HOMEOLOGOUS EPISTASIS IN WHEAT: THE SEARCH FOR AN IMMORTAL HYBRID

A Dissertation

Presented to the Faculty of the Graduate School

of Cornell University

in Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

by

Nicholas Santantonio

August 2018

HOMEOLOGOUS EPISTASIS IN WHEAT: THE SEARCH FOR AN IMMORTAL HYBRID

Nicholas Santantonio, Ph.D.

Cornell University 2018

The subgenomes of an allopolyploid crop will each contain complete, yet evolutionarily divergent, sets of genes. Like a diploid hybrid, allopolyploids will have two versions, or homeoalleles, for every gene. Partial functional redundancy between homeologous genes should result in a deviation from additivity. These epistatic interactions between homeoalleles are analogous to dominance effects, but are fixed across subgenomes through self pollination. An allopolyploid can therefore be viewed as an immortalized hybrid, with the opportunity to select and fix favorable homeoallelic interactions within inbred varieties. With the availability of affordable genotyping and a reference genome to locate markers, breeders of allopolyploids now have the opportunity to manipulate subgenomes independently and fix beneficial interactions across subgenomes. I present a statistical framework for partitioning genetic variance to individual subgenomes of an allopolyploid, predicting breeding values for each subgenome, and evaluating the magnitude of homeologous epistasis. I also present a subfunctionalization epistasis model to estimate the degree of functional redundancy between homeoallelic loci and to determine their importance within a population. I search for genome-wide patterns indicative of homeoallelic subfunctionalization in a winter wheat breeding population by anchoring homeologous marker sets to the IWGSC RefSeq v1.0 sequence. Some traits displayed a pattern indicative of homeoallelic subfunctionalization, while other traits showed a less clear pattern. Using genomic prediction accuracy to

evaluate importance of marker interactions, I show that homeologous interactions explain a significant portion of the non-additive genetic signal. Allopolyploids have traditionally been treated as diploids in breeding programs because they undergo disomic inheritance. With modern DNA marker technology and ever increasing computational power, I provide a new framework for breeders of allopolyploid crops to characterize the genetic architecture of existing populations, determine breeding goals, and develop new strategies for selection of additive effects and homeologous epistasis in these ancient immortal hybrids.

BIOGRAPHICAL SKETCH

Nicholas Santantonio grew up in New Mexico and attended New Mexico State University (NMSU) in 2005. He was the first graduate of a newly formed Bachelor's degree program in Genetics at NMSU in 2010. He stayed on to complete a Master's degree in Plant and Environmental Sciences at NMSU, under the advisement of the alfalfa breeder, Dr. Ian Ray. His Masters thesis focused on genetic mapping of transpirational efficiency and forage yield in drought stressed alfalfa. Nicholas came to Ithaca in 2013 to start a PhD in the small grains breeding program at Cornell under the advisement of Dr. Mark Sorrells. During his PhD at Cornell, Nicholas shifted his focus from applied plant breeding to more theoretical quantitative genetics. He became fascinated with allopolyploid genetics after reading a paper by James Mac Key (1970) in a literature review course lead by his adviser. A two page proposal written for that course was the inspiration that lead to the research presented here, and is included in its original form at the end of this document. Treating an allopolyploid as an immortalized hybrid, Nicholas developed a statistical framework to model homeologous interactions between the subgenomes of an allopolyploid. He will be starting a Postdoctoral Associate position under the the new quantitative genetics faculty member to the Plant Breeding section, Dr. Kelly Robbins, at Cornell in July, 2018.

This document is dedicated to the Hagermann family.

ACKNOWLEDGEMENTS

Funding of this research was provided by in part by the USDA National Needs Fellowship, which provided tuition and a stipend from fall 2013 through spring 2016. Funding for fall 2016 and spring 2017 was provided by the Biology teaching assistantship for BIOMG 1350 course, Introductory Biology: Cell and Developmental Biology supervised by Dr. Tim Huffaker. Dr. Mark Sorrells provided research assistantship funding for the summer of 2017. The section of Plant Breeding and Genetics provided a teaching assistantship for fall 2017 for the PLBRG 2010 course, Plants, Genes, and Global Food Production instructed by Dr. Susan McCouch. The WheatCAP project provided funding for a research assistantship as a data curator for the Triticeae Toolbox (T3) under Dr. Jean-Luc Jannink for spring 2018.

The field trials comprising the phenotypic data for the CNLM population were funded in part by the Hatch Project # 149-447. Genotyping was funded by the Wheat Coordinated Agricultural Project (WheatCAP).

I acknowledge and thank my committee members, chair Dr. Mark Sorrells, minor committee member Dr. Jason Mezey, and Dr. Jean-Luc Jannink for their valuable input to this research and this document.

I would like to acknowledge Jesse Poland's research group at Kansas State University for their contribution to genotyping of the CNLM materials. I want to thank the International Wheat Genome Sequencing Consortium for pre-publication access to IWGSC RefSeq v1.0 and Martin Mascher at the Leibniz-Institute of Plant Genetics and Crop Plant Research IPK (collaborating with IWGSC) for providing chromosome centromere positions.

I express gratitude to my adviser, Dr. Mark Sorrells, for allowing me to pursue my research interests, along with the magnitude of field plots we planted in the

Master population in 2015. I am also grateful to the Cornell small grains staff, particularly David Benscher and James Tanaka, who were vital in implementing, collecting and processing the materials used to build the CNLM dataset.

I would like to thank Roberto Jesus Lonzano Gonzalez del Valle for the suggestion of using the annotated coding sequences to identify homeologous genes. I give thanks to Uche Godfrey Okeke, Itaraju Brum and Marnin Wolfe for engaging conversation of various quantitative genetics topics and the whiteboard scribbling that ensued.

I want to thank my parents Elaine and Dan; without their love and support I could not have achieved such an accomplishment. To my friends in New York, New Mexico and beyond, you have been instrumental in my success and well being. Finally, I want to give special thanks to Amy Zimmermann, who helped to copy edit this dissertation and kept me afloat through the hardest parts of my degree. She is a true partner, and my best friend... even though she gives weak high fives.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

1.1 Significance

Identification, characterization of, and selection for interactions between allelic genomic regions has been paramount in the exploitation of heterosis in crop species. However, interactions between homeologous genes in allopolyploids have been paid relatively less attention despite their obvious analogy to dominance effects in diploid hybrids. To my knowledge, I provide the first attempt at the capitalization of an inherent characteristic of allopolyploids in order to provide a method for breeders to identify, select for and fix favorable homeoallelic interactions across subgenomic syntenic regions on a genome-wide scale.

1.2 Introduction

Whole genome duplication events are ubiquitous in the plant kingdom. The impact of these duplications on angiosperm evolution was not truly appreciated until the ability to sequence entire genomes elucidated their omnipresence (Soltis et al. 2009). Gene duplication is known to be a primary driver of evolution by providing the raw genetic material for gene diversification through sub- and neofunctionalization (Haldane 1933; Ohno 1970). Haldane (1933) postulated that single gene duplication allowed one copy to diverge through mutation while metabolic function was maintained by the other copy. Ohno (1970) reintroduced this hypothesis, and it has since been validated both theoretically (Ohta 1987; Walsh 1995; Lynch and Conery 2000), and empirically (Blanc and Wolfe 2004; Duarte et al. 2005; Liu,

Baute, and Adams 2011; Assis and Bachtrog 2013). The duplicated gene hypothesis does not, however, generally explain the apparent advantage of duplicating an entire suite of genes. The necessity of genetic diversity for plant populations to survive and adapt to divergent or changing environments may help to explain this pervasive phenomenon.

The need for gene diversity can become more immediate in plants than in animals, where the latter can simply migrate to "greener pastures" when conditions become unfavorable. Plants lack substantial within generation mobility and must therefore change gene expression to cope with changing environmental conditions. Seed dispersion allows plants to move across generations, but this distance is dependent on seed structure and is often limited. As a consequence, plant populations often experience a greater rate of inbreeding as siblings will generally germinate in close proximity to one another. Plant populations can also become isolated, further restricting gene flow. Therefore, allelic diversity alone may be insufficient to maintain gene pathway diversity with high inbreeding pressure. Duplication of important genes can allow the plant to maintain multiple functional copies, such that loss or fixation of a deleterious allele does not preclude the plant population from thriving, as duplicate copies are available.

Many species maintain gene diversity through alternate splicing, but this has been shown to be less common in plants than in other eukaryotes (Nagasaki et al. 2005). Whole genome duplication can generate the raw materials for the maintenance of genetic diversity (Wendel 2000; Adams and Wendel 2005). Gault (2018) demonstrated that similar sets of duplicated genes were preserved in two related genera, *Zea* and *Tripsacum*, millions of years after a shared paleopolyploidization event. This conserved pattern in purifying selection suggests that, at least for some

genes, there is a clear advantage to maintaining two copies.

Whole genome duplication events occur either through duplication of the same genome (autopolyploidy) or the union of two closely related genomes (allopolyploidy). These duplication events are less frequent in animal species, supposedly due to the inability of most animals to produce fertile offspring without a genetically suitable partner of the opposite sex (Muller 1925; Orr 1990). Many plants, on the other hand, can typically reproduce sexually in isolation due to the expression of both male and female sex organs on the same plant. The ability to self pollinate is particularly important for survival of a newly formed species if the hybridization event is rare, where the allopolyploid will find itself instantly sexually isolated.

The union of two complete, yet evolutionarily divergent, genomes during the formation of an allopolyploid can introduce manifold new gene pathways that can specialize to specific tissues or environments (Blanc and Wolfe 2004). Mac Key (1970) postulated a trade off between new-creating (allogamous) and self preserving (autogamous) mating systems, where allopolyploids favor self pollination to preserve diverse sets of alleles across their subgenomes. As such, an allopolyploid may be thought of as an immortalized hybrid, with heterosis fixed across syntenic subgenome regions (Ellstrand and Schierenbeck 2000; Feldman et al. 2012). While still hotly debated, evidence is mounting that allopolyploids exhibit a true heterotic response across homeologous regions as traditional hybrids have demonstrated across homologous regions (Wendel 2000; Adams and Wendel 2005; Chen 2010; Chen 2013).

Birchler et al. (2010) note that newly synthesized allopolyploids often outperform their subgenome progenitors, and that the heterotic response appears to be exaggerated in wider inter-specific crosses. This seems to hold true even within

species, where autopolyploids from wider crosses tend to exhibit higher vigor (Bingham et al. 1994; Segovia-Lerma et al. 2004). Complementation of deleterious recessive alleles (or pseudo-dominance) has long been the primary explanation of the heterotic response (Stuber et al. 1992; Cockerham and Zeng 1996). However, Birchler et al. (2010) indicate evidence against this, where purging deleterious alleles has increased the additive value of inbred maize parents but has not reduced the heterotic response observed in the hybrid (Duvick 1999). Complementation also seems an unlikely driver of a heterotic response in allopolyploids, as the inbred subgenome progenitors would supposedly need functional copies of these genes to survive.

The overwhelming prevalence of allopolyploidy to autopolyploidy in plant species (Soltis and Soltis 2009) may suggest that it is the increase in allelic diversity *per se* that is the primary driver for this observed tendency toward genome duplication. Instead of allowing genes to change function after a duplication event, alleles may develop novel function prior to their reunion during allopolyploidization. The branched gene networks of the allopolyploid may provide the organism with the versatility to thrive in a broader ecological landscape than those of its subgenome ancestors (Mac Key 1970; Ellstrand and Schierenbeck 2000; Osborn et al. 2003).

Statistical deviations from additivity (i.e. interactions) are important contributors to genetic variation, particularly in hybrids. Homologous gene interactions, also known as dominance, are deviations from an additive expectation due to different allele combinations at one locus. Non-homologous gene interactions, commonly referred to as epistasis, are deviations from an additive expectation due to different allele combinations at two or more loci (Fisher 1919; Cheverud and Routman

1995). When epistasis occurs between non-homologous loci with similar function, such as across orthologs or paralogs, these interactions are comparable to dominance effects. If interactions occur between homeologous orthologs on separate subgenomes of an allopolyploid, should we call this epistasis or dominance?

In classical hybrid variety production, divergent sets of alleles are intentionally isolated into heterotic groups and then brought back together to form the hybrid. This establishes heterozygosity (by decent) at all loci to form a homogeneous population. The union of two divergent suites of genes during the formation of an allopolyploid also results in a homogeneous population, but heterozygosity is established across homeologs rather than homologs. Diploid hybrids lose heterozygosity through segregation in following filial generations, but heterozygosity across homeologous genes is subsequently preserved through selfing in the allopolyploid (Mac Key 1970; Ozkan, Levy, and Feldman 2001; Abel, Möllers, and Becker 2005). Allelic interactions contribute to dominance variance in the diploid hybrid, whereas homeoallelic interactions will be present as part of the additive by additive epistatic variance in an inbred allopolyploid population. As such, allopolyploids may be thought of as an immortalized hybrid (Ellstrand and Schierenbeck 2000; Feldman et al. 2012), although it is not yet clear that these exhibit a true heterotic response as traditional hybrids have demonstrated.

Common wheat (*Triticum aestivum*) is a staple allopolyploid crop, accounting for about 20% of the calories consumed worldwide. Hexaploid wheat has undergone two allopolyploid events, resulting in three genomes, denoted A, B and D. The A genome ancestor, *Triticum uratu*, still exists today and was an early domesticate from the fertile crescent important in the neolithic revolution (Dvořák et al. 1993). The B genome ancestor (an *Aegilops spp.*) is believed to have since gone extinct

(Blake et al. 1999), but the tetraploid formed by these two genomes, *Triticum turgidum*, is still cultivated today primarily as emmer wheat. The D genome comes from a goat grass, *Aegilops tauschii*, which was may have been incorporated in a single hybridization event as recently as 10,000 years ago (Salamini et al. 2002). However, recent evidence based on sequence divergence of the D genome from the A and B genome has suggested a much earlier D genome incorporation around 400,000 years ago (Marcussen et al. 2014). Other evidence suggests that limited gene flow into the D genome may have occurred after the polyploidization event, but appears to be from a single lineage (Wang et al. 2013). As a result, the D genome has significantly lower genetic variation than either the A or B genome. The gene diversity provided by these three genome ancestors may explain why allohexaploid wheat has adapted to such wide spread cultivation from its source in the fertile crescent to significant crop production around the globe (Dubcovsky and Dvořák 2007).

In this report, I investigate the importance of homeologous interactions in allohexaploid wheat. Three data sets are used for this investigation: the Cornell small grains Master winter wheat breeding population (CNLM), a biparental winter wheat recombinant inbred line population formed from a cross between varieties 'Caledonia' and 'NY91017-8080' (RIL), and a CIMMYT wheat data set (W-GY) from Crossa et al. (Crossa et al. 2010). The CNLM dataset is the primary focus and is featured in all chapters.

CHAPTER 2

DATA SETS AND SOFTWARE

2.1 Cornell Master - CNLM Population

The CNLM dataset consists of 8,692 phenotypic records of 1,447 soft winter wheat inbred lines evaluated at four locations near Ithaca, NY from 2007 to 2016, representing 26 environments (Table 2.1). These phenotypic evaluations serve primarily as a first round of selection for grain yield and other agronomic traits before relatively few are selected for replicated regional trials around New York State. Lines are introduced and then removed after they are deemed either fit for advanced field trials or to be discarded or recycled in the breeding program. As such, this dataset is unbalanced in nature. Most lines were not replicated within a given trial (i.e. year and location), but various check varieties were used throughout these years and are typically replicated several times within a given trial.

Field plots 1.5 m by 3 m in size were planted with 100 g of seed in September or October of the year prior to the harvest year. Data was recorded for four agronomic traits: grain yield (GY), plant height (PH), test weight (TW) and heading date (HD). Plots were harvested for grain with a plot combine after physiological maturity and oven dried to a grain moisture of approximately 12%. Dried grain was cleaned, weighed and measured for moisture content using a grain moisture analyzer (GAC 2100, Dickey-John). GY was standardized to a uniform grain moisture of 12%. PH was measured as the distance from the ground to the top of the grain head at full extension. TW is used as a measure of grain quality and was measured as the mass of a volume of grain (g L^{-1}) using the grain moisture analyzer (GAC 2100, Dickey-John) which corrects TW for moisture content. HD is a

Table 2.1: Number of phenotypic observations for each location across 10 years.

Year	Helfer	Ketola	McGowan	Snyder
2007	0	0	246	0
2008	0	432	0	426
2009	409	0	431	0
2010	311	304	0	305
2011	310	319	318	0
2012	307	306	0	302
2013	127	130	338	0
2014	232	232	0	233
2015	469	464	461	0
2016	427	425	0	428

proxy for flowering time and was defined as the number of Julian days until 50% of the primary grain heads have extended out of the boot.

The data set initially consisted of 1,552 lines. Thirty one lines from 2007 were not harvested for GY, nor were they genotyped, and were dropped from the data set. Because GY was of primary interest from a breeding perspective, plots that were not harvested or had missing values for GY were dropped, resulting in 9,090 plots with GY measurements. This caused two additional lines with missing GY measurements to be dropped from the dataset. Due to the reasonable size of the dataset, small physical area of most trials, lack of replication within environment for most lines and the availability of genetic markers, raw plot observations were used. No attempt was made to correct plot level data for various spatial effects or otherwise. Preliminary results had indicated relatively high genomic prediction accuracy, suggesting that spatial correction, such as an AR1 $\times$ AR1 row column autocorrelation structure, would be unlikely to reduce error variance drastically and would complicate analysis. Instead, 59 plots that included breeder comments about bad seed or significant damage to the plot, via animal or otherwise, were removed from the dataset. Observations outside of a four standard deviation in-

Table 2.2: Means (μ) and standard deviations (σ) of four traits in the CNLM population.

	units	μ	σ
GY	kg ha^{-1}	5315.20	1015.76
PH	cm	90.84	11.99
HD	Julian days	151.64	3.87
TW	g L^{-1}	74.95	3.09

terval from the grand mean of uncorrected GY phenotypic observations were also removed to account for any significant undocumented damage, grain spillage or other undocumented mistakes. This included two observations that were deemed too high, and 20 observations that were deemed too low.

Observations of 11 lines lacking at least one phenotypic record in at least two separate trials and 61 lines that failed to pass the marker genotype filtering procedure described below were also removed from the dataset. This resulted in 8,692 phenotypic observations of 1,447 lines across 26 environments, representing 96.6% of the plots with grain yield measurements. HD was not recorded for the 246 observations from 2007, and PH was not recorded for the 840 observations from 2009. Two additional plots were missing PH measurements from the Ketola location, a probable height recording mistake of 2 meters made in 2008 that was set to missing, and another in 2010 which was mysteriously not recorded. While most of the genotypes were directly from the Cornell small grains breeding program, a few varieties and breeding lines from other breeding programs that had been genotyped and evaluated were not excluded from the dataset as long as they met the previous criteria. This included 75 lines from The Ohio State University wheat breeding program and 93 lines from the Michigan State University wheat breeding program that were part of the Allele Based Breeding initiative, among other lines from various breeding programs.

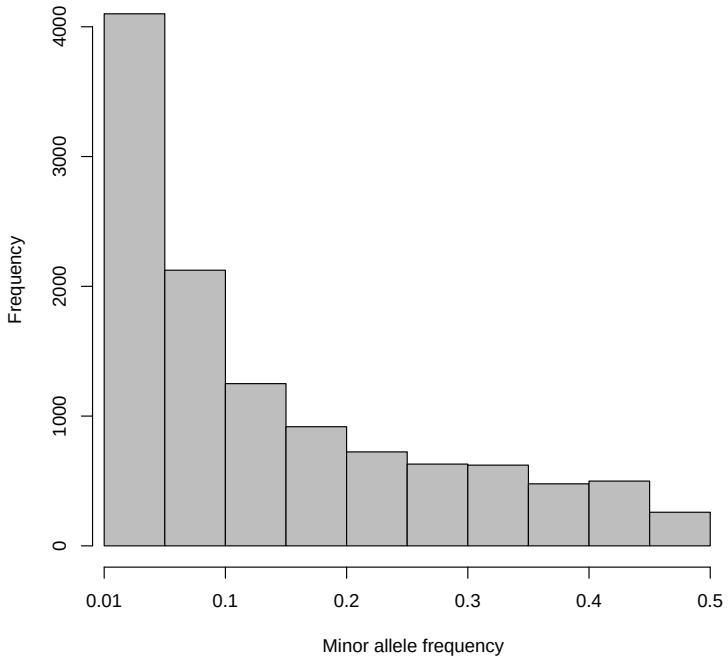


Figure 2.1: Distribution of minor allele frequencies for 11,604 GBS markers in the CNLM population.

Genotyping by sequencing (GBS) libraries (Elshire et al. 2011) of 1,521 CNLM were developed using the protocol described by Poland et al. (2012) at Kansas State University, and subsequently sequenced at the Genomic Diversity Facility at Cornell University. Genotyping calls were accomplished using standard parameters of the Tassel 5.0 GBS v2 Pipeline (Glaubitz et al. 2014) and were aligned to the International Wheat Genome Sequencing Consortium (IWGSC) RefSeq v1.0 wheat genome sequence of 'Chinese Spring' (IWGSC 2018, accepted). Following Poland et al. (Poland et al. 2012), 64 bp sequence tags containing no more than three Single Nucleotide Polymorphisms (SNPs) per tag were included to increase the likelihood of obtaining subgenome specific markers. Only markers with a minor allele frequency of at least 0.01 (Figure 2.1), no more than 30% missing scores, and

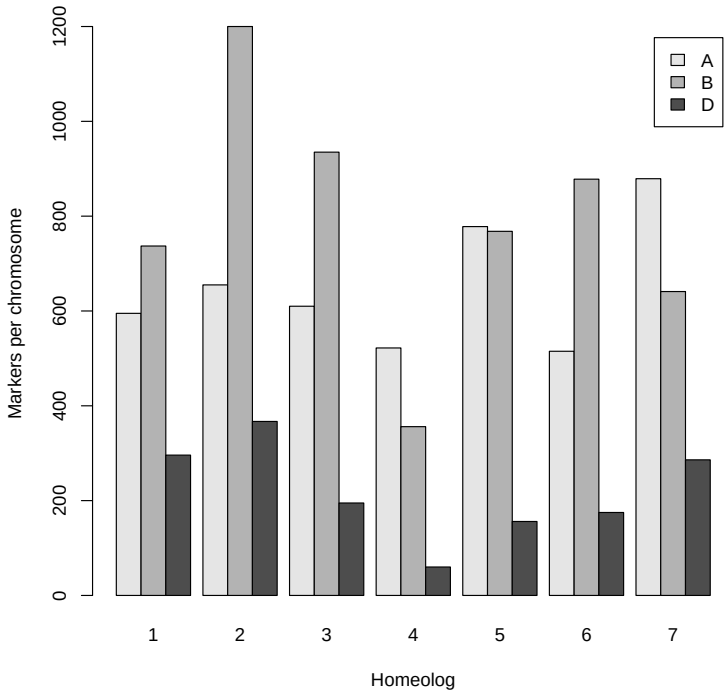


Figure 2.2: Distribution of 11,604 GBS markers on the 21 wheat chromosomes comprised of 7 homeologs of three subgenomes, A, B and D, for the CNLM population.

no more than 10% heterozygous calls were kept for the following analyses. Then individuals with greater than 20% heterozygous calls and individuals with more than 50% missing genotype calls were excluded from the dataset. The process was repeated iteratively, starting by filtering on markers until the number of markers and genotypes converged. This resulted in 11,604 available GBS markers distributed throughout the three subgenomes (Figure 2.2). Of the 61 lines removed, 60 were removed due to missing marker information and 1 was removed due to high heterozygosity in the final iteration. The 11 lines with a single phenotypic observation and the two lines without grain yield observations were subsequently removed to produce genotypic information for the 1,447 lines.

Table 2.3: Estimated genetic correlation of traits with additive (below diagonal) and independent genetic relationships (above diagonal). Standard deviations of scaled traits estimated with a realized additive covariance between individuals and assuming independence are shown in parentheses on the diagonal, respectively.

	GY	PH	TW	HD
GY	(0.29, 0.36)	-0.39	-0.24	0.16
PH	-0.44	(0.72, 0.65)	0.31	0.05
HD	-0.05	0.11	(0.44, 0.44)	-0.28
TW	-0.04	0.3	-0.22	(0.5, 0.49)

Marker scores were coded using $\{-1, 0, 1\}$ for homozygous major allele, heterozygous and homozygous minor allele, respectively. Categorical marker imputation was done independently for each chromosome using random forest imputation via the R package 'missForest' (Stekhoven and Bühlmann 2011) which relies on the R package 'randomForest' (Liaw and Wiener 2002). Random forest has been shown to be effective for genotype imputation in wheat (Rutkoski et al. 2013). To allow all individuals to be considered completely inbred, the remaining heterozygous calls ($< 2\%$ of all marker scores) were conservatively replaced with the population mode for that marker (i.e. the homozygous major allele, -1). Marker scores were then converted to $\{0, 1\}$ coding for presence of the minor allele.

Genetic correlations of traits were estimated in a multivariate model fit (Table 2.3). This was accomplished by treating genotypes as independent, or having a realized additive covariance structure calculated from genetic markers.

2.2 CIMMYT - W-GY Population

The W-GY wheat dataset of 599 historical wheat lines from the CIMMYT Global Wheat Breeding program reported in Crossa et al. (2010) was included in Chapter 3 due to its importance in genomic prediction of an inbred population with non-additive variation (Crossa et al. 2010; Martini et al. 2016). The W-GY dataset consists of genotypic values of all lines for grain yield in each of four environments. The genetic correlations between these environments ranges from -0.19 to 0.66 and can be found in Martini (2016). As performance between these environments is not highly correlated, I refer to grain yield performance in each environment as a trait. The dataset was used in its entirety with one exception. Of the 1,279 available DArT markers, only the 1,188 with known chromosomal positions as denoted by Crossa et al. (Crossa et al. 2010) were utilized in this study. This information was required to know which markers belonged to which subgenome such that subgenome specific relationship matrices could be calculated. The W-GY dataset was not used for the analysis in Chapters 4 or 5 due to the lack of known chromosomal positions. I attempted to obtain positions for these markers by aligning them to the RefSeq v1.0 wheat genome sequence with BLAST+, but found positions for less than half of the markers. Therefore analyses requiring known positions were restricted to the CNLM dataset.

2.3 NY91017-8080×Caledonia - RIL Population

The bi-parental recombinant inbred line (RIL) population was formed from a cross between two Cornell soft winter wheat lines, NY91017-8080 and Caledonia. The parental lines are semi-dwarfs, each containing one *Reduced Height*-1 (*Rht-1*)

dwarfing allele lacking in the the other parent line. *Rht-1* mutants are insensitive to the plant hormone gibberellic acid (GA), resulting in a semi-dwarf plant stature (Peng et al. 1999; Pearce et al. 2011). The semi-dwarf plant architecture is less susceptible to lodging, particularly under high nitrogen conditions, and was key to implementation of high yielding varieties produced during the Green Revolution. Plants with both dwarfing alleles are agronomically inviable due to extremely short stature. Caledonia contains a GA-insensitive *Rht-1D* allele, d, on chromosome 4D and a wildtype *Rht-1B* allele, B, on chromosome 4B, while NY91017-8080 has a GA-insensitive *Rht-1B* allele, b, on chromosome 4B and the wild type *Rht-1D* allele, D, on chromosome 4D.

The RIL population consisting of 192 individuals was planted in single row plots (i.e. headrows) in Ithaca NY and measured for plant height in 2008. The population was screened for loci influencing plant height on chromosomes 4B and 4D using genotyping by sequencing (GBS) markers. The markers with the lowest p-value on the short arms of 4B and 4D were used to indicate the *Rht-1* gene in this study. Only individuals with homozygous genotype calls for both loci were included to test for epistasis. This resulted in 19 double dwarfs ($bbdd$), 51 D genome semi-dwarfs ($BBdd$), 35 B genome semi-dwarfs($bbDD$), and 53 tall ($BBDD$), for a total of 158 individuals. It appeared that the Caledonia parent plant used in the cross was heterozygous for the D genome dwarfing allele, resulting in the 1:2 segregation ratio for the $d : D$ alleles, and was confirmed by the genotype call for that plant.

2.4 Software and File Descriptions

Models were fit using Restricted Maximum Likelihood (REML) for variance component estimation with the software ASReml (Gilmour 1997) implemented in R (Butler 2009). BLAST+ (Camacho et al. 2009) was used for coding sequence alignment, Tassel 5.0 GBS pipeline v2 (Glaubitz et al. 2014) along with the 'bwa' alignment tool (Li and Durbin 2009) were used for aligning GBS markers to the reference genome. All additional computation, analyses and figures were made using base R (R Core Team 2015) implemented in the Microsoft Open R environment 3.3.2 (Microsoft 2017) unless noted otherwise. Figures 4.2 and 4.1 were created using the 'tikz' package (Tantau 2018) for LaTeX. Figures 4.4, 5.4 and 5.8 were made with the 'circlize' R package (Gu et al. 2014). The R package 'xtable' (Dahl 2016) was used to generate LaTeXtables in R. The 'txtplot' R package (Bornkamp 2012) was also used regularly to visualize data during various data analyses, and merits recognition as a text friendly R plotting function. This document was compiled using the TexLive 2017 (`https://www.tug.org/texlive/`) distribution of LaTeX.

All data has been included as supplementary files for transparency and reproducibility. Phenotypes for the CNLM population are included in the file 'pheno.txt'. Marker information and imputed marker scores for the CNLM population are included in files 'snpInfo.txt' and 'snpMatrix.txt', respectively. Best Linear Unbiased Predictors (BLUPs) for whole and subgenome additive effects (GEBVs and SGEBVS, respectively), as well as non-additive whole and subgenome interaction effects can be found in the 'effectTable.txt' file.

A list of homeologous genes can be found in 'homeoGeneList.txt'. The file 'HomeoMarkerSet.txt' contains non-unique marker sets anchored to each homeologous gene set. Unique marker sets used for the analysis in Chapter 4 can be found

in 'uniqueHomeoMarkerSet.txt', 'WithinMarkerSet.txt', 'AcrossMarkerSet.txt' for the Homeo, Within and Across marker sets, respectively. Marker and marker interaction estimates and p-values for the Homeo set can be found in 'twoWayInteractions.txt' and 'threeWayInteractions.txt' for two- and three-way marker interactions, respectively.

Results from the arm test are not included due to the sheer volume of data associated with each pair. However, data can be requested from Nicholas Santantonio at ns722@cornell.edu.

Phenotypes and genotypes used in the RIL population are included in the 'NY8080Cal.txt' file. Genotype and phenotype data for the W-GY population can be found in the 'BGLR' package of R (Campos and Pérez Rodriguez 2015), and marker chromosome information can be found in Crossa et al. (2010)

CHAPTER 3

PREDICTION OF SUBGENOME ADDITIVE AND INTERACTION EFFECTS IN ALLOHEXAPLOID WHEAT

3.1 Introduction

The availability of affordable genome-wide markers has sparked a revolution in selection on additive variation through the use of genomic prediction models. The additive genetic merit of an individual can be estimated as the sum of its additive marker effects to produce a genomic estimated breeding value (GEBV) (Meuwissen, Hayes, and Goddard 2001) . When the number of markers is large, marker effects are typically considered random and normally distributed such that only one parameter need be estimated. Alternatively, the additive genetic covariance between individuals can be estimated from the same genome-wide markers and used to predict additive genetic values of individuals based on relatedness (Nejati-Javaremi, Smith, and Gibson 1997; VanRaden 2008). These models are equivalent for prediction under the same set of assumptions (Garrick 2007; VanRaden 2008; Strandén and Garrick 2009). Genomic prediction models have since become popular for their ability to predict the performance of genotyped individuals with no phenotypic observations. Selections on unobserved individuals allows for reduction in the cost of phenotyping and breeding cycle time, increasing the rate of genetic gain (Goddard and Hayes 2007; Heffner, Sorrells, and Jannink 2009; Jannink, Lorenz, and Iwata 2010; Heslot, Jannink, and Sorrells 2015).

The potential utility of genome-wide markers has also drawn renewed interest in non-additive genetic variation in recent years (Vitezica, Varona, and Legarra 2013; Martini et al. 2016; Jiang and Reif 2015; Huang and Mackay 2016; Jiang et al.

2017). Genomic prediction models that use genome-wide markers can incorporate non-additive genetic components to obtain better estimates of individual performance than based on additivity alone (Technow et al. 2012; Vitezica, Varona, and Legarra 2013; Jiang and Reif 2015; Akdemir and Jannink 2015; Akdemir, Jannink, and Isidro-Sánchez 2017; Wolfe et al. 2016). In outcrossing species such as maize, prediction of dominance effects is key to harnessing heterosis in unobserved hybrids (Technow et al. 2012). In inbred species, additive by additive epistatic effects have been shown to significantly increase genomic prediction accuracy (Crossa et al. 2010; Martini et al. 2016). Epistatic effects can be added to the prediction model by extending Henderson's (1985) method of expected epistatic covariance estimation to marker based covariance estimation (Jiang and Reif 2015; Martini et al. 2016).

The use of genome-wide markers has allowed for the partitioning of genetic variance to specific units of chromatin, previously infeasible with phenotypes alone (Bernardo and Thompson 2016). Allopolyploids have been traditionally treated as diploids because they undergo disomic inheritance (Mac Key 1970), such that the contribution of each subgenome to the genetic variance is ignored. By assigning markers to each subgenome, an additive genetic covariance based on each subgenome can be calculated. Using these covariances in a genomic prediction model, the genetic merit of an allopolyploid individual can be assigned to each of its subgenomes. These subgenomic estimated breeding values (SGEBV) can then be used to identify parents with complementary subgenome effects for crossing.

Under Hardy Weinburg equilibrium, subgenomes segregate independently, and realized estimates of additive covariance of individuals based on each subgenome will be independent. However, this does not generally hold true in breeding pro-

grams, where population structure from non-random mating is inherent. As a consequence, the estimates of additive covariance between individuals based on different subgenomes will not be independent, potentially leading to confounding of effects from each subgenome and problems partitioning variance reliably. In an attempt to circumvent this obstacle, I present an approach for removing the largest sources of genetic variance (i.e. population structure) using singular value decomposition of the matrix of marker scores.

I demonstrate this methodology using two allohexaploid wheat data sets, CNLM (Section 2.1) and W-GY (Section 2.2) described in Chapter 2.

