The Association of Indels with Meiotic Recombination Sites in Maize

Presented to the College of Arts and Sciences, Cornell University

in Partial Fulfillment of the Requirements for the Biological Sciences Honors Program

By Nikita Sajai

November 2022

Advisor: Wojtek Pawlowski

Abstract

The processes that occur during meiotic recombination, from the initiation of DNA double stand breaks (DSBs) to the completion of crossing over (CO) events, have numerous opportunities for inaccuracy. Occurrences of double DSBs (dDSBs) and other missteps during the process could create mutation in the genome. In this study, I assessed the indel generation potential of meiotic recombination in maize by mapping indels of various sizes to different sets of recombination-sites. I found that there is an enrichment of small indels (1-50 bp) at these sites, providing the first strong evidence of mutagenicity of meiotic recombination in plants.

I. Introduction

Crossover (CO) Formation

Meiotic recombination is the process of exchanging genetic material between homologous chromosomes. This process ensures maximum genetic diversity by exchanging genetic material from both parents which segregates out into progeny. Recombination events are initiated during the leptotene stage of meiotic prophase I by double-stranded breaks (DSBs) along chromosomes (Figure 1), which are catalyzed by the topoisomerase-like protein, SPO11 (Keeney et al., 2020). 5' to 3' resection of the DSB exposes 3' single-stranded ends that are subsequently bound and protected by replication protein A (RPA) (Rodgers and McVey, 2015). However, when the recombinase proteins RAD51 and DMC1 replace RPA, the 3' ends search for a homologous repair template and undergo single-end invasion, forming a displacement (D)-loop intermediate that is then primed for DNA synthesis (Wang and Copenhaver, 2018).

From this point, DSBs can either form COs or be repaired as non-COs depending on the proteins involved in their repair (Taagen et al., 2020). Non-crossovers (NCOs) may form through synthesis-dependent strand annealing (SDSA), by which the extended invading strand breaks off and reanneals to the 3' end on the opposite side of the original break, followed by DNA gap-filling and ligation (Wang and Copenhaver, 2018). Otherwise, class I COs, controlled by MLH3 and HEI10, or class II COs, controlled by MMS4 and MUS81, may form as DNA synthesis occurs and then second-end capture leads to the formation of double Holliday junction (dHJ) intermediates that induce DNA exchange between homologous chromosomes (Taagen et al., 2020).

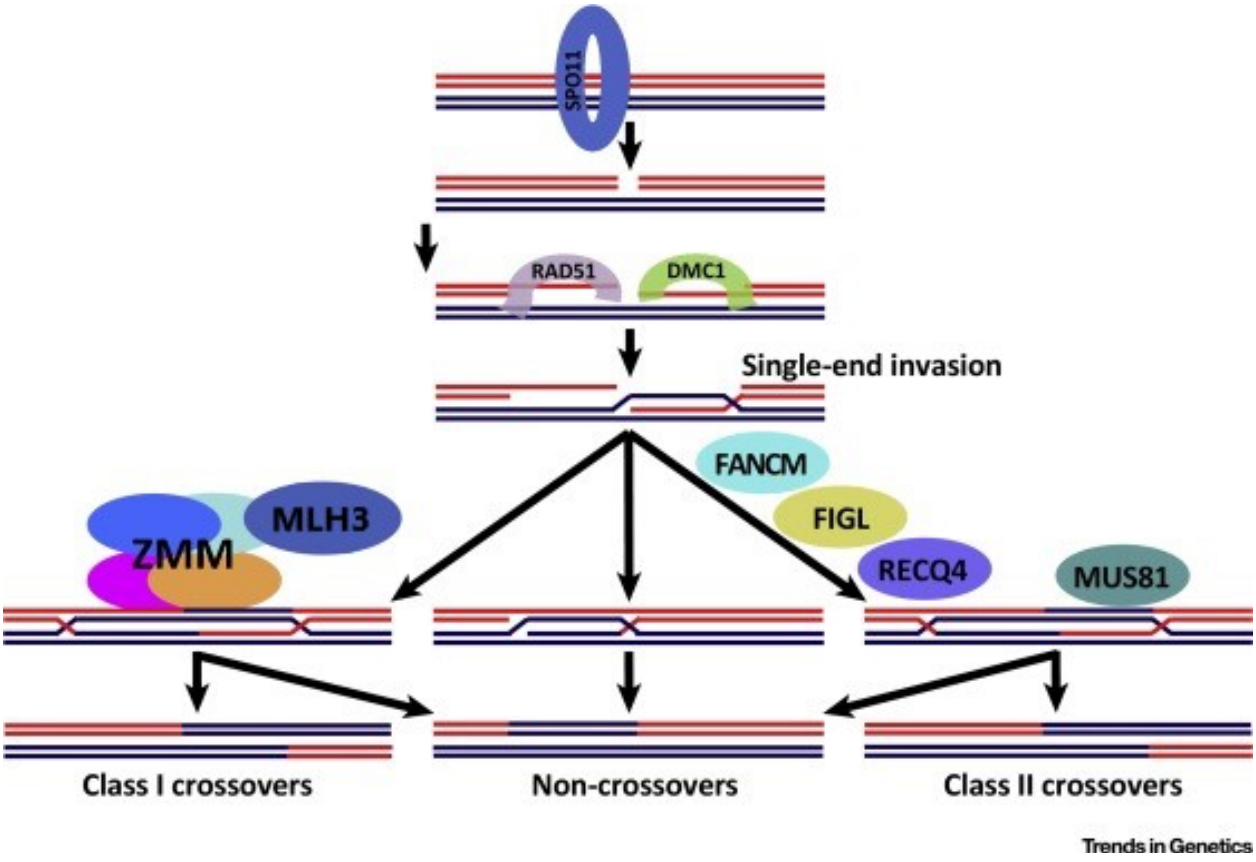


Figure 1: Diagram of homologous recombination pathway showing key proteins involved at each step. (Zelkowski et al., 2019)

Only a limited number of DSBs proceed to crossovers; in each meiosis event in maize, around ~300 to 500 DSBs form of which only ~20 are resolved as COs (Zhao et al., 2021). CO hotspots, key sources of genetic variation, frequent the gene promoters and terminators, but there are differing proposals as to what factors dictates hotspot formation. Markers of active chromatin, including decreased nucleosome density and trimethylation of lysine 4 in histone H3 (H3K4me3), are associated with increased recombination rates but not absolute factors in determining hotspot locations in maize (He et al., 2017). The Arabidopsis mutant chromatin-remodeler *decreased DNA methylation* (*ddm1*) affected higher recombination rates in

euchromatin, revealing the link between DNA methylation and crossover frequency (Taagen et al., 2020). In maize and other plants with large genomes, COs are suppressed at the heterochromatic centromere and telomeres, because of the evolutionarily conserved kinetochore assembly function of the centromere and inaccessible chromatin and epigenetically silenced pericentromeric regions (Taagen et al., 2020). CG DNA methylation and dimethylation of histone 3 lysine 9 function in suppressing hotspots in the pericentromeric region (Taagen et al., 2020).

Recombination and indel formation

In principle, meiotic recombination creates ample opportunity for indel creation (Figure 2). A study in humans found that there was no significant enrichment of indels in recombination hotspots (Montgomery et al., 2013). In contrast, a recent study found that short deletions (<500 bp) have higher rates of overlap with recombination hotspots, than would be expected by chance, in wild an cultivated tomato, including *Solanum pimpinellifolium*, *Solanum lycopersicum* var. *cerasiforme*, and *Solanum lycopersicum* var. *lycopersicum* (Fuentes et al., 2021). These data suggest that recombination may play a role in small indel generation and further implies that the study in humans may have lacked sensitivity in terms of indel detection. In this study, I examined indels, because previously, there has been limited focus on indels as a recombination outcome in maize.

