Gene Evolutionary Trajectories and GC Patterns Driven by Recombination in Zea mays

Anitha Sundararajan1, Stefanie Dukowic-Schulze2, Madeline Kwicklis1, Kayla Engstrom1, Nathan Garcia1, Oliver J. Oviedo1, Thiruvarangan Ramaraj1, Michael D. Gonzales1, Yan He1, Minghui Wang4, Qi Sun4, Jaroslaw Pillardy4, Shahryar F. Kianian5, Wojciech P. Pawlowski3, Changbin Chen2 and Joann Mudge1*

1 National Center for Genome Resources, Santa Fe, NM, USA, 2 Department of Horticultural Science, University of Minnesota, St. Paul, MN, USA, 3 Section of Plant Biology, School of Integrative Plant Science, Cornell University, Ithaca, NY, USA, 4 Biotechnology Resource Center Bioinformatics Facility, Cornell University, Ithaca, NY, USA, 5 Cereal Disease Laboratory, United States Department of Agriculture – Agricultural Research Service, St. Paul, MN, USA

Recombination occurring during meiosis is critical for creating genetic variation and plays an essential role in plant evolution. In addition to creating novel gene combinations, recombination can affect genome structure through altering GC patterns. In maize (Zea mays) and other grasses, another intriguing GC pattern exists. Maize genes show a bimodal GC content distribution that has been attributed to nucleotide bias in the third, or wobble, position of the codon. Recombination may be an underlying driving force given that recombination sites are often associated with high GC content. Here we explore the relationship between recombination and genomic GC patterns by comparing GC gene content at each of the three codon positions (GC1, GC2, and GC3, collectively termed GCx) to instances of a variable GC-rich motif that underlies double strand break (DSB) hotspots and to meiotic gene expression. Surprisingly, GCx bimodality in maize cannot be fully explained by the codon wobble hypothesis. High GCx genes show a strong overlap with the DSB hotspot motif, possibly providing a mechanism for the high evolutionary rates seen in these genes. On the other hand, genes that are turned on in meiosis (early prophase I) are biased against both high GCx genes and genes with the DSB hotspot motif, possibly allowing important meiotic genes to avoid DSBs. Our data suggests a strong link between the GC-rich motif underlying DSB hotspots and high GCx genes.

Keywords: recombination, GC, meiosis, meiocytes, maize, codon usage, wobble, gene expression

INTRODUCTION

In eukaryotes, meiotic exchange of genetic information, or recombination, between homologous chromosomes is a critical step in generating genetic diversity required for adaptation. Recombination is also a crucial tool in plant improvement efforts. Local genome architecture is sculpted by the recombination process, and genome architecture, in turn, drives recombination. This interplay helps to create variability in genomic space, defining relatively stable and plastic genomic regions. This fluctuation in genomic stability is critical for balancing adaptation and stability on the phenotypic level.
Recombination has direct implications for GC patterns and *vice versa*. GC content refers to the percentage of guanine and cytosine bases in a DNA sequence, as opposed to adenine and thymidine bases. There have been many studies substantiating the positive correlation between recombination and GC content (Ikemura and Wada, 1991; Eyre-Walker, 1993; Fullerton et al., 2001; Galtier et al., 2001; Marais et al., 2001; Duret and Arndt, 2008; Haudry et al., 2008; Escobar et al., 2010; Muyle et al., 2011). Crossovers have been found to be correlated with high GC content in rat, mouse, human, zebrafish, bee, and maize at a broad scale (Jensen-Seaman et al., 2004; Beve et al., 2006; Gore et al., 2009; Backstrom et al., 2010; Giraut et al., 2011), while other studies detected strong correlation only at a fine scale (~5 kb for yeast, ~15-128 kb for human) and rather weak correlation at a broad scale (~30 kb for yeast, ~1 Mb for human; Gerton et al., 2000; Myers et al., 2006; Marsolier-Kergoat and Yeramian, 2009).

In plants, correlation of recombination and GC content has been demonstrated in multiple species (Haudry et al., 2008; Escobar et al., 2010; Muyle et al., 2011). A study examining three different grasses (rice, maize, and *Brachypodium*) revealed significant correlation of the local recombination rate with high GC content, especially in the wobble codon position (third position in the codon; Serres-Giardini et al., 2012). This was in contrast to prior studies performed with lower levels of resolution, which found only weak correlation in maize and none in rice or *Brachypodium* (Gore et al., 2009; Tian et al., 2009; Huo et al., 2011). However, negative correlation between crossovers and high GC content was reported for *Arabidopsis* (Drouaud et al., 2006) in spite of the fact that a crossover motif has been identified that has high GC content every third nucleotide (Wijnker et al., 2013).

While most studies agreed on a positive correlation between recombination and high GC content, the cause for this has been disputed (Duret and Arndt, 2008; Marsolier-Kergoat and Yeramian, 2009; Giraut et al., 2011). Possible reasons suggested for the high GC/crossover correlation include selection on codon production is optimized (Gouy and Gautier, 1982; Sharp and Mrazek, 1996; Sueoka and Kawanishi, 2000). By now, it has been well documented that synonymous codon usage varies significantly among genomes and, in addition, among different genes within any given genome (Grantham et al., 1980; Sharp et al., 1988; Wang and Hickey, 2007). One implication of higher GC content from silent codon positions can be a general increased expression due to higher stability of mRNA, as demonstrated in mammals (Kudla et al., 2006). Intra- and inter-genomic deviations in codon usage can be attributed to many factors. In the case of prokaryotes and unicellular organisms, it is thought to be a result of natural selection during which protein production is optimized (Gouy and Gautier, 1982; Sharp and Li, 1986; Shields et al., 1988; Sharp et al., 2005). Likewise, in some higher eukaryotes, there is also evidence that codon bias may occur due to selection for translational efficiency (Shields et al., 1988; Stenico et al., 1994). Particularly, there is a positive correlation between the complementarity of the codons in highly expressed genes and the anticodons of the most abundant tRNAs (Ikemura, 1981; Moriyama and Powell, 1997; Kanaya et al., 1999).

