch, C.N., Foerster, J.M., Johnson, J.M., et al. (2014) Insights into the maize pan-genome and pan-transcriptome. *Plant Cell*. 26. 121–135.

Hirsch, C.N., Hirsch, C.D., Brohammer, A.B., et al. (2016) Draft Assembly of Elite Inbred Line PH207 Provides Insights into Genomic and Transcriptome Diversity in Maize. *Plant Cell*. 28, 2700–2714.

Hirsch, C.D. and Springer, N.M. (2017) Transposable element influences on gene expression in plants. *Biochim. Biophys. Acta - Gene Regul. Mech.* 1860, 157–165.

Hogg JS, Hu FZ, Janto B, et al (2007) Characterization and modeling of the Haemophilus influenzae core and supragenomes based on the complete genomic sequences of Rd and 12 clinical nontypeable strains. *Genome Biol.* 8, R103.

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

Hunter, CT, Saunders, JW, Magallanes-Lundback, M., Christensen, SA, Willett, D., Stinard, PS, Li, Q.-B., Lee, K., DellaPenna, D. and Koch, KE. (2018), Maize w3 disrupts homogentisate solanesyl transferase (ZmHst) and reveals a

plastoquinone-9 independent path for phytoene desaturation and tocopherol accumulation in kernels. *Plant J.* 93, 799–813.

Ito, H., Gaubert, H., Bucher, E., Mirouze, M., Vaillant, I. and Paszkowski, J. (2011) An siRNA pathway prevents transgenerational retrotransposition in plants subjected to stress. Nature. 472, 115–120.

Ito, H., Kim, J.M., Matsunaga, W., et al. (2016) A Stress-Activated Transposon in Arabidopsis Induces Transgenerational Abscisic Acid Insensitivity. *Sci. Rep*. 6, 23181.

Ivashuta, S., Naumkina, M., Gau, M., Uchiyama, K., Isobe, S., Mizukami, Y. and Shimamoto, Y. (2002) Genotype-dependent transcriptional activation of novel repetitive elements during cold acclimation of alfalfa (Medicago sativa). *Plant J.* 31, 615–627.

Jacq C, Miller JR, Brownlee GG (1977) A pseudogene structure in 5S DNA of Xenopus laevis. *Cell*. 12, 109–120.

Jeffares DC, Jolly C, Hoti M, et al (2017) Transient structural variations have strong effects on quantitative traits and reproductive isolation in fission yeast. *Nat Commun* 8, 14061.

Jiao, Yinping, Wang, B., Campbell, M.S., et al. (2017) Improved maize reference genome with single-molecule technologies. *Nature*, 546, 524–527.

Jiao, Yuannian, Wickett, N.J., Ayyampalayam, S., et al. (2011) Ancestral polyploidy in seed plants and angiosperms. *Nature*, 473, 97–100.

Jordan, I.K., Rogozin, I.B., Glazko, G. V. and Koonin, E. V. (2003) Origin of a substantial fraction of human regulatory sequences from transposable elements. *Trends Genet.* 19, 68–72.

Joshi N.A. and Fass J.N. (2011). Sickle: A sliding-window, adaptive, quality-based trimming tool for FastQ files (version 1.33) [Software] https://github.com/najoshi/sickle.

Kaeppler SM (2012) Heterosis: many genes, many mechanisms—end the search for an undiscovered unifying theory. ISRN Bot. 2012:682824.

Kato A, Lamb JC, Birchler JA (2004) Chromosome painting using repetitive DNA sequences as probes for somatic chromosome identification in maize. *Proc. Natl. Acad. Sci. USA.* 101, 13554–13559.

Kato TAY (1976) Cytological studies of maize [Zea mays L.] and teosinte [Zea mexicana Schrader Kuntze] in relation to their origin and evolution. Univ. Mas. Agric. Expt. Sta.

Keller, B. and Feuillet, C. (2000) Colinearity and gene density in grass genomes. *Trends Plant Sci.* 5, 246–251.

Kiełbasa, S.M., Wan, R., Sato, K., Horton, P. and Frith, M.C. (2011) Adaptive seeds tame genomic sequence comparison. *Genome Res.* 21, 487–493.

Kim, D., Pertea, G., Trapnell, C., Pimentel, H., Kelley, R. and Salzberg, S.L. (2013) TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol.* 14, R36.

Knox AK, Dhillon T, Cheng H, et al (2010) CBF gene copy number variation at Frost Resistance-2 is associated with levels of freezing tolerance in temperate-climate cereals. *Theor Appl Genet.* 121, 21–35.

Koonin EV, Wolf YI (2008) Genomics of bacteria and archaea: the emerging dynamic view of the prokaryotic world. *Nucleic Acids Res.* 36, 6688–6719.

Kyndt T, Quispe D, Zhai H, et al (2015) The genome of cultivated sweet potato contains Agrobacterium T-DNAs with expressed genes: An example of a naturally transgenic food crop. *Proc Natl Acad Sci USA.* 112, 5844–5849.

Lai, J., Li, R., Xu, X., et al. (2010) Genome-wide patterns of genetic variation among elite maize inbred lines. *Nat Genet*. 42, 1027–1030.

Lai J, Li Y, Messing J, Dooner HK (2005) Gene movement by Helitron transposons contributes to the haplotype variability of maize. *Proc. Natl. Acad. Sci. USA.* 102:9068–9073.

Langham, R.J., Walsh, J., Dunn, M., Ko, C., Goff, S.A. and Freeling, M. (2004) Genomic duplication, fractionation and the origin of regulatory novelty. *Genetics*, 166, 935–945.

Laurie DA, Bennett MD (1985) Nuclear DNA content in the genera Zea and Sorghum. Intergeneric, interspecific and intraspecific variation. *Heredity.* 55, 307–313.

Langmead, B. and Salzberg, S.L. (2012) Fast gapped-read alignment with Bowtie 2. *Nat. Methods*, 9, 357–359.

Lee H, Gurtowski J, Yoo S et al (2016) Third-generation sequencing and the future of genomics. bioRxiv.