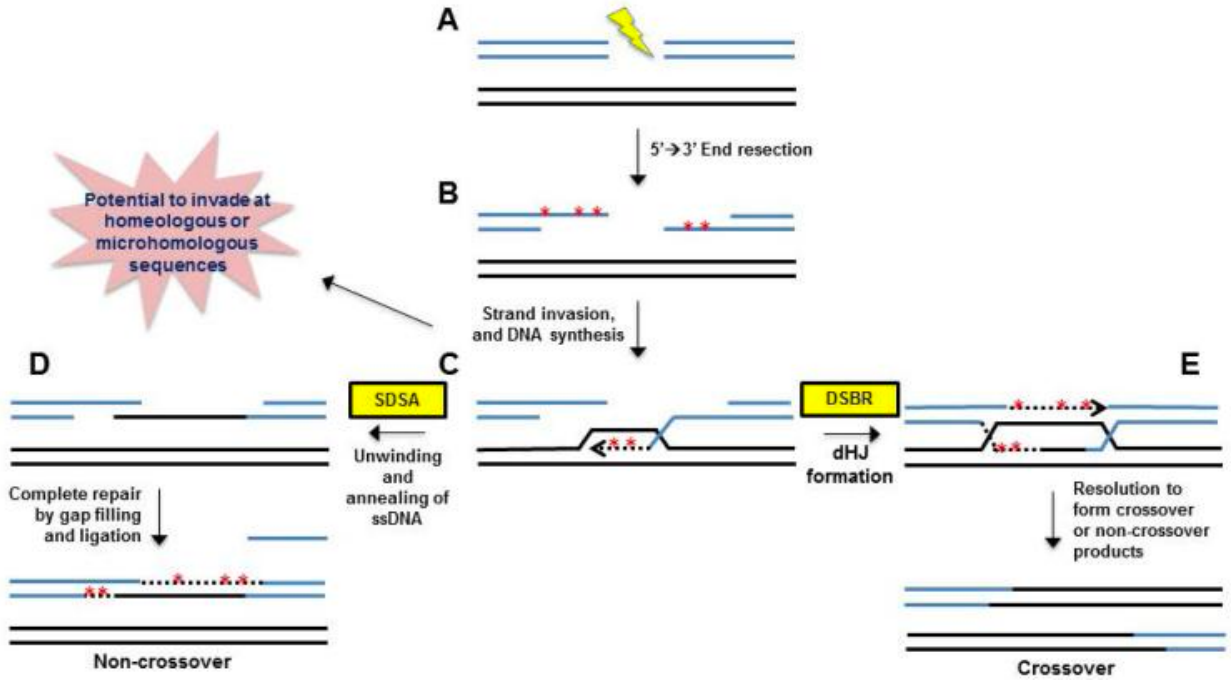


Figure 2: Schema of double-stranded break (DSB) repair with two pathways of resolution leading to non-crossovers and crossovers. Red asterisks represent points at which mutagenesis, including indel formation, may occur. (Rodgers and Mcvey, 2015.)

Potential Sources of Indels

a) Double-DSBs

A recent study in mice found that Spo11, a topoisomerase-like protein involved in initiating DSBs, can generate sizable, 34 to several hundred base-pair long gaps (Prieler et al, 2021). These deletions are formed by concerted cuts by Spo11 that release fragments bound by Spo11 at both ends, evidenced by the finding that these fragments had lengths of constant multiples (Prieler et al, 2021). These concerted cuts, known as double-DSBs (dDSBs), are widely distributed across the genome and enriched at DSB hotspots. Double-DSBs also coincide with topoisomerase II-binding sites in regions of topological stress, especially at promoter sites with negative supercoiling, demonstrating the relationship between transcriptional stress and dDSBs (Prieler et

al, 2021). Double-DSBs may be regulated by ATM kinase, which limits DSB numbers and controls DSB hotspot formation; there was an increased frequency of dDSB fragments in ATM-deficient mutant models in mice (Lukaszewicz et al., 2021). The excised fragments can integrate into ectopic positions or integrate with reversed orientation as inverted insertions (Lukaszewicz et al., 2021; Prieler et al, 2021). Gaps left by the fragment leave four DNA ends exposed that can serve as substrates for deletion mutagenesis; microdeletions may form from double-DSBs at a single hotspot site, as demonstrated in the study of ATM-deficient mutant mice that showed a peak in 30 bp and 60 bp microdeletions at hotspots (Lukaszewicz et al., 2021; Prieler et al, 2021).

b) Non-Homologous End-Joining Mediated DSB Repair

DSB repair may occur through two pathways: homologous recombination (HR)-mediated repair or non-homologous end-joining (NHEJ) repair. NHEJ can proceed through two sub-pathways known as classical NHEJ (C-NHEJ) and alternative NHEJ (A-NHEJ): 1) the heterodimer KU70/KU80 binds the broken ends of DNA and recruits DNA ligase 4 to re-ligate the ends without a homologous template; or 2) poly(ADP-ribose)-polymerase 1 (PARP1) binds and create microhomologies that anneal (Gehrke et al., 2022). C-NHEJ may create small indels, while A-NHEJ generates larger deletions because of the loss of the intermediate sequence after annealing (Gehrke et al., 2022).

A-NHEJ can be further classified into two sub-mechanisms, known as microhomology-mediated end joining (MMEJ) and theta-mediated end joining (TMEJ). MMEJ is a repair pathway involving 5’ to 3’ end resection to expose 6-20 nucleotide-long microhomologies (Rodgers and

McVey, 2015). These tiny overhangs then anneal to direct repeats flanking the DSB, and re-ligated after removal of the 3' flaps (Allen and Smith, 1996; Rodgers and McVey, 2015). Single-strand annealing (SSA) is an extreme procession of MMEJ, involving extensive resection (Rodgers and McVey, 2015). Though there are slight distinctions in the proteins involved in MMEJ and SSA, both result in simple deletions (Rodgers and McVey, 2015). In TMEJ, DNA polymerase theta (Polθ) promotes annealing and template-dependent and independent synthesis of possible microhomologies, creating complex deletions and insertions (Rodgers and McVey, 2015).

c) Homologous Recombination Mediated DSB Repair

As I describe above, homologous recombination (HR) is the mechanism leading to CO formation. Each step during the HR-mediated repair of DSBs is susceptible to error, suggesting that CO and NCO formation have high mutagenic capacity (Rodgers and McVey, 2015). DNA polymerase is not highly processive, so slippage and pausing may occur, increasing the possibility of template switching (Hicks et al, 2010). The invasion of the dissociated D-loop intermediate during template-switching may result in coupled deletions and insertions (Rodgers and McVey, 2015). During the homology search prior to strand invasion, there is a significant opportunity for inappropriate pairing and mutagenesis. Synthesis during D-loop extension can produce mutations such as base-pair substitutions and frameshift mutations (deletions) due to the action of error-prone polymerases (Hicks et al, 2010). A study of a mutated replicative polymerase in yeast, Pol δ, resulted in a suppression of all (−1) frameshift mutations, demonstrating Pol δ's mutational power (Hicks et al, 2010). In the NCO pathway, the nascent strand can incorrectly pair with sequences outside the DSB during SDSA, and errors during

single-stranded gap filling and ligation can be mutagenic. Unequal crossovers in the CO pathway, which refers to homologous recombination between separated direct sequence repeats, have the potential for short indel mutagenesis. Break-induced replication (BIR), another HR repair mechanism for one-ended double-stranded breaks, can be highly mutagenic. BIR also exposes ssDNA to damage that may be repaired in an error-prone manner, and it is also susceptible to mutational D-loop extension and template switching (Rodgers and McVey, 2015).

The Importance of Indels

Indels are an important class of structural variations (SVs) that are widely distributed in plant genomes with an average density comparable to that of SNPs (Yuan et al., 2021). Insertions and deletions refer to stretches of DNA containing genes or regulatory elements that were added to or removed from the genome, respectively (Mabire et al., 2019). An analysis of indel markers in rice demonstrated that there are five groups of indels: (i) indels of single-base pairs, (ii) monomeric base-pair expansions, (iii) multi-base-pair expansions of 2–15 bp repeat units, (iv) transposon insertions, and (v) indels containing random DNA sequences (Yuan et al., 2021). Another study of indels in the maize genome showed that a majority of indels were in intergenic space; the regions 0.5 kb upstream of transcription start sites (TSS) sites have the highest polymorphism rate and CDS regions have the lowest (Liu et al, 2015). Intriguingly, polymorphism levels in genic regions of the maize genome were higher than that in intergenic regions (Liu et al, 2015). This demonstrates that insertion and deletion mutations likely alter phenotype and function, marking their importance as a source of genetic diversity. Understanding mechanisms of indel origination and their relationship to the recombination mechanism is central to future studies that exploit genetic diversity for crop improvement.

In agreement with recent studies of recombination pathways in mice and yeast, I hypothesized that meiotic recombination in maize contributes to indel formation. Thus, I predicted enrichment of indels at meiotic recombination sites. To test this hypothesis, I mapped indels to maize DSB and CO sites and measured the presence and degree of enrichment of indels of different sizes. I discovered large increases in the occurrence of small indels (1bp-50bp) within all recombination sites, but only minimal differences in medium (100bp-500bp) and large (500bp-50kb) indel density.

II. Materials & Methods

Datasets

To examine the association of indels with recombination hotspots, I used two datasets of maize indels identified in the set of 26 diverse inbred lines (Hufford et al., 2022). One of the datasets contained small indels up to 500bp, which was produced after sequencing the inbred lines using Illumina sequencing and examining short reads. The indels were detected using the GATK pipeline HaplotypeCaller tool, which runs multiple sequence alignment on reads and forms consensus indels (Hufford et al, 2021). The other dataset included larger indels, or structural variation (SVs), of 500bp to 50kb in size. To generate this dataset, the 25 inbreds were sequenced with the PacBio long-read technology and indels were discovered through comparative analyses of the 26 assembled genomes. All indels were mapped to the v5 B73 maize reference genome.

In the analysis, I used a high-resolution dataset of ~3100 DSB hotspots identified as 2kb intervals in the maize B73 inbred (He et al., 2017). In contrast to DSBs, a saturated, high-resolution CO hotspot dataset does not yet exist in maize. Thus, I used three complementary datasets:

(1) A dataset of ~1100 COs mapped in the maize B73 x Mo17 hybrid (Kianan et al., 2018). This dataset is relatively small but of high resolution, as all COs have been mapped to intervals of 2kb or less.