Numerous studies have been conducted on codon usage in plant species including the model plants *Arabidopsis thaliana*, rice (*Oryza sativa*), and a moss (*Physcomitrella patens*), as well as in poplar, citrus and true grass (Poaceae/Graminaceae) species (Chiapello et al., 1998; Liu et al., 2003; Guo et al., 2007; Ingvarsson, 2008; Xu et al., 2013). In addition to the reported factors involved in codon usage bias, higher average GC content in monocot compared to dicot genomes plays an important role in dictating codon usage differences between the two groups of plants (Fenoy and Bailey-Serres, 1993; Carels and Bernardi, 2000). These compositional variations are illustrated by an abundance of genes with high GC levels in the third codon position in monocots (Matassi et al., 1989). The third
position in the codon, also referred to as the wobble position, is less biased for the amino acid than the other two bases and accounts for most of the degeneracy of the genetic code. The wobble position functions as a marker for GC richness, and the frequency of GC nucleotides at the third position is defined as GC\(_3\) (Chiapello et al., 1998; Tatarinova et al., 2010; Elhaik et al., 2014).

In some organisms, two classes of genes can be distinguished by their high or low GC\(_3\) content (Carels and Bernardi, 2000; Tatarinova et al., 2010). This trait is thought to be ancestral to monocots, with some lineages losing the bimodal GC\(_3\) distribution (Clément et al., 2014). These two GC\(_3\) classes in monocots experience divergent evolutionary pressure and contain different functional categories of genes. High GC\(_3\) genes within monocots were mainly categorized into functions involving electron transport or energy pathways, response to biotic and abiotic stressors and signal transduction (Carels and Bernardi, 2000; Tatarinova et al., 2010). High GC\(_3\) genes can be turned on quickly and experience accelerated evolution (Tatarinova et al., 2010).

Kawabe and Miyashita (2003) went beyond GC\(_3\) content and described GC content in the first, second, and third codon positions for seven plant species, including monocots and dicots, with some lineages losing the bimodal GC\(_3\) distribution (Clément et al., 2014). These two GC\(_3\) classes in monocots experience divergent evolutionary pressure and contain different functional categories of genes. High GC\(_3\) genes within monocots were mainly categorized into functions involving electron transport or energy pathways, response to biotic and abiotic stressors and signal transduction (Carels and Bernardi, 2000; Tatarinova et al., 2010). High GC\(_3\) genes can be turned on quickly and experience accelerated evolution (Tatarinova et al., 2010).

Recombination occurs during meiosis, specifically during prophase I when DSBs are formed and homologous chromosomes pair and recombine (Padmore et al., 1991; Sheehan and Pawlowski, 2012; Dukowic-Schulze et al., 2014b). While DSBs initiate recombination and are a necessary prerequisite, most DSBs will not be resolved into crossovers. Two Arabidopsis motifs associated with crossovers have been identified, including one that showed high GC content every three nucleotides (Wijnker et al., 2013). In maize, a previously identified variable motif underlying genic DSB hotspots is GC-rich and also shows high GC periodicity every three nucleotides, reminiscent of GC periodicity within the codon (He et al., in review).

In this study, we used maize as a model to better understand the relationship between genome architecture and recombination. We examine the interplay between genome evolution (including divergent evolutionary trajectories within a single genome), GC patterns, and recombination initiation in maize. Specifically, we address whether the GC-rich, three nucleotide-periodic motif underlying DSB hotspots in maize correlates with GC\(_3\) or other codon-driven GC patterns. In addition, we address how meiotic genes fit into the DSB and GC landscapes. Concurrently, we extend present knowledge of GC\(_1\), GC\(_2\), and GC\(_3\), collectively termed GC\(_x\), in Zea mays, by examining the relationships among them. In short, we aim to learn whether DSB-associated motifs with high GC content, particularly at every third nucleotide, could be the driving force behind bimodal GC patterns that split the maize genome into labile and stable evolutionary trajectories.

**MATERIALS AND METHODS**

**Reference Genome**

All analyses used the B73 maize reference genome version RefGen_v2 to match previous expression analysis. The filtered
gene set (annotation set 5b, gff format) was used for gene annotation.

**Differential Expression Analysis**
Samples, sequence, and differential expression analyses are described in Dukowic-Schulze et al. (2014a,b,c). For this analysis, a less stringent dataset was used with a significance cutoff for calling differential expression increased from \( p = 0.01 \) to \( p = 0.05 \).

**Double Strand Break Hotspots**
Using ChIP-seq with antibodies against the RAD51 protein as described in He et al. (2013), the DSB hotspot motif was identified with the sequence GVSGRSGNSGRSGVSGRSG (He et al., in review). The motif was identified from \( \sim 900 \) genic hotspot regions that did not contain transposable elements. Copies of the motif were identified using the rGADEM package (Li, 2009) to re-scan these genic hotspot regions for matches to the position weight matrix of the motif using a stringency of \( 80\% \).

**GC Calculations**
GC, GC\(_1\), GC\(_2\), and GC\(_3\) were calculated using custom Perl scripts. For GC\(_1\), GC\(_2\), and GC\(_3\), calculations for each gene were performed on the sequence that contributes to the protein (coding domain sequences (CDSs), and redundancies removed where CDSs overlapped. The phase of each CDS, defined as the number of nucleotides that need be removed from the beginning of the CDS to find the first base of the next codon, was taken into account. GC\(_1\) represents the GC content of the first nucleotides, GC\(_2\) the content of the second nucleotides, and GC\(_3\) the content of the third nucleotides of all codons in a gene. Genic GC was calculated for exons only (CDSs) as well as for exons together with introns in the pre-mRNA.

**Pathway Enrichment Analysis**
agriGO was used to perform gene ontology (GO) enrichment studies (Du et al., 2010) using singular enrichment analysis to identify enrichment compared to the \( Z. mays \) reference. Advanced statistical options include Fisher's exact test and, in order to perform multi-comparison adjustment with the large input dataset, the Benjamini–Hochberg correction method (Benjamini and Hochberg, 1995). A significance value of 0.05 was used to obtain lists of enriched GO terms unless the input gene list was large, in which case we focused on the most significant terms (\( p = 0.01 \)). This did not alter the nature of the functionalities that were enriched for within the analyses. In order to consolidate the large list of GO terms, REVIGO was used (Supek et al., 2011). REVIGO uses a simple hierarchical clustering procedure to remove redundant terms, summarize related terms, and visualize the final set of GO terms.