Li, H., Handsaker, B., Wysoker, A., et al. (2009) The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, 25, 2078–2079.

Li R, Li Y, Zheng H, et al (2010) Building the sequence map of the human pan-genome. *Nat Biotechnol.* 28, 57–63.

Li Y, Xiao J, Wu J, et al (2012) A tandem segmental duplication (TSD) in green revolution gene Rht-D1b region underlies plant height variation. *New Phytol.* 196, 282–291.

Li Y-H, Zhou G, Ma J, et al (2014) De novo assembly of soybean wild relatives for pan-genome analysis of diversity and agronomic traits. *Nat Biotechnol.* 32, 1045–1052.

Li, Q., Gent, J.I., Zynda, G., et al. (2015) RNA-directed DNA methylation enforces boundaries between heterochromatin and euchromatin in the maize genome. *Proc. Natl. Acad. Sci.* 32, 14728–14733.

Liao, Y., Smyth, G.K. and Shi, W. (2014) FeatureCounts: An efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics*. 30, 923–930.

Liu, S., Ying, K., Yeh, C-T, et al. (2012) Changes in genome content generated via segregation of non-allelic homologs. *Plant J.* 72, 390–399.

Liu S, Zheng J, Migeon P, et al (2017) Unbiased K-mer Analysis Reveals Changes in Copy Number of Highly Repetitive Sequences During Maize Domestication and Improvement. *Sci Rep.* 7:42444.

Love, M.I., Huber, W., Anders, S., et al. (2014) Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* 15, 550.

Lu, F., Romay, M.C., Glaubitz, J.C., et al. (2015) High-resolution genetic mapping of maize pan-genome sequence anchors. *Nat. Commun.*, 6, 6914.

Lyons, E., Pedersen, B., Kane, J. and Freeling, M. (2008) The Value of Nonmodel Genomes and an Example Using SynMap Within CoGe to Dissect the Hexaploidy that Predates the Rosids. *Trop Plant Biol*. 1, 181–190.

Maize Genetics and Genomics Database. Information about maize assembly Zm-Mo17-REFERENCE-NRGENE-1.0. In: maizegdb.org. http://maizegdb.org/genome/genome_assembly/Zm-Mo17-REFERENCE-NRGENE-1.0. Accessed 2 Jun 2017.

Maize Genetics and Genomics Database. Information about maize assembly Zm-B104-DRAFT-ISU_USDA-0.1. In: maizegdb.org. https://maizegdb.org/genome/genome_assembly/Zm-B104-DRAFT-ISU_USDA-0.1. Accessed 2 Jun 2017.

Maize Genetics and Genomics Database. Maize CML247. In: maizegdb.org. http://maizegdb.org/gbrowse/maize_cml247. Accessed 2 Jun 2017.

Makarevitch I, Waters AJ, West PT, et al (2015) Transposable elements contribute to activation of maize genes in response to abiotic stress. *PLoS Genet.* 11:e1004915.

Maron, L.G., Guimarães, C.T., Kirst, M., et al. (2013) Aluminum tolerance in maize is associated with higher MATE1 gene copy number. *Proc. Natl. Acad. Sci. USA.* 110, 5241–5246.

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

McClintock, B. (1948) Mutable loci in maize. Carnegie Inst. Wash. Yearb., 47, 155–169.

McClintock B (1950) The origin and behavior of mutable loci in maize. *Proc. Natl. Acad. Sci. USA.* 36, 344–355.

McClintock B, Kato A, Blumenschein A (1981) Chromosome constitution of races of maize: its significance in the interpretation of relationship between races and varieties in the Americas. Mexico: Colegio de Postgraduados 517.

Medini D, Donati C, Tettelin H, et al (2005) The microbial pan-genome. *Curr Opin Genet Dev.* 15, 589–594.

Messing J, Bharti AK, Karlowski WM, et al (2004) Sequence composition and genome organization of maize. *Proc. Natl. Acad. Sci. USA.* 101, 14349–14354.

Meyers BC, Tingey SV, Morgante M (2001) Abundance, distribution, and transcriptional activity of repetitive elements in the maize genome. *Genome Res.* 11, 1660–1676.

Mhiri, C., Morel, J.B., Vernhettes, S., Casacuberta, J.M., Lucas, H. and Grandbastien, M.A. (1997) The promoter of the tobacco Tnt1 retrotransposon is induced by wounding and by abiotic stress. *Plant Mol. Biol.* 33, 257–266.

Michael TP, Jackson S (2013) The first 50 plant genomes. Plant Genome 6:2.

Montenegro JD, Golicz AA, Bayer PE, et al (2017) The pangenome of hexaploid bread wheat. *Plant J.* 90:1007–1013.

Morgante M, Brunner S, Pea G, et al (2005) Gene duplication and exon shuffling by helitron-like transposons generate intraspecies diversity in maize. *Nat Genet.* 37:997–1002.

Muehlbauer, G.J., Fowler, J.E., Girard, L., Tyers, R., Harper, L. and Freeling, M. (1999) Ectopic Expression of the Maize Homeobox Gene Liguleless3 Alters Cell Fates in the Leaf. *Plant Physiol.* 119, 651–662.

Naito, K., Zhang, F., Tsukiyama, T., Saito, H., Hancock, C.N., Richardson, A.O., Okumoto, Y., Tanisaka, T. and Wessler, S.R. (2009) Unexpected consequences of a sudden and massive transposon amplification on rice gene expression. *Nature.* 461, 1130–1134.

Näsvall J, Sun L, Roth JR, Andersson DI (2012) Real-time evolution of new genes by innovation, amplification, and divergence. *Science,* 338, 384–387.

Nitcher, R., Distelfeld, A., Tan, C., Yan, L. and Dubcovsky, J. (2013) Increased copy number at the HvFT1 locus is associated with accelerated flowering time in barley. *Mol. Genet. Genomics*, 288, 261–275.

Ohno S (1970) Evolution by gene duplication. Springer-Verlag, New York, New York, USA.