3.2 Materials and Methods

3.2.1 Subgenome additive effects

To illustrate, I begin with a linear mixed model depicting environments as fixed effects and genotypes as random.

$$y_{ijk} = \mu + E_i + G_j + \varepsilon_{ijk} \tag{3.1}$$

where μ is the population mean, E_i and G_j are the fixed environmental and random genetic effects, respectively, of the j^{th} genotype evaluated in the i^{th} environment, and ε is the error associated with the k^{th} observation. Using matrix notation, model (3.1) can be rewritten as

$$\mathbf{y} = \mathbf{1}\mu + \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{g}_G + \boldsymbol{\varepsilon} \tag{3.2}$$

where $\mathbf{1}_n$ is a vector of ones, $\mathbf{X}$ is the design matrix, and $\boldsymbol{\beta}$ is the vector of fixed environmental effects. $\mathbf{Z}$ is the incidence matrix linking observations in the vector $\mathbf{y}$ to their respective genotype effects, in the vector $\mathbf{g}_G$. Normality was assumed for genotype effects and the residuals, where $\mathbf{g}_G \sim N(0, \sigma^2_G \mathbf{K}_G)$ and $\boldsymbol{\varepsilon} \sim \mathcal{N}(0, \sigma^2 \mathbf{I})$. The genetic covariance, $\mathbf{K}_G$, can be derived from the expectation (or coefficient) of co-ancestry between individuals from a pedigree (Henderson 1985), or by an empirical estimation of the realized genetic relationship calculated with genome-wide markers (VanRaden 2008). When genome-wide markers are used to estimate $\mathbf{K}_G$, the genomic prediction model initially suggested by Nejati-Javaremi, Smith and Gibson (1997) and Meuwissen, Hayes and Goddard (2001) is obtained.

Given an $n \times m$ matrix, $\mathbf{M}$, of m markers scored as reference allele counts (i.e. $\{0, 1, 2\}$) for n individuals, method I of Van Raden (2008) finds the genetic relationship $\mathbf{K}$ as,

$$\mathbf{K} = c^{-1}(\mathbf{M} - \mathbf{P})(\mathbf{M} - \mathbf{P})^{\mathrm{T}} + 0.01\mathbf{I} \tag{3.3}$$

where $\mathbf{P} = \mathbf{1}_n \otimes 2\mathbf{p}^{\mathrm{T}} - \mathbf{1}$, $c = 2\mathbf{p}^{\mathrm{T}}(\mathbf{1} - \mathbf{p})$ and $\mathbf{p}$ is the vector of allele frequencies. The small coefficient of 0.01 was added to the diagonal to recover full rank after centering the matrix, such that $\mathbf{K}_G$ is invertible.

I use allohexaploid wheat to illustrate, but this method is easily truncated to allotetraploids, or extended to higher level allopolyploids. If we allow the total genetic effect, G_j, to be decomposed into individual additive effects for each subgenome, such that $G_j = A_j + B_j + D_j$, the following model is obtained.

$$\mathbf{y} = \mathbf{1}\mu + \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{g}_A + \mathbf{Z}\mathbf{g}_B + \mathbf{Z}\mathbf{g}_D + \boldsymbol{\varepsilon} \tag{3.4}$$

In model 3.4, each subgenome is allowed to have its own additive genetic variance and covariance between individuals, such that $\mathbf{g}_A \sim \mathcal{N}(0, \sigma^2_A \mathbf{K}_A)$, $\mathbf{g}_B \sim \mathcal{N}(0, \sigma^2_B \mathbf{K}_B)$ and $\mathbf{g}_D \sim \mathcal{N}(0, \sigma^2_D \mathbf{K}_D)$. The realized additive genetic covariances for each subgenome, $\mathbf{K}_A$, $\mathbf{K}_B$ and $\mathbf{K}_D$, are estimated using only markers corresponding to the respective subgenome, and calculated as described above.

Subgenome epistatic interactions

Following Henderson (1985), the epistatic covariance of individuals can be calculated as the Hadamard product of the component covariance matrices. Martini et al.(2016) provide a proof of Henderson's method using genome-wide markers to estimate the additive by additive covariance matrix, $\mathbf{H}$. An additional linear kernel can then be added for an additive by additive epistatic interaction term, I_j, once the additive covariance is estimated to obtain the following model.

$$\mathbf{y} = \mathbf{1}\mu + \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{g}_G + \mathbf{Z}\mathbf{g}_I + \boldsymbol{\varepsilon} \tag{3.5}$$

where $\mathbf{g}_I \sim \mathcal{N}(0, \sigma^2_{g_I} \mathbf{H})$. Martini (2016) does not report scaling $\mathbf{H}$ by the sum of the joint marker variances. Although this scalar does affect the parameter estimate, it does not not affect the model fit or prediction. I use the square of the sum of the marker variances as an approximation to avoid calculating all joint marker variances. In this study, $\mathbf{H}$ is calculated as

$$\mathbf{H} = \mathbf{K} \odot \mathbf{K} - c^{-2}(\mathbf{W} \odot \mathbf{W})(\mathbf{W} \odot \mathbf{W})^{\mathrm{T}} \tag{3.6}$$

where $\mathbf{W} = \mathbf{M} - \mathbf{P}$.

The additive by additive epistatic interaction term, $\mathbf{g}_I$, can also be decomposed into across subgenome interactions and within subgenome epistatic interactions such that $I_j = AB_j + AD_j + BD_j + I_j^-$, where AB_j, AD_j and BD_j are the subgenome interaction effects and I_j^- is the remaining epistatic effects due to within subgenome epistasis. Since no markers are shared across subgenomes, subgenome interaction covariances can be estimated by extending Henderson's method of using the Hadamard product of their component covariance matrices (Martini et al. 2016). These interactions can then be incorporated in the following model.

$$\mathbf{y} = \mathbf{1}\mu + \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{g}_A + \mathbf{Z}\mathbf{g}_B + \mathbf{Z}\mathbf{g}_D + \mathbf{Z}\mathbf{g}_{AB} + \mathbf{Z}\mathbf{g}_{AD} + \mathbf{Z}\mathbf{g}_{BD} + \mathbf{Z}\mathbf{g}_{ABD} + \boldsymbol{\varepsilon} \tag{3.7}$$

where $\mathbf{g}_{AB} \sim \mathcal{N}(0, \sigma^2_{g_{AB}}(\mathbf{K}_A \odot \mathbf{K}_B))$, $\mathbf{g}_{AD} \sim \mathcal{N}(0, \sigma^2_{g_{AD}}(\mathbf{K}_A \odot \mathbf{K}_D))$, $\mathbf{g}_{BD} \sim \mathcal{N}(0, \sigma^2_{g_{BD}}(\mathbf{K}_B \odot \mathbf{K}_D))$ and $\mathbf{g}_{ABD} \sim \mathcal{N}(0, \sigma^2_{g_{ABD}}(\mathbf{K}_A \odot \mathbf{K}_B \odot \mathbf{K}_D))$. The three-way interaction is included here for biological completeness, but was found to be estimated on the boundary (i.e. zero) for all traits, and was therefore dropped from further analyses.

3.2.2 Accounting for population structure

Under Hardy Weinberg equilibrium, subgenomes segregate independently, such that for subgenome effects, $\mathrm{Cov}(A, B) = \mathrm{Cov}(A, D) = \mathrm{Cov}(B, D) = 0$ and

$\text{Var}(G) = \text{Var}(A) + \text{Var}(B) + \text{Var}(D)$. A breeding program, however, intentionally violates this assumption, and therefore may contain significant population structure. Price et al. (2006) demonstrated that the first k largest principal components (PCs) of the kinship matrix can be used to control for population structure in genome-wide association studies, and its use has since become wide spread. Because most realized estimates of additive covariance are proportional to $\mathbf{MM}^{\mathrm{T}}$, singular value decomposition of $\mathbf{M}$, instead of $\mathbf{MM}^{\mathrm{T}}$, can be used to separate the population structure from the entire matrix of marker scores before it is divided into its subgenome components. This is accomplished by first extracting the first k principal components in the $n \times k$ matrix $\mathbf{Q}$. The marker matrix can then be reconstructed by setting the first k singular values of the diagonal matrix to zero and multiplying to produce a matrix $\widetilde{\mathbf{M}}$ with the population structure removed from each subgenome simultaneously.

To illustrate, let $\mathbf{M}$ be the $n \times m$ matrix of m marker scores for n genotypes. Markers can be sorted into their respective subgenome, such that

$$\mathbf{M} = \begin{bmatrix} \mathbf{M}_A & \mathbf{M}_B & \mathbf{M}_D \end{bmatrix} \tag{3.8}$$

$\mathbf{M}$ can be factored using singular value decomposition as follows:

$$\mathbf{M} = \mathbf{UDV}^{\mathrm{T}} \tag{3.9}$$

where $\mathbf{U}$, and $\mathbf{V}$ are unitary matrices of left and right singular vectors, and $\mathbf{D}$ is a diagonal matrix of singular values.

The first k principal components of the marker matrix can be extracted by selecting the first k columns of $\mathbf{U}$ and the first k rows and columns of $\mathbf{D}$ and

multiplying.

$$\mathbf{Q} = \mathbf{U}_{n\times k}\mathbf{D}_{k\times k} \tag{3.10}$$

In a manner similar to Eckart and Young (Eckart and Young 1936), an approximation, $\widetilde{\mathbf{M}}$, of the marker matrix, $\mathbf{M}$ with the first q principal components removed can be reconstructed by setting the first k singular values in $\mathbf{D}$ to zero (denoted $\widetilde{\mathbf{D}}$).

$$\widetilde{\mathbf{M}} = \mathbf{U}\widetilde{\mathbf{D}}\mathbf{V}^{\mathrm{T}} = \begin{bmatrix}\widetilde{\mathbf{M}}_A & \widetilde{\mathbf{M}}_B & \widetilde{\mathbf{M}}_D\end{bmatrix} \tag{3.11}$$

Additive covariance matrices with reduced collinearity can then be calculated for each subgenome from $\widetilde{\mathbf{M}}$ and incorporated into the model as previously described. $\mathbf{Q}$ can then be added to the model as a set of fixed covariates, with slopes $\boldsymbol{\gamma}$, such that the model will now be of the form

$$\mathbf{y} = \mathbf{1}\mu + \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{Q}\boldsymbol{\gamma} + \sum_{l}\mathbf{Z}\mathbf{g}_l + \boldsymbol{\varepsilon} \tag{3.12}$$

for all l genetic terms in the model. Genomic estimated breeding values are then predicted by summing the centered population structure and genetic effects. For this study, a population structure of dimension $k = 5$ was chosen for both the CNLM and W-GY datasets, and used to compare to the $k = 0$ models that do not correct for population structure.

3.2.3 Genomic prediction

To determine the predictability of genetic effects and the variability of variance component estimates, k-fold cross-validation with 5 folds was performed with 10

replications. For each replicate, the set of individuals was randomly split into five groups, with 4 groups of 289 and one of 291. For each fold, records of individuals in the fold were removed (i.e. masked) from the dataset. Each model was subsequently fit with the remaining lines and used to predict the whole genetic effect of the masked lines in the fold. Predictions for all five folds were gathered and correlated to the "true" genetic values once for each replicate. In this way, prediction results are directly comparable between the different models, and not subject to differences in the individuals sampled. The whole genome values were calculated as the sum of the genotypic additive and epistatic effects in the model as previously described. Due to the unbalanced nature of the CNLM dataset, "true" genetic values were calculated as in equation 3.2 but were considered independent with a covariance $\mathbf{K}_G = \mathbf{I}$.

3.3 Results

3.3.1 Model fit and variance components

Model fit was assessed using Akaike's Information Criterion (AIC). Whole genome models tended to have the lowest AIC values, with the exception of the PH and HD traits for the epistatic ABD×ABD models in the CNLM population. When whole genome models had lower AIC values, the comparable subgenome models had only marginally higher AIC values (Tables 3.1 and 3.2). Whole genome predictions between comparable whole genome and subgenome models were correlated at $\rho > 0.999$ or $\rho > 0.993$ for traits within the CNLM and W-GY populations, respectively. This indicates that little, if any, genetic information was lost by splitting the whole

Table 3.1: Table of model fit statistics for whole genome and subgenome prediction models in the CNLM population.

Model	Terms	GY (0.30)	PH (0.73)	TW (0.53)	HD (0.79)
G	log$\mathcal{L}$	-48	2237	1547	6343
	parameters	28	26	28	27
	AIC	153	-4423	-3037	-12631
	G	0.268[a] (12.59)[b]	3.823 (20.75)	1.067 (16.66)	3.9 (21.16)
	R	0.324 (61.86)[c]	0.135 (56.17)	0.2 (60.12)	0.054 (58.76)
G×G	log$\mathcal{L}$	-43	2360	1630	6432
	parameters	29	27	29	28
	AIC	144	-4665	-3203	-12808
	G	0.203 (7.86)	0.889 (6.46)	0.194 (4.47)	1.121 (7.3)
	H	0.018 (3.04)	0.478 (11.95)	0.184 (11.33)	0.451 (11.13)
	R	0.322 (61.39)	0.132 (56.5)	0.195 (60.25)	0.053 (58.98)
ABD	log$\mathcal{L}$	-48	2242	1549	6366
	parameters	30	28	30	29
	AIC	155	-4428	-3037	-12673
	A	0.098 (5.86)	1.28 (8.41)	0.503 (8.24)	0.907 (7.41)
	B	0.153 (6.88)	1.585 (8.7)	0.38 (7.1)	1.534 (9.04)
	D	0.021 (2.47)	0.892 (6.24)	0.181 (4.38)	1.387 (8.03)
	R	0.324 (61.87)	0.135 (56.16)	0.199 (60.12)	0.054 (58.75)
ABD×ABD	log$\mathcal{L}$	-41	2375	1634	6451
	parameters	33	31	33	32
	AIC	149	-4687	-3201	-12839
	A	0.076 (4.66)	0.292 (3.7)	0.079 (2.79)	0.104 (1.71)
	B	0.119 (5.4)	0.429 (4.21)	0.114 (3.44)	0.587 (5.38)
	D	0.015 (1.95)	0.279 (3.31)	0.007 (0.5)	0.664 (5.26)
	AB	0.005 (0.49)	0.007 (0.11)	0.073 (2.38)	0.223 (3.29)
	AD	0.012 (1.36)	0.276 (4.16)	0.073 (2.51)	0.167 (2.49)
	BD	0	0.149 (2.19)	0.034 (1.15)	0.005 (0.08)
	R	0.322 (61.4)	0.132 (56.51)	0.195 (60.26)	0.053 (58.98)

[a] Variance component estimates reported for additive main effects (G, A, B and D) and epistatic interactions (H, A×B, A×D, B×D) are the ratios of the actual variance component to the residual variance component for ease of comparison.
[b] The variance component divided by their respective standard errors are shown in parentheses.
[c] The residual variance components are the actual estimates from the centered and scaled data (refer to Table 2.2 for scaling coefficients) with their associated standard errors in parentheses.

genome into biologically relevant subgenome effects. The lack of perfect correlation is at least partially due to floating point rounding errors during model fitting and summation of genotype effects.

Subgenome additive variance parameter estimates were positive for all models, but subgenome interaction variance parameter estimates were often estimated on

Table 3.2: Table of model fit statistics for whole genome and subgenome prediction models in the W-GY population.

Model	Terms	E1	E2	E3	E4
G	log$\mathcal{L}$	-243	-243	-264	-248
	parameters	2	2	2	2
	AIC	489	490	533	500
	G	0.55[a] (5.47)[b]	0.456 (5.1)	0.295 (4.23)	0.369 (4.69)
	R	0.541 (11.88)[c]	0.568 (12.17)	0.669 (12.73)	0.606 (12.5)
G×G	log$\mathcal{L}$	-222	-242	-249	-233
	parameters	3	3	3	3
	AIC	451	491	504	472
	G	0.426 (3.25)	0.44 (4.39)	0.303 (2.73)	0.314 (3.04)
	H	0.272 (4.59)	0.033 (1.02)	0.325 (4.81)	0.231 (4.34)
	R	0.343 (7.31)	0.532 (9.34)	0.368 (6.61)	0.391 (7.68)
ABD	log$\mathcal{L}$	-242	-242	-264	-247
	parameters	4	4	4	4
	AIC	492	492	536	503
	A	0.241 (3.38)	0.09 (1.9)	0.062 (1.66)	0.097 (2.1)
	B	0.215 (2.88)	0.267 (3.48)	0.222 (3.23)	0.188 (2.89)
	D	0.067 (1.72)	0.093 (1.92)	0.022 (0.93)	0.064 (1.86)
	R	0.547 (11.97)	0.571 (12.28)	0.668 (12.79)	0.613 (12.59)
ABD×ABD	log$\mathcal{L}$	-222	-241	-247	-232
	parameters	7	7	7	7
	AIC	458	495	509	478
	A	0.124 (1.56)	0.07 (1.52)	0.061 (1.15)	0.08 (1.47)
	B	0.269 (2.43)	0.273 (3.25)	0.239 (2.23)	0.138 (1.79)
	D	0.07 (1.26)	0.093 (1.81)	0.041 (0.94)	0.087 (1.74)
	AB	0.219 (3.39)	0.04 (1.33)	0.215 (2.69)	0.167 (2.88)
	AD	0.029 (0.62)	0	0.097 (1.52)	0.007 (0.14)
	BD	0	0	0	0.027 (0.57)
	R	0.349 (7.38)	0.526 (9.57)	0.366 (6.74)	0.409 (7.92)

[a] Variance component estimates reported for additive main effects (G, A, B and D) and epistatic interactions (H, A×B, A×D, B×D) are the ratios of the actual variance component to the residual variance component for ease of comparison.
[b]The variance component divided by their respective standard errors are shown in parentheses.
[c] The residual variance components are the actual estimates from the centered and scaled data (refer to Crossa et al. (2010) for scaling coefficients).

the boundary (i.e. near zero). Variance parameters estimated on the boundary were thus considered to be exactly zero. Shifts in variance component importance were seen when the epistatic terms were added in the model. For example, for the TW and E1 traits in the CNLM and W-GY populations, respectively, the A subgenome component was the largest in the additive only model, but was reduced to less than that of the B subgenome component in the epistatic model.

Additive variance components were generally reduced in epistatic models compared to additive only models, but this reduction in additive variance was accompanied by non-zero subgenome interaction components. The B subgenome contributed the greatest amount of additive variance in the epistatic ABD×ABD models for all traits accept HD. While the D subgenome variance component was far smaller than the A subgenome component for GY and TW in the CNLM population, it was comparable to the A subgenome component for all traits in the W-GY population.

The A×B component was particularly important for the W-GY traits, E1, E3 and E4, as well as the HD and TW in the CNLM population. The A×D component also featured prominently for the PH and TW traits in the CNLM population. The B×D component appeared to be less important, having the largest effect for PH. No epistatic terms were significantly greater than zero for the E2 trait in the W-GY population. Addition of epistatic interactions resulted in a significant likelihood ratio test at $p < 10^{-6}$ for all traits except GY, which was significant at $p < 10^{-2}$. Despite the significant addition of epistatic terms, additive only GEBVs were highly correlated with whole genome predictions from the epistatic models, at $\rho \geq 0.988$ for the CNLM population and $\rho \geq 0.869$ for the W-GY population. A model containing the three-way subgenome epistatic term was fit for all traits, but estimates of the three-way interaction variance parameter were zero for all traits.

The distributions of variance component estimates from repeated sub-sampling of the data during k-fold cross-validation were centered near the point estimate from the full model fits. These distributions were either as wide ($\approx$ 2 standard errors from the center) or tighter than expected based on the standard error from the full model fit (Figures 3.1 and 3.2). Standard errors were generally larger for epistatic variance components relative to their magnitude than additive vari-

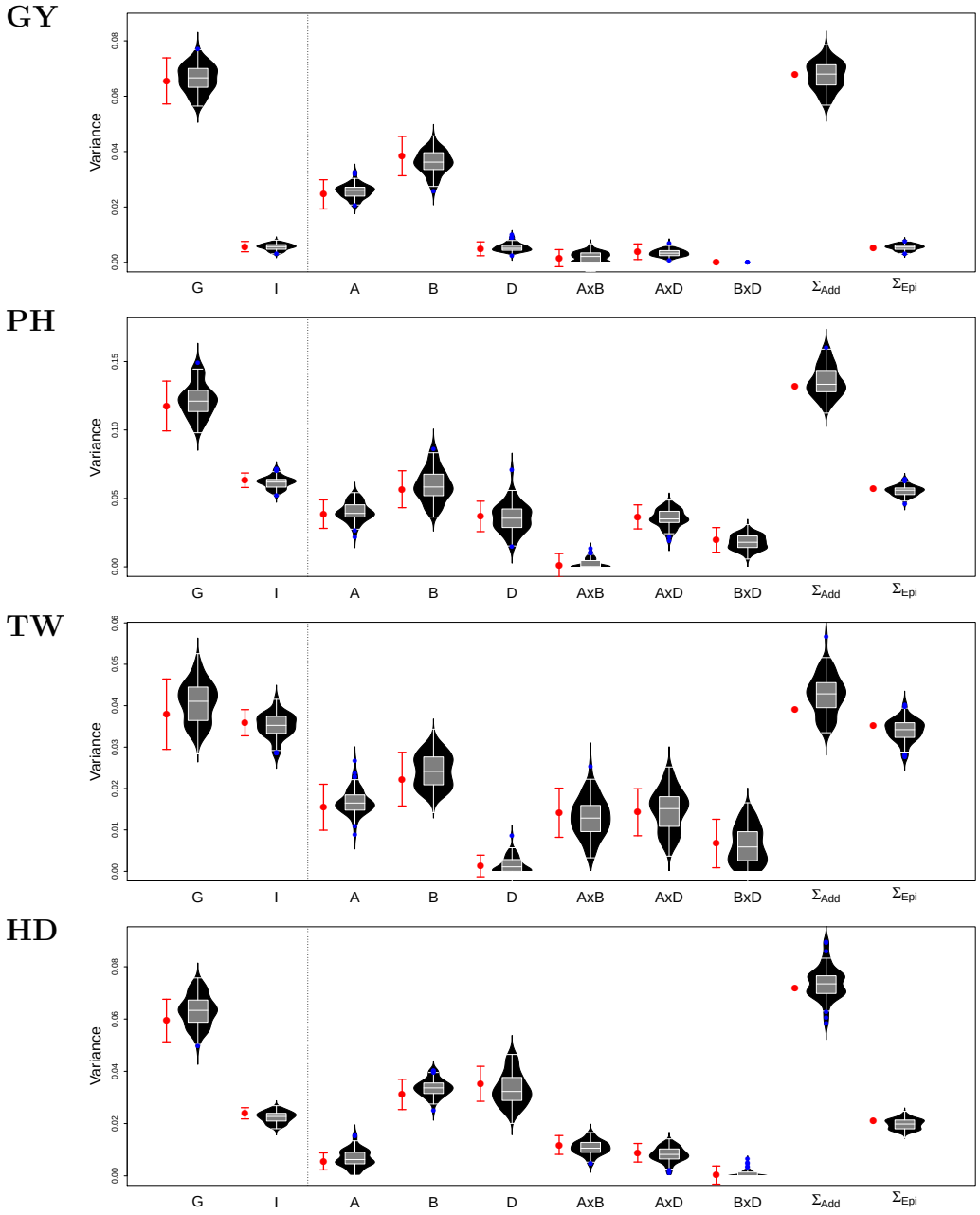


Figure 3.1: Estimates and standard errors of variance components for four traits in the CNLM populations from the full model (red) compared to the sampling distribution of variance component estimates from the cross-validation scheme (black violins). G×G and ABD×ABD models are shown to the left and right of the dotted line, respectively. The sum of the additive and interaction variance components is also shown for the ABD×ABD model.

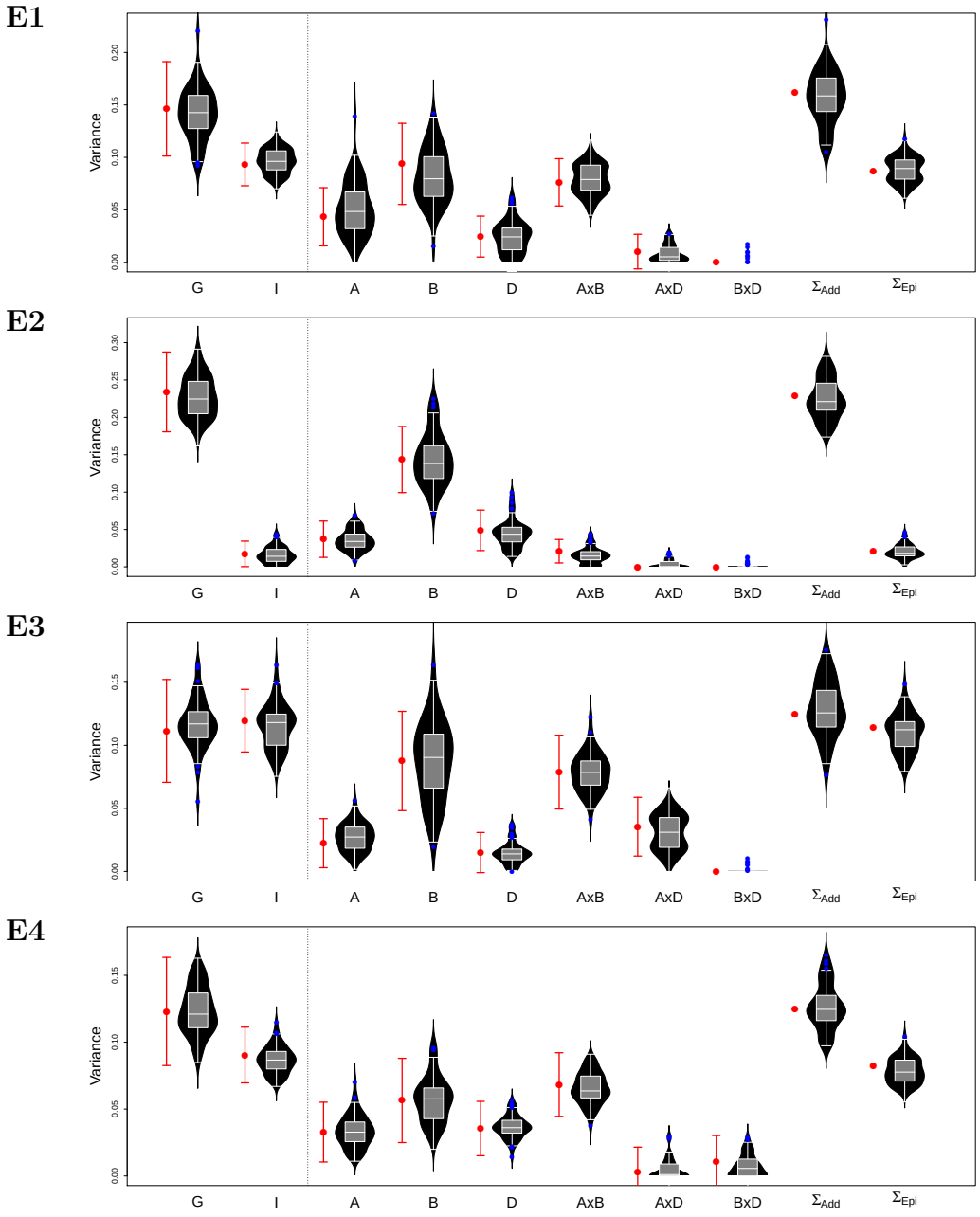


Figure 3.2: Estimates and standard errors of variance components for four traits in the W-GY populations from the full model (red) compared to the sampling distribution of variance component estimates from the cross-validation scheme (black violins). G×G and ABD×ABD models are shown to the left and right of the dotted line, respectively. The sum of the additive and interaction variance components is also shown for the ABD×ABD model.

ance components. Standard errors relative to their respective parameter estimates tended to be larger for all terms in models with more estimated variance parameters (Tables 3.1 and 3.2).

3.3.2 Subgenome additive effects

Subgenome estimated breeding values (SGEBVs) were moderately correlated with the whole genome effect, but weakly correlated with one another (Tables 3.3 and 3.4). The individuals with the highest SGEBV for one subgenome never had the highest SGEBV for the other two subgenomes, and were often not in the top 95% quantile of the population based on the other two subgenomes (Figures 3.3 and 3.4). For example, the individuals with the highest A, B and D SGEBV for GY in the CNLM population ranked 43^{rd}, 39^{th} and 60^{th} for the whole genome effect, respectively. In contrast, the individual with the best A SGEBV for GY ranked 1067^{th}, 952^{nd} for the B and D SGEBV, respectively. The individual with the highest B SGEBV for GY ranked 221^{th} and 1393^{rd} for the A and D SGEBV, respectively. The individual with the highest D SGEBV for GY ranked 347^{th} and 123^{rd} for the A and B subgenome, respectively. The individual with the highest whole genome GEBV for grain yield ranked 6^{th}, 22^{nd} and 519^{th} for the A, B and D SGEBV, respectively. In several cases, the top individual based on a SGEBV was not in the top 95% quantile based on their whole genome effect, particularly in the W-GY population.

Table 3.3: Correlation of whole genome and subgenome additive effects in the CNLM population. Correlations of additive random effects without correcting for population structure are shown above the diagonal, while correlations of effects correcting for populations structure using the first $k = 5$ PCs is shown below the diagonal.

	GY				PH				TW				HD			
	G	A	B	D	G	A	B	D	G	A	B	D	G	A	B	D
G		0.80	0.82	0.48		0.74	0.68	0.60		0.78	0.74	0.50		0.60	0.74	0.70
A	0.75		0.36	0.22	0.73		0.15	0.25	0.75		0.26	0.25	0.60		0.20	0.14
B	0.83	0.31		0.31	0.70	0.19		0.17	0.73	0.20		0.13	0.71	0.18		0.26
D	0.46	0.18	0.26		0.59	0.23	0.14		0.49	0.23	0.12		0.65	0.14	0.14	

Table 3.4: Correlation of whole genome and subgenome additive effects in the W-GY population. Correlations of additive random effects without correcting for population structure are shown above the diagonal, while correlations of effects correcting for populations structure using the first $k = 5$ PCs is shown below the diagonal.

	E1				E2				E3				E4			
	G	A	B	D	G	A	B	D	G	A	B	D	G	A	B	D
G		0.83	0.80	0.51		0.69	0.84	0.57		0.68	0.90	0.50		0.73	0.80	0.52
A	0.83		0.41	0.28	0.65		0.31	0.25	0.74		0.36	0.07	0.63		0.34	0.09
B	0.79	0.39		0.20	0.91	0.39		0.27	0.91	0.45		0.37	0.86	0.33		0.21
D	0.57	0.36	0.24		0.57	0.22	0.32		0.48	0.28	0.25		0.63	0.22	0.30	

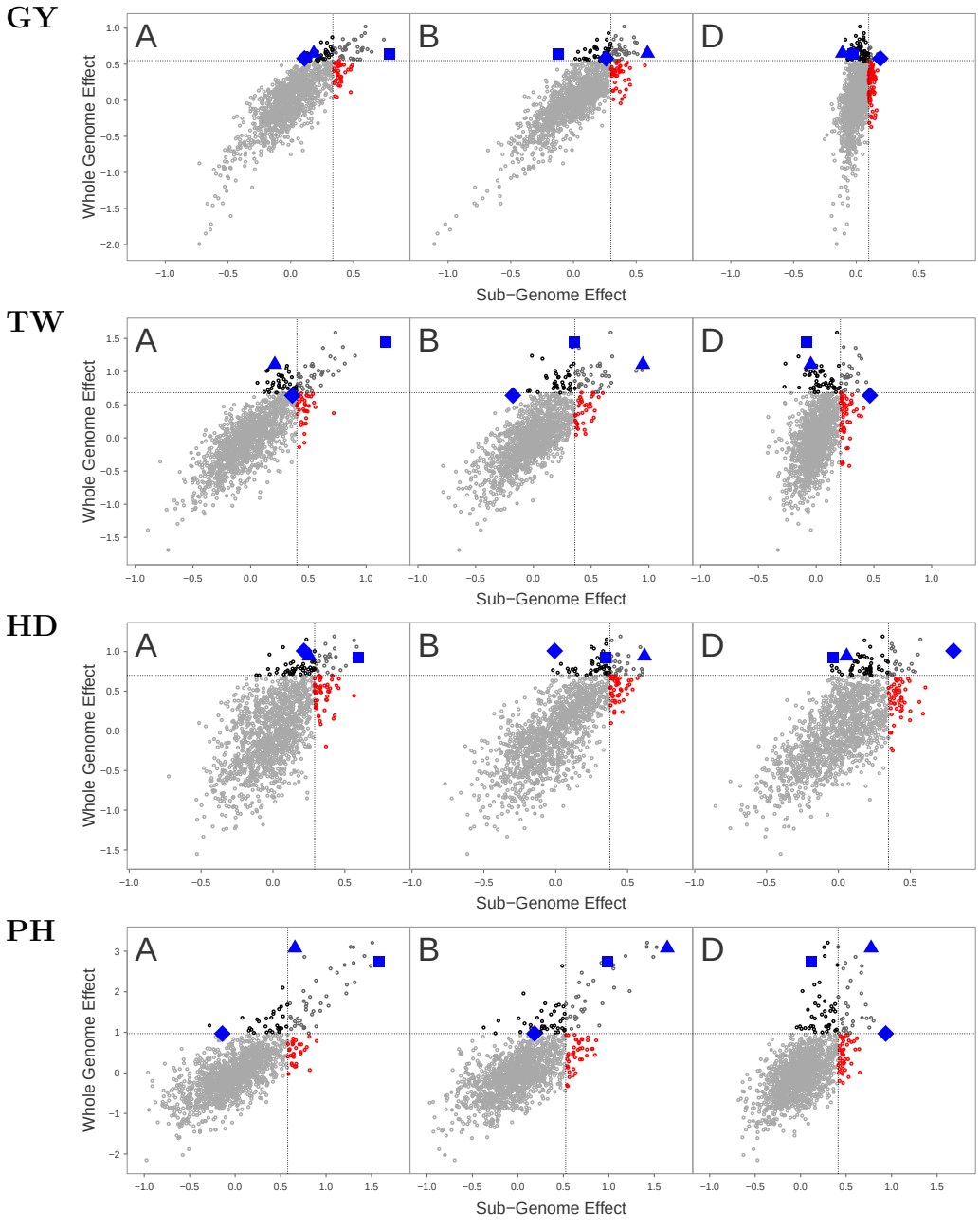


Figure 3.3: Plot of whole genome additive effects (GEBV) by subgenome additive effects (SGEBV) for four traits in the CNLM populations. The dotted line indicates the 95% quantiles for whole or subgenome effects. Blue squares, triangles and diamonds indicate the line with the highest SGEBV for each of the A, B and D subgenomes, respectively.