**Plotting and Statistical Analyses**
Plotting was done in R Statistical Package 3.2.0 and two-sided chi-square tests performed in Microsoft Excel v 14.6.4.

**RESULTS**

**GC Patterns in Maize Genes Show Bimodal Peaks with a Strong Bias in the Third Codon Position**
We examined the GC content of maize genes and their CDSs (Figure 2). The GC content of maize genes shows a bimodal peak, indicating that there are two classes of genes in the maize genome that are differentiated by GC content. This matches previous observations (Duret et al., 1995; Carels and Bernardi, 2000; Lescot et al., 2008; Paterson et al., 2009) and hold true both when calculated across genes, including introns (Figure 2A), and when calculated just across CDSs (Figure 2B). Including introns and untranslated regions (UTRs; Figure 2A) appears to bias genes toward the lower GC content class compared to the pattern using only CDS, though both classes are clearly visible. This suggests that the high GC content is maintained preferentially in the CDS regions. As discussed above, a strong bias in GC content at the third codon position has been shown to contribute to the bimodality of maize genic GC content but bimodality has been shown for all three codon positions. We took a comprehensive approach and examined GC content in the first (GC\(_1\)), second (GC\(_2\)), and third (GC\(_3\)) codon positions, collectively termed GC\(_x\).

GC\(_1\), GC\(_2\), and GC\(_3\) were calculated for all genes and the distribution plotted (Figure 3). A bimodal distribution was easily apparent for GC\(_1\) and GC\(_3\) (GC content peaks around 50–55% and 90–95%) where both GC content classes had higher GC content than the overall genic GC content (see Figures 2 and 3), regarding both the intron-inclusive (40–45% and around 60%) and the intron-exclusive CDS analysis (∼50% and 70–75%; Figure 3). In contrast, GC\(_2\) analysis showed had only one discernible peak of far lower GC content (around 45%), with a long tail to the right, with possibly a very shallow second peak showing high GC content in the same range as that seen in GC\(_1\) and GC\(_3\). As expected, the peak containing high GC content genes was most pronounced in the third position. A cutoff of \( \geq 80\% \) (Tatarinova et al., 2010) was used for all GC\(_x\) conditions to identify the class of high GC\(_x\) content genes among the 39,656 genes present in \( Z. mays \) (Table 1). Among the high GC\(_x\) content classes, the high GC\(_3\) content class had the most genes at 5,719, followed by the high GC\(_1\) content class at 3,647 genes. GC\(_2\) had the fewest with 629 genes (Table 1). There was almost no overlap between each of the high GC\(_x\) classes (Figure 4).

**GC Patterns in Genic Double Strand Break Hotspot Motifs Show High GC Every Third Nucleotide**
To gain insight on patterns possibly shaping bimodality and GC\(_x\) classes, we looked at copies of the DSB hotspot motif (hereafter referred to as “motifs”) in genic regions (He et al., in review). Of 3,423 DSB genic hotspot motifs, 2,143 fell into 544 genes. Of these, 1,909 motifs fell into 811 CDSs of 511 genes, and the remaining 234 motifs occurred in introns and 5′ and 3′
UTR regions. Maize genes are, on average, less than 50% coding sequence (1,153.7 nt of intron sequence for every 1 kb of CDS), yet 89% of the motifs occurred in the CDS. This indicates a strong preference for copies of the DSB hotspot motifs to occur in the protein-coding portion of the gene (chi-square p-value = 0). Of the 544 genes containing hotspot motifs, 84% contained more than one motif instance indicating the propensity for them to cluster (Table 2).

GC content of the hotspot motif (GVSGRSG-NSGRSG) demonstrates a 3 nt-based periodicity (He et al., in review). While none of the positions are perfectly conserved, the G's at every third nucleotide have information content values that are much higher than at any other residue and range from approximately 1.2 to 1.7. Indeed, every third position starting with the first nucleotide is nearly always “G.” We determined GC content across positions 1–19 of all underlying motifs. GC content per nucleotide position, averaged across all motif occurrences, was calculated for all hotspot motifs as well as only those hotspot motifs falling into genes. Every third nucleotide was grouped into a separate periodicity group (i.e., group 1 = nucleotide positions 1, 4, 7, 10, 13, 16, 19; group 2 = nucleotide positions 2, 5, 8, 11, 14, 17; group 3 = nucleotide positions 3, 6, 9, 12, 15, 18). These groups are framed relative to the start position of the motif and not necessarily to the coding frame of the containing gene (Figures 5A,B). Pairwise differences between GC content of these groups all showed significant differences (Table 3).

Because the periodic groups of the DSB hotspot motifs are not necessarily in frame with a gene’s coding frame, we analyzed GCx content of the motifs by identifying its coding frame and adjusting it from motif coordinates (relating to the beginning of the motif) to GCx space (relating to the coding frame of the gene that contains the motif). Indeed, motifs fell into all three possible coding frames, with 868 matching the gene’s coding frame (i.e., the first base of the motif is also the first position of a codon), 389 starting on the second position of the codon, and 1,094 starting on the third position of the codon. Nucleotides within motifs were then placed into GCx groups based on their position within
FIGURE 4 | Comparison of GCx distributions. (A) GC1 vs. GC2. (B) GC1 vs. GC3. (C) GC2 vs. GC3. Blue lines show the 80% cutoff separating the high and low GCx classes.