Ohtsu, K., Smith, M.B., Emrich, S.J., et al. (2007) Global gene expression analysis of the shoot apical meristem of maize (Zea mays L.). Plant J. 62, 391–404.

Oka, R., Zicola, J., Weber, B., et al. (2017) Genome-wide mapping of transcriptional enhancer candidates using DNA and chromatin features in maize. Genome Biol., 18, 137.

Ouyang, S., Zhu, W., Hamilton, J., et al. (2007) The TIGR Rice Genome Annotation Resource: improvements and new features. *Nucleic Acids Res.*, 35, D883–7.

Peacock WJ, Dennis ES (1981) Highly repeated DNA sequence limited to knob heterochromatin in maize. *Proc. Natl. Acad. Sci. USA.* 78, 4490–4494.

Phillips RL, Weber DF, Kleese RA, Wang SS (1974) The Nucleolus Organizer Region of Maize (Zea mays L.): Tests for Ribosomal Gene Compensation or Magnification. *Genetics.* 77, 285–297.

Poggio L, Rosato M, Chiavarino AM, Naranjo CA (1998) Genome Size and Environmental Correlations in Maize (Zea mays ssp. mays, Poaceae). Ann Bot 82:107–115.

Pophaly, S.D. and Tellier, A. (2015) Population Level Purifying Selection and Gene Expression Shape Subgenome Evolution in Maize. *Mol. Biol. Evol.*, 32, 3226–3235.

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

R Core Team. R: A Language and Environment for Statistical Computing [Internet]. 3rd ed. Vienna, Austria; 2014. http://www.R-project.org/.

Rayburn AL, Price HJ, Smith JD, Gold JR (1985) C-Band Heterochromatin and DNA Content in Zea mays. *Am J Bot.* 72, 1610–1617.

Renny-Byfield, S., Rodgers-Melnick, E. and Ross-Ibarra, J. (2017) Gene fractionation and function in the ancient subgenomes of maize. *Mol. Biol. Evol*. 8 1825–1832.

Rivin CJ, Cullis CA, Walbot V (1986) Evaluating quantitative variation in the genome of Zea mays. *Genetics.* 113, 1009–1019.

Romero Navarro JA, Wilcox M, Burgueño J, et al (2017) A study of allelic diversity underlying flowering-time adaptation in maize landraces. *Nat Genet.* 49, 476–480.

Salman-Minkov, A., Sabath, N. and Mayrose, I. (2016) Whole-genome duplication as a key factor in crop domestication. *Nat Plants*, 2, 16115.

Salvi, S., Sponza, G., Morgante, M., et al. (2007) Conserved noncoding genomic sequences associated with a flowering-time quantitative trait locus in maize. *Proc. Natl. Acad. Sci. U.S.A.* 104, 11376–11381.

SanMiguel, P., Tikhonov, A., Jin, Y.K., et al. (1996) Nested retrotransposons in the intergenic regions of the maize genome. *Science.* 274, 765–768.

SanMiguel PJ, Bennetzen JL (1998) Evidence that a Recent Increase in Maize Genome Size was Caused by the Massive Amplification of Intergene Retrotransposons. *Ann Bot.* 82, 37–44.

Schatz MC, Maron LG, Stein JC, et al (2014) Whole genome de novo assemblies of three divergent strains of rice, Oryza sativa, document novel gene space of aus and indica. *Genome Biol.* 15, 506.

Schnable, J.C. and Freeling, M. (2011) Genes identified by visible mutant phenotypes show increased bias toward one of two subgenomes of maize.. *PLoS ONE*. 6, e17855.

Schnable, J.C., Springer, N.M. and Freeling, M. (2011) Differentiation of the maize subgenomes by genome dominance and both ancient and ongoing gene loss. *Proc. Natl. Acad. Sci. U.S.A.* 108, 4069–4074.

Schnable, J.C., Freeling, M. and Lyons, E. (2012) Genome-wide analysis of syntenic gene deletion in the grasses. *Genome Biol Evol*, 4, 265–277.

Schnable, P.S., Ware, D.H., Fulton, R.S., et al. (2009) The B73 maize genome: complexity, diversity, and dynamics. *Science*, 326, 1112–1115.

Schnable PS, Springer NM (2013) Progress toward understanding heterosis in crop plants. *Annu Rev Plant Biol.* 64, 71–88.

Schneeberger, R.G., Becraft, P.W., Hake, S. and Freeling, M. (1995) Ectopic expression of the knox homeobox gene rough sheath1 alters cell fate in the maize leaf. Genes Dev. 9, 2292–2304.

Sekhon, R., H. Lin, K. Childs, C. Hansey, C.R. Buell, N. de Leon and S. Kaeppler. (2011). Genome-wide atlas of transcription through maize development. Plant Journal. 66, 553–563.

Slotkin, R.K. and Martienssen, R. (2007) Transposable elements and the epigenetic regulation of the genome. *Nat. Rev. Genet.* 8, 272–285.

Sievers, F., Wilm, A., Dineen, D., et al. (2011) Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol. Syst. Biol.* 7, 539–539.

Soltis, P.S. and Soltis, D.E. eds. (2012) *Polyploidy and Genome Evolution*, Berlin, Heidelberg: Springer Berlin Heidelberg.

Song R, Messing J (2003) Gene expression of a gene family in maize based on noncollinear haplotypes. *Proc. Natl. Acad. Sci. USA.* 100, 9055–9060.

Springer, N.M., Ying, K., Fu, Y., et al. (2009) Maize Inbreds Exhibit High Levels of Copy Number Variation (CNV) and Presence/Absence Variation (PAV) in Genome Content J. R. Ecker, ed. *PLoS Genet*. 5, e1000734.

Springer, N.M., Lisch, D. and Li, Q. (2016) Creating order from chaos: epigenome dynamics in plants with complex genomes. *Plant Cell*. 28, 314–325.

Stelpflug, S., Sekhon, R.S., Vaillancourt, B., Hirsch, C.N., Buell, C.R., de Leon, N., Kaeppler, S.M. (2016). An expanded maize gene expression atlas based on RNA-sequencing and its use to explore root development. *Plant Genome*. 9, 1–16.