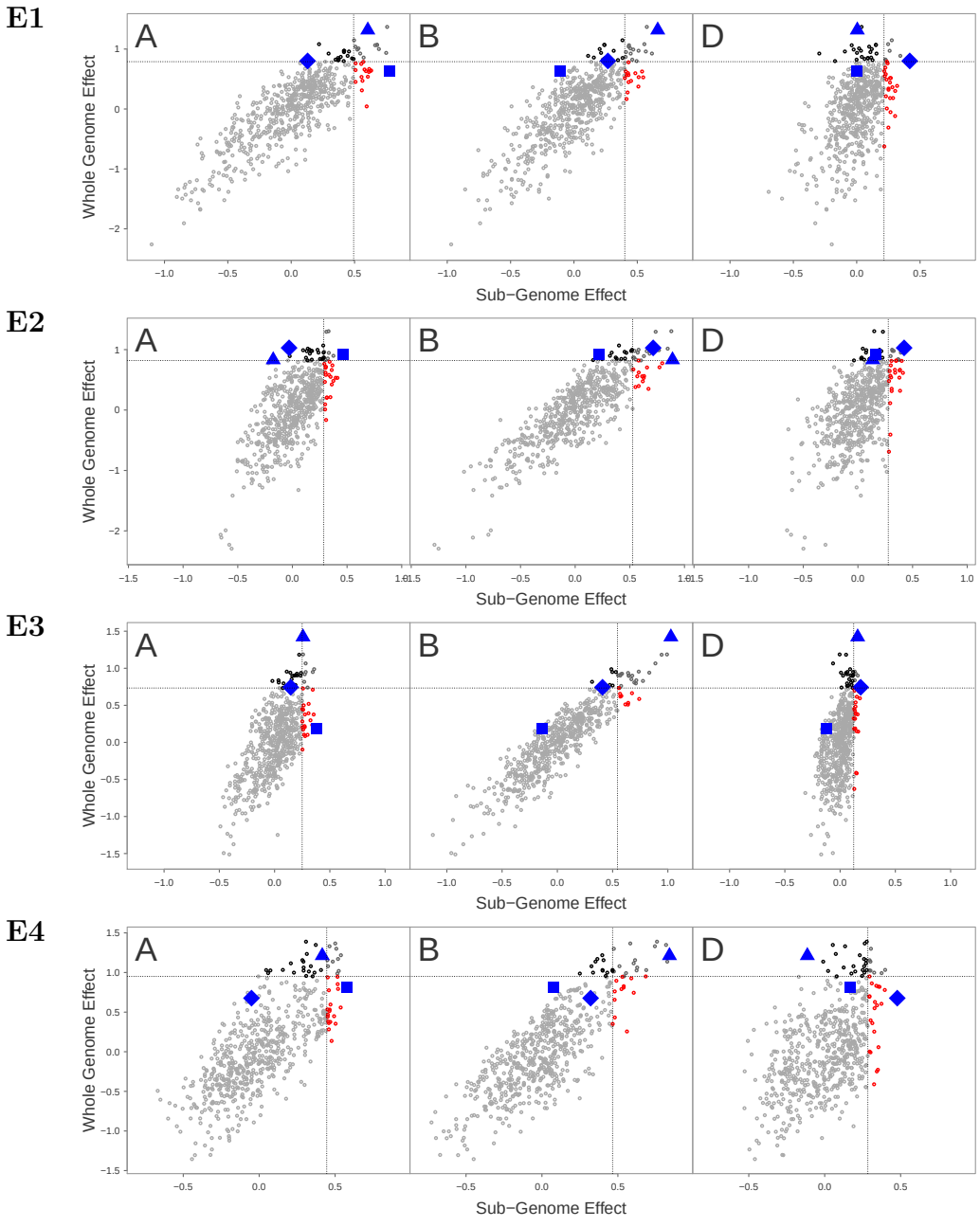


Figure 3.4: Plot of whole genome additive effects (GEBV) by subgenome additive effects (SGEBV) for four traits in the W-GY populations. The dotted line indicates the 95% quantiles for whole or subgenome effects. Blue squares, triangles and diamonds indicate the line with the highest SGEBV for each of the A, B and D subgenomes, respectively.

3.3.3 Prediction accuracy

Table 3.5: Table of genomic prediction accuracies for eight traits in the CNLM (GY, PH, TW and HD) or W-GY (E1, E2, E3, E4) populations with $k = 0$ and $k = 5$.

CNLM	k	GY	PH	TW	HD
G	0	0.601[a] (0.008)[b]	0.559 (0.007)	0.515 (0.010)	0.664 (0.009)
ABD		0.600 (0.008)	0.557 (0.008)	0.514 (0.011)	0.679 (0.007)
G×G		0.604 (0.008)	0.637 (0.004)	0.576 (0.010)	0.712 (0.008)
ABD×ABD		0.603 (0.008)	0.638 (0.005)	0.569 (0.011)	0.720 (0.006)
G	5	0.600 (0.009)	0.558 (0.007)	0.514 (0.011)	0.663 (0.010)
ABD		0.600 (0.009)	0.556 (0.008)	0.513 (0.011)	0.678 (0.008)
G×G		0.602 (0.008)	0.624 (0.005)	0.560 (0.010)	0.701 (0.008)
ABD×ABD		0.602 (0.007)	0.618 (0.005)	0.557 (0.010)	0.708 (0.006)
W-GY	k	E1	E2	E3	E4
G	0	0.501 (0.010)	0.493 (0.016)	0.356 (0.008)	0.457 (0.010)
ABD		0.492 (0.012)	0.481 (0.023)	0.346 (0.010)	0.449 (0.011)
G×G		0.568 (0.010)	0.494 (0.017)	0.396 (0.013)	0.520 (0.010)
ABD×ABD		0.549 (0.011)	0.484 (0.023)	0.393 (0.015)	0.509 (0.013)
G	5	0.502 (0.010)	0.491 (0.017)	0.354 (0.007)	0.458 (0.010)
ABD		0.495 (0.011)	0.475 (0.024)	0.345 (0.010)	0.453 (0.011)
G×G		0.526 (0.010)	0.491 (0.017)	0.381 (0.007)	0.493 (0.011)
ABD×ABD		0.520 (0.012)	0.475 (0.023)	0.373 (0.013)	0.486 (0.012)

[a] Mean genomic prediction accuracy across ten replicates of five fold cross validation.

[b] Standard deviation of prediction accuracy across ten replicates are shown in parentheses.

Including epistasis kernels significantly improved genomic prediction accuracy for all traits except GY and E2 (Table 3.5). Subgenome models had either comparable or slightly lower mean prediction accuracy than whole genome models for all traits except HD, for which subgenome models had superior accuracy. The variability in the prediction accuracy based on the individuals sampled was either the same (GY and TW) or lower (PH and HD) for the epistatic models compared to the additive models in the CNLM population, but was similar in the W-GY population (Table 3.5). The variability in prediction accuracy was increased for

the subgenome models compared to the whole genome models in the W-GY population for some traits (E2 and E3), but was either the same or decreased in the the CNLM population.

3.3.4 Adjustment for population structure

The first two principal components explained 17% and 19% of the variance of M in the CNLM and W-GY populations, respectively, indicating that some population structure exists in both populations (Supplementary Figure 3.7). The correlation of additive genetic covariance estimates between individuals based on the three subgenomes declined as PCs were removed from $\mathbf{M}$, but appeared to level out between 5 to 10 PCs (Figure 3.5). Correlation of whole genetic effects between additive models, G and ABD, for $k = 0$ and $k = 5$ was ≥ 0.999 and ≥ 0.996 for the CNLM and W-GY populations respectively. Whole genome effect correlations were lower between epistatic models G×G and ABD×ABD, with coefficients of ≥ 0.998 in the CNLM population and ≥ 0.980 in the W-GY population.

Removing population structure with $k = 5$ reduced most of the SGEBV effect correlation coefficients by up to 0.06 in the CNLM population, but there was one instance in which one correlation coefficient increased from 0.14 to 0.19 between A and B SGEBVs for PH (Table 3.5). This was not the case for the W-GY population, where many of the SGEBV effect correlations increased by up to 0.21.

Variance components generally decreased as k was increased from 0 to 10 (Figure 3.6). Ranks of additive variance components relative to one another were stable for most traits, while epistatic variance components were more sensitive to changes in k. Significant epistatic variance component rank changes occurred for the PH,

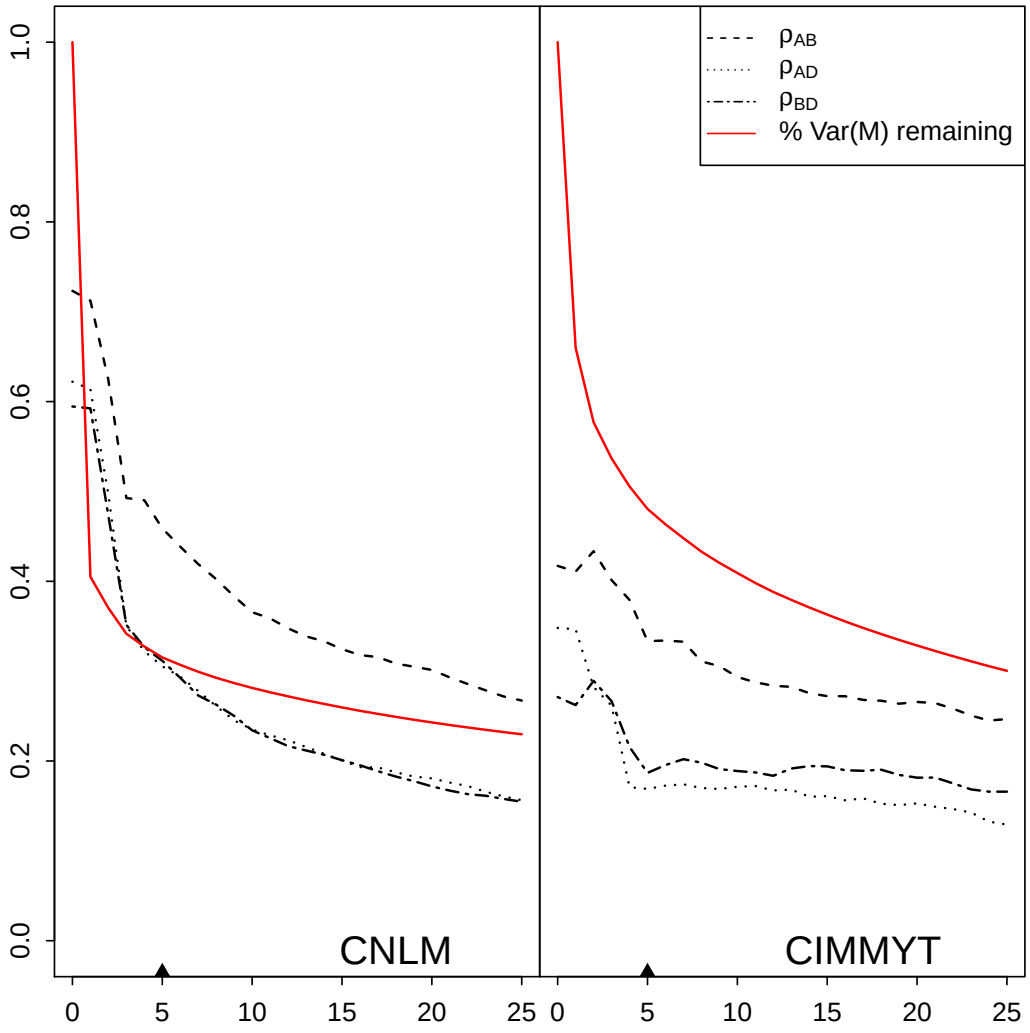


Figure 3.5: Correlation coefficient, ρ, of off-diagonal elements of estimated additive covariance matrices $\mathbf{K}_A$, $\mathbf{K}_B$ and $\mathbf{K}_D$. The percent genotype marker variance remaining in the marker matrix after removing k dimensions is shown in red. The chosen population structure dimension $k = 5$, is indicated by a ▲.

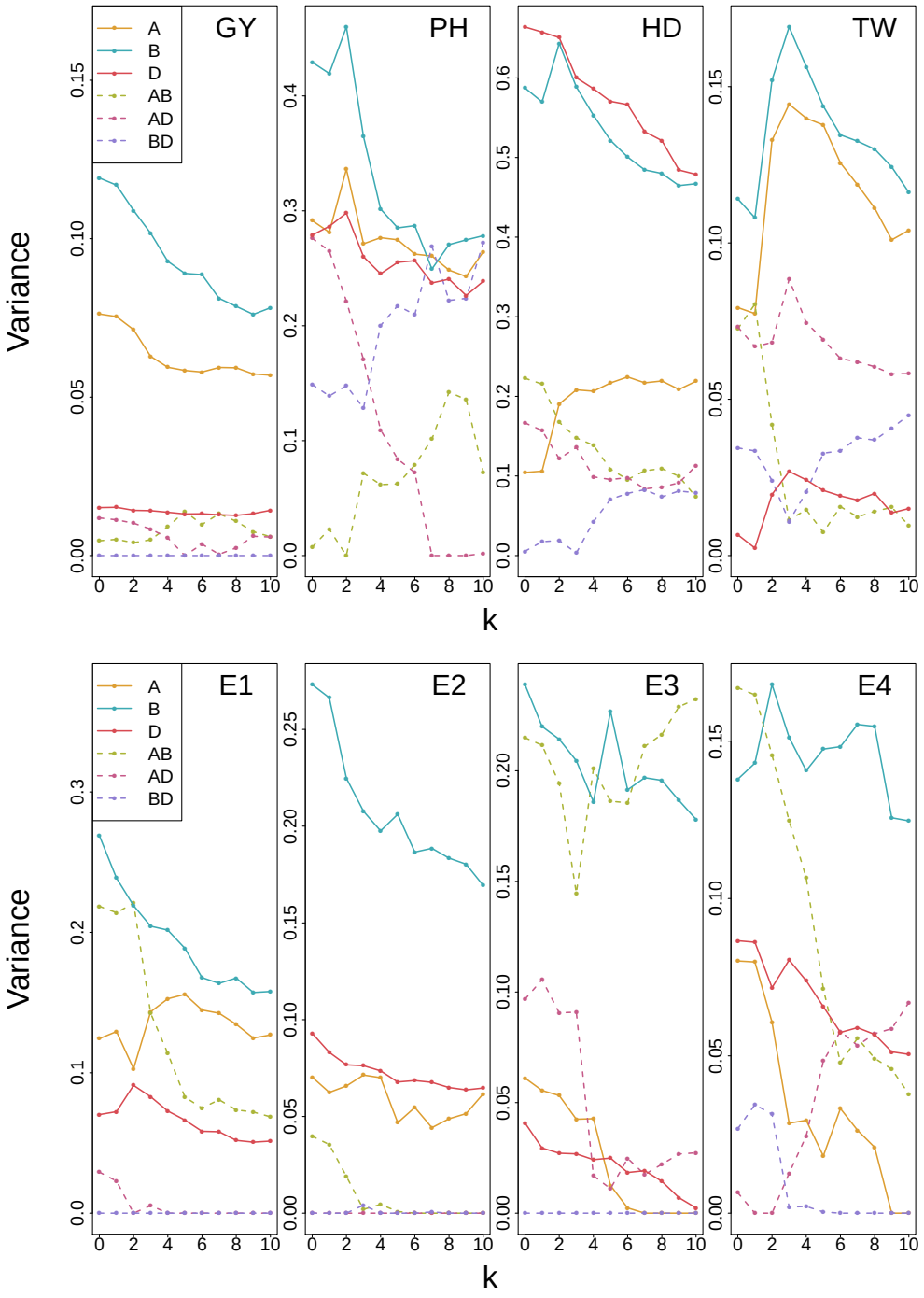


Figure 3.6: Subgenome additive and interaction variance parameter estimates from the ABD×ABD model correcting for population structure with $k \in \{0, 1, \ldots, 10\}$ principal components as fixed effects. Models were fit with four traits for the CNLM population and four traits for the W-GY population.

TW and E4 traits. For PH, the A×D term was comparable in magnitude with the additive variance components for A and D when $k = 0$, but declined as k increased. The reduction in A×D variance for PH was accompanied by an increase in both the A×B and B×D terms. Similarly, a decline in A×B variance was followed by an increase in B×D for TW and A×D for E4 as k was increased.

Correlations of variance component estimates were calculated from the average information matrix for models $k \in \{0, 1, \ldots, 10\}$ (Supplementary Figures 3.8 and 3.9). Correlations between subgenome additive variance estimates were generally low (0.2-0.4), while correlations of subgenome interaction variance estimates were high (0.8-0.95), and correlations between the two were moderate (0.4-0.6). Despite a small reduction in the correlation of SGEBVs as k was increased from 0 to 5, little reduction in variance component estimate correlations was observed as k was increased from 0. Generally, correlations of additive variance parameter estimates were slightly reduced while correlations between interaction variance parameter estimates increased slightly.

3.4 Discussion

3.4.1 Model fit and variance components

While whole genome models tended to be the most parsimonious, subgenome models are worth consideration because they provide insight into the biology of the allopolyploid organism. Given the stability of variance component estimation and that no genetic information appears to be lost by partitioning the whole genome into its individual subgenome additive effects, such a partition is informative.

The method presented here could be used for any set of independent loci, such as estimating a variance component and breeding value for each chromosome. However, this will become computationally burdensome as the number of variance components to be estimated increases. If the number of variance parameters to estimate is large and the data set is small this may become infeasible. It is also unclear if the estimates from larger numbers of additive kernels would be reliable.

Bernardo and Thompson (2016) assigned a breeding value for each of the 10 maize chromosomes by fitting a single ridge regression model to estimate marker effects. They subsequently summed marker effects by chromosome to produce a breeding value for each chromosome. However, this method does not allow for direct estimation of variance components for each unit of chromatin. By fitting each unit simultaneously, variance attributable to sets of loci will be split, and the sum of the variance estimates should not exceed the total genetic variance. It is unclear what effect LD across chromosomes has on the variance parameters estimated.

Here I assumed that the subgenome effects are independent, but this is clearly not the case. Generally, we can express the genetic variance due to the three subgenomes as $\text{Var}\left(\text{vec}([\mathbf{g}_A\ \mathbf{g}_B\ \mathbf{g}_D])\right) = \mathbf{S} \otimes \mathbf{J}_n \odot \mathbf{K}$, where $\mathbf{S}$ is the subgenome covariance matrix, $\mathbf{J}$ is an $n \times n$ matrix of ones for n genotypes, and $\mathbf{K}$ is the additive relationship matrix for within and across subgenomes. In this report, I have assumed that $\mathbf{S}$ is diagonal with $\mathbf{S}_{ii} = \sigma_i^2$ for the i^{th} subgenome, and $\mathbf{K}$ is a block diagonal with the i^{th} diagonal block represented by the subgenome additive covariance matrix.

$$\mathbf{S} = \begin{bmatrix} \sigma^2_A & 0 & 0 \\ 0 & \sigma^2_B & 0 \\ 0 & 0 & \sigma^2_D \end{bmatrix} \text{ and } \mathbf{K} = \begin{bmatrix} \mathbf{K}_A & 0 & 0 \\ 0 & \mathbf{K}_B & 0 \\ 0 & 0 & \mathbf{K}_D \end{bmatrix} \tag{3.13}$$

An unstructured covariance matrix, $\mathbf{S}$, could be estimated, with correlation coefficients between subgenomes, similar to the way that models are fit for maternal and paternal effects in animals. The subgenome effects would be allowed to have a correlation such that

$$\mathbf{S} = \begin{bmatrix} \sigma^2_A & \sigma_{AB} & \sigma_{AD} \\ \sigma_{AB} & \sigma^2_B & \sigma_{BD} \\ \sigma_{AD} & \sigma_{BD} & \sigma^2_D \end{bmatrix} \text{ and } \mathbf{K} = \begin{bmatrix} \mathbf{K}_A & \mathbf{K}_{AB} & \mathbf{K}_{AD} \\ \mathbf{K}_{AB} & \mathbf{K}_B & \mathbf{K}_{BD} \\ \mathbf{K}_{AD} & \mathbf{K}_{BD} & \mathbf{K}_D \end{bmatrix} \tag{3.14}$$

However, it is unclear what the covariance structure should be between subgenomes (e.g. $\mathbf{K}_{AB}$). If consensus haplotypes from uniquely identifiable sequences could be determined with two or more alleles segregating in at least two subgenomes, a covariance across the subgenomes could be constructed. Polymorphisms that predate speciation would be used to identify the consensus haplotypes, while post speciation polymorphisms would be used to identity the subgenome origin. Individuals would then receive a score based on the number of consensus haplotypes they have in common between two subgenomes. This could prove to be a formidable challenge given the evolutionary time between the subgenome ancestors. The Hadamard product of the two additive covariance matrices is a tempting candidate for these off diagonal blocks, however, this would substitute a correlation coefficient between additive effects in place of an epistatic variance. It is unclear to this author if epistasis variance can be thought of and modeled as a correlation between additive effects.

3.4.2 Genetic architecture

The genetic architecture of grain yield (GY, E1, E2, E3, E4) in the two populations investigated here are markedly similar, despite the divergent genetic backgrounds of the two populations. The CNLM population is primarily comprised of breeding lines and varieties derived from germplasm historically grown in the North East, in contrast to the W-GY population which has a broader pedigree.

The D genome is known to have low genetic diversity due to limited gene flow from a single *Ae. tauschii* lineage after the most recent allopolyploidization event (Wang et al. 2013), estimated to have taken place as recently as 10,000 years ago (Salamini et al. 2002; Marcussen et al. 2014). The International Maize and Wheat Improvement Center (Centro Internacional de Mejoramiento de Maíz y Trigo, CIMMYT) has introgressed some genetic material from the D genome ancestor, *Ae. tauschii*, through the use of synthetically produced hexaploid wheat to increase the genetic diversity of the historically bottle-necked D genome. The higher proportion of D genome variance in the W-GY population may be due to the increased use of wild *Ae. tauschii* in their breeding program, highlighting the merit of the strategy.

Many of the subgenome epistatic variance parameters were estimated at zero, possibly due to a lack of power to detect them. Greater genetic diversity, larger numbers of individuals, and higher allele frequencies would allow for increased power to detect true interactions. Hill et al. (2008) emphasized the effect of low allele frequencies on epistatic interactions, proving that as allele frequencies (and therefore joint frequencies of alleles at two loci) approach zero or one, most of the epistatic variance becomes additive. For example, suppose two loci have a large interaction, such that one pair of alleles is selected. Once one locus becomes

fixed, all remaining variance is due to the presence of the two alleles at the other locus, and becomes strictly additive. The low joint frequency is magnified in the three-way interactions, likely causing the inability to detect any three-way epistatic interaction signal between the three subgenomes.

This is also apparent in the reduction of additive variance components upon the addition of epistatic terms to the model. These components were often estimated to be rather large compared to the additive components, but did not change the final whole genome value drastically. This suggests that the additive terms absorb much of the epistatic variance in the absence of the epistatic kernels.

The A×B epistatic terms were the most important for many of the traits, reflecting the greater genetic variation of these two subgenomes. Subgenome interaction terms including the D genome were notably more important for traits known to have important loci on the D genome. PH is partially governed by two dwarfing genes, *Rht-1D* and *Rht-1B* on 4B and 4D, respectively. These two genes have been shown to exhibit a less than additive epistatic interaction, where the double wildtype is less tall than expected based on the additive effects of the two semi dwarfs from the double dwarf (see section 4.4.1 of Chapter 4). The B×D term was large for PH, particularly after correction for population structure. Population structure is common for these genes, as breeding programs primarily utilize one or the other dwarfing gene to avoid producing double dwarfs during crossing, which are agronomically undesirable.

3.4.3 Selection on SGEBVs

Partitioning genetic variance to the subgenomes of an allopolyploid provides a method for identifying individuals with complementary subgenomes as potential parents for crossing. If we consider the upper 95% quantiles as candidates for parental selection, many of the top candidates based on subgenome breeding values would not be considered candidates based on their whole genome breeding value. When they would be considered, they were typically not the top candidates. The low correlation between SGEBVs highlights the opportunity to identify individuals with complementary subgenomes for crossing. These individuals may or may not be among the top performing selection candidates, demonstrating that the optimum set of crosses are not always between the top performing individuals (Akdemir and Sánchez 2016).

The low correlation and high predictability of SGEBVs suggests that individual subgenomes may be directly manipulated as never before. Prior to the discovery and use of genetic markers to track genomic regions, the phenotype (or some summary statistic thereof) was the only indicator of the genetic structure of a genetically distinct individual. Variety releases still demonstrate this legacy, with phenotypic descriptors that define a new variety as genetically distinct from other similar varieties. One breeding strategy will be selecting parents for crossing that have complementary SGEBVs to increase the potential of transgressive segregation in the resulting offspring. I envision other breeding strategies beyond simply choosing parents with complementary subgenomes, and see an opportunity to weight SGEBVs according to some breeding goal.

For example, a newly formed population could undergo several rounds of genomic selection only on the D genome SGEBVs (i.e. weights of 0, 0 and 1 for the

A, B and D subgenomes respectively) before phenotypic or whole genome selection. Because the D genome contributes the least to the total genetic variance, phenotypic selection on D genome loci is challenging. Selection will act on the largest sources of genetic variance first, potentially leading to fixation of small effect loci in the D genome by drift, while selection acts on the large effect loci on the A and B genomes first. By selecting on D genome SGEBVs, gains can be made to the D genome directly with little to no selection on the A and B genomes, a feat previously impossible with phenotypic selection.

3.4.4 Subgenome interactions

Genomic prediction of GY and E2 did not appear to benefit from including epistatic interactions as it did for the other six traits. This may be due in part to the highly polygenic nature of grain yield, which is the culmination of essentially all functional genetic variants subjected to stress throughout the growth cycle. The E2 trait in the W-GY population has previously been shown to be invariant to the addition of various epistatic terms (Crossa et al. 2010; Martini et al. 2016), and it is unclear why this population does not exhibit non-additive variation in this environment as it does in the others. It may be that important epistatic interactions of GY in the CNLM population are too small to detect or are involved with differing performance across years or locations, such that they are lacking in a model that does not include genotype by environment interactions.

Subgenome epistatic terms increased genomic prediction accuracy equivalent to modeling all pairwise interactions across the subgenomes, suggesting that the most important interacting loci are on different subgenomes. This result is consistent with the observation that newly formed allopolyploids undergo considerable

changes in gene expression, known as genome shock (McClintock 1984). This shock has been suggested to be caused by incompatibilities of genetic pathways across the subgenomes (Comai et al. 2003). Residual subgenome incompatibility may still be affecting the germplasm pool, even thousands of years after the last polyploidization event. Decay of negative gene interactions has been shown to take hundreds to thousands of generations before all interacting genes are lost or silenced (Lynch and Conery 2003).

It is unclear what proportion of this non-additive signal is due to homeoallelic interactions. The proposed method models all pairwise interactions across subgenomes, of which homeoallelic gene interactions are a small minority in number. Smaller homeoallelic regions or homeoallele specific marker sets would need to be constructed to determine the relative importance of these interactions relative to other gene interactions across the subgenomes. The usefulness of the epistatic subgenome interactions is currently unclear and warrants further investigation.

Regardless of the source of the epistasis, I suggest that a breeding scheme should be designed to take advantage of beneficial subgenome interactions. If a suitable training set related to the breeding material can be established, subgenome interactions can be predicted in new, genotyped breeding materials. I suggest that a series of small bi-parental populations be constructed from important contributors to the breeding program, and be used in the development of a training population to balance high genetic diversity and high allele frequencies. This training population will be used to predict SGBEVs and subgenome interactions in individuals formed from new crosses. Individuals that contain favorable interactions can then be selected such that they are fixed in early filial generations. After fixation in a given line, phenotypic, whole genome, or subgenome selection can be used for

further line development until complete homozygosity is reached.

3.4.5 Adjustment for population structure

The efficacy of the proposed method to handle population structure may need to be improved, or a different approach may need to be taken. While this method reduced the correlation of additive genetic subgenome covariance estimates across the three genomes, variance parameter estimate correlations were not drastically reduced. The correlation of subgenome interaction variance parameter estimates tended to increase slightly when accounting for higher levels of population structure, counter to the assumption that removing this structure should result in better estimates of subgenome interactions.

The lower correlation between epistatic models that correct and do not correct for population structure is likely due to removing Q from the marker matrix. Correcting for population structure also had a small, but negative effect on genomic prediction accuracy for epistatic models. The population structure fixed effect predictors are strictly additive and the loss of accuracy may be due to epistasis variance associated with these PCs (i.e. population structure epistasis). Epistatic variance related to these PCs may be recovered by using the squares of the PC scores, although this was not done in this study. At least for the additive models, it appears little to no genetic information is lost using the population structure adjustment proposed here.

Determining the best value for k will be at the crux for implementing this methodology for various traits and populations. The same population may need different values of k for different traits, depending on how the population is struc-

tured. Traits such as PH or HD may have less complex structure than traits such as TW or GY, due to different marker effect distributions and the history of the breeding population. Several methods might be used to determine k empirically from the marker matrix (Patterson, Price, and Reich 2006, e.g.), however, these methods may not capture subtle differences in the population structure of a given trait. I used the first one to k PCs in this study, but there is no reason why we must include all PCs up to some value k. There may be certain PCs that are important for a given trait, and could be tested as fixed effects for inclusion or exclusion.

This method may have better performance in populations with greater degrees of population structure than in the populations presented here. Use of this method for partitioning genetic variance to biologically important sets of chromatin and estimating epistatic interactions will need further testing and validation before widespread use.

3.5 Conclusion

To my knowledge, I provide the first attempt to assign a breeding value to each subgenome of an allopolyploid crop. With estimates of subgenome additive effects, parents with complementary subgenomes can be selected for crossing. Weighted selection of subgenomes using genomic prediction could be key to increasing the diversity of the D genome in wheat germplasm. Direct selection on the D genome may allow targeted introgression from *Ae. tauschii* while mitigating the effects of introducing unimproved alleles. Subgenome additive genetic variances appear to be estimated well, and no genetic information appears to be lost partitioning the

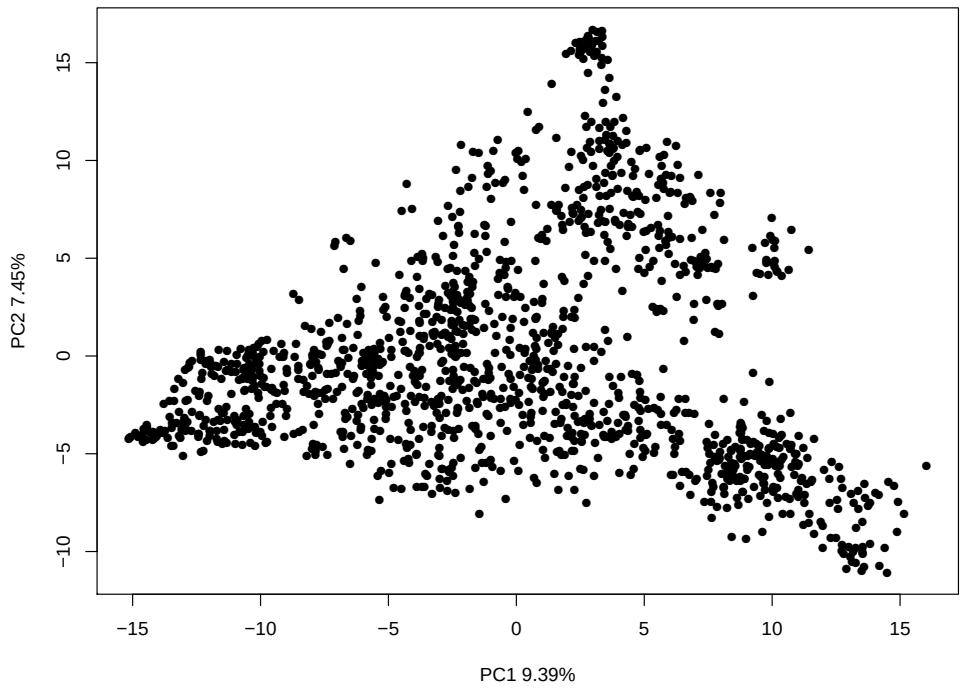
genome into its subgenome components. This demonstrates that partitioning genetic variance to the subgenomes of an allopolyploid can provide useful information for genomic assisted breeding efforts.

Subgenome interactions increase prediction accuracy, but it is unclear how well the epistatic variance is partitioned to the three interaction terms and what proportion of that variance is due to homeologous gene interactions. Because the homeologous interactions make up relatively few of the possible interactions across subgenomes, they may only explain a small portion of the observed epistatic variance. Yet, seeing as how homeologous genes likely operate in the same or similar physiological pathway, the likelihood for interactions between homeologous loci is high. Further research is needed to investigate the efficacy of modeling subgenome interaction terms, and to what degree this is explained by interactions between homeologous orthologs.

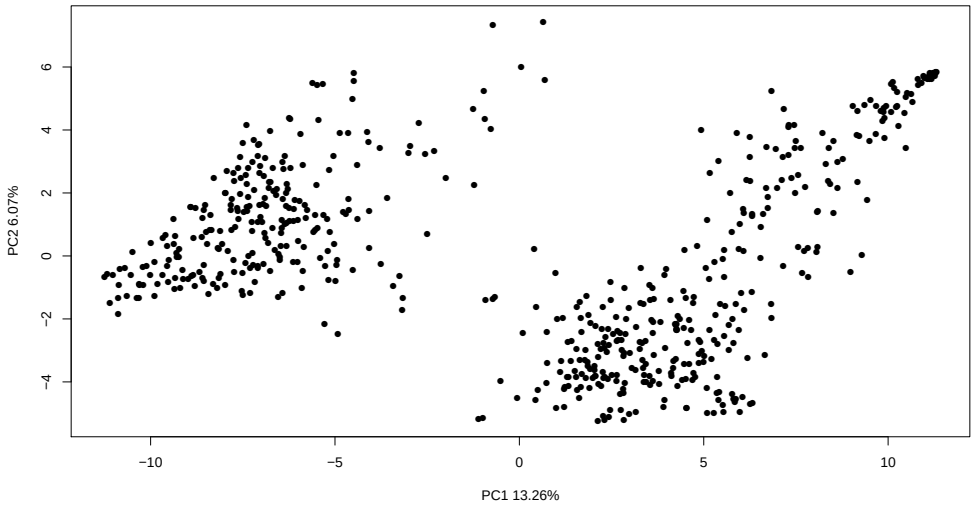
3.6 Supplementary Materials

Figure 3.7: Plot of the first two principal components of the marker matrix M in the CNLM and W-GY populations.