TABLE 2 | Frequencies of double strand break hotspot motifs per gene.

| Motifs per Gene | Number of Genes |
|-----------------|-----------------|
| 1               | 85              |
| 2               | 107             |
| 3               | 92              |
| 4               | 75              |
| 5               | 58              |
| 6               | 45              |
| 7               | 32              |
| 8               | 20              |
| 9               | 13              |
| 10              | 5               |
| 11              | 6               |
| 12              | 5               |
| 15              | 1               |
| 16              | 1               |

the codon rather than in the motif. The GC2 group within motifs showed significantly higher GC content than either the GC1 or GC3 groups and was the only group that overlapped the high GCx range (≥80%; Figure 5C). This is particularly interesting when considering that genes with high GC2 content are far less frequent than those with high GC1 or GC3 (Figure 3). Given these patterns, we tested whether high GCx on a gene level is correlated with high GCx of the contained motif. Genes containing DSB hotspot motifs were indeed significantly overrepresented for high GCx genes (Table 4). Around 70% (380/544) of the genes containing DSB hotspot motifs fall into at least one peak of high GCx content (i.e., GC1, GC2, and/or GC3). Of these, only three had high GCx content in more than one category, and only one had high GCx content in all three. The number of DSBs that fall into each of the high GCx categories mirrors the size of each of the high GCx content peaks. Twenty-four percent of genes containing DSB hotspots have high GC1 content, 5% have high GC2 content, and 42% have high GC3. When looking at the GCx content of individual DSB hotspot motifs, as opposed to looking at them collectively as done above, a large fraction of the motifs are GC-rich (≥80%), including many with 100% GC content (Figure 6). Indeed, more than a quarter of motifs have 100% GC1 and GC3 content and more than half have 100% GC2 content. It is thus not surprising that genes containing the motifs also tend to be have high GCx content.

Genes with High and Low GCx Content Show Functional Differentiation

Among the 39,656 genes annotated in the maize reference genome, 9,861 genes (25%) are in a high GCx content peak. However, there is very little overlap between high GCx genes in the GC1, GC2, and GC3 categories (Figure 4). Indeed, overlap rates are significantly less than expected given random sampling (Table 5). The lack of overlap is especially stark between the high GC content peaks of GC1 and GC3 (Figure 4B) where very few genes simultaneously have high GC1 and high GC3 content (≥80%), while many genes have a strong positive correlation when being of both low GC1 and GC3 content (<80%). Of 526 genes that would be expected to fall into the high GC3 content peak given random sampling of 3,647 high GC1 genes from the maize genome, only 96 (18% of expectation) fall into the high GC3 content peak. Furthermore, most of the genes that do show high GC1 and GC3 content are limited to GCx contents close to the 80% cutoff in one or both GCx categories, indicating they may be part of the upper tail of the lower GCx content peak that merges into and is indistinguishable from the high GCx content.
peak (Figure 3). Taken together, there seems to be a selection against high GCx content in more than one codon position.

When genes in the low GC3 class were subjected to agriGO analysis, housekeeping pathways were enriched, such as genes involved in cellular processes, metabolic processes, protein metabolism, protein modifications, biosynthetic pathways, localization, and transport. Genes responsible for transcription and transcriptional regulation processes were also categorized as low GC3. Genes in the high GC3 class fell into functions implicated in the immune system and stress responses including adaptive immunity and response to wounding. Genes involved in sexual reproduction were also enriched in the high GC3 class though this did not appear to translate specifically to enrichment of meiotic recombination genes. In short, our results indicate more generalized or housekeeping functions for low GC3 genes, SOS-type functions for high GC3 genes, and are very similar to those seen in other studies (Carels and Bernardi, 2000; Tatarinova et al., 2010).

In the case of GC content of the first nucleotide of each codon, 36,009 genes were categorized with low GC1 (<80%) and 3,648 genes with high (≥80%) GC1 content. GO analysis on GC1 gene classes showed similar trends to the GC3 classes. Low GC1 genes described functions such as cellular protein modification process, protein localization, macromolecule metabolism (primary and cellular metabolism), including lipid and phosphorus metabolism, gene expression, and cell death. High GC1 genes were implicated in functions such as response to stress, DNA packaging, regulation of biological quality, adaptive immune response, fatty acid metabolism, and response to stimulus.

When all genes in Z. mays were examined for the GC content in the second nucleotide of each codon, even though there were significantly fewer genes with high GC content compared to high GC1 and GC3 content, the general trends observed in GC1 and GC3 were seen in GC2 for the low (39,027 genes) and high (630 genes) GC2 classes with some subtle exceptions. To elaborate, low GC2 genes were implicated in ontologies such as protein metabolism, localization, RNA metabolism, lipid metabolism, and intracellular transport. The 630 high GC2 genes, however, were categorized into different ontologies compared to the high GC1 and high GC3 gene sets. High GC2 genes were involved in cell-wall organization, transmembrane transport, lipid, carbohydrate, and phosphorus metabolism, G-coupled signaling pathways, protein phosphorylation, post-translational modification, sexual reproduction, and localization type functions. The low number of genes in the high GC2 class may have lowered our power to detect some of the classes seen in high GC1 and GC3 genes.

Overall, high GCx genes tend to play a role in stress and adaptive responses as well as sexual reproduction. The low GCx genes tend to be biased toward more generalized functions.

**Genes Up-Regulated in Meiocytes Are Underrepresented for DSB Motifs**

To study correlations between DSB motif presence, GCx patterns and gene expression, we examined genes that contained DSB hotspot motifs regarding their expression level in meiocytes compared to anthers or seedlings (Table 6). Genes with DSB hotspot motifs were down-regulated in meiocytes at a rate expected given independence of down-regulated genes and motifs. However, meiocyte up-regulated expression of DSB motifs.
TABLE 3 | Comparison of GC content between 3 nt-based periodic nucleotide groups of DSB motifs within DSB hotspots and genic DSB hotspots.

| Comparisons | F-test for equal variance | Equal variance (homoscedastic), two-tailed t-tests | Unequal variance (heteroscedastic), two-tailed t-tests |
|-------------|---------------------------|--------------------------------------------------|--------------------------------------------------|
| All motifs (position within motif) | nt1 vs. nt2 | 0.06315758 | 0.141343E-05 | 0.000426348 |
| | nt2 vs. nt3 | 0.548540329 | 3.77664E-06 | 0.000329802 |
| | nt1 vs. nt3 | 0.016068929 | 5.4902E-06 | 0.38307282 |
| Genic motifs (position within motif) | nt1 vs. nt2 | 0.012207227 | 1.7E-06 | 0.000099458 |
| | nt2 vs. nt3 | 0.596027462 | 1.87802E-73 | 0.000329802 |
| | nt1 vs. nt3 | 0.003212561 | 0.000426348 | 0.00329802 |
| Genic motifs (position within codon) | GC1 vs. GC2 | 0.348471769 | 0.000199458 | 0.000426348 |
| | GC2 vs. GC3 | 0.288858144 | 2.97987E-11 | 0.000329802 |
| | GC1 vs. GC3 | 0.058998065 | 1.7E-06 | 0.38307282 |