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

Sundaram, V., Cheng, Y., Ma, Z., Li, D., Xing, X., Edge, P., Snyder, M.P. and Wang, T. (2014) Widespread contribution of transposable elements to the innovation of gene regulatory networks. *Genome Res*. 12, 1963–1976.

Sutton T, Baumann U, Hayes J, et al (2007) Boron-toxicity tolerance in barley arising from efflux transporter amplification. *Science.* 318, 1446–1449.

Swanson-Wagner, R.A., Eichten, S.R., Kumari, S., Tiffin, P., Stein, J.C., Ware, D.H. and Springer, N.M. (2010) Pervasive gene content variation and copy number variation in maize and its undomesticated progenitor. *Genome Res.* 20, 1689–1699.

Swigonová, Z., Lai, J., Ma, J., Ramakrishna, W., Llaca, V., Bennetzen, J.L. and Messing, J. (2004) Close split of sorghum and maize genome progenitors. *Genome Res.* 14, 1916–1923.

Syvanen M (2012) Evolutionary implications of horizontal gene transfer. *Annu Rev Genet.* 46, 341–358.

Tan B-C, Guan J-C, Ding S, et al (2017) Structure and Origin of the White Cap Locus and Its Role in Evolution of Grain Color in Maize. *Genetics.* 206, 135–150.

Tang H, Bowers JE, Wang X, et al (2008) Synteny and collinearity in plant genomes. *Science*. 320, 486–488.

Tank, D.C., Eastman, J.M., Pennell, M.W., et al. (2015) Nested radiations and the pulse of angiosperm diversification: increased diversification rates often follow whole genome duplications. *New Phytol.* 207, 454–467.

Tarasov, A., Vilella, A.J., Cuppen, E., Nijman, I.J. and Prins, P. (2015) Sambamba: fast processing of NGS alignment formats. *Bioinformatics*. 31, 2032–2034.

Tenaillon MI, Hollister JD, Gaut BS (2010) A triptych of the evolution of plant transposable elements. *Trends Plant Sci.* 15, 471–478.

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

Tettelin H, Masignani V, Cieslewicz MJ, et al (2005) Genome analysis of multiple pathogenic isolates of Streptococcus agalactiae: implications for the microbial "pan-genome". *Proc. Natl. Acad. Sci. USA.* 102, 13950–13955.

Tettelin H, Riley D, Cattuto C, Medini D (2008) Comparative genomics: the bacterial pan-genome. *Curr Opin Microbiol.* 11, 472–477.

Thornton, K. (2003) Libsequence: a C++ class library for evolutionary genetic analysis. *Bioinformatics*, 19, 2325–2327.

Tikhonov AP, SanMiguel PJ, Nakajima Y, et al (1999) Colinearity and its exceptions in orthologous adh regions of maize and sorghum. *Proc. Natl. Acad. Sci. USA.* 96, 7409–7414.

Torres EM, Williams BR, Amon A (2008) Aneuploidy: cells losing their balance. *Genetics.* 179, 737–746.

Unterseer S, Seidel MA, Bauer E, Haberer G (2017) European Flint reference sequences complement the maize pan-genome. bioRxiv.

Varagona, M.J., Purugganan, M. and Wessler, S.R. (1992) Alternative Splicing Induced by Insertion of Retrotransposons into the Maize waxy Gene. *Plant Cell.* 4, 811–820.

Van de Peer Y, Mizrachi E, Marchal K (2017) The evolutionary significance of polyploidy. *Nat Rev Genet.* 408:796.

Vielle-Calzada J-P, Martínez de la Vega O, Hernández-Guzmán G, et al (2009) The Palomero genome suggests metal effects on domestication. *Science.* 326, 1078–1078.

Wallace, J.G., Bradbury, P.J., Zhang, N., Gibon, Y., Stitt, M. and Buckler, E.S. (2014) Association mapping across numerous traits reveals patterns of functional variation in maize. J. O. Borevitz, ed. *PLoS Genet*. 10, e1004845.

Walley, J.W., Sartor, R.C., Shen, Z., et al. (2016) Integration of omic networks in a developmental atlas of maize. *Science*, 353, 814–818.

Wei, F., Stein, J.C., Liang, C., et al. (2009) Detailed analysis of a contiguous 22-Mb region of the maize genome.*PLoS Genet.* 11, e1000728.

Wesler, S.R. and Varagona, M.J. (1985) Molecular basis of mutations at the waxy locus of maize: correlation with the fine structure genetic map. *Proc. Natl. Acad. Sci.*, 82, 4177–4181.

Wetterstrand KA DNA Sequencing Costs: Data from the NHGRI Genome Sequencing Program (GSP). In: www.genome.gov/sequencingcostsdata. Accessed 2 Jun 2017

White, S.E., Habera, L.F. and Wessler, S.R. (1994) Retrotransposons in the flanking regions of normal plant genes: A role for copia-like elements in the evolution of gene structure and expression. *Proc. Natl. Acad. Sci. USA.*, 91, 11792–11796.

Whitkus R, Doebley J, Lee M (1992) Comparative genome mapping of Sorghum and maize. *Genetics.* 132, 1119–1130.

Wicker, T., Sabot, F., Hua-Van, A., et al. (2007) A unified classification system for eukaryotic transposable elements. Nat. Rev. Genet. 12, 973–982.

Wingen LU, Münster T, Faigl W, et al. (2012) Molecular genetic basis of pod corn (Tunicate maize). *Proc Natl Acad Sci USA.* 109, 7115–7120.

Winzer, T., Gazda, V., He, Z., et al. (2012) A Papaver somniferum 10-gene cluster for synthesis of the anticancer alkaloid noscapine. *Science*. 336, 1704–1708.

Woodhouse, M.R., Schnable, J.C., Pedersen, B.S., et al. (2010) Following Tetraploidy in Maize, a Short Deletion Mechanism Removed Genes Preferentially from One of the Two Homeologs K. H. Wolfe, ed. PLoS Biol. 8, e1000409–15.

