University of Nebraska - Lincoln
DigitalCommons@University of Nebraska - Lincoln

Theses, Dissertations, and Student Research in Agronomy and Horticulture

Agronomy and Horticulture Department

6-2022

# RECOMBINATION HOTSPOTS IN SOYBEAN [GLYCINE MAX (L.) MERR.]

Samantha J. McConaughy
*University of Nebraska-Lincoln*

Follow this and additional works at: https://digitalcommons.unl.edu/agronhortdiss

 Part of the Agricultural Science Commons, Agriculture Commons, Agronomy and Crop Sciences Commons, Horticulture Commons, Other Plant Sciences Commons, and the Plant Biology Commons

McConaughy, Samantha J., "RECOMBINATION HOTSPOTS IN SOYBEAN [GLYCINE MAX (L.) MERR.]" (2022). *Theses, Dissertations, and Student Research in Agronomy and Horticulture*. 236.
https://digitalcommons.unl.edu/agronhortdiss/236

This Article is brought to you for free and open access by the Agronomy and Horticulture Department at DigitalCommons@University of Nebraska - Lincoln. It has been accepted for inclusion in Theses, Dissertations, and Student Research in Agronomy and Horticulture by an authorized administrator of DigitalCommons@University of Nebraska - Lincoln.

# RECOMBINATION HOTSPOTS IN SOYBEAN [*GLYCINE MAX* (L.) MERR.]

by

Samantha J. McConaughy

A DISSERTATION

Presented to the Faculty of

The Graduate College at the University of Nebraska

In Partial Fulfillment of Requirements

For the Degree of Doctor of Philosophy

Major: Agronomy

(Plant Breeding and Genetics)

Under the Supervision of Professor David L. Hyten

Lincoln, Nebraska

June, 2022

RECOMBINATION HOTSPOTS IN SOYBEAN [*GLYCINE MAX* (L.) MERR.]

Samantha J. McConaughy, Ph.D.

University of Nebraska, 2022

Advisor: David L. Hyten

Recombination allows for the exchange of genetic material between two parents which plant breeders exploit to make new and improved varieties. This recombination is not distributed evenly across the chromosome. In crops, it mostly occurs in the euchromatic regions of the genome and even then, recombination is focused into recombination hotspots flanked by recombination cold spots. Understanding the distribution of these hotspots along with the sequence motifs associated with them may lead to methods that enable breeders to better exploit recombination in breeding.

In chapter 1 background information on recombination, recombination hotspots detection methods, landscape of recombination (describe recombination patterns along the genome), and environmental influence on recombination hotspot locations are outlined. In chapter 2 recombination hotspots were mapped in two-biparental soybean [*Glycine max* (L.) Merr.] recombinant inbred line (RIL) populations, Williams crossed by Essex (WE) and Williams 82 crossed by PI479752 (WP). These populations consist of 922 RIL(WE) and 1,086 RIL (WP) and were genotyped with 50,000 SNP markers using the SoySNP50k Illumina Infinium assay. In chapter 3 the location of recombination hotspots in the USDA Soybean Germplasm Collection in three populations: wild (806),

landraces (5396), and North American cultivars (563) are reported. Genotyping was conducted using the SoySNP50k Illumina Infinium assay. Germplasm hotspot locations were compared to results in chapter 2, two-biparental soybean recombinant inbred line (RIL) populations. In chapter 2 and 3 statistical tests were conducted for genome features association with hotspot locations based on logistical regression, discovered nucleotide motifs surrounding hotspot regions across the genome.

## Acknowledgments

I would like to thank Dr. David Hyten for providing the opportunity to join his soybean genomics research team to complete my Ph.D research. I could not have asked for a better person to not only teach me molecular plant breeding but guide my professional development. His dedication to student development and commitment to providing resources to the soybean community has profoundly influenced and encouraged me as a person and a scientist.

I would also like to thank my committee members, Dr. Reka Howard, Dr. Keenan Amundsen, Dr. George Graef, and Dr. David Holding, who have helped guide my research and provided their expertise. Special thanks to Dr. Keenan Amundsen who has not only provided access to his resources but insightful conversations. I am grateful for my committee's guidance, dedication, and encouragement.

Thank you to John Wang, the laboratory technician for the soybean genomics research team. He helped me with techniques and assisting with greenhouse/laboratory work. I would also like to thank the graduate students in the program including Mary Happ, Erika Sanchez Betancourt, and Sarah Johnson. Thank you for the laughter, sharing of notes, and advice on work. And I would be remiss if I didn't acknowledge Piyaporn (Bee) Phansak who helped collect and dissect numerous flowers during her visits.

Lastly, I would like to thank my family: my parents, Jean and Scott McConaughy, my brother Benjamin, his wife Shanon, their three children, Charlie, Carter, and Parker

my husband Benjamin Fisher and our daughter, Viola. Their endless support, love, and advice have been my motivation.

**Table of Contents**

Page

## List of Tables

Page

## List of Figures

Page

# Introduction

The non-random distribution of crossovers, the exchange of genetic between homologous chromosomes, offers an opportunity to engineer new allelic combinations; however, high-density genetic maps are needed to first identify and define characteristics of recombination hotspots. In soybean, two recombinant inbred populations with high-resolution genetic maps are publicly available, 'Williams 82' [*Glycine max* (L.) Merr.] crossed to 'Essex' (*G. max*) as well as PI479752 (*Glycine soja)* crossed to Essex. Additionally, high-density genomics information is available on the USDA soybean germplasm collection that consists of 1,168 wild accession (*G. soja*) and 18,480 domesticated lines (*G. max*). These available data sets are resources that can be used for the characterization of recombination hotspots.

In chapter 1 background information on recombination, recombination hotspots detection methods, landscape of recombination (describe recombination patterns along the genome), and environmental influence on recombination hotspot locations are outlined. In chapter 2 recombination hotspots were mapped in two-biparental soybean [*Glycine max* (L.) Merr.] recombinant inbred line (RIL) populations, Williams crossed by Essex (WE) and Williams 82 crossed by PI479752 (WP). In chapter 3 the location of recombination hotspots in the USDA Soybean Germplasm Collection in three populations: wild populations (806), landraces (5396), and North American cultivars (563) are reported. Germplasm hotspot locations were compared to results in chapter 2, two-biparental soybean recombinant inbred line (RIL) populations. In chapter 2 and 3 statistical tests were conducted for genome features association with hotspot locations

based on logistical regression, discovered nucleotide motifs surrounding hotspot regions across the genome.

# CHAPTER ONE

## Literature Review

### Recombination

Crossovers allow for the exchange of genetic material, which plant breeders exploit to make new and improved varieties. Plant breeders have been utilizing recombination for generations. They use recombination to map quantitative traits by development of recombinant inbred lines, to backcross in traits of interest, to create near isogenic lines, and create new and improved varieties. Recombination occurs during the pachytene stage of prophase I of meiosis, where recombination nodules appear along the synaptonemal complex. At these nodules, crossover occurs between non-sister chromatids, which results in the exchange of genetic material, termed recombination.

There are two known pathways that produce crossovers; Class I, interference and Class II, no interference (Berchowitz and Copenhaver, 2010). In Class I, the interference of one crossover decreases the probability of another crossover occurring near the first crossover in the same pair of chromosomes. In Class II, crossover events do not affect the probability of another crossover event happening in a nearby region. A group of proteins called ZMMs (Zip1, 2,3, Msh 4,5, and Mer3) are responsible for Class I. Class II involves the endonuclease MUS81 (Youds and Boulton, 2011).

The formation of crossovers is a highly regulated process that involves DNA double stranded breaks (DSBs). Until the discovery of recombination hotspots, many researchers believed recombination happened randomly across chromosomes (Lichten and Goldman, 1995). A recombination hotspot is a location in the genome where there is

an increased amount of crossovers that occur in comparison to other regions throughout the genome. Cold spots also exist, where crossovers seldomly occur. An initial approach to study recombination hotspots began with the use of molecular techniques to detect double stranded breaks (Keeney, 2008). Murakami and Keeney (2008) showed that DSB are not randomly distributed but rather occur in specific areas of chromosomes, through a process that is not well understood. Additionally, there are only a few crossovers that occur per meiosis despite the large amount of double stranded breaks (Crismani and Mercier, 2012). One example is Arabidopsis where approximately 200 DSB are observed but only 10 crossovers on average are detected per meiosis (Chelysheva *et al.*, 2007; Mercier *et al.*, 2005; Sanchez-Moran *et al.*, 2007) Therefore, just observing DSB alone is not an approach to advancing the understanding of recombination.

Coupling DSB and genomic information provides insight into recombination hotspots. One such aspect is to look at what genes affect recombination. Other aspects include looking at the distribution of crossovers, understanding the machinery of DSB formation, studying mutants that affect crossover frequency and positioning, and better understanding biotic and abiotic impact on crossover positioning. The connection between information associated with crossovers creates a foundation for better understanding of the mechanisms governing recombination.

Identifying genes that modulate recombination has been successful (Crismani and Mercier, 2012). An approach to understanding recombination is to characterize factors that limit meiotic crossovers (Crismani and Mercier, 2012). In order to study what limits meiotic crossover, one group hypothesized that crossover defective mutants could regain crossover functionality by mutations that increase crossovers, also known as gain of

function mutants (Crismani *et al.*, 2012). *Arabidopsis* was mutated with ethylmethane sulfonate to create *zmm zip4* mutant seeds, lacking meiotic crossovers, (Crismani *et al.*, 2012). Zip4 mutants have mis-segregation of homologs, which reduces fertility. The mutated plants contain different properties than the original mutation; eight of the lines showed an increase in fertility These lines showed single recessive mutations that were apart of three complementation groups This discovery led to a gene that is a homolog of *Fanconi anemia complementation group M* (FANCM) in humans. This was the first time that FANCM was observed in association to meiotic crossovers; FANCM is known to be involved with genome stability (Crismani *et al.*, 2012). Moreover, the research suggests that FANCM is involved in meiotic double strand repair mechanisms and allows for hyperrecombination (Crismani *et al.*, 2012). That crossover frequency is maintained through generations by natural selection (Crismani *et al.*, 2012).

Two other proteins have been identified in association with a decreasing the number of crossovers, MHF1 and MHF2 (Girard *et al.*, 2014). MHF1 and MHF2 were identified from a group of other Fanconi Anemia (FA) proteins, through similar approaches as what led to the discovery of FANCM, by identifying the restoration of crossovers in defective mutants (Crismani *et al.*, 2012; Girard *et al.*, 2014). Fanconi Anemia is a rare genetic disorder that is characterized by bone marrow failure and physical abnormalities. The pathway in which FA works is known to be present in all eukaryotes and promotes genome stability (Kottemann and Smogorzewska, 2013). Additionally, MHF1 and MHF2 act in the same pathway as FANCM to limit meiotic recombination in the Class II pathway (Girard *et al.*, 2014). The single mutants *Atmhf2-2* and *Atmhf2-1* overal genetic map length increased by ~60% in comparison to wild type

(Girard *et al.*, 2014). FANCM is not dependent on the presence of MHF1 and MHF2, however MHF1 and MHF2 provide FANCM with the formation of a heterotetramer that limits crossover occurrence during meiosis (Girard *et al.*, 2014).

Recombination detection methods

Crossover distribution can be studied by observing the coinheritance of heterozygous markers in a population. Markers physically closer to each other are more likely to be co-inherited than markers that are more separated. Crossover events that occur historically in populations, between two markers, break the genetic association that they are linked. Therefore, historical crossover events can be calculated based on linkage disequilibrium (LD) analysis, which detects the association of a set of markers on chromosomes in a population (Lambing *et al.*, 2017).

LD analysis has been used in human genetics to create a fine scale map of recombination rates and hotspots in humans (Myers *et al.*, 2005). The model is based on coalescent patterns of LD, this is often referred to as coalescent-based methods. The method begins by fitting a statistical model based on the associations of SNPs and estimate recombination rates within a Bayesian framework Priors are used in the Bayesian framework to smooth and reduce over fitting of recombination rates. The software program LDHat has been developed to combine the coalescent-based methods within a Bayesian framework. (McVean *et al.*, 2004). LDHat is the product of PHASE, the original model for linkage disequilibrium to identify recombination hotspots using SNPs (Li and Stephens, 2003).

Recombination landscape

There are few similarities in the distribution of crossovers across humans, plants, or animals but some similarities have been found in DNA motifs. In humans, crossovers occur at intergenic CCN repeat motifs, 13-mer CCNCCNTNNCCNC. These motifs are bound by PRDM9 (PR/SET Domain 9) a zinc finger protein with histone methyltransferace activity (Myers *et al.*, 2008). A similar CCN repeat motif has also been found in wheat and Arabidopsis but homolog for PRDM9 has not been found in plants (Darrier *et al.*, 2017; Shilo *et al.*, 2015). In plants, poly-A stretches have been associated with recombination hotspots. (Comeron *et al.*, 2012; Darrier *et al.*, 2017; Myers *et al.*, 2008). In wheat, poly-A motif was found in 51.6% of hotspots (Darrier *et al.* 2017).

Crossover frequency is higher in sub-telomeric regions and lower in interstitial regions in maize and barely (Gore *et al.*, 2009; Saintenac *et al.*, 2009). Recombination isrepressed at the centromeres and telomeres (Gore *et al.*, 2009; Saintenac *et al.*, 2009). In Arabidopsis, crossovers are detected along the chromosomes except at the centromere (Salome *et al.*, 2012). Lambing *et al.* (2017) hypothesize that the difference in crossover frequency between maize and barley versus Arabidopsis is due to genome organization, specifically in relation to transposons. When DNA methylation and transposon frequencies are increased there is a decrease in the amount of recombination in Arabidopsis, rice, and maize (Lambing *et al.*, 2017). Also, the DNA methylation pattern is more stable across the maize chromosomes than compared to Arabidopsis (Lambing *et al.*, 2017). On a smaller scale, recombination hotspots tend to be located toward the 5' and 3' ends of genes in maize populations (Lambing *et al.*, 2017). This has also been observed in Arabidopsis with most recombination occurring at gene promoters and terminators regions (Choi and Henderson, 2015). In Arabidopsis, historical

recombination is highest immediately downstream of gene transcriptional start sites (Choi and Henderson, 2015). In general, for plants, recombination hotspots are the highest among gene-rich euchromatic regions and lowest in repeat-rich heterochromatic regions (Choulet *et al.*, 2014; Gore *et al.*, 2009; Li *et al.*, 2015; Liu *et al.*, 2009; Rodgers-Melnick *et al.*, 2015; Song *et al.*, 2016; Wei *et al.*, 2009).

Recombination frequency is modulated by the environment

Extrinsic and intrinsic factors can modulate the recombination rate (Modliszewski and Copenhaver, 2017). Plants are bound to their environment; under stressful conditions, a defense mechanism that plants can utilize is changing the recombination rate. Any methods that would increase crossovers in low recombination regions would increase genetic diversity which is beneficial to plant breeders (Crismani *et al.*, 2013; Lambing *et al.*, 2017). Temperature changes from 22°C to 30°C caused a reduction in chiasmata and seed set in barley (Higgins *et al.*, 2012). In Arabidopsis a temperature change from 18°C to 28°C was associated with an increase in recombination frequency (Francis *et al.*, 2007). In many species, recombinaton rate in response to temperature follows a U-shaped pattern (Modliszewski and Copenhaver, 2017). Also, an increase in recombination is observed between stress and non-stress treatments with water. Genetic recombination maps of stressed (drought treated) maize were 15% longer than non-stressed maize (Verde, 2003).

Pathogen infections can cause an increased recombination rate (Lambing *et al.*, 2017). In Arabidopsis, a systemic increase in recombination frequency was associated with infection from the tobacco mosaic virus and oilseed rape mosaic virus (Kovalchuk *et*

*al.*, 2003; Yao *et al.*, 2013). Similar results have been observed in barley and tomato plants infected with mosaic viruses, a shift in the position of the chiasmata toward interstitial regions (Andronic, 2012). Even though abnormal tetrads were detected, the number of total chiasma was not significantly increased (Andronic, 2012). Heat stress and pathogen infections also increased the recombination rate in rice (Si *et al.*, 2015).

Plant breeders' impact can be measured based on their ability to combine favorable alleles into new varieties (Crismani *et al.*, 2013; Wijnker and de Jong, 2008). Two factors that control the maximum meiotic recombination are the number of chromosomes and the number and positions of crossovers on the pairs of homologous chromosomes (crossovers recombination) (Wijnker and de Jong, 2008). Crossovers recombination is a tightly constrained process in terms of the number and distribution (Wijnker and de Jong, 2008). Since the number of crossovers is usually one or two per chromosome, a method to increase recombination would be a powerful tool for plant breeders (Sturtevant, 1915).

## References

**Andronic L**. 2012. Viruses as triggers of DNA rearrangements in host plants. *Canadian Journal of Plant Science* **92**, 1083-1091.
**Berchowitz LE, Copenhaver GP**. 2010. Genetic interference: don't stand so close to me. *Curr Genomics* **11**, 91-102.
**Chelysheva L, Gendrot G, Vezon D, Doutriaux MP, Mercier R, Grelon M**. 2007. Zip4/Spo22 is required for class I CO formation but not for synapsis completion in Arabidopsis thaliana. *Plos Genetics* **3**, e83.
**Choi K, Henderson IR**. 2015. Meiotic recombination hotspots - a comparative view. *Plant Journal* **83**, 52-61.
**Choulet F, Alberti A, Theil S, Glover N, Barbe V, Daron J, Pingault L, Sourdille P, Couloux A, Paux E, Leroy P, Mangenot S, Guilhot N, Le Gouis J, Balfourier F, Alaux M, Jamilloux V, Poulain J, Durand C, Bellec A, Gaspin C, Safar J, Dolezel J, Rogers J, Vandepoele K, Aury JM, Mayer K, Berges H, Quesneville H, Wincker P, Feuillet C**. 2014. Structural and functional partitioning of bread wheat chromosome 3B. *Science* **345**, 1249721.
**Comeron JM, Ratnappan R, Bailin S**. 2012. The many landscapes of recombination in Drosophila melanogaster. *Plos Genetics* **8**, e1002905.
**Crismani W, Girard C, Froger N, Pradillo M, Santos JL, Chelysheva L, Copenhaver GP, Horlow C, Mercier R**. 2012. FANCM Limits Meiotic Crossovers. *Science* **336**, 1588-1590.
**Crismani W, Girard C, Mercier R**. 2013. Tinkering with meiosis. *Journal of Experimental Botany* **64**, 55-65.
**Crismani W, Mercier R**. 2012. What limits meiotic crossovers? *Cell Cycle* **11**, 3527-3528.
**Darrier B, Rimbert H, Balfourier F, Pingault L, Josselin AA, Servin B, Navarro J, Choulet F, Paux E, Sourdille P**. 2017. High-Resolution Mapping of Crossover Events in the Hexaploid Wheat Genome Suggests a Universal Recombination Mechanism. *Genetics* **206**, 1373-1388.
**Francis KE, Lam SY, Harrison BD, Bey AL, Berchowitz LE, Copenhaver GP**. 2007. Pollen tetrad-based visual assay for meiotic recombination in Arabidopsis. *Proceedings of the National Academy of Sciences of the United States of America* **104**, 3913-3918.
**Girard C, Crismani W, Froger N, Mazel J, Lemhemdi A, Horlow C, Mercier R**. 2014. FANCM-associated proteins MHF1 and MHF2, but not the other Fanconi anemia factors, limit meiotic crossovers. *Nucleic Acids Research* **42**, 9087-9095.
**Gore MA, Chia JM, Elshire RJ, Sun Q, Ersoz ES, Hurwitz BL, Peiffer JA, McMullen MD, Grills GS, Ross-Ibarra J, Ware DH, Buckler ES**. 2009. A first-generation haplotype map of maize. *Science* **326**, 1115-1117.
**Higgins JD, Perry RM, Barakate A, Ramsay L, Waugh R, Halpin C, Armstrong SJ, Franklin FC**. 2012. Spatiotemporal asymmetry of the meiotic program underlies the predominantly distal distribution of meiotic crossovers in barley. *Plant Cell* **24**, 4096-4109.
**Keeney S**. 2008. Spo11 and the Formation of DNA Double-Strand Breaks in Meiosis. *Genome Dyn Stab* **2**, 81-123.

**Kottemann MC, Smogorzewska A**. 2013. Fanconi anaemia and the repair of Watson and Crick DNA crosslinks. *Nature* **493**, 356-363.
**Kovalchuk I, Kovalchuk O, Kalck V, Boyko V, Filkowski J, Heinlein M, Hohn B**. 2003. Pathogen-induced systemic plant signal triggers DNA rearrangements. *Nature* **423**, 760-762.
**Lambing C, Franklin FCH, Wang CJR**. 2017. Understanding and Manipulating Meiotic Recombination in Plants. *Plant Physiology* **173**, 1530-1542.
**Li N, Stephens M**. 2003. Modeling linkage disequilibrium and identifying recombination hotspots using single-nucleotide polymorphism data. *Genetics* **165**, 2213-2233.
**Li X, Li L, Yan J**. 2015. Dissecting meiotic recombination based on tetrad analysis by single-microspore sequencing in maize. *Nature Communications* **6**, 6648.
**Lichten M, Goldman AS**. 1995. Meiotic recombination hotspots. *Annu Rev Genet* **29**, 423-444.
**Liu S, Yeh CT, Ji T, Ying K, Wu H, Tang HM, Fu Y, Nettleton D, Schnable PS**. 2009. Mu transposon insertion sites and meiotic recombination events co-localize with epigenetic marks for open chromatin across the maize genome. *Plos Genetics* **5**, e1000733.
**McVean GAT, Myers SR, Hunt S, Deloukas P, Bentley DR, Donnelly P**. 2004. The fine-scale structure of recombination rate variation in the human genome. *Science* **304**, 581-584.
**Mercier R, Jolivet S, Vezon D, Huppe E, Chelysheva L, Giovanni M, Nogue F, Doutriaux MP, Horlow C, Grelon M, Mezard C**. 2005. Two meiotic crossover classes cohabit in Arabidopsis: one is dependent on MER3,whereas the other one is not. *Current Biology* **15**, 692-701.
**Modliszewski JL, Copenhaver GP**. 2017. Meiotic recombination gets stressed out: CO frequency is plastic under pressure. *Current Opinion in Plant Biology* **36**, 95-102.
**Myers S, Bottolo L, Freeman C, McVean G, Donnelly P**. 2005. A fine-scale map of recombination rates and hotspots across the human genome. *Science* **310**, 321-324.
**Myers S, Freeman C, Auton A, Donnelly P, McVean G**. 2008. A common sequence motif associated with recombination hot spots and genome instability in humans. *Nature Genetics* **40**, 1124-1129.
**Rodgers-Melnick E, Bradbury PJ, Elshire RJ, Glaubitz JC, Acharya CB, Mitchell SE, Li CH, Li YX, Buckler ES**. 2015. Recombination in diverse maize is stable, predictable, and associated with genetic load. *Proceedings of the National Academy of Sciences of the United States of America* **112**, 3823-3828.
**Saintenac C, Falque M, Martin OC, Paux E, Feuillet C, Sourdille P**. 2009. Detailed recombination studies along chromosome 3B provide new insights on crossover distribution in wheat (Triticum aestivum L.). *Genetics* **181**, 393-403.
**Salome PA, Bomblies K, Fitz J, Laitinen RA, Warthmann N, Yant L, Weigel D**. 2012. The recombination landscape in Arabidopsis thaliana F2 populations. *Heredity (Edinb)* **108**, 447-455.
**Sanchez-Moran E, Santos JL, Jones GH, Franklin FC**. 2007. ASY1 mediates AtDMC1-dependent interhomolog recombination during meiosis in Arabidopsis. *Genes Dev* **21**, 2220-2233.

