The recombination landscape of diploid wheat from a high-density genetic map of *Triticum monococcum* subsp. *monococcum* x *Triticum monococcum* subsp. *aegilopoides*

by

Cameron Henning Amos

B.S., Kansas State University, 2020

A THESIS

submitted in partial fulfillment of the requirements for the degree

MASTER OF SCIENCE

Interdepartmental Genetics Program
College of Agriculture

KANSAS STATE UNIVERSITY
Manhattan, Kansas

2022

Approved by:

Co-Major Professor
Dr. Dal-Hoe Koo

Approved by:

Co-Major Professor
Dr. Jesse Poland

Copyright

Abstract

Einkorn wheat (*Triticum monococcum*), one of the oldest cultivated grains in the world, was domesticated during the early Neolithic period in South-Eastern Turkey. Einkorn played a critical role in supporting the unrivaled societal and population growth in and around the 'Cradle of Civilization'. Einkorn wheat includes both wild einkorn (*Triticum monococcum subsp. aegilopoides)* and its domesticated form (*Triticum monococcum subsp. monococcum*), which is the only domesticated diploid wheat species. Another diploid A genome species morphologically similar to *Triticum monococcum* is *Triticum urartu*, being the direct source of the A-genome in tetraploid and hexaploid wheat. *Triticum monococcum* was cultivated for several centuries before free-threshing durum wheat (*Triticum durum*) and hexaploid bread wheat (*Triticum aestivum*) emerged from later hybridization events. The A-genome holds impressive nutritional content and high genetic polymorphism, and its genetic potential for wheat improvement has sparked renewed interest for this ancient crop. Here we investigate a mapping population of a cross between domesticated (*Triticum monococcum subsp. monococcum*) and wild einkorn (*Triticum monococcum subsp. aegilopoides*) to understand the structure, function, and evolution of the A genome. Two parents of a recombinant inbred line (RIL) population were sequenced to ~5x coverage to discover variants between the domesticated einkorn parent and the wild ancient progenitor parent. Two sets of derived RILs from these parents were skim-sequenced, including 93 RILs at 0.2x coverage and 786 RILs at 0.028x coverage. These populations were aligned to a newly developed reference for *T. monococcum* to identify recombination breakpoints in the RILs and evaluate the genome-wide recombination landscape in einkorn. We identified recombination hotspots and investigated the positional trends and associations within each chromosome between crossovers, genes, and variant SNPS. We

observed a clear and strong positive correlation between the recombination density, gene density and distance from the centromere. There was a more drastic increase for both crossover and gene density with distance from the centromere observed than previously described in hexaploid wheat, rice, and maize. Regions with high crossover frequency reached ~3 cM/Mb in the short arms and ~1.5 cM/Mb in the long arms near the telomeres. Likewise, gene-rich regions were ~3x as gene-dense compared to proximal region of the same chromosome arm. Spikes in crossover frequency and gene density were also positionally correlated, especially within the distal halves of the arms. In this study we have demonstrated the utility of using very low coverage 'skim-sequencing' for accurate genotyping of genetic populations while revealing and comprehensive picture for the recombination and gene landscape of einkorn.

Table of Contents

List of Figures

List of Tables

Chapter 1 - Introduction

The domestication of cultivated tetraploid (*Triticum turgidum,* AABB) and hexaploid wheat (*Triticum aestivum,* AABBDD) has caused significant genetic bottlenecks within each subgenome (Feldman & Levy, 2012). To combat modern wheat's genetic erosion and increase diversity for breeding, einkorn (*Triticum monoccum*) and other A-genome species (2n = 2x = 14, AA) have gained increased attention due to their higher levels of diversity. Such diploid species could be used to introgress genes into tetraploid and hexaploid wheat to increase the A-genome diversity (Brunazzi et al., 2018; Mondal et al., 60 2016). Desirable traits can be found in einkorn that are of interest for introgression into bread wheat lines, including disease and pest resistance, higher protein and mineral content, and abiotic stress tolerance (Zaharieva 2014). Importantly, an increased knowledge of the A genome from diploid wheat species would hasten the creation of healthier and more diverse cultivars in breeding programs.

Einkorn also holds interest as a food crop itself. The Neolithic Revolution was spurred by the domestication and cultivation of the first crops, with einkorn being the first domesticated and cultivated cereal crop. It was domesticated in the Karacadağ mountain region circa ~9600 BCE (Brandolini et al., 2016), with wild einkorn still growing in The Near-East to this day. Cereal grains were the staple food source of the first ancient civilizations including the Sumerians, as well as a central cultural aspect of the oldest known civilization that enabled food production at a level suitable for the first urbanization and larger organization of society. In fact, the first writing system that allowed the administrative cohesion of multiple cities may not have developed without the logistical necessity of tallying agricultural yields (Pae, 2020). So

important and conjoined were agriculture and writing in Sumerian culture that scribes commonly ended their tablets with a praise to Nisaba, a dedicated god to both scribal arts and grain. In the sympatric association of domestication and the first cultivation of cereal grains, the artificial selection of wild einkorn into domesticated einkorn brought larger seeds and an earlier heading time (Brandolini & Heun, 2019). Importantly, the development of a non-shattering indehiscent head, caused by a loss-of-function mutation of the *Brt1* gene on chromosome 3A, allowed for much easier harvests (Pourkheirandish et al., 2018).

A genetically distinct undomesticated A-genome wheat species, *T. urartu* (Middleton et al., 2014), is the diploid ancestor of the modern polyploid wheats. Hybridization of *T. urartu* with a close relative of *Aegilops speltoides* ~500,000-150,000 years ago (Charmet, 2011) formed the wild tetraploid Emmer wheat (*T. turgidum L. subsp. dicoccoides*) (AABB, 2n = 4x = 28). This grass was domesticated into the free-threshing durum wheat (*T. turgidum subsp. durum*), and later naturally hybridized with goat grass (*Aegilops tauschii*, DD, 2n=2x=14), the D genome donor, ~10,000 years ago to form hexaploid bread wheat (*Triticum aestivum* L., 2n=6x=42) with AABBDD genome composition (Charmet, 2011).

Within the A-genome species lie important genetic resources for a myriad of beneficial traits ranging from nutritional content to disease, pest, and abiotic stress resistances (Zaharieva 2014). This includes the first gene cloned in polypoid wheat that confers resistance to the recent and devastating Ug99 stem rust fungus-- the Sr35 gene derived from the long arm of

chromosome 3A of einkorn (Saintenac et al., 2013). Thus, it is of great importance to survey A-genome species for wheat improvement.

Genetic Recombination

Knowledge how and where crossovers occur in a chromosome-wide scope is key to understanding how genomes differentially evolve along the chromosomes as well as important implications for creation and selection of new haplotypes in breeding and crop improvement. Recombination rates are known to vary substantially throughout the genome. In a well-studied and wide-sweeping trend of higher plant species it was shown that crossovers increase with distance from the centromere within chromosome arms (Anderson et al. 2006). In wheat and rice, this skewed distribution has been found to cause an uneven loss of synteny in the distal regions of the chromosome (Anderson et al. 2006). Additionally, crossover occurring in the pericentromeric region or within the centromere itself can cause a failure of proper chromatid separation, producing aneuploidy and infertile gametes (Nambiar and Smith 2016). Overall, it is well known that crossovers are highly suppressed in the centromeric regions of plant chromosomes, with 50% to 75% of the central region of chromosomes in *Triticeae* species having highly restricted crossover recombination (Anderson et al. 2006).

Meiotic recombination is initiated by a double-stranded break (DSB), created by a pair of topoisomerase-related Spo11 endonucleases. Although hexaploid wheat is known to have two homologues (Spo11-1 and Spo11-2), in *Triticum urartu* only one homologue is present, similar to yeast and humans (Benyahya et al. 2020). A 5'-3' end resection occurs via the MRX complex--

a chimeric protein complex consisting of Mre11, Rad50, and Xrs2-- and removes the Spo11, which generates single strand dangling ends up to thousands of bp long. Rad51 and DMC1 guide the dangling ends to similar sequences on the recipient DNA on the opposing chromatid. DNA polymerase extends the dangling end after invasion to create a Holliday junction. To resolve the junction, the yen1 (plant ortholog of human gen1) resolvase cuts the substrates to form either crossover or non-crossover events (Wang and Copenhaver 2018). Upon a successful crossover, all distal chromatin is switched between the two homologous chromosomes. Most non-crossover events exchange only 500 bp or less of genetic material between the two homologous chromosomes (Wang and Copenhaver 2018).

Where the DSBs can occur and to what state they resolve is influenced by several factors, including DNA content and methylation, chromatin and histone structure and layout, large-scale chromosome features such as chromatin loops, the availability of gene promoters, as well as external recruiters of Spo11 (Wang and Copenhaver 2018, Pan et al. 2011). It has been shown in wheat that recombination rate and gene density increase with physical distance from the centromere (Anderson et al. 2006). This trend is found across plants, but is more pronounced in the large chromosomes of Triticeae than other plant species. The strong suppression of crossover events in the pericentromeric region is to ensure a successful anaphase, separating chromatids evenly between the daughter cells (Floutsakou et al. 2013). Correlations between gene density and breakpoints and between gene density and distance from the centromere of the mitotic chromosome were found in hexaploid wheat (Akhunov et al. 2003). The local availability of gene promoters that Spo11 can targets is among the most

significant contributors to DSB formation, and accounts for ~50% of the variation in recombination rates in maize (Anderson and Stack 2006). Although methods have been described to identify recombination breakpoints indirectly, such as detecting proteins present in recombination nodules which are correlated with crossovers (Anderson and Stack 2005), directly identifying meiotic recombination hotspots on a genome-wide scale requires a high number of offspring lines to capture genetic crossovers as well as sufficient marker density to grant satisfactory resolution for crossover detection.

Here, we undertook a detailed study of recombination landscape in the diploid wheat *Triticum monococcum*. We have sequenced two parents of a recombinant inbred line (RIL) population to ~5x coverage to discover variants between the domesticated einkorn parent and the wild ancient progenitor parent (subsp. *aegilopoides*). Two sets of derived RILs from these parents were skim-sequenced, including 93 RILs at 0.2x coverage and 786 RILs at 0.028x coverage, respectively. With genome annotations from *T. monococcum* reference accessions, we investigate the distribution of crossover recombination sites and genes, the localized correlations between them, as well as the effect of chromatin content and other factors impacting crossover breakpoints within the seven chromosomes.

Chapter 2 - Materials and Methods

Einkorn Lines Utilized

The Recombinant Inbred Line (RIL) population studied here was derived from a cross of the same two parental accessions of *Triticum monococcum subsp. aegilopoides* (accession TA4342-L95, referred to as P1) and *Triticum monococcum subsp. monococcum* (accession TA4342-L96, referred to as P2). The parental lines were crossed and inbred via single-seed-descent to form the RIL populations. Seeds from the original parental accessions were grown in 2020 and sequenced to >5x coverage.

The first, original RIL population contained 93 lines and were advanced via single-seed-descent to the F_6 generation, planted as described by Singh et al. 2007. The second RIL population contains 768 lines which were initiated in 2013 and advanced to the F_8 generation. The original and new populations were skim-sequenced at 0.2x and 0.028x coverage, respectively, with corresponding Nextera library IDs NEX0021 and NEX0025.

We utilized reference genome assemblies from a larger project to assemble and annotate the *T. monococcum* genome (Ahmed et al. 2022). The reference genome assemblies were constructed for two accessions including subspecies *T. monococcum subsp. aegilopoides* (accession number TA299) which was originally collected in Dahuk, Iraq, and an accession for the domesticated subspecies *T. monococcum subsp. monococcum* (accession number TA10622) originally collected in Permet, Albania.

Genotyping RIL Population

Multiplexed DNA libraries were constructed using low-volume Nextera libraries for the RIL populations (NEX0021 and NEX0025), which were demultiplexed using a custom Perl script which utilized the dual index reads from the Illumina sequencing (Adhikari et al. 2021). Fastp v0.23 was used for adapter trimming, with qualified_quality_phred as 20, length required as 75. A custom python script was used to count unknown reads and contrast the amount of reads in a blank versus a non-blank well to ensure DNA plates were correctly assigned to libraries and sequencing files. Hisat2 2-2.2.1 with flags --no-spliced-alignment --no-unal --rg-id, was used to align the paired end reads to the *T. monococcum subsp. aegilopoides* assembly of accession number TA299. Samtools 1.6 view and sort command were then used to process the .sam file into a sorted and indexed .bam file. Samtools view was then used again to filter out broken reads in conjunction with the awk command.

Variant Discovery and Genotype Calling

From the high-coverage parental fastq files, Hisat2 and Samtools was again used to align reads to the reference genome (the *T. monococcum subsp. aegilopoides* parent). In order to filter for uniquely mapped and concordant read pairs, that is, to filter out reads that are mapped multiple times to the reference as well as pairs who are misoriented or mapped with unsuitable distances between them, samtools' view, sort, and index commands were used to create a .bam file of unique and concordant reads. Bcftools' 1.10.2 mpileup command (with flag –skip-indels) was utilized in junction with the call command (flags -m –variants-only –skip-variants indels) to identified variants between the two parents using the reference genome associated.

We then utilized a custom Perl script to filter the resulting .vcf file and kept variants with minor allele frequency (MAF) > 0.1, less than 30% missing, and heterozygosity > 5%. Incorrect genotype calls on the parents were identified— reads who were under a skim-seq repetition of a parental line but were found to align to the contrasting parent—and filtered out as well. Bcftools was used again to filter out SNPs that have a depth less than 3x coverage. This filtered variant set was then called on the entire RIL population into two final *.vcf* files, which were then merged.

Allele Identification and Processing

From the *.vcf*, a combination of bash and python scripts were used to format the information into a tab-delimited file, with the header row detailing the columns underneath them: chromosome number, marker position in base pairs, parent 1 (L95), parent 2 (L96), followed by each RIL ID. Underneath all sample line headers were the nucleotides found at that variant site. The nucleotides of the RILs were compared against the nucleotides for the parental lines, and would be labelled as which parent they derived that variant from. This information in turn was used to generate an individual file for each RIL containing every variant marker. All SNPs that were heterozygous or missing-- those that were found to not belong to either parent-- were filtered out using the sed command. This final filtered dataset included the parental genotype for each position in each RIL.

Recombination Breakpoint Detection

To determine the most accurate position of recombination breakpoints, both the macroscopic region layout-- its general allelic consensus, and obvious trends of regions at a chromosome-scale-- must be utilized to account for misidentified markers and thin data as well as the microscopic region-- marker to marker precision-- to ensure the precise location of the recombination break is identified between the two differing markers with as much resolution possible. Additionally, even at the F_6 to F_7 inbreeding generation, heterozygous regions should still constitute roughly 1-2% of the genome for any given RIL, making the ability for the algorithm to detect heterozygous sites crucial. Although there are previously described crossover detection methods (Huang et al. 2009), these lack the robustness to properly identify breakpoints in low marker density regions, identify heterozygous regions with precision, and maintain resolution in sequencing error-prone data. A novel and robust breakpoint identification method was thus developed to account for low marker density, high sequencing error, and physical position of the markers (Amos, Cameron. 2022). First, a roaming score was sequentially calculated along an entire chromosome: an allele belonging to parent 1 (L95), would be summed as +1 point. An allele belonging to the second parent, parent 2 (L96), would sum as -1 point (figure 1).

Each variant site with a neighboring allele of the alternate parents were identified as ‘bends’, and a moving window was used to analyze markers to the left and right of these positions. A dual window of a combined width (or diameter) of 6,500,000 bp was used instead of a fixed number of markers to account for varying marker density. After markers had been selected

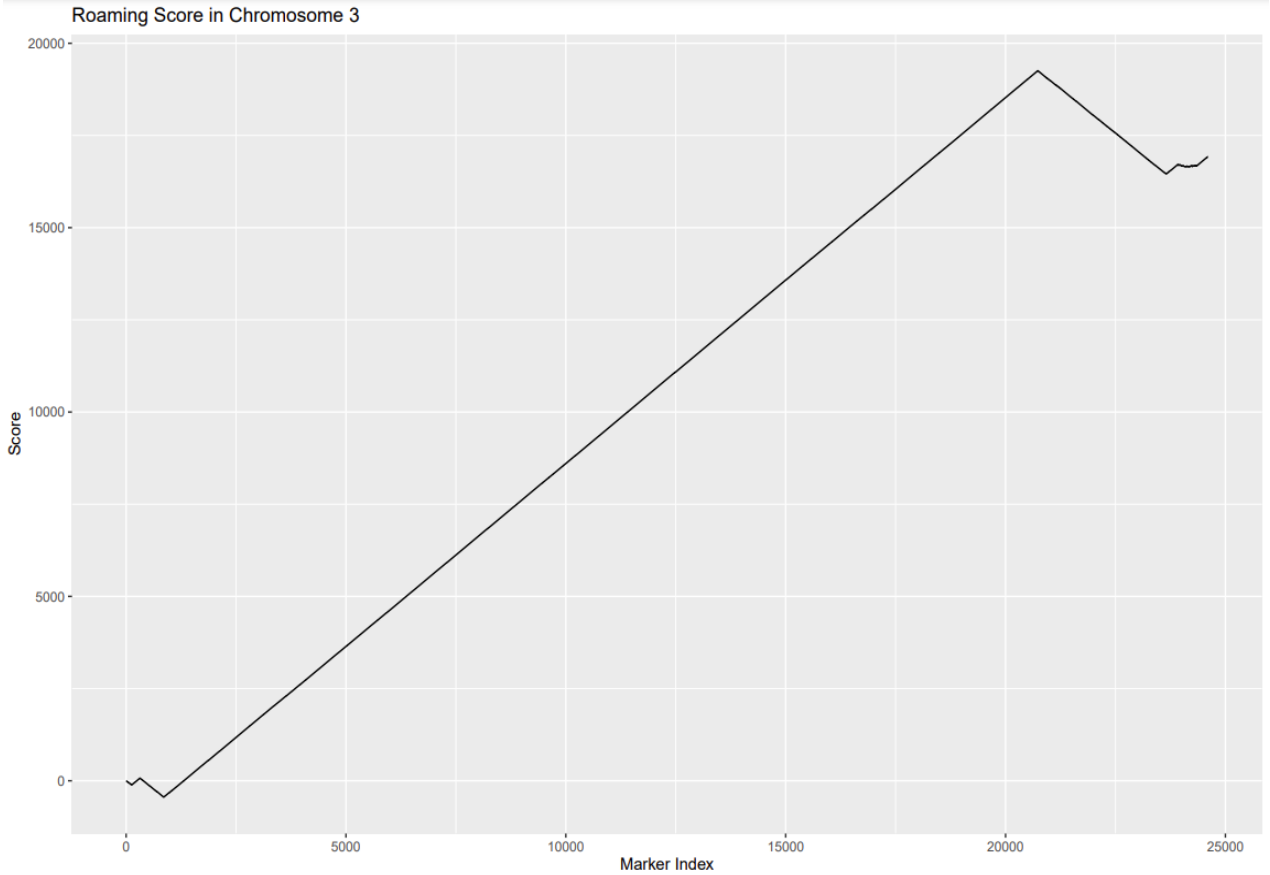

Figure 1.--Roaming score of Chr 3 on RIL 2013-18-107.