TABLE 4 | Chi-square test to determine whether GCx high and low categorization is randomly distributed within the 544 DSB motif-containing genes.

| High GCx observed (expected) | Low GCx observed (expected) | p-value |
|------------------------------|-----------------------------|--------|
| GC1 130 (50) | 414 (494) | 1.79998E-32 |
| GC2 28 (9) | 516 (535) | 2.97987E-11 |
| GC3 227 (78) | 317 (466) | 1.87802E-73 |

An F-test was performed to test the hypothesis of equal variances followed by one of two t-tests based on the outcome of the F-test. See the legend for Figure 5 for a detailed description of the groups. All comparisons were done either as a homoscedastic or a heteroscedastic t-test, depending on the outcome of the F-test.

DISCUSSION

We explored the correlations of genes with instances of a motif found at meiotic DSB hotspots in maize with GC patterns as well as meiotic expression levels. GC-rich motifs with three-nucleotide periodicity underlie hotspots of DSBs that are precursors to recombination between homologous chromosomes. We looked at the genomic landscape of these motifs, including motifs in regions outside of the DSB hotspots that were identified in our dataset. While most DSB hotspots occurred near the transcription start or stop sites, GCx content is, by definition, within the codons. Nevertheless, genes tend to be either high or low GCx as a whole though there is a slight increase in GC content in the 5′–3′ direction (Tatarinova et al., 2010) and GCx determination included bases between the transcription start and stop sites.

These DSB hotspot motifs tend to occur in genes that frequently have high GC1, GC2, or GC3 content. High GC1, GC2, and GC3 content genes are negatively correlated with each other, indicating that the extremely high GC content (≥80%) that occurs in genes with high GCx almost exclusively derive from only one frame.

To incorporate meiotic gene expression, we sequenced RNA from meiocytes (the cells where meiosis takes place) that were just beginning the meiotic process. Specifically, the cells were in the early stages of prophase I, which itself consists of five sequential stages (leptotene, zygotene, pachytene, diplotene, and diakinesis) and is the platform of DSBs and crossovers. Early prophase I includes the formation of DSBs (leptotene), resection and single-end invasion (zygotene), and resolution of some of the DSBs into crossovers (pachytene) and therefore is highly relevant to the generation of novel genetic combinations. Notably, genes that are up-regulated in early meiosis tend to have low GCx and are biased against containing DSB hotspot motifs. Our data thus indicates that genes up-regulated in early meiosis might be protected against DSBs. Our observation that genes that are up-regulated in meiocytes are significantly biased against genes with DSB hotspot motifs makes sense given the environment in early prophase I. Meiotic genes induced during early prophase I, by necessity, need to be genes that are not simultaneously undergoing DSBs, though
we emphasize that the presence of a motif does not necessarily indicate the presence of a recombination hotspot. As a key point, our analysis suggests that DSB formation could be targeted to or avoided in specific functional classes, regulated by the periodic GC-rich motif underlying DSBs.

In general, the low GC\(_x\) gene classes in our dataset were underrepresented for the DSB hotspot motif (Table 4; Figure 7) and showed bias toward housekeeping functions. In contrast, we found genes with high GC\(_x\) that are enriched for more specialized functions, including sexual reproduction, though genes did not appear to be specific to meiotic recombination. Possibly, these latter genes are needed during other stages of development or meiosis, and not around the time when DSBs are generated, but benefit from recombination-driven evolution.

GC bimodality refers to the occurrence of two classes of genes, distinguished by their GC content. They have been reported in monocot but not in dicot plants (Clément et al., 2014). To date, there have been several reports explaining the differences in relative distribution of genes based on GC\(_x\) between monocots and dicots. Additionally, pronounced differences were observed in GC\(_x\) between close relatives in plant families. For example, a comparative study between *A. thaliana*, *Raphanus sativus*, *Brassica rapa*, and *Brassica napus* revealed that the GC\(_x\) values of *R. sativus*, *B. rapa*, and *B. napus* are on an average 5% higher than that of *A. thaliana* orthologs (Villagomez and Kuleck, 2009; Tatarinova et al., 2010). Further, many plants, including grasses are prone to genome duplications, which provides redundancy that can relax selection pressure for individual genes though genome-wide some stabilization must occur after a polyploid event. Nevertheless, the flexibility provided by redundancy that occurs after polyploidization may enable the evolution of genomic regions that have varied recombination rates and GC contents.

Degeneracy in the genetic code in the third codon position gave rise to the wobble hypothesis (Crick, 1966) which has

---

### TABLE 5 | Chi-square test to determine whether other GC\(_x\) classes are randomly distributed within each of the high GC\(_x\) classes.

| High GC\(_x\) observed (expected) | Low GC\(_x\) observed (expected) | p-value |
|----------------------------------|---------------------------------|---------|
| Among high GC\(_1\) genes        |                                 |         |
| GC\(_2\)                         | 26 (58) 3,621 (3,589)           | 2.43E-05|
| GC\(_3\)                         | 96 (526) 3,551 (3,121)          | 2.57E-91|
| Among high GC\(_2\) genes        |                                 |         |
| GC\(_1\)                         | 26 (58) 603 (571)               | 1.11E-05|
| GC\(_3\)                         | 26 (91) 603 (538)               | 2.06E-13|
| Among high GC\(_3\) genes        |                                 |         |
| GC\(_1\)                         | 96 (526) 5,623 (5,193)          | 3.59E-86|
| GC\(_2\)                         | 26 (91) 5,693 (5,628)           | 7.44E-12|

Significant p-values allowing rejection of the null hypothesis (p ≤ 0.05) are in bold.