**Shilo S, Melamed-Bessudo C, Dorone Y, Barkai N, Levy AA**. 2015. DNA Crossover Motifs Associated with Epigenetic Modifications Delineate Open Chromatin Regions in Arabidopsis. *Plant Cell* **27**, 2427-2436.
**Si WN, Yuan Y, Huang J, Zhang XH, Zhang YC, Zhang YD, Tian DC, Wang CL, Yang YH, Yang SH**. 2015. Widely distributed hot and cold spots in meiotic recombination as shown by the sequencing of rice F-2 plants. *New Phytologist* **206**, 1491-1502.
**Song Q, Jenkins J, Jia G, Hyten DL, Pantalone V, Jackson SA, Schmutz J, Cregan PB**. 2016. Construction of high resolution genetic linkage maps to improve the soybean genome sequence assembly Glyma1.01. *Bmc Genomics* **17**, 33.
**Sturtevant AH**. 1915. The behavior of the chromosomes as studied through linkage. *Molecular and General Genetics MGG* **13**, 234-287.
**Verde LA**. 2003. The effect of stress on meiotic recombination in maize (Zea mays L).
**Wei F, Zhang J, Zhou S, He R, Schaeffer M, Collura K, Kudrna D, Faga BP, Wissotski M, Golser W, Rock SM, Graves TA, Fulton RS, Coe E, Schnable PS, Schwartz DC, Ware D, Clifton SW, Wilson RK, Wing RA**. 2009. The physical and genetic framework of the maize B73 genome. *Plos Genetics* **5**, e1000715.
**Wijnker E, de Jong H**. 2008. Managing meiotic recombination in plant breeding. *Trends in Plant Science* **13**, 640-646.
**Yao YL, Kathiria P, Kovalchuk I**. 2013. A systemic increase in the recombination frequency upon local infection of Arabidopsis thaliana plants with oilseed rape mosaic virus depends on plant age, the initial inoculum concentration and the time for virus replication. *Frontiers in Plant Science* **4**.
**Youds JL, Boulton SJ**. 2011. The choice in meiosis - defining the factors that influence crossover or non-crossover formation. *J Cell Sci* **124**, 501-513.

# CHAPTER TWO

## RECOMBINATION HOTSPOTS IN SOYBEAN [*GLYCINE MAX* (L.) MERR.]

### Abstract

Recombination allows for the exchange of genetic material between two parents which plant breeders exploit to make new and improved varieties. This recombination is not distributed evenly across the chromosome. In crops, recombination mostly occurs in euchromatic regions of the genome and even then, recombination is focused into recombination hotspots, clusters of crossovers. Understanding the distribution of these hotspots along with the sequence motifs associated with them may lead to methods that enable breeders to better exploit recombination in breeding. To map recombination hotspots and identify sequence motifs associated with hotspots in soybean [*Glycine max* (L.) Merr.], two bi-parental recombinant inbred lines (RILs) populations were genotyped with 50,000 SNP markers using the SoySNP50k Illumina Infinium assay. A total of 451 recombination hotspots were identified in the two populations. Despite being half-sib populations, only 18 hotspots were in common between the two populations. While pericentromeric regions did exhibit extreme suppression of recombination, twenty-seven percent of the detected hotspots were located in the pericentromic regions of the chromosomes. Two genomic motifs associated with hotspots are similar to human, dog, rice, wheat, drosophila, and arabidopsis. These motifs were a CCN repeat motif and a poly-A motif. Genomic regions spanning other hotspots were significantly enriched with the tourist family of mini-inverted-repeat transposable elements (MITEs) that resides in

less than 0.34% of the soybean genome. The characterization of recombination hotspots in these two large soybean bi-parental populations demonstrates that hotspots do occur throughout the soybean genome and are enriched for specific motifs but their locations may not be conserved between different populations.

## Introduction

Recombination is a fundamental process that drives the rearrangement of alleles resulting in new combinations of those alleles. This can lead to favorable allelic combinations and break up undesirable ones, such as decoupling linked deleterious alleles (Barton and Charlesworth, 1998). However, recombination is a highly regulated process that involves DNA double stranded breaks, resulting in crossover or non-crossover events (Baudat *et al.*, 2013) Crossovers, at least one per chromosome, are required for proper chromosome segregation (Mercier *et al.*, 2015a). The distribution, associated motifs, strength, and size of recombination vary across all eukaryotes, potentially implying a lack of unifying characteristics (Choi and Henderson, 2015; Choulet *et al.*, 2014; Darrier *et al.*, 2017; Marand *et al.*, 2017; Mercier *et al.*, 2015b; Saintenac *et al.*, 2009; Tenaillon *et al.*, 2002).

While the genomic landscape of recombination varies across plant species, it transpires in a small percentage of the genome. Recombination rates are the highest in gene-rich euchromatic regions and lowest in repeat-rich heterochromatic regions (Choulet *et al.*, 2014; Gore *et al.*, 2009; Li *et al.*, 2015; Liu *et al.*, 2009; Paterson *et al.*, 2009; Rodgers-Melnick *et al.*, 2015; Song *et al.*, 2016; Wei *et al.*, 2009). In soybean [*Glycine max* (L.) Merr.], 93% of recombination occurs in the euchromatic DNA which only accounts for 43% of the genome (Schmutz *et al.*, 2010). Recombination rates across the euchromatic regions in the soybean genome average one centimorgan (cM) per 197 kilobase (kb) while in pericentromeric regions the average increases to one cM per 3.5 megabase (Mb) (Schmutz *et al.*, 2010).

Due to the non-random pattern of recombination, the relationship between the physical distance and genetic map distance results in it not being a one-to-one ratio across the chromosome. Crossovers are clustered into small physical distances that are defined as recombination hotspot regions. These recombination hotspots are often flanked by regions of DNA that have lower than expected recombination called recombination cold spots. In plants, hotspot regions have been discovered to vary in size from 500 to 23,000 base pairs (bp), whereas cold spots can range from 5,000 to millions of bp in length (Drouaud *et al.*, 2013b; Fu *et al.*, 2002; Patterson *et al.*, 1995; Saintenac *et al.*, 2009; Yao and Schnable, 2005; Yelina *et al.*, 2012). The resolution to identify recombination hotspots and cold spots is often limited by marker density and population size.

While the size of recombination hotspots and cold spots are an important attribute, the distribution of recombination influences the probability of combining new allelic combinations and introgressing diversity in crops. In maize and wheat, recombination hotspots are located in sub-telomeric regions, whereas cold spots are concentrated near centromeres, telomeres, and interstitial regions (Lambing et al. 2017). In contrast, recombination hotspots in Arabidopsis are not concentrated in specific regions, but are dispersed throughout the chromosome, except in the centromere (Salome *et al.*, 2012). Lambing *et al.* (2017) hypothesized the difference between Arabidopsis, maize, and wheat hotspot locations may be due to transposable elements.

While the distribution of hotspots varies by species and genome locations, the relationship between transposable elements along with sequence motif may help explain this difference. Several experiments have identified closely associated repeat motifs to recombination hotspots. The first discovery was in humans with a CCN-like motif bound

by PRDM9 (PR/SET Domain 9) a zinc finger protein with histone methyltransferace activity (Myers *et al.*, 2008). The 13-mer motif, CCNCCNTNNCCNC, was reported in 40% of human hotspots (Myers *et al.*, 2008). In plants, three common motifs have been found in associations with hotspots including a CCN-repeat, CTT-repeat, and poly-A stretch motif (Choi *et al.*, 2013; Shilo *et al.*, 2015; Wijnker and de Jong, 2008). The annotation of the CCN-like motif discovered in wheat is related with the terminal inverted repeat (TIR)-Mariner sequence. These motifs are suggested to be involved in chromatin structure within promoter regions due to the location near transcription start sites and the similarity to PRDM9 protein (Darrier *et al.*, 2017). Since plants do not have a known homolog for PRDM9, the DNA transposons from the Mariner family along with the associated motifs could be functionally related to recombination due to their high association with recombination hotspots (Darrier *et al.*, 2017) These experiments provide insight into a potential relationship between transposable elements and recombination.

The development of the high-density genotyping array, SoySNP50K, has made it possible to conduct recombination hotspot mapping in soybean. The objectives of this work were to determine the location of recombination hotspots in soybean, identify genomic features associated with hotspots, and explore if hotspots are stable across two large recombinant inbred line (RIL) populations.

# Materials And Methods

## Plant Material

The two RIL populations used in this study have been previously described by Song et al. (2016). The first RIL population was a cross between 'Williams 82' (*G. max*) (Bernard and Cremeens, 1988) to 'Essex' (*G. max*) (Smith and Camper, 1973). The second population was a cross between Williams 82 and PI479752 (*Glycine soja* Sieb. & Zucc.). The Williams 82 by Essex population (WE) consists of 922 $F_5$ derived RILs developed by single seed decent (SSD) at the University of Tennessee, Knoxville, TN. The Williams 82 x PI479752 population (WP) consists of 1,083 $F_5$-derved RILs and were developed by SSD at USDA-ARS, Beltsville, MD. DNA was extracted from the single $F_5$ plants that derived each RIL.

## Genotyping RILs and Linkage Map Construction

Genotyping of the two RIL populations, and their linkage map construction has been previously described by Song *et al.* (2016). In brief, the WP and WE populations were genotyped with the SoySNP50K Beadchip (Song *et al.*, 2013). There are 11,922 polymorphic SNPs for the WE population and 21,478 polymorphic SNPs for the WP population (Song *et al.*, 2016). Since DNA was extracted from single plants, genotype calls produced distinct clusters for the two homozygous and the heterozygous alleles. This allowed for highly accurate calling of each allele. Quality control steps used were the removal of markers with >10% missing data, markers with significant segregation distortion ($p<0.01$), and only retaining one marker if multiple markers had identical allele segregation patterns. To construct the linkage maps, MSTMAP software was used (Wu *et*

*al.*, 2008). The distance between polymorphic SNPs was calculated using JoinMap 4.0 (Van Oojien *et al.*, 2006). The pericentromeric and euchromatic regions in both populations were previously defined by Song *et al.* (2016).

**Recombination Estimation and Hotspot Detection**

Recombination hotspots were detected using three different methods; spline model in MareyMap (Rezvoy *et al.*, 2007), a frequency metric (the ratio of the statistical genetic map to physical map distances) (Petes, 1991), and a pedigree/linkage disequilibrium (LD) method using PHASE (Li and Stephens, 2004; Rezvoy *et al.*, 2007). Only MareyMap results are reported here since all three methods had comparable results (data not included). The cubic spline function in MareyMap, MMSpline 3 was used with default settings. A cubic spline consciously interpolates through second derivatives and passes through all the data points. A customized perl script was used to identify peaks (recombination hotspots) within the recombination estimate windows. Uneven marker spacing (particularly in heterochromatic regions) led to right skewed distribution in identifying hotspot size. Therefore, the median and average differ drastically. Therefore, outliers in hotspot size distributions were identified in the data set within euchromatic and heterochromatic regions using the Tukey method (Tukey, 1977). Hotspot distribution values that were more than 1.5 times the interquartile range above the third quartile were removed from the data set.

**Correlations & Motif Discovery**

To test if genomic features are associated with recombination hotspots, statistical tests for correlations were based on binomial logistical regression. The statistical tests for

logistical regression were performed with Students t-test on the covariate effect using the GLM function in R 3.2.2. Soybean transposable elements were downloaded from SoyTE database, www.soybase.org (Du *et al.*, 2010). The MEME suite 5.1.0 software was used to discover motifs associated with nucleotides surrounding and including the recombination hotspots within 200 b.p. upstream (Bailey *et al.*, 2006). For the motif discovery, a 1st order Markov model was selected to look at both nucleotide and dinucleotide repeats across the genome and to search for motifs on both strands (Bailey *et al.*, 2006). Sequence logos for each discovered motif as well as E-values were generated. Additionally, for each generated sequence motif a gene ontology analysis was completed resulting in the top 5 specific predictions (Buske *et al.*, 2010). Also, a SpaMO analysis was performed in conjunction with the MEME suite pipeline, generating reports using the Yeast and Arabidopsis databases (Buske *et al.*, 2010).

## Results

The SoySNP50K Beadchip provided a high density of polymorphic markers in each of the two large RIL populations. The SoySNP50K allowed genotyping of 11,922 high quality SNPs in the WE population and 21,478 high quality SNPs in the WP population. Of the 11,922 markers in the WE population, 9,514 SNPs were located in euchromatic DNA resulting with one polymorphic marker approximately every 47 kb. The remaining 2,408 markers were located in the heterochromatic DNA resulting in an average of one polymorphic marker every 208 kb. The distribution of SNPs in the WP population between euchromatic and heterochromatic regions were similar with 17,955 markers in the euchromatic regions and 3,523 markers in the heterochromatic regions. The higher density of markers gave an average of approximately one marker per 25 kb for euchromatic regions and one marker per 142 kb for heterochromatic regions. This distribution closely resembles the original design of the SoySNP50k which has 76% of its SNPs located in the euchromatic regions and 24% of its SNPs located in heterochromatic regions.

The high density of markers genotyped with the SoySNP50K Beadchip enabled high-resolution identification of 451 recombination hotspots across both populations. The WE population contains 245 hotspots across the genome while the WP population has 206 recombination hotspots across the genome (Supplemental Table 1). Of the 451 hotspots identified, the majority (73%), are located within euchromatic DNA. Despite heterochromatic regions having severely reduced recombination (Schmutz *et al.*, 2010), there were 122 hotspots identified within the heterochromatic regions of the chromosomes. To determine if these hotspots are shared between the two populations, the

locations of the hotspots were also compared between the WE and WP populations. Of the 245 WE hotspots and 206 WP hotspots less than 8% occur in the same genomic locations (Supplemental Figures 1-20). This indicates that hotspot locations might not be widely conserved in bi-parental populations despite both these populations sharing a common parent.

Plotting all 451 hotspots by chromosome revealed recombination rates were highest towards the distal regions of the chromosomes in euchromatic DNA and lowest in the pericentromeric regions near the centromere which are heterochromatic DNA (Figure 2.1). The range of recombination rate within the hotspots for both populations ranged from 0.01-20.82 cM/Mbp. Due to the significant difference in recombination between euchromatic DNA and heterochromatic DNA, the data was split based on chromatin state (euchromatic or heterochromatic regions), and analyzed independently. Hotspot intensity for WE averaged 6.08 cM/Mbp in euchromatic DNA with a similar rate within the WP population at 6.47 cM/Mbp, not significantly different (Tukey HSD; $p$ =0.356). Although, recombination rates were lower in heterochromatic, hotspots were still identified in these highly dense regions. (Supplemental 1-20). One example is the pericentromeric region on Chromosome 19, which contains two hotspot regions ~329 kb apart with an average recombination rate of 16 cM/Mbp (Supplemental 19). Despite this one pericentromeric hotspot having a high recombination rate the overall the pericentromeric recombination rate was much lower, averaging 1.5 cM/Mbp across both populations. The WE population had a recombination rate of 0.89 cM/Mbp for hotspots located in pericentromeric DNA while the WP population had a higher rate of 2.1 cM/Mbp. While hotspots do occur in the highly heterochromatic pericentromeric DNA,

the average strength (cM/Mbp) of the hotspot is reduced by 4.78 cM/Mbp averaged across both populations compared to hotspots that occur in the euchromatic regions of the chromosomes (Tukey HSD <2e-16).

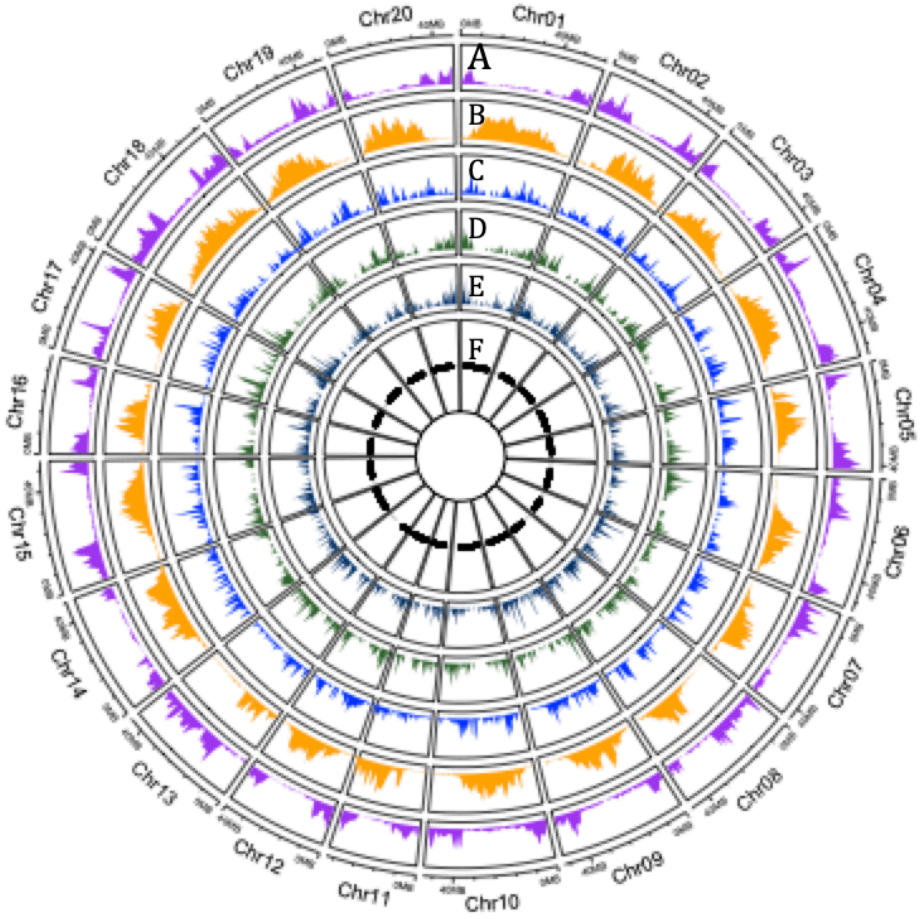



**Figure 2.1 Genome wide recombination rates in two biparental populations and transposable elements associations. The two outer rings represent the average biparental populations by physical distance along chromosomes for Williams 82 x PI479752 and Williams 82 x Essex. A) directly under the physical distance ring purple displays the recombination rates in cM/Mbp (y-axis). B) The next circle in orange represents a density plot of retrotranposon (class I). C) TE Class type II MITE/Tourists element frequencies are in blue. D) TE Class type II the mite stowaway elements frequencies are in green. E) The inner circle represents heterochromatic regions in black and euchromatic regions in white**

The average size of the recombination hotspots was smaller in euchromatic DNA in both populations when compared to heterochromatic regions (Tukey HSD <2e-16). The average hotspot size in euchromatic regions spans 190 kb while in heterochromatic regions the average is 493 kb. Both distributions for hotspot size are skewed to the smaller size (Figure 2.2). This skewedness results in a median hotspot size of 85kb for euchromatic DNA and 258 kb for heterochromatic DNA. In addition, 53% of the euchromatic hotspots are less than 75 kb in length and range from 3.44 kb up to 4,317 kb (Figure 2.2). Heterochromatic hotspots have a range in size from 5.86 kb to 8,966 kb (Figure 2.2).

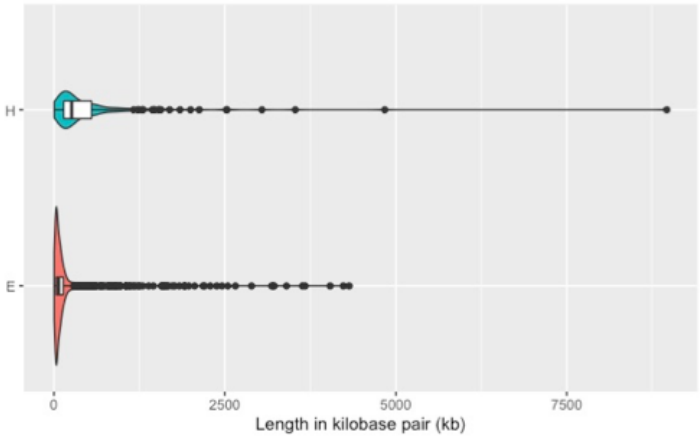



**Figure 2.2 Density of hotspot length in kilobase pairs (kb) by chromatin state, heterochromatic (H) or euchromatic (E) regions. Kernel density estimates of hotspot length are represented by the plot outline. Box plot information is embedded within the violin plot. A vertical solid black line within the box represents the median and whiskers as the black horizontal line.**

With the location and size of each hotspot determined on a high-resolution scale, it was possible to make associations of these hotspots with genome features such as gene regions. The gene regions consist of six categories: 3' untranslated region (UTR), 3'UTR/coding sequence (CDS), 5'UTR, 5'UTR/CDS, CDS and introns. Despite 78% of gene regions being concentrated in chromosome ends, ~28% of the hotspots were located

within a gene region. Of the 28% hotspots in gene regions, half were associated with introns (50.4%). The other half were distributed among the other categories with 27.9% within CDS, 13.2% within 5'UTR, 6.9% within 3'UTR, and 1.6% within 5'UTR/CDS regions.

Additional genomic features such as class I and class II transposable elements, order, super family, and other genomic descriptions from the soybean transposable element database (SoyTEeb) were tested for significant association with recombination hotspots and cold spots. Recombination hotspots were positively associated with a class of type II transposable elements, MITE/Tourist transposable elements (p-value <0.01) (Figure 2.3).

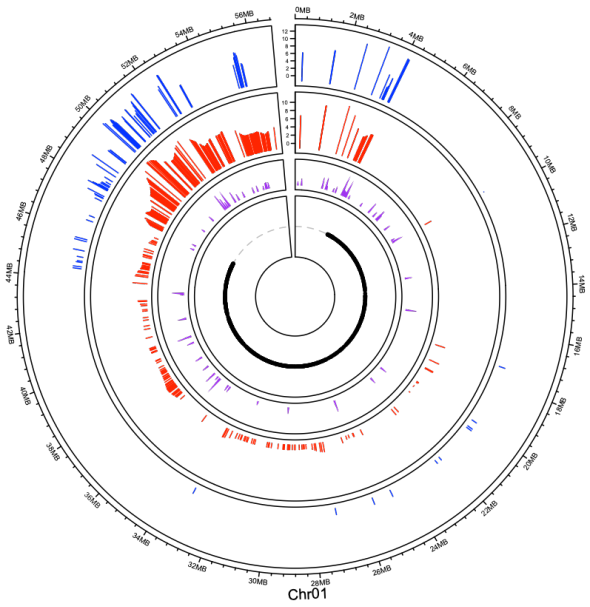



**Figure 2.1 Chromosome 1 recombination rates, MITE/Tourists and heterochromatic regions. Williams 82 x PI479752 (blue) and Williams 82 x Essex (red) recombination rates are displayed in cM/Mbp (y-axis). The type II transposons subclass, MITE/Tourist are shown in purple. The black dots on the inner circle represent heterochromatic regions of the chromosome. Euchromatin regions are displayed by the grey dashed lines**

Cold spots regions were positively associated with the long terminal repeat (LTR) order of class I transposable elements (p-value <0.005).

Recombination associated motifs have been consistently reported in previous hotspot mapping studies (Choi and Henderson, 2015; Darrier *et al.*, 2017; Marand *et al.*, 2017; Mercier *et al.*, 2015b). To determine if motifs are associated with recombination rates in soybean, MEME suite was used to search for over-represented motifs in the 200 bp of sequence flanking the hotspot regions. Two motifs were identified as having an association with hotspots across the genome, a poly-A motif stretch and a CCN repeat (Figure 2.4).

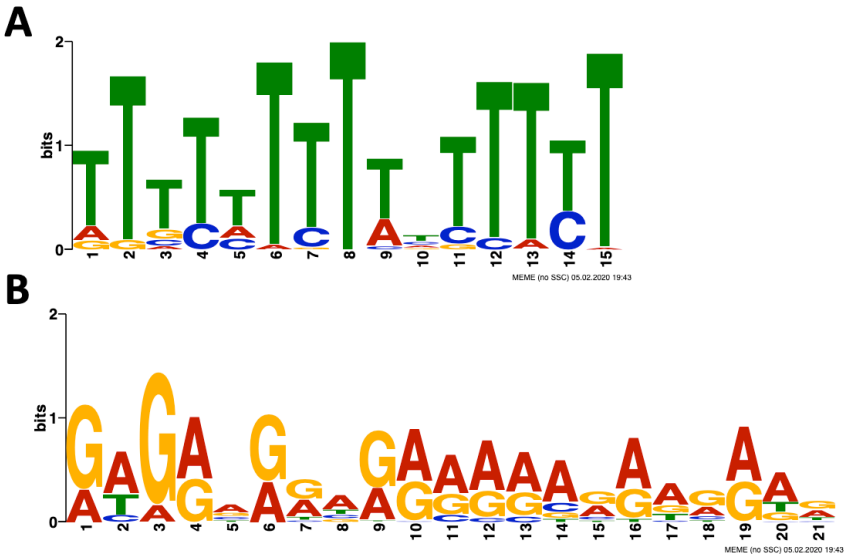



**Figure 2.4 Two most common discovered motifs on recombination hotspots using MEME Suite. The Poly T/A repeat (A) and a second motif which is CCN-like (B) were detected within 200 b.p. of recombination hotspots..**

The A stretch motif is associated with 32% of the hotspots in soybean. These hotspots have an average hotspot recombination rate of 7.24 cm/Mbp with the majority located in euchromatic DNA (81.8%). The hotspot intensity associated with the poly-A motif, 7.24 cM/Mbp (p-value < 2.2e-16) motifs are significantly different compared to hotspot region not associated with motifs. The second motif, CNCCNCCACAACCAANNCANNA is similar to the CCN repeat that has been associated with H3K4me3 modified nucleosome (Shilo *et al.*, 2015). This CCN like motif is associated with 54% of the hotspots in soybean with the majority in euchromatic DNA (78.8%). The average hotspot recombination rate associated with the CCN like motif was significantly higher (7.11 cM/Mbp) but is statistically significant (p-value = 8.633e-05) than the average recombination rate for hotspots in the two populations (WE euchromatic regions average = 6.08 cM/Mbp, WP euchromatic regions average = 6.47 cM/Mbp). Ten percent of the hotspots were found to be associated with both motifs. Having both motifs present in a hotspot did not increase the average recombination rate of the hotspot (p-value = 0.5281). The average rate with both motifs present within a hotspot was 7.44 cM/Mbp.