X-axis representing the marker index of the variant, y-axis representing the roaming score based upon the parental allele. Positively sloped regions signify chromosomal regions belonging to P1, negatively sloped regions to P2, and flat or 0-sloped regions signifying a heterozygous site.

inside both windows, a simple linear fit using least squares regression was performed on each side to determine the slope of the line of best fit. If either paired window had too low of marker density as defined as less than 25% of the average marker density of the chromosome or a window was represented by only 3 markers, it was skipped and marked as low confidence. If the slope of the line was within a given tolerance of +1, -1, or 0, and the collinearity of the points were above a threshold, then that side of the window would be designated as a P1, P2, or Heterozygous region respectively. A tolerance of ± 0.2 was allowed for a homozygous window, and a tolerance of ±0.4 was allowed for a heterozygous window. This leaves the ranges -0.8 to -0.4 and 0.4 to 0.8 unassigned.

Each window was given a score based on the difference between the two slopes of each side. If the moving window pair had differing regions on either side, it was considered as a Candidate Recombination Breakpoint (CRB). If the moving window had a score below 0.1 for the difference in slope between the two lines of best fit, then that position would be marked an 'anchor' site, as its sides represent very similar regions. If two bends were more than half the distance of the sliding window away from each other, that position was also tested and would result in an anchor site between them to ensure no region was unrepresented.

These CRBs and anchor sites (with their respective information) were arranged in a list by physical position. Sites were then filtered so that only the highest scoring CRBs and lowest scoring anchor sites exist within their local neighborhood (determined by the window radius of 2.25 Mb), so that only the highest-scoring candidates of their region-to-region transition remain. Anchor sites were merged if adjacent. Finally, using the anchor sites, the continuity between each site was tested-- that is, that the same type of region (P1, P2, or Heterozygous) would be matching from the right side of a CRB to the left of the next CRB continuously through the chromosome.

A recursive logic tree then filtered the CRBs down to the candidates that best represent the marker data. First, CRBs were removed if either of the two regions that were determined to be on either side of it could not match with any other CRB between two anchor sites. The highest scoring duplicate CRB between two anchor sites whose left and right regions matched the left and right anchor sites adjacent was kept, the duplicate being removed. After filtering, if there

were no surviving CRBs between two differing anchor sites (there was not a recombination breakpoint found that fits the region transition required between the two anchor sites), a secondary macroscopic search algorithm begins. This approach is useful to finding a transition point between two differing regions in a wider scale. It first identifies the width of the uncertain region. The algorithm determines if it is searching for an upward or downward bend from the region types of the anchors. For example, a region transition from P1 to P2 would mean that the region between the +1 and -1 sloped anchor sites must bend downwards at some point. From between the closest two anchor sites, a sliding window of diameter n markers, with n equal to the marker distance between the two anchor sites, would investigate every bend of the data. Each score is recorded. The window with the highest score, and with the expected bend (upwards or downwards), would be inserted as a CRB between the two anchor sites. Upon reaching continuity through every anchor site and to each end of the chromosome, the CRBs were then considered high-confidence recombination breakpoints (RBs), and the recombination breakpoints of each chromosome under each RIL were recorded.

Recombination Landscape

Each crossover interval was defined by the informative flanking markers. These intervals were aggregated and sorted into a single dataframe. For each crossover recombination size, the nearest informative flanking markers were recorded with respective position on the genome assembly. The physical distance between the two markers in which the breakpoint was detected was calculated for all crossover sites to represent the width of the recombination interval.

It has been shown previously that a major factor that determines recombination breakpoint distribution is the presence of genes, and actively transcribed genes in particular. Finding a higher frequency of recombination intervals containing genes than would be expected from a random distribution would corroborate this thought. To test if crossovers more often occur in gene-rich regions, we utilized the gene annotations for the *T. monococcum subsp. aegilopoides* reference (TA299) (Ahmed *et al*., *submitted*). The gene models were classified as high-confidence (HC) genes if a start and stop codon was identified in the UniViri/UniPoa/BUSCO poales database, and was classified as low-confidence (LC) genes if not, as described by Adhikari et al 2022.

To statistically test if crossover intervals are associated with the presence of a gene in the respective interval, we first calculated the percentage of crossover recombination intervals that contained at least one gene. To determine if this was significantly greater than a random intersection of recombination intervals and genes, we created two control groups consisting of (1) random recombination intervals with observed gene positions and (2) observed recombination interval positions with random gene positions. For each of these permutation sets we conducted random sampling of the same observed interval size and gene sizes from its respective group. Each group was randomly sampled 100 times. We also sought to test if recombination breakpoints occur at a higher rate around genes annotated as high-confidence (HC) compared to low-confidence (LC) gene annotations, that is, genes with annotations that lack support from additional sources and only partially supported by gene models, including gene fragments, and genes without or low homology support. To do this, a unique approach

was required to normalize the positional information of the genes, instead of the resulting breakpoint to gene associations. As LC genes outnumbered HC genes by nearly 3-1 in every chromosome, LC genes were randomly selected and removed until the LC gene count equaled that of the HC genes in a particular chromosome. This process was repeated 100 times across each chromosome, and the resulting association scores for each breakpoint were averaged. The results were deemed significant by utilizing 95% confidence interval.

Chapter 3 - Results

Population Statistics

The parental lines were sequenced to approximately 8.0-fold coverage (44 Gb) for the *aegilopoides* parent and 8.5-fold coverage (46.6 Gb) for the *monococcum* parent. The *aegilopoides* reference contained a GC content of 0.462. A total of 728,706 filtered variants were detected between the parents, with an average of 335 lines represented in any given variant. 812 RILs remained after removing parental repetitions and blanks and filtering out lines which were mixed from a different RIL or contaminated. These lines were identified by excess heterozygosity or by a distinct and consistent 2:1 allele ratio that can only arise from the mixing of two RILs.

Recombination and Gene Landscape

Using the developed algorithm, we identified the positions of 15,919 putative crossover breakpoints. The algorithm was able to accurately denote crossover events using skim-seq data of down to 0.028x coverage within an average interval of 219 kbp. Based on a population size of 812 RILs, this corresponds to a total map length of 1,960 cM, consistent with the expected genetic map length of a diploid wheat for an RIL population. In total, between the 812 lines, 50.0% of the observed alleles were attributed to L95, 44.9% to L96, and the remaining 5.1% of genotypes were heterozygous. Consistent with the known landscape of recombination in wheat and related species, low recombination rates were observed across all seven chromosomes in the centromeric and pericentromeric regions (Figure 2).

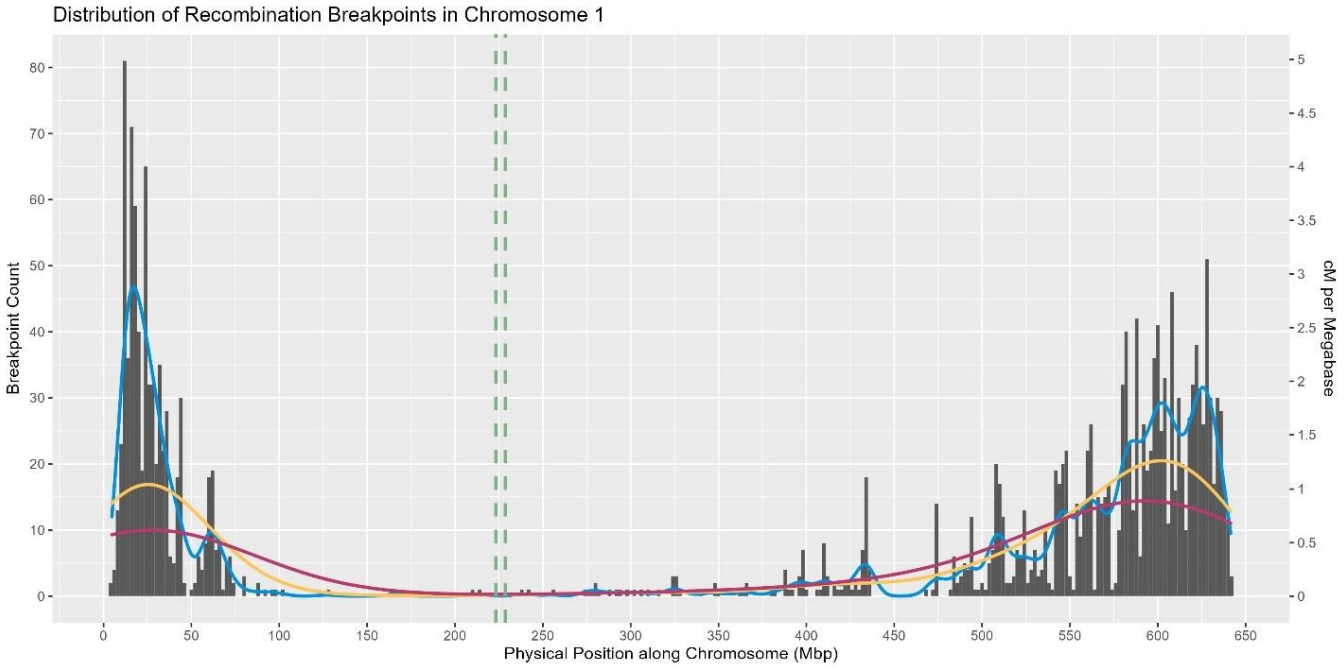

Figure 2.--Recombination landscape of chromosome 1.

Plotted in 2 Mb bins with three moving averages of varying bandwidths. The green dotted lines mark the bounds of the centromeric region.

Accuracy of Mapping Recombination Intervals

The genotypes for the RIL population were obtained using low coverage (0.028x to 0.2x) skim-sequencing. With extremely high-density variants on the parents and low-coverage sequencing on the progeny, we still expect a relatively high marker density for delineating breakpoints. We observed on average 264,825 genotype calls per RIL. Based on this genotype density there would be on average one typed variant every 19.3 Kbp, mostly uniform through ought the chromosomes save for highly repetitive regions where sequences struggle to align. Thus, we investigated the accuracy of precisely mapping crossovers with the skim-seq data. For each

recombination breakpoint interval, we identified the informative flanking markers and calculated the size of defined intervals (Figure 3). We observed a median interval length of 114 Kbp and a mean of 219 Kbp for the delineated crossover breakpoints.

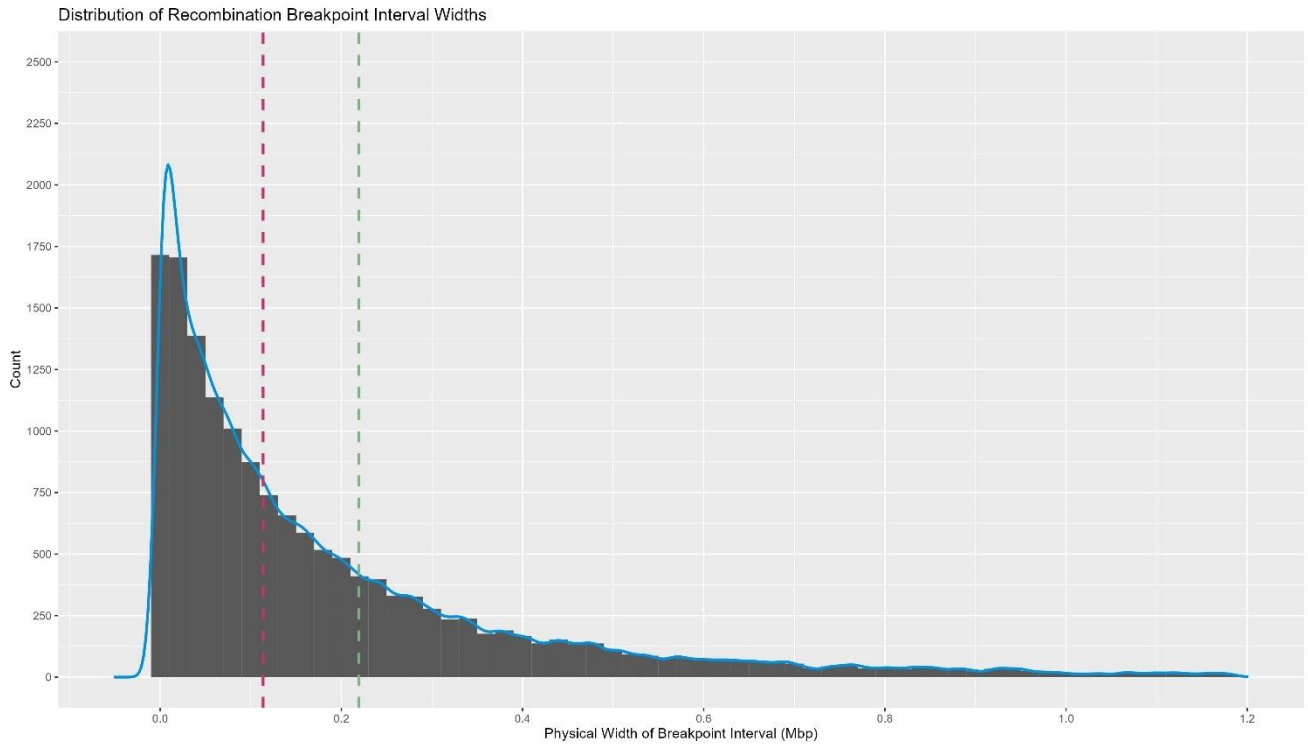

Figure 3.--Distribution of recombination breakpoint interval widths.

The red vertical line denotes the median of 0.114 Mbp, the green vertical line denoting the mean of 0.219 Mbp.

A handful of recombination events were detected within a centromeric region across all lines, and upon manual investigation, each detected a short homozygous region just within the borders of the centromere. This was likely caused by a high number of errant variant markers misaligning into the repetitive region of the centromere, and is not evidence of genuine recombination within the centromere. Therefore, these putative crossovers were removed from further analysis.

Gene Density and Recombination Rate

There is a well-studied correlation between genes and recombination breakpoints, for both crossovers and gene conversion (Pan et al. 20, Baudat and Nicholas 1997). Therefore, we examined the gene density distribution across the chromosome against the relative recombination rate. Crossover frequency across all seven chromosomes were significantly higher in the distal regions of the arms, the trend increasing by over 100-fold in crossover frequency toward the distal ends as compared to centromeric regions.

LC gene density across all seven chromosomes maintained a density of roughly 15 genes/Mb within the proximal and central arm, increasing significantly and doubling the inner-arm gene density roughly 40 Mb from the telomeres. This was a similar U-shaped trend as observed with the crossovers. The change in gene density from the pericentromeric region to the distal arms increased more drastically in the high confidence (HC) genes (Figure 4) than in the low confidence (LC) genes (Figure 5). There were 2.75x as many LC genes identified as HC genes. Chromosome 4 had the lowest gene densities, with 3.830 HC genes/Mb and 16.171 LC genes/Mb, and chromosome 5 with the highest with 4.915 HC genes/Mb and 21.823 LC genes/Mb, mirroring the RB densities. Variant count and recombination statistics reveal 4A having the highest number of variants per Gb and the lowest number of crossovers per RIL per Gb (Table 1).

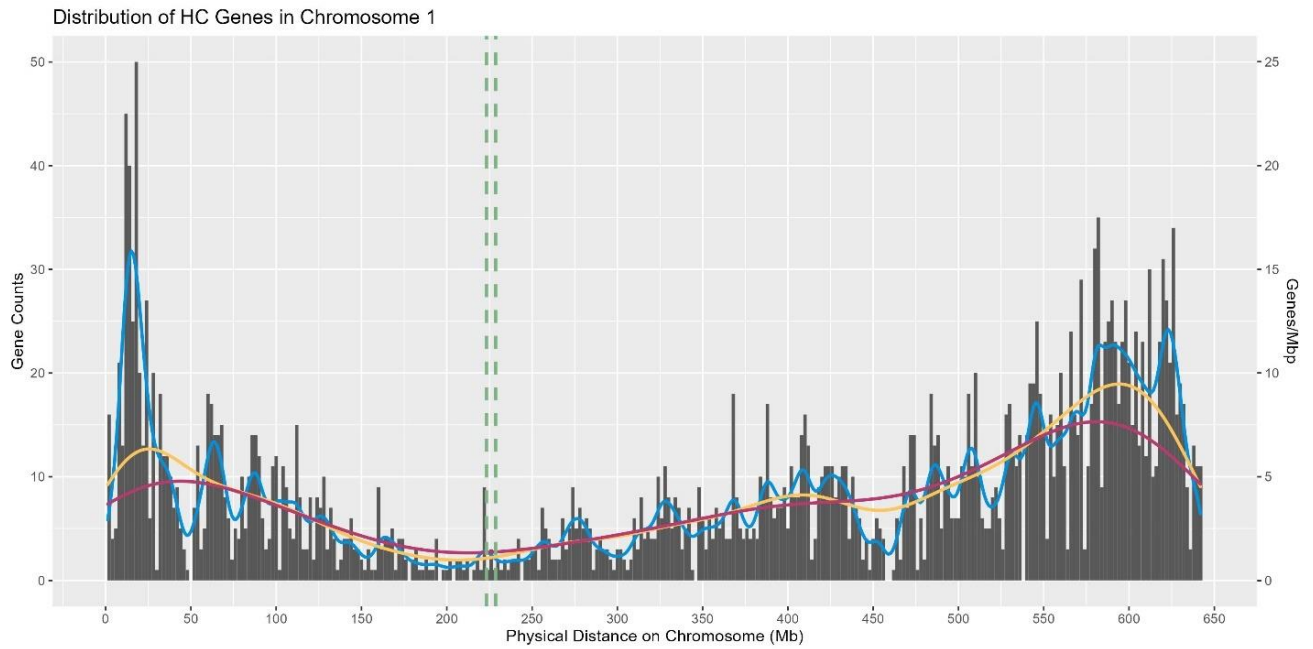

Figure 5.-- HC gene distribution in chromosome 1.