Würschum T, Longin CFH, Hahn V, et al (2017) Copy number variations of CBF genes at the Fr-A2 locus are essential components of winter hardiness in wheat. *Plant J.* 89, 764–773.

Xiao H, Jiang N, Schaffner E, et al (2008) A retrotransposon-mediated gene duplication underlies morphological variation of tomato fruit. *Science.* 319, 1527–1530.

Xu X, Liu X, Ge S, et al (2011) Resequencing 50 accessions of cultivated and wild rice yields markers for identifying agronomically important genes. *Nat Biotechnol.* 30, 105–111.

Yandeau-Nelson MD, Zhou Q, Yao H, et al (2005) MuDR transposase increases the frequency of meiotic crossovers in the vicinity of a Mu insertion in the maize a1 gene. *Genetics.* 169, 917–929.

Yandeau-Nelson MD, Xia Y, Li J, et al (2006) Unequal sister chromatid and homolog recombination at a tandem duplication of the A1 locus in maize. *Genetics.* 173, 2211–2226.

Yang, Z. (2007) PAML 4: phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol.*, 24, 1586–1591.

Yang, Q., Li, Z., Li, W., et al. (2013) CACTA-like transposable element in ZmCCT attenuated photoperiod sensitivity and accelerated the postdomestication spread of maize. *Proc. Natl. Acad. Sci.*, 110, 16969–16974.

Yao, H., Zhou, Q., Li, J., Smith, H., Yandeau, M., Nikolau, B.J. and Schnable, P.S. (2002) Molecular characterization of meiotic recombination across the 140-kb multigenic a1-sh2 interval of maize. *Proc. Natl. Acad. Sci. USA.* 9, 6157–6162.

Yao W, Li G, Zhao H, et al (2015) Exploring the rice dispensable genome using a metagenome-like assembly strategy. *Genome Biol.* 16, 187.

Yasuda, K., Ito, M., Sugita, T., Tsukiyama, T., Saito, H., Naito, K., Teraishi, M., Tanisaka, T. and Okumoto, Y. (2013) Utilization of transposable element mPing as a novel genetic tool for modification of the stress response in rice. *Mol. Breed.*, 32, 505–516.

Yona AH, Manor YS, Herbst RH, et al (2012) Chromosomal duplication is a transient evolutionary solution to stress. *Proc Natl Acad Sci USA.* 109, 21010–21015.

Yoshida S, Maruyama S, Nozaki H, Shirasu K (2010) Horizontal gene transfer by the parasitic plant Striga hermonthica. Science 328, 1128–1128.

Young ND, Zhou P, Silverstein KA (2016) Exploring structural variants in environmentally sensitive gene families. *Curr Opin Plant Biol.* 30, 19–24.

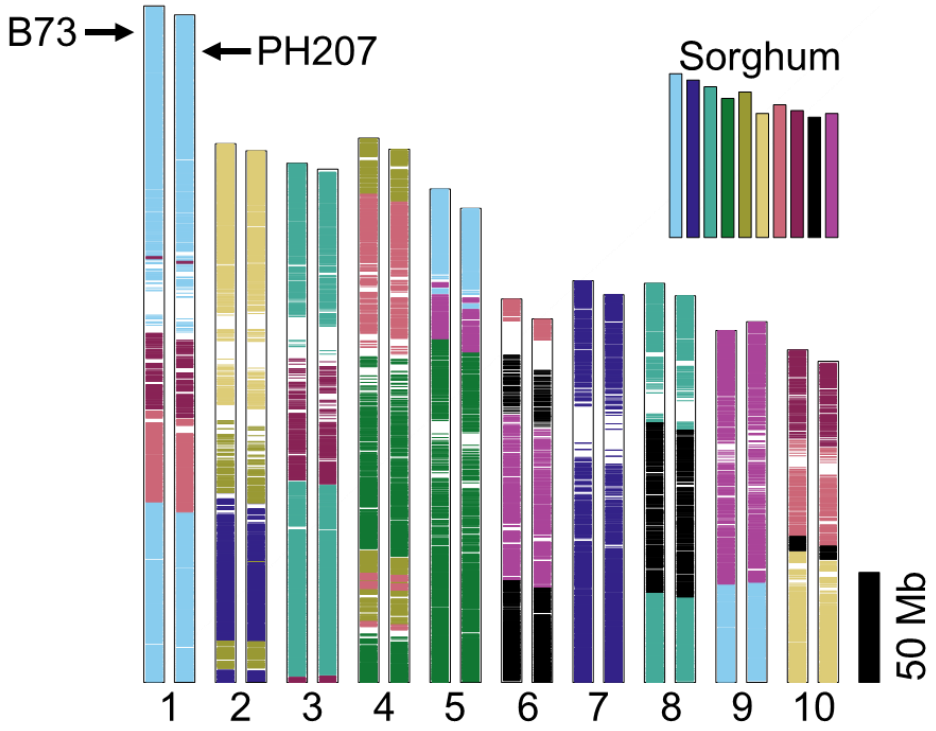
H. Zhao, W. Zhang, L. Chen, L. Wang, A. P. Marand, Y. Wu, J. Jiang. (2018) Proliferation of regulatory DNA elements derived from transposable elements in the maize genome. *Plant Physiol.* 176, 2789–2803.

Zhang Z, Mao L, Chen H, et al (2015) Genome-Wide Mapping of Structural Variations Reveals a Copy Number Variant That Determines Reproductive Morphology in Cucumber. *Plant Cell.* 27, 1595–1604.