A spaced motif analysis was conducted and located an overlapping sequence with the Cha4p protein for the hotspot that contains the CCN like motif. This protein encodes DNA binding transcriptional activator and mediates serine/threonine activation of the catabolic L-serine deaminase (CHA1), Zinc-finger protein Zn[2]-Cys[6] fungal type binuclear. Additionally, the gene ontology results illustrate an ~83% association with transcriptional factor activity which is consistent with other species. In humans, Zinc-finger proteins are the largest class of transcription factors, however, they are not as

common in plants. The CCCH Zinc-Finger proteins have been associated with abiotic stress tolerance (Han *et al.*, 2021).

## Discussion

High-density SNP data previously used to create high resolution genetic maps for two large soybean RIL populations was successfully used to map and characterize recombination hotspots in soybean. The recombination hotspot average length of 25 kb is twice as long as the 5-10 kb hotspots lengths reported in the majority of other species (Choi and Henderson, 2015; Darrier *et al.*, 2017). While in horses, the recombination hotspot average size of 23.8 kb is similar to soybean (Beeson *et al.*, 2019). The hotspots detected in soybean heterochromatic DNA had lengths that spanned much larger distances. The 136 kb average length of hotspots in soybean are 128 kb larger than the 8 kb average length reported in Arabidopsis heterochromatic (Drouaud *et al.*, 2013a). This could be due to the hotspots in Arabidopsis being less influenced by DNA methylation (Melamed-Bessudo and Levy, 2012).

The heterochromatic DNA in soybean has been shown to have suppressed recombination (Schmutz *et al.*, 2010). This led to the surprising result that 27% of the recombination hotspots discovered in the two soybean populations were discovered in heterochromatic DNA. The recombination fractions for the hotspots in the heterochromatic regions were lower than for the hotspots in euchromatic regions. However, there were exceptions. Chromosome 19 had a recombination fraction of 16 cM/Mbp, which is comparable to the euchromatic regions. The models for meiotic recombination are not fully understood or the exact role of chromatin in recombination. Recently, it was demonstrated that histone mutation can increase recombination in pericentromeric regions (Underwood *et al.*, 2018). The presence of hotspots in recombination-poor regions provides future possibilities of breaking unfavorable alleles

that tend to be inherited together in the large pericentromeric regions of the soybean genome.

Between the two biparental populations, less than 8% of the hotspots were located in the same genomic location. The lack of shared hotspots could be due to biased gene conversion. The potential effects of biased gene conversion have been noted by the lack of hotspots between humans and chimpanzees, despite a 99% genetic sequence similarity (Coop and Myers, 2007). The authors discovered biased gene conversion to be a strong force in rapid decline of intense hotspots; however, dim hotspots, reduction in cM/Mbp compared to the average recombination for a genomic region, are more likely to be shared (Coop and Myers, 2007). Dim hotspots can arise by the competition between local and ancestral recombination meaning the intensity (cM/Mbp) would be suppressed, thus reducing selection against alleles causing them, as well as some alleles to rise in frequency due to genetic drift (Coop and Myers, 2007). The turnover rate of recombination hotspots was further investigated in $F_1$ hybrids coming from four subspecies of *Mus musculus* with different Prdm9 alleles (Smagulova *et al.*, 2016). They found preferential use of Prdm9 alleles as well as hotspots that become active in hybrids have a greater sequence diversity (Smagulova *et al.*, 2016). Structural variance can influence the location of recombination hotspots along chromosomes. For example, in maize, hotspots were suppressed in regions believed to have an inversion (Rodgers-Melnick *et al.*, 2015). Since structural variance in soybean is relatively low; gene content variation likely doesn't explain the lack of hotspots shared between the two populations (McHale *et al.*, 2012).

Two DNA motifs were discovered in association with recombination hotspots in soybean. One associated DNA motif was a poly-A motif and the second a CCN like repeat. The poly-A motif has been previously reported to be associated with recombination hotspots in humans, wheat, Drosophila, and Arabidopsis (Comeron *et al.*, 2012; Darrier *et al.*, 2017; Myers *et al.*, 2008; Shilo *et al.*, 2015). In wheat, the poly-A motif was found to be associated with 51.6% of the crossovers, which was the highest percentage of the reported motifs (Darrier *et al.*, 2017). Poly-A motifs are thought to influence recombination hotspots due to their association with heterochromatic regions and resilience to nucleosome folding but do not directly prompt recombination (Comeron *et al.*, 2012; Darrier *et al.*, 2017; Myers *et al.*, 2008; Shilo *et al.*, 2015). The second associated motif, a CNN-like repeat, shares similar characteristics to previously identified hotspot motifs in wheat, Arabidopsis, dog, Drosophila (Auton *et al.*, 2013; Comeron *et al.*, 2012; Darrier *et al.*, 2017; Shilo *et al.*, 2015). However, the CCN repeat motif was not associated with recombination hotspot in maize (Rodgers-Melnick *et al.*, 2015). The two motifs did have significantly higher rates of recombination when compared to hotspots without the motifs (p-value < 2.2e-16). This could indicate that the motifs play a role in increasing the probability of a recombination when compared to other hotspots without the motifs. Although the two motifs have been associated with hotspots in other species, the association of recombination intensity as a separate metric was not reported.

Short motifs do not completely explain variation in recombination hotspots; other genomic features such as transposable elements heavily influence recombinant positions. Darrier *et al.* (2017) showed the importance of transposable elements in wheat recombination hotspots by reporting retroransposons (GYPSY and COPIA elements) are

63.7% higher in non-recombinant regions. In general, plant genomes contain a large number of retrotransposons. The soybean genome contains 42.24% retrotransposons (Schmutz *et al.*, 2010). Therefore, it is not surprising that the LTR subclass (41.99% of the genome) is positivity associated with recombination cold spots (p-value <0.005).

DNA transposons (Type II) are less abundant than type I. Soybean contains only 16.5% of type II transposable elements (Schmutz *et al.*, 2010). Terminal inverted repeats make up the largest order within the DNA transposons in soybean and the largest subclass belongs to CACTA (10.16%) (Schmutz *et al.*, 2010). One of the smallest groups is PIF/Harbinger covering only 0.29% and the MITE tourist subclass with 0.33% of the genome (Schmutz *et al.*, 2010). In soybean, recombination hotspots display significant association with and MITE Tourist elements (p-value <0.01). Potato hotspots have been associated with MITE Stowaway elements while rice hotspots have been associated with PIF/Harbinger (Darrier *et al.*, 2017; Marand *et al.*, 2017). Soybean and potato are both palaeopolyploid. Wheat hotspots have been associated with TC1-mariner elements, (Marand *et al.*, 2019). Marand *et al.* (2019) hypothesized an indirect role for Stowaway and PIF/Harbinger elements in promoting long AT-repeat regions over time, creating regions of DNA instability and susceptibility to double strand breaks. The TC1-Mariner elements insertion site is in a TA sequence and are found close to genic regions, which is very similar to the Tourist target repeat region, TAA (Darrier *et al.*, 2017; Zhao *et al.*, 2016). MITE's capabilities to alter sequences near gene regions could attract DNA binding domains of meiotic factors, similar to PRDM9 (Myers *et al.*, 2008).

## Conclusion

Within two large segregating soybean populations, recombination hotspots were identified and characterized. Hotspots in these populations were more commonly found in euchromatic regions with a quarter of them located in heterochromatic DNA. The small percentage of hotspot shared between the populations could allude to a role of gene content variation affecting the location of hotspots. Recombination cold spots were found to be associated with LTR transposable elements. Two common motifs, an A stretch and a CCN, were found in association with recombination hotspots that had a higher rate of recombination than hotspots without the motifs. Uniquely, soybean hotspots are found in areas enriched with Tourist MITEs, which reside in a very small percentage of the genome. While other classifications of MITEs were observed in association with hotspots for potato, rice, and wheat, each one classifies in the TIR order. Further investigation of individual MITE elements such as Tourist, Stowaway, and Tc1-mariner elements near recombination hotspots will be necessary to identify a potential mechanism. Collectively, this study provides insights into the distinctive and shared genomic features of soybean recombination hotspots.

## Acknowledgement

Funding for this project was provided by University of Nebraska Lincoln Agriculture Research Division and the University of Nebraska-Lincoln Agronomy and Horticulture Department. The Holland Computing Center (HCC) at the University of Nebraska provided computational resources.

## References

**Auton A, Li YR, Kidd J, Oliveira K, Nadel J, Holloway JK, Hayward JJ, Cohen PE, Greally JM, Wang J**. 2013. Genetic recombination is targeted towards gene promoter regions in dogs. *Plos Genetics* **9**, e1003984.

**Bailey TL, Williams N, Misleh C, Li WW**. 2006. MEME: discovering and analyzing DNA and protein sequence motifs. *Nucleic Acids Research* **34**, W369-373.

**Barton NH, Charlesworth B**. 1998. Why sex and recombination? *Science* **281**, 1986-1990.

**Baudat F, Imai Y, de Massy B**. 2013. Meiotic recombination in mammals: localization and regulation. *Nature Reviews Genetics* **14**, 794-806.

**Beeson SK, Mickelson JR, McCue ME**. 2019. Exploration of fine-scale recombination rate variation in the domestic horse. *Genome Research* **29**, 1744-1752.

**Bernard R, Cremeens C**. 1988. 2475381. Registration of'Williams 82'soybean. *Crop Science* **28**, 1027-1028.

**Buske FA, Bodén M, Bauer DC, Bailey TL**. 2010. Assigning roles to DNA regulatory motifs using comparative genomics. *Bioinformatics* **26**, 860-866.

**Choi K, Henderson IR**. 2015. Meiotic recombination hotspots - a comparative view. *Plant Journal* **83**, 52-61.

**Choi K, Zhao X, Kelly KA, Venn O, Higgins JD, Yelina NE, Hardcastle TJ, Ziolkowski PA, Copenhaver GP, Franklin FC, McVean G, Henderson IR**. 2013. Arabidopsis meiotic crossover hot spots overlap with H2A.Z nucleosomes at gene promoters. *Nature Genetics* **45**, 1327-1336.

**Choulet F, Alberti A, Theil S, Glover N, Barbe V, Daron J, Pingault L, Sourdille P, Couloux A, Paux E, Leroy P, Mangenot S, Guilhot N, Le Gouis J, Balfourier F, Alaux M, Jamilloux V, Poulain J, Durand C, Bellec A, Gaspin C, Safar J, Dolezel J, Rogers J, Vandepoele K, Aury JM, Mayer K, Berges H, Quesneville H, Wincker P, Feuillet C**. 2014. Structural and functional partitioning of bread wheat chromosome 3B. *Science* **345**, 1249721.

**Comeron JM, Ratnappan R, Bailin S**. 2012. The many landscapes of recombination in Drosophila melanogaster. *Plos Genetics* **8**, e1002905.

**Coop G, Myers SR**. 2007. Live hot, die young: transmission distortion in recombination hotspots. *Plos Genetics* **3**, e35.

**Darrier B, Rimbert H, Balfourier F, Pingault L, Josselin AA, Servin B, Navarro J, Choulet F, Paux E, Sourdille P**. 2017. High-Resolution Mapping of Crossover Events in the Hexaploid Wheat Genome Suggests a Universal Recombination Mechanism. *Genetics* **206**, 1373-1388.

**Drouaud J, Khademian H, Giraut L, Zanni V, Bellalou S, Henderson IR, Falque M, Mezard C**. 2013a. Contrasted Patterns of Crossover and Non-crossover at Arabidopsis thaliana Meiotic Recombination Hotspots. *Plos Genetics* **9**.

**Drouaud J, Khademian H, Giraut L, Zanni V, Bellalou S, Henderson IR, Falque M, Mezard C**. 2013b. Contrasted patterns of crossover and non-crossover at Arabidopsis thaliana meiotic recombination hotspots. *Plos Genetics* **9**, e1003922.

**Du J, Grant D, Tian Z, Nelson RT, Zhu L, Shoemaker RC, Ma J**. 2010. SoyTEdb: a comprehensive database of transposable elements in the soybean genome. *Bmc Genomics* **11**, 113.

**Fu H, Zheng Z, Dooner HK**. 2002. Recombination rates between adjacent genic and retrotransposon regions in maize vary by 2 orders of magnitude. *Proc Natl Acad Sci U S A* **99**, 1082-1087.
**Gore MA, Chia JM, Elshire RJ, Sun Q, Ersoz ES, Hurwitz BL, Peiffer JA, McMullen MD, Grills GS, Ross-Ibarra J, Ware DH, Buckler ES**. 2009. A first-generation haplotype map of maize. *Science* **326**, 1115-1117.
**Han G, Qiao Z, Li Y, Wang C, Wang B**. 2021. The Roles of CCCH Zinc-Finger Proteins in Plant Abiotic Stress Tolerance. *International Journal of Molecular Sciences* **22**, 8327.
**Lambing C, Franklin FCH, Wang CJR**. 2017. Understanding and Manipulating Meiotic Recombination in Plants. *Plant Physiology* **173**, 1530-1542.
**Li N, Stephens M**. 2004. Modeling linkage disequilibrium and identifying recombination hotspots using single-nucleotide polymorphism data (vol 165, pg 2213, 2004). *Genetics* **167**, 1039-1039.
**Li X, Li L, Yan J**. 2015. Dissecting meiotic recombination based on tetrad analysis by single-microspore sequencing in maize. *Nat Commun* **6**, 6648.
**Liu S, Yeh CT, Ji T, Ying K, Wu H, Tang HM, Fu Y, Nettleton D, Schnable PS**. 2009. Mu transposon insertion sites and meiotic recombination events co-localize with epigenetic marks for open chromatin across the maize genome. *PLoS Genet* **5**, e1000733.
**Marand AP, Jansky SH, Zhao H, Leisner CP, Zhu X, Zeng Z, Crisovan E, Newton L, Hamernik AJ, Veilleux RE, Buell CR, Jiang J**. 2017. Meiotic crossovers are associated with open chromatin and enriched with Stowaway transposons in potato. *Genome Biology* **18**, 203.
**Marand AP, 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**, 645-662.
**McHale LK, Haun WJ, Xu WW, Bhaskar PB, Anderson JE, Hyten DL, Gerhardt DJ, Jeddeloh JA, Stupar RM**. 2012. Structural Variants in the Soybean Genome Localize to Clusters of Biotic Stress-Response Genes. *Plant Physiology* **159**, 1295-1308.
**Melamed-Bessudo C, Levy AA**. 2012. Deficiency in DNA methylation increases meiotic crossover rates in euchromatic but not in heterochromatic regions in Arabidopsis. *Proceedings of the National Academy of Sciences* **109**, E981-E988.
**Mercier R, Mezard C, Jenczewski E, Macaisne N, Grelon M**. 2015a. The molecular biology of meiosis in plants. *Annual Review of Plant Biology, Vol 66* **66**, 297-327.
**Mercier R, Mezard C, Jenczewski E, Macaisne N, Grelon M**. 2015b. The Molecular Biology of Meiosis in Plants. *Annual Review of Plant Biology, Vol 66* **66**, 297-327.
**Myers S, Freeman C, Auton A, Donnelly P, McVean G**. 2008. A common sequence motif associated with recombination hot spots and genome instability in humans. *Nature Genetics* **40**, 1124-1129.
**Paterson AH, Bowers JE, Bruggmann R, Dubchak I, Grimwood J, Gundlach H, Haberer G, Hellsten U, Mitros T, Poliakov A, Schmutz J, Spannagl M, Tang H, Wang X, Wicker T, Bharti AK, Chapman J, Feltus FA, Gowik U, Grigoriev IV, Lyons E, Maher CA, Martis M, Narechania A, Otillar RP, Penning BW, Salamov AA, Wang Y, Zhang L, Carpita NC, Freeling M, Gingle AR, Hash CT, Keller B, Klein P, Kresovich S, McCann MC, Ming R, Peterson DG, Mehboob ur R, Ware D,**

**Westhoff P, Mayer KF, Messing J, Rokhsar DS**. 2009. The Sorghum bicolor genome and the diversification of grasses. *Nature* **457**, 551-556.
**Patterson GI, Kubo KM, Shroyer T, Chandler VL**. 1995. Sequences required for paramutation of the maize b gene map to a region containing the promoter and upstream sequences. *Genetics* **140**, 1389-1406.
**Petes TD**. 1991. The Molecular and Cellular Biology of the Yeast Saccharomyces: Genome Dynamics, Protein Synthesis, and Energetics Vol. 1 edited by JR Broach, JR Pringle and EW Jones (Cold Spring Harbor Laboratory Press, 1991). In: Hoekstra MF, ed: Elsevier Current Trends.
**Rezvoy C, Charif D, Gueguen L, Marais GA**. 2007. MareyMap: an R-based tool with graphical interface for estimating recombination rates. *Bioinformatics* **23**, 2188-2189.
**Rodgers-Melnick E, Bradbury PJ, Elshire RJ, Glaubitz JC, Acharya CB, Mitchell SE, Li CH, Li YX, Buckler ES**. 2015. Recombination in diverse maize is stable, predictable, and associated with genetic load. *Proceedings of the National Academy of Sciences of the United States of America* **112**, 3823-3828.
**Saintenac C, Falque M, Martin OC, Paux E, Feuillet C, Sourdille P**. 2009. Detailed recombination studies along chromosome 3B provide new insights on crossover distribution in wheat (Triticum aestivum L.). *Genetics* **181**, 393-403.
**Salome PA, Bomblies K, Fitz J, Laitinen RA, Warthmann N, Yant L, Weigel D**. 2012. The recombination landscape in Arabidopsis thaliana F2 populations. *Heredity (Edinb)* **108**, 447-455.
**Schmutz J, Cannon SB, Schlueter J, Ma J, Mitros T, Nelson W, Hyten DL, Song Q, Thelen JJ, Cheng J, Xu D, Hellsten U, May GD, Yu Y, Sakurai T, Umezawa T, Bhattacharyya MK, Sandhu D, Valliyodan B, Lindquist E, Peto M, Grant D, Shu S, Goodstein D, Barry K, Futrell-Griggs M, Abernathy B, Du J, Tian Z, Zhu L, Gill N, Joshi T, Libault M, Sethuraman A, Zhang XC, Shinozaki K, Nguyen HT, Wing RA, Cregan P, Specht J, Grimwood J, Rokhsar D, Stacey G, Shoemaker RC, Jackson SA**. 2010. Genome sequence of the palaeopolyploid soybean. *Nature* **463**, 178-183.
**Shilo S, Melamed-Bessudo C, Dorone Y, Barkai N, Levy AA**. 2015. DNA Crossover Motifs Associated with Epigenetic Modifications Delineate Open Chromatin Regions in Arabidopsis. *Plant Cell* **27**, 2427-2436.
**Smagulova F, Brick K, Pu Y, Camerini-Otero RD, Petukhova GV**. 2016. The evolutionary turnover of recombination hot spots contributes to speciation in mice. *Genes & Development* **30**, 266-280.
**Smith T, Camper H**. 1973. Registration of Essex soybean (Reg. no. 97). *Crop Science* **13**, 495-495.
**Song Q, Hyten DL, Jia G, Quigley CV, Fickus EW, Nelson RL, Cregan PB**. 2013. Development and evaluation of SoySNP50K, a high-density genotyping array for soybean. *Plos One* **8**, e54985.
**Song Q, Jenkins J, Jia G, Hyten DL, Pantalone V, Jackson SA, Schmutz J, Cregan PB**. 2016. Construction of high resolution genetic linkage maps to improve the soybean genome sequence assembly Glyma1.01. *Bmc Genomics* **17**, 33.
**Tenaillon MI, Sawkins MC, Anderson LK, Stack SM, Doebley J, Gaut BS**. 2002. Patterns of diversity and recombination along chromosome 1 of maize (Zea mays ssp. mays L.). *Genetics* **162**, 1401-1413.
**Tukey JW**. 1977. *Exploratory data analysis*: Reading, MA.

**Underwood CJ, Choi K, Lambing C, Zhao X, Serra H, Borges F, Simorowski J, Ernst E, Jacob Y, Henderson IR**. 2018. Epigenetic activation of meiotic recombination in Arabidopsis centromeres by disruption of H3K9me2 and non-CG DNA methylation.
**Van Oojien J, Van Ooijen J, Van Ooijen J, Van Ooijen J, Kyazmay B, Ooijen J, Riel J, van OOIJEN J, Camp N, van't Verlaat J**. 2006. JoinMap 4, software for the calculation of genetic linkage maps in experimental population.
**Wei F, Zhang J, Zhou S, He R, Schaeffer M, Collura K, Kudrna D, Faga BP, Wissotski M, Golser W, Rock SM, Graves TA, Fulton RS, Coe E, Schnable PS, Schwartz DC, Ware D, Clifton SW, Wilson RK, Wing RA**. 2009. The physical and genetic framework of the maize B73 genome. *PLoS Genet* **5**, e1000715.
**Wijnker E, de Jong H**. 2008. Managing meiotic recombination in plant breeding. *Trends in Plant Science* **13**, 640-646.
**Wu Y, Bhat PR, Close TJ, Lonardi S**. 2008. Efficient and accurate construction of genetic linkage maps from the minimum spanning tree of a graph. *Plos Genetics* **4**, e1000212.
**Yao H, Schnable PS**. 2005. Cis-effects on meiotic recombination across distinct a1-sh2 intervals in a common Zea genetic background. *Genetics* **170**, 1929-1944.
**Yelina NE, Choi K, Chelysheva L, Macaulay M, de Snoo B, Wijnker E, Miller N, Drouaud J, Grelon M, Copenhaver GP, Mezard C, Kelly KA, Henderson IR**. 2012. Epigenetic Remodeling of Meiotic Crossover Frequency in Arabidopsis thaliana DNA Methyltransferase Mutants. *Plos Genetics* **8**.
**Zhao D, Ferguson AA, Jiang N**. 2016. What makes up plant genomes: The vanishing line between transposable elements and genes. *Biochimica et Biophysica Acta (BBA)-Gene Regulatory Mechanisms* **1859**, 366-380.

# CHAPTER THREE

## HISTORICAL RECOMBINATION HOTSPOTS IN SOYBEAN [*GLYCINE MAX* (L.) MERR]


### Abstract

Recombination is the primary mechanism for introducing variation in a population and allows plant breeders to identify novel allelic combinations for agronomic improvement. Soybean [*Glycine max* (L.) Merr.] has undergone several genetic bottlenecks that have affected its genetic diversity and altered allele frequencies. To understand the effect of genetic bottlenecks on recombination hotspots in soybean three large populations were used to characterize historical recombination hotspots. The three populations were selected from the USDA soybean germplasm collection to investigate historical recombination hotspots that span the historical genetic bottlenecks of domestication and the introduction of soybean into North America (N.A.) followed by modern plant breeding. Approximately, 13,000 hotspots were found among each population, wild soybean, landraces, and N.A. elite. A reduction of 10 cM/Mbp in the average recombination rate was observed between wild soybean and landraces and could be a result of the domestication bottleneck. Previously identified genomic motifs were found among all the historical populations suggesting a common mechanism. This characterization of historical recombination hotspots in soybean helps to further understand the relationship between hotspots and genetic bottlenecks.

## Introduction

Meiotic recombination is the basic genetic mechanism that generates new allelic combinations in crops. It is tightly regulated and occurs at low frequency, averaging one crossover per chromosome per generation. The locations of where recombination occurs along the chromosome are not random. It is severely repressed in heterochromatic regions and does not occur randomly in euchromatic regions. Recombination is often clustered in small genomic regions that are termed recombination hotspots. The older a population, the more chance it has had to accumulate historical recombination. This historical recombination can be used to map recombination hotspots in a population and learn the genomic characteristics that define them.