Plotted in 2 Mb bins with density curves of varying widths. The green dotted lines mark the bounds of the centromeric region.

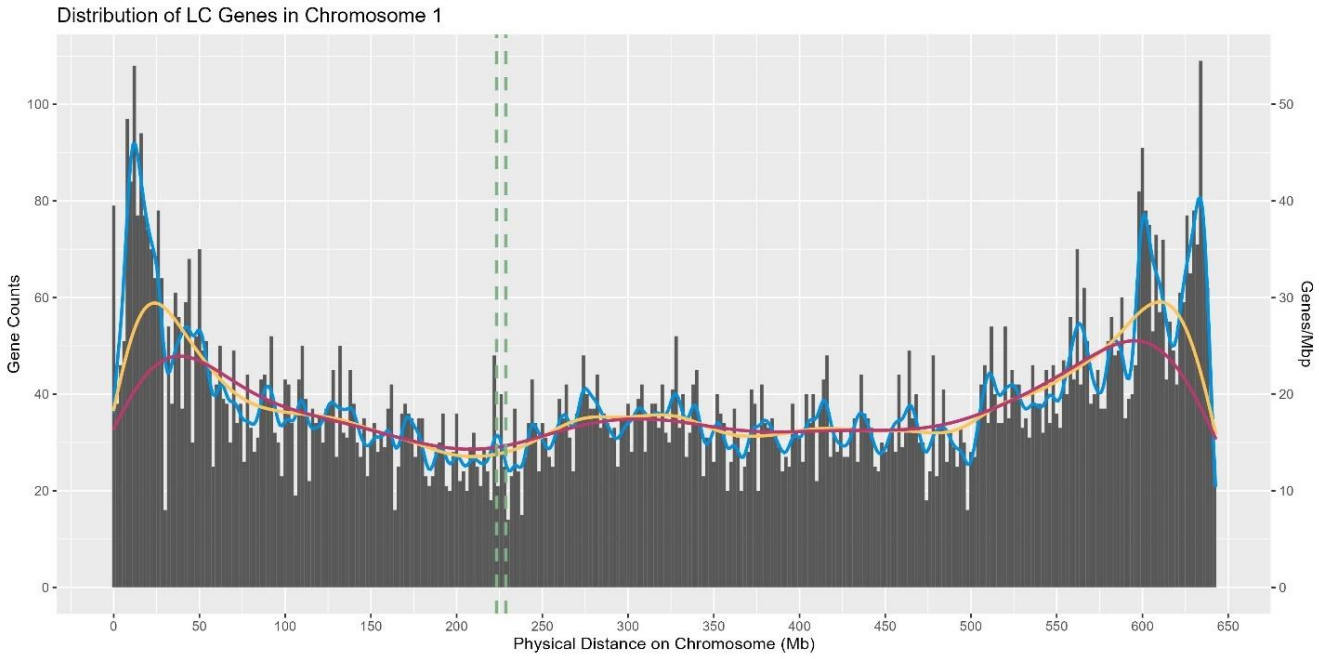

Figure 4.--LC gene distribution in chromosome 1.

Plotted in 2 Mb bins with density curves of varying widths. The green dotted lines mark the bounds of the centromeric region.

Chr	Length (Mb)	Total Variants	Variants / Mb	Total RBs	RBs / Mb	Total Genes	Genes / Mb	HC Genes	HC / Mb	LC Genes	LC / Mb
1A	642.6	1,794k	2,792	2,104	3.274	15,329	23.855	4029	6.270	11806	18.372
2A	827.8	2,203k	2,662	2,340	2.827	18,456	22.295	5358	6.473	13831	16.708
3A	819.6	2,260k	2,758	2,450	2.989	17,989	21.949	4992	6.090	13582	16.571
4A	657.9	2,099k	3,191	1,700	2.584	13,159	20.002	3717	5.650	10046	15.270
5A	735.2	1,974k	2,685	2,791	3.796	19,659	26.740	5042	6.858	15181	20.649
6A	640.6	1,767k	2,760	1,860	2.904	14,017	21.881	3908	6.101	10693	16.692
7A	793.0	2,013k	2,540	2,674	3.372	18,312	23.092	5183	6.536	13673	17.242
Ave	731.0	2,016k	2,770	2,274	3.107	16,703	22.831	4604	6.282	12687	17.358

Table 1.--Chromosome Statistics.

Variant, recombination breakpoints, and gene count statistics for each chromosome and the average derived from the combined 812 lines.

Centromere-Telomere Recombination and Gene Distribution

We found that both detected crossovers and genes of high and low confidence positively correlated with distance from the centromere. The positional information for all seven chromosomes for both datasets were translated to relative distance from the centromere and merged for both long and short arms. Although in a study by Akhunov (2003) this trend for crossovers was fitted to a squaring function, we found that an exponential function calculated via a lossless smoothing function was a better fit for both crossover (Figure 6) and gene distributions for HC and LC genes. A scalar was added to the gene distribution exponential equations that was not needed in the RB distribution due to the recombination rate in the pericentromeric region dropping to virtually zero, while the average gene density never fell below a density of ~20 genes/Mb. Across the seven chromosomes, there was a general increase of 3.38 recombinations per Mb of chromosome length (R^2 = 0.488).

The recombination distribution for 4A was an outlier among the chromosomes, possessing a more significant difference between its pericentromeric region and its distal regions, with an x coefficient of 12.1 (Figure 7). This is potentially due to its high number of variants or a stronger pericentric suppression of crossovers. The recombination density slopes of the short arms were also steeper than that of the long arms, with an x-coefficient of 7.06 compared to 4.99 respectively, potentially due to the centromere exerting a similar range of influence on both arms, which would comprise a larger proportion of the short arm.

HC genes were scarcely found in the proximal third of both arms, creating more of a V-shape than a U-shaped distribution (Figure 8). The HC density slope was more gradual, with an x coefficient of 2.87 compared to LC genes' 6.40 (Figure 9), due to the higher pericentromeric distribution of the latter.

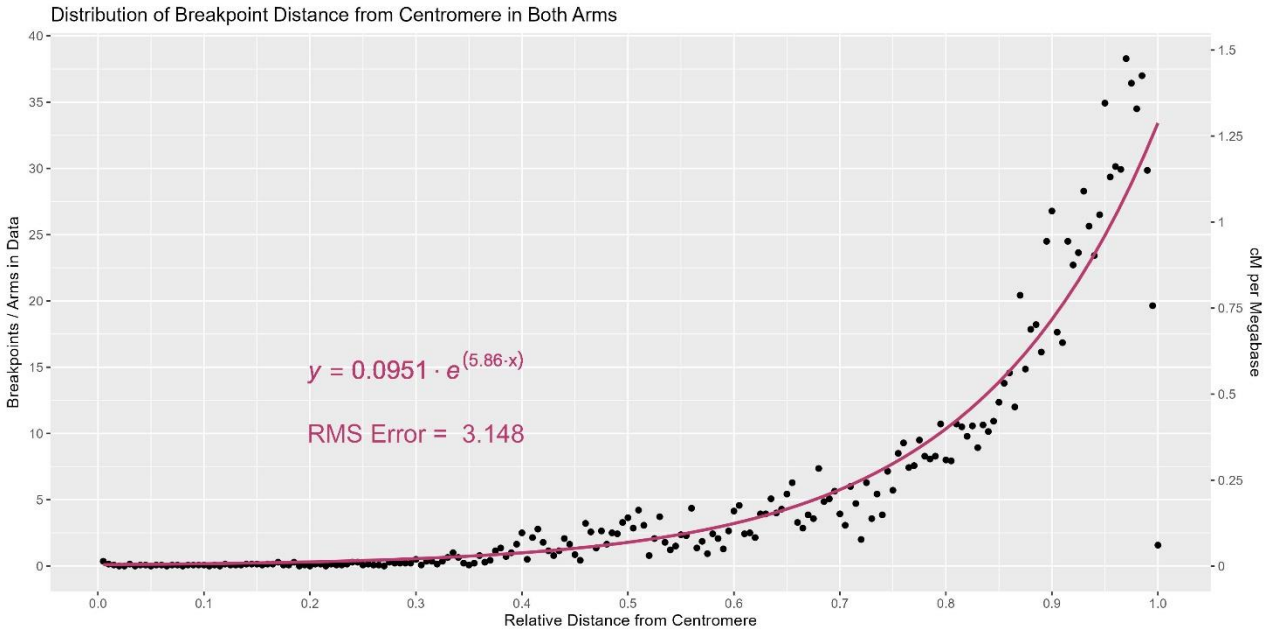

Figure 7.--Relative crossover positions of all chromosomes and arms.

The relative position of chromosome bins along the arm, demarcated into 0.5% arm length bins with the number of crossover breakpoints in the respective bin. On the horizontal axis, 0.0 represents the centromere, with 1.0 representing the telomere. On the left y axis, the number of observed breakpoints divided by the number of arms in the data (2x7 = 14). On the right y axis, the cM/Mb mapping distance.

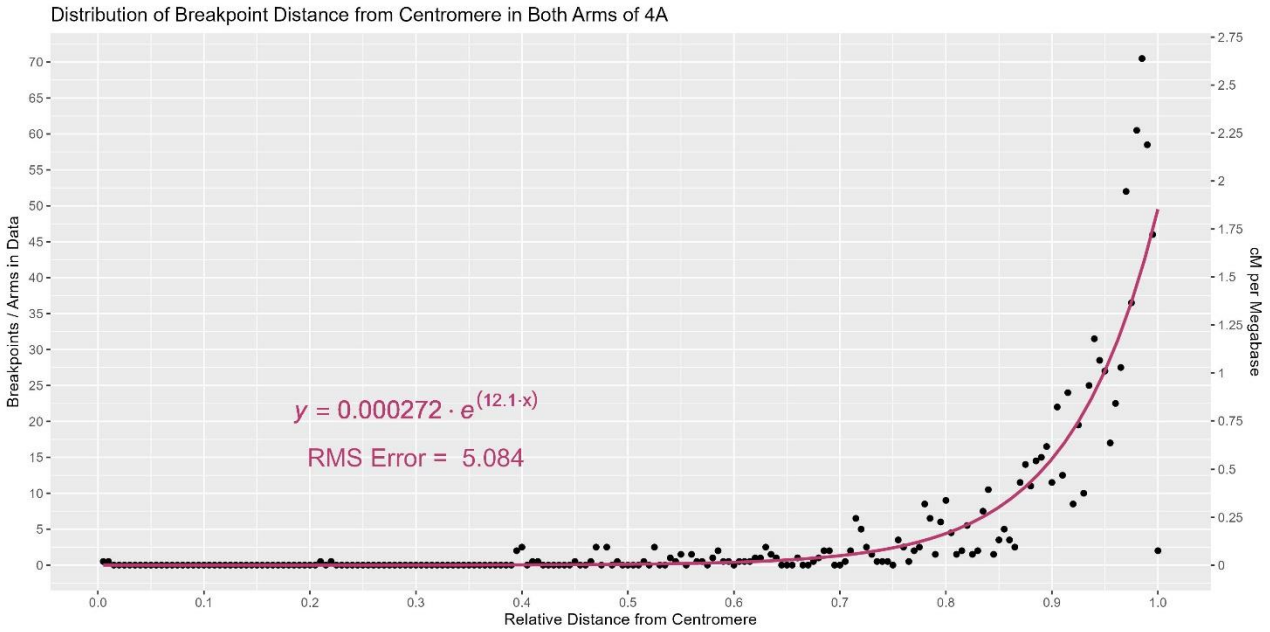

Figure 6.--Relative breakpoint positions of both arms of chromosome 4A.

The relative position of chromosome bins along both arms of chromosome 4A, demarcated into 0.5% arm length bins with the number of crossover breakpoints in the respective bin. On the horizontal axis, 0.0 represents the centromere, with 1.0 representing the telomere. On the left y axis, the number of observed breakpoints divided by the number of arms in the data (2x7 = 14). On the right y axis, the cM/Mb mapping distance.

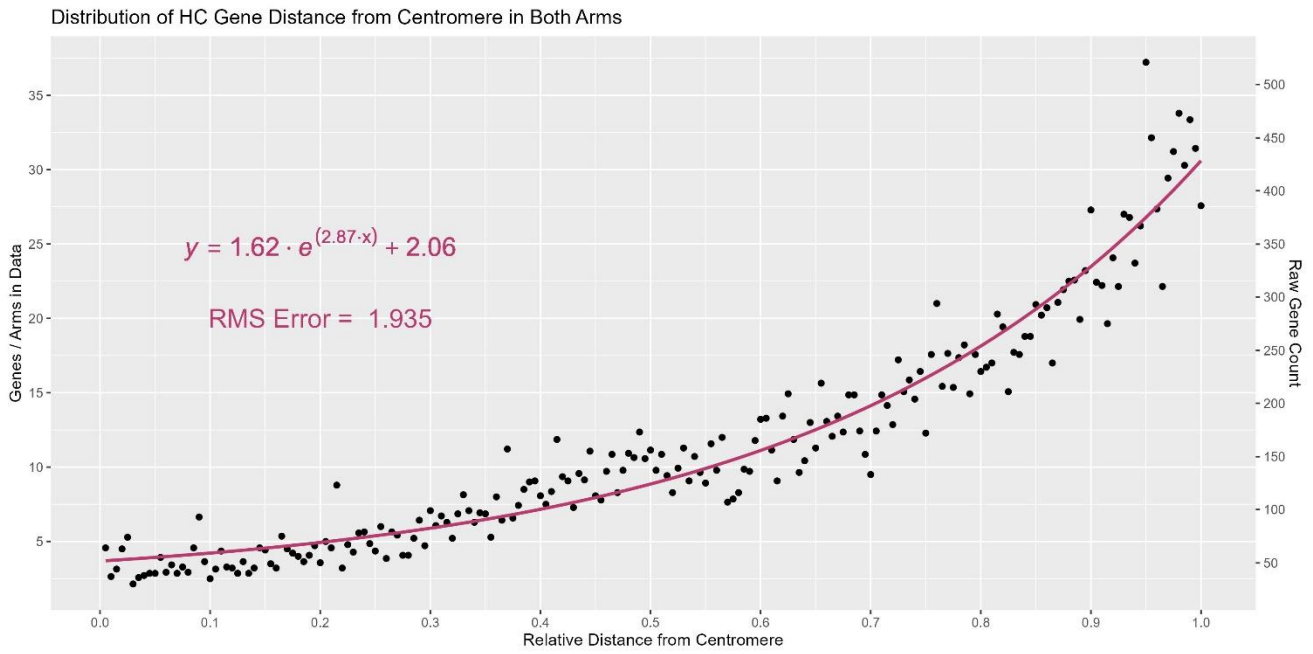

Figure 8.-- Relative HC gene positions of all chromosomes and arms.

The relative position of chromosome bins along the arm, demarcated into 0.5% arm length bins with the number of HC genes found in the respective bin. On the horizontal axis, 0.0 represents the centromere, with 1.0 representing the telomere. On the left y axis, the number of observed HC genes divided by the number of arms in the data (2x7 = 14). On the right y axis, the raw HC gene count within the bin.

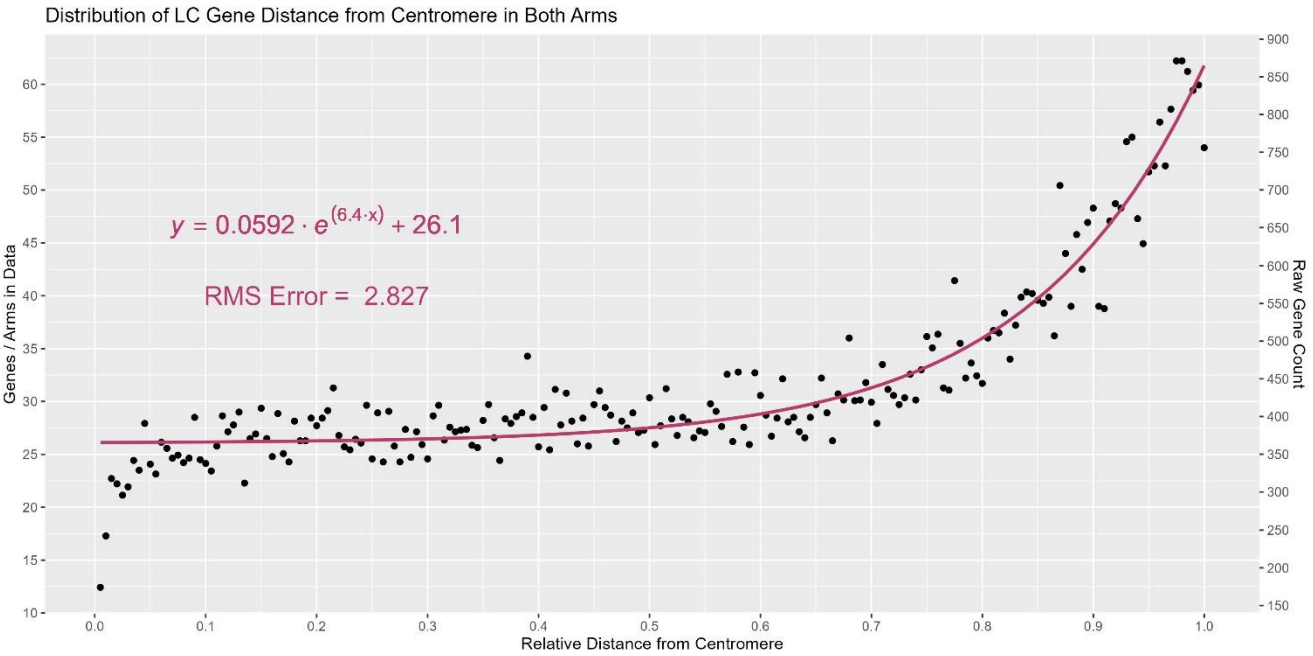

Figure 9.-- Relative LC gene positions of all chromosomes and arms.