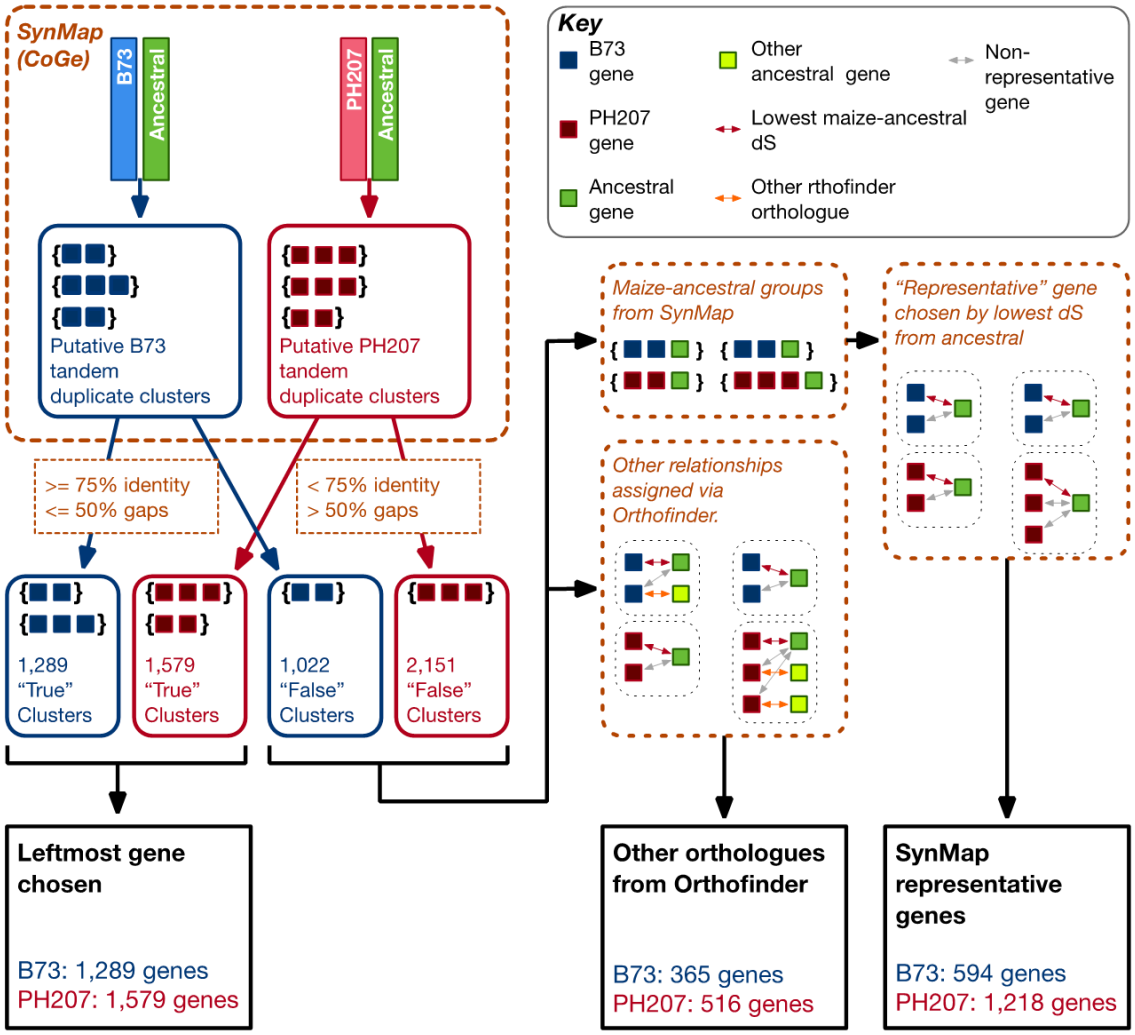
Zhou P, Silverstein KAT, Ramaraj T, et al (2017) Exploring structural variation and gene family architecture with De Novo assemblies of 15 Medicago genomes. *BMC Genomics*. 18, 261.

Zhu J, Pearce S, Burke A, et al (2014) Copy number and haplotype variation at the VRN-A1 and central FR-A2 loci are associated with frost tolerance in hexaploid wheat. *Theor Appl Genet* 127, 1183–1197.

Appendix 1: Chapter 2 Supplementary Material



Supplementary Figure 1. Maize ancestral syntenic blocks in the B73 and PH207 genomes. The 10 maize chromosomes are represented with B37 on the left and PH207 on the right. Syntenic blocks for the maize1 and maize2 subgenomes were determined based on comparison with the ancestral state from sorghum and rice. Colors of the maize chromosomes represent the ancestral chromosome relative to sorghum. The inset depicts the sorghum chromosomal karyotype colored by chromosome.