(2) A CO hotspot dataset (Melnick et al., 2015) generated from the maize Nested Association Mapping (NAM) population, which was constructed by crossing 25 maize inbred lines selected to represent the diversity of maize, to a common parent, the B73 inbred, and deriving 200 recombinant inbred lines (RILs) from each cross (McMullen et al., 2009). This dataset is based on ~75000 COs but has relatively low resolution, with sites ranging up to almost 100kb in length.

(3) A dataset of CO sites, all 2kb in length, predicted in the genome of the maize B73 inbred using a Machine Learning algorithm (Wang et al., 2022) based on chromatin characteristics of empirical maize COs.

I performed a clean-up and filtering of all datasets in Linux and R. This entailed excluding data points in the indel datasets that failed preset filters (Quality Depth < 2.0 || FS > 200.0 || ReadPosRankSum < -20.0) using vcftools, version control using CrossMap (transforming all datasets to the version 4 of the maize genome draft), as well as consolidating bookend or overlapping intervals in these datasets using bedtools. After clean-up, the recombination datasets, DSB hotspots, COs, CO hotspots, and predicted COs, contained 2924, 58847, 33568, and 294

sites, respectively. The small indel dataset was divided into a “small” and “medium” dataset, with all 1-50 base pair (bp) long indels in the “small” dataset and 100-500 bp long indels in the “medium” dataset. The large indel dataset was modified to only include 10kb-50kb long indels.

To prepare all DSB and CO datasets for analysis, I subsetted each to include intervals <10 kb in size, and I created datasets of the regions 2kb upstream and 2kb downstream from these intervals in R. Coldspot datasets, termed cold due to the location of these datasets within recombination desert regions, that corresponded to each of the recombination datasets were generated in R to serve as a negative controls. These datasets were created by finding DSB, CO and CO hotspot desert regions through identifying the top 5% longest regions where there were none of these recombination features were present. Following this, I wrote a script in R to randomly sample random sites from these desert regions without overlap and excluding pericentromeric regions. I ensured that each negative control dataset contained the same number of total sites and same number of sites on each chromosome as their corresponding dataset. All datasets were transformed to the version 4 of the maize reference genome using CrossMap.

Characterizing Indel Distribution

Distribution plots of each of the indel datasets were generated in R using the package ChromoMap to assess the general relationship between chromosomal location and indel density. Next, I computed Spearman rank correlation with log transformed data to assess the relationship between nucleosome occupancy and indel density. The nucleosome occupancy data was obtained from Mateusz Zelkowski and prepared by Ruth Epstein in the Pawlowski Lab (Zelkowski et al.,

2022). Indel location was then investigated in terms of overlap with genes, transcription start sites (TSS), and transcription termination sites (TTS) sites. I also examined indel location in terms of overlap with exons and introns. Indel density in each of these regions was calculated in R. The statistical significance of indel density within exonic and intronic regions was evaluated with a two-way ANOVA test using a Poisson regression model with interaction effects between the region and indel size, followed by the Bonferroni correction for posthoc analysis.

Analyzing Indel Presence at Recombination Sites

To assess the overlap between the indel datasets and recombination datasets, as well as their complementary coldspot datasets, scripts using the module bedtools in Linux were used. To assess the overlap between the indels and the NCOs, the DSB dataset was processed in R to exclude DSBs that resolve to COs, and bedtools in Linux was used to calculate the overlap. DNA methylation at CG, CHG (H = A, C, or T) and CHH sites was analyzed with bedtools and R. The overlap for each of these datasets was visualized in R by calculating three measures:

1. the percentage of indels that overlap with the feature (DSB hotspot, CO, CO hotspot, predicted CO,),

$$\text{Indel overlap (\%)} = \frac{\#\ of\ indels\ that\ intersect\ with\ each\ interval\ (CO/DSB)}{total\ \#\ of\ indels\ in\ dataset} * 100\%$$

2. the percentage difference of each feature that overlaps with indels,

$$\text{Feature overlap (\%)} = \frac{\#\ of\ features\ \ (CO/DSB)\ that\ intersect\ with\ indels}{total\ \#\ of\ features\ in\ dataset} * 100\%$$

3. and the indel density at each of the features.

$$\text{Indel Density (indels/bp)} = \frac{\#\ of\ indels\ that\ intersect\ with\ each\ feature}{size\ of\ feature\ (bp)}$$

The statistical significance of the average differences among the first two measures, indel overlap and feature overlap, were evaluated by first using a Shapiro-Wilk's and Levene test in R to ensure a normal distribution and homogeneity among groups, followed by a one-way ANOVA. The statistical significance of the differences in indel density within each feature and the negative control regions was evaluated with a two-way ANOVA test using a Poisson regression model with interaction effects between the region and indel size, followed by the Bonferroni correction for posthoc analysis of not normally-distributed data. Additionally, the DSB hotspots, COs, CO hotspots, and predicted COs that fall within DNA sequence features (gene body, TSS, TTS) and their respective indel densities were calculated to assess whether indels are more frequent among any specific chromosomal features. The differences among the indel densities were compared with a one-way ANOVA test.

Methylation and Transposable Element (TE) Analysis

Using bedtools in Linux and R, I calculated the CG, CHG, and CHH, site methylation levels of indels that intersect with DSB hotspots, COs, CO hotspots, and predicted COs and at recombination sites that did not intersect with indels. The datasets of recombination sites without indel presence were produced with bedtools subtract. The methylation levels were assessed using the intersect function in bedtools Linux, and calculations were completed in R. The statistical significance of differences in CG, CHG, and CHH, methylation levels was evaluated with a nested ANOVA, followed by the Bonferroni correction for posthoc analysis. Metaplots to visualize each comparison were created in R. Additionally, the transposable element (TE) density in both datasets was compared using bedtools intersect and R, and the significance of the

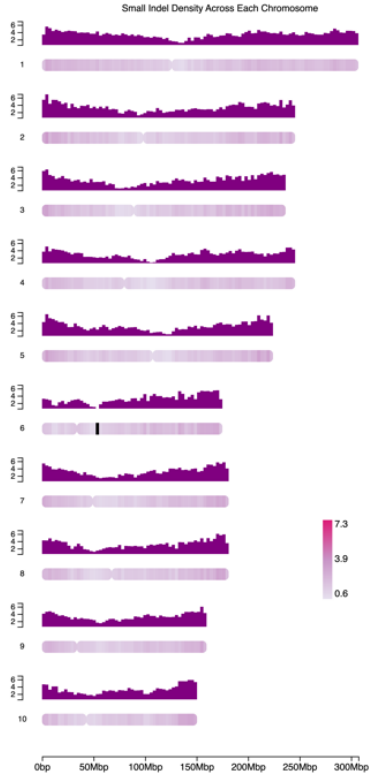
differences was evaluated with the Kruskal-Wallis chi-squared rank-sum test for not normally-distributed data and the one-way ANOVA test for normally distributed data ($P < 0.05$).

III. Results and Discussion

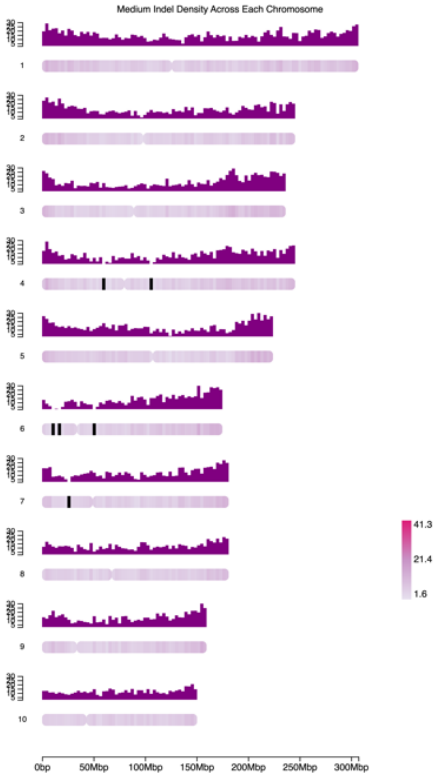
Where do indels lie?

Small (1-50 bp), medium (100-500bp), and large (500bp-50kb) indel distribution exhibit a general U-shaped pattern of increased indel density in subtelomeric regions and decreases around pericentromeric regions on each chromosome (Figure 3). The indel distribution pattern on chromosome 6 is an exception to this pattern; there is a stark decrease in indel density on the small arm, contributing to a distribution that slopes upward towards the subtelomeric region of the large arm. This pattern could be explained by the presence of a heterochromatic knob in this location. Heterochromatic knobs are a stretch of DNA repeats of two types, the 180 bp repeat and the TR-1 repeat (Ananiev, Phillips, & Rines, 1998). Knobs have higher methylation levels than the average CO site and can influence recombination patterns, as CO suppression was discovered in three of the five knobs in the genome of the B73 maize inbred line (Kianan et al., 2018). The U-shaped indel distribution patterns resemble CO distribution patterns, which provides a preliminary indication of a potential relationship between recombination and indel genesis.

A


Small Indel Density Across Each Chromosome
7.3
3.9
0.6
0bp
50Mbp
100Mbp
150Mbp
200Mbp
250Mbp
300Mbp

B


Medium Indel Density Across Each Chromosome
41.3
21.4
1.6
0bp
50Mbp
100Mbp
150Mbp
200Mbp
250Mbp
300Mbp

C


Large Indel Density Across Each Chromosome
98.3
53.1
8.0
0bp
50Mbp
100Mbp
150Mbp
200Mbp
250Mbp
300Mbp

Figure 3: Indel density distribution across chromosomes 1-10. Density calculated as the number of indels per megabase (Mb).