The relative position of chromosome bins along the arm, demarcated into 0.5% arm length bins with the number of LC genes found in the respective bin. On the horizontal axis, 0.0 represents the centromere, with 1.0 representing the telomere. On the left y axis, the number of observed LC genes divided by the number of arms in the data (2x7 = 14). On the right y axis, the raw LC gene count within the bin.

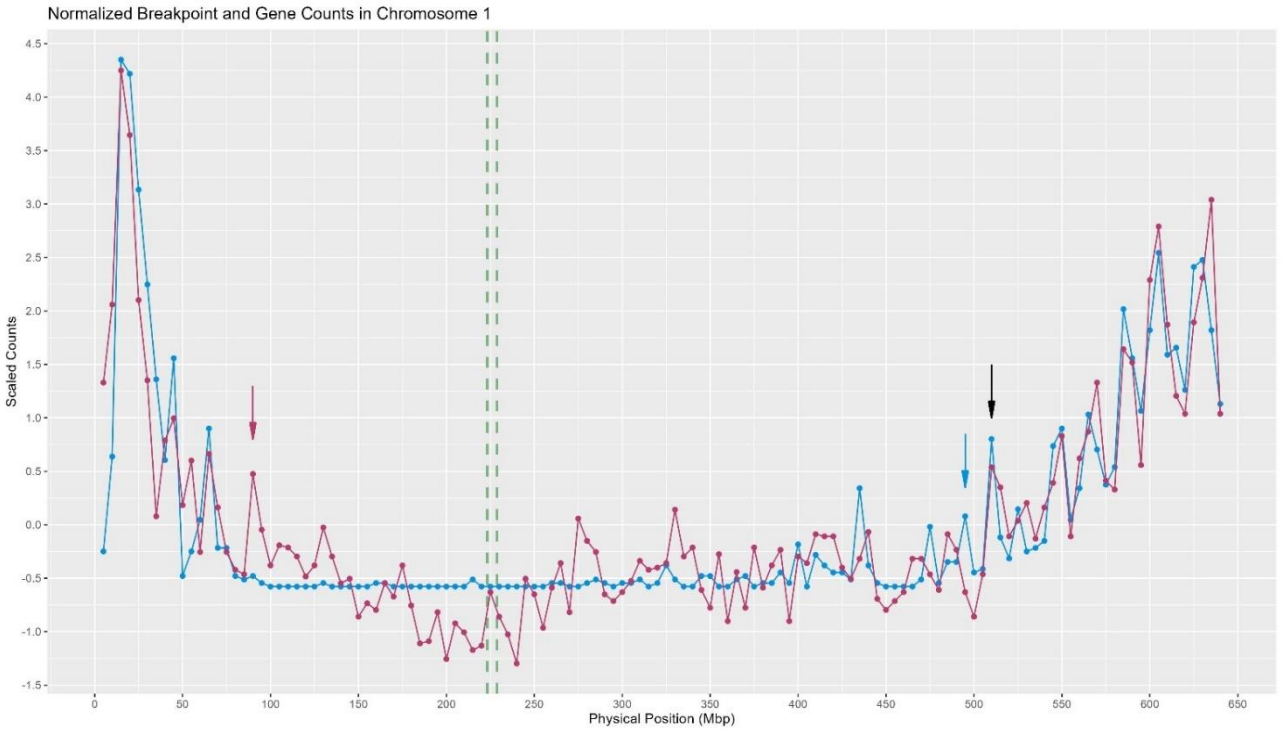

Figure 10.-- Standard scaled distribution of crossovers and gene counts in chromosome 1A.

Crossovers are represented as blue data and gene counts as red data. The vertical green lines delimit the centromeric region. The red arrow indicates a spike in gene density with no corresponding spike in recombination density, the blue arrow indicates a recombination hotspot with no corresponding spike in gene density, and the black arrow indicates one of many examples of a region where the density spikes occur in the same region of the chromosome. The counts for both crossovers and genes were normalized and scaled so that the mean and standard deviation of each function is 1 and binned in 5 Mb intervals to better visualize the proximities between wide-scale spikes in gene and crossover density.

Recombination + Gene Proximity Associations

We observed that many spikes in gene density were closely aligned with regions of high crossover rates compared to their immediate surroundings and vice versa in non-pericentromeric regions (Figure 10, Figure S8- S12). These recombination 'hotspots' can have at least 5x the crossover density as compared to their immediate neighborhood. Their width can range from only a few kilobases to a larger region of 1.5 megabases. A distinct positive correlation was found between both HC and LC gene density and crossover density within every chromosome (Figure 11A, 11B). The trend was found to be stronger for HC gene positions than LC gene positions.

To investigate in higher resolution the physical proximities between crossovers and genes, each breakpoint interval (determined by the positions of the left and right markers where the region transition occurred) was inspected to find how many intervals contained one or more genes within the interval (Table 2). To test if the overlap between crossover intervals and genes was not random, we utilized a permutation approach with randomly sampled recombination intervals and genes were utilized as described previously. A 95% confidence interval was used in comparing the observed data's breakpoint associations against both randomly sampled datasets' breakpoint associations (the red highlighted lines of Table 2) for both HC (Figure 12A) and LC (Figure 12B) genes, confirming that recombination sites tend to form in higher

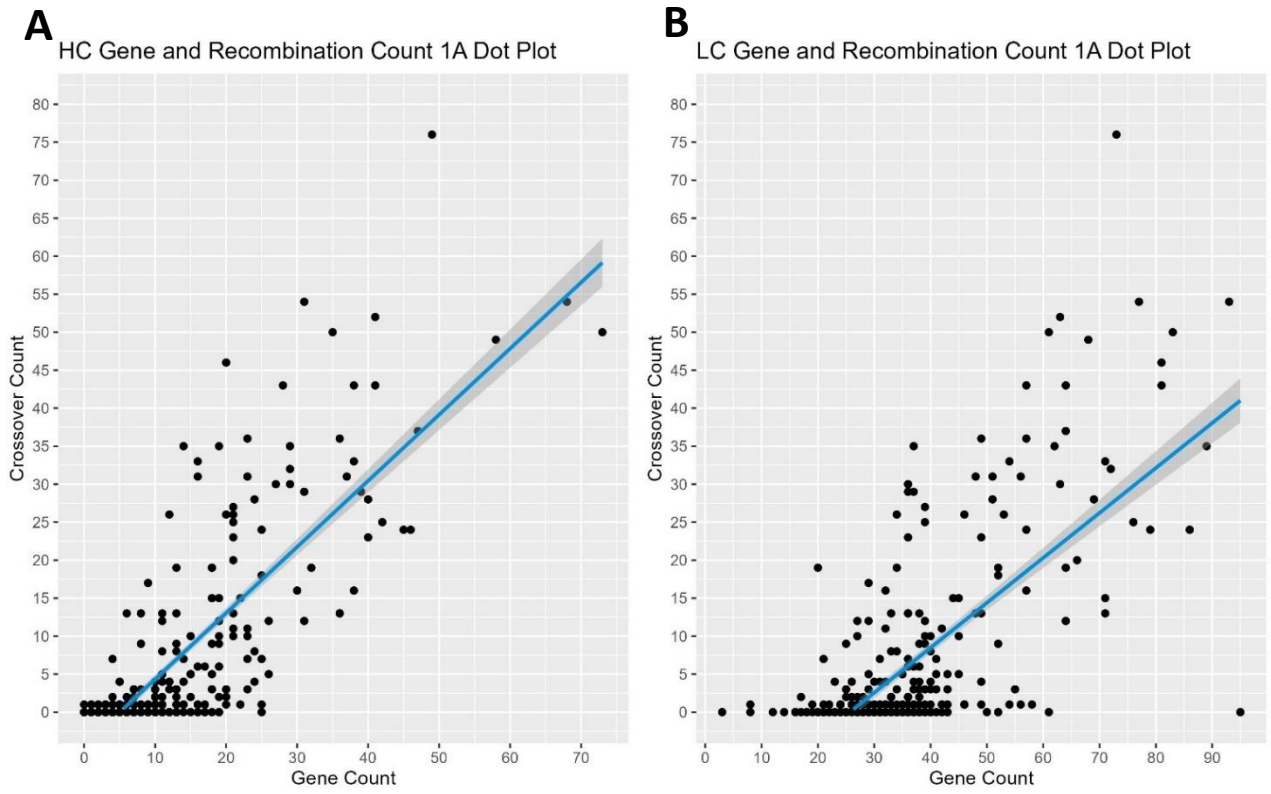

Figure 11.--HC/LC gene count VS crossover count dot plot for chromosome 1A.

2 Mb bin-demarcated dot plot of HC gene count vs crossover count (A) and LC gene count vs crossover count (B) within in the same bin for all lines across chromosome 1A. In blue is a linear regression with a 95% confidence interval. A spearman correlation coefficient of 0.706 and 0.533 was found for HC and LC genes vs crossovers, respectively.

frequencies in areas surrounding genes than gene-poor regions. Additionally, while comparing the associations of the observed HC gene breakpoint associations against the same breakpoints with normalized LC gene counts (the yellow highlighted lines of Table 2) (Figure 12C), we find that the percentage of the number of RBs associated with at least one gene amongst the observed HC genes is well outside of the 95% confidence interval of the normalized LC percentages. This suggests that breakpoints tend to form at even higher frequencies around actively transcribing genes (or at least genes that likely have the capacity to transcribe), rather than pseudo genes.

Breakpoint Condition	Gene Condition	% RBs Associated With At Least One Gene	Median Genes Associated	Mean Genes Associated
Obs. RBs	Obs. HC+LC	87.6%	5	10.037
Obs. RBs	Obs. HC	72.0%	2	3.734
Obs. RBs	Obs. LC	80.2%	3	6.302
Obs. RBs	Normalized LC	62.1%	1	2.286
Obs. RBs	Rand. HC+LC	77.5%	3	5.290
Obs. RBs	Rand. HC	52.4%	1	1.414
Obs. RBs	Rand. LC	72.3%	2	3.876
Rand. RBs	Obs. HC+LC	72.7%	2	5.283
Rand. RBs	Obs. HC	49.3%	0	1.409
Rand. RBs	Obs. LC	68.7%	2	3.875

Table 2.-- Association of crossover intervals and genes.

Randomized sets of genes and crossover intervals were permuted and calculated the average percentage of crossover intervals with a gene from the 100 permutations. Each double line demarcates a different set of permuted data, first taken from observed positions and overlap of crossover site and HC genes, normalized LC gene counts to match HC gene counts so they may be compared, randomly sampled gene positions, and lastly randomly sampled recombination interval positions. The corresponding colors match observed (Obs.) intersection of crossover and gene positions with permuted datasets of these intervals. The white line corresponds to the LC sampled dataset in order to normalize its counts so it may be compared to the HC genes.

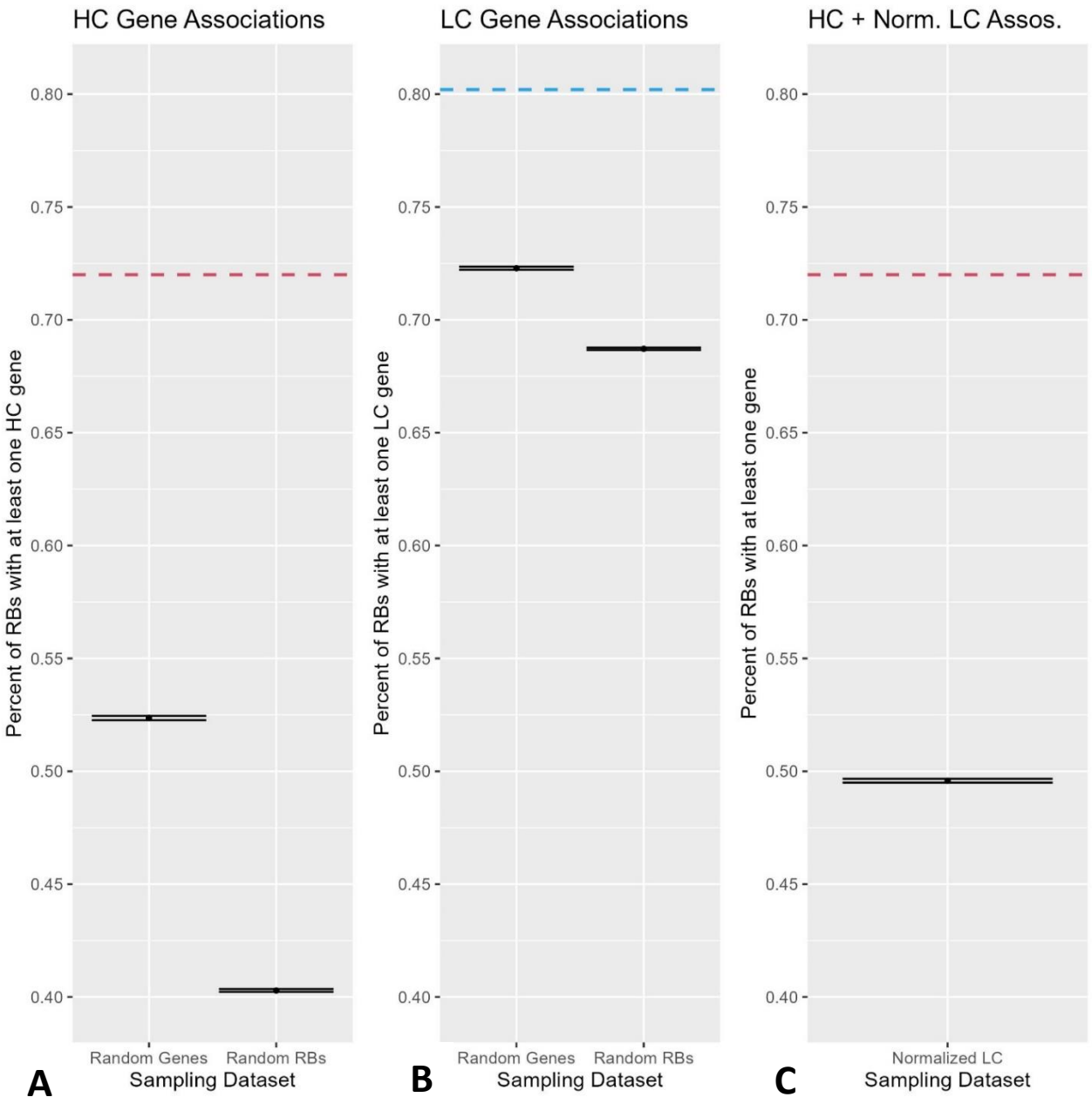

Figure 12.--HC vs LC gene associations compared to random sampling.

The dashed horizontal lines represent the observed percent of RB intervals with at least one HC (red) or LC (blue) gene within it. Sampled permutation data is drawn from 100 iterations with 95% confidence intervals drawn from the sampled distribution of 100 permutations. The associations of RBs with at least one HC (A) and LC (B) gene, compared against the randomly sampled gene dataset and the randomly sampled RB interval dataset. (C) the percentages of associated RBs between the sampled normalized LC dataset are also compared against the actual HC dataset.

Recombination Redistribution in the 1A and 5A Short Arm

Within the final 5% of the arm, crossover density across all chromosomes dropped significantly, a trend that was not found nearly as significantly in the gene densities. However, a significantly greater drop was found in the recombination of the short arm peritelomeric region in the 1A and 5A chromosomes (Figure 13). Although a drop in the recombination rate within several Mb of the telomere can be explained by the technical limitations of the recombination detection algorithm, there is a distinct low crossover density region some 12 Mb proximal from the 1A short arm telomere, and 7 Mb proximal from the 5A short arm telomere. This observed absence of crossover in the short-arm telomeric region is consistent with the position of nucleolus organizing regions (NORs), constitutive heterochromatic regions with highly repetitive rRNA genes that cause structural polymorphism between haplotypes.

Nor9 and Nor10, known NORs of *T. monococcum* located at the short arm telomeres of 1A and 5A respectively, are the most likely factors of this reduced rate of crossovers. The einkorn-derived NORs contain an order of magnitude more rRNA gene units than the corresponding regions in Chinese spring. In a study by Luo et al.1998, a notable drop in recombination near the einkorn-derived Nor9 locus was observed. The same effect was suspected for Nor10 on Chr. 5A, although there was insufficient marker density on chromosome 5A to determine the effects of the Nor10 locus. These NORs do not simply limit recombination but redistribute the homologous meiotic exchanges away from the NOR locus. Nor9's range of influence was previously reported at around 17 cM proximal from the Nor9 locus. This should mean the short arm of chromosome 1A and potentially 5A should have a higher-than-average RB density in the

intra-arm region up to roughly 17 cM from the telomere, with a drop to lower-than-average RB density at the distal region. Indeed, a clear region of low recombination density is present in the distal 5% of 1A and distal 2.5% of 5A. In the low-density region, 1A possessed only 25.4% the RB density as the following 10% of the arm, while the most distal 2.5% of 5A possessed only 59.5% the RB density as the following 10% of its respective arm. Comparing the averages of the non-NOR carrying chromosomes, the distal 5% of 1A had only 19.0% the crossover density, while the distal 2.5% of 5A had only 51.8% the crossover density. However, this transition caused by cessation of Nor9's influence occurs much closer to the telomere than previously estimated, at ~10.17 Mb proximal from the telomere (roughly 6 cM). Regardless, the distinct redistribution of RBs on the telomeric short arm remains unique to the 1A and 5A, enough so that we may attribute the dramatic change in crossover density to NORs.