Mapping ancestral recombination rates has provided insight into the mechanism of recombination hotspots. Human and yeast studies have shown hotspots to have a high turnover rate along with an association of a common mutation in the PRDM9 gene, (PR/SET Domain 9) a zinc finger protein with histone methyltransferace activity (Myers *et al.*, 2008). It is commonly believed that the high mutation rates in the zinc finger domains of PRDM9 explain the co-evolution of hotspots within specific lineages (Thomas *et al.*, 2009). While this motif has been frequently associated with hotspots, there is a lack of hotspot concordance between humans and chimpanzees and the PRDM9 homolog has not been associated with hotspots in plants (Auton and McVean, 2012). While not associated with PRDM9, some motifs associated with recombination hotspots have been associated with zinc finger domains in plants suggesting a similar mechanism. In addition, across plant species including potato, rice, tomato, and wheat there has been repetitive motifs and transposable elements associated with recombination hotspots

(Darrier *et al.*, 2017; Fuentes *et al.*, 2022; Marand *et al.*, 2017; Marand *et al.*, 2019; Shilo *et al.*, 2015).

Mapping recombination hotspots in plants to discover their genomic features can also be accomplished by looking at populations that have undergone domestication. An increased rate of historical recombination is observed with intense selection during domestication (Moyers *et al.*, 2018; Ross-Ibarra, 2004). In tomato, while the general landscape of recombination was conserved, there were significantly higher recombination rates at distal chromosome regions (Fuentes *et al.*, 2022). In early domesticated tomato varieties there was an increase in local recombination when calculating recombination in shorter, 50 kilobase windows across the genome compared to 1 megabase windows (Fuentes *et al.*, 2022). Barley displayed a similar conservation during domestication and resulted in increased local recombination that was shaped to environmental conditions (Dreissig *et al.*, 2019). For cocoa populations, wild populations had lower rates of recombination than domesticated populations (Schwarzkopf *et al.*, 2020).

In soybean, [*Glycine max* (L.) Merr.], current elite cultivars in North America (N.A.) have a genetic landscape that has been shaped by three major historical events, i) domestication, ii) introduction into N. America via a small number of plant introductions, and iii) modern breeder selection to develop the current elite cultivars. Linkage disequilibrium (LD) and haplotype structure analysis between wild populations, landrace, and N.A. elite cultivars discovered haplotype blocks of varying sizes indicating uneven historical recombination in the USDA soybean germplasm collection (Song *et al.*, 2015). In the N.A. elite cultivars, the genetic bottleneck not only affected diversity but also the haplotype block size. Larger haplotype block sizes and more extensive LD is present in

the N.A. elite as compared to the landraces and from landraces compared to wild populations (Song *et al.*, 2015). Haplotype block sizes can be used as a measurement of historical recombination; a larger block indicates that longer regions in the genome have not been subjected to historical recombination. On average, wild haplotype block sizes were 10.7 kb, landraces were 39.6 kb, and N.A. elite block sizes were 79.6 kb, in euchromatic regions. In euchromatic regions, this result indicates there have been more historical recombination in the wild populations compared to the landraces and more in the landraces compared to the N.A. elite. While this range in haplotype block size was observed in euchromatic DNA, it was not as dramatic in heterochromatic regions. Heterochromatic regions average blocks sizes in the populations ranged from 511 kb to 675 kb for the wild populations, landraces, and N.A. elite (Song *et al.*, 2015).

The United States Department of Agriculture (USDA) Soybean Germplasm collection contains more than 23,000 diverse accessions including wild populations, landraces, and N.A. elite. It is a resource that can be used for the mapping and comparison of recombinant hotspots. The objectives of this study were to i) map historical recombination hotspot in wild populations, landraces, and N.A. elite and ii) identify any genomic association with hotspot regions.

# Materials And Methods

## Plant Material and Genotyping

Three populations contained within the USDA soybean germplasm collection were used for this study. The populations consisted of 806 wild *G. soja* accessions, 5396 landrace accessions, and 562 N.A. elite soybean cultivars. All accessions have been previously genotyped with the SoySNP50K BeadChip as described by Song *et al.* (2015). The vcf and bcf files were downloaded from www.soybase.org. Single nucleotide polymorphisms (SNP) were filtered based on rate of missing and heterozygous allele calls, greater then 0.1 among the soybean and wild soybean accessions were eliminated (Song *et al.*, 2015).

## Recombination estimation and Hotspot detection

Recombination hotspots were estimated using a topological data analysis (TDA) (Humphreys *et al.*, 2019). This approach was selected due to its computational efficiency and connection to coalescent models (Humphreys *et al.*, 2019). The relative accuracy of TREE is comparable to other coalescent models such as LDhelmet. The correlation between absolute predicted values of TREE and LDhelmet were positive and with p-values<0.0001 (Humphreys *et al.*, 2019). The input files were created by converting the VCF file to FASTA using the VCF-kit phylo fasta function. Recombination rates were calculated in 25 SNP windows on each chromosome for the subset data groups (wild populations, landraces, and N.A. elite) with 42,509 markers. To identify peak regions (hotspots) a customized perl script was used to identity peaks with the recombination estimates. Uneven marker spacing (particularly in heterochromatic regions) led to severe

right skewed distribution in calculated hotspot size. Therefore, hotspot size outliers were identified in the data set within euchromatic and heterochromatic regions using the Tukey method (Tukey, 1977). Hotspot size values that were more than 1.5 times the interquartile range and above the third quartile were removed from the data set. The hotspot regions were compared by physical location using SNP IDs to previously reported hotspots within the bi-parental populations using the common parent Williams 82 crossed to Essex and PI479752. To determine significant differences among population hotspots a Tukey honest significant difference (HSD) test was applied (Tukey, 1977).

**Correlations & Motif Discovery**

Statistical tests for correlations are based on logistical regression with Students t-test on the covariate effect using the GLM function in R 3.2.2. Soybean transposable elements were downloaded from SoyTE database, www.soybase.org (Du *et al.*, 2010). The MEME suite 5.1.0 software was used to discover motifs associated with nucleotides within 200 bp upstream or downstream surrounding the recombination hotspots (Bailey *et al.*, 2006). The 1$^{st}$ order Markov model was selected for motif discovery to look at both nucleotide and dinucleotide repeats across the genome and to search for motifs on both strands (Bailey *et al.*, 2006). Sequence logos for each discovered motifs as well as E-values were generated.

# Results

## Comparing recombination rates among populations

The genetic landscape in soybean is attributed to three major events; domestication, introduction into North America by a small number of plant introductions, and selection on elite cultivars. The three soybean populations; wild, landraces, and the N.A. elite represent key historical genetic milestones. Landraces were domesticated from wild populations and this domestication bottleneck appears to have reduced the average recombination rate of wild populations (37.68 cM/Mbp) by 10 cM/Mbp when compared to the landraces (27.7 cM/Mbp), all populations are significantly different (Tukey HSD; $p < 2.2e\text{-}16$) Figure 3.1. The bottleneck created from the development of the elite cultivars from the landraces appears to have further reduced the recombination rate by another 15.39 cM/Mbp (Figure 3.1). Moreover, the distribution of recombination rate highlights the relationship between the populations.

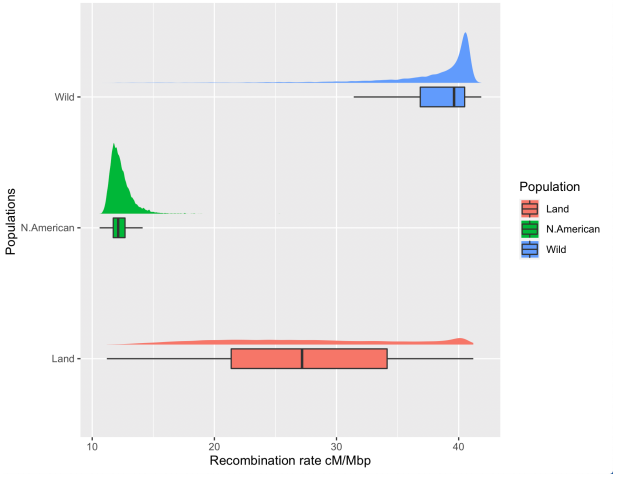



**Figure 3.1 Recombination Hotspot intensity distribution in centimorgans (cM) over Megabase pairs (Mbp) by soybean population, wild populations (blue), landraces (red) and N.A. elite (green). Box plot information is adjacent to the recombination frequency distributions.**

The wild population displaying an upper ward bound peak while a bimodal distribution is observed in the landraces. The domestication bottleneck from wild populations and the selection intensity on the N.A. elite permit the bimodal observation.

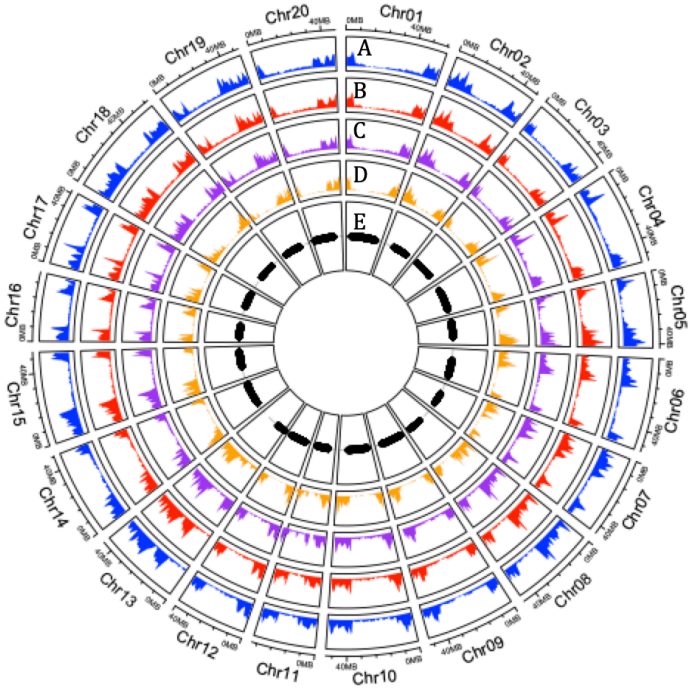



**Figure 3.3. Genome wide recombination rates in the USDA Germplasm Collection divided by population, A) wild populations (blue), B) landraces (red) and C) N. A. elite (purple). D) Gene density is displayed in orange. E) The inner circle represents heterochromatic in black and euchromatin in white**

The N.A. elite cultivars have a lower bound distribution centered on the lowest recombination rate among the populations. The number of recombination hotspots is similar for all populations ranging from 13,019 to 13,808. While domestication and selection have affected recombination intensity, the number of historical recombination hotspots is unaffected by the genetic bottlenecks.

## Comparing recombination hotspot locations and size between populations

The pattern of recombination rate varied across the genome displaying peaks and valleys (Figure 2.2 and Sup 21-40). In general, hotspots are consistent across heterochromatic regions state with the majority occurring in the euchromatic regions, averaging 80.8% of all hotspots across all populations (Table 3.1).

**Table 3.1 Summary of the number of hotspot length by soybean population, landraces, wild populations, and North America (N.A.) elite. The number of lines in each population are represented in the first column. The last four columns contain information on the number and percentage of hotspots in heterochromatic and euchromatic regions**

| # of lines | Population | Total # of Hotspots | Total # HS shared with BiParental | Average HS Size in kb | HS in Heterochromatin | HS in Euchromatin | % Het | %E |
|---|---|---|---|---|---|---|---|---|
| 806 | Wild | 13607 | 129 | 36.1 | 2634 | 10973 | 0.1935 | 0.8064 |
| 5396 | Landraces | 13808 | 143 | 45.5 | 2631 | 11177 | 0.1905 | 0.8094 |
| 562 | N.American | 13019 | 131 | 21.9 | 2502 | 10517 | 0.1921 | 0.8078 |

With fifty-seven percent of the soybean genome occurs in repeat rich, heterochromatic regions, historical recombination hotspots were only observed in 19.2% of heterochromatic regions. Since hotspot marker density averaged 110.5 markers per Mb for euchromatic regions and 20.5 markers per Mb in heterochromatic regions, fine scale hotspots can be detected in both categories of chromatin. When examining hotspot regions within heterochromatic DNA, the peaks in one population were typically present in another. (Supplemental Figures 21-40)

The high density of markers provides several anchor points to calculate the size of recombination hotspots. The N.A. elite hotspots were on average 10.8 kb smaller than the landraces (significant Tukey HSD $p < 0.007$) but only 1.3kb smaller than the wild

populations (not significant Tukey HSD $p$ =0.93) (Table 1). The landrace hotspots were also significantly smaller than the wild populations (Tukey HSD $p < 0.01$). The average and median hotspot values vary drastically due to the skewed distribution of hotspot size.

**Identifying DNA sequence motifs and genomic features associated with hotspots**

In plant species, two common motifs have been identified within recombination hotspots, a poly - motif and a CCN repeat motif (Darrier *et al.*, 2017; McConaughy, 2022; Shilo *et al.*, 2015). In this study, among the 1,381 shared recombination hotspots between the wild populations, landrace, and elite populations, a poly-A motif and CCN repeat were discovered. In the wild populations, landraces, and N.A. elite cultivar populations, 47.6% (658 out of 1,381) were found in a poly-A repeat sequence and 37.4% (516 out of 1,381) in the CCN-like repeat (Figure 3.3). The CCN-like repeat aligned with a known Zinc finger super family protein motif in *Arabidopsis*.

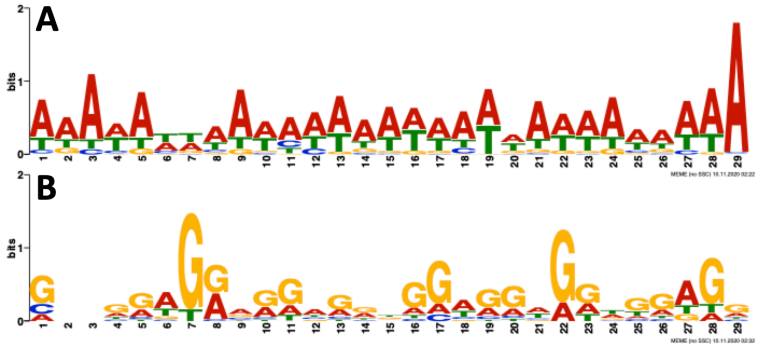



**Figure 3.3. Two most common discovered motifs on recombination hotspots using MEME Suite. The Poly-T/A repeat (A) found in 47.6% of ancestral recombination hotspots and a second motif which is CCN like repeat (B) occurred in 37.4% of hotspots were detected within 200 b.p. of recombination hotspots**

Individual motif analysis for the wild populations, landraces, and N.A. elite cultivars did not reveal common associations that were found across populations,

averaging less than 11% in the hotspots per population. In the wild populations, the poly-A motif was observed in 10.56% (1,437 out of 13,607) of the hotspots.

Hotspots have been previously associated with gene regions and transposable elements (Choi *et al.*, 2013; Darrier *et al.*, 2017; McConaughy, 2022; Shilo *et al.*, 2015). In the wild populations, landraces, and N.A. elite cultivars, an overwhelming majority (78.6%) of hotspots were located near Type I transposable elements (Figure 3.4A). Of the different elements of Type I, the hotspots were mostly associated with LTR, which are the most abundant of the Type I elements (Figure 3.4C). Gypsy elements represent over half of the hotspot regions for super families with the most prevalent element being the LTR-intact elements.

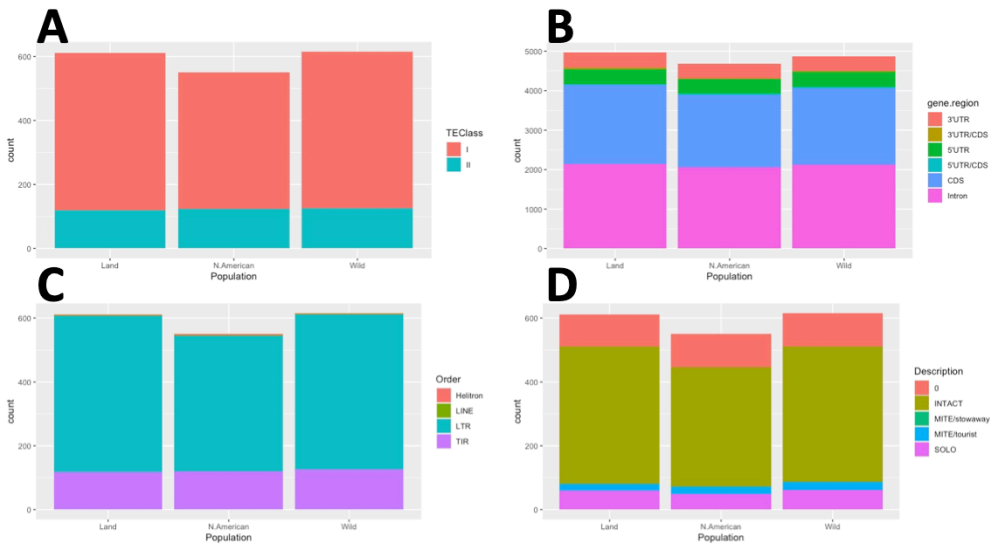



**Figure 3.4. Counts of genomic elements found with Hotspots by population, landraces, wild, and N. A. elite. Genomic information is displayed in four charts, TE class information (A), gene region (B), gene order (C), and gene descriptors (D)**

With a small percentage of recombination hotspots associated with Type II Transposable elements (21.4%), two elements were notable, MITE/Tourist and Stowaway element (Figure 3.4D). Of the multifaceted logistic regression analysis, Type I transposable elements, MITE Stowaway elements presented the strongest results approaching significance. This is notable given the very small percentage of hotspot regions that occur in Type II transposable elements. With the majority of hotspots found in Type I transposable elements there is not statistical evidence to support that these regions are more enriched for hotspots compared to the rest of the soybean genome. MITE Stowaway and other similar elements have been associated with hotspots in biparental soybean populations and in potato (Marand *et al.*, 2017; McConaughy, 2022).

## Discussion

The wild populations, landraces, and N.A. elite cultivars are key populations contained within the USDA Soybean Germplasm Collection that have been extensively genotyped and can be used to identify historical recombination. These populations provide an opportunity to study historical recombination hotspots through soybean domestication, introduction into North America, and selection on elite cultivars. While the average recombination rate was significantly affected by these events, the total number of recombination hotspots has not been affected. One hypothesis is that bottlenecks would reduce nucleotide diversity in soybean and this reduction in diversity reduces the ability to detect historical recombination in the resulting populations. However, greater than 72% of the genetic diversity was retained across populations (Hyten *et al.*, 2006). Similarly, the bottlenecks did not reduce the number of hotspots that could be detected in each population.

In soybean, recombination rates have been reported in two large biparental populations, one population within North American germplasm and another crossing North American material to wild accession, PI479752. The two populations were $F_5$ derived recombinant inbred lines. Comparing the association of recombination hotspot locations between soybean biparental populations (McConaughy, 2022) and ancestral populations provides the opportunity to identify global recombination hotspots (across historical and recent population recombination). Soybean recombination hotspot average size in the biparental populations (25kb) are comparable to the USDA germplasm collection populations, wild (36 kb), landraces (45.5 kb), and N.A. elite (21.9kb).

Both soybean historical and recent population recombination is larger on average compared to other crops that have been reported to average 5-10 kb (Kauppi *et al.*, 2004; Mercier *et al.*, 2015; Mezard, 2006). A fine scale recombination mapping study in horse did identify hotspots closer in size to soybean with an average hotspot size of 23.8 Mb (Beeson *et al.*, 2019). Time has an affect on hotspot size. The wild populations, landraces, and N.A. elite have had more time to allow for more recombination events across the hotspot to be sampled in comparison to a recombinant inbred line populations. This may indicate that recombination hotspots span larger regions than the initial reports.

Two DNA motifs (poly-A and CCN repeat) were discovered to be associated with recombination hotspots in biparental populations and historical populations from the germplasm collection in soybean. These motifs have also been reported in humans, wheat, *Drosophila*, and *Arabidopsis* (Comeron *et al.*, 2012; Darrier *et al.*, 2017; Myers *et al.*, 2008; Shilo *et al.*, 2015). Unique to soybean recombination hotspots, MITE Stowaway elements were identified in more than one population from the USDA Germplasm collection and in the biparental populations. The enrichment for MITE Stowaway elements in soybean recombination hotspots provides more emphasis on the importance of transposable elements affecting the recombination landscape.

Stowaway elements have been hypothesized to have an indirect role in promoting long AT repeat regions over time, which may create genome instability and thus, susceptibility to double stranded breaks (Marand *et al.*, 2019). The capability of MITE to alter sequences near gene regions could attract DNA binding domains of meiotic factors, similar to PRDM9 (Myers *et al.*, 2008). In humans, simple tandem repeats have been accountable for nearly 50 diseases in humans (Khristich and Mirkin, 2020). The same

repeat unit can be responsible for different diseases in human thus displaying the diverse impact of a repeat classified as repeats expansion diseases (REDs). One unique feature of REDs is the number of inherited repeats is positively correlated with disease severity (Khristich and Mirkin, 2020). Although a different species, a similar trend is displayed with recombination hotspots and repeat regions.

Soybean Chromosome 20 contains a major pleiotropic seed protein and oil quantitative trait locus that has been well documented including a haplotype analysis showcasing linkage disequilibrium blocks starting at Glyma20g19620 and ending at Glyma20g22650 (Bandillo *et al.*, 2015). Interestingly, the recombination hotspot corresponding between haplotype block 3 (987 kb) and haplotype block 4 is predominately displayed in wild and landraces populations. In the N.A. elite no hotspot exist across the entire QTL that spans 30 Mb. A transposable element insertion within the CCT domain protein accounts for the seed protein alleles (Fliege *et al.*, 2022). We hypothesized the transposable element insertion happened during U.S. plant breeding on the Landraces. If a breeding program would like to add variation to this region, our results suggest less linkage drag around the haplotype in the landraces and wild populations versus the N.A. elite. Therefore, to change the allele frequency it would be best to use a landrace or wild accessions than a N.A. elite. This example demonstrates one application of using the historical hotspot locations. The historical hotspot locations will accelerate the introgression of traits, genomic editing, and guide the creation of novel allelic variation in soybean.

## References

**Auton A, McVean G**. 2012. Estimating Recombination Rates from Genetic Variation in Humans. *Evolutionary Genomics: Statistical and Computational Methods, Vol 2* **856**, 217-237.
**Bailey TL, Williams N, Misleh C, Li WW**. 2006. MEME: discovering and analyzing DNA and protein sequence motifs. *Nucleic Acids Research* **34**, W369-373.
**Bandillo N, Jarquin D, Song Q, Nelson R, Cregan P, Specht J, Lorenz A**. 2015. A population structure and genome‐wide association analysis on the USDA soybean germplasm collection. *The Plant Genome* **8**, 1-13.
**Beeson SK, Mickelson JR, McCue ME**. 2019. Exploration of fine-scale recombination rate variation in the domestic horse. *Genome Research* **29**, 1744-1752.
**Choi K, Zhao X, Kelly KA, Venn O, Higgins JD, Yelina NE, Hardcastle TJ, Ziolkowski PA, Copenhaver GP, Franklin FC, McVean G, Henderson IR**. 2013. Arabidopsis meiotic crossover hot spots overlap with H2A.Z nucleosomes at gene promoters. *Nature Genetics* **45**, 1327-1336.
**Comeron JM, Ratnappan R, Bailin S**. 2012. The many landscapes of recombination in Drosophila melanogaster. *Plos Genetics* **8**, e1002905.
**Darrier B, Rimbert H, Balfourier F, Pingault L, Josselin AA, Servin B, Navarro J, Choulet F, Paux E, Sourdille P**. 2017. High-Resolution Mapping of Crossover Events in the Hexaploid Wheat Genome Suggests a Universal Recombination Mechanism. *Genetics* **206**, 1373-1388.
**Dreissig S, Mascher M, Heckmann S**. 2019. Variation in recombination rate is shaped by domestication and environmental conditions in barley. *Molecular Biology and Evolution* **36**, 2029-2039.
**Du J, Grant D, Tian Z, Nelson RT, Zhu L, Shoemaker RC, Ma J**. 2010. SoyTEdb: a comprehensive database of transposable elements in the soybean genome. *Bmc Genomics* **11**, 113.
**Fliege CE, Ward RA, Vogel P, Nguyen H, Quach T, Guo M, Viana JPG, Dos Santos LB, Specht JE, Clemente TE**. 2022. Fine mapping and cloning of the major seed protein quantitative trait loci on soybean chromosome 20. *The Plant Journal* **110**, 114-128.
**Fuentes RR, de Ridder D, van Dijk AD, Peters SA**. 2022. Domestication shapes recombination patterns in tomato. *Molecular Biology and Evolution* **39**, msab287.
**Humphreys DP, McGuirl MR, Miyagi M, Blumberg AJ**. 2019. Fast Estimation of Recombination Rates Using Topological Data Analysis. *Genetics* **211**, 1191-1204.
**Hyten DL, Song Q, Zhu Y, Choi IY, Nelson RL, Costa JM, Specht JE, Shoemaker RC, Cregan PB**. 2006. Impacts of genetic bottlenecks on soybean genome diversity. *Proc Natl Acad Sci U S A* **103**, 16666-16671.
**Kauppi L, Jeffreys AJ, Keeney S**. 2004. Where the crossovers are: recombination distributions in mammals. *Nature Reviews Genetics* **5**, 413-424.
**Khristich AN, Mirkin SM**. 2020. On the wrong DNA track: Molecular mechanisms of repeat-mediated genome instability. *Journal of Biological Chemistry* **295**, 4134-4170.
**Marand AP, Jansky SH, Zhao H, Leisner CP, Zhu X, Zeng Z, Crisovan E, Newton L, Hamernik AJ, Veilleux RE, Buell CR, Jiang J**. 2017. Meiotic crossovers are associated with open chromatin and enriched with Stowaway transposons in potato. *Genome Biology* **18**, 203.