A: Small indel distribution, shading across chromosome correlates with indel density (darker color represents higher density). Scaled by 1000 (7.3 represents 7300 indels).

B: Medium indel distribution. Unscaled.

C: Large indel distribution. Unscaled.

There did not appear to be a correlation between small indel density and nucleosome occupancy ($r = -0.0413$, $P = 1$). However, there appeared to be a significant moderately positive relationship between medium indel density and nucleosome occupancy ($r = 0.4317$, $P < 2e\text{-}16$) and between large indel density and nucleosome occupancy ($r = 0.3407$, $P < 2e\text{-}16$). Higher nucleosome occupancy levels correspond with heterochromatin structure. This suggests that small indel location is not dependent on chromatin openness, while medium and large indels tend to be tolerated in genomic regions, such as heterochromatin, with fewer highly expressed genes. A deeper analysis of indel location revealed that a majority of small and large indels are in intergenic regions, while a majority of medium indels are in genic regions (Table 1).

Table 1: Percent of small (1-50bp), medium (100-500bp), and large (500bp-50kb) indels located within non-genic and genic regions. Genic regions encompass 2kb-long transcription start site regions (TSS), 2kb-long transcription termination site regions (TTS), and gene bodies. The genome outside of these regions is termed non-genic.

	TSS regions(%)	TTS regions (%)	Gene bodies (%)	Non-Genic (%)

Small Indels	11.68	10.53	14.16	63.63
Medium Indels	18.09	16.40	22.47	43.04
Large Indels	15.50	13.56	9.27	61.66

The average small indel density was higher within introns (0.0181 indels/bp) than exons (0.0103 indels/bp). In contrast, densities of medium and large indels in exons and introns were more similar (Figure 4). A two-factor ANOVA test indicates that region and indel size are associated with significant differences in indel count ($P < 2e\text{-}16$) and the interaction effects were significant, suggesting that the relationship between indel size and indel count depends on indel location ($P = 6.32e\text{-}8$). Posthoc comparison with Bonferroni correction indicated that the mean small indel count within exons was significantly different than within introns ($P < 2e\text{-}16$), but the mean medium and large indel counts did not significantly differ between the regions ($P = 1$).

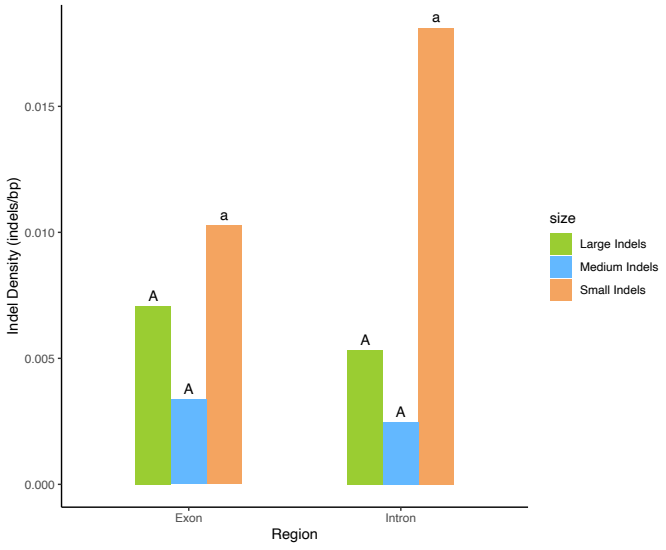


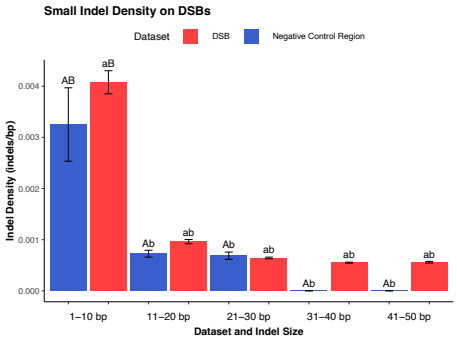
Figure 4: Comparison of indel density in intronic and exonic regions. Lowercase of letters above the bars indicate a statistically significant difference at $P < 0.05$, i.e. bars marked with lowercase 'a' are significantly different from bars marked with uppercase 'A'.

Indel Density at Recombination Sites

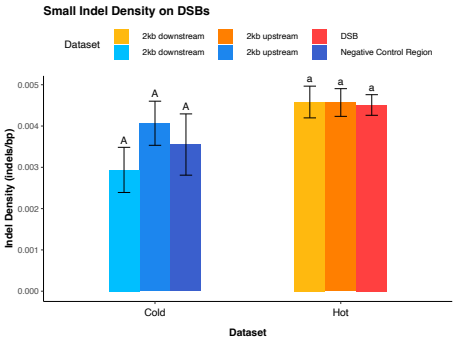
DSB hotspots

The small indel density at DSBs hotspots was on average higher than in the negative control recombination desert regions, and a two-way ANOVA and posthoc Bonferroni correction revealed that the difference in indel density was significant ($P <$ 2e-16, Figure 5A). The difference in indel density was significant in all indel size groups ($P <$ 2e-16, Figure 5A). However, there were no statistically significant differences in indel density between the 2kb upstream and 2kb downstream regions of DSB hotspots and their respective negative control regions ($P = 0.937$, Figure 5B).

A



B



C

D

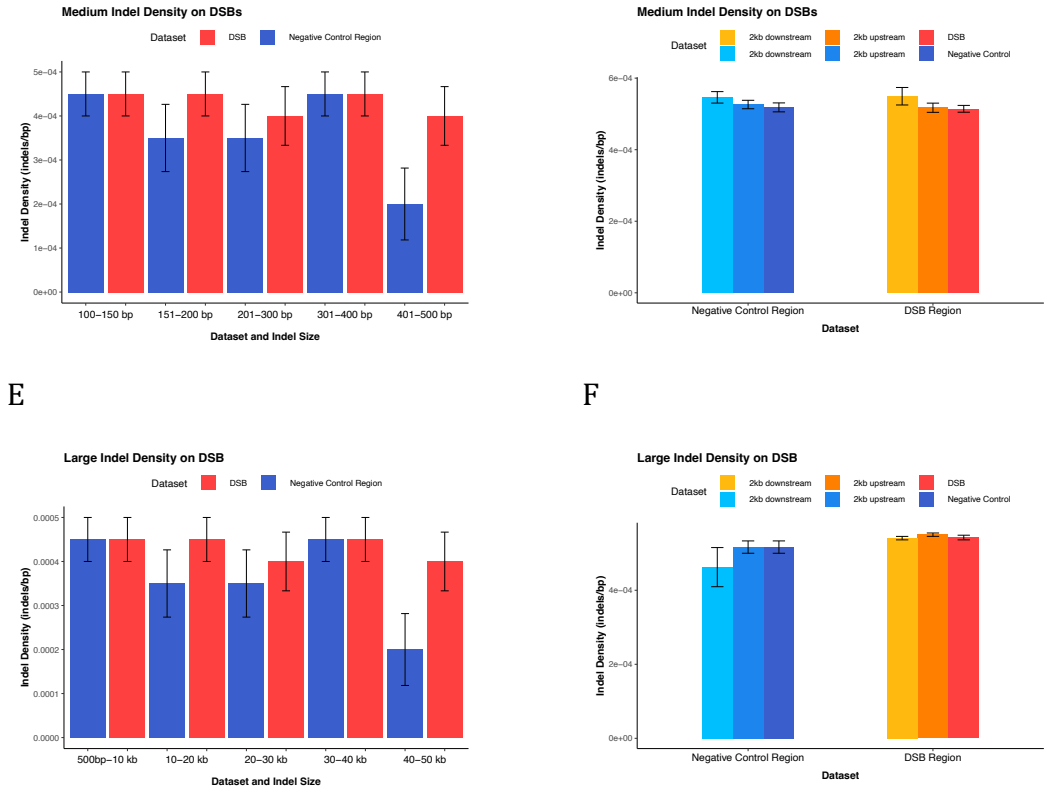


Figure 5: Comparison of small, medium, and large indel density at DSB hotspots compared to DSB desert regions. Lowercase of letters above the bars indicate a statistically significant difference at $P < 0.05$, i.e. bars marked with lowercase 'a' are significantly different from bars marked with uppercase 'A'. Error bars represent +/- 1 SE, N=10.

A: Density of indels at DSB hotspots compared to DSB deserts.

B: Density of all small indels at DSB hotspots compared to DSB deserts and their respective 2kb upstream and downstream regions.

C: Density of medium size indels at DSB hotspots compared to DSB deserts.

D: Density of all medium size indels at DSB hotspots compared to DSB desert and their respective 2kb upstream and downstream regions.

E: Density of large indels at DSB hotspots compared to DSB deserts.

F: Density of all large indels at DSB hotspots compared to DSB desert and their respective 2kb upstream and downstream regions.