### TABLE 6 | Chi-square test to determine whether up- or down-regulated genes in meiocytes are randomly distributed within the 544 motif-containing genes.

| Differential expression in meiocytes | Yes | No | p-value |
|-------------------------------------|-----|----|---------|
| Meiocyte-specific expression        |     |    |         |
| Comparison tissue                   |     |    |         |
| Observed (expected)                 |     |    |         |
| Up-regulated                        |     |    |         |
| Anther                              | 2 [3] 542 (541) | 0.712147988 |
| Seedling                            | 18 [44] 526 (500) | 3.57589E-05 |
| Down-regulated                      |     |    |         |
| Anther                              | 2 [3] 542 (541) | 0.506157854 |
| Seedling                            | 62 [57] 482 (487) | 0.465276348 |

Significant p-values allowing rejection of the null hypothesis (p ≤ 0.05) are in bold.
been invoked to explain the bimodal GC$_x$ distribution of genes in maize and other monocots (Tatarinova et al., 2010). In the wobble position, tRNA modifications allow pairing of tRNA with multiple codons, allowing multiple codons to code for a single amino acid (reviewed in Agris et al., 2007). This degeneracy of the genetic code gives freedom, mainly in the third codon, for GC shifts to take place.

While the wobble hypothesis is often only applied to the third codon position, there is degeneracy in the genetic code in the first and second positions of the codon as well, though much less than is seen in the third position (Supplementary Table 1). Given the much larger amount of wobble in the third codon position compared to the first and second, however, it is surprising that the size of the peak with high GC$_x$ content genes is nearly two-thirds of the corresponding high GC$_3$ peak in our dataset of DSB hotspot motif containing genes (Figure 3; Table 1). Further, the amount of wobble in the first and second positions is quite similar (Supplementary Table 1), yet the second position has only 17% the number of genes with high GC$_2$ content compared to GC$_1$ (Figure 3; Table 1). If GC bimodality patterns in maize were enabled simply by the freedom that codon wobble gives for GC shifts, than nearly all of the high GC$_x$ genes would be in the high GC$_3$ category. The fact that nearly half of the high GC$_x$ genes fall into the GC$_1$ and GC$_2$ categories clearly indicates that the wobble hypothesis, on its own, is insufficient for explaining the GC$_x$ bimodality across all three frames.

The ability in many parts of a protein to interchange amino acids with amino acids of similar biochemical properties (Henikoff and Henikoff, 1992) might help explain the discrepancy between the numbers of genes in each high GC$_x$ class and the codon position's corresponding wobble. Substitution of amino acids of similar biochemical properties may allow a genome to shift GC content at all the codon positions. More work needs to be done to test this hypothesis and discover its effect on the varying sizes of the high GC$_x$ classes across all genes but is not in the scope of this work.

The importance of GC$_x$ bimodality is underscored by its maintenance through mutational pressure and selective restraint though the mechanism is yet unknown (Liu et al., 2012). The propensity for recombination to increase GC content (Figure 1) may be a contributing mechanism of positive reinforcement for maintaining genes in high GC$_x$ content and DSB hotspot pools (Galtier et al., 2001). Further, the typically high evolutionary rates associated with high GC$_x$ genes (Tatarinova et al., 2010) could be due to their association with the GC-rich DSB motifs, leading

---

**TABLE 7 | Chi-square test on B73 genes that are expressed in or up- or down-regulated in meiocytes compared to anthers or seedlings.**

| Meiocyte expression | GC$_x$ | Low GC$_x$ observed (expected) | High GC$_x$ observed (expected) | p-value |
|---------------------|-------|-------------------------------|-------------------------------|---------|
| All expressed genes | GC$_1$ | 22,023 (21,981) | 2,184 (2,226) | 0.3477  |
| All expressed genes | GC$_2$ | 23,852 (23,823) | 355 (384) | 0.1363  |
| All expressed genes | GC$_3$ | 20,648 (20,716) | 3,559 (3,491) | 0.2136  |
| Up vs. anthers      | GC$_1$ | 180 (172) | 9 (17) | 0.0349  |
| Up vs. seedlings    | GC$_1$ | 3,002 (2,938) | 234 (298) | 3.7800E-09 |
| Down vs. anthers    | GC$_1$ | 192 (211) | 40 (21) | 8.5100E-27 |
| Down vs. seedlings  | GC$_1$ | 3,560 (3,759) | 580 (381) | 0.5596  |
| Up vs. anthers      | GC$_2$ | 186 (186) | 4 (3) | 3.1000E-02 |
| Up vs. seedlings    | GC$_2$ | 3,200 (3,185) | 36 (51) | 0.7010E-02 |
| Down vs. anthers    | GC$_2$ | 224 (228) | 8 (4) | 1.7100E-08 |
| Down vs. seedlings  | GC$_2$ | 4,029 (4,074) | 111 (66) | 2.3200E-02 |
| Up vs. anthers      | GC$_3$ | 179 (162) | 10 (27) | 0.0004  |
| Up vs. seedlings    | GC$_3$ | 2,855 (2,769) | 381 (467) | 1.7800E-05 |
| Down vs. anthers    | GC$_3$ | 167 (199) | 65 (33) | 3.7800E-09 |
| Down vs. seedlings  | GC$_3$ | 3,117 (3,543) | 1,023 (597) | 3.6400E-79 |

The null hypothesis is that there is no relation between these sets of genes and the high and low GC$_x$ categories. Significant p-values allowing rejection of the null hypothesis (p ≤ 0.05) are in bold.

---

**FIGURE 7** | Overview of findings. High GC$_x$ genes are overrepresented for DSB hotspots and genes down-regulated in meiocytes but underrepresented for genes up-regulated in meiocytes. Individual high GC$_x$ classes (GC$_1$, GC$_2$, and GC$_3$) show less overlap than expected given the number of genes in each class. Genes up-regulated in meiocytes are underrepresented for DSB hotspots.
to increased rates of recombination. Genetic recombination is crucial to an organism’s ability to survive, adapt, and evolve. Genetic recombination occurs through crossing over during meiosis and, furthermore, gene conversion occurs with a frequent GC bias in the case of non-crossovers.