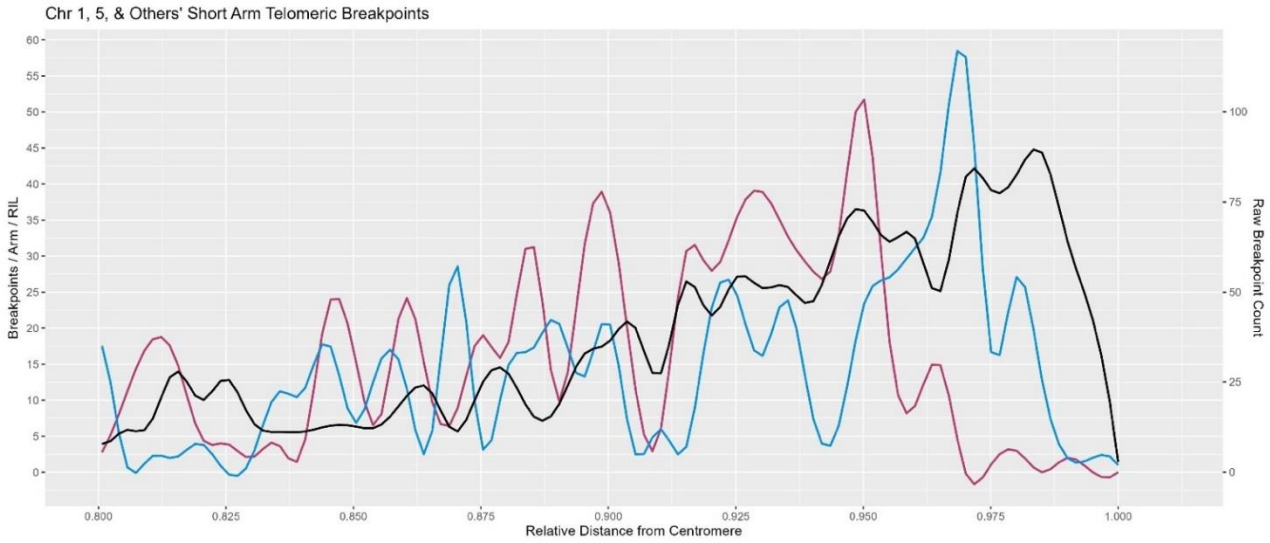

Figure 13.-- Crossover rate of chromosome arms with and without NORs.

Plotted on a centromere-to-telomere scale in relative 0.05-unit bins on the short arms of chromosome 1A (red) under Nor9's influence, and chromosome 5A (blue) under Nor10's influence. The average of chromosomes without major influences of any known NORs (2A, 3A, 4A, 6A, and 7A) are plotted in black.

Chapter 4 - Discussion

Utilizing a novel approach to crossover detection possessing multiple checks to ensure that the accuracy of smaller regions is properly identified as well as ensuring the representation of larger trends in region homology. The formation of crossover events is affected by a multitude of factors, including where the prior double-stranded break forms in the context of large- and small- scale genetic and epigenetic features. These double-stranded breaks may resolve to a crossover or a non-crossover. These non-crossover events exchange little genetic information, usually less than 500bp, and are much more common than crossovers in budding yeast (Vincenten et al. 2015). In *Arabidopsis*, nucleosome organization, such as H2A.Z histone subunits, histone trimethylation (H3K4me3) as well as DNA regions that were generally unmethylated and especially A-T rich were found to be associated with recombination hotspots (Choi et al 2013). H2A.Z and DMC1 are also both regulated by the SWR1 complex which regulates expression in Arabidopsis meiosis, the former of which was found to be related to increased recombination due to cold stress (Wang and Copenhaver 2018). In yeast, active promoter sites can create localized preferences for Spo11 due to enrichment of H3K4me3, relevant due to the widespread conservation of these functions. Chromatin super-organization can also greatly impact Spo11 binding and hence double-stranded breaks. This superstructure includes 10-110 kb chromatin loops consisting of exposed hills and cohesin-rich valleys, where Spo11 oligos struggle cleaving DNA in local proximity to the Rec8 cohesin subunit. The densely packed heterochromatin and repetitive regions within NORs as well as pericentromeric kinetochores can strongly suppress crossover formation. Genes that do transcribe within facultative heterochromatin have been found to be additionally post-transcriptionally regulated

and silenced by RNAi, which in yeast was seen to maintain proper centromere function accompanied by histone H3 lysine-9 methylation (Volpe et al. 2002). DNA methylation, including CpG, CHG, and CHH residues (H representing adenine, cytosine, or thymine) has a strong, negative association upon recombination (Wang and Copenhaver 2018), but were found to be evenly distributed throughout all hexaploid wheat sub-genomes (a trend also shared in rice). (Gardiner et al. 2015). Thus, it is likely that DNA methylation plays a much stronger role in regulating local crossover placement in specific domains than affecting global crossover trends. Additionally, many biotic and abiotic factors can also impact crossover rates. For example, male meiosis was found to have less recombination than female meiosis, and in terms of abiotic factors, higher temperatures have been shown to shift the distribution of recombination events towards the centromeres (Coulton et al. 2020). Altogether, a myriad of factors contributes to the resulting meiotic crossover points on a chromosome.

Recombination and Gene Distribution Throughout the Chromosome Arms

We find a strong increase in recombination breakpoint density (Figure 6) and HC and LC gene density (Figure 8, Figure 9) with increasing physical distance from the centromere. Both the recombination (Figure 2, Supplementary Figure 1-6) and gene distribution overall was a uniform U-shape, the latter maintaining a lower average of 20 genes/Mb around the centromere while reaching over double that density in the sub-telomeric regions (Figure 10, Supplementary Figure 7-12). The most notable exception was chromosome 5, where the long arm had a slightly more even spread of recombinations and genes than any of the other arms, though the upwards trend remained (Supplemental Figure 4, Supplemental Figure 10). Across all

chromosomes, the recombination distribution is particularly dramatic following a steep U-shape with virtually no crossovers within the pericentric regions, including the proximal 40% of the chromosome arms. We observed 91% of all crossovers occurred in the distal 40% of the chromosome arms, consistent with crossover studies in chromosome 3B of *Triticum aestivum* (Saintenac et al. 2009). The recombination density trend is amplified as compared to the following species, listed in decreasing homogeneity of crossovers throughout the arm: rice, maize, (Anderson et al. 2006), hexaploid wheat, and even *Aegilops speltoides* (Luo et al. 2005), which was noted as itself having particularly extreme localizations of crossovers within the sub telomeric regions.

Although the details involved in eukaryotic pericentromeric crossover repression remains somewhat enigmatic, multiple major players are known to shape the chromosome-wide recombination landscape. The cohesin protein complex, and by extension the highly conserved centromeric Ctf19 kinetochore complex plays a major role in shaping a cell's meiotic crossover events to prevent aneuploidy commonly associated with pericentromeric crossovers (Vincenten et al. 2015). A secondary role of the kinetochore outside of its critical role in anaphase lies in inhibiting double-stranded break formation within ~6kb of the centromere, as well as suppressing homolog interactions in longer ranges (60+ kb) via the centromeric recruitment of the synaptonemal complex component Zip1 led by Ctf19 and the Scc2/4 cohesin loader in early prophase. This spread of cohesion past the centromere suppresses meiotic crossover events within a significant radius of the centromere, lessening the danger of improper separation during anaphase. However, the wide-spanning pericentromeric suppression observed in the

proximal 40% of a given arm is likely due to the distribution of heterochromatin which densely inhabits the pericentromeric regions in the *T. aestivum* A-genome (WGSC 2018), and secondarily repetitive transposable element (TE) distribution. HC gene-rich euchromatin in contrast is found mostly localized in the distal subtelomeric regions. However, it is worth noting that euchromatin is not necessarily associated with higher gene density and that chromosome accessibility is mostly retained in a translocated region (Jordan et al. 2020). The centromere to telomere-wide trend did develop from multiple contributing factors within the wheat genomes, most notably chromatin structure and gene placement. However, local recombination is largely due to the local genetic landscape, although significant changes can occur over long periods of time. It was found in hexaploid wheat that although the distal regions of 4A had significantly less accessible chromatin structure than the same region on 4B, they had very similar gene densities (Jordan et al. 2020). Importantly, a significant increase in Gypsy transposable elements (TEs) and decrease in CACTA TEs was also found in the same 4A region as compared to 4B. In fact, homoeologous group 4 displayed the least accessible chromatin compared to the other chromosome groups, which could be an explanation for the lack of observed RBs in 4A. There was also a strong TE density gradient observed across the chromosomes, with 73-89% less TE content in the distal ends of chromosomes compared to centromeric regions. These highly repetitive TE elements, which makes up 67% of the wheat genome, have been suggested to affect nearby gene expression through epigenetic regulatory mechanisms, as well as distant cis-regulatory elements that regulate genes via shaping of the 3D chromatin structure. These repetitive regions are also known to carry significantly reduced numbers of successful crossovers due to the reduced chromatin accessibility they are associated with. Chromatin

density, along with the prolific tandem repeats in the pericentromeric region of all A-genome chromosomes orchestrate a suppression of recombination up to 200 Mb from the centromere.

The weak positive trend between chromosome length and number of crossovers supports the well-described mechanism of crossover interference, a mechanism that limits the number and distribution of crossovers throughout a chromosome (Jones and Franklin 2006). The stress-based mechanism as proposed by Wang et al. 2015, called the beam-film model, proposes that adjacent unresolved Holliday junctions undergo mechanical stress from the compacted chromatin fiber during meiosis, until one commits to becoming a CO and outwardly relieves the tension in that region, disfavoring other crossovers in its surroundings. Altogether, multiple overlapping recombination-altering processes help ensure consistent and successful meiotic exchange and chromosome segregation.

Gene-Recombination Proximity Associations

We found that crossovers localize strongly with genes in both across the length of the chromosome (Figure 11, Supplemental Figure 13-18) as well as at the site of crossovers (Table 2). This data aligns with previous observations that crossover events are associated with genes, at least outside of the centromeric region (Anderson et al. 2006). Across all chromosome arms, there were significant sharp spikes in crossover frequency found outside of the pericentromeric region that reach crossover rates >5x the surrounding regions. These recombination hotspots usually aligned with spikes in gene density, especially in more distal locations on an arm. Although there are spikes in gene concentration in the pericentromeric regions, centromeric

suppression limits the maturation of DSB into crossover events, instead usually resolving into non-crossovers, which significantly limits the average crossover density of the region. However, weaker recombination hotspots do occasionally occur in the proximal 40% region of an arm, usually aligned with at least a partial increase in gene density. This aligns with previous thought that crossover events are associated with gene positions, although there are other macro and micro factors that can significantly alter recombination frequencies (Anderson et al. 2006). This breakpoint to gene proximity is most likely caused by the highly conserved Spo11 protein and several other factors that are responsible for catalyzing double-stranded breaks, and the fact that an active gene is most likely found in the more accessible euchromatin. They are associated with chromatin regions whose nucleosomes contain the trimethylation of histone H3 Lysine 4 (H3K4) (reviewed in de Massy 2013), which in budding yeast are spatially correlated near active promoters (Vincenten et al. 2015). These factors incentivize crossover formation in the surrounding intergenic regions, especially in genes that carry an active promoter.

Lack of Recombination in Short Arms of 1A and 5A likely due to Nucleolus Organizing Regions

The lack of recombination in the distal regions of the 1A and 5A short arms (Figure 13) was likely due to previously described recombination-suppressing Nor9 and Nor10 respectively. NORs have been described as containing three distinct regions: a proximal region, tandem repeats of rDNA genes, and a distal region (Floutsakou et al. 2013). Although the flanking regions are highly repetitive facultative heterochromatin, the proximal region sees frequent recombination, and the distal junction transcriptionally active. This distal junction is used to anchor homologous chromatids in close proximity within the nucleolus to better facilitate

homolog synapsis. Although the area of effect was shorter than the expected 17 cM (or roughly 22 Mb), (Luo et al.1998), Nor9 caused the telomeric 10 Mb of the 1A short arm to be roughly ~8% the recombination density as the following spike in rates at ~15 Mb on the same arm. Nor10 caused the telomeric 5 Mb of the 5A short arm to be roughly ~20% the RB density as the following spike at 8 Mb. These NORs caused a definitive drop in recombination as compared to the short-arm telomere-proximal regions of the other chromosomes. Due to the reduced recombination area from Nor10 being almost half that of Nor9, Nor9 could potentially be located more proximally in the arm than Nor10, could contain more gene units and therefore a stronger effect, or both. Although the gene unit counts are lower than NORs recorded in other cereal species, they leave a significant mark on the Einkorn recombination landscape within their territory, highlighting the importance of understanding the effects of these regions in other species.

Significant Pericentromeric Suppression and Lack of Overall Recombination in 4A

Chromosome 4A interestingly was an outlier on multiple fronts (Table 1)— importantly, it had by far the lowest number of detected crossovers (34% lower) and crossover per Mb (20% lower). It also had the highest variants per Mb at 15% greater than the average, and the lowest gene density (14% lower than the average). Additionally, its recombination landscape lacked any significant amount of crossover in the proximal third of the arms (Figure 7), and had the largest proportion of its crossovers in the distal 80% of the genome (88.9%). The lack of recombinations in 4A is certainly due to multiple overlapping factors. Although length potentially is a contributing factor, it is the 2nd shortest chromosome, and 6A, the shortest, has

only slightly below average crossovers considering its length. One potential cause is the lack of genes within 4A, which might limit the effect of promoter-based crossover enrichment. It has also been shown that a lack of homology between sister chromatids in a meiotic event between einkorn and the A genome chromosomes of *T. aestivum* can decrease recombination across the chromosomes (Luo et al. 2000), and due to the high rate of variation, this could certainly lessen the recombination in the chromosome. It is also worth noting that the 4A centromeric region underwent significant rearrangement during Einkorn's domestication, with a neocentromere now fully developed some 6.5 Mb upstream of the ancient centromere. It is unknown how this affects the function of the new centromere, or if any aspect of this rearrangement contributes to the enhanced pericentromeric suppressive effect. Homoeologous group 4 in *T. aestivum* as a whole was detected to have less euchromatin content compared to other chromosomes, with the chromatin of the short arm of 4A subtelomeric region especially compact and inaccessible (Jordan et al. 2020), which is known to reduce successful crossover events. The lack of recombination in 4A is due to multiple factors, some of them poorly understood, although it is believed that its small size, lack of sequence homology and gene density as well as the overall chromatin structure play major roles.

Chapter 5 - Conclusions

The strengths of skim-sequencing a large RIL population, along with a novel approach to breakpoint detection in low-coverage skim-seq data, enabled us to obtain a detailed picture of the unique recombination landscape of Einkorn and the myriad of factors that affect it from a large to small scope. This method was able to properly and accurately identify recombinations on a genome-wide scale while utilizing ultra-low coverage skim-seq lines. *Triticeae* species especially are known to have a strongly skewed recombination density towards the distal regions of its chromosome, which is affirmed by our findings. Distance from the centromere plays a large role in the general trends of crossover density within a centromere, a trend caused in large part by wide-scope chromatin structure as well as minorly the suppressive effects of the Ctf19 kinetochore complex and the cohesin that propagates outwards from the centromere during meiosis. The crossover density at its extremes varies by more than 100-fold and closely follow an exponential function, due in large part to the chromatin structure of *Triticeae* chromosomes, in which accessible euchromatin preferentially locates in distal regions of the chromosomes while the tightly packed heterochromatin is found in high frequency in the wider pericentromeric region, much more so than in maize or rice. This is especially apparent in chromosome 4A, where the reduced crossovers observed are at least in part due to the higher concentration of heterochromatin it possesses. These heterochromatic regions also contain a significant number of repetitive TEs, which also are known to correlate with suppressed recombination frequency. We confirm the recombination-suppressing presence of Nor9 and Nor10 in the short arms of 1A and 5A previously described in Einkorn. Genes are another major player in the resolution of DSB into full-fledged crossover events, especially genes that are likely

transcribing. The highly conserved DSB initiator Spo11 has been shown to be locally associated with H3K4 methylation in yeast, which itself is found near active promoters. The implication to this is confirmed in this study-- significantly more crossovers were found to form around HC genes than LC genes. These gene distributions and their activity are also affected by chromatin structure distribution, but both uniquely affect crossover formation tendencies in their own right. The complete picture of how recombination rates are defined within a chromosome as a whole remains to be further studied, but knowledge of these important factors, and their scope and synergies within the chromosome, is necessary for demystifying this crucial mechanism to better understand the evolution and potential of the Einkorn genome, and by extension, the A sub-genome as a whole.

References

Adhikari, Laxman et al. 2022. “Genetic Characterization and Curation of Diploid A-Genome Wheat Species.”

Adhikari, Laxman, and Amos, Cameron. 2022. “Skim-Seq-Population-Genotyping-for-Linkage-Mapping.” https://github.com/laxmangene7/Skim-Seq-Population-Genotyping (October 18, 2022).

Ahmed et al. 2022 "Einkorn genomics sheds light on the evolutionary history of the oldest domesticated wheat" (reference number: 2022-10-16176), recently submitted to Nature.

Akhunov, Eduard D. et al. 2003. “The Organization and Rate of Evolution of Wheat Genomes Are Correlated With Recombination Rates Along Chromosome Arms.” *Genome Research* 13(5): 753–63.