There was no significant difference in medium indel density at DSB hotspots ($P = 0.410$) or when comparing 2kb regions upstream and downstream hotspots form DSBs to the respective negative control regions ($P = 0.104$, Figures 5C and D). There was also no significant difference in large indel density among different indel size groups ($P = 0.417$) or when comparing 2kb regions upstream and downstream hotspots form DSBs to the respective negative control regions ($P = 0.344$) (Figures 5E and F).

COs

The indel density at COs exhibited similar trends to DSB hotspots, in the sense that average small indel density at COs was higher than at the negative control regions, and a two-way ANOVA and posthoc Bonferroni correction revealed that the surge in indel density was significant ($P < 2e\text{-}16$, Figures 6A, 6B, 6C). Across all CO datasets, the difference between the density of 1-10bp indels and other indel size groups was significant ($P < 2e\text{-}16$), and the interaction effects between indel size and indel region were also significant ($P < 2e\text{-}16$, Figures 6A, 6B, 6C). There was also a significant increase in the density of 11-20 bp (small) indels as compared to other indel size groups in empirical and predicted CO regions ($P < 2e\text{-}16$) (Figures 6B and 6C).

Furthermore, the disparity in small indel density among the CO hotspots and predicted COs and their respective 2kb upstream and 2kb downstream regions was significant ($P < 2e\text{-}16$), as were

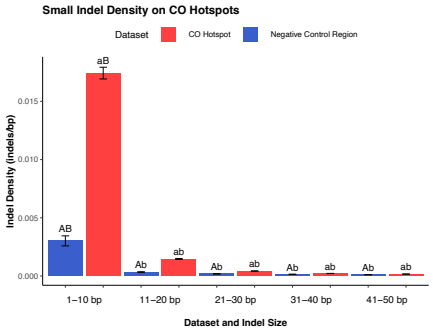
the differences in indel density among negative control regions and 2kb upstream ($P = 0.011$) and downstream ($P = 0.026$) (Figures 7A and 7C). The small indel density was also significantly higher at empirical COs than in 2kb downstream ($P = 4.7e-12$) but not in 2kb upstream regions ($P = 0.35$; Figure 7B).

The trends for medium and large indel density across the different CO sites were much more nuanced. At CO hotspots and predicted COs, the average medium and large indel density were significantly higher than at the negative control regions ($P = 0.00223$, $P < 2e-16$, $P < 2E-16$, $P < 2e-16$) (Figures 6D, 6G, 6F, 6I). However, at predicted CO sites, the medium indel density was slightly higher in negative control regions than at predicted CO sites ($P < 2E-16$; Figure 6F). This increase in medium indel density in negative control regions was not observed in other datasets. On the other hand, the medium and large indel density at empirical COs did not differ significantly from negative control regions ($P = 0.16$ and $P = 0.059$), (Figures 6E and 6H).

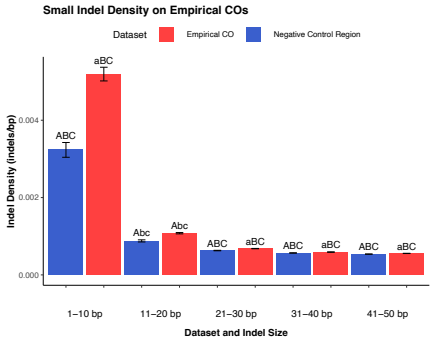
In terms of different indel size groups, there was no significant difference in medium indel density at COs and CO hotspots ($P = 0.110$, Figures 6D and 6E). The density of medium indels at predicted COs did not differ significantly by size group except when comparing 400-500 bp indels to other medium-sized indels (Figure 6F). Interestingly, the mean density of 500bp – 10kb (large) indels at CO hotspots and empirical COs was higher than other large indel size groups ($P = 8.2e-09$, $P = 5.7e-11$, $P = 4.1865e-8$, $P=8.30895e-7$, $P=3.2e-4$) (Figures 6G and 6H). Meanwhile, the density of 500bp–10kb indels at COs was significantly lower than other indel size groups ($P < 2e-16$; Figure 6I).

There was no significant difference in the change in indel density in 2kb upstream and downstream regions at CO hotspots (P=0.93482, Figures 7D and 7G) or at CO (P=0.699 and P = 0.8081, Figures 7E and 7H). Unexpectedly, the medium and large indel density was lower at predicted COs and negative control regions compared to 2kb upstream and downstream regions. These differences in medium and large indel density were significant within predicted COs and 2kb upstream and downstream regions (P< 2E-16), but not significant within negative control regions and their respective 2kb upstream and downstream regions (Figures 7F and 7I).

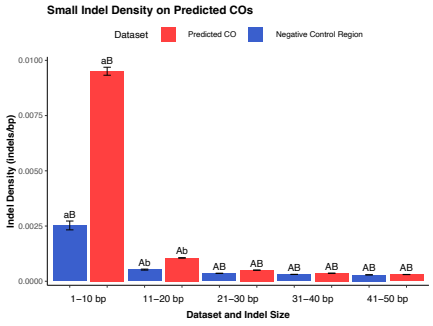
A



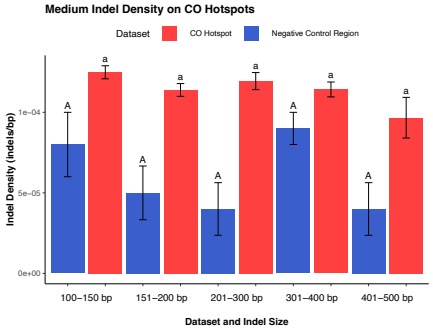
B



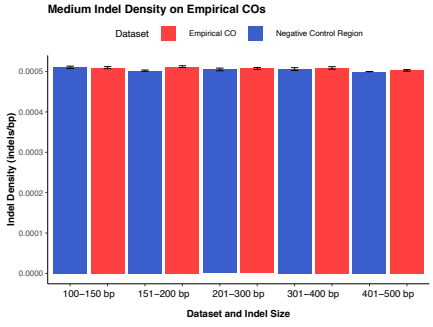
C



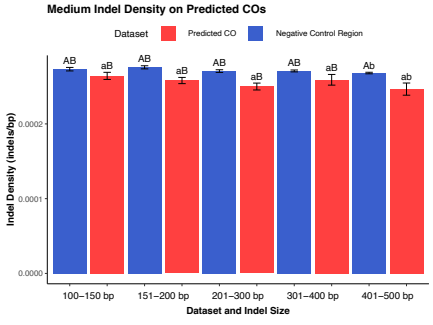
D



E



F



G H I

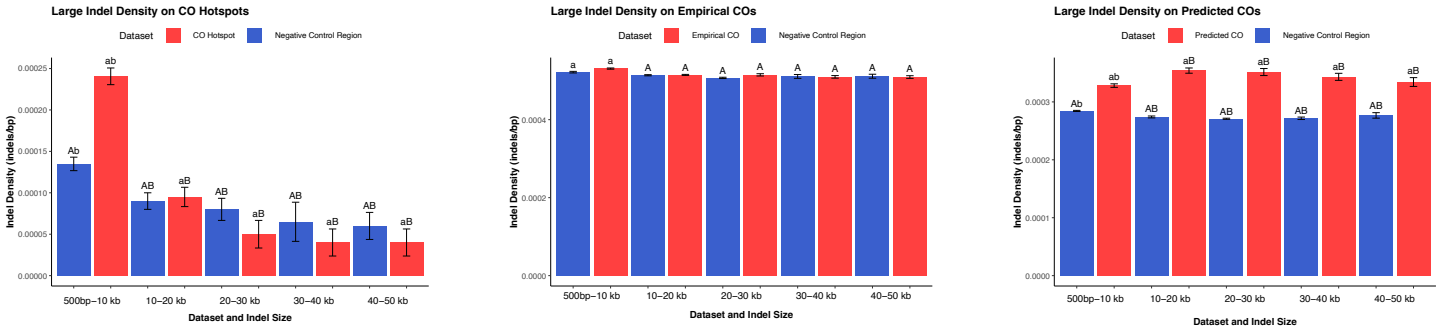


Figure 6: Comparison of indel density of different indel size groups at CO sites and negative control regions. Lowercase of letters above the bars indicate a statistically significant difference and different letters indicate a different statistical group at $P < 0.05$. Error bars represent +/- 1 SE, N=10.

A: Density of small indels at CO hotspots and negative control regions.

B: Density of small indels at empirical COs and negative control regions.

C: Density of small indels at predicted COs and negative control regions.

D: Density of medium indels at CO hotspots and negative control regions.

E: Density of medium indels at empirical COs and negative control regions.

F: Density of medium indels at predicted COs and negative control regions.

G: Density of large indels at CO hotspots and negative control regions.

H: Density of large indels at empirical COs and negative control regions.

I: Density of large indels at predicted COs and negative control regions.