CNLM



W-GY



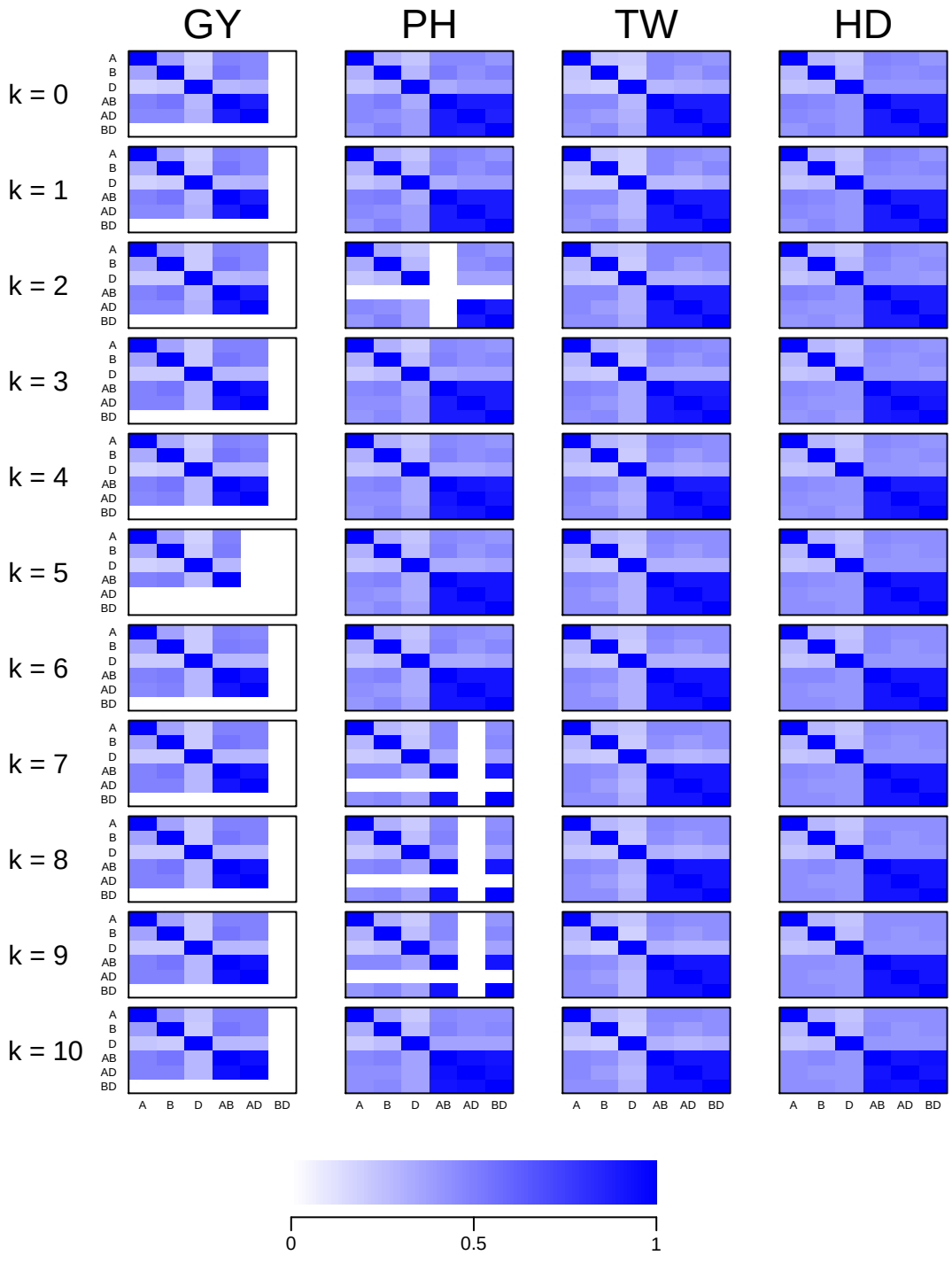


Figure 3.8: Correlation of variance component estimates derived from the average information from the model fit for models correcting for population structure with $k \in \{0, 1, \ldots, 10\}$ principal components for four traits in the CNLM population.

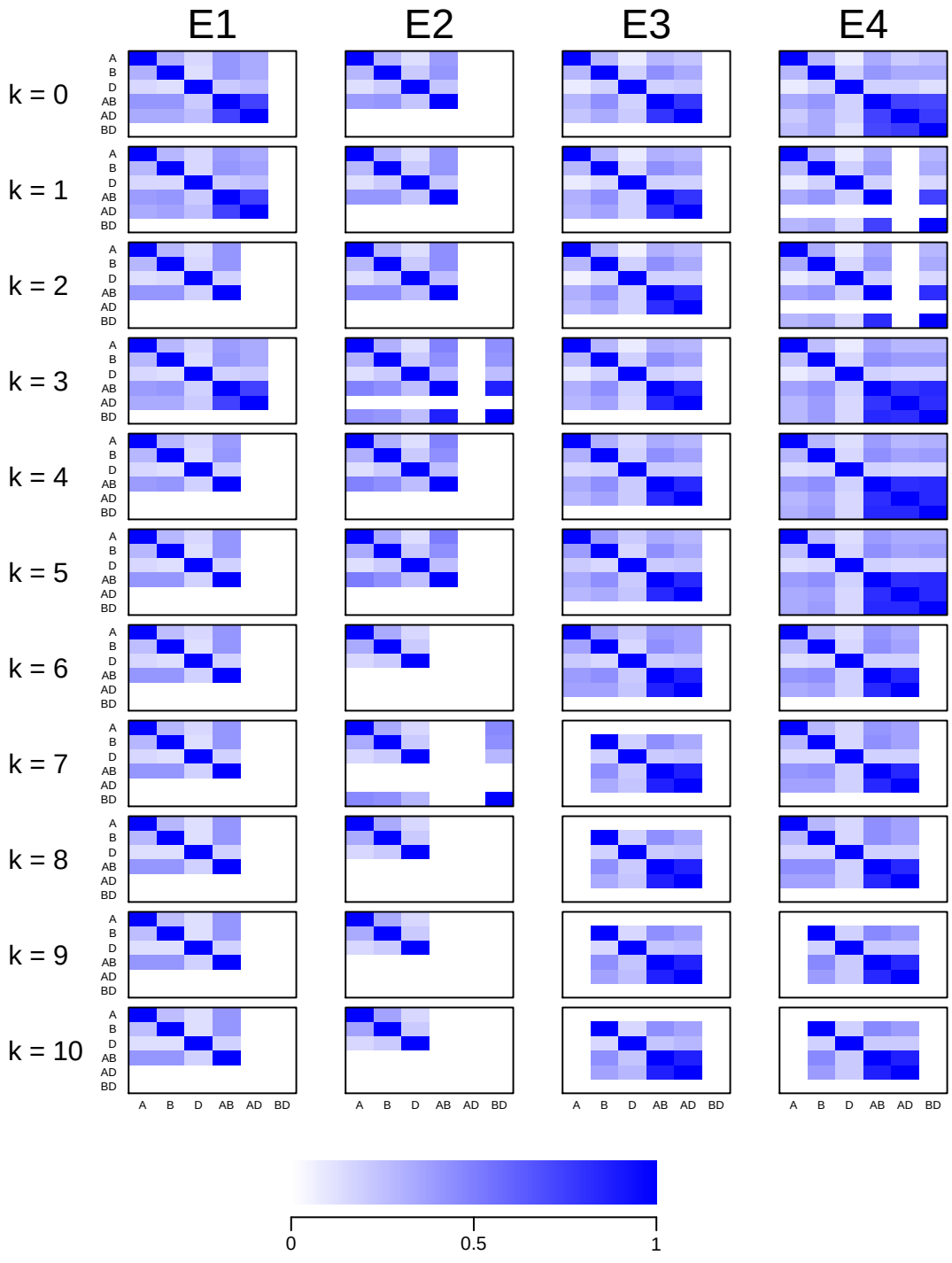


Figure 3.9: Correlation of variance component estimates derived from the average information from the model fit for models correcting for population structure with $k \in \{0, 1, \ldots, 10\}$ principal components for four traits in the W-GY population.

CHAPTER 4

A SUBFUNCTIONALIZATION EPISTASIS MODEL TO EVALUATE HOMEOLOGOUS GENE INTERACTIONS IN ALLOPOLYPLOID WHEAT

4.1 Introduction

Subfunctionalization and neofunctionalization are often described as distinct evolutionary processes. Neofunctionalization implies the duplicated genes have completely novel non-redundant function (Ohno 1970). Subfunctionalization is described as a partitioning of ancestral function through degenerate mutations in both copies, such that both genes must be expressed for physiological function (Stoltzfus 1999; Force et al. 1999; Lynch and Force 2000). However, barring total functional gene loss, many mutations will have some quantitative effect on protein kinetics or expression (Zeng and Cockerham 1993). Duplicated genes will demonstrate some quantitative degree of functional redundancy until the ultimate fate of neofunctionalization (i.e. complete additivity) or gene loss (pseudogenization) of one copy. It has been proposed that essentially all neofunctionalization processes undergo a subfunctionalization transition state (Rastogi and Liberles 2005).

If the mutations occur before the duplication event, as in allopolyploidy, the two variants are unlikely to have degenerate mutations. Instead, they may have differing optimal conditions in which they function or are expressed. The advantage of different variants at a single locus (Allard and Bradshaw 1964, alleles) or at duplicated loci (Mac Key 1970, homeoalleles) can result in greater plasticity to environmental changes. Allopolyploidization has been suggested as an evolutionary strategy to obtain the genic diversity necessary for invasive plant species to adapt

to the new environments they invade (Ellstrand and Schierenbeck 2000; Beest et al. 2011).

Adams et al. (2003) showed that some homeoallelic genes in cotton were expressed in an organ specific manner, such that expression of one homeolog effectively suppressed the expression of the other in some tissues. These results have since been confirmed in other crops such as wheat (Pumphrey et al. 2009; Akhunova et al. 2010; Feldman et al. 2012; Pfeifer et al. 2014), and evidence for neofunctionalization of homeoallelic genes has been observed (Chaudhary et al. 2009). Differential expression of homeologous gene transcripts has also been shown to shift upon challenge with heat, drought (Liu et al. 2015) and salt stress (Zhang et al. 2016) in wheat, as well as water submersion and cold in cotton (Liu and Adams 2007).

In the absence of outcrossing in inbred populations, selection can only act on individuals, changing their frequency within the population. If the selection pressure changes (e.g. for modern agriculture), combinations of homeoalleles within existing individuals may not be ideal for the new set of environments and traits. This presents an opportunity for plant breeders to capitalize on this feature of allopolyploids by making crosses to form new individuals with complementary sets of homeoalleles. Many of these advantageous combinations have likely been indirectly selected throughout the history of wheat domestication and modern breeding.

A crucial example involves the two homeologous dwarfing genes (Börner et al. 1996) important in the Green Revolution, which implemented semi-dwarf varieties to combat crop loss due to nitrogen application and subsequent lodging. I discuss this example in detail, and use it as a starting point to justify the search for homeologous interactions. While the effect of allopolyploidy has been demonstrated

at both the transcript level and whole plant level, I am unaware of attempts to use genome-wide homeologous interaction predictors to model whole plant level phenotypes such as growth, phenology and grain yield traits.

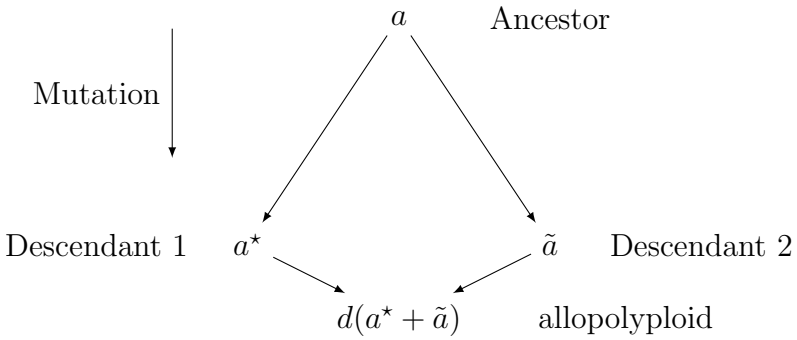
Using a soft winter wheat breeding population, I demonstrate that epistatic interactions account for a significant portion of genetic variance and are abundant throughout the genome. Some of these interactions occur between homeoallelic regions and I demonstrate their potential as targets for selection. If advantageous homeoallelic interactions can be identified, they could be directly selected to increase homeoallelic diversity, with the potential to expand the environmental landscape to which a variety is adapted.

I hypothesize that the presence of two evolutionarily divergent genes with partially redundant function leads to a less than additive gene interaction, and introduce this as a subfunctionalization model of epistasis.

4.2 Subfunctionalization Epistasis

I generalize the duplicate factor model of epistasis from Hill et al. (2008), by introducing a subfunctionalization coefficient d, that allows the interaction to shift between the duplicate factor and additive models. Let us consider an ancestral allele with an effect a. Through mutation, the effect of this locus is allowed to diverge from the ancestral allele to have effects $a^\star$ and $\tilde{a}$ in the two descendant species. When the two divergent loci are brought back together in the same nucleus, the effect of combining these two alleles is modeled as the sum of their individual effects multiplied by the subfunctionalization factor d, such that the effect of having both alleles becomes $d(a^\star + \tilde{a})$ (Figure 4.1).

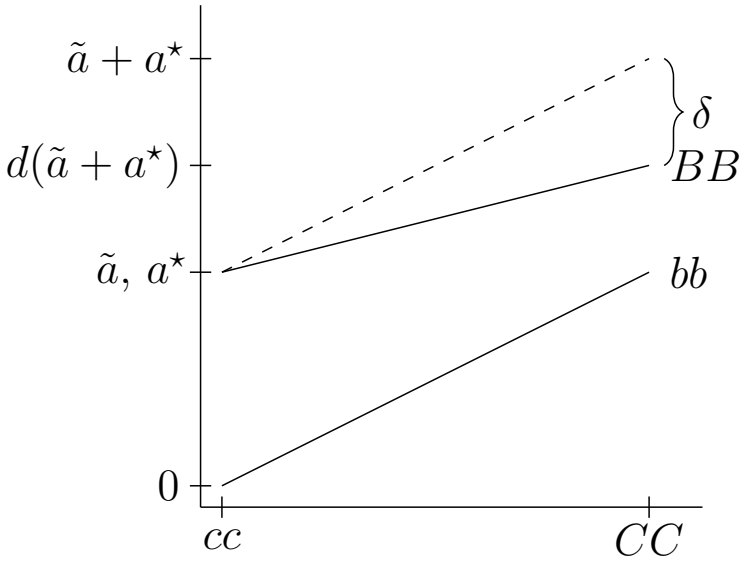
Figure 4.1: Diagram of subfunctionalization where a is the effect of a functional allele, $a^\star$ and $\tilde{a}$ are the effects of the descendant alleles, and d is the subfunctionalization (or divergence) coefficient.



Values of $d < 1$, indicate a less-than additive epistasis (Eshed and Zamir 1996), in this case, resulting from redundant gene function. When $d = 1/2$, and $a^\star = \tilde{a}$, the descendant alleles have maintained the same function and the duplicate factor model is obtained. As d exceeds $1/2$, the descendant alleles diverge in function (i.e. subfunctionalization), until d reaches 1, implying that the two genes evolved completely non-redundant function (i.e neofunctionalization). At the point where $d = 1$, the effect becomes completely additive.

For values of $d > 1/2$, the benefit of multiple alleles is realized in a model analogous to overdominance in traditional hybrids. As alleles diverge they can pick up advantageous function under certain environmental conditions. The homeo-heterozygote then gains an advantage if it experiences conditions of both adapted homeoalleles. Values of $d < 1/2$ may indicate allelic interference (Herskowitz 1987), a phenomenon that has been observed in many newly formed allopolyploids (Comai et al. 2003; McClintock 1984). Allelic interference, also referred to as dominant negative mutation, can result from the formation of non-functional homeodimers, while homodimers from the same ancestor continue to function properly. This

Figure 4.2: Epistatic interaction of two loci, B and C, with the expected effects for the $\{0, 1\}$ parameterization. δ indicates the deviation of the $BBCC$ genotype from an additive model for the $\{0, 1\}$ parameterization, where $d = 1 + \frac{\delta}{\tilde{a}+a^\star}$. The dotted line indicates the expectation under the additive model.



interference effectively reduces the number of active dimers by half (Herskowitz 1987; Veitia 2007).

4.2.1 Epistasis models

Let us consider the two locus model, with loci B and C. Using the notation of Hill, Goddard and Visscher (2008), the phenotype, y, is modeled as

$$y = B\alpha_B + C\alpha_C + BC\alpha\alpha_{BC} \tag{4.1}$$

where B and C are the marker allele scores, BC is the pairwise product of those

Table 4.1: Three types of epistatic interactions for inbred populations for two loci, B and C. The Additive $\times$ Additive and Duplicate factor are adapted from Hill, Goddard and Visscher (2008) with the heterozygous genotypes removed.

Additive $\times$ Additive

	CC	cc
BB	$2a$	0
bb	0	$2a$

Duplicate Factor

	CC	cc
BB	a	a
bb	a	0

Subfunctionalization

	CC	cc
BB	$d(a^\star + \tilde{a})$	$a^\star$
bb	$\tilde{a}$	0

marker scores, α_B and α_C are the additive effects of the B and C loci and $\alpha\alpha_{BC}$ is the interaction effect.

I will revisit two epistatic models, the "Additive $\times$ Additive Model without Dominance or Interactions Including Dominance" (called Additive $\times$ Additive hence forth) and the "Duplicate Factor" considered by Hill, Goddard and Visscher (2008) that are relevant for this discussion. I will also propose a generalized Duplicate Factor epistatic model to estimate the degree of gene functional redundancy, or subfunctionalization. Letting a be the effect on the phenotype, we can represent these epistatic models in Table 4.1.

When markers are coded $\{0, 1\}$ for presence of the functional allele, the deviation from the additive expectation, δ, is estimated by $\alpha\alpha_{BC}$. δ can then be used to calculate the subfunctionalization coefficient, $d = 1 + \frac{\delta}{a^\star + \tilde{a}}$ (Figure 4.2). Omitting the population mean, the least squares expectation of additive and epistatic effects is then

$$\mathrm{E}\begin{bmatrix} B\alpha_B \\ C\alpha_C \\ BC\alpha\alpha_{BC} \end{bmatrix} = \begin{bmatrix} a^\star \\ \tilde{a} \\ \delta \end{bmatrix} = \begin{bmatrix} a^\star \\ \tilde{a} \\ (d-1)(a^\star + \tilde{a}) \end{bmatrix}$$

4.2.2 Epistatic contrasts

Epistatic interaction predictors must be formed from marker scores in order to estimate interaction parameters. These interaction predictors are typically calculated as the pairwise product of the genotype scores for their respective loci. This can lead to ambiguity in the meaning of those interaction effects depending on how the marker scores are coded. Different marker parameterizations can center the problem at different reference points (i.e. different intercepts), and can scale the predictors based on allele or genotype effects (i.e. different slopes).

Table 4.2: Epistatic interaction tables resulting from $\{-1, 1\}$ and $\{0, 1\}$ marker coding for inbreds.

	CC	cc
BB	1	-1
bb	-1	1

	CC	cc
BB	1	0
bb	0	0

When loci B and C are coded as $\{-1, 1\}$ for inbred genotypes, including the product of the marker scores, BC, corresponds to the Additive $\times$ Additive model. Changing the reference allele at either locus does not change the magnitude of effect estimates but will change their signs. Using $\{0, 1\}$ coding, BC corresponds to the subfunctionalization model and estimates δ directly. For this coding scheme, the magnitude and sign can change depending on the reference allele at the two loci. This highlights one of the difficulties of effect sign interpretation, as it is not

clear which marker orientations should be paired. That is, which allele should be B as opposed to b, and which should be C as opposed to c? Marker alleles can be oriented to have either all positive or all negative additive effects, but the question remains: which direction should the more biologically active allele have on the phenotype?

Marker scores are typically assigned as either presence (or absence) of the reference, major, or minor allele, which may or may not be biologically relevant. For the $\{-1, 1\}$ parameterization, marker orientation does not change the magnitude of the effects, but will change their sign. It is therefore not possible to determine the sign of the epistatic effect relative to the additive effects without biological knowledge of marker orientation.

While it has been noted that the two different marker encoding methods do not result in the same contrasts of genotypic classes (He, Wang, and Parida 2015; Martini et al. 2016; Martini et al. 2017), coding does not affect the least squares model fit (Zeng, Wang, and Zou 2005; Álvarez-Castro and Carlborg 2007). Álverez-Castro and Carlborg (2007) show that there exists a linear transformation to shift between multiple parameterizations using a change-of-reference operation. This is convenient because all marker orientation combinations can be easily generated by changing the effect signs of a single marker orientation fit for the $\{-1, 1\}$ marker coding. These effect estimates can subsequently be transformed to the $\{0, 1\}$ coding effect estimates using the change-of-reference operation for all marker orientation combinations.

Following Álverez-Castro and Carlborg (2007), I demonstrate the change-of-reference operation simplified for inbred populations. For $\{0, 1\}$ marker coding and allowing G_1 to be the reference genotype, the genotypic values at a single

locus can be represented as

$$\mathbf{G} = \begin{bmatrix} G_1 \\ G_2 \end{bmatrix} = \mathbf{S}_{01}\mathbf{E}_{01} = \begin{bmatrix} 1 & 0 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} \mu \\ a \end{bmatrix} \tag{4.2}$$

where $\mathbf{S}_{01}$ is the marker score matrix using the $\{0, 1\}$ marker parameterization and $\mathbf{E}_{01}$ is the vector of expected values. For the two locus epistasis model, the four genotypic values are then

$$\mathbf{G} = \begin{bmatrix} G_{11} \\ G_{12} \\ G_{21} \\ G_{22} \end{bmatrix} = (\mathbf{S}_{01} \otimes \mathbf{S}_{01})\mathbf{E}_{01} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 1 & 1 \end{bmatrix} \begin{bmatrix} \mu \\ a_1 \\ a_2 \\ a_1a_2 \end{bmatrix} \tag{4.3}$$

The three locus interaction is extended by

$$\mathbf{G} = (\mathbf{S}_{01} \otimes \mathbf{S}_{01} \otimes \mathbf{S}_{01}) \left[\mu \; a_1 \; a_2 \; a_1a_2 \; a_3 \; a_1a_3 \; a_2a_3 \; a_1a_2a_3\right]^{\mathrm{T}} \tag{4.4}$$

To shift from $\{-1, 1\}$ coding estimates, $\boldsymbol{\beta}_{101}$, to $\{0, 1\}$ coding estimates, $\boldsymbol{\beta}_{01}$ the following transformation exists (Álvarez-Castro and Carlborg 2007). Let $\mathbf{S}_{\text{-}11}$ indicate the $\{-1, 1\}$ marker parameterization

$$\mathbf{S}_{\text{-}11} = \begin{bmatrix} 1 & -1 \\ 1 & 1 \end{bmatrix}$$

then $\mathbf{E}_{01} = (\mathbf{S}_{01}^{-1} \otimes \mathbf{S}_{01}^{-1})(\mathbf{S}_{\text{-}11} \otimes \mathbf{S}_{\text{-}11})\mathbf{E}_{\text{-}11}$.

4.3 Materials and Methods

4.3.1 RIL population

A bi-parental recombinant inbred line (RIL) population of 158 lines segregating for two dwarfing genes was used to illustrate an epistatic interaction between the well known homeologous genes on chromosomes 4B and 4D, *Rht-B1* and *Rht-D1*, important in the Green Revolution. Two genotyping by sequencing (GBS) markers linked to these genes were used to track the segregating mutant (b and d) and wildtype (B and D) alleles. Only one test for epistasis between these two markers was run. This homeologous marker pair was denoted RIL_Rht1. Details of the population can be found in Chapter 2.3.

4.3.2 CNLM population

The Cornell small grains soft winter wheat breeding population (CNLM) was used to investigate the importance of homeologous gene interactions in a large adapted breeding population. The dataset consists of 1,447 lines evaluated in 26 environments around Ithaca, NY. Because the data were collected from a breeding population, only 21% of the genotype/environment combinations were observed, totaling 8,692 phenotypic records. Standardized phenotypes of four traits, grain yield (GY), plant height (PH), heading date (HD) and test weight (TW) were recorded. All lines were genotyped with 11,604 GBS markers aligned to the International Wheat Genome Sequencing Consortium (IWGSC) RefSeq v1.0 wheat genome sequence of 'Chinese Spring' (IWGSC 2018, accepted), and subsequently imputed. Further details can be found in Chapter 2.1.

4.3.3 Homeologous marker sets

Using the IWGSC RefSeq v1.0 'Chinese Spring' wheat genome sequence (IWGSC 2018, accepted), homeologous sets of genes were constructed by aligning the coding sequences back onto themselves. Alignments of coding sequences was accomplished with BLAST+, allowing up to 10 alignments with an e-value cutoff of 1e-5. Alignments were only considered if they aligned to 80% or more of the query gene. Of the 110,790 coding sequences, 13,111 triplicate sets with one gene on each homeologous chromosome (representing 39,333 genes) were identified with no other alignments meeting the criterion. An additional 5,073 triplicates (representing 15,219 genes) were added by selecting the top 2 alignments if they were on the corresponding homeologous chromosomes. Duplicate sets were also included if there was not a third alignment to one of the three subgenomes, adding an additional 5,612 duplicates. The coding sequences for which I did not identify homeologous genes either appeared to be singletons (24,695 coding sequences) that did not have a good alignment to a gene on a homeologous chromosome, or had many alignments across the genome making it impossible to determine with certainty which alignments were truly homeologous (20,319 coding sequences). The known 4A, 5A, and 7B translocation in wheat (Devos et al. 1995) was ignored for simplicity in this study, but could easily be accounted for by allowing homeologous pairs across these regions.

The resulting 23,796 homeologous gene sets, comprised of 18,184 triplicate and 5,612 duplicate gene sets, sampled roughly 59% of the gene space of hexaploid wheat. Each homeologous gene was then anchored to the nearest marker by physical distance (Supplementary Figures 4.7 and 4.8), and homeologous sets of markers were constructed from the anchor markers to each homeologous gene set.

Redundant marker sets due to homeologous genes anchored by the same markers were removed, resulting in 6,142 triplicate and 3,985 duplicate marker sets for a total of 10,127 unique homeologous marker sets. These marker sets were then used to calculate marker interactions as pairwise products of the marker score vectors.

As a control, two additional marker sets were produced by sampling the same number of duplicate and triplicate marker sets as the homeologous set (Homeo) described above. These markers sets were sampled either from chromosomes within a subgenome (Within, e.g. markers on 1A, 2A and 3A), or across non-syntenic chromosomes of different subgenomes (Across, e.g. markers on 1A, 2B and 3D). Samples were taken to reflect the same marker distribution of the Homeo set with regard to their native genome, which has a larger proportion of D genome markers relative to their abundance (see Figure 2.2). To illustrate, note that three-way homeologous interactions have equal proportions of markers belonging to the A, B and D genomes, whereas D genome markers only account for 13% of all markers in the CNLM population. All analyses conducted on the Homeo marker set were conducted on the Within and Across marker sets to determine the importance of homeologous gene interactions relative to non-homeologous gene interactions.

4.3.4 Determining marker orientation

For each homeologous marker set, additive homeologous marker effects and their multiplicative interaction effects were estimated as fixed effects in the following linear mixed model while correcting for background additive and epistatic effects. The $\{-1, 1\}$ marker parameterization was used for fixed marker additive and interaction effect estimates.

$$\mathbf{y} = \tilde{\mathbf{Z}}\mathbf{S}_{\text{-}11}\mathbf{E}_{\text{-}11} + \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{g}_{G+I} + \boldsymbol{\varepsilon} \tag{4.5}$$

where $\mathbf{1}_n$ is a vector of ones, μ is the general mean, $\mathbf{X}$ is the design matrix, $\boldsymbol{\beta}$ is the vector of fixed environmental effects, and $\mathbf{Z}$ is the line incidence matrix. $\mathbf{S}_{\text{-}11}$ is the matrix of genotype marker scores and interactions, while $\mathbf{E}_{\text{-}11}$ is the additive and interaction effects that need to be estimated. $\tilde{\mathbf{Z}}$ is the incidence matrix for the two- or three-way genotype of each homeologous marker set. $\mathbf{Z}$ and $\tilde{\mathbf{Z}}$ differ in that the former links observations to a specific line, whereas the latter links observations to one of the two- or three-way genotype classes for the homeologous marker set. The background genetic effects were assumed to be $\mathbf{g}_{G+I} \sim \mathcal{N}(0, \sigma^2_G\mathbf{K}_G + \sigma^2_I\mathbf{H}_I)$ with population parameters previously determined (Zhang et al. 2010). The additive, $\mathbf{K}_G$ and epistatic, $\mathbf{H}_I$, covariances were calculated as described in Van Raden (2008, method I) and Martini et al. (2016), respectively. This weighted covariance matrix was used to reduce computational burden associated with estimating two variance components in the same fit.

A Wald test was used to obtain a p-value for the null hypothesis that the marker effects or interactions are zero. All marker orientation combinations were generated by changing the estimated effect signs, and then transformed to the $\{0, 1\}$ marker effect estimates using the change-of-reference operation (Álvarez-Castro and Carlborg 2007). Only marker orientations with all positive or all negative effects were considered.

Four marker orientation schemes were used in further analyses. Markers were oriented to have either all positive (POS) effects, all negative (NEG) effects, or to have the direction chosen for each marker set by one of two methods. The first method, 'low additive variance high additive effect' (LAVHAE), selected the

marker orientation that minimized the difference (or variance for three-way sets) of the additive main effects while maximizing the mean of the absolute values of the additive main effects. Only additive effects were used to select the marker orientation to keep from systematically selecting marker orientations with a specific interaction pattern for the LAVHAE scheme. A second method was used to select marker orientations solely based on the highest variance of estimated additive and interaction effects, denoted 'high total effect variance' (HTEV). The HTEV method biases the interaction term to be in the opposite direction to that of the additive effects, therefore discussion of results using this method is limited.

4.3.5 Additive only simulated controls

Marker effect and interaction estimates using $\{0, 1\}$ are not orthogonal, so care must be taken when interpreting the direction and magnitude of the effects estimates. The positive covariance between the marker scores and their interaction leads to a multicollinearity problem, and results in a negative relationship between additive and interaction effects if both additive effects are oriented in the same direction. The $\{-1, 1\}$ marker parameterization does not result in orthogonal contrasts either, but only the magnitude of the estimates are affected and not their sign. Because magnitudes do not change with the $\{-1, 1\}$ marker parameterization, sign relationships between additive and interaction effects cannot be determined.

To determine if any observed pattern was due entirely to multicollinearity, a new phenotype with no epistatic effects was simulated from the data for each trait. The estimate of the marker variance was calculated from the additive genetic variance estimate as $\hat{\sigma}^2_m = \hat{\sigma}^2_G(2\mathbf{p}^{\mathrm{T}}(\mathbf{1} - \mathbf{p}))^{-1}$, where $\mathbf{p}$ is the vector of marker

allele frequencies. Then for each trait, a new additive phenotype was simulated as $\mathbf{y}_{sim} = \mathbf{1}\mu + \mathbf{X}\boldsymbol{\beta} + \mathbf{ZMu}_{sim} + \boldsymbol{\varepsilon}_{sim}$ where $\mathbf{M}$ is the matrix of marker scores, $\mathbf{u}_{sim}$ was sampled from $\mathcal{N}(0, \hat{\sigma}^2_m)$ and $\boldsymbol{\varepsilon}$ was sampled from $\mathcal{N}(0, \hat{\sigma}^2)$. A Kolmogorov-Smirnov (KS) test was used to determine if the distribution of the estimated interaction effects from the actual data differed from the distribution of effects estimated from simulated data. An additional simulated phenotype was also produced by first permuting each column of the marker matrix to remove any effects due to LD structure.

4.3.6 Genomic prediction

To determine the importance of epistatic interactions to the predictability of a genotype, a genomic prediction model was fit as

$$\mathbf{y} = \mathbf{1}\mu + \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{g}_G + \mathbf{Z}\mathbf{g}_I + \boldsymbol{\varepsilon} \tag{4.6}$$

The random vectors of additive genotype, epistatic interactions, and errors were assumed to be distributed as $\mathbf{g}_G \sim \mathcal{N}(0, \sigma^2_G\mathbf{K})$, $\mathbf{g}_I \sim \mathcal{N}(0, \sigma^2_I\mathbf{H})$ and $\boldsymbol{\varepsilon} \sim \mathcal{N}(0, \sigma^2)$, respectively.

The additive covariance matrix, $\mathbf{K}$, was calculated using Van Raden (2008), method I. The epistatic covariance matrix $\mathbf{H}$ was calculated either as defined by Martini et al. (2016) to model all pairwise epistatic interactions (Pairwise), or in a similar fashion as $\mathbf{K}$ for oriented marker sets, where only unique products of marker variables were included instead of the marker variables. For the latter, the matrix was scaled with the sum of the joint marker variances as $(2\mathbf{q}^{\mathrm{T}}(\mathbf{1} - \mathbf{q}))^{-1}$,

where **q** is the frequency of individuals containing both the non-reference marker alleles.

A small coefficient of 0.01, was added to the diagonals of the covariance matrix to recover full rank lost in centering the matrix of scores prior to calculating the covariance. Five-fold cross validation was performed by randomly assigning individuals to one of five folds for 10 replications. Four folds were used to train the model and predict the fifth fold for all five combinations. All models were fit to the same sampled folds so that models would be directly comparable to one another, and not subject to sampling differences. Prediction accuracy was assessed by collecting genetic predictions for all five folds, then calculating the Pearson correlation coefficient between the predicted genetic values for all individuals and a "true" genetic value. The "true" genetic values were obtained by fitting a mixed model to all the data with fixed effects for environments and a random effect for genotypes, assuming genotype independence with a genetic covariance **I**.