**Marand AP, 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**, 645-662.
**McConaughy S**. 2022. Recombination Hotspots in Soybean [Glycine max (L.) Merr.].
**Mercier R, Mezard C, Jenczewski E, Macaisne N, Grelon M**. 2015. The Molecular Biology of Meiosis in Plants. *Annual Review of Plant Biology, Vol 66* **66**, 297-327.
**Mezard C**. 2006. Meiotic recombination hotspots in plants. *Biochemical Society Transactions* **34**, 531-534.
**Moyers BT, Morrell PL, McKay JK**. 2018. Genetic costs of domestication and improvement. *Journal of Heredity* **109**, 103-116.
**Myers S, Freeman C, Auton A, Donnelly P, McVean G**. 2008. A common sequence motif associated with recombination hot spots and genome instability in humans. *Nature Genetics* **40**, 1124-1129.
**Ross-Ibarra J**. 2004. The evolution of recombination under domestication: a test of two hypotheses. *The American Naturalist* **163**, 105-112.
**Schwarzkopf EJ, Motamayor JC, Cornejo OE**. 2020. Genetic differentiation and intrinsic genomic features explain variation in recombination hotspots among cocoa tree populations. *Bmc Genomics* **21**, 1-16.
**Shilo S, Melamed-Bessudo C, Dorone Y, Barkai N, Levy AA**. 2015. DNA Crossover Motifs Associated with Epigenetic Modifications Delineate Open Chromatin Regions in Arabidopsis. *Plant Cell* **27**, 2427-2436.
**Song Q, Hyten DL, Jia G, Quigley CV, Fickus EW, Nelson RL, Cregan PB**. 2015. Fingerprinting Soybean Germplasm and Its Utility in Genomic Research. *G3 (Bethesda)* **5**, 1999-2006.
**Thomas JH, Emerson RO, Shendure J**. 2009. Extraordinary molecular evolution in the PRDM9 fertility gene. *Plos One* **4**, e8505.
**Tukey JW**. 1977. *Exploratory data analysis*: Reading, MA.

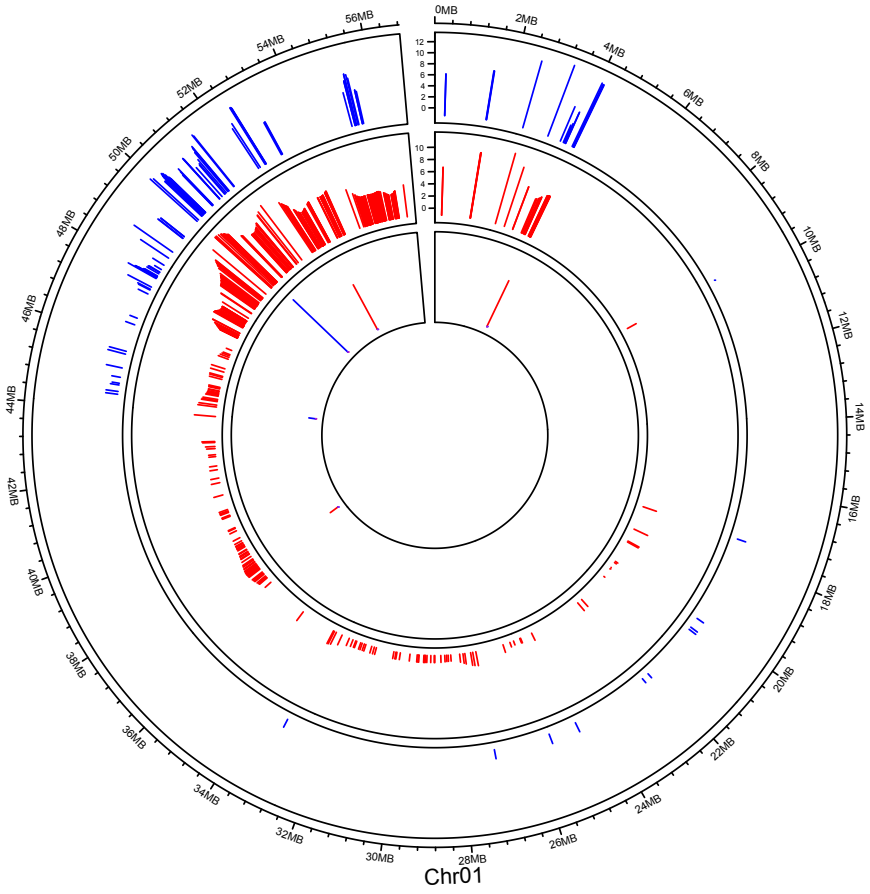
# Appendix

## Supplemental Information

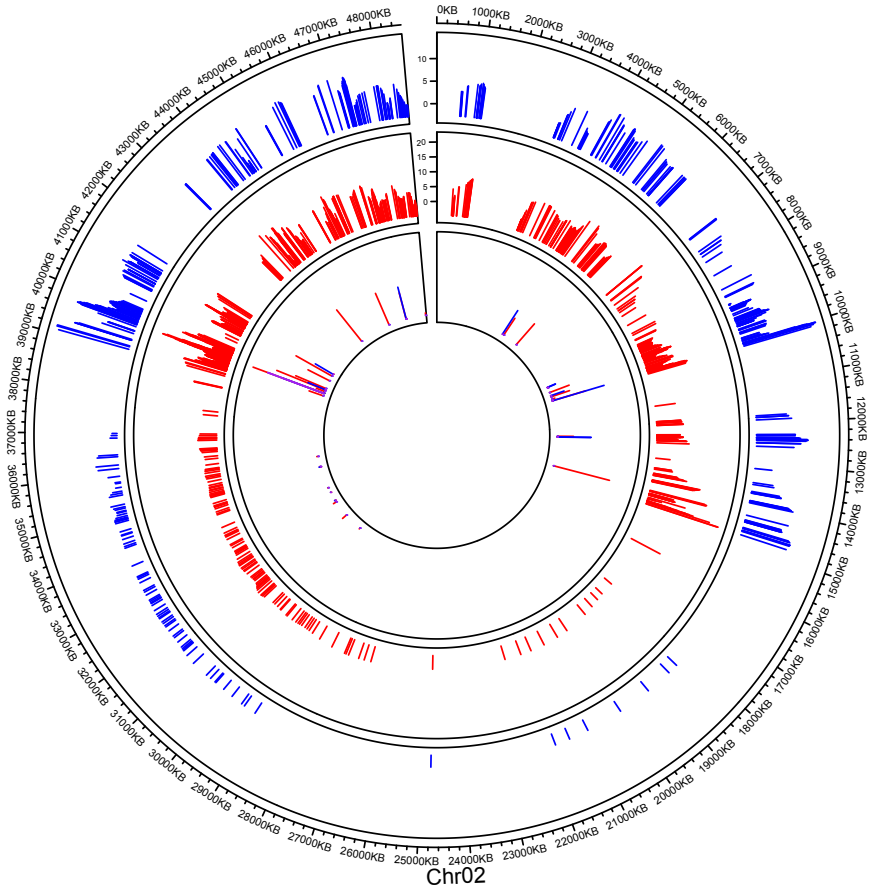
Supplemental figures 1-20 contain one figure per chromosome, each figure displays the same information with the two outer rings represent the biparental populations by physical distance along chromosomes, Williams 82 x PI479752 (blue) and Williams 82 x Essex (red) display recombination rates in cM/Mbp (y-axis). The inner most circle contains the hotspots for each population in their respective color

Supplemental figures 21-40 contain one figure per chromosome, each figure displays the same information with the two outer rings represent the biparental populations by physical distance along chromosomes, Williams 82 x PI479752 (blue) and Williams 82 x Essex (red) display recombination rates in cM/Mbp (y-axis). The inner most circle contains the hotspots for each population in their respective color

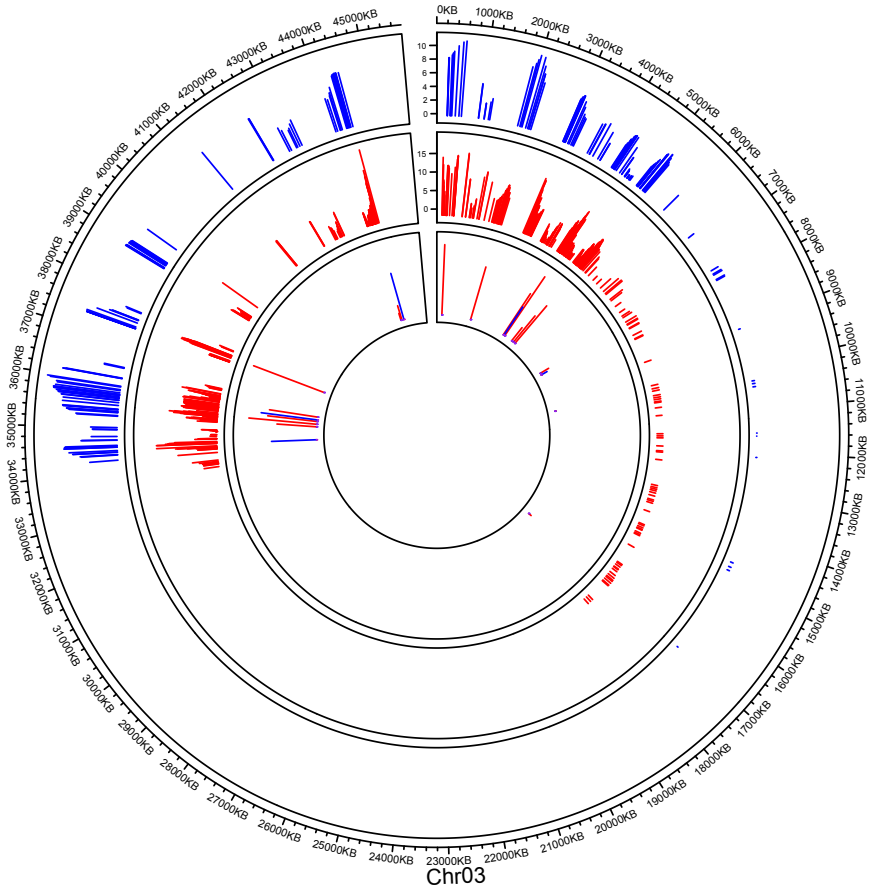
**Supplementary Fig S1. Chromosome 1 Recombination Hotspots.**



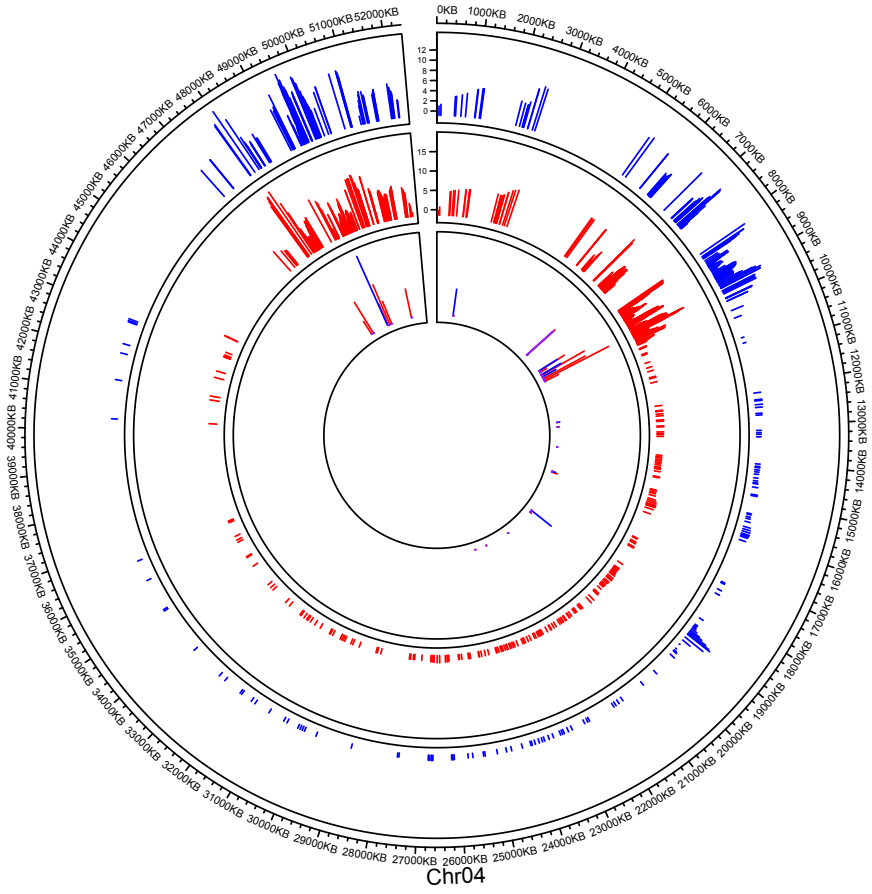
**Supplementary Fig S2. Chromosome 2 Recombination Hotspots.**



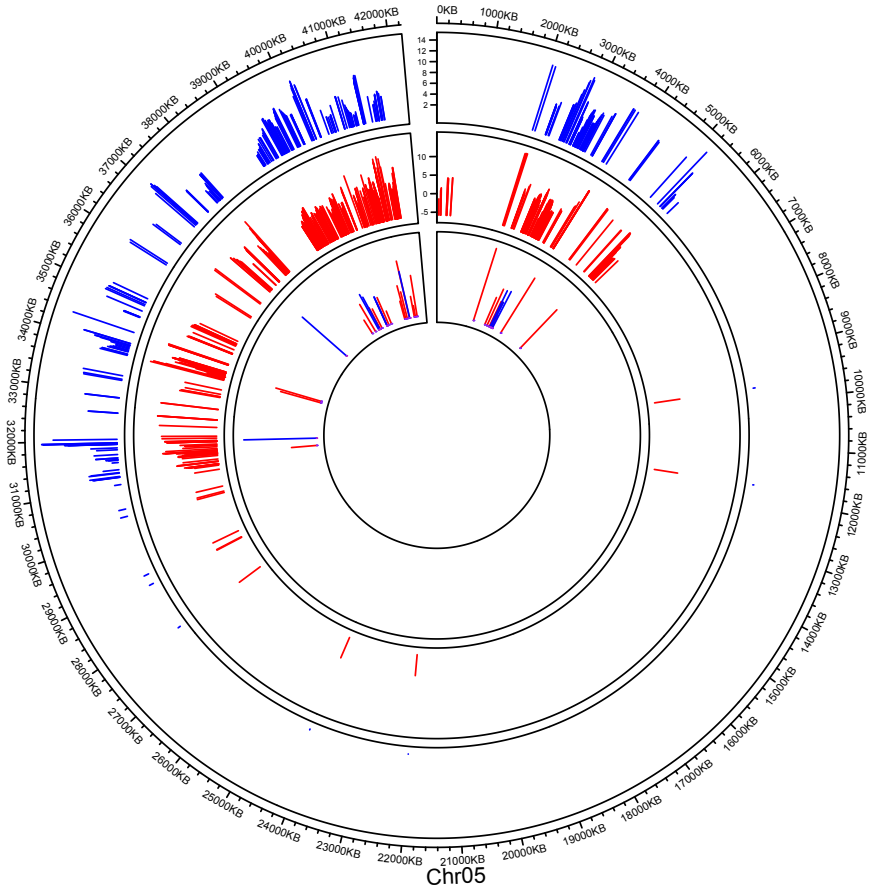
**Supplementary Fig S3. Chromosome 3 Recombination Hotspots.**



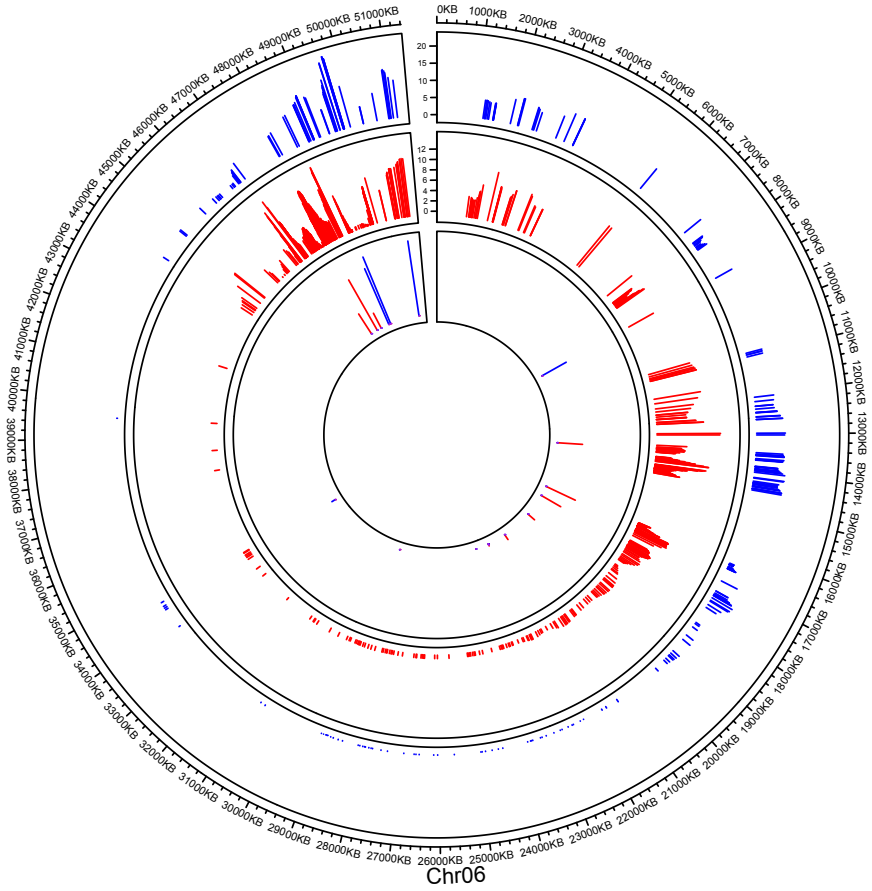
**Supplementary Fig S4. Chromosome 4 Recombination Hotspots.**



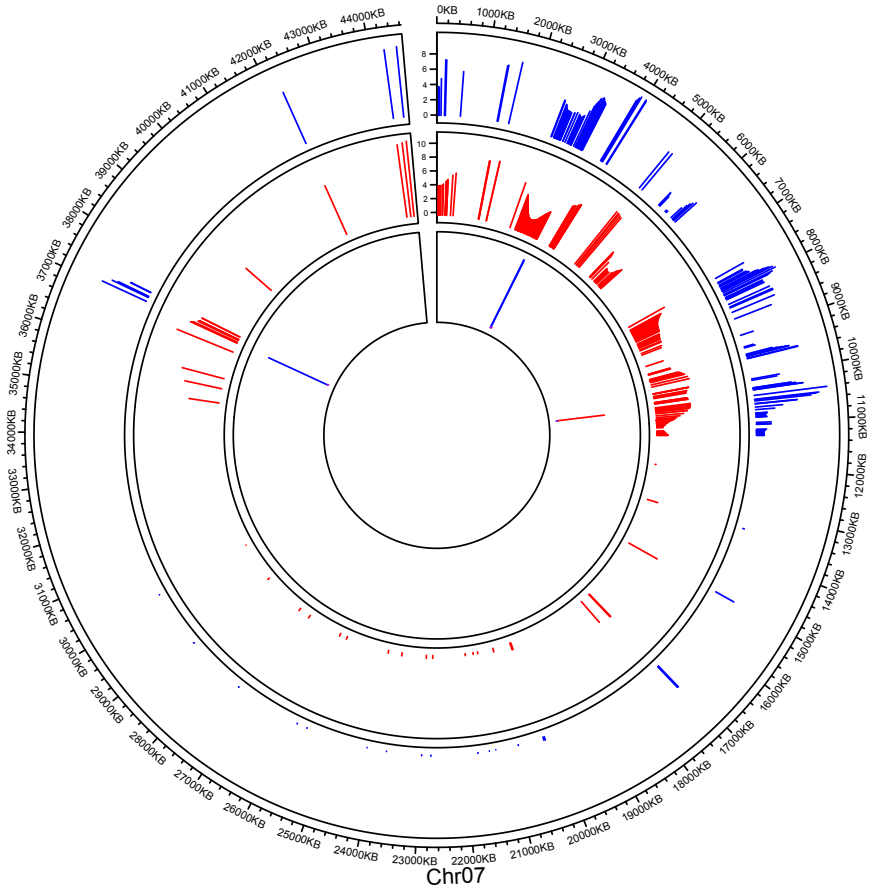
**Supplementary Fig S5. Chromosome 5 Recombination Hotspots.**



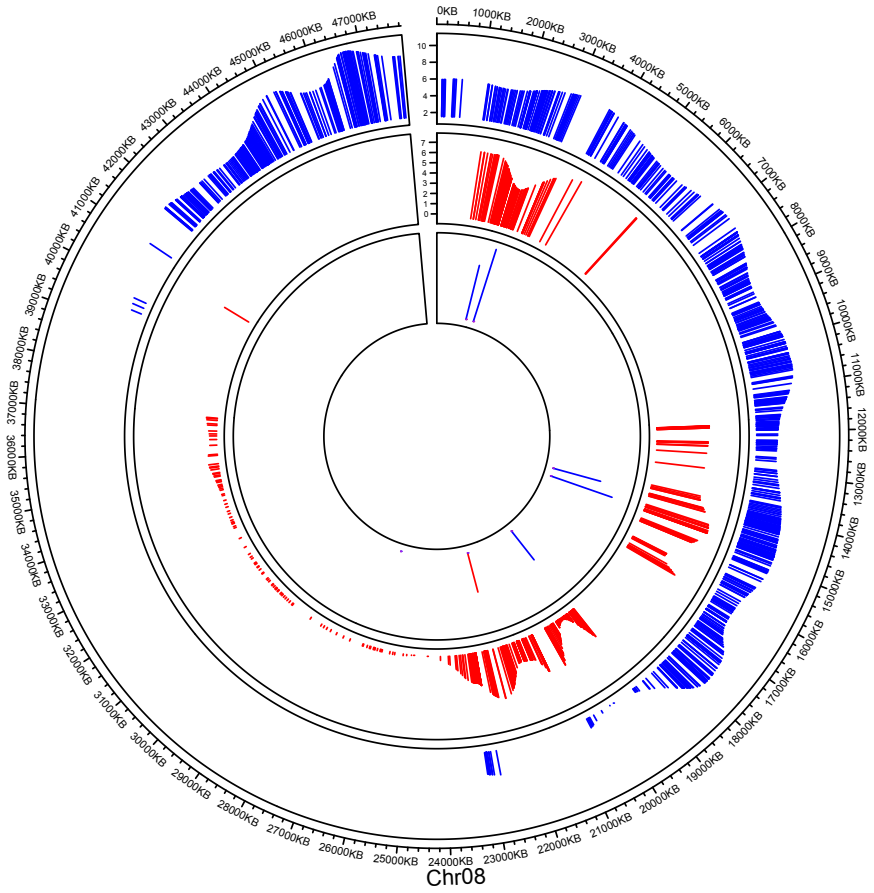
**Supplementary Fig S6. Chromosome 6 Recombination Hotspots.**