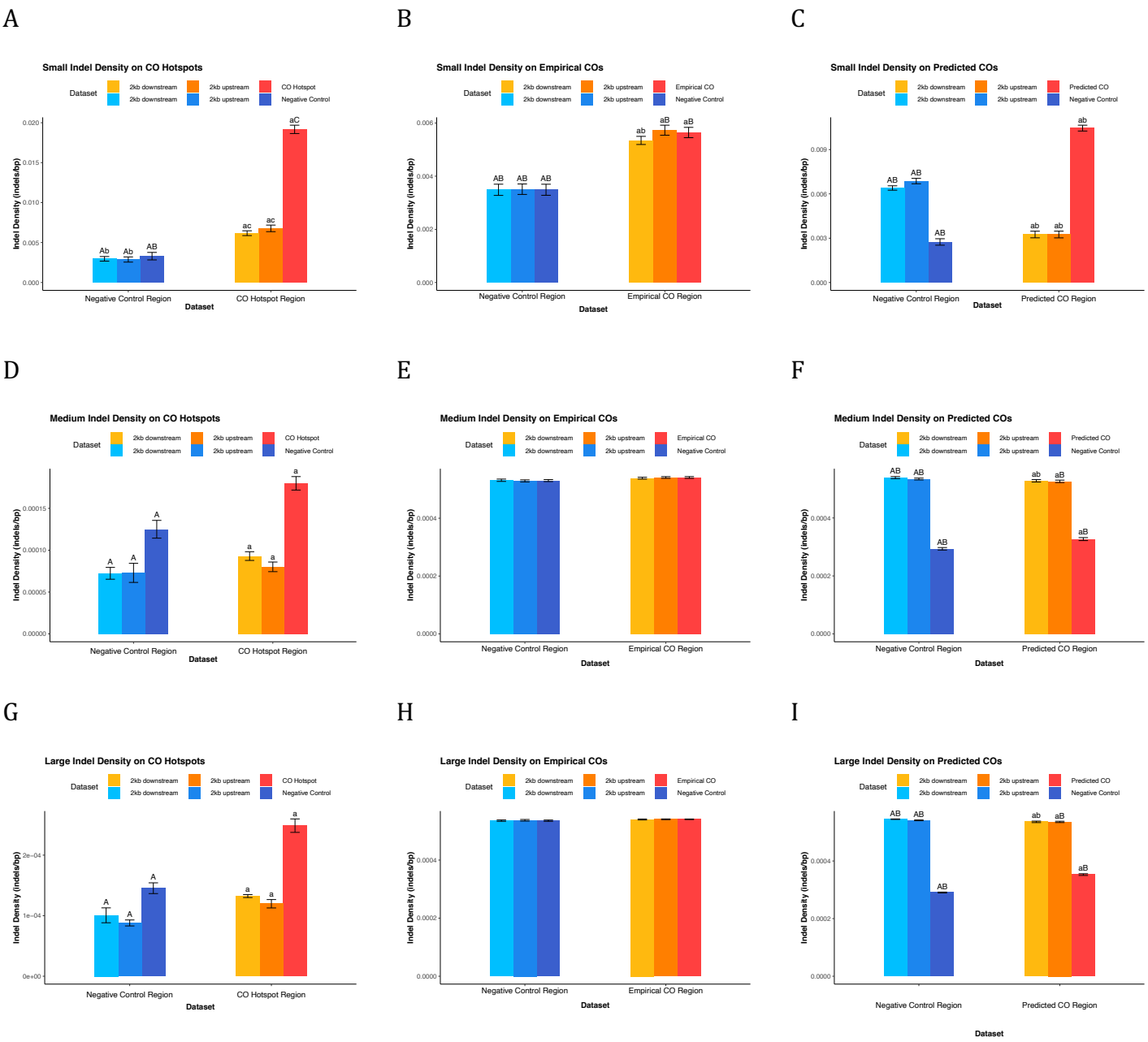


Figure 7: Comparison of small, medium, and large indel density at CO sites and negative control regions and their respective 2kb upstream and downstream regions. Lowercase of letters above the bars indicate a statistically significant difference at $P < 0.05$. Error bars represent +/- 1 SE, N=10.

A: Density of small indels at CO hotspots and negative control regions and their respective 2kb upstream and downstream regions.

B: Density of small indels at empirical COs and negative control regions and their respective 2kb upstream and downstream regions.

C: Density of small indels at predicted COs and negative control regions and their respective 2kb upstream and downstream regions.

D: Density of medium indels at CO hotspots and negative control regions and their respective 2kb upstream and downstream regions.

E: Density of medium indels at empirical COs and negative control regions and their respective 2kb upstream and downstream regions.

F: Density of medium indels at predicted COs and negative control regions and their respective 2kb upstream and downstream regions.

G: Density of large indels at CO hotspots and negative control regions and their respective 2kb upstream and downstream regions.

H: Density of large indels at empirical COs and negative control regions and their respective 2kb upstream and downstream regions.

I: Density of large indels at predicted COs and negative control regions and their respective 2kb upstream and downstream regions.

MLH3 sites

I found that the average small indel density at MLH3 hotpot sites was 0.00792 indels per bp, which was significantly higher than at DSB hotspots and empirical COs, but lower than at predicted COs and CO hotspots ($P < 2e\text{-}16$) (Figure 9). These differences are in accordance with observations about MLH3 distribution patterns and may demonstrate how indels are produced at a molecular level, but they do not fully explain the elevated indel density at CO hotspots.

Altogether, I found enrichment of small indel density at various recombination sites compared to their respective recombination desert regions, indicating the mutagenic effect of meiosis. This enrichment was highly focused within the CO hotspot and predicted CO sites compared to their 2kb upstream and downstream regions, indicating that the effect of recombination in creating indels is relatively localized. However, there was only a minimal increase in medium and large indels at these sites compared to small indels. Meiotic recombination is more likely to generate more minor scale errors. In contrast, larger indels may be produced in somatic cells through mitotic DNA repair processes (Symington, Rothstein, & Lisby, 2014). I found differences between the different types of sites as well; DSB hotspots exhibited the smallest increase in small indel density and the lowest small indel density overall as compared to COs, CO hotspots, and predicted COs (Figure 9). COs had, on average, lower small indel densities than predicted COs sites, possibly because the predicted CO dataset represented a more complete repertoire of CO sites than what was detected empirically. In contrast, CO hotspots have the greatest increase in small indel density. CO hotspots represent sites with higher CO rates than predicted and empirical CO sites, which further illustrates the indel generation potential of the meiotic recombination pathway.

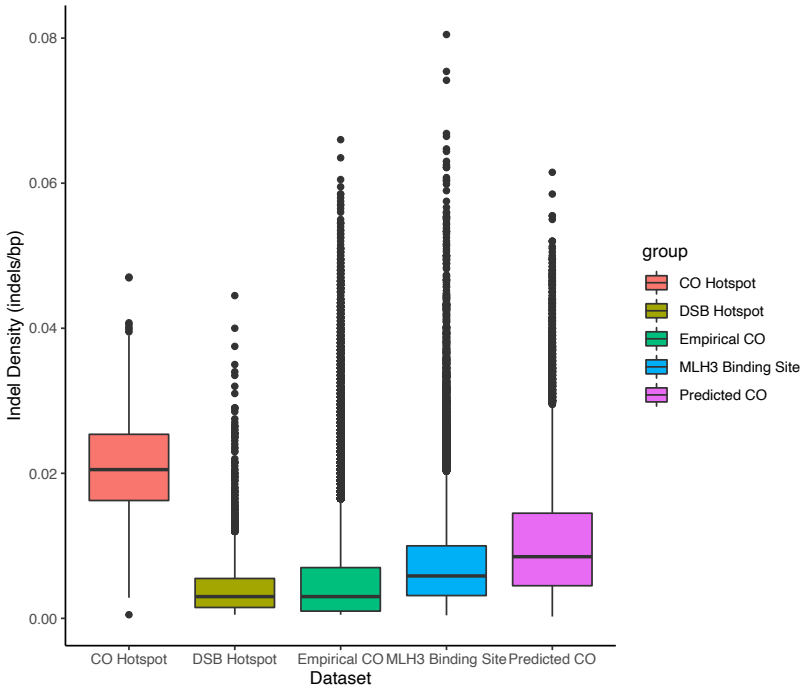


Figure 9: Summary of small indel density among all recombination sites, including DSB hotspots, empirical COs, predicted COs, CO hotspots, and MLH3 binding sites.

What fraction of indels are at recombination sites? Conversely, what fraction of recombination sites overlap with indels?

I also wanted to know what percent of all indels across the genome could have been generated by these recombination mechanisms. Overall, fewer than 50% of all indels were located within recombination sites, demonstrating that recombination is just one of the possible mechanisms for indel generation. However, a significant proportion of all the indels were encompassed by the predicted and empirical COs and their negative control regions, which can be explained by the differences in the size of the datasets. The CO and predicted CO datasets contained more data points than DSB and CO hotspot datasets, increasing the chances of an indel falling within either of these features. However, despite the DSB hotspot dataset containing 2,924 intervals, DSB

hotpots overlapped with a slightly lower proportion of small and medium indels and a slightly greater proportion of large indels than the CO hotspots, which contain only 294 intervals. DSBs can either proceed to COs or not, and the findings that more DSBs intersected with indels larger than 500 bp while more CO hotspots intersected with indels smaller than 500 bp emphasize the mutagenic potential of the CO mechanism. This result illustrates the stronger role of DSB hotspots in larger indel genesis and CO formation in smaller indel genesis, which may be because larger indels forming at CO hotspots would lead to the extinction of those hotspots.