Increase in genomic prediction accuracy from the additive model was used as a proxy to assess the relative importance of oriented marker interaction sets. To determine the proportion of non-additive genetic signal attributable to each interaction set, the ratio of the prediction accuracy increase from the additive model using the interaction set (Homeo, Within and Across) to the prediction accuracy increase from the additive model modeling all pairwise epistatic interactions (Pairwise) was used for comparison of models. The percentage of non-additive predictability was calculated as follows for each interaction set.

$$\frac{\text{accuracy(Interaction Set)} - \text{accuracy(Additive)}}{\text{accuracy(Pairwise)} - \text{accuracy(Additive)}} \tag{4.7}$$

Table 4.3: Marker and epistatic effect estimates for *Rht-1D* and *Rht-1B* linked GBS markers for plant height (cm) in 158 RIL lines derived from NY91017-8080 × Caledonia. Least squares effect estimates are for markers coded either using $\{0, 1\}$ coding or $\{-1, 1\}$, and then oriented such that the two marker main effects are either both positive (+) or both negative (−)

Marker Coding	Effect Orientation	Intercept	*Rht-1B*	*Rht-1D*	*Rht-1B*×*Rht-1D*
$\{0, 1\}$	+	69.9	23.4	22.2	-12.2
$\{0, 1\}$	−	103.3	-11.2	-10.0	-12.2
$\{-1, 1\}$	+	89.7	8.6	8.0	-3.0
$\{-1, 1\}$	−	89.7	-8.6	-8.0	-3.0

4.4 Results and Discussion

4.4.1 *Rht-1*

RIL population

The markers linked to the *Rht-1B* and *Rht-1D* genes both had significant effects ($p < 10^{-10}$) and explained 19.6% and 20.5% of the variation in the height of the RIL population (Supplemental Table 4.7). The test for a homeoallelic epistatic interaction between these *Rht-1* linked loci was also significant ($p = 0.0025$), but only explained 3.5% of the variance after accounting for the additive effects. Had I tested all pairwise marker interactions in this population, this test would not have passed a Bonferroni corrected significance threshold.

Effect estimates for the *Rht-1* markers and their epistatic interaction are shown in Table 4.3, for $\{0, 1\}$ and $\{-1, 1\}$ marker parameterizations, and for orientations where the marker main effects are both positive or both negative. By plotting these values on a classic epistasis plot and indicating the genotype class means, it

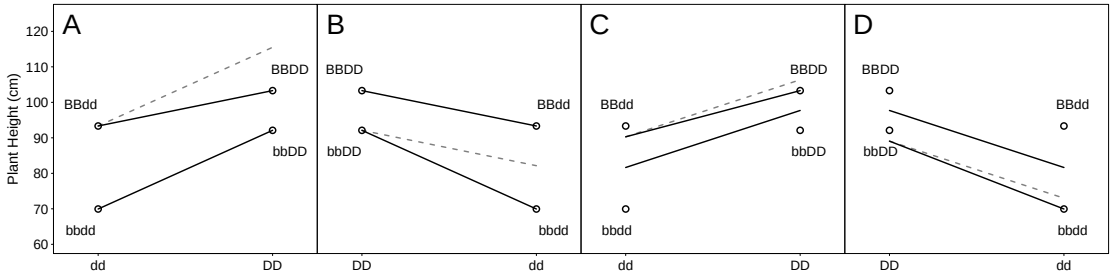


Figure 4.3: Epistasis plot of effects for *Rht-1B* and *Rht-1D* linked markers on plant height (cm) in 158 RIL lines derived from NY91017-8080 × Caledonia. Circles indicate genotype class means, and lines indicate the marker effect slopes. The dotted line indicates the expected slope based on the additive model. A) $\{0, 1\}$ marker coding with positive marker effect orientation. B) $\{0, 1\}$ marker coding with negative marker effect orientation. C) $\{-1, 1\}$ marker coding with positive marker effect orientation, D) $\{-1, 1\}$ marker coding with negative marker effect orientation.

it evident that the $\{0, 1\}$ parameterization is arguably more intuitive due to effects corresponding directly to differences in genotype values (Figure 4.3). They both contain the same information and are equivalent for prediction, but the interpretation of the $\{-1, 1\}$ marker coding is less obvious because the slopes are deviations from the expected double heterozygote, which does not exists in an inbred population. The $\{0, 1\}$ parameterization uses the double dwarf as the reference point, where the effects α_B and α_C are the two semi-dwarf genotypes. The tall genotype is the sum of the semi-dwarf allele effects plus the deviation coefficient, δ, which corresponds to $\alpha\alpha_{BC}$.

The estimated d parameter of 0.73 indicates a significant degree of redundancy between the wild type *Rht-1* homeoalleles. This suggests that either the gene products maintain partial redundancy in function, or the expression of the two homeoalleles is somewhat redundant. The latter is less likely given that the two functional wild type genes have comparable additive effects relative to the double

dwarf. If the two genes were expressed at different times or in different tissues based on their native subgenome, the additive effects would be likely to differ in magnitude. This demonstrates a functional change between homeoalleles that has been exploited for a specific goal: semi-dwarfism.

When the markers are oriented in the opposite direction, to indicate the GA insensitive mutant allele as opposed to the GA sensitive wildtype allele, the interpretation of the interaction effect changes. The additive effect estimates become indicators of the reduction in height by adding a GA insensitive mutant allele. The interaction effect becomes the additional height reduction from the additive expectation of having both GA insensitive mutant alleles, resulting in a d parameter of 1.58. The same interpretation can be made, but must be done so with care. Losing wildtype function at both alleles results in a more drastic reduction in height than expected because there is redundancy in the system. Therefore, the d parameter is most easily interpreted when the functional direction of the alleles is known. Simply put, when you add function on top of function, little is gained, but when you remove all function, catastrophe ensues.

CNLM population

For the CNLM population, the markers with the lowest p-values associated with plant height on the short arms of 4B and 4D did not show a significant interaction with their respective assigned homeologous marker. The marker S4B_PART1_38624956 of the homeologous marker set H4.16516 had the lowest p-value on the 4BS chromosome ($p = 5.7 \times 10^{-4}$), but its assigned homeologous marker, S4D_PART1_28548806, was not significant. Similarly, H4.23244 included the 4DS SNP with the lowest p-value, S4D_PART1_10982050, was significant at a

Bonferroni correction threshold ($p = 6.2 \times 10^{-8}$), but it's assigned homeologous marker on 4BS, S4D_PART1_10982050, did not have a low p-value. The significant markers were 8 Mbp downstream and 8 Mbp upstream of the *Rht-1B* and *Rht-1D* genes, respectively. While these distances would seem extreme to those working with a small genome such as that of *Arabidopsis*, they are less than $< 2\%$ of the chromosome length. Therefore, these distances are quite reasonable considering the size of the chromosomes and relatively high degree of LD in wheat. Chromosomes 4B and 4D are 674 Mbp and 510 Mbp in length based on the current reference sequence, RefSeqv1.0, respectively (IWGSC 2018, accepted). For comparison, chromosome 4B is about five times the size of the entire *A. thaliana* genome.

One marker set on homeolog 4, H4.7408 with markers S4B_PART1_60701027 and S4D_PART1_40659460, did have main effects and an epistatic interaction all significant at $p < 0.05$. However, these markers were located considerably further downstream, 30 Mbp and 22 Mbp from the *Rht-1B* and *Rht-1D* genes. This was one of only 30 homeologous marker sets that had two main effects and an interaction significant at $p < 0.05$ across all traits. These sets are discussed in further detail below (See Section 4.4.3).

A new homeologous marker set was constructed, called CNLM_Rht1, with the SNPs on 4BS and 4DS with the lowest p-values mentioned above (S4B_PART1_38624956 and S4D_PART1_10982050). The additive effects of markers S4B_PART1_38624956 and S4D_PART1_28548806 had p-vales of $p = 5.5 \times 10^{-4}$ and $p = 3.7 \times 10^{-8}$, respectively. The pair was tested for an epistatic interaction, but was only significant at a $\alpha = 0.05$ threshold ($p = 0.015$). This set was oriented in the same direction as the RIL_Rht1 set using the LAVHAE orientation method.

While the magnitude of these effects was reduced (7.13, 7.09 and -4.56 for the 4D, 4B and 4B×4D effects respectively), the CNLM_Rht1 set had a *d* parameter value of 0.68, similar to that of RIL_Rht1. Had this set alone been tested, I would have concluded that this was a significant homeologous interaction.

It appears that the S4B_PART1_38624956 marker is not in high LD with the *Rht-1B* gene perhaps due to physical distance, multiple alleles or separation of these mutations in time. This SNP may have occurred before the *Rht-1B* mutation, and is therefore flagging both the wild type and a GA-insensitive allele in the population. The 4B wild type allele had a high frequency of 0.88, leading to a higher incidence of tall wheat genotypes (0.36) than expected (Supplemental Table 4.8). While some of the important historical varieties in this population are known to lack either dwarfing allele, it is unlikely that the proportion would be as high as 36% of the lines. It appears that a GBS SNP in high LD with the *Rht-1B* was not sampled, or was filtered out of the marker set. The presence of the double dwarf marker genotype also indicates a lack of perfect LD with the functional polymorphism, as there are no phenotypic double dwarfs in the population.

These results highlight the difficulty in using single SNP markers to track homeologous regions. If a non-functional mutation that is not in LD with a functional mutation is identified as the homeologous marker, then there will appear to be no interaction, while a marker in high LD with the homeolog but physically further away from the gene was missed. Further research is required to determine if functionally important homeologous marker pairs can be identified based on LD signatures in the population.

One of the challenges of using diverse panels of individuals is that marker proximity to a functional mutation is not necessarily indicative of high LD between

the two sites. Significantly older or newer marker mutations may be in weak LD with a functional mutation despite close physical proximity, at least until a genetic bottleneck brings them back into high LD. The best markers are those mutations in close proximity and occurring at approximately the same historical time as the functional mutation in the same individual, such that both polymorphisms are inherited together and subsequently co-inherited.

Another strategy to determine functional homeologous regions is to relax which sets of markers are considered homeologous. This could be accomplished by allowing pairwise relationships with all markers across entire subgenomes (Chapter 3), on syntenic chromosomes or chromosome arms (Chapter 5). Haplotypes could be constructed in the population and used to model homeologous haplotype interactions. This strategy would likely have more power assuming the correct SNPs could be identified to assign functional haplotypes, but that is beyond the scope of this report.

4.4.2 Significant homeoallelic interactions

A trait-wise Bonferroni significance threshold of $0.05/20,641 = 2.4 \times 10^{-6}$ for the 20,641 testable interaction effects out of 28,553 total interaction terms was used to determine which interaction effects had a significant effect on the phenotype. The other 7,912 interaction terms were a linear combination of the marker additive effects scores, precluding the ability to test them as separate predictors. This was often the case when a marker allele was always co-inherited with its respective homeologous marker allele on a different subgenome (i.e. only 3 genotypic classes).

Few homeoallelic interactions were significant at the trait-wise Bonferroni cutoff

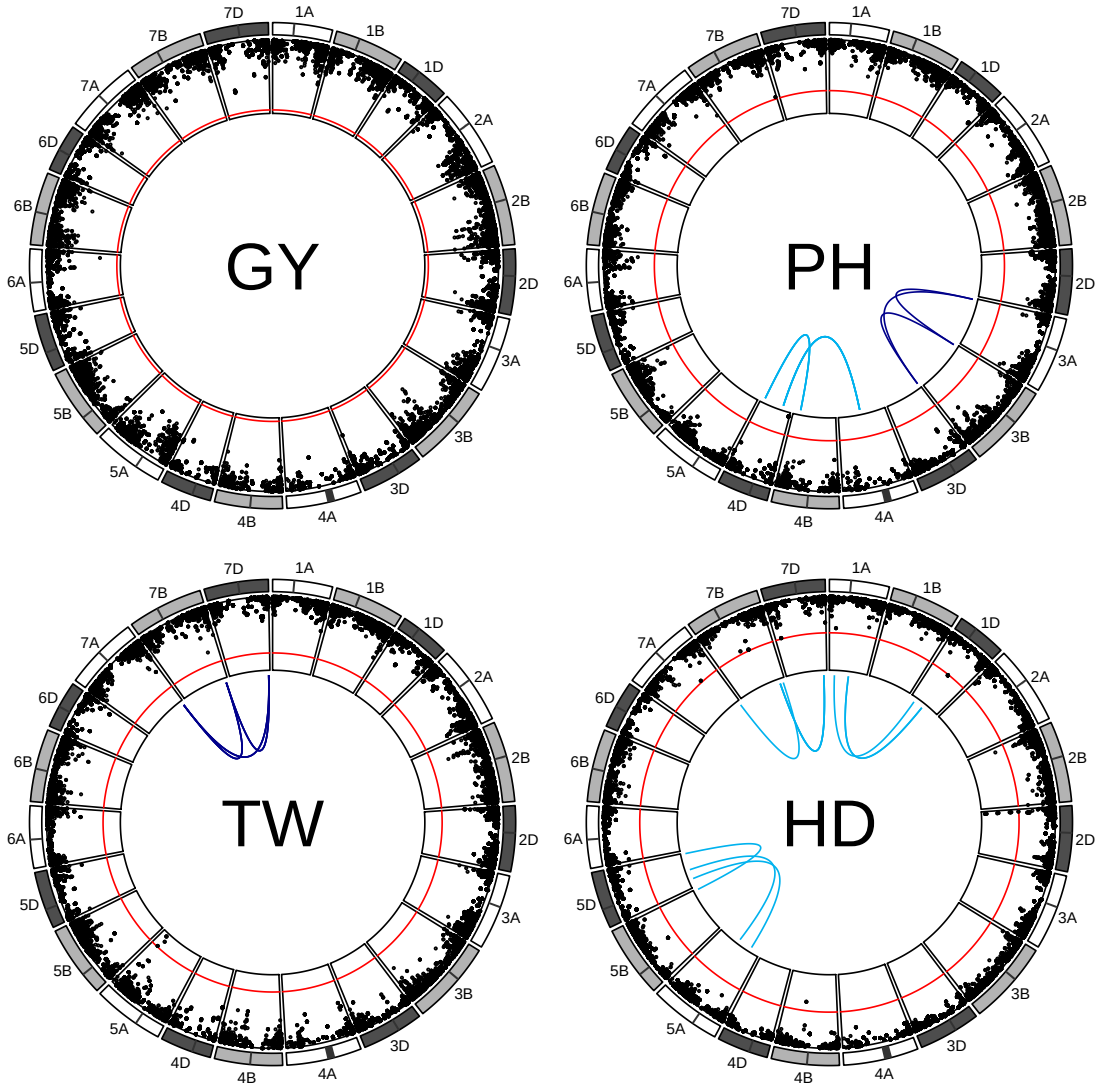


Figure 4.4: Manhattan plot of homeoallelic marker sets for each of the 21 chromosomes of wheat. The red line indicates a trait wise Bonferroni significance threshold for additive effects of $-\log_{10}(6.0 \times 10^{-}6) = 5.2$. Light blue lines indicate significant two-way homeoallelic marker interactions that exceeded a Bonferroni threshold for all testable interaction effects $-\log_{10}(2.4 \times 10^{-}6) = 5.6$. Dark blue lines indicate significant 3-way homeoallelic marker interactions that exceeded the same Bonferroni threshold.

(Figure 4.4). Significant homeoallelic interactions for PH were identified between 4AL and 4DS as well as 4BL and 4DL. Both of these locations were likely too far away from the *Rht-1* alleles to be tagging these genes directly, but they may be regulatory sites for these genes. Another set of interacting sites between the short arms of homeologous chromosomes 3AS, 3BS and 3DS was also identified for PH, but the additive effects were not significant. Two interacting regions on homeolog 1, between 1AS and 1DS and between 1AL and 1DL, and three interacting regions on homeolog 5 also appeared to be influencing HD. One region on the distal end of homeolog 7 affected both HD and TW, with significant two-way and three-way interactions. Although they were tagged with different marker sets for the two traits, these epistatic regions appeared to co-localize within 2 Mbp.

No significant additive or interaction effects were detected for GY, highlighting the highly polygenic nature of the grain yield trait. In several cases, one of the additive effects was significant but the other was not, and it is not clear if this is influencing the detection of interactions. It may be that the significant marker is simply in higher LD with the functional mutation conditional on the presence of the other marker, allowing the interaction to pick up the additional signal from the functional mutation (Wood et al. 2014). However, if this were the case, the interaction would be expected to be in the same direction as the additive effect, which was not generally observed.

I did not detect an interaction between the two significant additive regions on 2B and 2D for the HD trait. While these two markers were not grouped as a homeologous set, they were tested as such based on their proximity to the well described Photoperiod-1 genes, *Ppd-B1* and *Ppd-D1*, on chromosomes 2B and 2D respectively. These genes are known to influence photoperiod sensitivity, and

therefore transition to flowering and heading date (Welsh J.R. and R.D. 1973; Law, Sutka, and Worland 1978; Scarth and Law 1983). Certain allele pairs at these genes have been shown to exhibit a high degree of epistasis (Poland 2018, personal communication) in a bi-parental family. It is unclear why no interaction was observed in this population.

Jiang et al. (2017) also investigated the presence of homeologous interactions, but found little evidence in a large population of hybrid wheat. They did not attempt to tag homeologous loci, but instead considered interactions across any markers on homeologous chromosomes to be syntenic. It is possible that interactions at homologous (i.e. heterozygous) loci largely outweighed the interactions across homeologous loci in that population, given it was constructed from highly divergent parents. Additionally, they tested all pairwise marker combinations, resulting in a strict significance threshold that may have missed small effect homeologous interactions.

Homeologous interactions make up relatively few of the potential two-way interactions within an allopolyploid genome. Given a subgenome with k genes and alloploidy level p (i.e. the number of subgenomes), there are $k\binom{p}{2}$ two-way homeologous interactions versus $\binom{k}{2} - k\binom{p}{2}$ potential two-way non-homeologous gene interactions. For a subgenome size of 30,000 genes, this represents 0.02% and 0.006% of the possible two-way gene interactions for an allotetraploid and an allohexaploid, respectively. That said, homeoallelic interactions should be far more likely to have a true biological interaction than random pairs of genes because they should belong to the same or similar biochemical pathways.

Table 4.4: Estimates of d coefficients for marker sets where both additive and the two-way interaction effects were significant at $p < 0.05$, combined for all 4 traits. The expected number of non-zero additive and two-way interactions effects based on a 0.05 significance threshold by chance is 11 (i.e. 4 traits $\times$ 22,411 two-way interactions $\times$ 0.05^3). Coefficients have been grouped by categories related to the potential mode of epistasis, where $d < 0.5$ indicates a highly negative interaction, $0.5 \leq d < 1$ a less than additive interaction may be indicative of subfunctionalization for homeologous genes, and $d > 1$ which indicates positive, or greater than additive, epistasis. Three marker sets are shown, either across all homeologous loci (Homeo), sampled sets within (Within) and across (Across) non-syntenic subgenome regions. An additional phenotype was simulated to contain additive only phenotypes to contain no epistasis, and fit with the Homeo marker set (Simulated Additive).

Marker Set	$d < 0.5$	$0.5 \leq d < 1$	$d > 1$	Total
Homeo	8	14	8	30^{***}
Simulated Additive	1	1	4	6
Across	9	7	1	17^{*}
Within	6	3	4	13

*, **, *** indicate significantly greater than the expected number of significant sets at p = 0.05, 0.01 and 10^{-6} based the binomial distribution with 89,644 trials and a probability of 0.05^3.

4.4.3 Estimates of d

There were few cases where at least two additive effects and their corresponding interaction effect were all significantly different from zero. This may be due to the difficulty of assigning functional homeologous gene sets using single SNPs, as well as a lack of statistical power owing to low minor allele frequencies (Hill, Goddard, and Visscher 2008). The lack of a large number of significant interactions is not surprising given that allele frequencies near 0.5 are uncommon in both natural and breeding populations.

To determine whether more homeologous marker sets were displaying a pattern indicative of subfunctionalization than would be expected by chance, marker sets

where both additive and two-way interaction effects were significant at a threshold of $\alpha = 0.05$ were examined (Table 4.4). The expected number of two-way marker sets with significant additive and interaction effects is about 11 (i.e. 4 traits $\times$ 22,411 two-way interactions $\times$ 0.05^3), assuming independence of loci and true additive and interaction effects of zero. Only the Homeo and Across marker sets had more than the expected number. The homeologous marker set had a larger proportion of d coefficients estimated between 0.5 and 1 relative to the strictly additive simulated phenotypes as well as the other non-homeologous marker sets, suggesting that homeologous loci exhibit a pattern indicative of subfunctionalization more so than other marker sets tested. However, it is unclear if these few marker sets are indicative of any global pattern, and may simply be an artifact of sampling. When I looked at d statistics for all two-way interactions regardless of significance, the Homeo set had the lowest proportion of d parameters between 0.5 and 1.

Because the power to detect significant effects diminishes as more tests are accomplished, it may be prudent to look at global trends between homeologous additive effects and their interactions, regardless of statistical significance.

4.4.4 Evidence of subfunctionalization

A strong negative relationship between additive and interaction effects was observed when using the $\{0, 1\}$ marker parameterization (Supplementary Figure 4.9). This negative relationship was also observed in the phenotypes simulated to be strictly additive (Supplementary Figure 4.10). The multicollinearity of the additive and epistatic predictors at least partially drives this relationship, where positively correlated additive and epistatic predictors will tend to have effect estimates

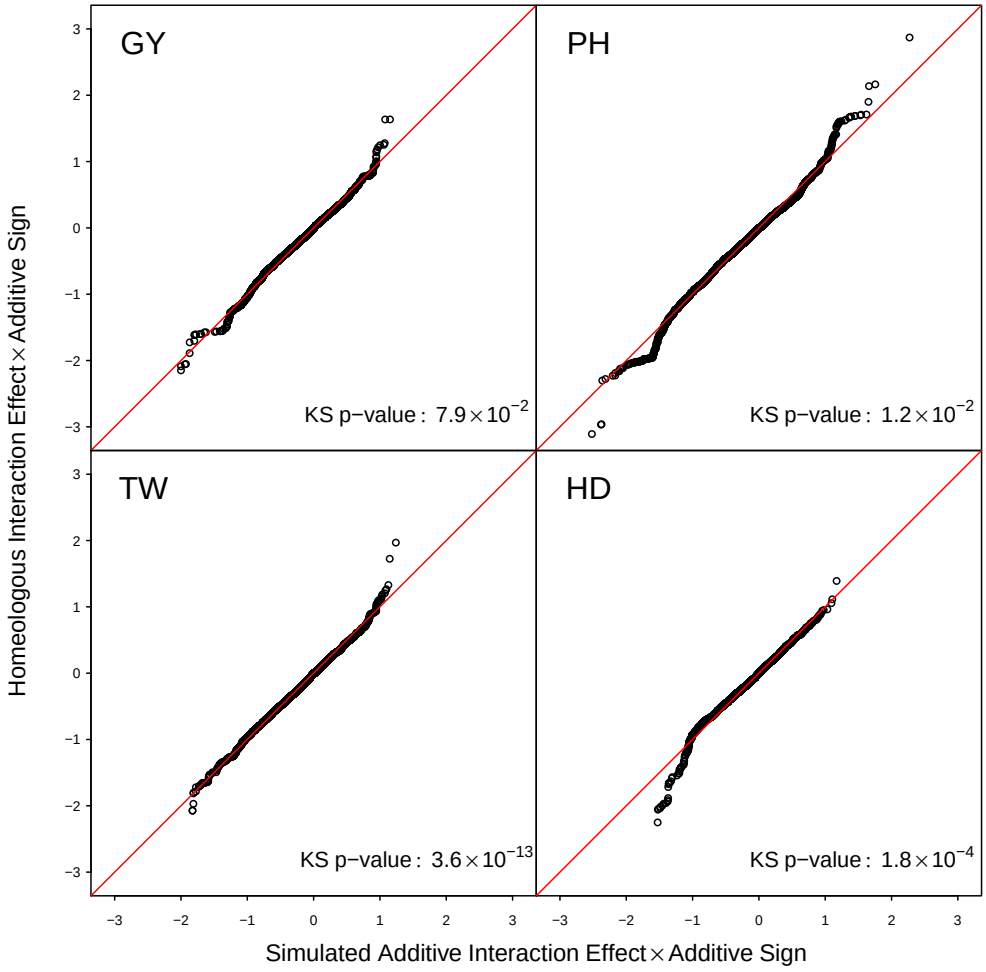


Figure 4.5: Quantile quantile plot of the ordered estimated homeologous interaction effects plotted against those from a simulated phenotype sampled to obtain no epistatic interactions. Interaction effects have been multiplied by the effect sign of the corresponding additive effects to emphasize the relationship between the additive and interaction effects. The p-value from a Kolmorgorov-Smirnov (KS) test is reported to determine if the sampled effect estimate distribution is different from that of the effect distribution estimated from the actual data. A deviation below the line on the bottom left of each graph (i.e. a low dropping tail) should indicate a less than additive epistatic pattern of subfunctionalization, whereas a deviation above the line in the upper right (i.e. a high rising head) should indicate a greater than additive epistasis pattern of homeologous overdominance.

in opposing directions.

To determine if the interaction effects were greater in magnitude than expected by chance, the ordered interaction effects from the true and simulated phenotypes were plotted against one another to form a quantile-quantile plot (Figure 4.5). The interaction effects were multiplied by the sign of the corresponding additive effects to highlight the direction of interaction effect relative to the additive effect. Interaction effect distributions were significantly different between the observed and strictly additive simulated data as determined by the Kolmogorov-Smirnov test (KS; $p < 0.05$) for all traits accept GY.

HD showed a pattern consistent with a subfunctionalization model, with a low dropping tail for interaction effects in the opposite direction than that of the corresponding additive effects. This indicates that the less than additive effects of some estimated interactions are greater than expected by additivity alone. PH showed some evidence of this pattern, but also demonstrated a greater than additive effect for positively related interaction effects. The LAVHAE orientation scheme may have selected the wrong marker coding for those marker sets, resulting in a d parameter greater than 1, or there are true greater than additive interaction responses for positive effect alleles. Greater than additive responses would be indicative of overdominance across homeologous loci. GY and TW showed little evidence of the less than additive pattern, yet TW did show this trend when the HTEV marker orientation was used (Supplemental Figures 4.12 and 4.13). These relationships were more pronounced when the markers were permuted to remove LD before simulating the data (Supplemental Figure 4.11). High LD between homeologous marker sets may result in dampening of the epistatic signal due to unbalanced or missing genotype classes.

These findings are further supported by comparing the homeologous interactions to the Within and Across interaction effect estimates. The Homeo marker set showed more severe less than additive epistasis than both Within and Across for HD but not the other traits (Supplementary Figures 4.14 and 4.15). The Within set had more severe less than additive interaction effects than the Homeo set for TW (Supplementary Figure 4.14), and the Across had more severe less than additive effects for PH (Supplementary Figure 4.15). Large or moderate effect negative epistasis is expected across subgenomes in allopolyploids, but it is unclear why this was also observed for the Within marker set for TW.

4.4.5 Homeologous model fit

Comparing variance component estimates across different unstructured covariance matrices can be misleading as variance components can be scaled by pulling a constant out of the covariance matrix. Additionally, variance partitioning is only reliable when the covariance matrices are truly independent (Vitezica et al. 2017; Huang and Mackay 2016; Jiang et al. 2017). Therefore, I do not make an attempt to discern meaning from the variance components *per se*, and instead focus the discussion on model fit diagnostics, as well as prediction accuracy from cross validation to determine the value of the predictive information included in the model.

All epistatic models using the $\{-1, 1\}$ marker parameterization provided a superior fit to the additive only model based on Akaike's Information Criterion for all traits (Table 4.5). These results were confirmed by a likelihood ratio test to determine if the epistatic variance component was zero for all traits. With the exception of the GY trait, all of the epistatic models using the $\{0, 1\}$ marker pa-

Table 4.5: Mixed model REML fit summaries of one additive and four epistasis models for four traits (GY, PH, TW and HD) in the CNLM population based on the $\{-1, 1\}$ marker parameterization using the LAVHAE marker orientation. Plot level heritabilities assuming genotype independence (i.i.d.) for each trait are shown underneath each trait name.

Trait		Additive	Pairwise	Homeo	Within	Across
GY	log$\mathcal{L}$	-48	-43	-42	-26	-23
h^2	parameters	28	29	29	29	29
0.30	AIC	153	144	141	110	104
	G	0.268[a] (12.59)[b]	0.203 (7.86)	0.204 (8.49)	0.133 (5.93)	0.13 (5.84)
	H		0.018 (3.04)	0.046 (3.29)***	0.093 (5.64)****	0.093 (5.77)****
	R	0.324 (61.86)[c]	0.322 (61.39)	0.323 (61.68)	0.321 (61.7)	0.321 (61.7)
PH	log$\mathcal{L}$	2237	2360	2314	2367	2374
h^2	parameters	26	27	27	27	27
0.73	AIC	-4423	-4665	-4574	-4680	-4694
	G	3.823 (20.75)	0.889 (6.46)	1.882 (11.66)	0.986 (7.35)	1.046 (7.81)
	H		0.478 (11.95)	0.914 (8.72)****	1.277 (11.67)****	1.253 (11.62)****
	R	0.135 (56.17)	0.132 (56.5)	0.133 (56.34)	0.133 (56.45)	0.133 (56.5)
HD	log$\mathcal{L}$	6343	6432	6404	6425	6444
h^2	parameters	27	28	28	28	28
0.53	AIC	-12631	-12808	-12751	-12794	-12831
	G	3.9 (21.16)	1.121 (7.3)	2.019 (12.03)	1.483 (9.25)	1.212 (8.29)
	H		0.451 (11.13)	0.857 (8.26)****	1.091 (10.01)****	1.202 (10.97)****
	R	0.054 (58.76)	0.053 (58.98)	0.053 (58.88)	0.053 (58.93)	0.053 (58.96)
TW	log$\mathcal{L}$	1547	1630	1608	1641	1632
h^2	parameters	28	29	29	29	29
0.79	AIC	-3037	-3203	-3159	-3224	-3205
	G	1.067 (16.66)	0.194 (4.47)	0.442 (8.35)	0.212 (4.81)	0.221 (4.79)
	H		0.184 (11.33)	0.346 (8.39)****	0.473 (10.95)****	0.473 (10.66)****
	R	0.2 (60.12)	0.195 (60.25)	0.198 (60.24)	0.197 (60.35)	0.197 (60.31)

[a]Variance component estimates reported for additive main effects (G) and epistatic interactions (H) are the ratios of the actual variance component to the residual variance component for ease of comparison.
[b]The variance component divided by their respective standard errors are shown in parentheses.
[c]The residual variance components, R, are the actual estimates from the centered and scaled data (refer to Table 2.2 for scaling coefficients).
*, **, ***, **** denote p-values of $p < 0.05$, $p < 0.01$, $p < 0.001$, $p < 10^{-6}$, respectively for the likelihood ratio test to determine if the epistatic variance component is zero.

rameterization also had non-zero variance components (Supplementary Table 4.9), but did not result in a better fit for any models or traits. All LAVHAE oriented marker sets had a better model fit than forcing allele effects to be either all positive (POS) or all negative (NEG), demonstrating that some information was gained from orienting markers (Supplementary Tables 4.10 and 4.11). The LAVHAE also resulted in a better model fit than the HTEV marker orientation scheme under

either marker parameterization (Supplementary Tables 4.12 and 4.13).

Despite resulting in a better fit, adding the epistatic interactions resulted in rather high correlations between additive and epistatic variance component estimates, as obtained from the average information matrix. Variance parameter estimate correlations between the additive and epistatic interactions consistently ranged from 0.57 to 0.65 for Pairwise, Within and Across models, but were notably lower for the Homeo epistatic set, ranging from 0.53 to 0.57. High correlations of variance estimates have been shown to be indicative of over fitting (Bates et al. 2015b; Bates et al. 2015a). However, if addition of epistatic interactions improves prediction accuracy of cross-validation, then the additional information they contain is warranted for inclusion in the model.

The Pairwise, Within and Across epistatic models outperformed the Homeo marker interaction set for all traits. This may be due to poor assignment of homeologous sets, or relatively fewer identifiable interactions and is discussed later (Section 4.4.7).

Table 4.6: Prediction accuracies of whole genome Additive and Pairwise epistasis, along with the Homeo, Within and Across genome marker sets for both $\{-1, 1\}$ and $\{0, 1\}$ marker coding using the LAVHAE marker orientation. The percentage of the non additive genetic predictability as relative to the the Pairwise model is shown in parentheses (equation 4.7).

LAVHAE	Additive	Pairwise	Homeo_{-11}	Homeo_{01}	Within_{-11}	Within_{01}	Across_{-11}	Across_{01}
GY	0.601[a]	0.604	0.606 (167%)	0.599 (-67%)	0.627 (867%)	0.600 (-33%)	0.630 (967%)	0.604 (100%)
PH	0.559	0.637	0.606 (60%)	0.580 (27%)	0.652 (119%)	0.570 (14%)	0.650 (117%)	0.584 (32%)
TW	0.515	0.576	0.560 (74%)	0.516 (2%)	0.596 (133%)	0.514 (-2%)	0.581 (108%)	0.525 (16%)
HD	0.664	0.712	0.692 (58%)	0.682 (38%)	0.710 (96%)	0.674 (21%)	0.722 (121%)	0.682 (38%)

[a] Mean Pearson correlation between predicted and observed genetic values across 10 random 5-fold cross-validation replications.

4.4.6 Genomic prediction

All epistatic models resulted in higher prediction accuracies for all traits other than GY, where only marginal increases were seen for certain marker interaction sets and parameterizations (Table 4.6).

The $\{-1, 1\}$ marker coding resulted in higher prediction accuracies with a mean increase of 0.045 over the $\{0, 1\}$ coding, and ranged from 0.007 to 0.084 higher accuracy. This increase may be due to choosing the wrong orientation using the $\{0, 1\}$ marker coding effects. While these two codings are equivalent for prediction using ordinary least squares, this does not appear to be the case for the mixed model genomic prediction environment. The discrepancy may lie in shrinkage of interaction effects, where the $\{0, 1\}$ marker coding should result in greater shrinkage than the $\{-1, 1\}$ marker coding. This can be seen from a simple example with one observation of each genotypic class in $\{bbcc, bbCC, BBcc, BBCC\}$. The $\{-1, 1\}$ coding would have an interaction predictor of $\{1, -1, -1, 1\}$, whereas the $\{0, 1\}$ coding would have an interaction predictor of $\{0, 0, 0, 1\}$. This results in different numbers of observations per interaction class, with the $\{0, 1\}$ coding contrasting 3 and 1, verses 2 and 2 for the $\{-1, 1\}$ coding. Therefore the shrinkage of the $\{0, 1\}$ coding should be greater than for the $\{-1, 1\}$ coding. Martini et al. (2017), also noted that the $\{-1, 1\}$ marker coding has a 50% chance of choosing the wrong marker orientation if chosen at random, whereas the $\{0, 1\}$ marker coding has a 75% chance of being the wrong marker orientation.