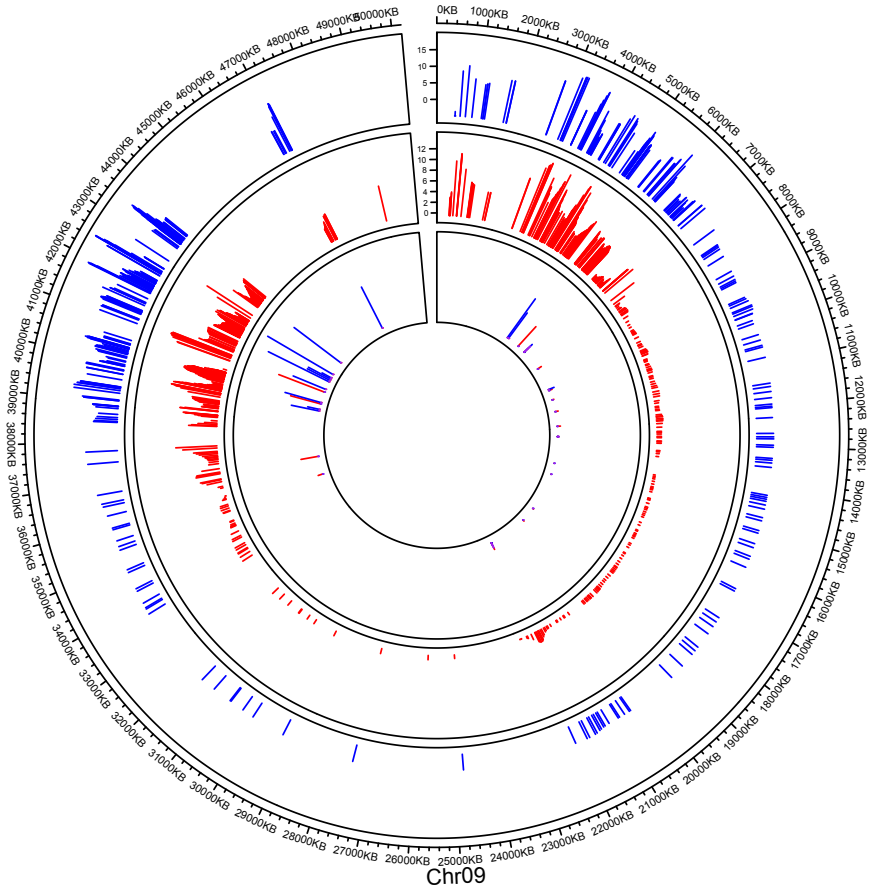
**Supplementary Fig S7. Chromosome 7 Recombination Hotspots.**



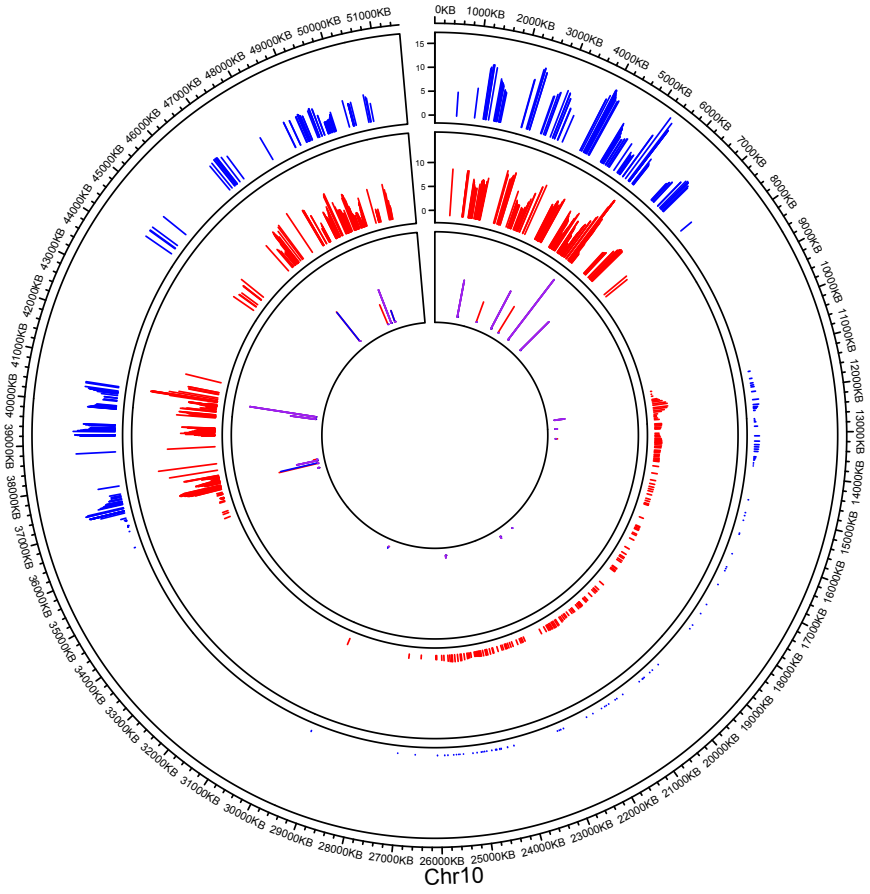
**Supplementary Fig S8. Chromosome 8 Recombination Hotspots.**



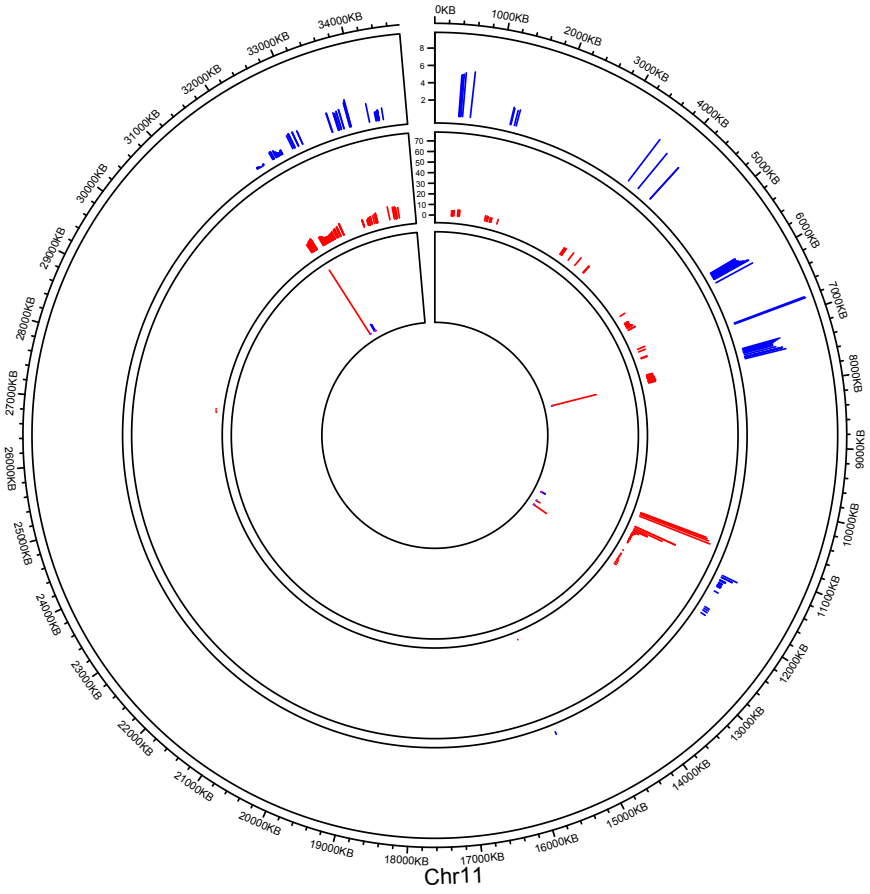
**Supplementary Fig S9. Chromosome 9 Recombination Hotspots.**



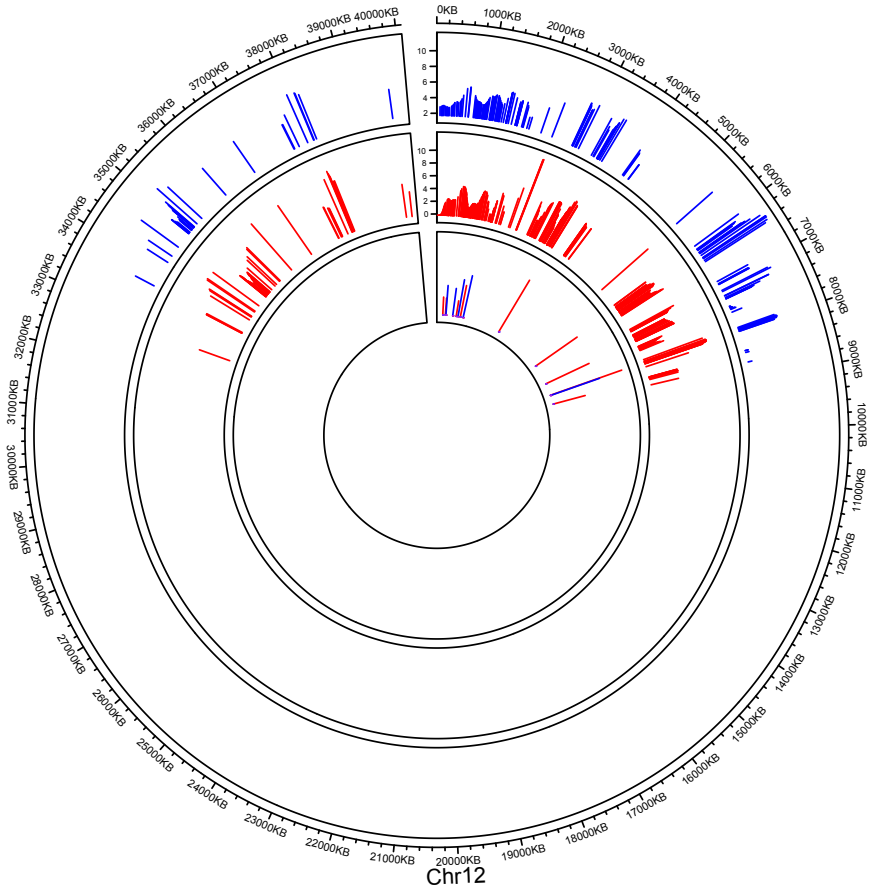
**Supplementary Fig S10. Chromosome 10 Recombination Hotspots.**



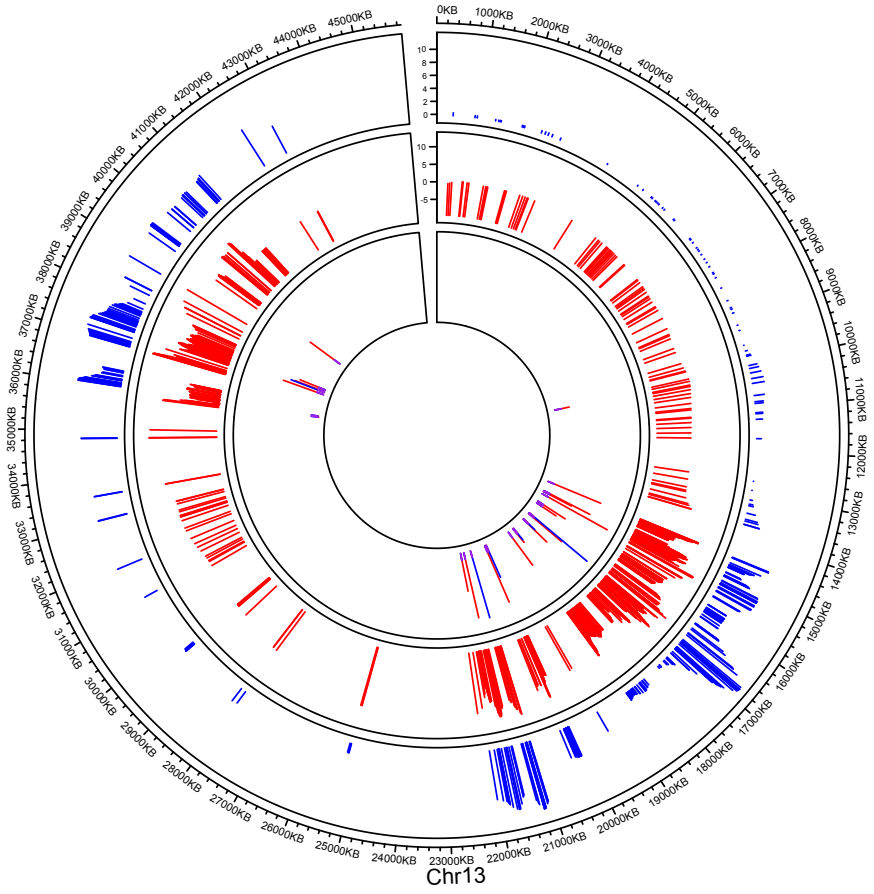
**Supplementary Fig S11. Chromosome 11 Recombination Hotspots.**



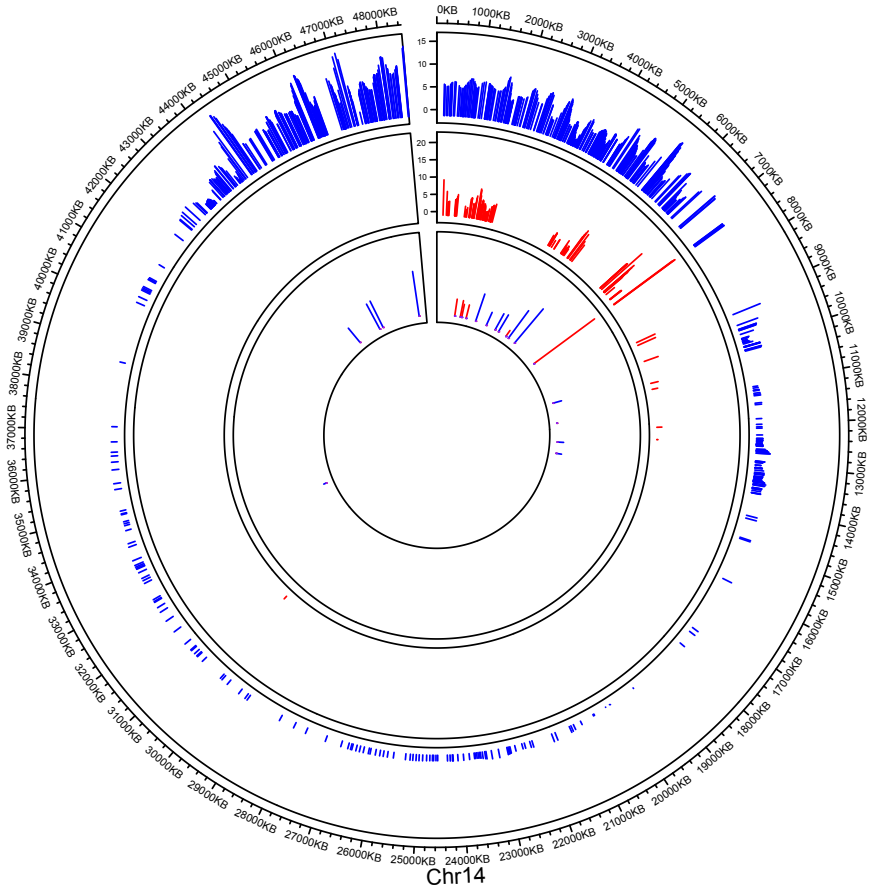
**Supplementary Fig S12. Chromosome 12 Recombination Hotspots.**



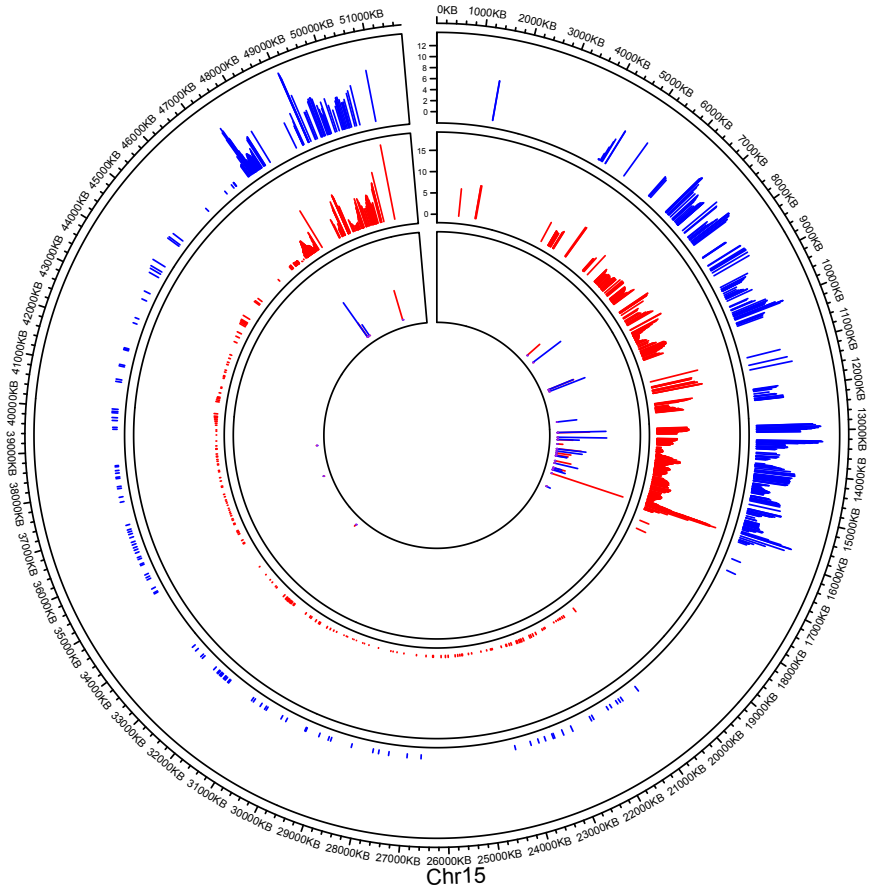
**Supplementary Fig S13. Chromosome 13 Recombination Hotspots.**



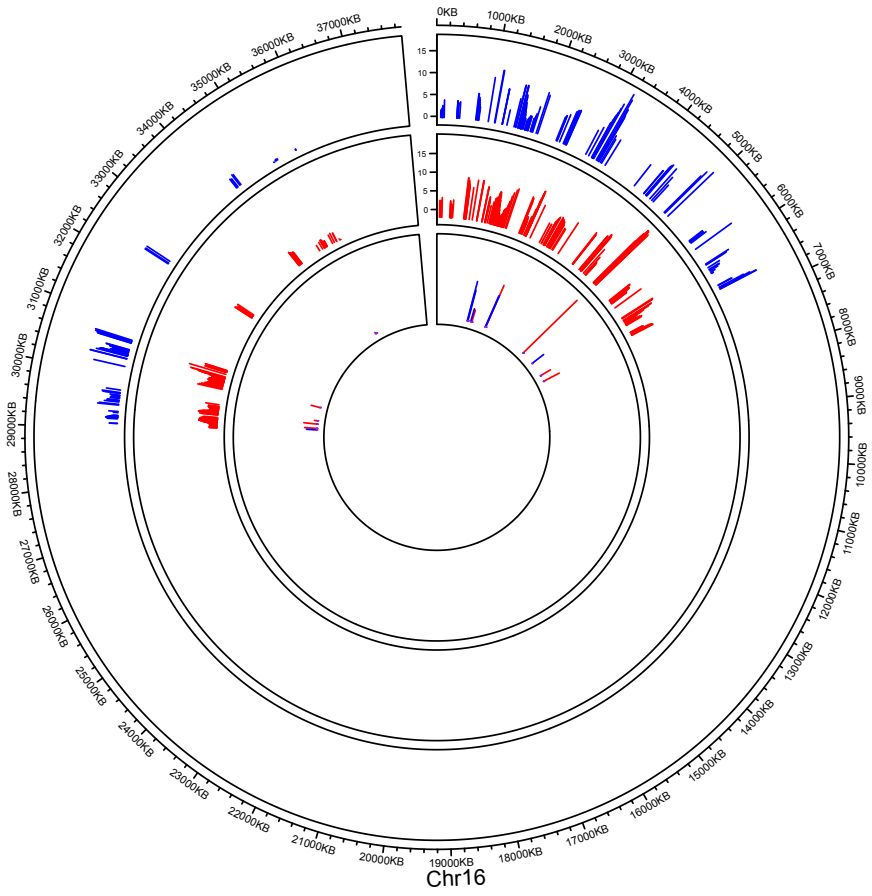
**Supplementary Fig S14. Chromosome 14 Recombination Hotspots.**



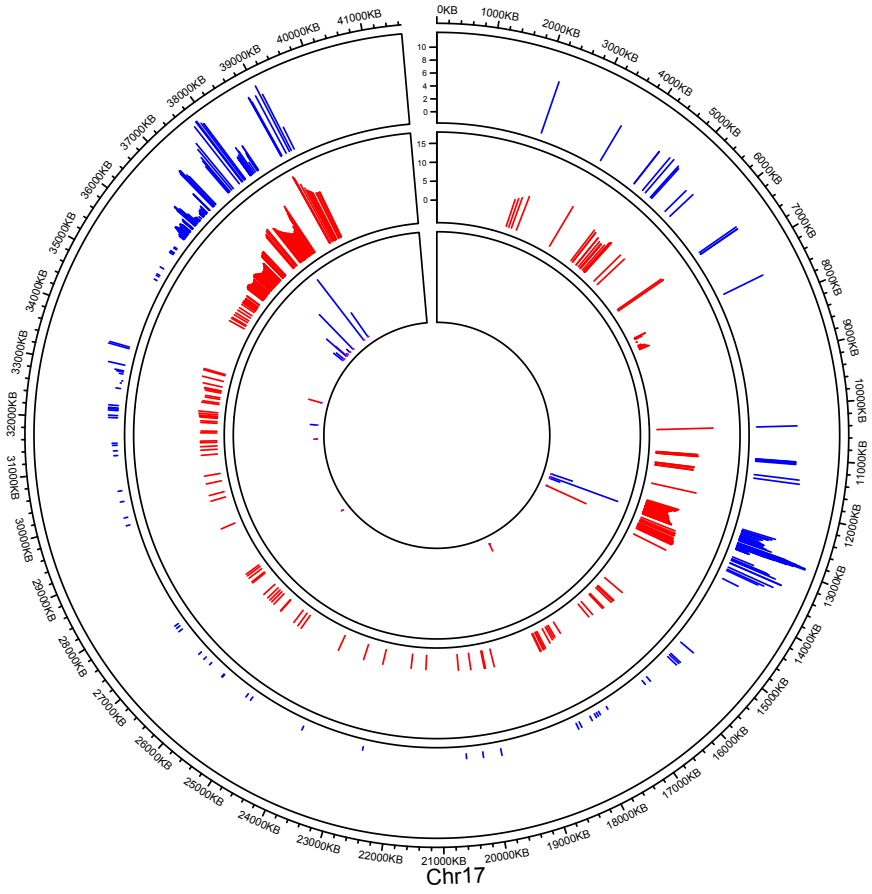
**Supplementary Fig S15. Chromosome 15 Recombination Hotspots.**



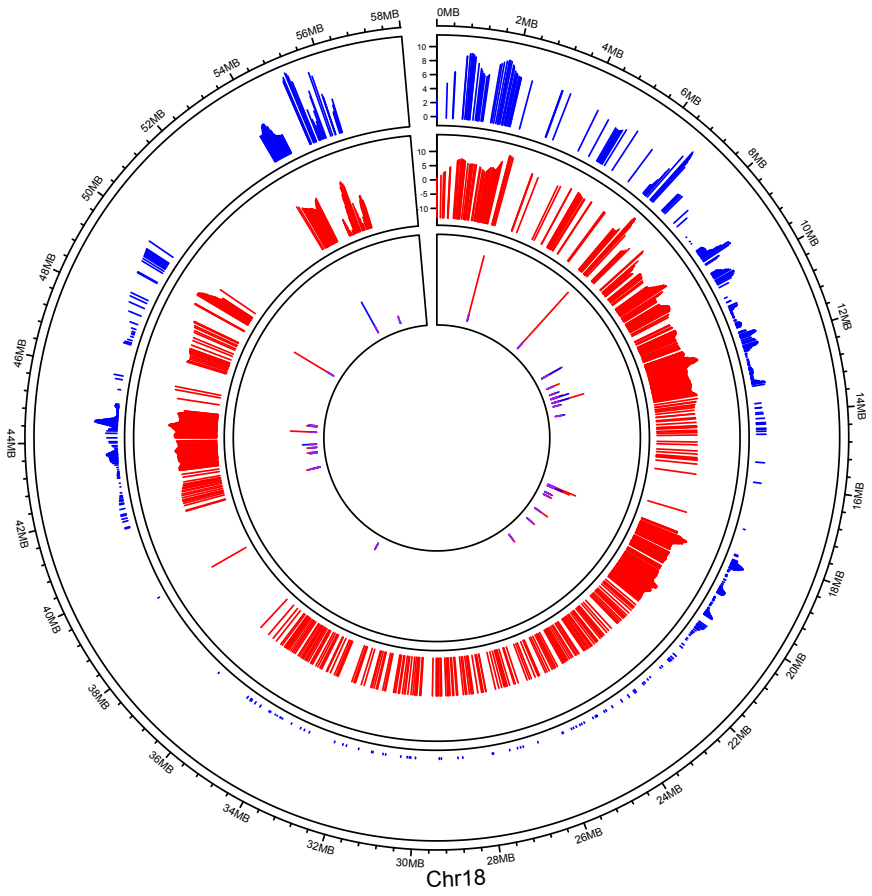
**Supplementary Fig S16. Chromosome 16 Recombination Hotspots.**