I also looked at the proportion of recombination sites that intersected with at least one indel to understand how frequently recombination processes resulted in indels. Likewise, a higher proportion of recombination sites than negative control sites overlapped with the three indel datasets (Table 2). The proportion of recombination sites that intersected with indels was the greatest among small indels, with nearly all of CO hotspots, COs, and DSBs intersected by at least one indel, but this measure may just reflect the ubiquitous nature of small indels, rather than providing an indication of how often recombination produces indels.

Table 2: Summary measures (percent of indels intersecting with recombination sites and percent of recombination sites intersecting with indels) for small (1-50bp), medium (100-500bp), and large (500bp-50kb) indel datasets. Significance assessed with one-way ANOVA.

Indel size	Dataset	Percent of Indels in Site(s) (%)	Percent of Indels in Site(s) (-) (%)	P-value	Percent of Sites Containing Indel(s) (%)	Percent of Sites (-) Containing Indel(s) (%)	P-value
Small Indels	**DSB**	0.380	0.013	<2E-16	92.582	85.873	0.116
	CO Hotspot	0.942	0.113	1.74E-07	100.000	100.000	N/A

	Predicted COs	23.014	6.825	3.38E-10	98.592	94.432	3.87E-12
	Empirical COs	9.160	4.975	9.31E-07	86.260	81.437	1.43E-06
Medium Indels	**DSB**	0.359	0.261	0.0291	3.288	2.404	0.0119
	CO Hotspot	1.197	0.143	5.97E-07	60.587	16.612	8.62E-09
	Predicted COs	32.034	8.021	3.36E-12	19.257	6.168	1.17E-10
	Empirical COs	10.595	6.011	3.58E-05	4.912	3.331	0.00488
Large Indels	**DSB**	1.142	0.035	3.40E-16	21.78	24.04	0.709
	CO Hotspot	1.127	0.279	9.60E-06	87.93	55.99	8.32E-05
	Predicted COs	21.582	14.568	0.0113	35.61	32.26	0.00882
	Empirical COs	18.495	14.819	0.0915	31.90	29.71	0.121

Are there more indels at recombination sites in different gene regions?

I investigated these increases in indel density by looking into potential nuances among recombination sites in varied locations. More specifically, I wanted to know whether indel density differed between gene bodies and their associated promoter and termination regions and to elucidate whether the recombination mechanism preferentially generated indels within any of these regions some functional purpose. To determine if there were distinctions, I compared the small indel density at DSB hotspots, COs, CO hotspots, and predicted COs that intersected with different gene regions. This analysis did not reveal statistically significant differences, as the small indel density was roughly equal across these regions and did not deviate greatly from the average indel density at these recombination features (Figure 8). This suggests that the recombination mechanism generates indels uniformly across gene space, with no detectable bias towards promoters or termination regions.

A B

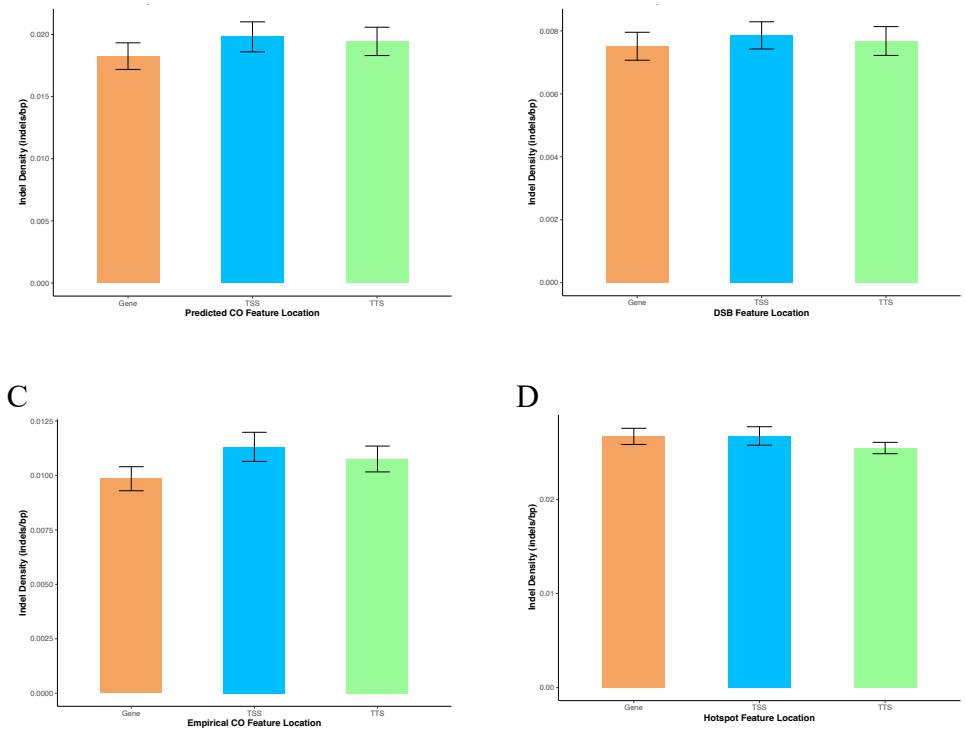


Figure 8: Analysis of small indel density among different features that intersect with genes (open reading frames), transcription start sites (TSS; gene promoter regions), and transcription termination sites (TTS). Error bars represent +/- 1 SE, N=10.

A: Comparison of small indel density among predicted COs that intersect with genes, TSS, and TTS regions. A one-way ANOVA test showed no significant difference in indel density among these features (P=0.609).

B: Comparison of small indel density among DSBs that intersect with genes, TSS, and TTS regions. A one-way ANOVA test showed no significant difference in indel density among these features (P = 0.861).

C: Comparison of small indel density among empirical COs that intersect with genes, TSS, and TTS regions. A one-way ANOVA test after log transformation of the data showed no significant difference in indel density among these features (P = 0.244).

D: Comparison of small indel density among CO hotspots that intersect with genes, TSS, and TTS regions. Levene's test indicated unequal variances (P=.4435) and a one-way ANOVA test after log transformation of the data showed no significant difference in indel density among these features (P = 0.537).

Exploring the elevated indel density in DSB and CO regions

Methylation Levels at Indels:

A comparison of CG, CHG, and CHH site methylation levels was used to determine whether DNA methylation differences may explain increased rates of indel generation. This analysis uncovered an average decrease in CG and CHG site methylation levels at recombination sites associated with indels compared to recombination sites without indel presence. The largest difference in mean CHG site methylation levels was at indels that intersected with empirical COs, with difference in methylation of 33.8% between sites with indels and sites without (Table 3). However, indels in predicted COs were characterized by the largest difference in mean CG site methylation levels with a difference of 38.1% between predicted COs with indels compared to those without (Table 3). CO hotspots had the smallest difference in mean CHG site methylation (Table 3). A nested ANOVA revealed that the feature intersecting with the indel and the binary indication of indel presence both have statistically significant effects on the CHG and CG site methylation levels (P < 2e-16). Posthoc comparison with Bonferroni correction indicated that CG and CHG site methylation levels at indels within DSBs, predicted COs, empirical COs, and CO hotspots were significantly different than within recombination sites without indels P < 0.05 (Table 3). CHH site methylation levels differed significantly at indels in predicted and

empirical COs and in CO hotspots, but they did not significantly differ between indels at DSBs and the average difference for all datasets was around 1% (Table 3).

These results suggest that CG and CHG methylation may explain the differences in the average indel density at recombination sites. Lower methylation levels at these sites increase the regions of active chromatin and increase recombination rates. Elevated indel density at these sites is a product of errors in the recombination pathway. Indels at CO hotspots had the lowest overall CHG and CG methylation levels, and this attribute may help explain why these sites had the largest small indel densities.

Table 3: CG, CHG, and CHH methylation levels (%) at indels intersecting DSB hotspots, COs, CO hotspots, and predicted COs compared to recombination sites that do not intersect with indels. Significance assessed using two-way ANOVA followed by Bonferroni correction for comparison of means.

	CHG Methylation	*P* value for CHG site methylation	CHH Methylation	*P* value for CHH site methylation	CG Methylation	*P* value for CG site methylation
DSB hotspots with Indels	49.469	< 2e-16	1.414	0.0867	67.847	< 2e-16
DSB hotspots without Indels	70.271		1.152		85.406	
CO hotspots with Indels	16.419	< 2e-16	1.306	< 2E-16	23.612	< 2e-16
CO hotspots without Indels	30.347		2.516		44.763	
COs with Indels	39.057	< 2e-16	6.559	< 2e-16	54.923	< 2e-16
COs without Indels	72.919		1.481		88.410	
Predicted COs with Indels	23.768	< 2e-16	1.719	2.7e-05	34.191	< 2e-16
Predicted COs without Indels	57.394		1.596		72.326	

Indel formation in Transposable Element (TE) regions:

TEs are ubiquitous in the maize genome (Stitzer et.al., 2021). As they formed repeats, TEs could act as a of indels. I wanted to know whether TE presence density could explain increase in indels at recombination sites. TE density in each CO and DSB dataset was calculated to determine whether TE density influences differences in indel density between COs and DSBs and their respective negative controls. Analysis using the Kruskal-Wallis indicated significant differences in TE density between empirical COs, predicted COs, and DSBs and their respective control regions, but the difference was less than 2-fold (P=0.008151, P=0.0001571, P=0.000211; Table 4). Analysis with the one-way ANOVA test showed that CO hotspots and the control regions did not exhibit significant differences in TE density (P=0.0724; Table 4). This suggests that TE density does not impact indel density to a large extent nor can it fully explain the elevated indel density at DSB or CO sites.