Amos, Cameron. 2022. “Skim-Seq Recombination Detector”, Github repository, https://github.com/CameronAmos/Skim-Seq-Recombination-Detector

Anderson, L. K., and S. M. Stack. 2005. “Recombination Nodules in Plants.” *Cytogenetic and Genome Research* 109(1–3): 198–204.

Anderson, Lorinda K. et al. 2006. “Uneven Distribution of Expressed Sequence Tag Loci on Maize Pachytene Chromosomes.” *Genome Research* 16(1): 115–22.

Baudat, Frédéric, and Alain Nicolas. 1997. “Clustering of Meiotic Double-Strand Breaks on Yeast Chromosome III.” *Proceedings of the National Academy of Sciences of the United States of America* 94(10): 5213–18.

Benyahya, Fatiha et al. 2020. “SPO11.2 Is Essential for Programmed Double-Strand Break Formation during Meiosis in Bread Wheat (Triticum Aestivum L.).” The Plant Journal 104(1): 30–43.

Brandolini, A., A. Volante, and M. Heun. 2016. “Geographic Differentiation of Domesticated Einkorn Wheat and Possible Neolithic Migration Routes.” *Heredity* 117(3): 135–41.

Brunazzi, Alice et al. 2018. “Molecular Diversity and Landscape Genomics of the Crop Wild Relative Triticum Urartu across the Fertile Crescent.” *The Plant Journal: For Cell and Molecular Biology* 94(4): 670–84.

Charmet, Gilles. 2011. “Wheat Domestication: Lessons for the Future.” *Comptes Rendus Biologies* 334(3): 212–20.

Choi, Kyuha et al. 2013. “Arabidopsis Meiotic Crossover Hot Spots Overlap with H2A.Z Nucleosomes at Gene Promoters.” *Nature Genetics* 45(11): 1327–36.

Coulton, Alexander, Amanda J. Burridge, and Keith J. Edwards. 2020. “Examining the Effects of Temperature on Recombination in Wheat.” *Frontiers in Plant Science* 11. https://www.frontiersin.org/articles/10.3389/fpls.2020.00230 (August 30, 2022).

Danguy des Déserts, Alice, Sophie Bouchet, Pierre Sourdille, and Bertrand Servin. 2021. “Evolution of Recombination Landscapes in Diverging Populations of Bread Wheat.” *Genome Biology and Evolution* 13(8): evab152.

Divashuk, Mikhail G. et al. 2016. “Variation in Copy Number of Ty3/Gypsy Centromeric Retrotransposons in the Genomes of Thinopyrum Intermedium and Its Diploid Progenitors.” *PLOS ONE* 11(4): e0154241.

Floutsakou, Ioanna et al. 2013. “The Shared Genomic Architecture of Human Nucleolar Organizer Regions.” *Genome Research* 23(12): 2003.

Gardiner, Laura-Jayne et al. 2015. “A Genome-Wide Survey of DNA Methylation in Hexaploid Wheat.” *Genome Biology* 16(1): 273.

Huang, Xuehui et al. 2009. “High-Throughput Genotyping by Whole-Genome Resequencing.” *Genome Research* 19(6): 1068–76.

International Wheat Genome Sequencing Consortium (IWGSC). 2018. “Shifting the Limits in Wheat Research and Breeding Using a Fully Annotated Reference Genome.” *Science (New York, N.Y.)* 361(6403): eaar7191.

Jones, Gareth H., and F. Chris H. Franklin. 2006. “Meiotic Crossing-over: Obligation and Interference.” *Cell* 126(2): 246–48.

Jordan, Katherine W. et al. 2020. “Differential Chromatin Accessibility Landscape Reveals Structural and Functional Features of the Allopolyploid Wheat Chromosomes.” *Genome Biology* 21(1): 176.

Luo, M. C., Z. L. Yang, and J. Dvorak. 1998. “Position Effects of Ribosomal RNA Multigene Loci on Meiotic Recombination in Wheat.” *Genetics (Austin)* 149(2): 1105–13.

Luo, Ming-Cheng, Karin R. Deal, Zu-Li Yang, and Jan Dvorak. 2005. “Comparative Genetic Maps Reveal Extreme Crossover Localization in the Aegilops Speltoides Chromosomes.” *Theoretical and Applied Genetics* 111(6): 1098–1106.

Luo, Ming-Cheng, Zu-Li Yang, Rama S Kota, and Jan Dvořák. 2000. “Recombination of Chromosomes 3Am and 5Am of Triticum Monococcum With Homeologous Chromosomes 3A and 5A of Wheat: The Distribution of Recombination Across Chromosomes.” *Genetics* 154(3): 1301–8.

de Massy, Bernard. 2013. “Initiation of Meiotic Recombination: How and Where? Conservation and Specificities Among Eukaryotes.” *Annual Review of Genetics* 47(1): 563–99.

Middleton, Christopher P. et al. 2014. “Sequencing of Chloroplast Genomes from Wheat, Barley, Rye and Their Relatives Provides a Detailed Insight into the Evolution of the Triticeae Tribe.” *PloS One* 9(3): e85761.

Nambiar, Mridula, and Gerald R. Smith. 2016. “Repression of Harmful Meiotic Recombination in Centromeric Regions.” *Seminars in cell & developmental biology* 54: 188–97.

Pae, Hye K. 2020. “The Emergence of Written Language: From Numeracy to Literacy.” In *Script Effects as the Hidden Drive of the Mind, Cognition, and Culture*, Literacy Studies, ed. Hye K. Pae. Cham: Springer International Publishing, 25–36. https://doi.org/10.1007/978-3-030-55152-0_2 (August 18, 2022).

Pan, Jing et al. 2011. “A Hierarchical Combination of Factors Shapes the Genome-Wide Topography of Yeast Meiotic Recombination Initiation.” *Cell* 144(5): 719–31.

Saintenac, Cyrille et al. 2009. “Detailed Recombination Studies along Chromosome 3B Provide New Insights on Crossover Distribution in Wheat (Triticum Aestivum L.).” Genetics 181(2): 393–403.

Saintenac, Cyrille et al. 2013. “Identification of Wheat Gene Sr35 That Confers Resistance to Ug99 Stem Rust Race Group.” Science (New York, N.Y.) 341(6147): 783–86.

Singh, Kuldeep et al. 2007. “An Integrated Molecular Linkage Map of Diploid Wheat Based on a Triticum Boeoticum × T. Monococcum RIL Population.” Theoretical and Applied Genetics 115(3): 301–12.

Vincenten, Nadine et al. 2015. “The Kinetochore Prevents Centromere-Proximal Crossover Recombination during Meiosis.” eLife 4: e10850.

Volpe, Thomas A. et al. 2002. “Regulation of Heterochromatic Silencing and Histone H3 Lysine-9 Methylation by RNAi.” *Science (New York, N.Y.)* 297(5588): 1833–37.

Wang, Shunxin, Denise Zickler, Nancy Kleckner, and Liangran Zhang. 2015. “Meiotic Crossover Patterns: Obligatory Crossover, Interference and Homeostasis in a Single Process.” *Cell Cycle (Georgetown, Tex.)* 14(3): 305–14.

Wang, Yingxiang, and Gregory P. Copenhaver. 2018. “Meiotic Recombination: Mixing It Up in Plants.” *Annual Review of Plant Biology* 69: 577–609.

Zaharieva, Maria, and Philippe Monneveux. 2014. “Cultivated Einkorn Wheat (Triticum Monococcum L. Subsp. Monococcum): The Long Life of a Founder Crop of Agriculture.” *Genetic Resources and Crop Evolution* 61(3): 677–706.

Chapter 6 - Supplemental Figures

The following set of figures are paired with Figure 2, presenting the recombination landscape in chromosomes 2-7 plotted in 2 Mb bins in addition to 3 moving averages of varying window widths. The green vertical lines delimit the centromeric region. (Figure S1- S6).

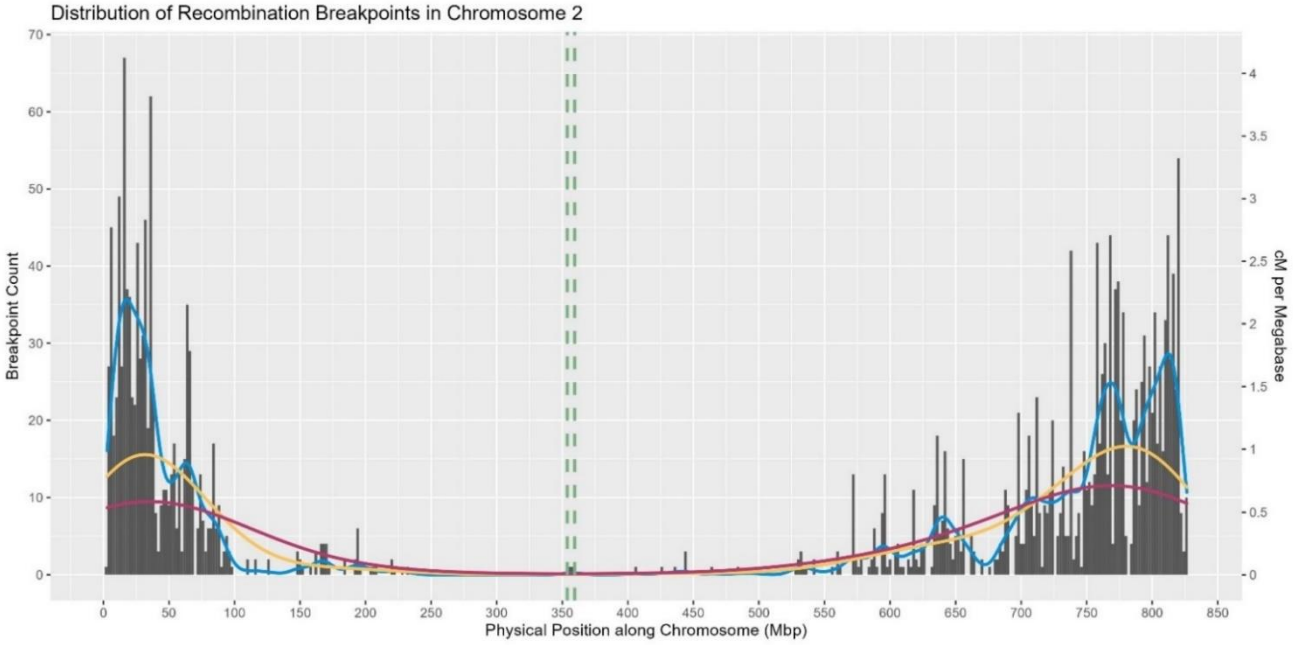

Figure S1.—Recombination landscape of chromosome 2.

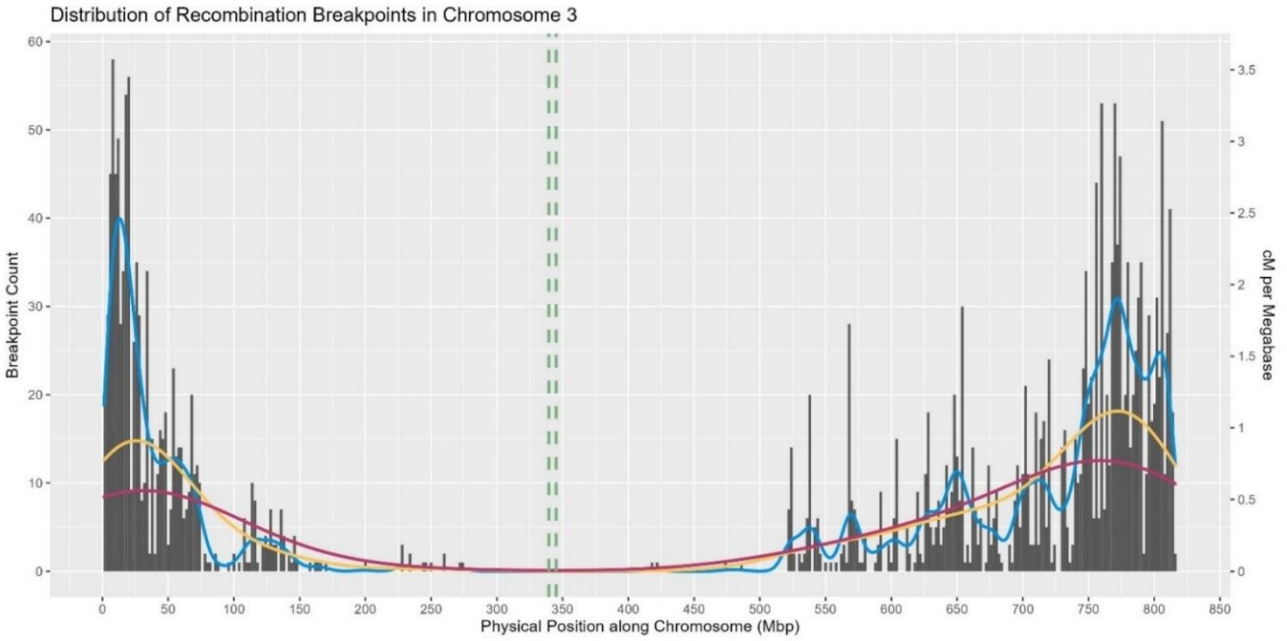

Figure S2.—Recombination landscape of chromosome 3.

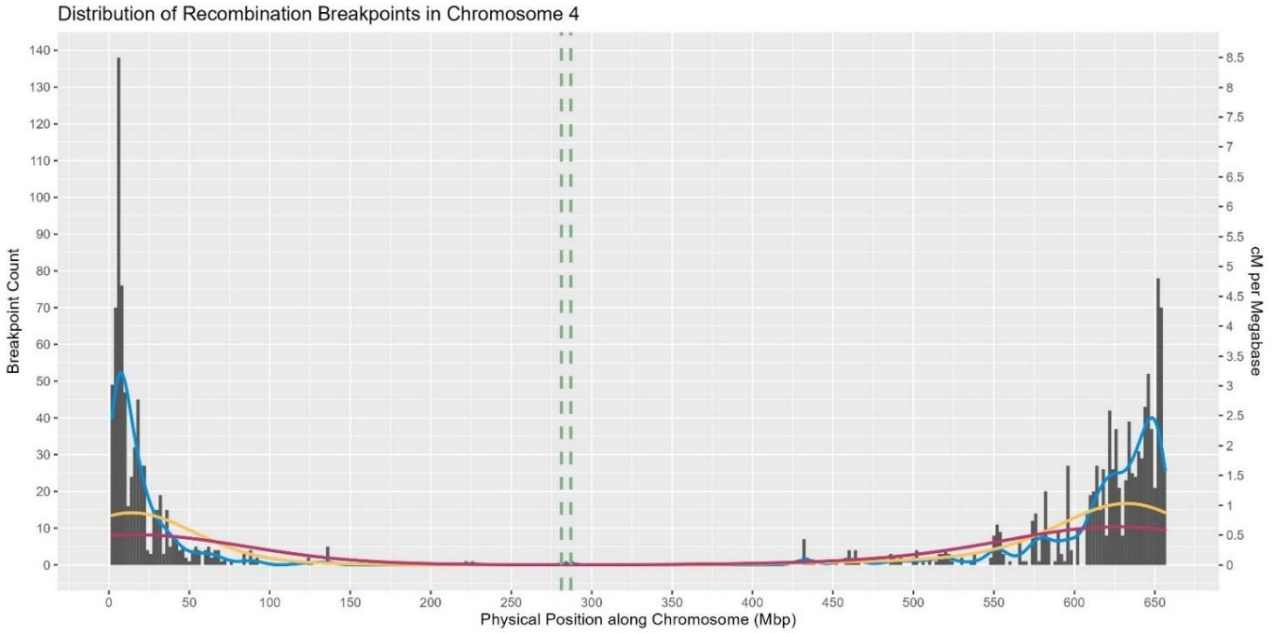

Figure S3.—Recombination landscape of chromosome 4.

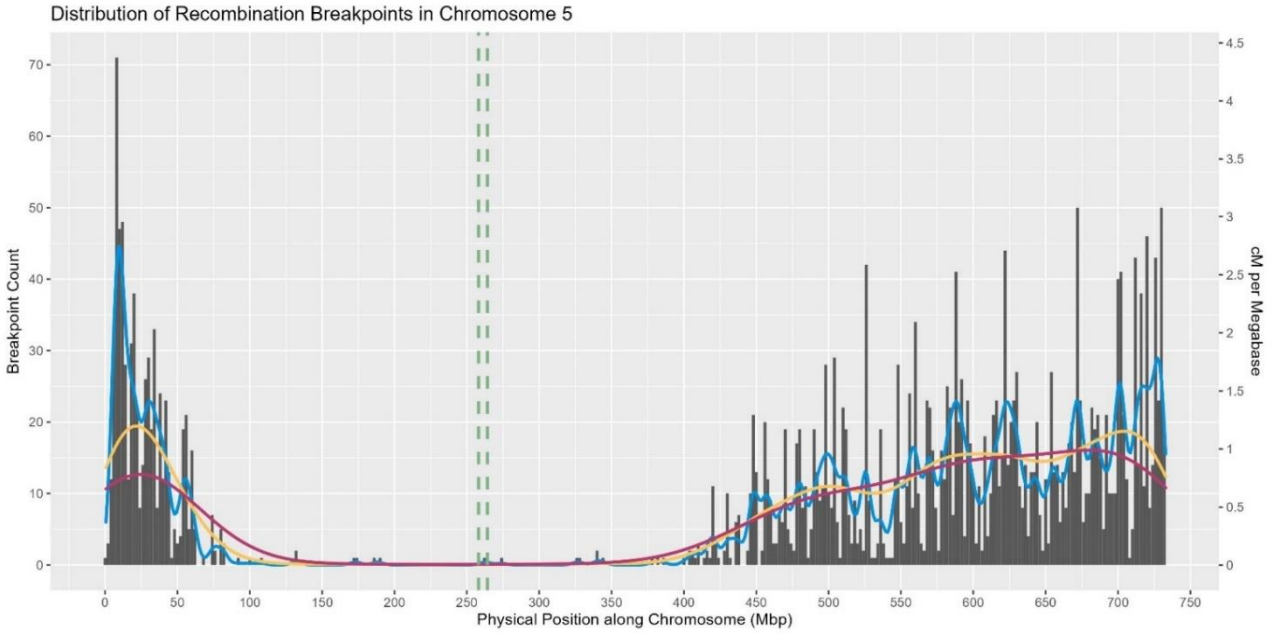

Figure S4.—Recombination landscape of chromosome 5.