DSBs were assayed using ChIP-seq targeting the RAD51 gene (He et al., 2013), which binds to DSBs during zygotene, resulting in the discovery of a maize DSB hotspot motif (He et al., in review). Given the three nucleotide-based GC periodicity found in the DSB hotspot motifs and the documented GC3 bias found in maize, it was natural to look for correlations between the two. High correlation was found between genes containing DSB hotspot motifs and GC3 levels, both of the motif itself and the containing gene. Interestingly, similar correlations were found for GC1 and GC2 levels. Indeed, 70% of all genic DSB hotspots fall into high GCx genes, a remarkable and significant proportion given that only one-quarter of genes overall fall into the high GCx category. Even more intriguingly, genes with high GC2 motifs make the biggest contribution, although high GC2 genes are far less prevalent than GC1 or GC3 genes (Figures 5C and 6B). Further work needs to be done in order to determine whether there is a direct or secondary relationship between copies of the motif, double-strand breaks, and, by extension, recombination. Nevertheless, this data suggests intriguing hypotheses that imply a role for recombination in maintaining bimodal GCx distributions that cannot fully be explained by the wobble hypothesis.

CONCLUSION

Sexual reproduction coupled with meiotic recombination is an important evolutionary strategy for generating genetic diversity crucial to adaptation and evolution of species. As the precursors to crossovers and non-crossovers, DSBs and their locations are critical to understanding recombination processes and biases. Here we showed that DSB hotspot motifs have a strong tendency to occur in genes with high GCx content. These genes have extremely high GC content in one frame, have high evolutionary rates, and are biased toward SOS-functional classes, implicating support for adaptation in these genes. Further, genes up-regulated in meiocytes have a strong bias against high GCx classes and DSB motifs, indicating the importance of these genes and their acquisition of protective strategies against DSBs. Intriguingly, wobble bases and degeneracy in the genetic code are not sufficient for explaining the high GCx classes and their sizes across all three frames. The presence of the three nucleotide-based periodic GC-rich motifs underlying double strand hotspot motifs may provide a first glimpse at additional selection pressure driving the generation of functionally biased high GCx classes in maize and related organisms. More work needs to be done to determine whether the motif provides recognition for the DSB machinery, thus providing a possible mechanism promoting high GCx content.

AUTHOR CONTRIBUTIONS

AS, SD-S, and JM analyzed the data and wrote the manuscript. MK, KE, OO, NG, TR, and MG helped with data analysis. MW, YH, QS, and WP performed experiments to identify DSB hotspot motifs and provided data. WP and CC edited the manuscript. WP, CC, SK, JP, and JM are investigators on the NSF IOS-1025881 grant of which this manuscript is a part.

FUNDING

This research was supported by the United States National Science Foundation grant IOS-1025881 and by an Institutional Development Award (IDeA) from the National Institute of General Medical Sciences of the National Institutes of Health under grant number P20GM103451.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: http://journal.frontiersin.org/article/10.3389/fpls.2016.01433

SUPPLEMENTAL TABLE 1 | Degeneracy of the genetic code in the first and second codon positions.

REFERENCES

Agris, P. F., Vendeix, F. A., and Graham, W. D. (2007). tRNAs wobble decoding of the genome: 40 years of modification. J. Mol. Biol. 366, 1–13. doi: 10.1016/j.jmb.2006.11.046

Backstrom, N., Forstmeier, W., Schieleth, H., Mellenius, H., Nam, K., Bolund, E., et al. (2010). The recombination landscape of the zebrafish Taeniopygia guttata genome. Genome Res. 20, 485–495. doi: 10.1101/gr.101410.109

Benjamini, Y., and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. J. R. Stat. Soc. 57, 289–300.

Beye, M., Gattermeier, I., Hasselmann, M., Genpe, T., Schioett, M., Baines, J. F., et al. (2006). Exceptionally high levels of recombination across the honey bee genome. Genome Res. 16, 1339–1344. doi: 10.1101/gr.5680406

Birdsell, J. A. (2002). Integrating genomics, bioinformatics, and classical genetics to study the effects of recombination on genome evolution. Mol. Biol. Evol. 19, 1181–1197. doi: 10.1093/oxfordjournals.molbev.a004176

Carrels, N., and Bernardi, G. (2000). Two classes of genes in plants. Genetics 154, 1819–1825.

Chen, C., Farmer, A. D., Langley, R. J., Mudge, J., Crow, J. A., May, G. D., et al. (2010). Meiosis-specific gene discovery in plants: RNA-Seq applied to isolated Arabidopsis male meiocytes. BMC Plant Biol. 10:280. doi: 10.1186/1471-2229-10-280

Chiapello, H., Lisacek, F., Caboche, M., and Henaut, A. (1998). Codon usage and gene function are related in sequences of Arabidopsis thaliana. Gene 209, GC1-GC38. doi: 10.1016/s0378-1119(97)00671-9

Clément, Y., Fustier, M.-A., Naholbol, Z., and Gléméin, S. (2014). The bimodal distribution of genic GC content is ancestral to Monocot species. Genome Biol. Ecol. 7, 336–348. doi: 10.1093/gbe/eva278

Crick, F. H. (1966). Codon–anticodon pairing: the wobble hypothesis. J. Mol. Biol. 19, 548–555. doi: 10.1016/S0022-2836(66)80022-0

Cruciver, S., D’Onofrio, G., and Bernardi, G. (2000). The compositional transition between the genomes of cold- and warm-blooded vertebrates:
Marsolier-Kergoat, M. C., and Yeramian, E. (2009). GC content and recombination: reassessing the causal effects for the *Saccharomyces cerevisiae* genome. *Genetics* 183, 31–38. doi: 10.1534/genetics.109.105049

Matassi, G., Montero, L. M., Salinas, J., and Bernardi, G. (1989). The isochore organization and the compositional distribution of homologous coding sequences in the nuclear genome of plants. *Nucleic Acids Res.* 17, 5273–5290. doi: 10.1093/nar/17.13.5273

Moriyama, E. N., and Powell, J. R. (1997). Codon usage bias and tRNA abundance in *Drosophila*. *J. Mol. Evol.* 45, 514–523. doi: 10.1007/P00006256