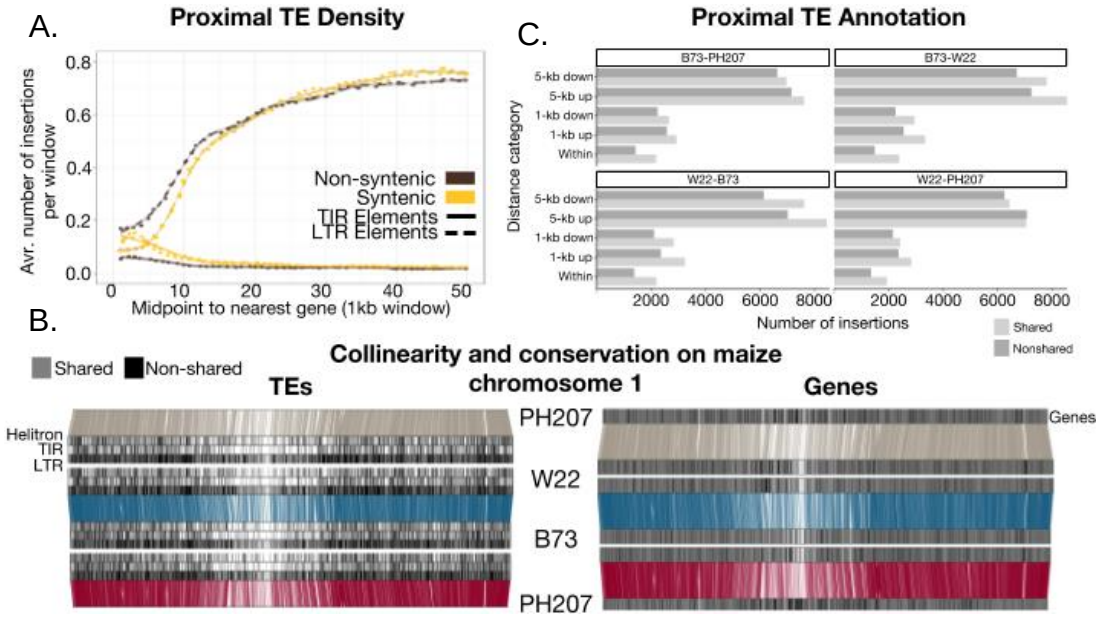


Supplementary Figure 2. Analysis pipeline for resolving tandem duplicates. Tandem duplicates were identified by SynMap (Lyons et al., 2008) and were resolved as true or false tandem duplicates. For false tandem duplicates the correct ancestral gene was resolved using a series of criteria. For true tandem duplicates the left-most gene was selected as the representative copy.

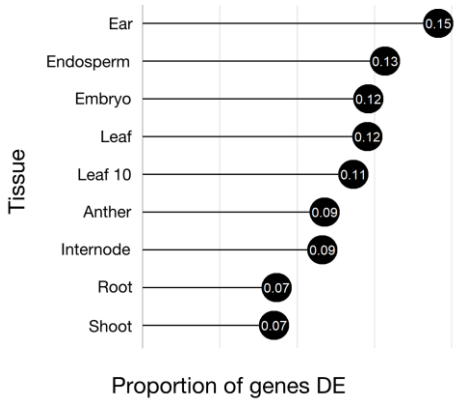
Supplemental Tables for this chapter are available with the original publication:

Brohammer AB, Kono TJY, Springer NM, McGaugh SE, Hirsch CN. 2018. The limited role of differential fractionation in genome content variation and function maize (Zea mays L.) inbred lines. *Plant Journal*. 93:131-141. doi:10.1111/tpj.13765

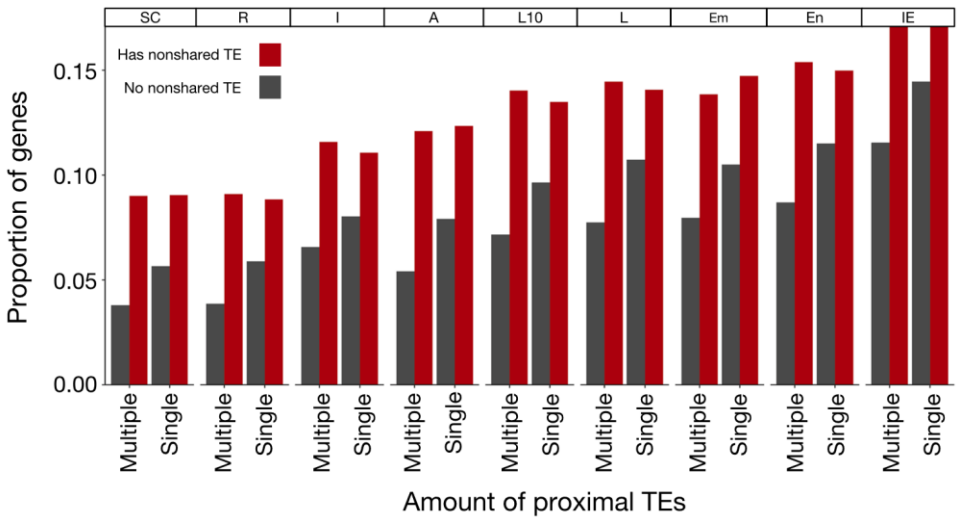
Appendix 2: Chapter 3 Supplementary Material



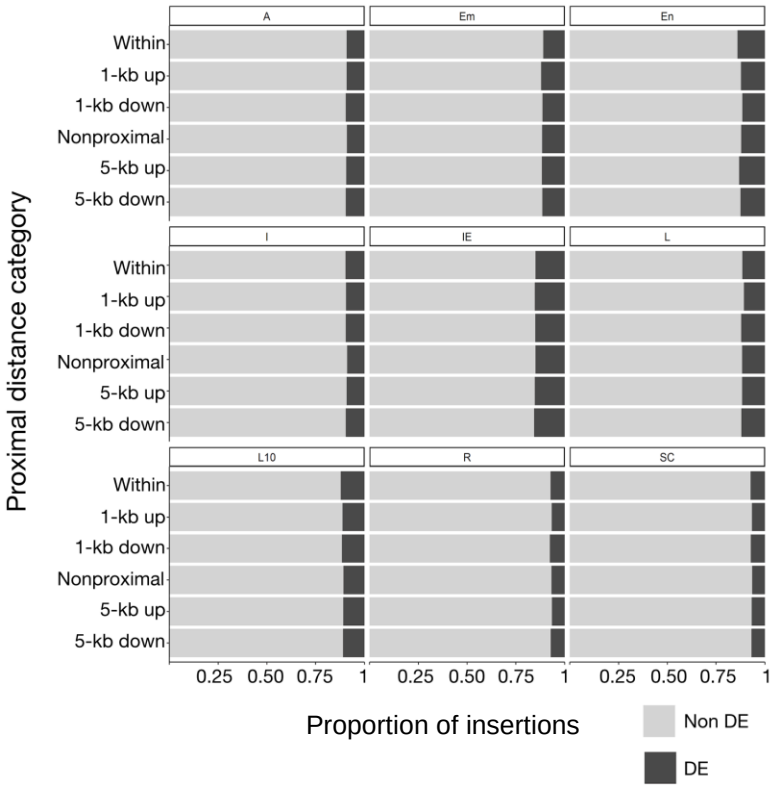
Supplemental Figure 1. Proximal TE characterization by density, distance category, and collinearity across the maize genome. A.) TE density as measured by calculating the number of insertions in non-overlapping 1-kb windows moving from a gene TSS to the midpoint of the nearest gene. The midpoint of each LTR and TIR element was used to determine its presence in a window and the average number of insertions for each 1-kb window interval was used for plotting. B.) Gene collinearity and the patterns of shared and nonshared TEs for pairwise comparison of B73, W22, and PH207 across maize chromosome 1. For both panels, collinear genes are depicted as tan, blue, and maroon ribbons that connect homologous genes in one genome with those in another. For the left panel, Helitron, LTR, and TIR elements are plotted based on the tracks above each ribbon, while genes are plotted on right panel. Conserved genes and TEs are shaded grey, while not conserved genes or TEs are shaded black for both panels. C.) The number of proximal TE insertions categorized by distance from gene. Each proximal insertion is classified as shared or nonshared for each pairwise genome comparison.