Table 4: Transposable element (TE) density (TEs/bp) within predicted COs, empirical COs, CO hotspots, and DSBs compared to their negative control regions (represented as -). Measured average, standard deviation (SD) and median of TE density within each region.

	Average TE Density	SD of TE Density	Median TE Density	Average TE Density (-)	SD TE Density (-)	Median TE Density (-)
Predicted COs	7.60E-04	9.67E-05	8.01E-04	1.17E-03	3.22E-05	1.18E-03
Empirical COs	1.03E-03	6.87E-05	1.05E-03	1.09E-03	2.59E-05	1.09E-03
CO Hotspots	6.63E-04	3.79E-05	6.66E-04	7.17E-04	8.06E-05	7.32E-04
DSBs	7.18E-04	1.38E-04	7.14E-04	1.02E-03	2.51E-05	1.02E-03

Concluding Thoughts

This study is a starting point in answering the question of the recombination pathway's indel generation potential in maize. Overall, my results suggest that the recombination pathway is, in fact, mutagenic, and exhibits specific and consistent trends across all datasets. Crossing-over in maize seems to generate the largest amount of small indels, producing them in a gradient pattern based on size, with an uptick in the density of the smallest-sized indels (1-10bp) that tapers off as indel size increases. Small indel density also decreases in regions 2 kb upstream and downstream from the recombination sites, implying that small indels are generated in a localized fashion. Nuances in indel density among different features, like between DSBs and CO hotspots, reflect the highly active CO pathways at CO hotspots and the corresponding surge in indel density.

IV. Acknowledgements

I would like to thank Dr. Wojtek Pawlowski for his consistent guidance and support throughout my time in the lab. I also thank Ruth Epstein for being an incredible mentor from when I first started in the lab until now. Both of their advice, encouragement, and general mentorship through the duration of this project made my work possible. I also would like to thank Minghui Wang for his advice and for his work on the datasets that were the basis of my project. Lastly, I would like to thank Ryan Chaffee, Quinn Johnson, and all the other members of the Pawlowski lab for their support throughout my project as well.

References

Allen, M. R., & Smith, L. A. (1996). Monte Carlo SSA: Detecting irregular oscillations in the presence of colored noise. *Journal of climate*, 9(12), 3373-3404.

Andersen, J. R., & Lübberstedt, T. (2003). Functional markers in plants. *Trends in Plant Science*, *8*(11), 554-560.

Ananiev, E. V., Phillips, R. L., & Rines, H. W. (1998). A knob-associated tandem repeat in maize capable of forming fold-back DNA segments: are chromosome knobs megatransposons?. *Proceedings of the National Academy of Sciences*, 95(18), 10785-10790.

Fuentes, R. R., de Ridder, D., van Dijk, A. D., & Peters, S. A. (2022). Domestication shapes recombination patterns in tomato. *Molecular Biology and Evolution*, 39(1), msab287.

Gehrke, F., Schindele, A., & Puchta, H. (2022). Nonhomologous end joining as key to CRISPR/Cas-mediated plant chromosome engineering. *Plant Physiology*, 188(4), 1769-1779.

He, Y., Wang, M., Dukowic-Schulze, S., Zhou, A., Tiang, C. L., Shilo, S., ... & Pawlowski, W. P. (2017). Genomic features shaping the landscape of meiotic double-strand-break hotspots in maize. *Proceedings of the National Academy of Sciences*, 114(46), 12231-12236.

Hicks, W. M., Kim, M., & Haber, J. E. (2010). Increased mutagenesis and unique mutation signature associated with mitotic gene conversion. *Science*, 329(5987), 82-85.

Hu, J., & Ng, P. C. (2012). Predicting the effects of frameshifting indels. *Genome Biology*, 13(2), 1-11.

Hufford, M. B., Seetharam, A. S., Woodhouse, M. R., Chougule, K. M., Ou, S., Liu, J., ... & Dawe, R. K. (2021). De novo assembly, annotation, and comparative analysis of 26 diverse maize genomes. *Science*, *373*(6555), 655-662.

Kianian, P., Wang, M., Simons, K., Ghavami, F., He, Y., Dukowic-Schulze, S., ... & Pawlowski, W. P. (2018). High-resolution crossover mapping reveals similarities and differences of male and female recombination in maize. *Nature Communication*s, 9(1), 1-10.

Kole, C., Muthamilarasan, M., Henry, R., Edwards, D., Sharma, R., Abberton, M., ... & Prasad, M. (2015). Application of genomics-assisted breeding for generation of climate resilient crops: progress and prospects. *Frontiers in Plant Science*, *6*, 563.

Lardy, G. (2018). Feeding corn to beef cattle. https://www.ag.ndsu.edu/publications/livestock/feeding-corn-to-beef-cattle

Liu, H. J., Wang, X., Xiao, Y., Luo, J., Qiao, F., Yang, W., ... & Yan, J. (2020). CUBIC: an atlas of genetic architecture promises directed maize improvement. *Genome Biology*, *21*(1), 1-17.

Liu, J., Qu, J., Yang, C., Tang, D., Li, J., Lan, H., & Rong, T. (2015). Development of genome-wide insertion and deletion markers for maize, based on next-generation sequencing data. *BMC Genomics*, *16*(1), 1-9.

Lukaszewicz, A., Lange, J., Keeney, S., & Jasin, M. (2021). De novo deletions and duplications at recombination hotspots in mouse germlines. *Cell*, 184(24), 5970-5984.

Mabire, C., Duarte, J., Darracq, A., Pirani, A., Rimbert, H., Madur, D., ... & Nicolas, S. D. (2019). High throughput genotyping of structural variations in a complex plant genome using an original Affymetrix® axiom® array. *BMC Genomics*, *20*(1), 1-25.

Marand, A. P., Zhao, H., Zhang, W., Zeng, Z., Fang, C., & Jiang, J. (2019). Historical meiotic crossover hotspots fueled patterns of evolutionary divergence in rice. *Plant Cell*, *31*(3), 645-662.

McMullen, M. D., Kresovich, S., Villeda, H. S., Bradbury, P., Li, H., Sun, Q., ... & Buckler, E. S. (2009). Genetic properties of the maize nested association mapping population. *Science*, 325(5941), 737-740.

Montgomery, S. B., Goode, D. L., Kvikstad, E., Albers, C. A., Zhang, Z. D., Mu, X. J., ... & 1000 Genomes Project Consortium. (2013). The origin, evolution, and functional impact of short insertion–deletion variants identified in 179 human genomes. *Genome Research*, *23*(5), 749-761.

Prieler, S., Chen, D., Huang, L., Mayrhofer, E., Zsótér, S., Vesely, M., ... & Klein, F. (2021). Spo11 generates gaps through concerted cuts at sites of topological stress. *Nature*, 594(7864), 577-582.

Rodgers, K., & McVey, M. (2016). Error-prone repair of DNA double-strand breaks. *Journal of Cellular Physiology*, *231*(1), 15-24.

Stitzer, M. C., Anderson, S. N., Springer, N. M., & Ross-Ibarra, J. (2021). The genomic ecosystem of transposable elements in maize. *PLoS Genetics*, 17(10), e1009768.

Symington, L. S., Rothstein, R., & Lisby, M. (2014). Mechanisms and regulation of mitotic recombination in Saccharomyces cerevisiae. *Genetics*, 198(3), 795-835.

Taagen, E., Bogdanove, A. J., & Sorrells, M. E. (2020). Counting on crossovers: controlled recombination for plant breeding. *Trends in Plant Science*, *25*(5), 455-465.

Wang, M., Shilo, S., Zhou, A., Zelkowski, M., Olson, M. A., Azuri, I., ... & Pawlowski, W. P. (2022). Machine learning reveals conserved chromatin patterns determining meiotic recombination sites in plants. *bioRxiv*.

Wu, D. H., Wu, H. P., Wang, C. S., Tseng, H. Y., & Hwu, K. K. (2013). Genome-wide InDel marker system for application in rice breeding and mapping studies. *Euphytica*, *192*(1), 131-143.

Yuan, H., Yang, W., Zou, J., Cheng, M., Fan, F., Liang, T., ... & Hu, J. (2021). InDel Markers Based on 3K Whole-Genome Re-Sequencing Data Characterise the Subspecies of Rice (*Oryza sativa* L.). *Agriculture*, *11*(7), 655.

Zelkowski, M., Olson, M. A., Wang, M., & Pawlowski, W. (2019). Diversity and determinants of meiotic recombination landscapes. *Trends in Genetics*, *35*(5), 359-370.

Zelkowski, M., Wang, M., Sun, Q., Pillardy, J., Kianian, P. M., Kianian, S., ... & Pawlowski, W. P. (2022). Crossing-over decision landscape in maize. *bioRxiv*.

Zhao, M., Ku, J. C., Liu, B., Yang, D., Yin, L., Ferrell, T. J., ... & Lisch, D. (2021). The mop1 mutation affects the recombination landscape in maize. *Proceedings of the National Academy of Sciences*, *118*(7), e2009475118.