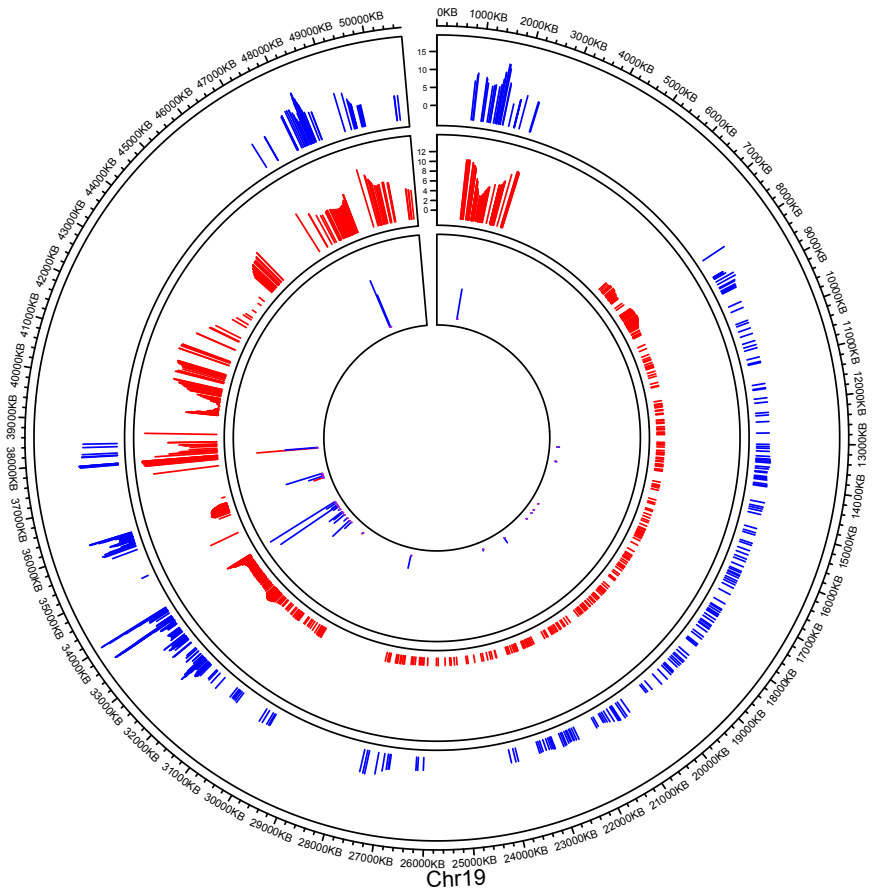
**Supplementary Fig S17. Chromosome 17 Recombination Hotspots.**



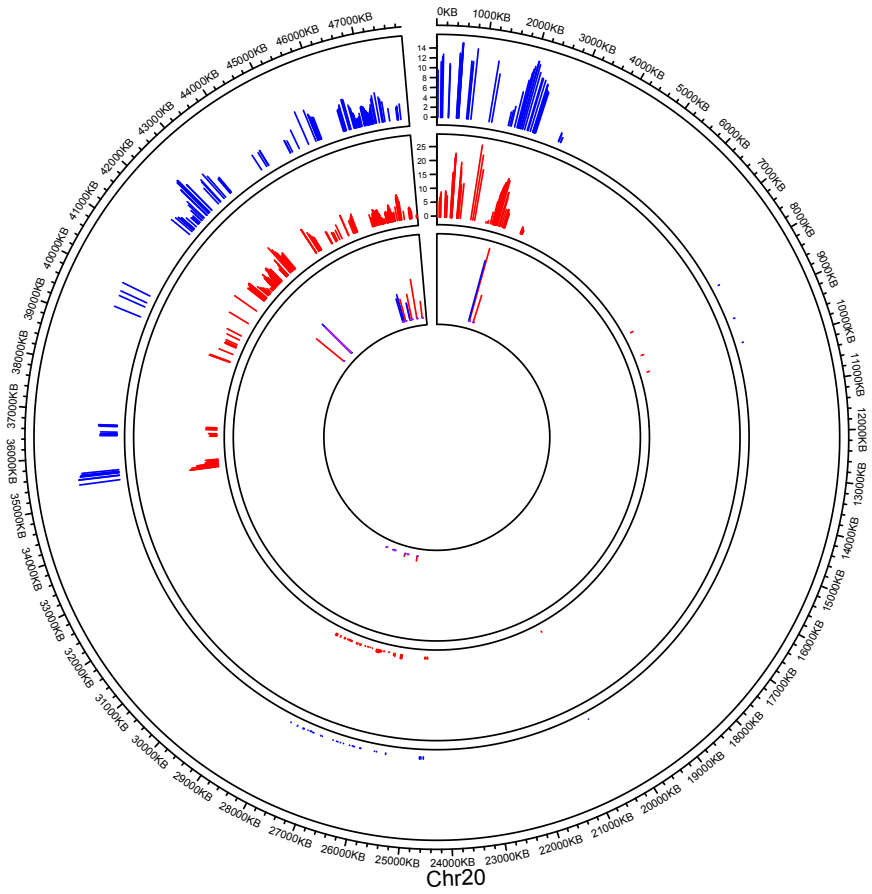
**Supplementary Fig S18. Chromosome 18 Recombination Hotspots.**



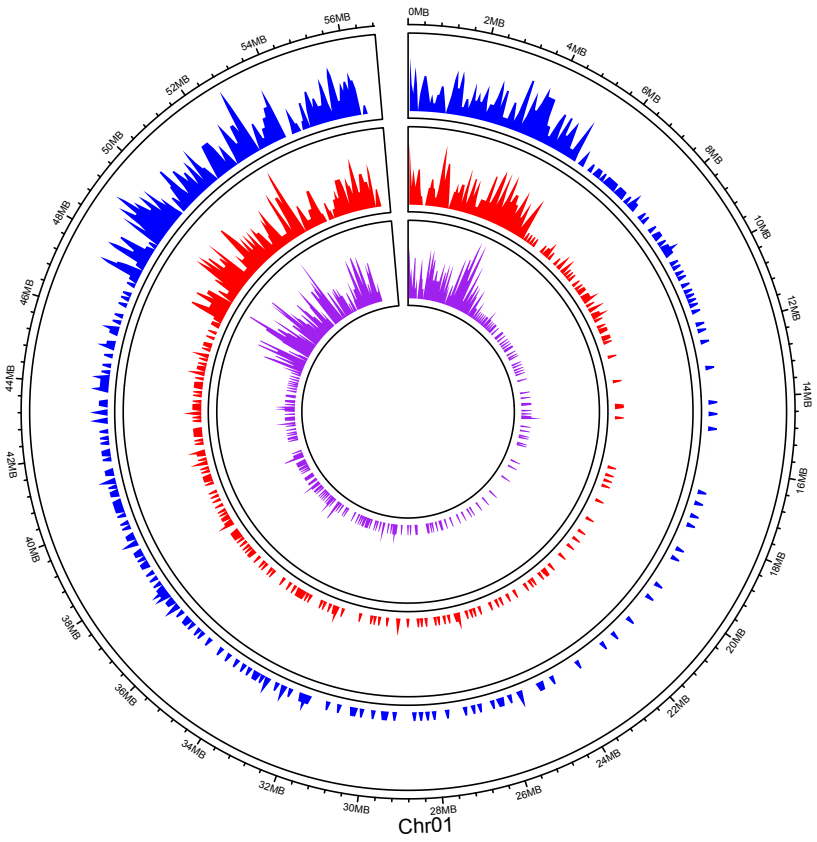
**Supplementary Fig S19. Chromosome 19 Recombination Hotspots.**



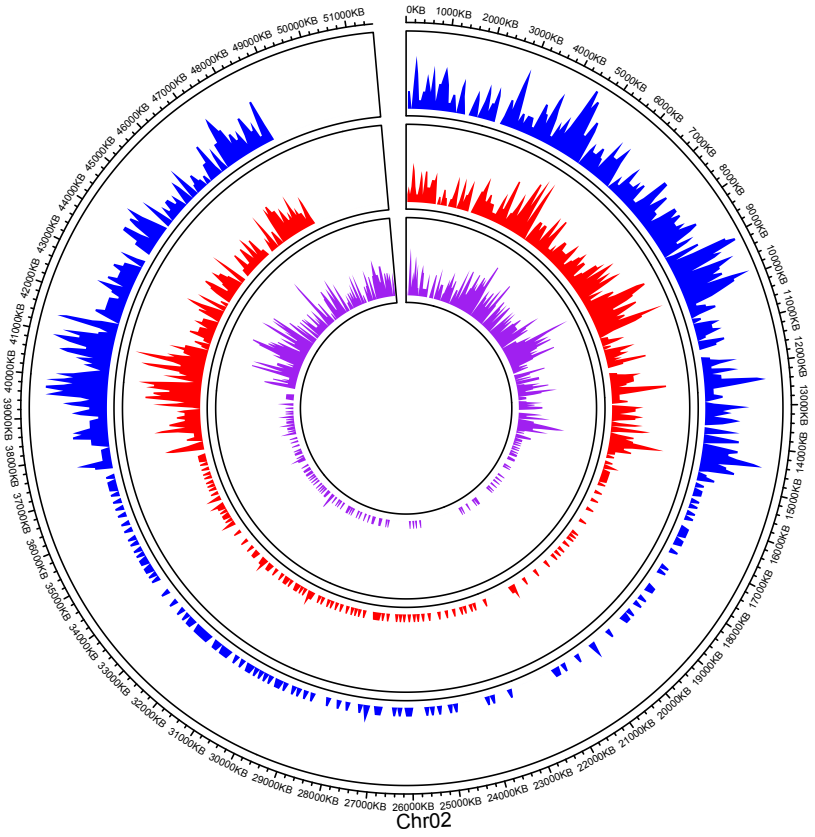
**Supplementary Fig S20. Chromosome 20 Recombination Hotspots.**



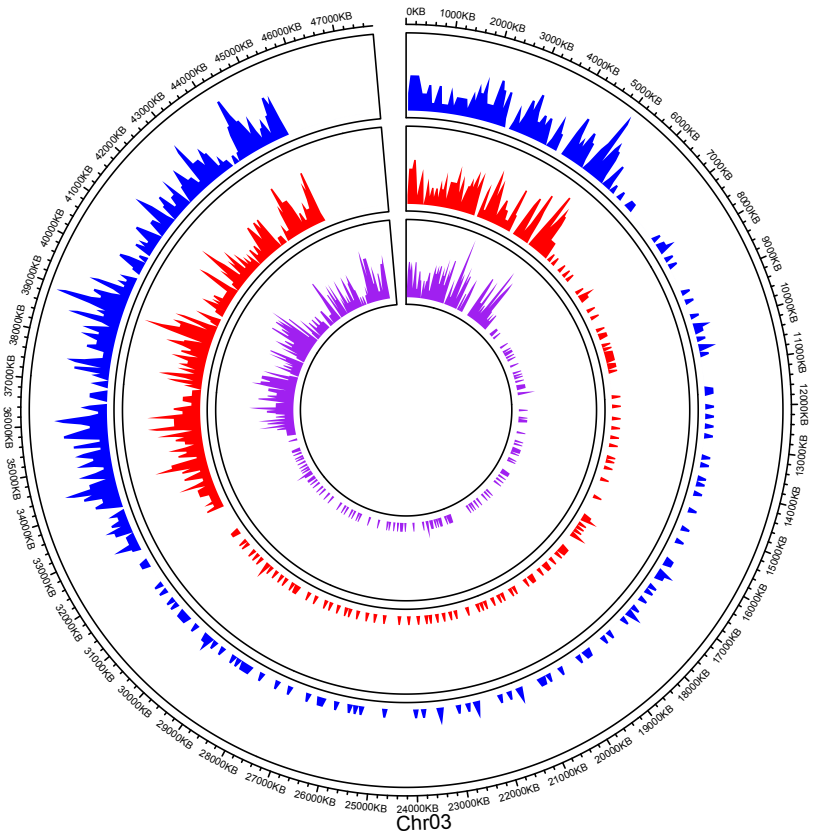
**Supplementary Fig S21. Chromosome 1 Ancestral Recombination Hotspots.**



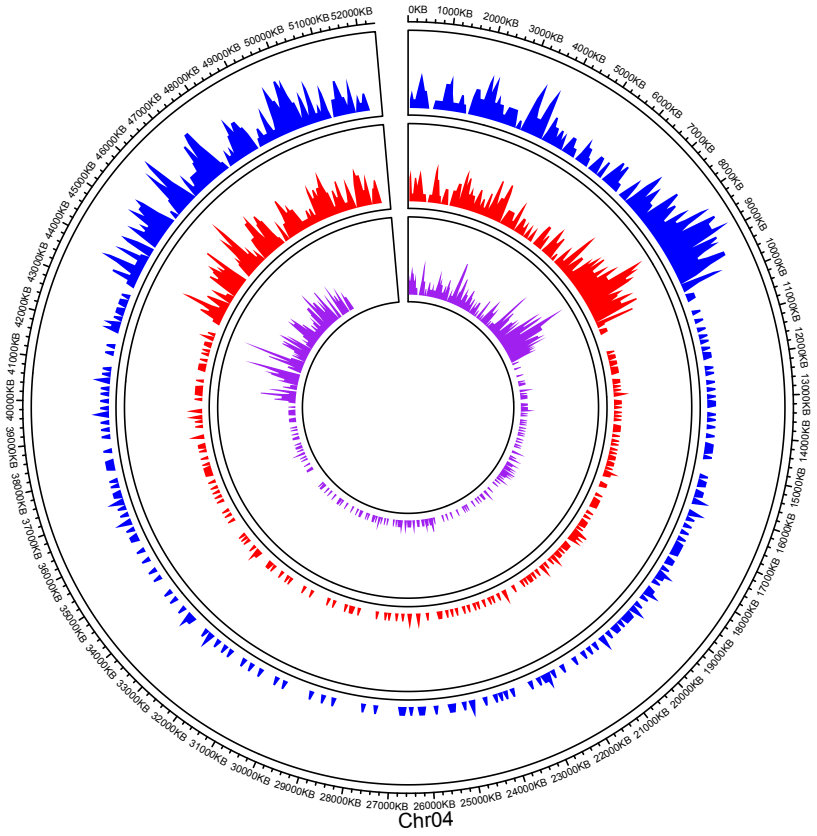
**Supplementary Fig S22. Chromosome 2 Ancestral Recombination Hotspots.**



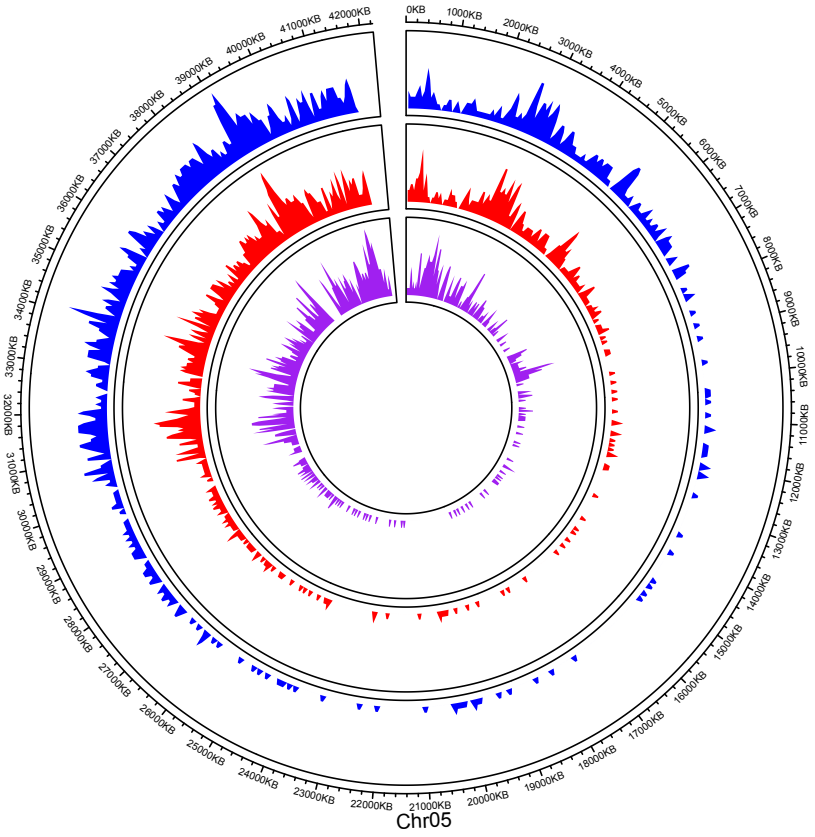
**Supplementary Fig S23. Chromosome 3 Ancestral Recombination Hotspots.**



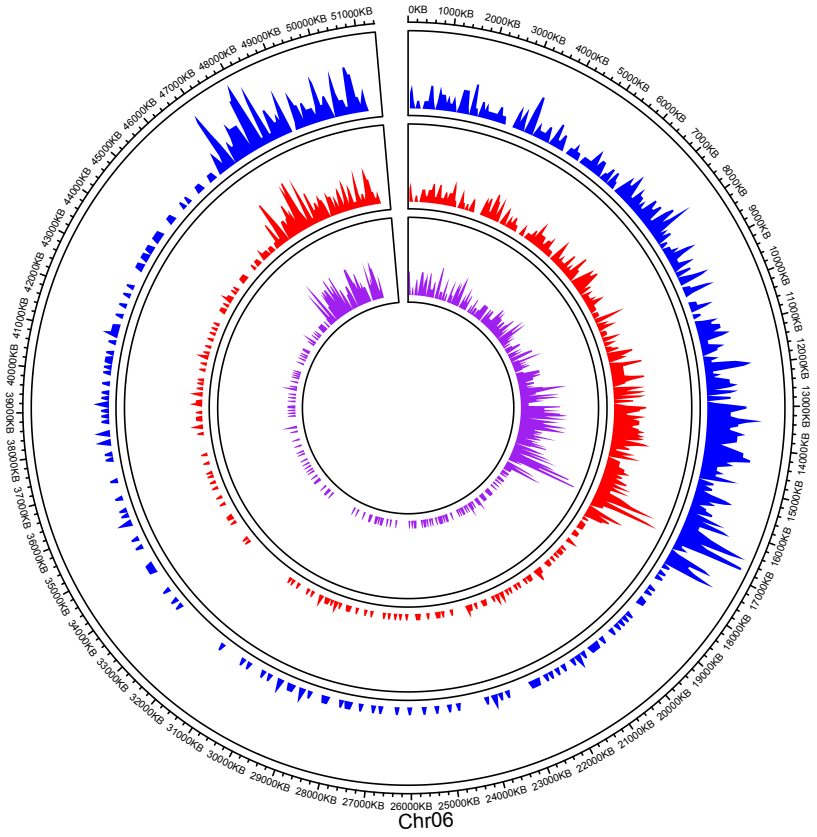
**Supplementary Fig S24. Chromosome 4 Ancestral Recombination Hotspots.**



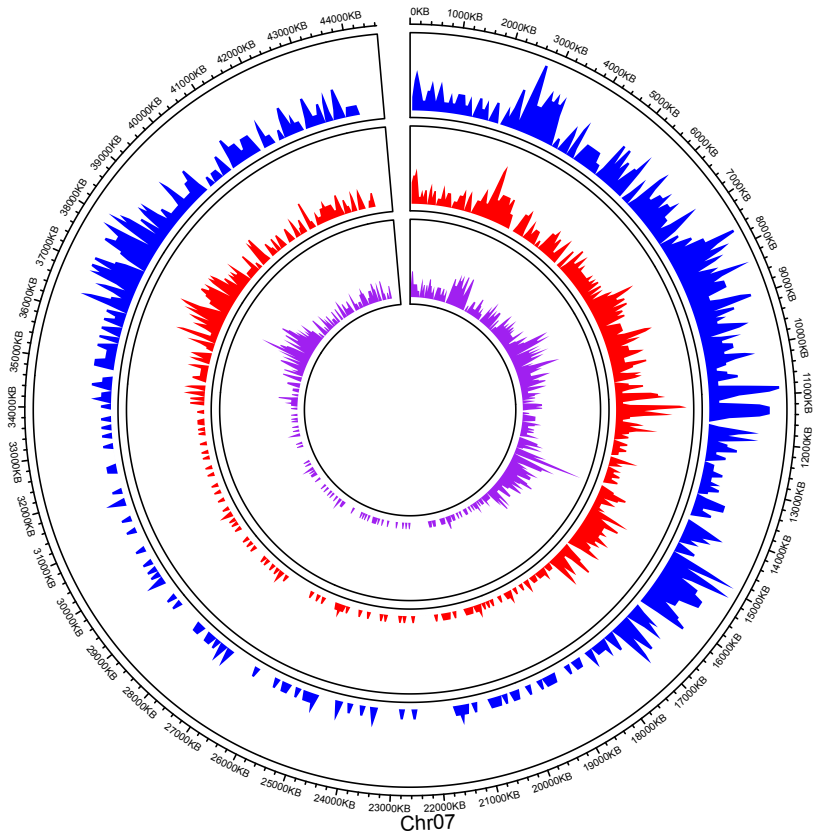
**Supplementary Fig S25. Chromosome 5 Ancestral Recombination Hotspots.**



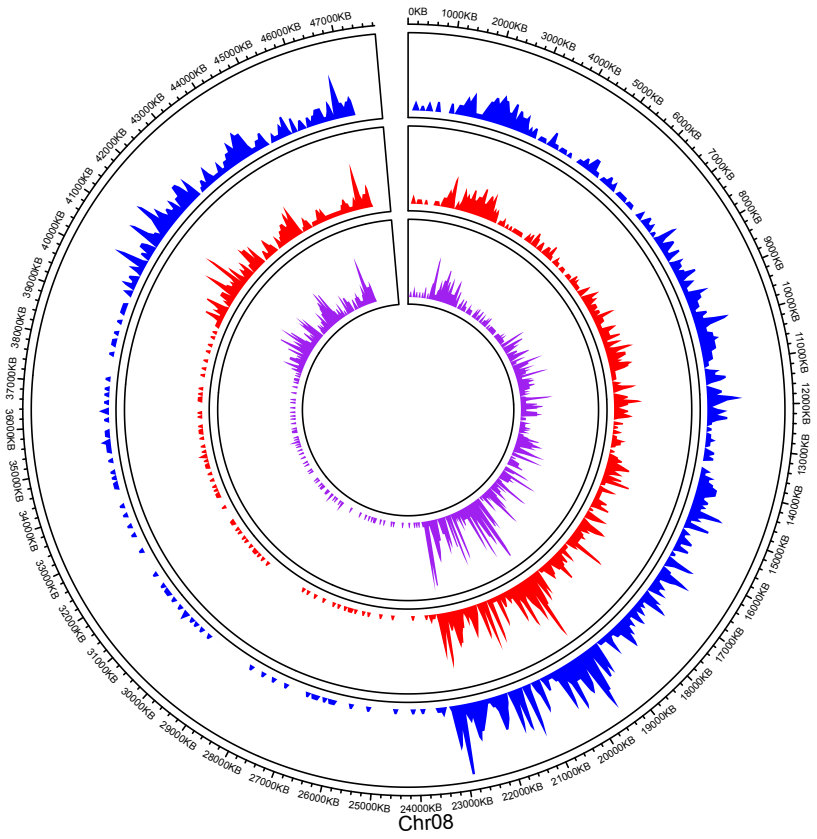
**Supplementary Fig S26. Chromosome 6 Ancestral Recombination Hotspots.**