Mugy, A., Serres-Giardi, L., Ressayre, A., Escobar, J., and Glemin, S. (2011). GC-biased gene conversion and selection affect GC content in the *Oryza* genus (rice). *Mol. Biol. Evol.* 28, 2695–2706. doi: 10.1093/molbev/msr104

Myers, S., Spencer, C. C., Auton, A., Bottolo, L., Freeman, C., Donnelly, P., et al. (2006). The distribution and causes of meiotic recombination in the human genome. *Biochem. Soc. Trans.* 34, 526–530. doi: 10.1042/BST0340526

Nabholz, B., Kunstner, A., Wang, R., Jarvis, E. D., and Ellegren, H. (2011). Dynamic evolution of base composition: causes and consequences in avian phylogenomics. *Mol. Biol. Evol.* 28, 2197–2210. doi: 10.1093/molbev/msr047

Padmore, R., Cao, L., and Kleckner, N. (1991). Temporal comparison of recombination and synaptonemal complex formation during meiosis in *S. cerevisiae*. *Cell* 66, 1239–1256. doi: 10.1016/0092-8674(91)90046-2

Paterson, A. H., Bowers, J. E., Feltus, F. A., Tang, H., Lin, L., and Wang, X. (2009). Comparative genomics of grasses promises a bountiful harvest. *Plant Physiol.* 149, 125–131. doi: 10.1104/pp.108.129262

Serres-Giardi, L., Belkhir, K., David, J., and Glemin, S. (2012). Patterns and evolution of nucleotide landscapes in seed plants. *Plant Cell* 24, 1379–1397. doi: 10.1105/tpc.112.1093674

Sharp, P. M., Bailes, E., Grocock, R. J., Peden, J. F., and Socke, R. E. (2005). Variation in the strength of selected codon usage bias among bacteria. *Nucleic Acids Res.* 33, 1141–1153. doi: 10.1093/nar/gki242

Sharp, P. M., Cowe, E., Higgins, D. G., Shields, D. C., Wolfe, K. H., and Wright, F. (1988). Codon usage patterns in *Escherichia coli*, *Bacillus subtilis*, *Saccharomyces cerevisiae*, *Schizosaccharomyces pombe*, *Drosophila melanogaster* and *Homo sapiens*: a review of the considerable within-species diversity. *Nucleic Acids Res.* 16, 8207–8211. doi: 10.1093/nar/16.17.8207

Sharp, P. M., and Li, W. H. (1986). An evolutionary perspective on synonymous codon usage in unicellular organisms. *J. Mol. Evol.* 24, 28–38. doi: 10.1007/BF02099948

Sharp, P. M., and Matassi, G. (1994). Codon usage and genome evolution. *Curr. Opin. Genet. Dev.* 4, 851–860. doi: 10.1016/0959-437X(94)90070-1

Sheehan, M. J., and Pawlowski, W. P. (2012). Imaging chromosome dynamics in meiosis in plants. *Methods Enzymol.* 505, 125–143. doi: 10.1016/B978-0-12-388448-0.00015-2

Shields, D. C., Sharp, P. M., Higgins, D. G., and Wright, F. (1988). "Silent" sites in *Drosophila* genes are not neutral: evidence of selection among synonymous codons. *Mol. Biol. Evol.* 5, 704–716.

Stenico, M., Lloyd, A. T., and Sharp, P. M. (1994). Codon usage in *Caenorhabditis elegans*: delineation of translational selection and mutational biases. *Nucleic Acids Res.* 22, 2437–2446. doi: 10.1093/nar/22.13.2437

Sueoka, N., and Kawanishi, Y. (2000). DNA G++C content of the third codon position and codon usage biases of human genes. *Gene* 261, 53–62. doi: 10.1016/S0378-1119(00)00480-7

Supek, F., Bosnjak, M., Skunca, N., and Smuc, T. (2011). REVIGO summarizes and visualizes long lists of gene ontology terms. *PLoS ONE* 6:e21800. doi: 10.1371/journal.pone.0021800

Tatarinova, T. V., Alexandrov, N. N., Bouck, J. B., and Feldmann, K. A. (2010). GC3 biology in corn, rice, sorghum and other grasses. *BMC Genomics* 11:308. doi: 10.1186/1471-2164-11-308

Tian, Z., Rizzon, C., Du, J., Zhu, L., Bennetzen, J. L., Jackson, S. A., et al. (2009). Do genetic recombination and gene density shape the pattern of DNA elimination in rice long terminal repeat retrotransposons? *Genome Res.* 19, 2221–2230. doi: 10.1101/gr.083899.108

Villagoméz, L. T. T. V., and Kuleck, G. (2009). Ecological Genomics: Construction of Molecular Pathways Responsible for Gene Regulation, and Adaptation to Heavy Metal Stress in *Arabidopsis thaliana*, and *Raphanus sativus*. Stockholm: ISMB/ECB.

Wang, H. C., and Hickey, D. A. (2007). Rapid divergence of codon usage patterns within the rice genome. *BMC Evol. Biol.* 7(Suppl 1):S6. doi: 10.1186/1471-2148-7-S1-S6

Webster, M. T., Axelsson, E., and Ellegren, H. (2006). Strong regional biases in nucleotide substitution in the chicken genome. *Mol. Biol. Evol.* 23, 1203–1216. doi: 10.1093/molbev/msk088

Winjker, E., Velikikakam James, G., Ding, J., Becker, F., Klasen, J. R., Rawat, V., et al. (2013). The genomic landscape of meiotic crossovers and gene conversions in *Arabidopsis thaliana*. *Elife* 2:e01426. doi: 10.7554/elife.01426

Xu, C., Dong, J., Tong, C., Gong, X., Wen, Q., and Zhuge, Q. (2013). Analysis of synonymous codon usage patterns in seven different citrus species. *Evol. Bioinform.* 9, 215–228. doi: 10.4137/EBO.S11930

**Conflict of Interest Statement:** The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2016 Sundararajan, Dukovic-Schulze, Kwicklis, Engstrom, Garcia, Oviedo, Ramaraj, Gonzales, He, Wang, Sun, Pillardy, Kianian, Pawlowski, Chen and Mudge. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.