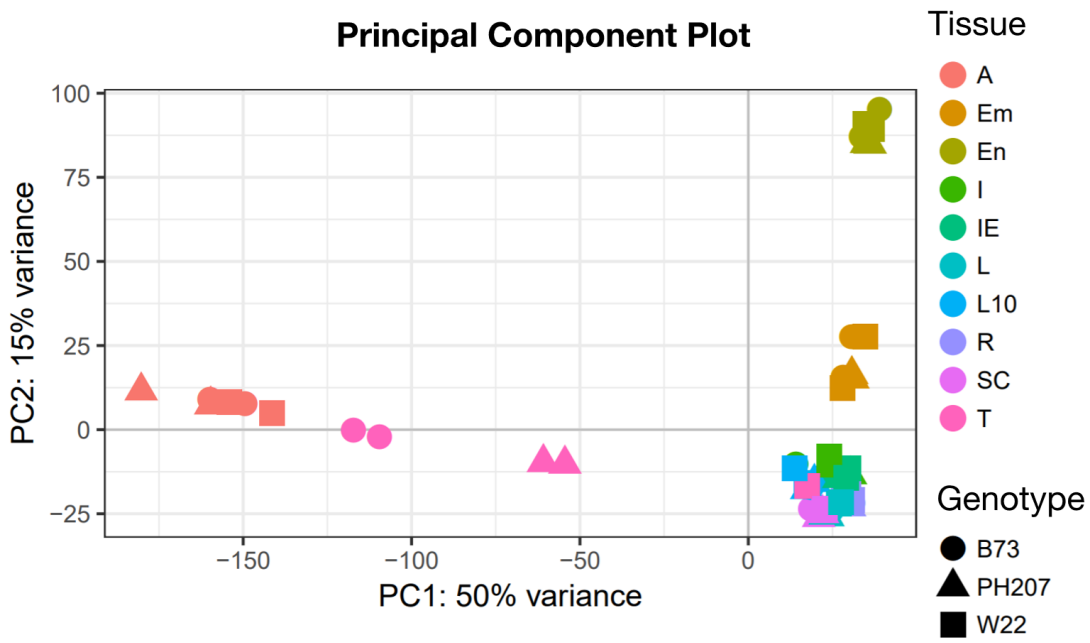
Supplemental Figure 2. Proportion of DE genes in each tissue.



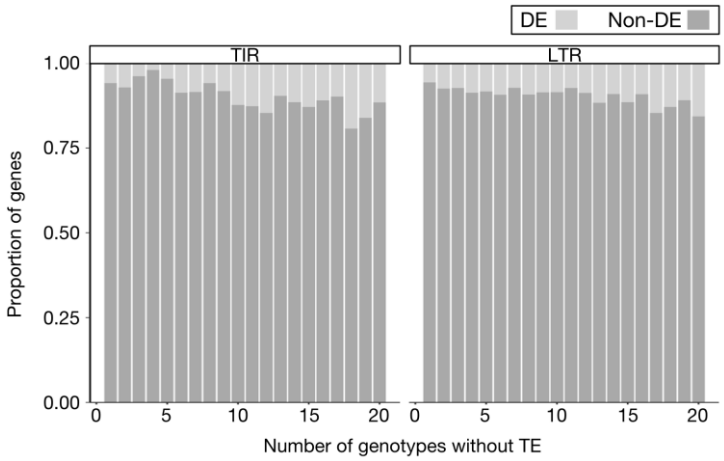
Supplemental Figure 3. Relationship of DE status with genes containing a single proximal TE and multiple proximal TEs.



Supplemental Figure 4. Relationship of proportion of DE genes with TE distance to gene.



Supplemental Figure 5. Principle component plot of RNA-seq data. Samples are clustered primarily by tissue and reps for the same genotype cluster close together confirming the high-quality of the RNA-seq data. Meiotic tassel (T) was removed from the analysis due to high variability of this tissue.



Supplemental Figure 6. Proportion of DE genes in diversity panel by TE class.

Supplemental Table 1. Number of shared versus nonshared TE insertions for each distance category. TE elements were filtered to those that could be confidently called shared or nonshared. Elements may be represented more than once if proximal to more than one gene, however each TE is represented only once for each TE/proximal gene combination.

	Shared elements	**Nonshared elements**	**Total elements**
Within gene TE elements	21,483	13,401	34,884
1-kb upstream TE elements	30,105	20,286	50,391
1-kb downstream TE elements	26,532	20,286	46,818
5-kb upstream TE elements	76,995	65,223	142,218
5-kb downstream TE elements	70,227	60,345	130,572

Supplemental Table 2. Proportions of DE and non-DE genes across genes with multiple proximal TEs and a single proximal TE

Distance category	DE, Multiple	Non-DE Multiple	DE, Single	Non-DE Single
Within gene	4,148	34,687	400	3,020
1-kb upstream	6,202	54,215	1,558	12,896
1-kb downstream	5,874	49,683	1,257	10,353
5-kb upstream	13,717	117,089	3,394	28,835
5-kb downstream	13,319	110,629	2,984	25,060

Supplemental Table 3. Relationship between allele bias and presence of a nonshared proximal TE for each consistent status.

Consistent Status	Has nonshared TE	No nonshared TE
Not expressed	0.22	0.21
Inconsistent	0.51	0.57
Consistent, B73	0.15	0.11
Consistent, W22	0.11	0.11