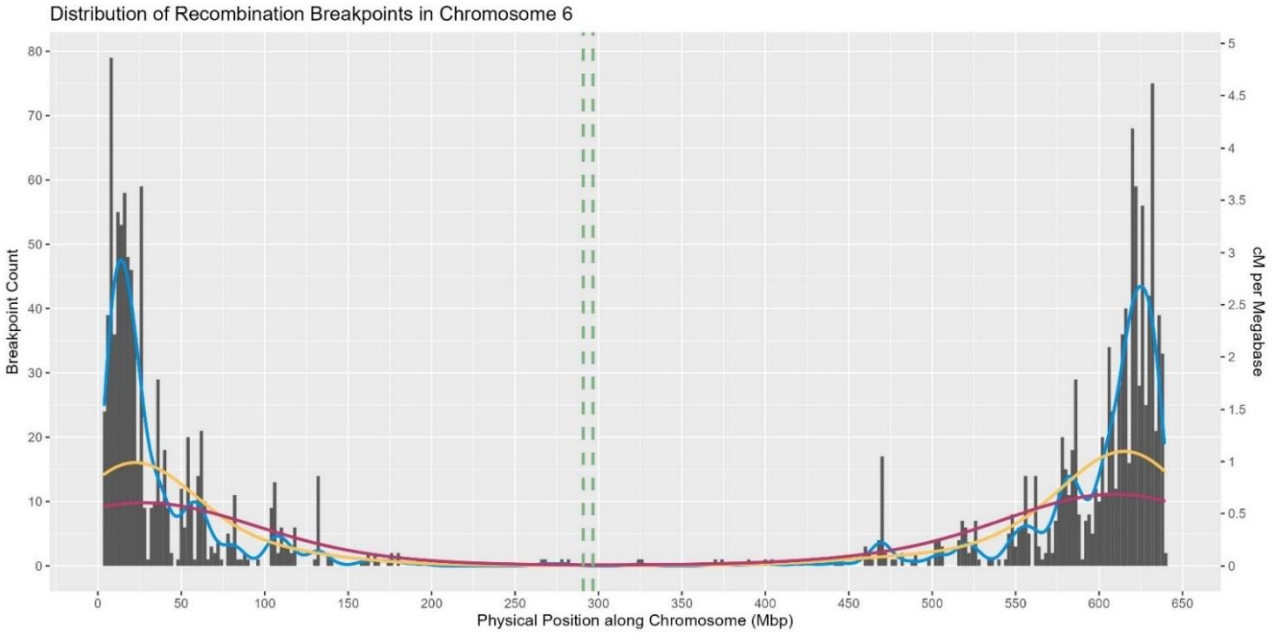

Figure S5.—Recombination landscape of chromosome 6.

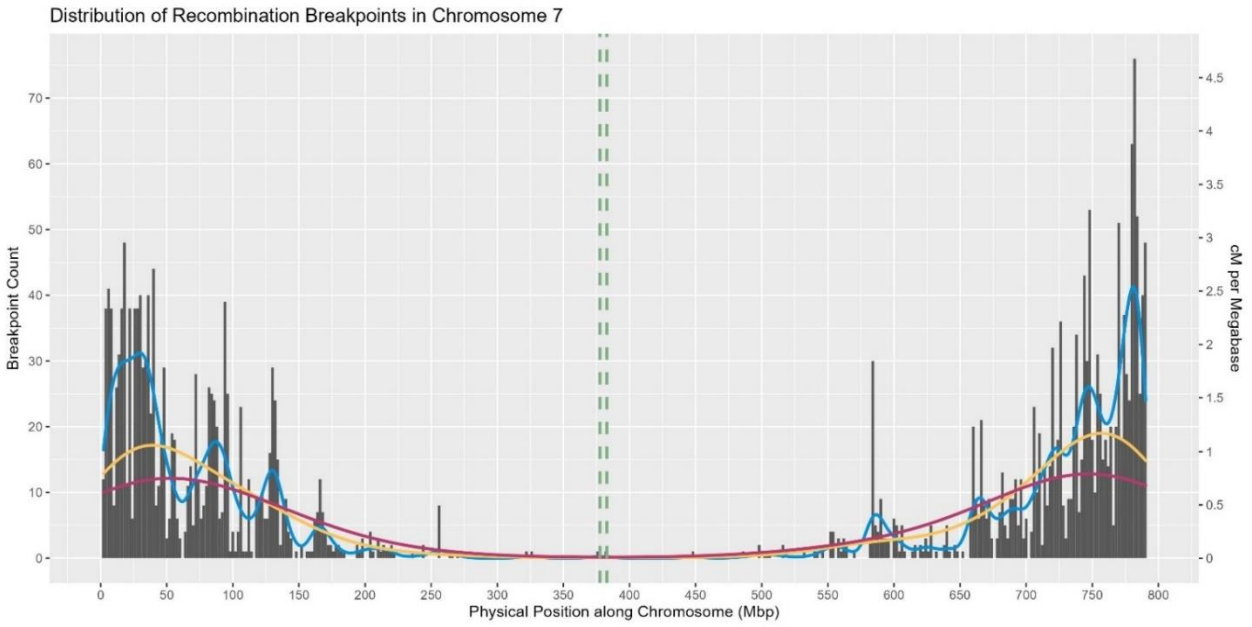

Figure S6.—Recombination landscape of chromosome 7.

The following set of figures are paired with Figure 10, presenting the standard scaled distributions of breakpoint (blue) and gene (red) counts in chromosomes 2-7 (Figure S7- S12).

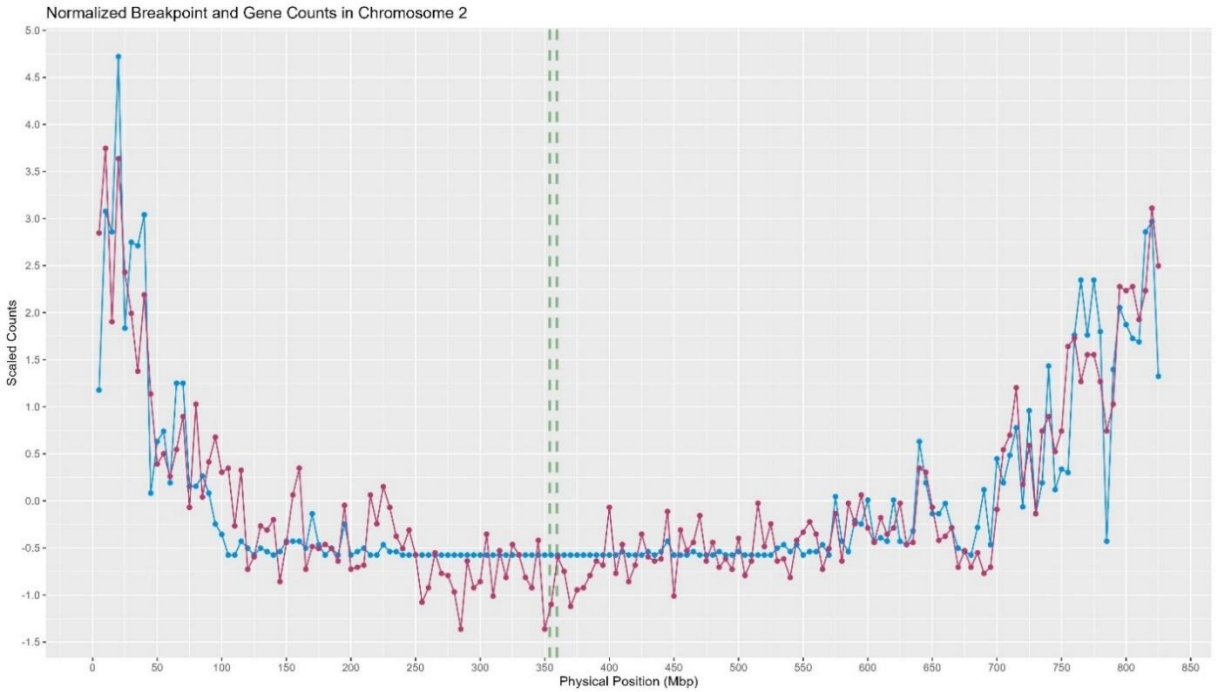

Figure S7.-- Standard scaled dist. of crossovers and gene counts in chromosome 2A.

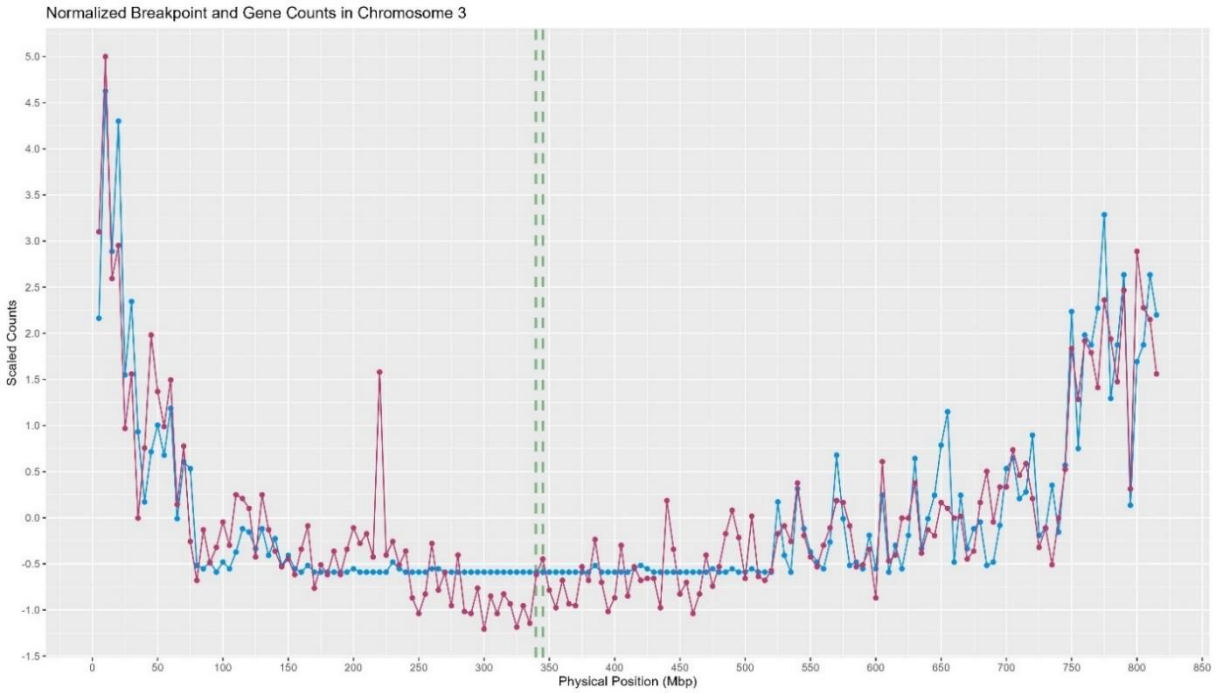

Figure S8.-- Standard scaled dist. of crossovers and gene counts in chromosome 3A.

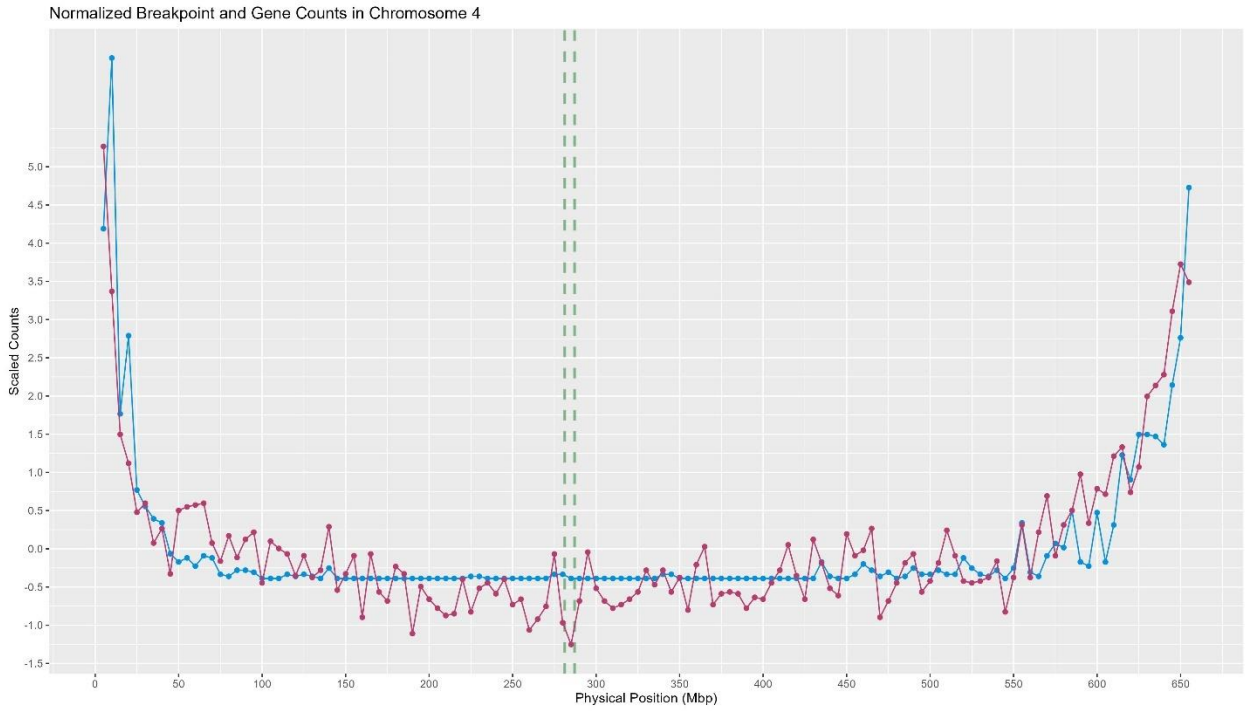

Figure S9.-- Standard scaled dist. of crossovers and gene counts in chromosome 4A.

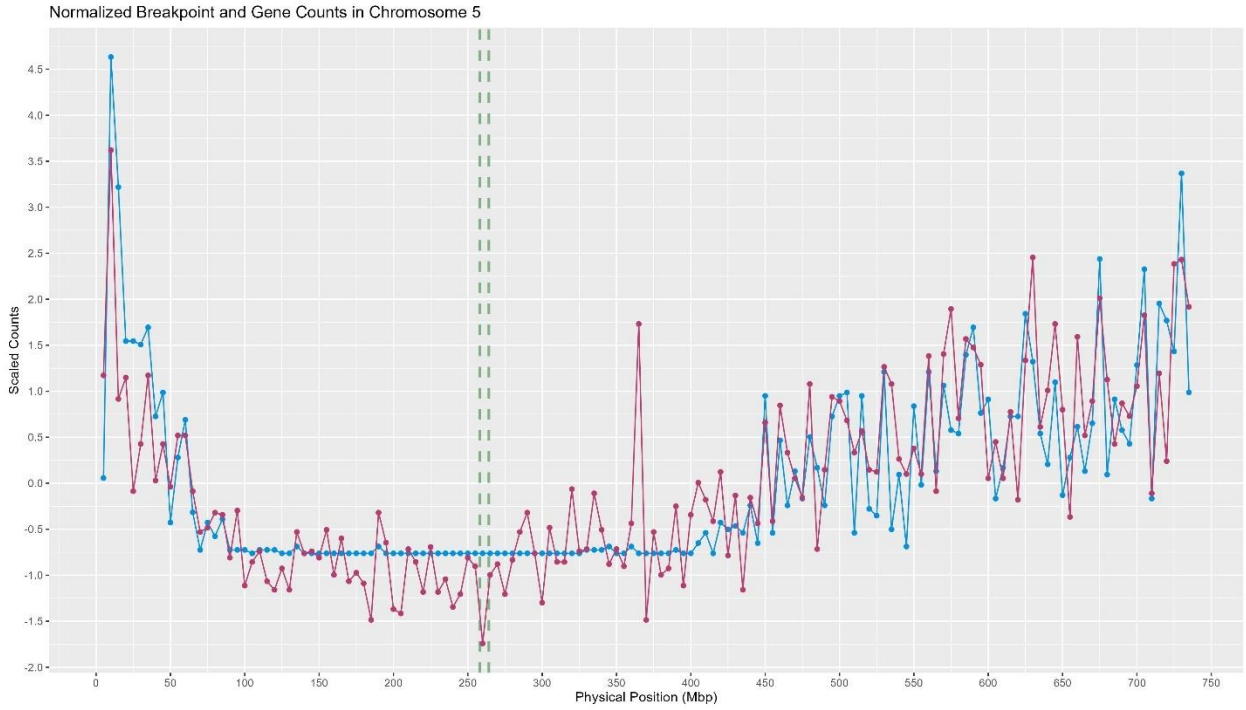

Figure S10.-- Standard scaled dist. of crossovers and gene counts in chromosome 5A.