Choosing the marker orientation based on the LAVHAE scheme was better for all traits and marker sets for the $\{-1, 1\}$ coding, with a mean prediction accuracy increase of 0.025 (ranging from 0.004 to 0.053) over the all positive (POS) or all negative (NEG) orientations (Supplemental Tables 4.14 and 4.15). Choosing the

orientation had little to no effect on the $\{0, 1\}$ marker coding (mean 0.003, ranging from -0.002 to 0.011). It is unclear if my attempt to choose the marker orientation resulted in the biologically relevant orientation more often than would be expected by chance, as this resulted in higher accuracies for the $\{-1, 1\}$ but not the $\{0, 1\}$ marker coding.

The increase in genomic prediction accuracy by choosing the orientation using the LAVHAE scheme over the all positive (POS) or all negative (NEG) scheme suggests that information can be gained from orienting markers relative to one another (Supplementary Tables 4.15 and 4.14). This is further supported by a better model fit for the LAVHAE orientation. However, it is unclear what strategy should be used to orient pairs of markers. In this report, marker additive effects were forced to be either all positive or all negative to model the homeologous subfunctionalization hypothesis, but there may be more biologically relevant orientations not explored here. Martini et al. (2017) used a categorical interaction that included a predictor for each pairwise genotype, but that model was shown to be less predictive than the $\{-1, 1\}$ multiplicative model, perhaps due to more linearly dependent predictors assumed to have non-zero effects. How an optimal set of orientations might be obtained without losing biological meaning of the orientation warrants further investigation.

An indirect estimate of the proportion of non-additive genetic signal attributable to homeologous gene interaction was determined by taking the ratio of the percent increase in prediction accuracy of the Homeo, Within or Across prediction models from the additive model to the increase in prediction accuracy due to all pairwise interactions (equation 4.7). All three marker sets resulted in higher genomic prediction accuracy than the additive only GBLUP model (G) when the

$\{-1, 1\}$ marker coding was used. The homeologous marker interaction set explained between 58% and 167% of the additional genetic signal from the additive model. This result supports the idea that homeologous interactions are an important feature in the wheat genome. Conversely, Within and Across epistatic marker sets always resulted in a higher increase in genomic prediction accuracy relative to the Homeo marker set for all traits. This may suggest that the homeologous marker interactions are the least important relative to other epistatic interactions within and across the subgenomes, but could also be due to the paucity of these interactions relative to all possible two-way interactions, as previously discussed.

Another explanation might be provided by the relatively higher degree of LD across Homeo marker sets than found for the Within or Across marker sets. The Within and Across marker interaction sets also resulted in a larger number of unique interaction predictors because they were randomly selected across chromosomes. Homeologous marker sets were selected next to one another along syntenic regions of homeologous chromosome, and more often shared two of the three homeoallelic markers (Supplemental Figure 4.16). The Within and Across sets appear to have sampled the entire genome better than selecting only homeologous loci, as they track more unique pairs of genomic regions.

4.4.7 Homeologous LD

The superiority of the Within and Across genomic prediction models to the Homeo genomic prediction model may indicate that homeologous interactions are relatively less important than other sets of interacting loci. However, homeologous marker sets had a much higher tendency to be co-inherited together, as seen by relatively higher standardized linkage disequilibrium values, D' (Lewontin 1964),

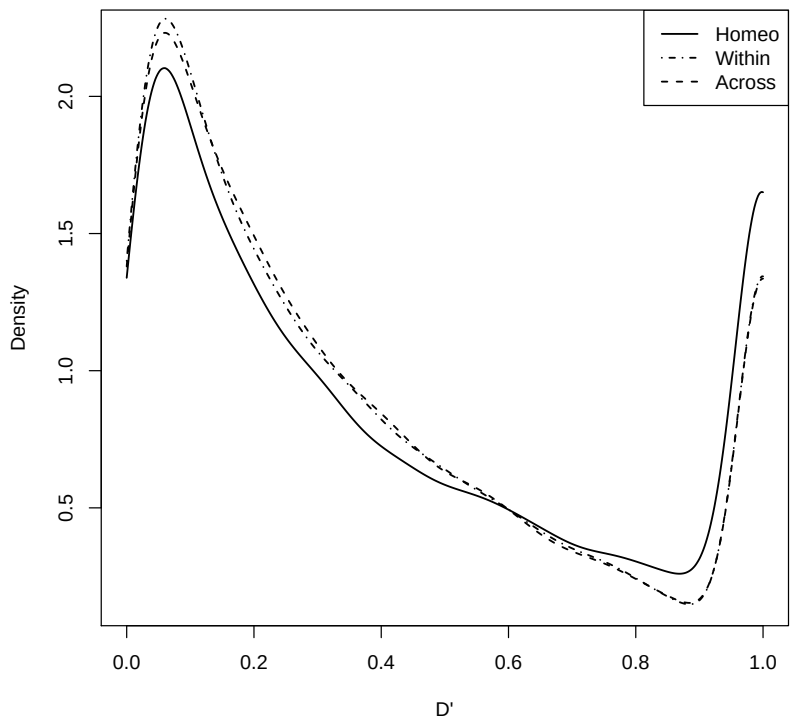


Figure 4.6: Smoothed densities of standardized D' statistics of linkage disequilibrium for expected and observed joint allele frequencies for Homeo, Within and Across marker sets. Kolmogorov-Smirnov (KS) tests were used to determine if the distribution of LD differed between Homeo and Within (KS test p-value = 1.1×10^{-6}) or Across (KS test p-value = 2.3×10^{-13}) marker sets.

than observed for either Within (KS test p-value = 1.1×10^{-6}) or Across (KS test p-value = 2.3×10^{-13}) marker sets (Figure 4.6). The greater fixation of allele pairs at homeologous regions may explain the lack of increased prediction accuracy of the Homeo marker set, but this may not diminish the importance of homeologous interactions. As sets of interactions are fixed within the population, the epistatic variance becomes additive (Hill, Goddard, and Visscher 2008). The higher degree of LD, *per se*, may indicate the importance of homeologous interactions.

The Green Revolution dwarfing genes are an excellent example of how pairs of homeoalleles may become fixed, or develop a tendency for co-inheritance under selection. In this example, the desirable phenotype is a semi-dwarf, due to its resistance to lodging. Therefore, wildtype *Rht-1B* alleles will usually be paired with a GA-insensitive *Rht-1D* dwarfing allele, while wildtype *Rht-1D* alleles will usually be found with a GA-insensitive *Rht-1B* dwarfing allele to confer the desirable semi-dwarf phenotype. The CNLM_Rht1 set had a large standardized D' value of 0.66, indicating that pairs of alleles were being fixed in the population, despite the apparent lack of high LD between the B genome marker and the *Rht-1B* gene.

I recognize that it is also possible that the higher degree of LD observed between homeologous marker pairs could be due to misalignment of markers to the wrong subgenome. Markers assigned to the wrong homeolog would appear in high LD simply because they are physically located near their homeologous partner on the same chromosome. I used strict filtering parameters to reduce the likelihood of misalignment. This included a threshold on observed heterozygosity in the population, which could indicate alignment to more than one subgenome.

4.5 Conclusion

Our results indicate that homeoallelic interactions do not account for a large portion of the non-additive genetic variance in the Cornell soft winter wheat breeding population. While they do to contribute to the genetic variance of the population as evidenced by a better model fit and higher prediction accuracy over the additive model, sampling interactions across non-syntenic regions was superior for all traits examined. Homeologous interactions appear to make up the minority of epistatic

interactions within this population.

Wagner (2005) suggested that there are two potential drivers of less than additive (Eshed and Zamir 1996) or synergistic (Segre et al. 2005) epistasis. These drivers are i) functional redundancy, as might be expected across homeologous loci, and ii) distributed robustness of function, in which there can be are many pathways that can acheive the same outcome. My observation that most epistasis is not due to homeologous interactions is supported by the findings of Jannink et al. (2009), who found the synergistic epistasis signal in a wheat dataset to be indicative of Wagner's distributed hypothesis, and not of the redundancy hypothesis.

HD showed the greatest evidence of subfunctionalization. The LAVHAE orientation appeared to be effective for the HD trait, but it is unclear if this marker orientation scheme was effective for PH, which also demonstrated a greater than additive effect. TW showed little pattern of subfunctionalization, despite having significant epistatic interactions. GY showed no evidence of being affected which may simply be due to the highly polygenic nature of the trait. Essentially all functional differences in the population should contribute to GY, and it may be that the effects of epistatic interactions are too small to detect. Large effect homeologous interactions for GY are likely to have allele pairs that have been fixed or are well on their way to fixation in the elite wheat genotypes as a consequence of modern plant breeding. The fixation of the semi-dwarf phenotype provides a profound example where specific pairs of homeoalleles result in drastic increases in grain production under modern agriculture.

The apparent lack of substantial interactions across homeoallelic loci may be explained by several factors. It may be that there are few differences in protein function or expression across the three subgenomes, although this seems unlikely

given mounting evidence that homeologous copies are differentially expressed in time, tissue and environment (Adams et al. 2003; Liu and Adams 2007; Liu, Baute, and Adams 2011; Chaudhary et al. 2009; Pfeifer et al. 2014; Liu et al. 2015; Mutti, Bhullar, and Gill 2017; Zhang et al. 2016). I was unable to assign homeologous pairs to all genes within the genome, suggesting that many of these potential sites for interacting loci were lost during, or shortly after, the polyploidization event. Rapid loss of genetic material due to genome shock (McClintock 1984) is common in newly synthesized allopolyploids (Chen and Ni 2006), as has been shown in synthetic allopolyploid wheat (Ozkan, Levy, and Feldman 2001; Kashkush, Feldman, and Levy 2002). Other interacting loci may have undergone epigenetic (Comai 2000; Lee and Chen 2001; Comai et al. 2003) or transposon induced silencing of one or more homeoalleles (Kashkush, Feldman, and Levy 2003; Wang et al. 2004).

It may also be that important homeoallele pairs have been effectively fixed in the CNLM breeding population, as evidenced by a higher degree of LD between homeologous markers compared to markers sampled from non-homeologous regions. Finally, determining homeologous regions is relatively simple using gene positions and orientation, but tagging those regions with markers that are informative of interacting loci appears to be a challenge, even for well known loci such as *Rht-1*. Marker imputation using a diverse panel of highly sequenced individuals may increase the marker density and the ability to identify interacting loci. As sequencing becomes more affordable, higher read coverage will allow for better genotyping, and should produce more high quality markers. Other approaches, such as the use of haplotypes, should be developed to assign informative homeologous locus indicators, as opposed to simply using physical position of markers or lumping markers on the same subgenome together. Smaller units of homeologous chromatin may have higher power to detect these interactions.

4.6 Supplementary Materials

Table 4.7: ANOVA table for *Rht-1B* and *Rht-1D* linked GBS markers and their epistatic interaction for plant height (cm) in 158 RIL lines derived from NY91017-8080 × Caledonia.

Source	df	SS	MS	F value	$-\log_{10}$(p-value)
SNP36427	1	7065	7065	53.5	10.9
SNP11172	1	7391	7391	56.0	11.3
SNP36427:SNP11172	1	1243	1243	9.4	2.6
Residuals	154	20323	132		

Table 4.8: Table of genotype frequencies for the *Rht-1* linked homeologous markers in the CNLM population. The margins indicate the marker allele frequencies.

	S4D_PART1_10982050$^-$	S4D_PART1_10982050$^+$	
S4B_PART1_38624956$^-$	0.022	0.095	0.117
S4B_PART1_38624956$^+$	0.525	0.357	0.883
	0.547	0.452	$D' = 0.66$

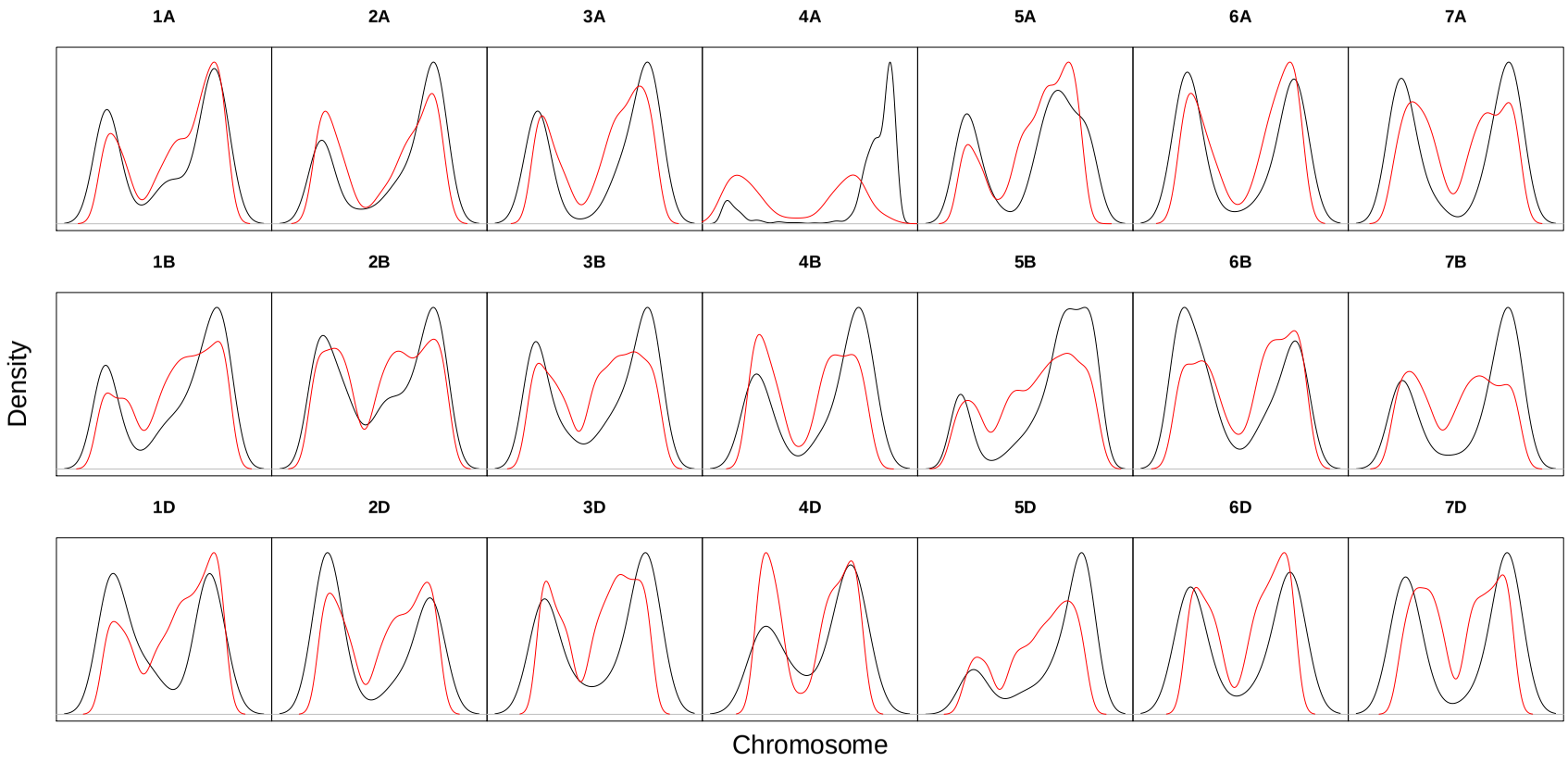


Figure 4.7: Smoothed densities of GBS markers (black) and genes (red) along the 21 wheat chromosomes in the CNLM population.

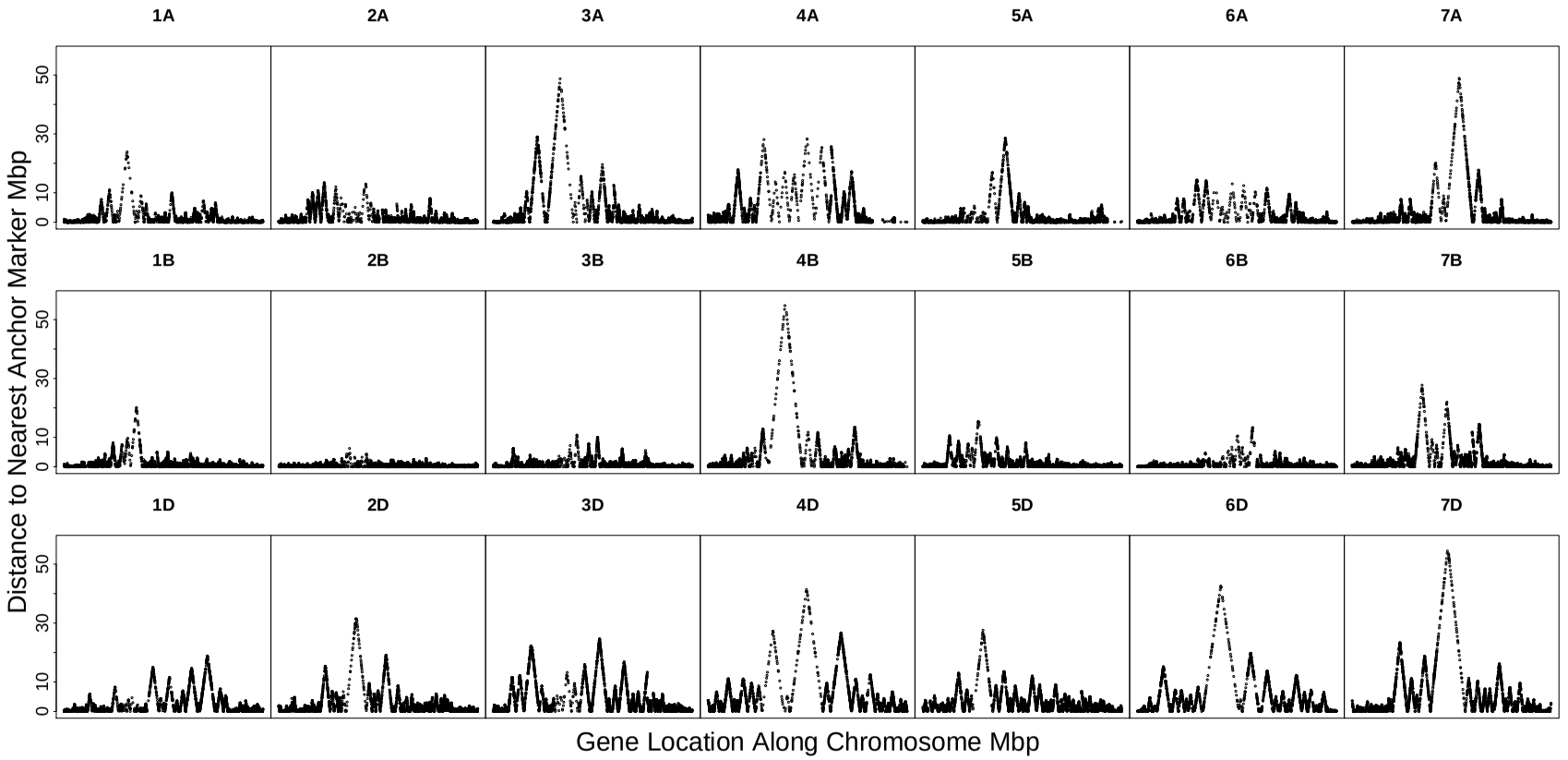


Figure 4.8: Distance of genes from their nearest GBS anchor marker along the 21 wheat chromosomes in the CNLM population.

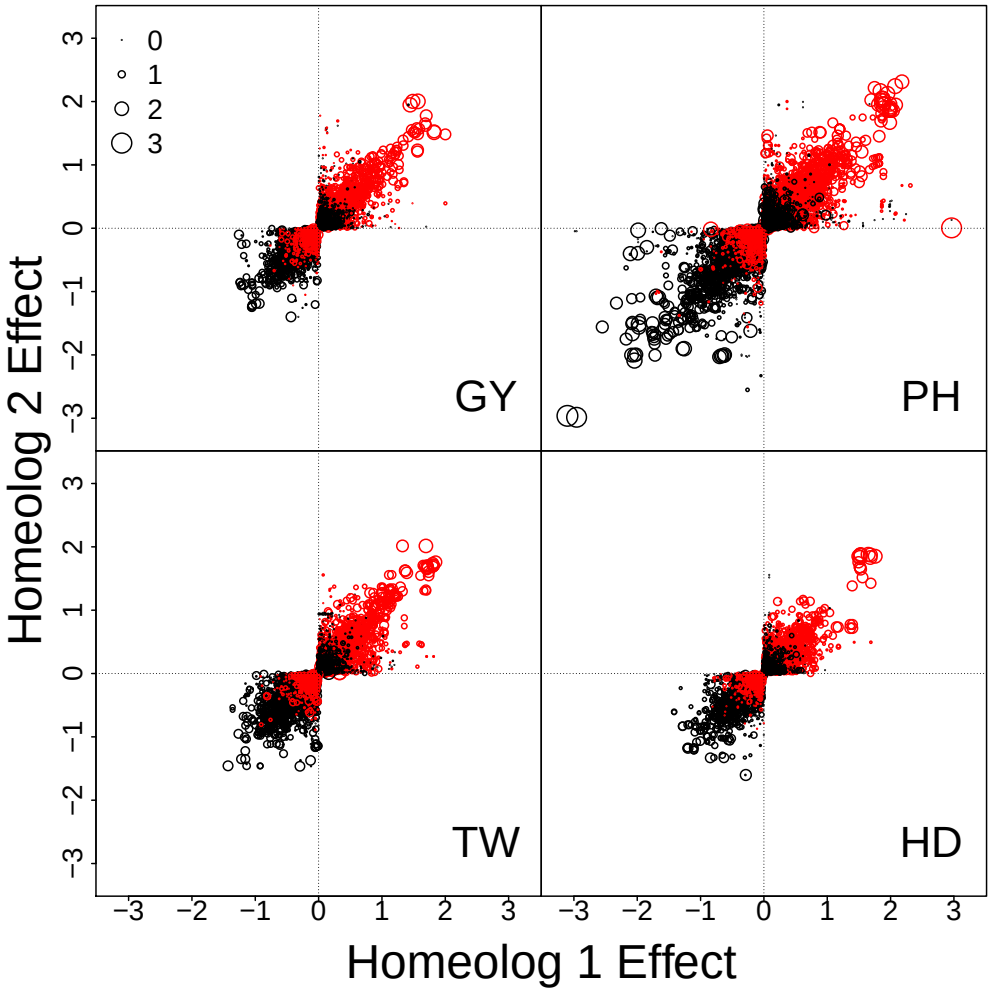


Figure 4.9: LAVHAE oriented homeologous marker pair additive effects with point size representing the magnitude of the two-way homeologous interaction effect, and the color denoting the direction of that effect where black is positive and red is negative. Four traits, GY, PH, TW and HD, are shown.

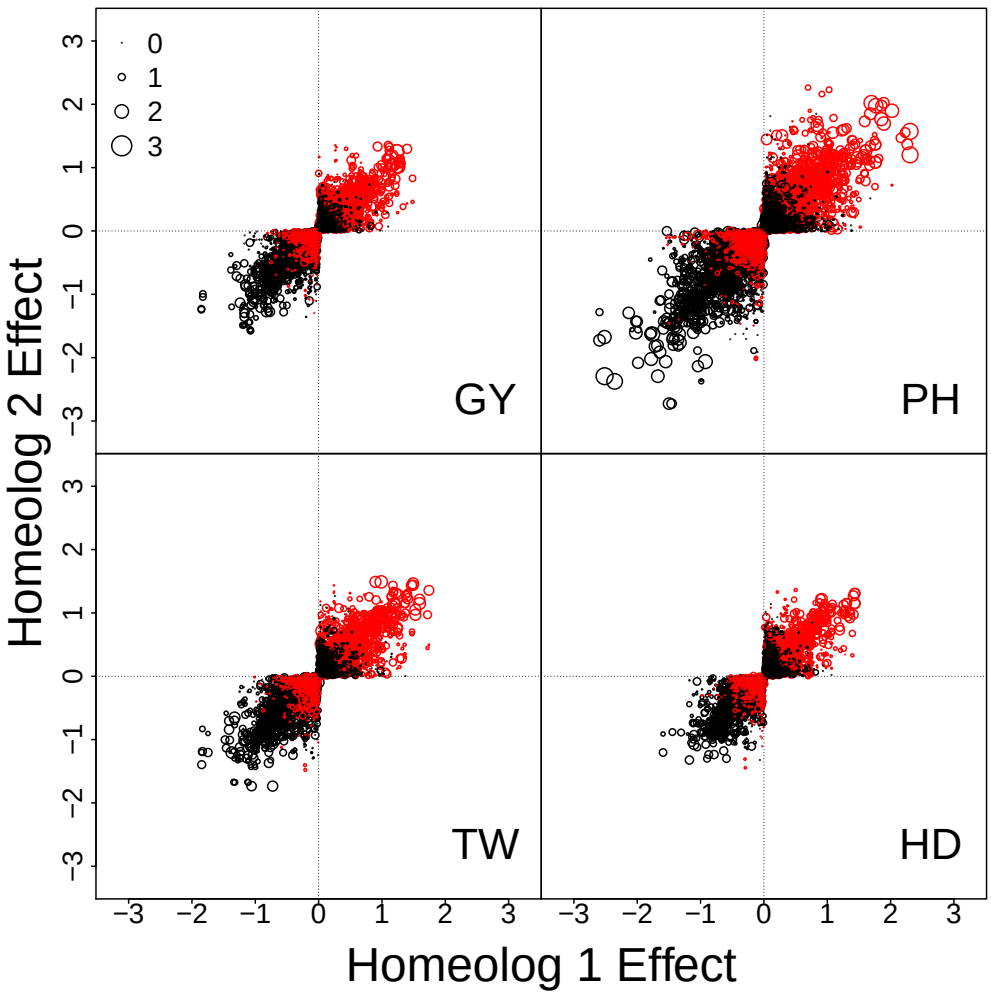


Figure 4.10: LAVHAE oriented homeologous marker pair additive effects with point size representing the magnitude of the two-way homeologous interaction effect, and the color denoting the direction of that effect where black is positive and red is negative. Four simulated phenotypes sampled to obtain no epistatic interactions, GY, PH, TW and HD, are shown.

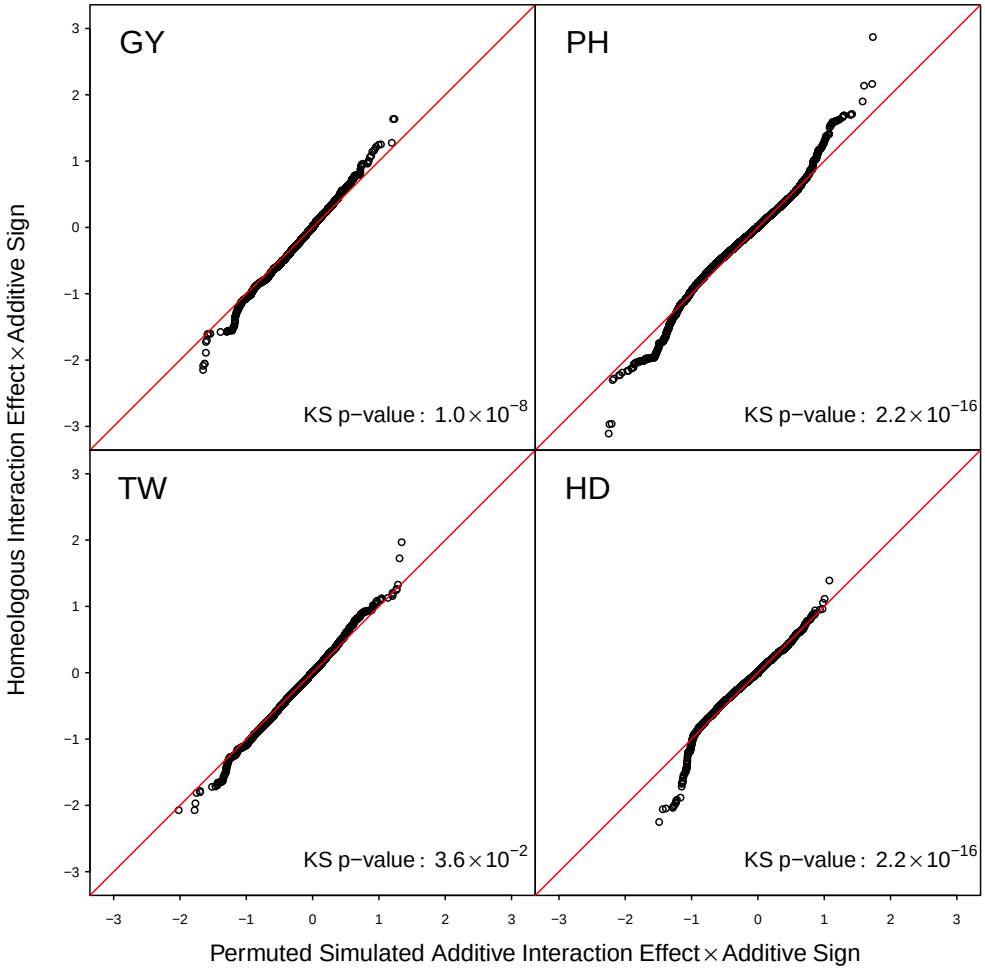


Figure 4.11: Quantile quantile plot of the ordered estimated homeologous interaction effects plotted against those from a simulated phenotype sampled to obtain no epistatic interactions using the LAVHAE marker orientation. Markers scores were permuted before simulation of the phenotype to remove LD between markers. Interaction effects have been multiplied by the effect sign of the corresponding additive effects to emphasize the relationship between the additive and interaction effects.

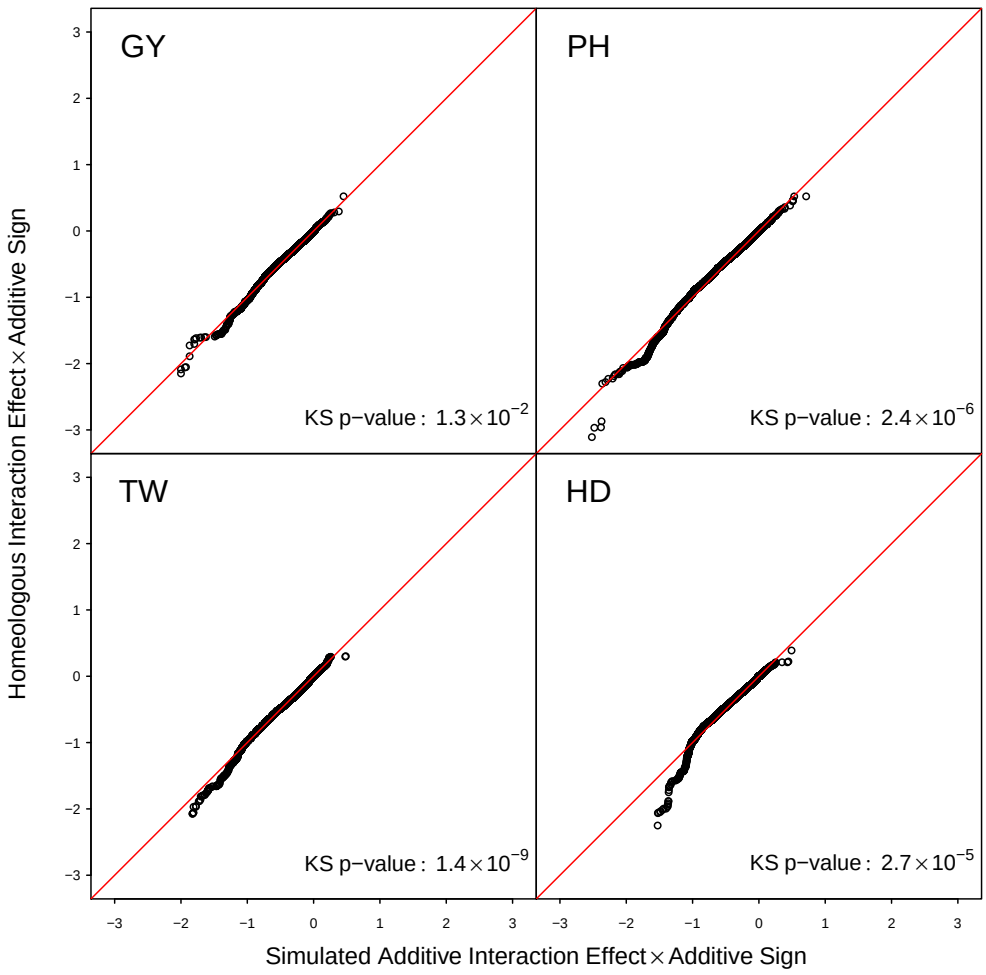


Figure 4.12: Quantile quantile plot of the ordered estimated homeologous interaction effects plotted against those from a simulated phenotype sampled to obtain no epistatic interactions using the HTEV marker orientation. Interaction effects have been multiplied by the effect sign of the corresponding additive effects to emphasize the relationship between the additive and interaction effects.