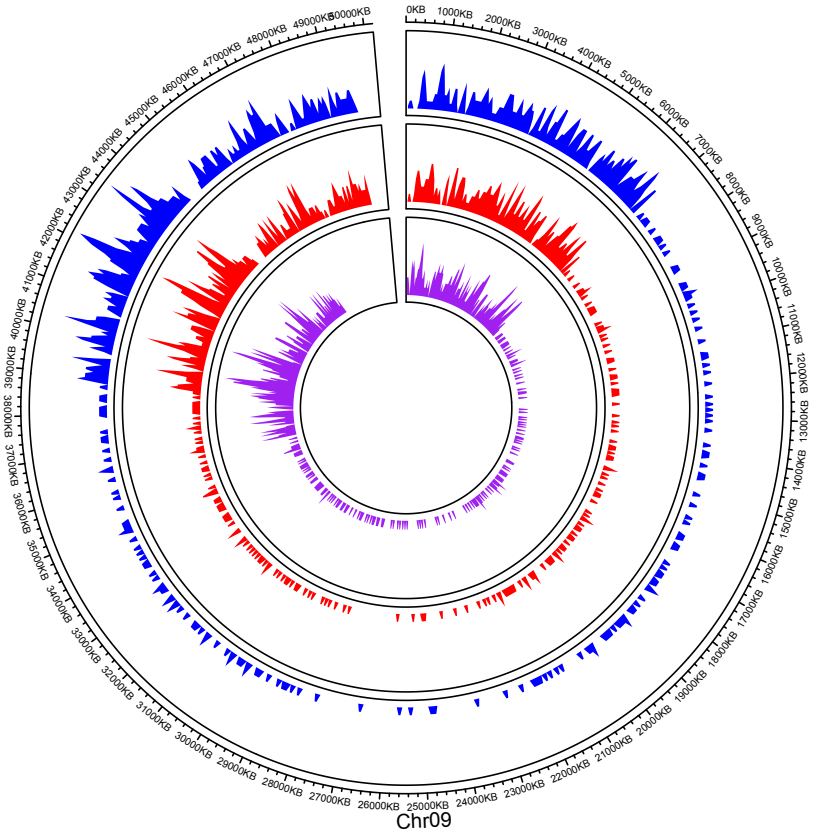
**Supplementary Fig S27. Chromosome 7 Ancestral Recombination Hotspots.**



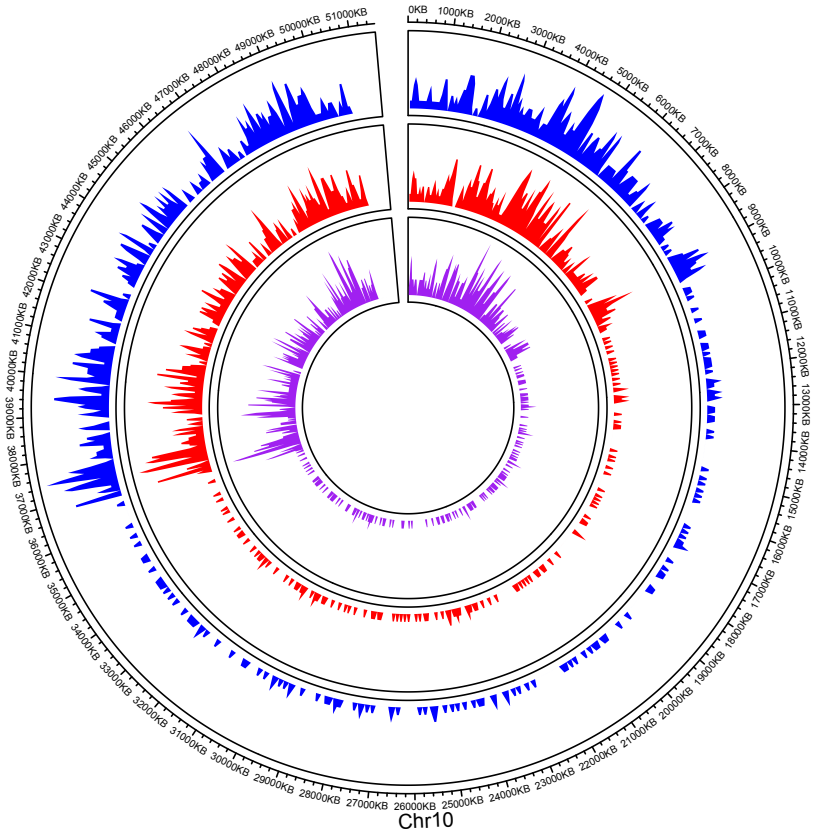
**Supplementary Fig S28. Chromosome 8 Ancestral Recombination Hotspots.**



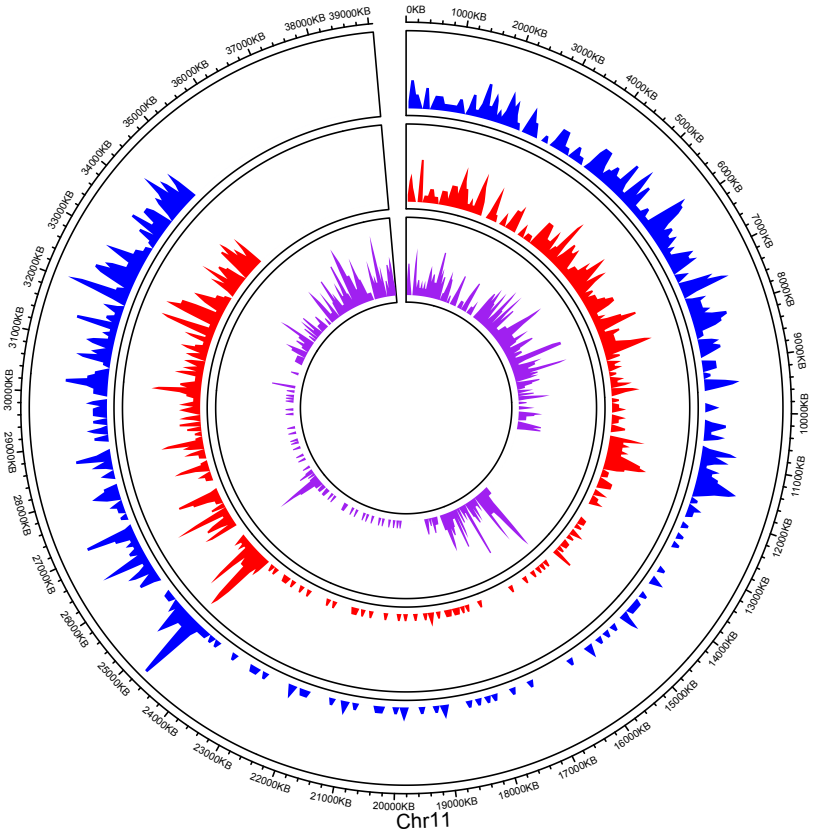
**Supplementary Fig S29. Chromosome 9 Ancestral Recombination Hotspots.**



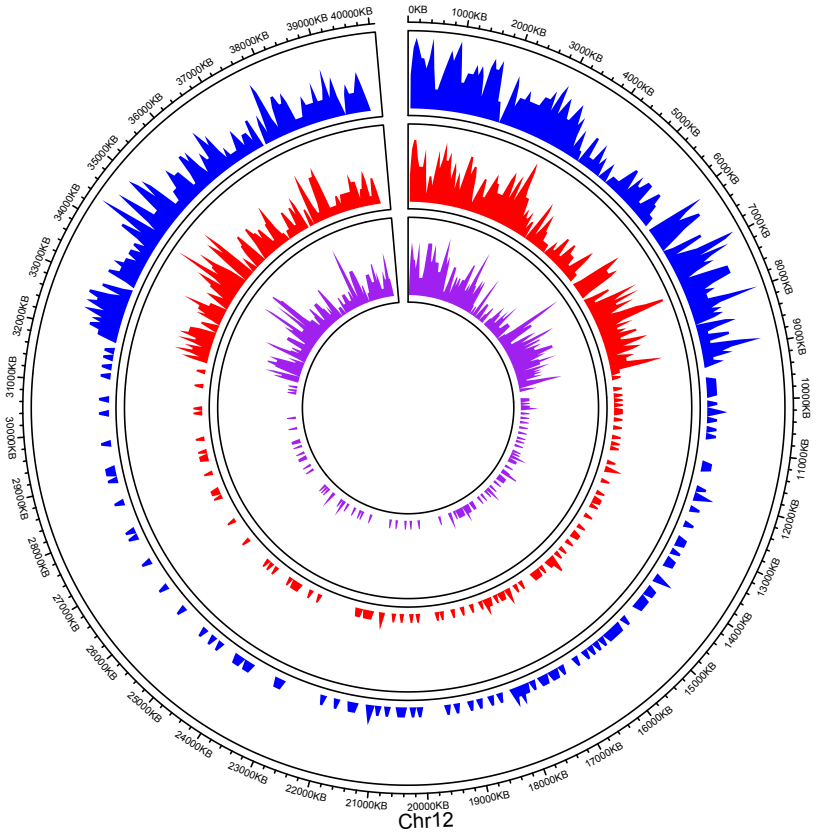
**Supplementary Fig S30. Chromosome 10 Ancestral Recombination Hotspots.**



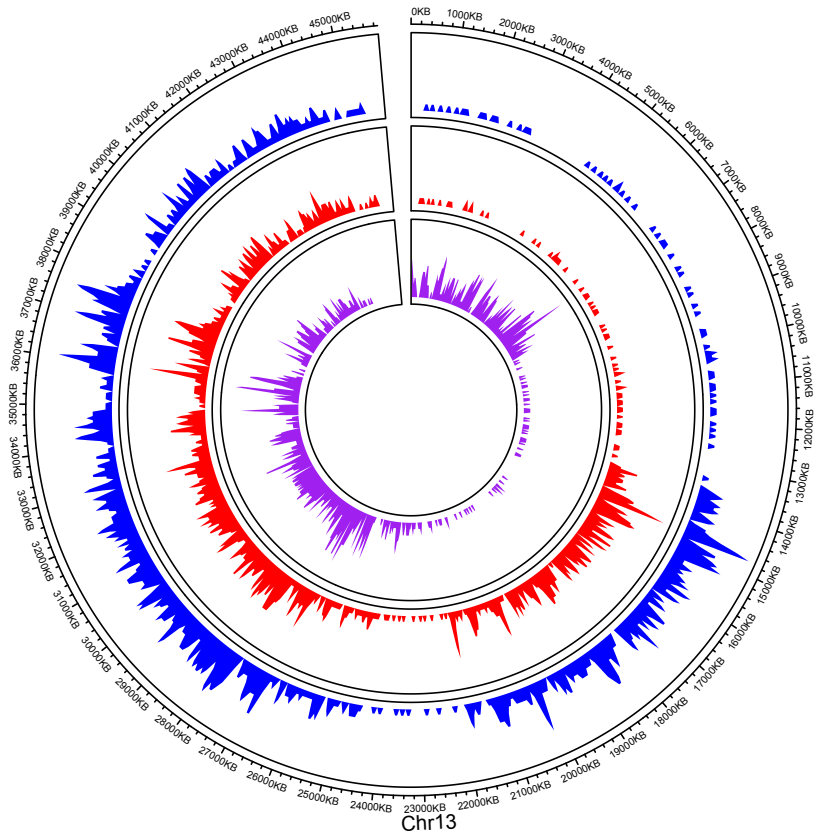
**Supplementary Fig S31. Chromosome 11 Ancestral Recombination Hotspots.**



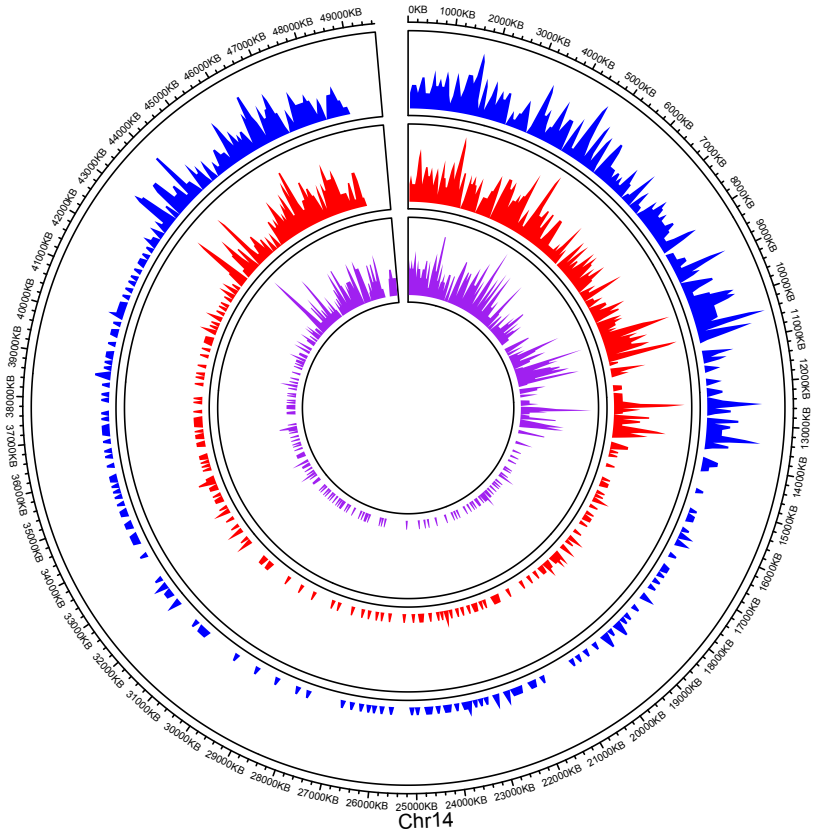
**Supplementary Fig S32. Chromosome 12 Ancestral Recombination Hotspots.**



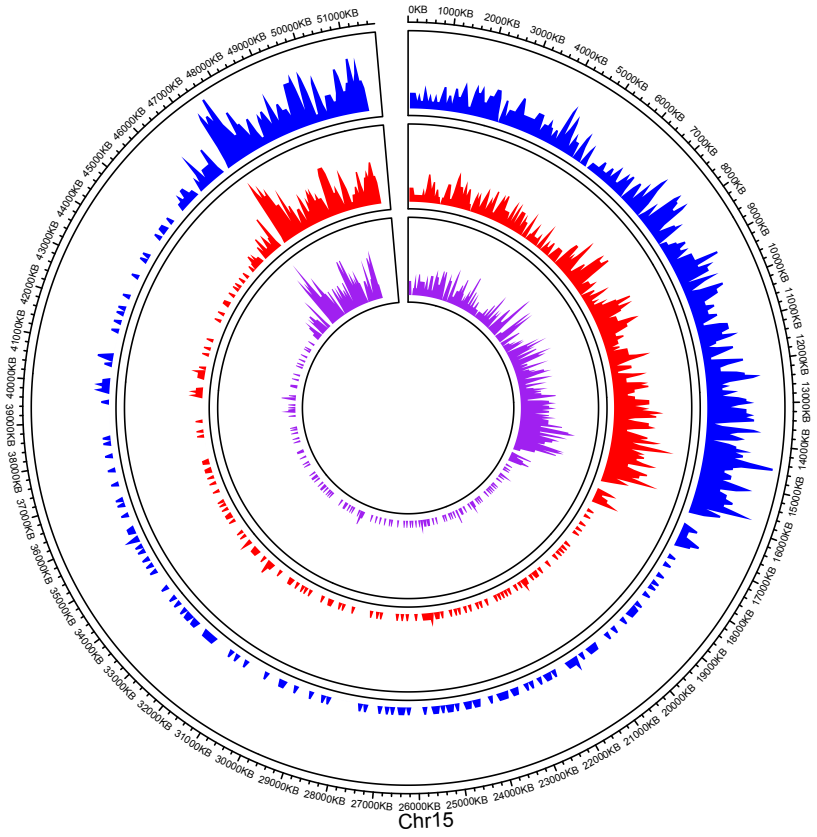
**Supplementary Fig S33. Chromosome 13 Ancestral Recombination Hotspots.**



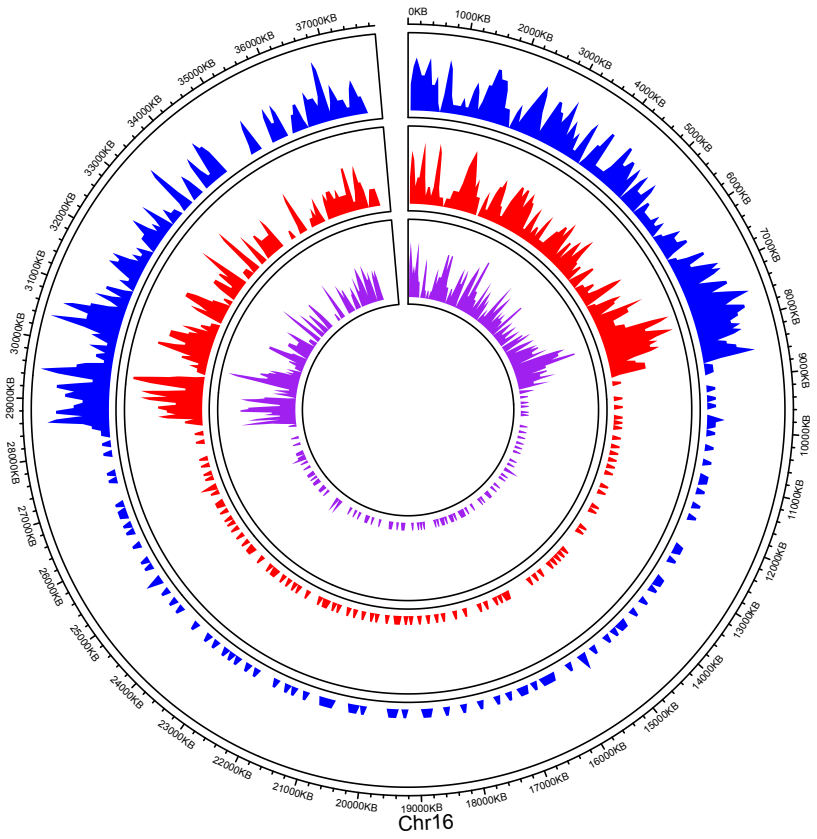
**Supplementary Fig S34. Chromosome 14 Ancestral Recombination Hotspots.**



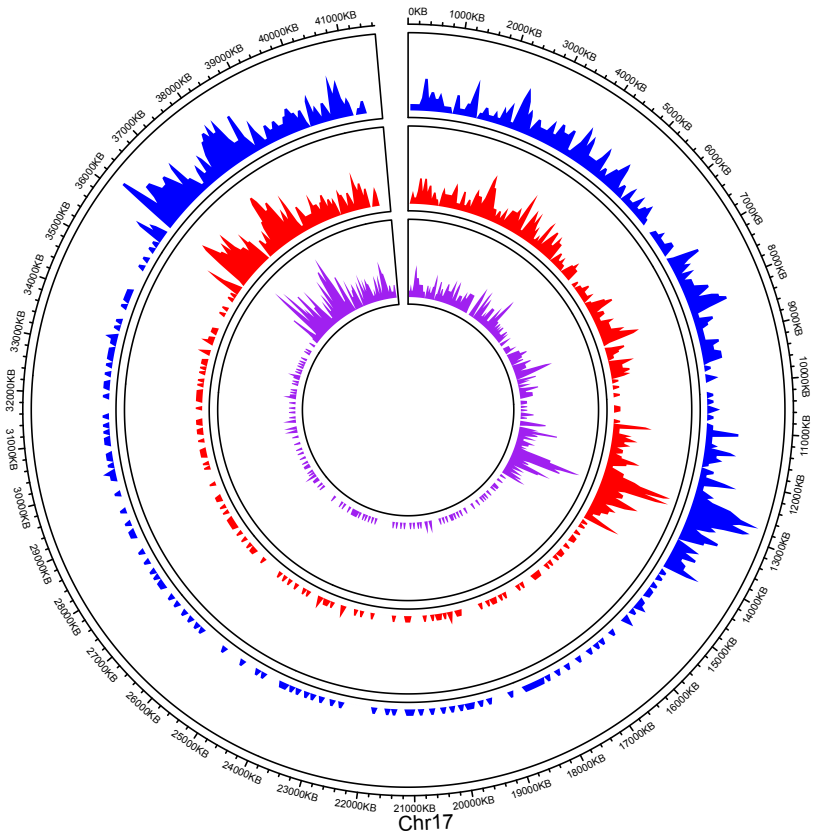
**Supplementary Fig S35. Chromosome 15 Ancestral Recombination Hotspots.**



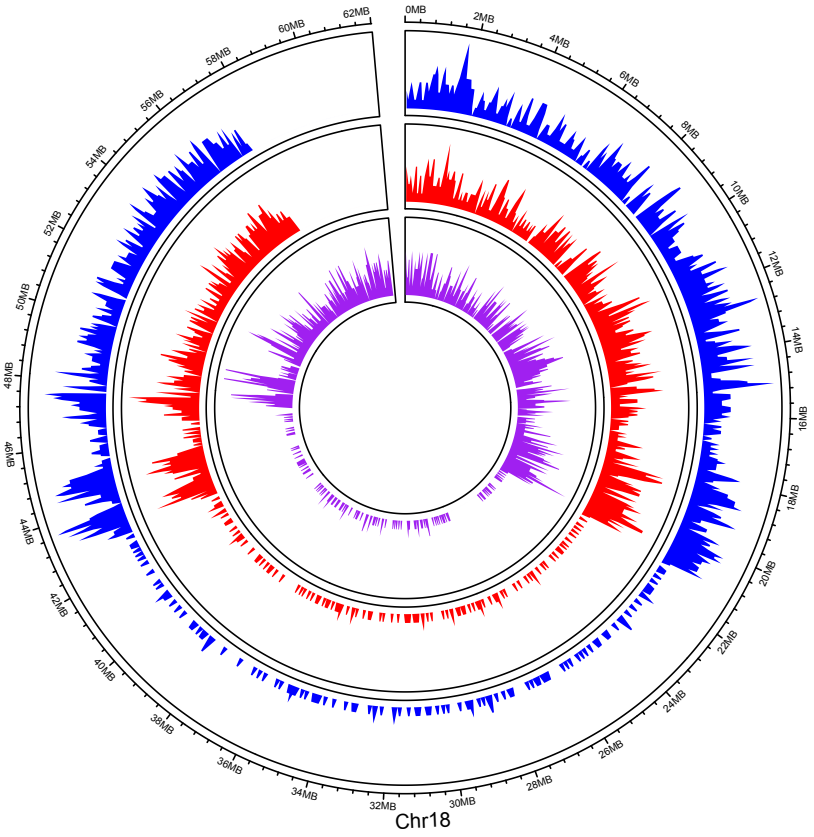
**Supplementary Fig S36. Chromosome 16 Ancestral Recombination Hotspots.**



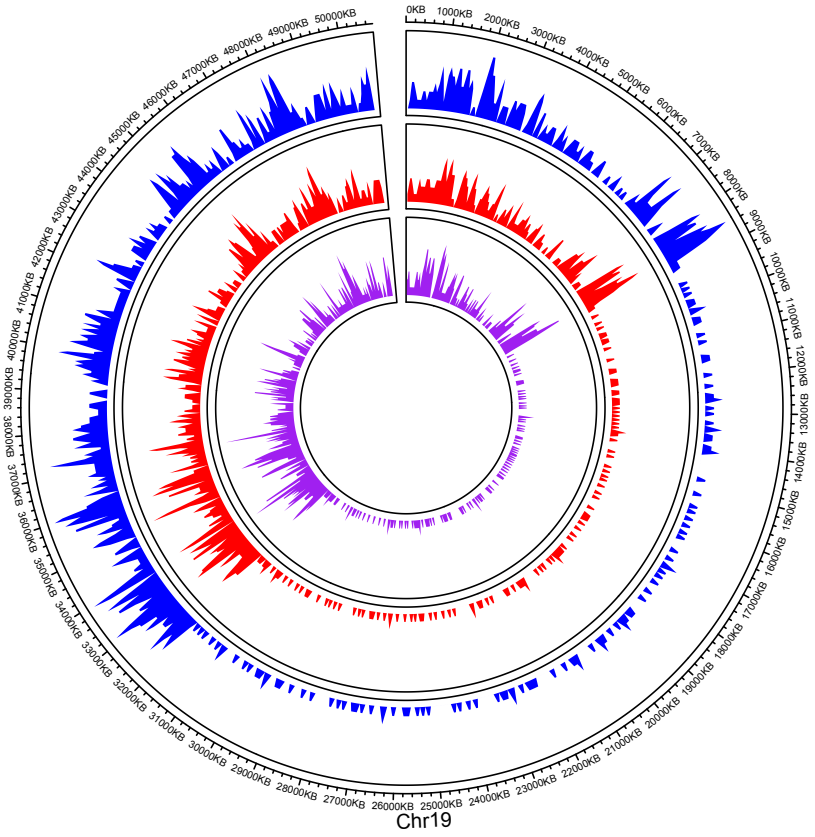
**Supplementary Fig S37. Chromosome 17 Ancestral Recombination Hotspots.**



**Supplementary Fig S38. Chromosome 18 Ancestral Recombination Hotspots.**



**Supplementary Fig S39. Chromosome 19 Ancestral Recombination Hotspots.**



**Supplementary Fig S40. Chromosome 20 Ancestral Recombination Hotspots.**