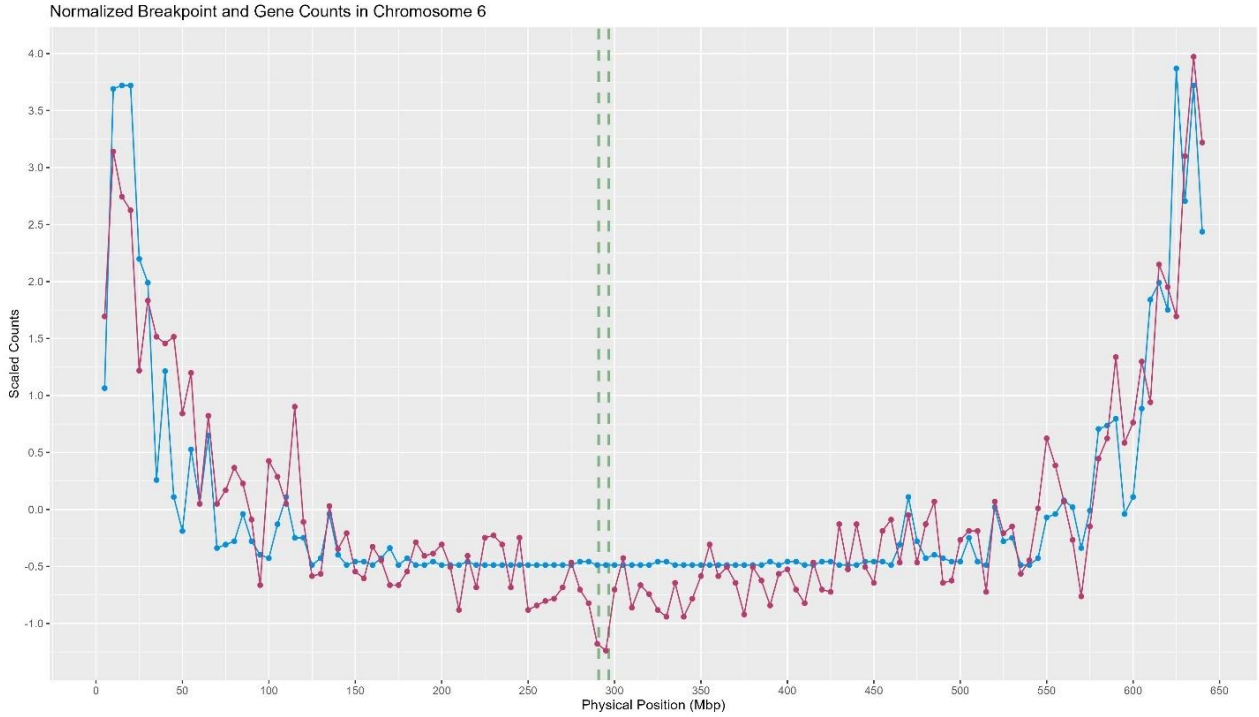

Figure S11.-- Standard scaled dist. of crossovers and gene counts in chromosome 6A.

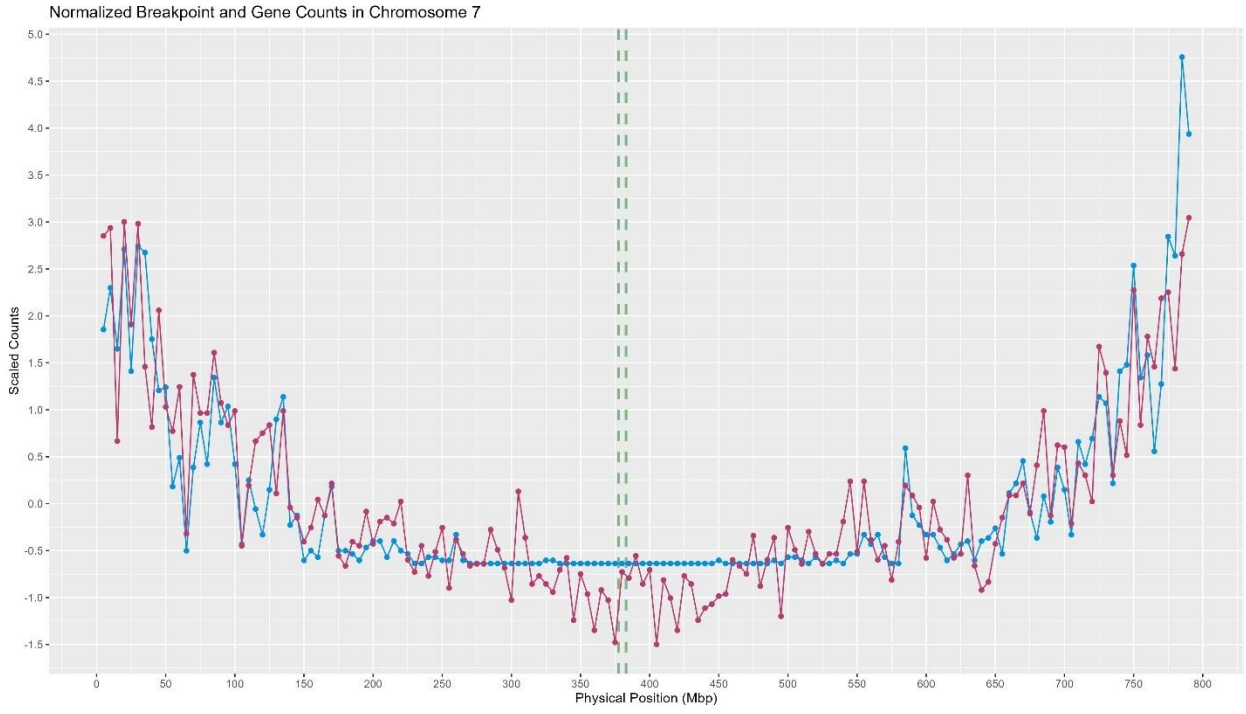

Figure S12.-- Standard scaled dist. of crossovers and gene counts in chromosome 7A.

The following set of supplemental figures are paired with Figure 11, presenting the dot plot of HC/LC gene count vs crossover count in chromosomes 2-7 (Supplemental Figure S13- S18). The blue line represents a linear regression with a 95% confidence interval.

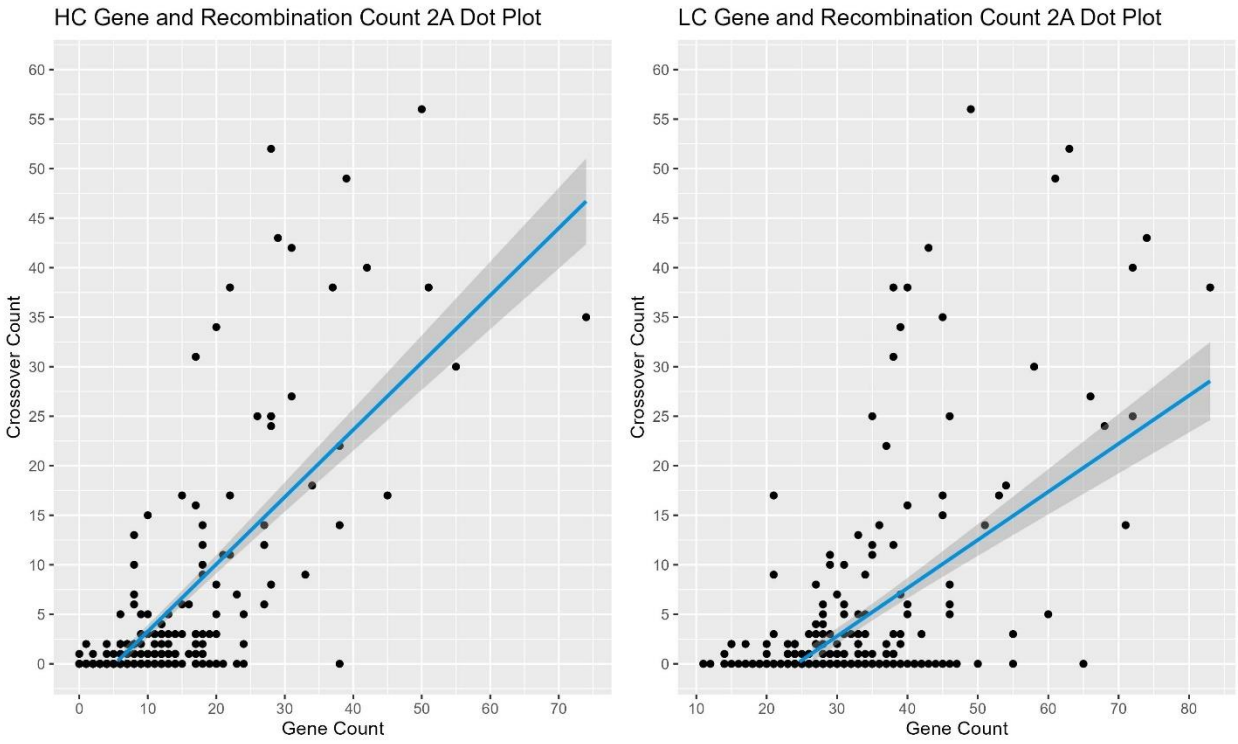

Figure S2.--HC/LC gene count VS crossover count dot plot for chromosome 2A.

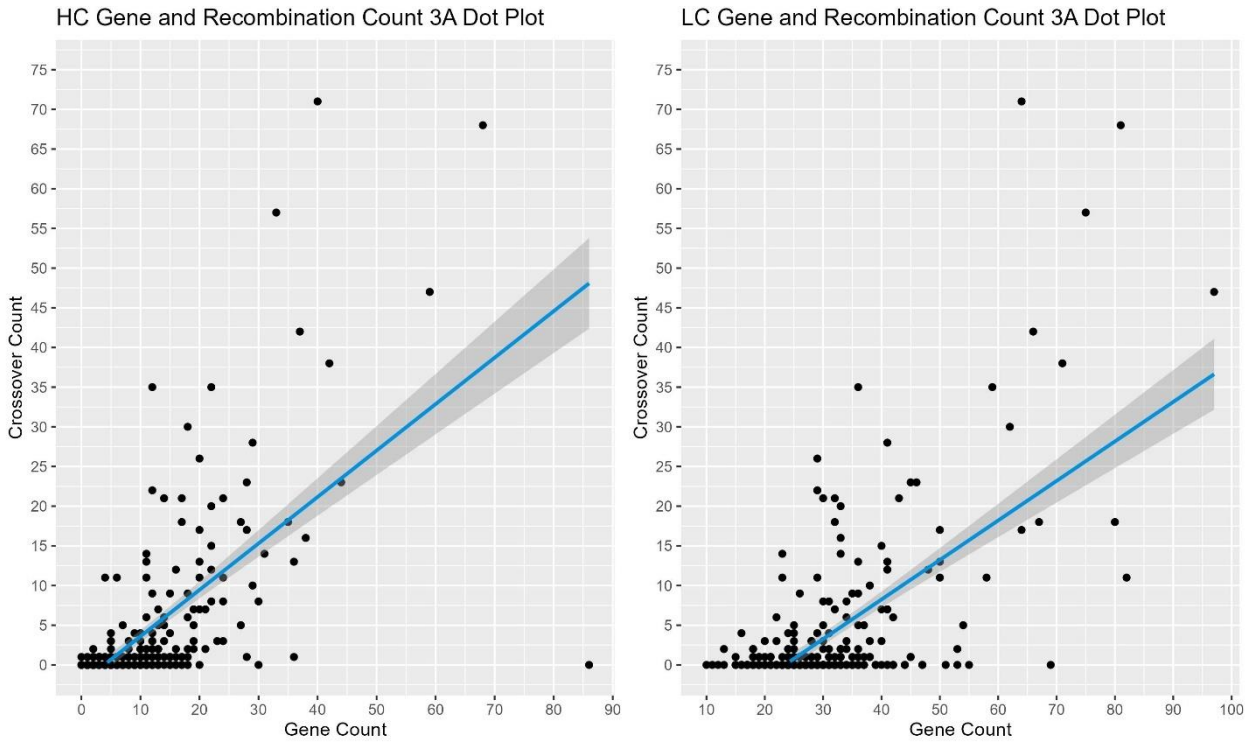

Figure S14.--HC/LC gene count VS crossover count dot plot for chromosome 3A.

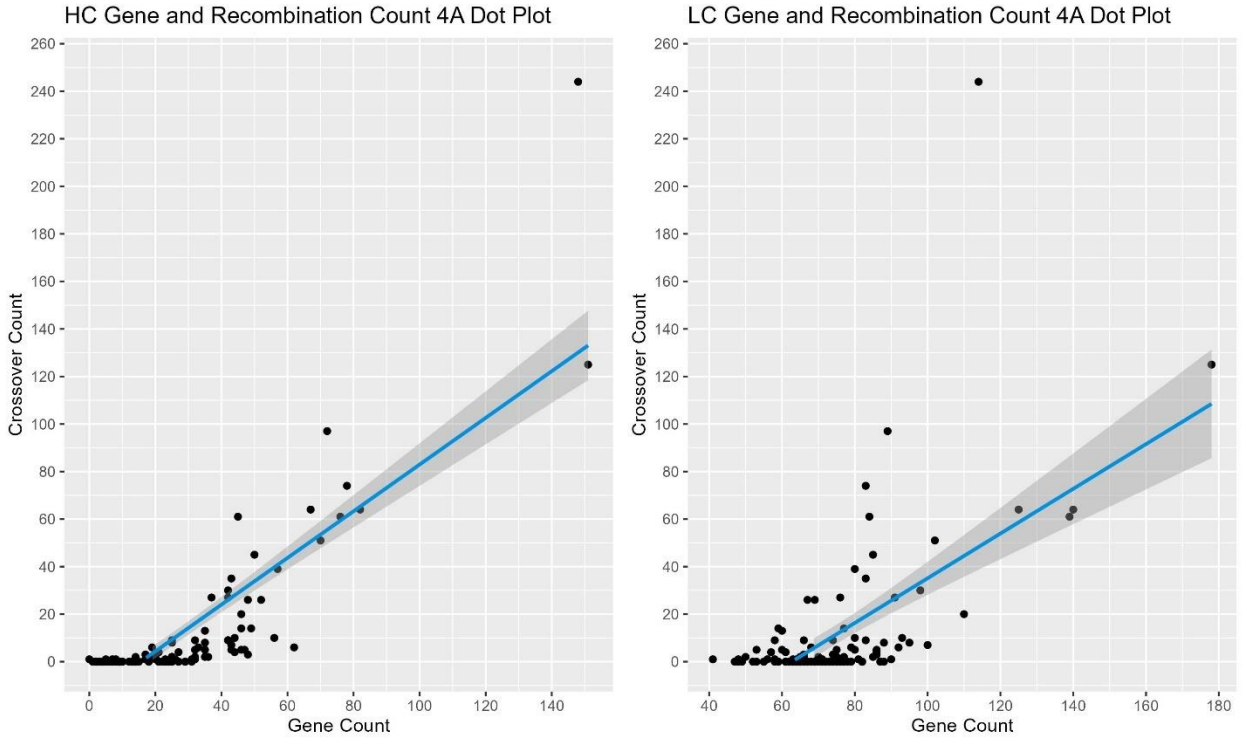

Figure S15.--HC/LC gene count VS crossover count dot plot for chromosome 4A.

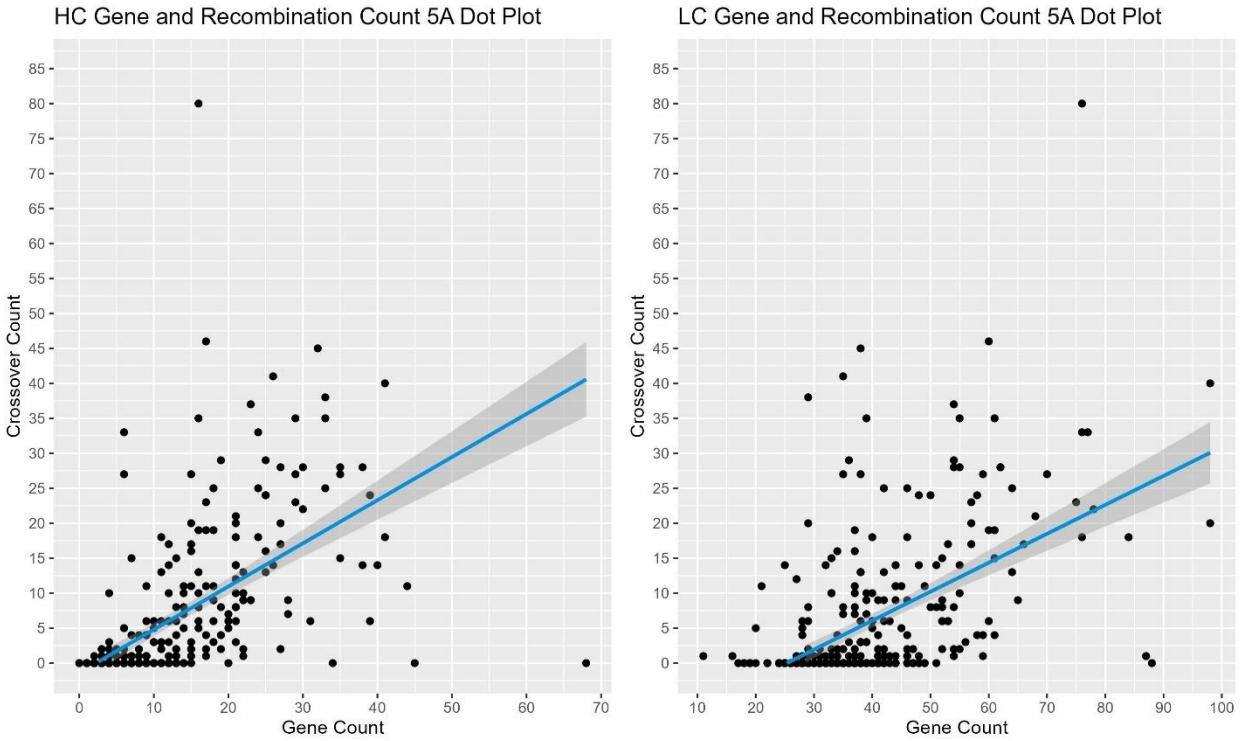

Figure S16.--HC/LC gene count VS crossover count dot plot for chromosome 5A.

Figure S17.--HC/LC gene count VS crossover count dot plot for chromosome 6A.

Figure S18.--HC/LC gene count VS crossover count dot plot for chromosome 7A.