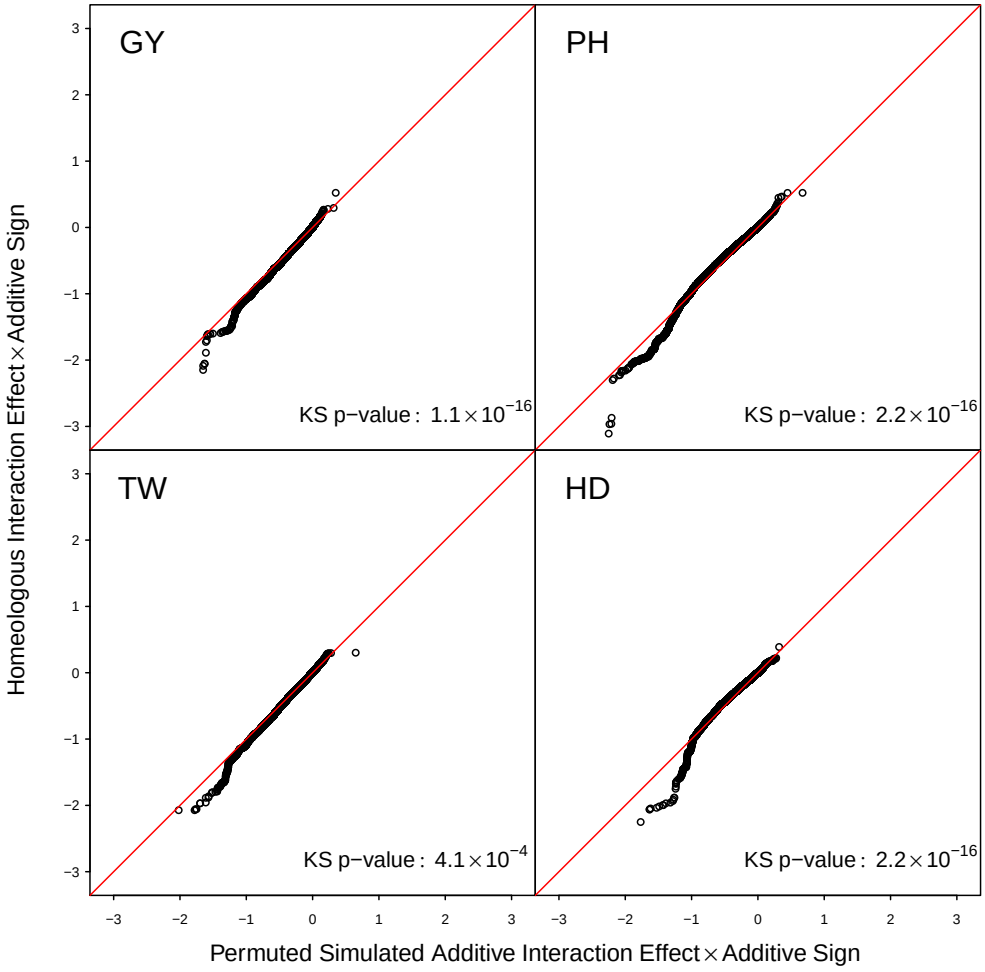


Figure 4.13: Quantile quantile plot of the ordered estimated homeologous interaction effects plotted against those from a simulated phenotype sampled to obtain no epistatic interactions using the HTEV marker orientation. Markers scores were permuted before simulation of the phenotype to remove LD between markers. Interaction effects have been multiplied by the effect sign of the corresponding additive effects to emphasize the relationship between the additive and interaction effects.

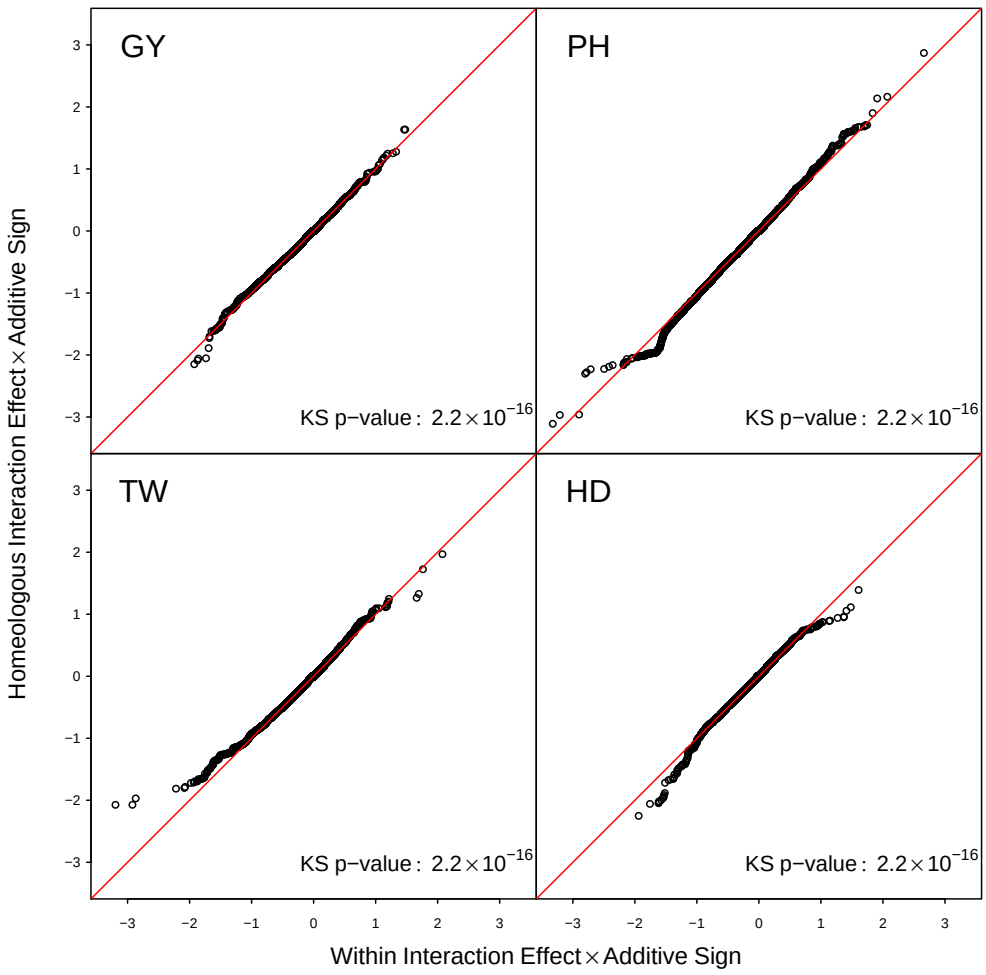


Figure 4.14: Quantile quantile plot of the ordered estimated homeologous interaction effects plotted against those from marker sets sampled within subgenome chromosomes (Within) using the LAVHAE . Interaction effects have been multiplied by the effect sign of the corresponding additive effects to emphasize the relationship between the additive and interaction effects.

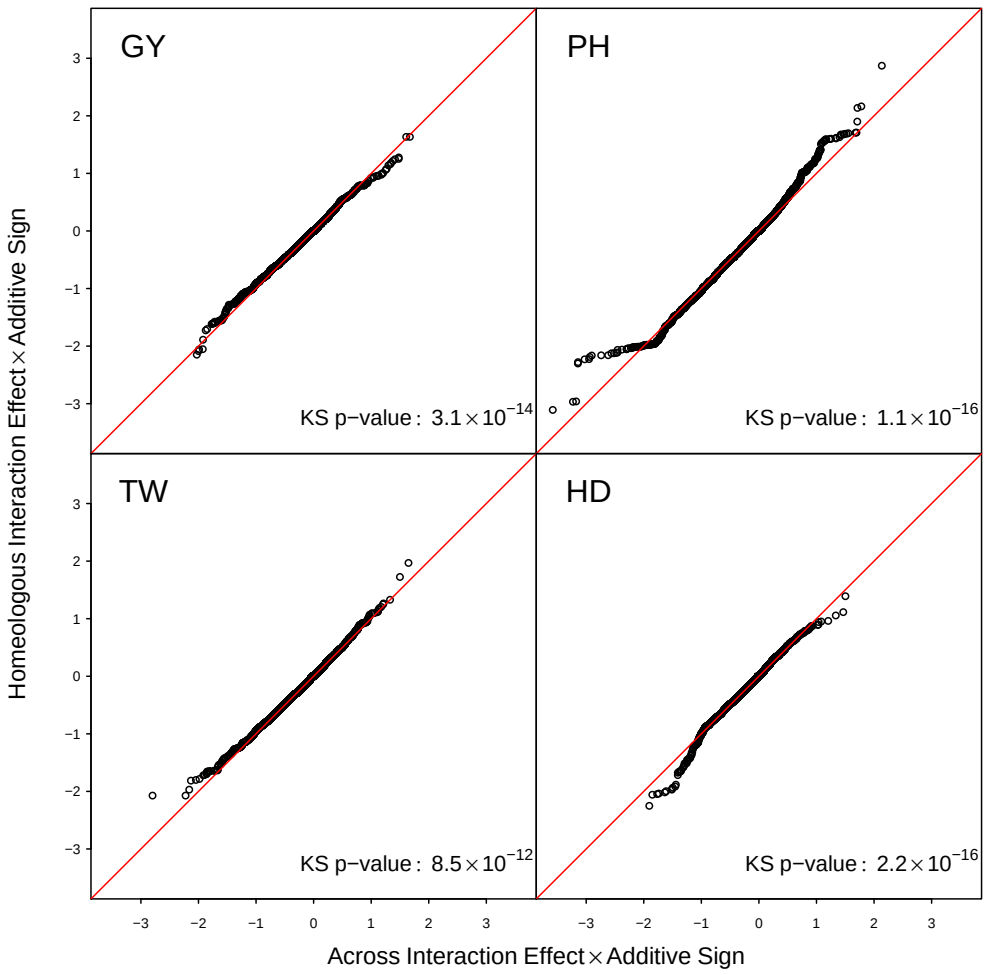


Figure 4.15: Quantile quantile plot of the ordered estimated homeologous interaction effects plotted against those from marker sets sampled across non-syntenic subgenome chromosomes (Across) using the LAVHAE marker orientation. Interaction effects have been multiplied by the effect sign of the corresponding additive effects to emphasize the relationship between the additive and interaction effects.

Table 4.9: Mixed model REML fit summaries of one additive and four epistasis models for 4 traits (GY, PH, TW and HD) in the CNLM population based on the $\{0, 1\}$ marker parameterization using the LAVHAE marker orientation.

Trait		Homeo	Within	Across
GY	log$\mathcal{L}$	-48	-47	-42
	parameters	29	29	29
	AIC	155	152	143
	G	0.267[a] (7.6)[b]	0.207 (5.5)	0.146 (4.16)
	H	0 (0.01)	0.054 (1.73)	0.108 (3.39)***
	R	0.324 (61.81)[c]	0.324 (61.77)	0.324 (61.8)
PH	log$\mathcal{L}$	2282	2268	2285
	parameters	27	27	27
	AIC	-4510	-4482	-4516
	G	1.198 (5.03)	1.766 (6.95)	1.177 (5.02)
	H	1.981 (8.36)****	1.592 (6.95)****	2.051 (8.66)****
	R	0.134 (56.23)	0.134 (56.24)	0.134 (56.24)
HD	log$\mathcal{L}$	6382	6364	6379
	parameters	28	28	28
	AIC	-12709	-12673	-12702
	G	1.51 (6.14)	2.077 (7.82)	1.659 (6.67)
	H	1.781 (7.73)****	1.358 (6.09)****	1.68 (7.36)****
	R	0.053 (58.84)	0.054 (58.78)	0.054 (58.81)
TW	log$\mathcal{L}$	1560	1555	1567
	parameters	29	29	29
	AIC	-3061	-3052	-3076
	G	0.553 (5.88)	0.659 (6.68)	0.498 (5.57)
	H	0.414 (5.04)****	0.331 (4.06)***	0.482 (5.85)****
	R	0.199 (60.11)	0.199 (60.1)	0.198 (60.13)

[a]Variance component estimates reported for additive main effects (G) and epistatic interactions (H) are the ratios of the actual variance component to the residual variance component for ease of comparison.
[b]The variance component divided by their respective standard errors are shown in parentheses.
[c]The residual variance components, R, are the actual estimates from the centered and scaled data (refer to Table 2.2 for scaling coefficients).
*, **, ***, **** denote p-values of $p < 0.05$, $p < 0.01$, $p < 0.001$, $p < 10^{-6}$, respectively for the likelihood ratio test to determine if the epistatic variance component is zero.

Table 4.10: Mixed model REML fit summaries of one additive and four epistasis models for 4 traits (GY, PH, TW and HD) in the CNLM population based on the $\{-1, 1\}$ marker parameterization using the POS marker orientation.

Trait		Homeo	Within	Across
GY	log$\mathcal{L}$	-48	-41	-40
	parameters	29	29	29
	AIC	154	140	138
	G	0.257[a] (10.31)[b]	0.191 (7.44)	0.186 (7.32)
	H	0.008 (0.75)	0.052 (3.45)***	0.054 (3.61)***
	R	0.324 (61.7)[c]	0.323 (61.64)	0.323 (61.64)
PH	log$\mathcal{L}$	2287	2323	2326
	parameters	27	27	27
	AIC	-4521	-4593	-4598
	G	2.316 (13.04)	1.507 (9.34)	1.551 (9.59)
	H	0.705 (7.3)****	1.056 (9.85)****	1.036 (9.72)****
	R	0.134 (56.29)	0.133 (56.38)	0.133 (56.4)
HD	log$\mathcal{L}$	6379	6393	6415
	parameters	28	28	28
	AIC	-12701	-12730	-12774
	G	2.547 (13.61)	2.017 (10.81)	1.689 (9.94)
	H	0.601 (6.43)****	0.848 (8.04)****	0.982 (9.26)****
	R	0.053 (58.83)	0.053 (58.87)	0.053 (58.92)
TW	log$\mathcal{L}$	1589	1599	1604
	parameters	29	29	29
	AIC	-3120	-3139	-3150
	G	0.554 (9.49)	0.437 (7.44)	0.395 (7.02)
	H	0.282 (7.22)****	0.354 (8.36)****	0.368 (8.71)****
	R	0.198 (60.18)	0.197 (60.2)	0.197 (60.21)

[a]Variance component estimates reported for additive main effects (G) and epistatic interactions (H) are the ratios of the actual variance component to the residual variance component for ease of comparison.
[b]The variance component divided by their respective standard errors are shown in parentheses.
[c]The residual variance components, R, are the actual estimates from the centered and scaled data (refer to Table 2.2 for scaling coefficients).
*, **, ***, **** denote p-values of $p < 0.05$, $p < 0.01$, $p < 0.001$, $p < 10^{-6}$, respectively for the likelihood ratio test to determine if the epistatic variance component is zero.

Table 4.11: Mixed model REML fit summaries of one additive and four epistasis models for 4 traits (GY, PH, TW and HD) in the CNLM population based on the $\{-1, 1\}$ marker parameterization using the NEG marker orientation.

Trait		Homeo	Within	Across
GY	log$\mathcal{L}$	-46	-38	-35
	parameters	29	29	29
	AIC	151	134	129
	G	0.236[a] (9.44)[b]	0.181 (7.35)	0.178 (7.35)
	H	0.022 (1.86)	0.058 (3.9)***	0.06 (4.1)****
	R	0.324 (61.71)[c]	0.323 (61.68)	0.322 (61.68)
PH	log$\mathcal{L}$	2293	2336	2342
	parameters	27	27	27
	AIC	-4532	-4619	-4629
	G	2.235 (12.79)	1.428 (9.19)	1.464 (9.46)
	H	0.746 (7.52)****	1.061 (10.06)****	1.038 (10.07)****
	R	0.134 (56.3)	0.133 (56.39)	0.133 (56.42)
HD	log$\mathcal{L}$	6380	6402	6409
	parameters	28	28	28
	AIC	-12704	-12747	-12762
	G	2.48 (13.41)	1.88 (10.5)	1.753 (10.09)
	H	0.626 (6.71)****	0.895 (8.59)****	0.95 (9.02)****
	R	0.053 (58.83)	0.053 (58.89)	0.053 (58.9)
TW	log$\mathcal{L}$	1580	1605	1601
	parameters	29	29	29
	AIC	-3101	-3153	-3144
	G	0.614 (9.96)	0.373 (6.78)	0.388 (6.71)
	H	0.241 (6.4)****	0.374 (8.94)****	0.367 (8.59)****
	R	0.199 (60.15)	0.198 (60.22)	0.197 (60.21)

[a]Variance component estimates reported for additive main effects (G) and epistatic interactions (H) are the ratios of the actual variance component to the residual variance component for ease of comparison.
[b]The variance component divided by their respective standard errors are shown in parentheses.
[c]The residual variance components, R, are the actual estimates from the centered and scaled data (refer to Table 2.2 for scaling coefficients).
*, **, ***, **** denote p-values of $p < 0.05$, $p < 0.01$, $p < 0.001$, $p < 10^{-6}$, respectively for the likelihood ratio test to determine if the epistatic variance component is zero.

Table 4.12: Mixed model REML fit summaries of one additive and four epistasis models for 4 traits (GY, PH, TW and HD) in the CNLM population based on the $\{-1, 1\}$ marker parameterization using the HTEV marker orientation.

trait		Homeo	Within	Across
GY	log$\mathcal{L}$	-46	-34	-30
	parameters	29	29	29
	AIC	151	127	118
	G	0.233[a] (9.23)[b]	0.165 (6.86)	0.151 (6.45)
	H	0.025 (1.97)	0.071 (4.56)****	0.079 (5)****
	R	0.323 (61.65)[c]	0.322 (61.66)	0.322 (61.67)
PH	log$\mathcal{L}$	2300	2355	2357
	parameters	27	27	27
	AIC	-4546	-4655	-4659
	G	2.052 (12.02)	1.101 (7.81)	1.142 (7.99)
	H	0.84 (8.12)****	1.227 (11.24)****	1.209 (11.09)****
	R	0.133 (56.32)	0.133 (56.43)	0.133 (56.46)
HD	log$\mathcal{L}$	6397	6410	6423
	parameters	28	28	28
	AIC	-12738	-12764	-12790
	G	2.13 (12.27)	1.62 (9.54)	1.395 (8.69)
	H	0.808 (7.9)****	1.029 (9.43)****	1.139 (10.18)****
	R	0.053 (58.88)	0.053 (58.91)	0.053 (58.94)
TW	log$\mathcal{L}$	1599	1623	1623
	parameters	29	29	29
	AIC	-3140	-3189	-3187
	G	0.476 (8.51)	0.283 (5.73)	0.267 (5.4)
	H	0.335 (7.92)****	0.435 (10.13)****	0.45 (10.15)****
	R	0.198 (60.2)	0.197 (60.29)	0.197 (60.28)

[a]Variance component estimates reported for additive main effects (G) and epistatic interactions (H) are the ratios of the actual variance component to the residual variance component for ease of comparison.
[b]The variance component divided by their respective standard errors are shown in parentheses.
[c]The residual variance components, R, are the actual estimates from the centered and scaled data (refer to Table 2.2 for scaling coefficients).
*, **, ***, **** denote p-values of $p < 0.05$, $p < 0.01$, $p < 0.001$, $p < 10^{-6}$, respectively for the likelihood ratio test to determine if the epistatic variance component is zero.

Table 4.13: Mixed model REML fit summaries of one additive and four epistasis models for 4 traits (GY, PH, TW and HD) in the CNLM population based on the $\{0, 1\}$ marker parameterization using the HTEV marker orientation.

trait		Homeo	Within	Across
GY	log$\mathcal{L}$	-48	-48	-48
	parameters	29	29	29
	AIC	155	155	155
	G	0.268[a] (12.59)[b]	0.268 (12.59)	0.268 (12.59)
	H	0	0	0
	R	0.324 (61.86)[c]	0.324 (61.86)	0.324 (61.86)
PH	log$\mathcal{L}$	2260	2246	2248
	parameters	27	27	27
	AIC	-4466	-4438	-4443
	G	1.981 (7.41)	2.84 (9.98)	2.502 (8.49)
	H	1.423 (6.05)****	0.806 (3.68)***	1.081 (4.44)***
	R	0.134 (56.2)	0.134 (56.19)	0.134 (56.19)
HD	log$\mathcal{L}$	6358	6350	6356
	parameters	28	28	28
	AIC	-12660	-12643	-12656
	G	2.528 (9.24)	2.937 (10.1)	2.468 (8.5)
	H	1.052 (4.83)****	0.749 (3.44)***	1.16 (4.78)****
	R	0.054 (58.79)	0.054 (58.76)	0.054 (58.78)
TW	log$\mathcal{L}$	1552	1547	1549
	parameters	29	29	29
	AIC	-3046	-3036	-3041
	G	0.746 (7.61)	0.992 (9.51)	0.857 (8.24)
	H	0.264 (3.38)***	0.064 (0.87)	0.183 (2.26)*
	R	0.199 (60.1)	0.199 (60.08)	0.199 (60.09)

[a]Variance component estimates reported for additive main effects (G) and epistatic interactions (H) are the ratios of the actual variance component to the residual variance component for ease of comparison.
[b]The variance component divided by their respective standard errors are shown in parentheses.
[c]The residual variance components, R, are the actual estimates from the centered and scaled data (refer to Table 2.2 for scaling coefficients).
*, **, ***, **** denote p-values of $p < 0.05$, $p < 0.01$, $p < 0.001$, $p < 10^{-6}$, respectively for the likelihood ratio test to determine if the epistatic variance component is zero.

Table 4.14: Prediction accuracies of Homeo, Within and Across genome marker sets for both $\{-1, 1\}$ and $\{0, 1\}$ marker coding using POS marker orientation.

POS	$\text{Homeo}_{\text{-}11}$	Homeo_{01}	$\text{Within}_{\text{-}11}$	Within_{01}	$\text{Across}_{\text{-}11}$	Across_{01}
GY	0.599^a	0.599	0.607	0.600	0.607	0.599
PH	0.583	0.573	0.607	0.568	0.612	0.576
TW	0.535	0.518	0.543	0.514	0.547	0.524
HD	0.681	0.681	0.688	0.670	0.698	0.671

[a] mean Pearson correlation between predicted and observed genetic values across 10 random 5-fold cross-validation replications.

Table 4.15: Prediction accuracies of Homeo, Within and Across genome marker sets for both $\{-1, 1\}$ and $\{0, 1\}$ marker coding using NEG marker orientation.

NEG	$\text{Homeo}_{\text{-}11}$	Homeo_{01}	$\text{Within}_{\text{-}11}$	Within_{01}	$\text{Across}_{\text{-}11}$	Across_{01}
GY	0.602^a	0.599	0.612	0.599	0.615	0.600
PH	0.589	0.582	0.620	0.565	0.615	0.579
TW	0.535	0.513	0.555	0.510	0.546	0.519
HD	0.676	0.671	0.698	0.671	0.697	0.680

[a] mean Pearson correlation between predicted and observed genetic values across 10 random 5-fold cross-validation replications.

Table 4.16: Prediction accuracies of Homeo, Within and Across genome marker sets for both $\{-1, 1\}$ and $\{0, 1\}$ marker coding using HTEV marker orientation.

HTEV	$\text{Homeo}_{\text{-}11}$	Homeo_{01}	$\text{Within}_{\text{-}11}$	Within_{01}	$\text{Across}_{\text{-}11}$	Across_{01}
GY	0.601^a	0.601	0.616	0.600	0.621	0.600
PH	0.591	0.565	0.640	0.557	0.633	0.558
TW	0.548	0.513	0.572	0.513	0.568	0.513
HD	0.688	0.669	0.700	0.666	0.706	0.667

[a] mean Pearson correlation between predicted and observed genetic values across 10 random 5-fold cross-validation replications.

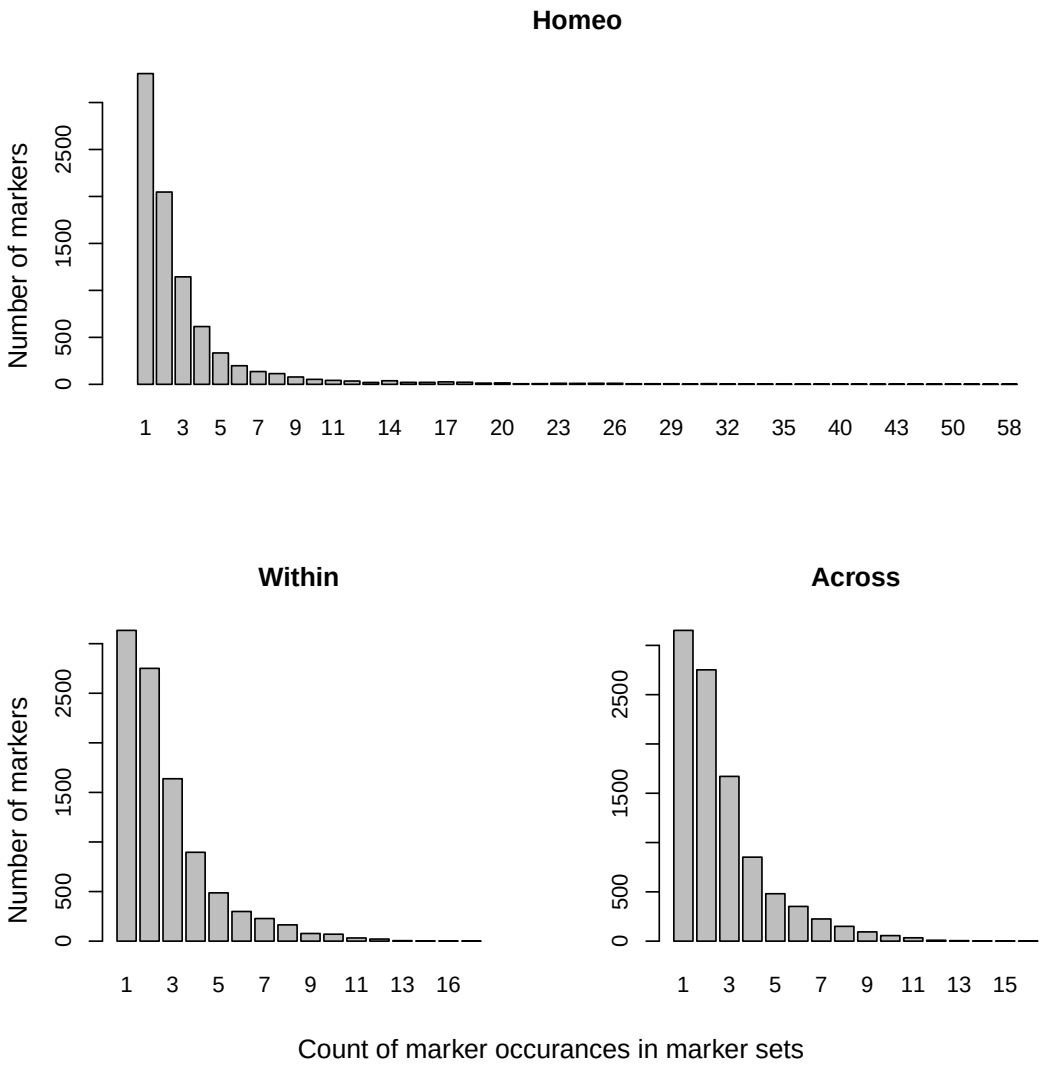


Figure 4.16: Distribution of the number of marker occurrences in marker sets. An occurrence of 1 indicates that a marker was only included in one marker set, whereas an occurrence of 10 would indicate that the marker was included in 10 marker sets.

CHAPTER 5

A LOW RESOLUTION EPISTASIS MAPPING APPROACH FOR IDENTIFYING CHROMOSOME ARM INTERACTIONS IN ALLOHEXAPLOID WHEAT

5.1 Introduction

Epistasis is the interaction of alleles, or variants, at two or more loci. Early observations of epistasis by William Bateson (2007) were mostly qualitative, noting that certain loci could mask the effects at other loci. Quantitative epistasis was first suggested and defined by Ronald Fisher (1919) who coined the term 'epistasy'. Statistically, epistasis is the deviation from an additive expectation of two or more loci, often described as a change in the slope of one locus based on the genotype at another locus (Fisher 1919). Variance due to quantitative epistasis has been shown to be an important contributor to the genetic variance in populations of model organisms such as *Arabidopsis* (Malmberg et al. 2005; Kusterer et al. 2007), as well as crop species such as maize (Stuber and Moll 1971; Melchinger, Geiger, and Schnell 1986; Lamkey, Schnicker, and Melchinger 1995; Wolf and Hallauer 1997; Lukens and Doebley 1999) and rice (Yu et al. 1997; Li et al. 2008; Shen et al. 2014). Significant epistasis has been reported in allopolyploid crops like cotton (Lee, Cockerham, and Smith 1968) and wheat (Crossa et al. 2010; Jiang et al. 2017). Epistasis across subgenomes may be indicative of interactions between homeologous loci, analogous to dominance in diploids, and a possible contributor to that adaptation of these crops to a wide landscape (Wendel 2000; Adams and Wendel 2005; Chen 2010; Chen 2013). However, there is still little direct evidence that epistasis between homeologous loci is a large contributor to the total genetic

variance in allopolyploids.

Epistasis has also been shown to be an important contributor to evolution (Doebley, Stec, and Gustus 1995; Lukens and Doebley 1999; Carlborg et al. 2006; Phillips 2008; Hansen 2013; Doust et al. 2014). There has been considerable effort over the past several decades to incorporate these non-additive genetic factors into the genotype to phenotype map. More recently these effects have been incorporated into whole genome prediction models (Vitezica, Varona, and Legarra 2013; Martini et al. 2016; Jiang and Reif 2015; Huang and Mackay 2016; Akdemir and Jannink 2015; Wolfe et al. 2016; Akdemir, Jannink, and Isidro-Sánchez 2017; Jiang et al. 2017).

In practice, detecting epistatic interactions is difficult. The pairwise search space is large even for modest numbers of markers. For example, a population genotyped with 100 markers would require 4,950 tests for pairwise epistasis. With advances in genotyping technologies, the number of DNA markers available is typically much larger, in the tens to hundreds of thousands, and more recently in the millions. In this study, 11,604 markers were available, which would result in approximately 67 million tests for pairwise epistasis. A 0.05 genome-wide Bonferroni significance threshold for all pairwise epistasis tests in this study would then be 7.4×10^{-10}.

Several methods have been proposed to reduce the multiple testing problem. Epistasis is partitioned in part to the additive variance, particularly when allele frequencies differ from 0.5 at either locus (Hill, Goddard, and Visscher 2008). Therefore, genome-wide scans can be used to first identify marker alleles (or variants) with a significant additive effect, then test only all pairwise variants identified in the scan (Carlson et al. 2004). This can greatly reduce the number of epistatic

tests performed, while increasing the likelihood that epistasis will be identified. Other methods include relaxing the multiple test correction threshold (Benjamini and Hochberg 1995), or reducing the marker pairs tested based on some criteria such as biological function or other filtering methods (Ritchie 2011; Cowman and Koyutürk 2017; Crawford et al. 2017).

The multiple test correction problem is not the only challenge to identifying epistatic interactions. Allele frequency, linkage disequilibrium and the number of alleles at a given locus can all reduce the efficacy of pairwise marker epistasis detection. Low allele frequencies at either locus reduce the epistatic effect, partitioning it to the additive instead (Hill, Goddard, and Visscher 2008). Less than perfect linkage disequilibrium between the markers and causal mutations also reduces the apparent effect size, limiting detection much as it does for additive effects (Carlson et al. 2004). SNP markers are typically considered bi-allelic, despite the potential for many different alleles in the population. The impact of these factors can be reduced by using multiple linked markers to determine haplotypes. Haplotypes have been shown to be powerful in the detection of additive and interaction effects by accurately tracking larger segments of DNA in high or perfect LD, and allowing multiple alleles at every locus (Lin and Zeng 2006; Zhang et al. 2012; Jiang, Schmidt, and Reif 2018). While allele frequencies are typically reduced using haplotypes (i.e. the frequency of two alleles will be higher than the frequency of three alleles), the added power from accurately tracking relevant LD blocks make these methods attractive.

Haplotypes do not need to be assigned directly to gain an advantage from using multiple markers to identify regions associated with complex traits. Regional heritability mapping (Nagamine et al. 2012; Riggio and Pong-Wong 2014) has been

used to identify additive effects of rare and common variants in humans (Nagamine et al. 2012; Shirali et al. 2016) as well as plants species like eucalyptus (Resende et al. 2017) and cassava (Okeke et al. 2018). These methods employ the estimation of additive covariance between individuals based on markers in a given region of chromatin, and are used in a mixed model to estimate the genetic variance attributable to the region. Variance components can then be tested to determine if they are greater than zero using a likelihood ratio test.

I propose a method to greatly reduce the number of statistical tests while taking advantage of multiple markers to determine importance of epistatic interactions across chromosome arms of an allohexaploid wheat population. This method is similar to the "divide and conquer" method of Akdemir and Jannink (2015), but models interactions across chromosomes instead of local epistasis. Epistatic covariances can be formed using the Hadamard product of component additive or dominance covariance matrices (Henderson 1985; Jiang and Reif 2015; Martini et al. 2016). Additive by additive epistatic interactions between disjoint sets of related (i.e. linked) markers can be modeled by first calculating an additive covariance for each marker set, K_i and $K_{i'}$, and using $K_i \odot K_{i'}$ as the covariance estimate of the epistatic term between these sets. I define marker sets by the chromosome arm to which they belong, and estimate the epistatic variance component between the two arms using Restricted maximum likelihood (REML) while correcting for background additive and epistatic effects.

Common wheat is an important allohexaploid crop with three subgenomes, A, B and D, resulting from hybridization events approximately 500 thousand and 10 thousand years ago. Due to the allopolyploid nature of wheat, I was particularly interested in identifying interactions across homeologous loci. Interactions at

homeologous loci are analogous to dominance effects in diploid hybrids, and could be used to fix favorable homeoallelic interactions in inbred lines (Wendel 2000; Adams and Wendel 2005; Birchler et al. 2010; Chen 2010; Chen 2013). Of the 21 chromosomes of wheat, chromosome arms pairs include $\binom{3}{2}14 = 42$ homeologous pairs, $\binom{14}{2}3 = 273$ within subgenome pairs, and $\binom{14}{2}6 = 546$ across subgenome arm pairs.

Each chromosome arm of the wheat genome was sequenced independently using flow cytometry to assist in the assembly of the large complex genome (International Wheat Genome Sequencing Consortium 2014). The lone exception was chromosome 3B, which was sequenced and assembled in its entirety before the other chromosomes of wheat (Paux et al. 2008; Choulet et al. 2014). Therefore, assigning markers to a chromosome arm is feasible, but their position along that arm may not be well defined if the number of scaffolds is large, as was the case with the first wheat survey sequence (International Wheat Genome Sequencing Consortium 2014). Using markers across an entire chromosome arm known to be homeologous to other chromosome arms may therefore be a better strategy than attempting to assign single homeologous marker pairs. If interactions are detected across homeologous regions, this may provide evidence of beneficial homeoallelic interactions indicative of inter-genomic heterosis.

I demonstrate the low resolution epistasis mapping methodology using the CNLM population, and show that epistasis can be detected between homeologous and non-homeologous chromosome arms.

5.2 Materials and Methods

5.2.1 Chromosome centromere positions

Chromosome centromere positions were provided by the IWGSC for all chromosomes except 3B (IWGSC, personal communication, March 1, 2017). Those positions were assigned by determining where chromosome arm library reads aligned to the final assembly. Each chromosome arm was sequenced independently using flow-cytometry to remove the chromosome arm from a series of aneuploid stocks, each containing an extra arm. The lone exception was chromosome 3B, which was sequenced in its entirety, so no centromere position was available for the 3B chromosome. Centromere start and stop addresses provided by IWGSC are shown in Table 5.1.

While restriction sites are expected to be uniformly distributed throughout the genome, methylation of cytosine is not. One of the restriction enzyme used to generate GBS libraries, MspI, is sensitive to DNA methylation, digesting unmethylated DNA at a much higher rate than methylated DNA (McClelland, Nelson, and Raschke 1994). Methylation is an important regulator of chromatin structure, where euchromatin tends to contain few methylation sites relative to heterochromatin (Keshet, Lieman-Hurwitz, and Cedar 1986). Therefore restriction sites in heterochromatin with high levels of methylation, such as at the centromere, are less likely to be retained as GBS markers because digestion is less likely to happen at these sites. This means that the GBS markers can be used to roughly assign a centromere position using the density of GBS markers along the chromosome.

To determine the centromere position of 3B, I used kernel density estimation

of the `density()` function of the 'stats' package in R to determine the smoothed density of GBS marker positions. I then assigned the 3B centromere interval to the chromosome positions flanking the second position for which the derivative of the density was zero. I performed this operation for all chromosomes to determine the efficacy of this method for determining the centromere position.

5.2.2 Chromosome arm resolution epistasis

The low resolution epistasis mapping approach employed here uses markers from two defined regions, i and i', to calculate additive covariance between individuals based on those regions (i.e $\mathbf{K}_i$ and $\mathbf{K}_{i'}$ $\forall$ $i \neq i'$). The Hadamard product of these additive covariance matrices can be used to produce the pairwise additive by additive epistatic relationship, $\mathbf{K}_{i\times i'} = \mathbf{K}_i \odot \mathbf{K}_{i'}$, between these two regions (Henderson 1985; Martini et al. 2016). In this study, I defined regions as the short (S) and long (L) arms of each chromosome, where $i \in \{1AS, 1AL, 1BS, \ldots, 7BL, 7DS, 7DL\}$. Variance components for each region and their respective interaction were estimated by fitting the following nested models

$$\mathbf{y} = \mathbf{1}\mu + \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{g}_{G^-} + \mathbf{Z}\mathbf{g}_{I^-} + \boldsymbol{\varepsilon} \tag{5.1}$$

$$\mathbf{y} = \mathbf{1}\mu + \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{g}_{G^-} + \mathbf{Z}\mathbf{g}_{I^-} + \mathbf{Z}\mathbf{g}_{\mathrm{A}_i} + \mathbf{Z}\mathbf{g}_{\mathrm{A}_{i'}} + \boldsymbol{\varepsilon} \tag{5.2}$$

$$\mathbf{y} = \mathbf{1}\mu + \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{g}_{G^-} + \mathbf{Z}\mathbf{g}_{I^-} + \mathbf{Z}\mathbf{g}_{\mathrm{A}_i} + \mathbf{Z}\mathbf{g}_{\mathrm{A}_{i'}} + \mathbf{Z}\mathbf{g}_{\mathrm{A}_i\times\mathrm{A}_{i'}} + \boldsymbol{\varepsilon} \tag{5.3}$$

where $\mathbf{g}_{\mathrm{A}_i} \sim \mathcal{N}(0, \sigma^2_{a_i}\mathbf{K}_i)$, $\mathbf{g}_{\mathrm{A}_{i'}} \sim \mathcal{N}(0, \sigma^2_{a_{i'}}\mathbf{K}_{i'})$ and $\mathbf{g}_{\mathrm{A}_i\times\mathrm{A}_{i'}} \sim \mathcal{N}(0, \sigma^2_{a_i\times a_{i'}}\mathbf{K}_{i\times i'})$. $\mathbf{g}_{G^-}$ and $\mathbf{g}_{I^-}$ were modeled as previously described in Chapter 3 equation 3.5, but with markers belonging to region i and i' removed prior to calculating the

covariances.

Sequential nested likelihood ratio tests were used to determine if the additive (model 5.2 versus model 5.1) and interaction (model 5.3 versus model 5.2) variance estimates of the chromosome arms were greater than zero. From the Neyman-Pearson lemma (Neyman and Pearson 1933), the likelihood ratio test statistic, is defined as $D = -2(\log\mathcal{L}_{alternative} - \log\mathcal{L}_{null})$, where $D \sim \chi^2_{df_{H_1} - df_{H_0}}$, and is the uniformly most powerful test.

BLUPs were subsequently used to look for patterns between additive and interaction effects for the chromosome arm pair. The pairwise product of the additive chromosome arm BLUPs was then compared to the chromosome arm interaction BLUPs, in a manner analogous to the Additive $\times$ Additive single locus model (see Table 4.1). Negative associations should indicate a less than additive model, whereas positive relationships would demonstrate a more than additive epistatic effect.

When a variance parameter is very close to zero, parameter estimates are unreliable, and were therefore considered to be zero. Markers were oriented by minor allele frequency as it was unclear how to apply the previous orientation schemes to multiple sets of markers. It is not clear what effect, if any, this orientation had on the estimates of interaction effects.

For the 14 three-way homeologous arm sets, a three-way interaction was included and tested against a model with only the three two-way interaction terms. I did not attempt to run all three-way chromosome arm combinations, as this would have been computationally infeasible, with $\binom{42}{3} = 11,480$ combinations. The Hadamard product of the three additive covariance matrices was used to

produce the three-way additive by additive by additive epistatic relationship, $\mathbf{K}_{i\times i'\times i''} = \mathbf{K}_i \odot \mathbf{K}_{i'} \odot \mathbf{K}_{i''}$. The following two models were fit to test the three-way interaction.

$$\begin{aligned}\mathbf{y} = \mathbf{1}\mu + \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{g}_{G^-} + \mathbf{Z}\mathbf{g}_{I^-} &+ \mathbf{Z}\mathbf{g}_{\mathrm{A}_i} + \mathbf{Z}\mathbf{g}_{\mathrm{A}_{i'}} + \mathbf{Z}\mathbf{g}_{\mathrm{A}_{i''}} \\ &+ \mathbf{Z}\mathbf{g}_{\mathrm{A}_i\times\mathrm{A}_{i'}} + \mathbf{Z}\mathbf{g}_{\mathrm{A}_i\times\mathrm{A}_{i''}} + \mathbf{Z}\mathbf{g}_{\mathrm{A}_{i'}\times\mathrm{A}_{i''}} + \boldsymbol{\varepsilon}\end{aligned} \tag{5.4}$$

$$\begin{aligned}\mathbf{y} = \mathbf{1}\mu + \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{g}_{G^-} + \mathbf{Z}\mathbf{g}_{I^-} &+ \mathbf{Z}\mathbf{g}_{\mathrm{A}_i} + \mathbf{Z}\mathbf{g}_{\mathrm{A}_{i'}} + \mathbf{Z}\mathbf{g}_{\mathrm{A}_{i''}} \\ &+ \mathbf{Z}\mathbf{g}_{\mathrm{A}_i\times\mathrm{A}_{i'}} + \mathbf{Z}\mathbf{g}_{\mathrm{A}_i\times\mathrm{A}_{i''}} + \mathbf{Z}\mathbf{g}_{\mathrm{A}_{i'}\times\mathrm{A}_{i''}} \\ &+ \mathbf{Z}\mathbf{g}_{\mathrm{A}_i\times\mathrm{A}_{i'}\times\mathrm{A}_{i''}} + \boldsymbol{\varepsilon}\end{aligned} \tag{5.5}$$

The likelihood ratio test was then used to determine if adding the three-way interaction term significantly improved the model fit beyond the two-way interaction terms.

5.3 Results

5.3.1 Centromere positions

Most of the GBS marker density estimates of centromere locations agreed well with the positions provided by the IWGSC (Figure 5.1). Chromosomes 1D and 4A were exceptions. I estimated the 3B centromere to be positioned between 347.3 Mbp and 347.9 Mbp (Table 5.1).

Table 5.1: Table of centromere positions for the 21 chromosomes of hexaploid wheat based on the RefSeq v1.0 of 'Chinese Spring' (IWGSC 2018, accepted). These positions were provided by the IWGSC (IWGSC, personal communication, March 1, 2017)

Chromosome	length[a]	start[b]	end[c]	single gap (T/F)[d]
1A	594,102,056	213,545,945	213,546,046	T
2A	780,798,557	340,034,816	340,034,917	T
3A	750,843,639	319,010,276	319,010,377	T
4A	744,588,157	265,465,435	343,405,303	F
5A	709,773,743	253,779,933	253,780,034	T
6A	618,079,260	285,321,675	285,321,776	T
7A	736,706,236	359,432,051	359,432,152	T
1B	689,851,870	236,742,047	236,742,148	T
2B	801,256,715	349,410,174	349,410,275	T
3B[e]	830,829,764	347,366,424	347,944,259	F
4B	673,617,499	319,324,823	319,324,924	T
5B	713,149,757	198,851,987	218,709,746	F
6B	720,988,478	325,245,204	325,245,305	T
7B	750,620,385	296,411,983	296,412,084	T
1D	495,453,186	172,519,511	172,519,612	T
2D	651,852,609	268,023,149	268,023,250	T
3D	615,552,423	242,690,774	242,690,875	T
4D	509,857,067	185,780,323	185,780,424	T
5D	566,080,677	188,798,562	188,798,663	T
6D	473,592,718	214,085,311	214,085,412	T
7D	638,686,055	339,371,184	339,371,285	T

[a] Total chromosome length

[b] Centromere start address

[c] Centromere end address

[d] Single gap is a boolean indicator referring to whether a clear position was determined for the centromere of each chromosome.

[e] The chromosome 3B centromere position was estimated using a kernel density estimate of GBS marker positions on 3B.

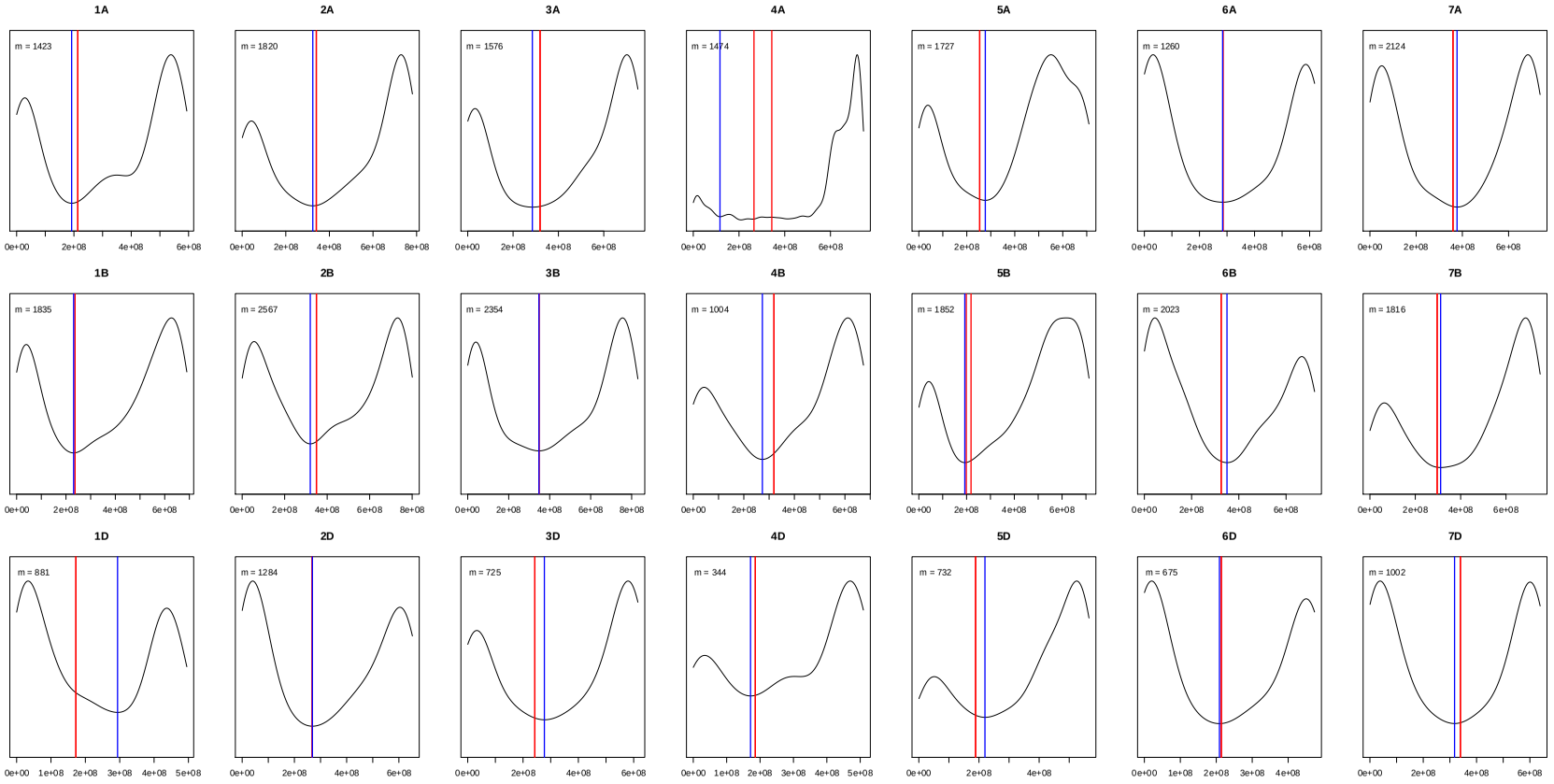


Figure 5.1: Kernel density estimation of GBS marker distribution across the 21 chromosomes of wheat. Red lines indicate the centromere interval provided by IWGSC (personal communication, March 1, 2017), and blue line indicate the centromere interval estimate based on the first derivative of the density estimate.

5.3.2 Model fit and p-value distribution

Homeologous chromosome arm pair models each had five random genetic effects and therefore five covariance structures for the two-way interaction models. All models converged, but some variance parameter estimates were often close to the parameter boundary and were considered to be zero. Variance component estimates on the boundary did not occur for the background additive or epistatic effects, but often occurred for one or both of the additive chromosome arm effects or the interaction effect. This resulted in a relatively large number of additive and interaction variance component tests with a p-value of 1. As a result, p-value distributions were heavily skewed toward 0 and 1 (Figures 5.2 and 5.3). Most chromosome arms had low additive effect p-values, whereas most interaction p-values were high, indicating that the majority of chromosome arm pairs do not have effect interactions large enough to detect.

5.3.3 Homeologous arm test

The U-shaped distribution of the p-values suggested that when the true variance was very small or zero, the average information algorithm estimated the parameter on the boundary (i.e. 0), and when it was positive, the p-value tended to be low. Larger sample sizes may be necessary to obtain uniform p-value distributions when the null hypothesis is true. I therefore considered homeologous arm interaction effects with a p-value less than 0.05 that also had positive additive variance component estimates to determine the relationship between additive chromosome arm effects and their interaction.

Seventeen homeologous chromosome arm pairs had significant interaction ef-

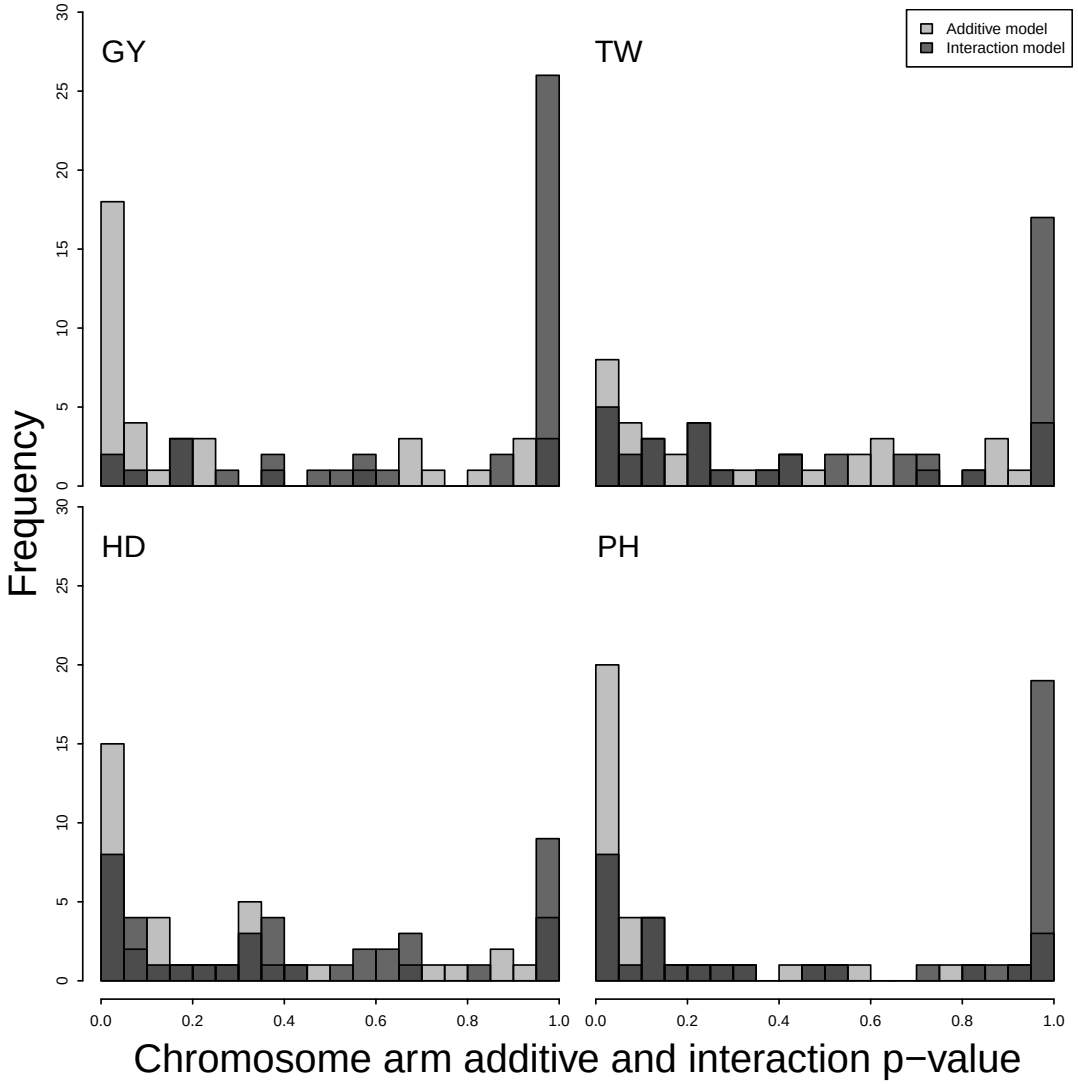


Figure 5.2: Distribution of p-values for 42 homeologous chromosome arm pair models for four traits, GY, TW, PH and HD. The p-value from the likelihood ratio test for the additive chromosome arm model is plotted in light gray, whereas the p-value from the the interaction model test is shown in dark gray.

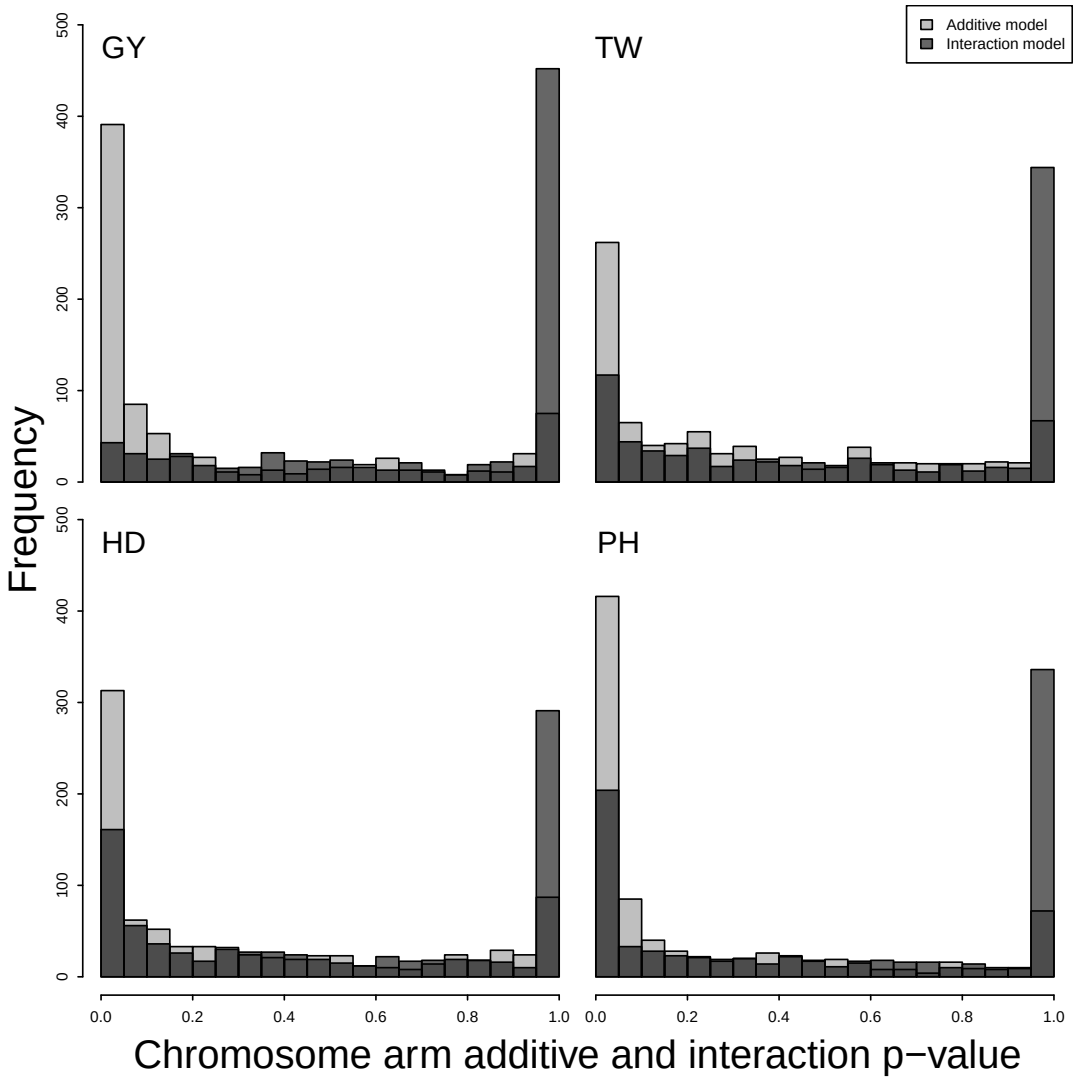


Figure 5.3: Distribution of p-values for all 861 possible chromosome arm pair models for four traits, GY, TW, PH and HD. The p-value from the likelihood ratio test for the additive chromosome arm model is plotted in light gray, whereas the p-value from the the interaction model test is shown in dark gray.

Table 5.2: Table of significant homeologous chromosome arm interactions. The proportion of genetic variance attributed to each arm and their corresponding interaction are shown with statistical significance from a nested likelihood ratio test.

Trait	arm_i	$\text{arm}_{i'}$	$(h^2_{\text{arm}_i}, h^2_{\text{arm}_{i'}})^a$	$h^2_{\text{arm}_i \times \text{arm}_{i'}}$	ρ^b
GY	5BS	5DS	$(0.038, 0)$	0.028^{**}	0.27^{***}
GY	7AL	7BL	$(0.018, 0)$	0.041^{*}	0.1^{***}
PH	2AS	2DS	$(0.021, 0.079)^{***}$	0.033^{***}	-0.04
PH	4AS	4DS	$(0, 0.039)^{***}$	0.017^{*}	0.19^{***}
PH	4AL	4BL	$(0.013, 0.034)^{*}$	0.029^{*}	0.1^{***}
PH	4AL	4DL	$(0.015, 0.0038)$	0.027^{***}	0.07^{**}
PH	4BS	4DS	$(0.0021, 0.031)^{***}$	0.049^{***}	0.18^{***}
PH	4BL	4DL	$(0.048, 0.0033)^{*}$	0.058^{***}	-0.65^{***}
PH	6AL	6DL	$(0.11, 0.0053)^{**}$	0.024^{*}	0.06^{*}
PH	7AL	7BL	$(0, 0.07)$	0.029^{**}	0.45^{***}
TW	1BS	1DS	$(0, 0)$	0.073^{***}	0
TW	4BL	4DL	$(0.096, 0.049)^{***}$	0.013^{*}	0.14^{***}
TW	6AL	6BL	$(0.031, 0)$	0.047^{*}	0.12^{***}
TW	7AL	7DL	$(0.019, 0.03)$	0.14^{***}	-0.04
TW	7BL	7DL	$(0.043, 0.061)$	0.092^{***}	0.16^{***}
HD	1BS	1DS	$(0, 0)$	0.018^{*}	0
HD	4BS	4DS	$(0, 0.0023)$	0.014^{**}	0.15^{***}
HD	6AS	6BS	$(0.0078, 0.041)^{*}$	0.049^{***}	0.02
HD	6AS	6DS	$(0.014, 0)$	0.046^{***}	-0.03
HD	6AL	6BL	$(0.0087, 0.11)^{*}$	0.013^{*}	-0.21^{***}
HD	7AS	7DS	$(0.013, 0.045)^{**}$	0.032^{*}	-0.05^{*}
HD	7AL	7BL	$(0, 0.045)^{***}$	0.025^{*}	0.14^{***}
HD	7BS	7DS	$(0.013, 0.054)^{***}$	0.012^{*}	0.29^{***}

[a] h^2 represents the proportion of the chromosome arm additive or interaction variance component estimates to the total genetic variance.
[b] ρ indicates the correlation between the product of the additive arm effects and their interaction effect with correlation coefficients significantly different from zero indicated by stars. If only one additive effect had a non-zero variance, the correlation coefficient shown is the correlation between the additive effect with the non-zero variance and the interaction effect.
*, **, and *** correspond to p-values < 0.05, 0.01, and a Bonferroni correction of $0.05/42 = 0.0012$, respectively.

fects for at least one of the four traits (Table 5.2 and Figure 5.4). Interactions involving homeologs 4 and 7 were overrepresented, with 14 of the 22 significant interactions identified between one of these two homeologs. Despite significant pairwise homeologous marker interactions found on chromosome homeologs 1 and 5 for HD and homeolog 3 for PH (section 4.4.2 of Chapter 4), chromosome arm pair tests failed to detect the significant homeologous marker set interactions on

those arms. The failure to detect these significant regions using the chromosome arm test suggests that these are either spurious associations, or their signal is being washed out by the abundance of uninformative markers on those chromosome arms. The lack of a two-way arm interactions PH interaction on chromosome arm 3S agrees with the homeologous marker set identified there, where only the three-way homeologous marker set interaction term was significant.

The test for three-way homeologous chromosome arm interactions only revealed three sets of homeologous arms that had a significant three-way interaction (Table 5.3). The three-way 3S chromosome arm interaction was found to have a positive three-way arm interaction variance parameter estimate with a p-value of $p = 0.02$, supporting the evidence from the homeologous marker set on 3S. The 7L three-way arm interaction term was also found to have a low p-value for TW of $p = 0.006$, also confirming the significant three-way homeologous marker set found there.

Many interactions were detected on chromosome arms where no homeologous marker sets were identified with a significant interaction effect. Notably, a strong interaction effect was identified on homeolog 6S for HD, and two regions for GY on 5S and 7L, where no significant homeologous interaction sets were identified. Neither of the interacting pairs for GY had a p-value lower than a homeologous arm Bonferroni correction of $0.05 / 42 = 0.0012$.

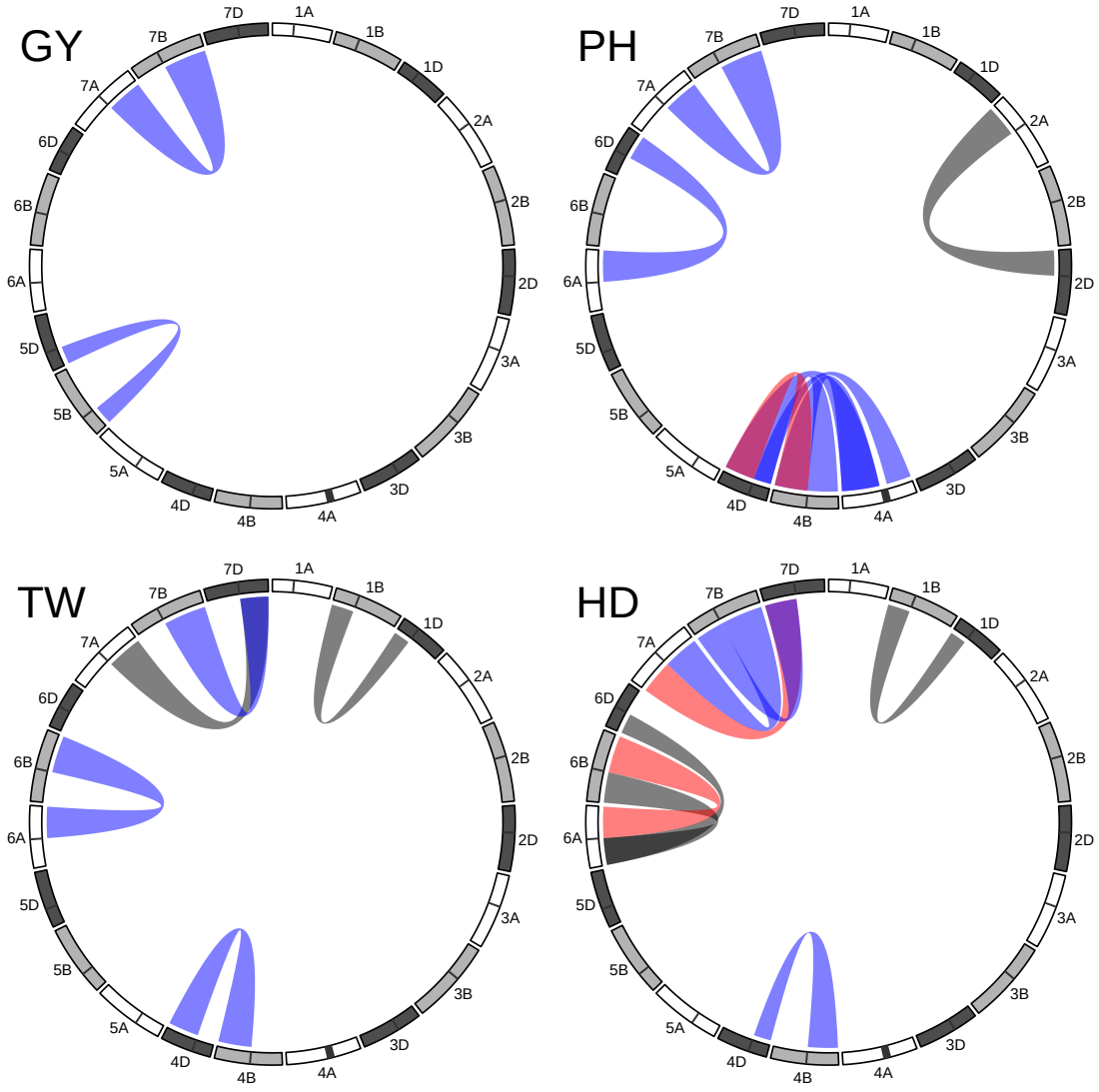


Figure 5.4: Homeologous chromosome arm interactions significant at $p < 0.05$. Blue and red bridges indicate interactions with a significant positive or negative correlation between the product of the additive effects and their interaction effect, respectively. Black bridges indicate significant interactions that did not have a significant correlation between additive products and the interaction effect.

Table 5.3: Table of significant three-way homeologous chromosome arm interactions. The proportion of genetic variance attributed to each arm and their corresponding interaction are shown with statistical significance from a nested likelihood ratio test indicated by stars.

Trait	arm_i	$\text{arm}_{i'}$	$\text{arm}_{i''}$	$(h^2_{\text{arm}_i}, h^2_{\text{arm}_{i'}}, h^2_{\text{arm}_{i''}})$	$(h^2_{\text{arm}_i \times \text{arm}_{i'}}, h^2_{\text{arm}_i \times \text{arm}_{i''}}, h^2_{\text{arm}_{i'} \times \text{arm}_{i''}})$	$h^2_{\text{arm}_i \times \text{arm}_{i'} \times \text{arm}_{i''}}$
PH	3AS	3BS	3DS	$(0.017, 0.017, 0.054)^{**a}$	$(0, 0, 0.007)^{**}$	0.010^{*}
TW	7AL	7BL	7DL	$(0.017, 0.044, 0.035)$	$(0.005, 0.057, 0)$	0.051^{**}
HD	6AS	6BS	6DS	$(0.005, 0.031, 0)^{*}$	$(0.035, 0.021, 0)^{*}$	0.019^{*}

[a] h^2 represents the proportion of the chromosome arm additive or interaction variance component estimates to the total genetic variance.
*, **, and *** correspond to p-values < 0.05, 0.01, and a Bonferroni correction of $0.05/42 = 0.0012$, respectively.

5.3.4 Homeologous additive and interaction effect relationships

Relationships between chromosome arm additive and interaction effects were only considered for the ten chromosome arm pair trait combinations that had all chromosome arm additive and interaction effects with significant non-zero variance components. Of these ten, six had significant correlations between the additive product and the interaction with an absolute value ≥ 0.1. Four of these showed positive relationships, while the other two showed negative relationships. By far the strongest relationship detected was between 4BL and 4DL for PH ($\rho = -0.65$, Figure 5.5), indicating that individuals with high or low additive values for both arms tended to have genotypic values less than expected by additivity. Conversely, the same 4BL/4DL pair had a weak, yet positive relationship for TW ($\rho = 0.14$, Figure 5.5). The 4BS/4DS pair, where the *Rht-1* genes are known to reside, had a weak, yet significant, positive correlation for PH (Figure 5.7).

5.3.5 All pairwise arm tests

For all $\binom{42}{2} = 861$ pairwise chromosome arm pairs, I only consider those tests that passed a Bonferroni threshold of $0.05/861 = 5.8 \times 10^{-5}$ in this section. Seventy nine chromosome arm interaction variance components were declared significantly greater than zero for at least one trait, representing about 2% of the number tested (Table 5.5). Of these, interactions for the PH trait were the most prevalent, representing 49 (62%), of the interactions detected. HD and TW accounted for the remaining 13 (16%) and 17 (22%) interactions. No chromosome arm interactions were detected for GY at the Bonferroni significance threshold. No interactions

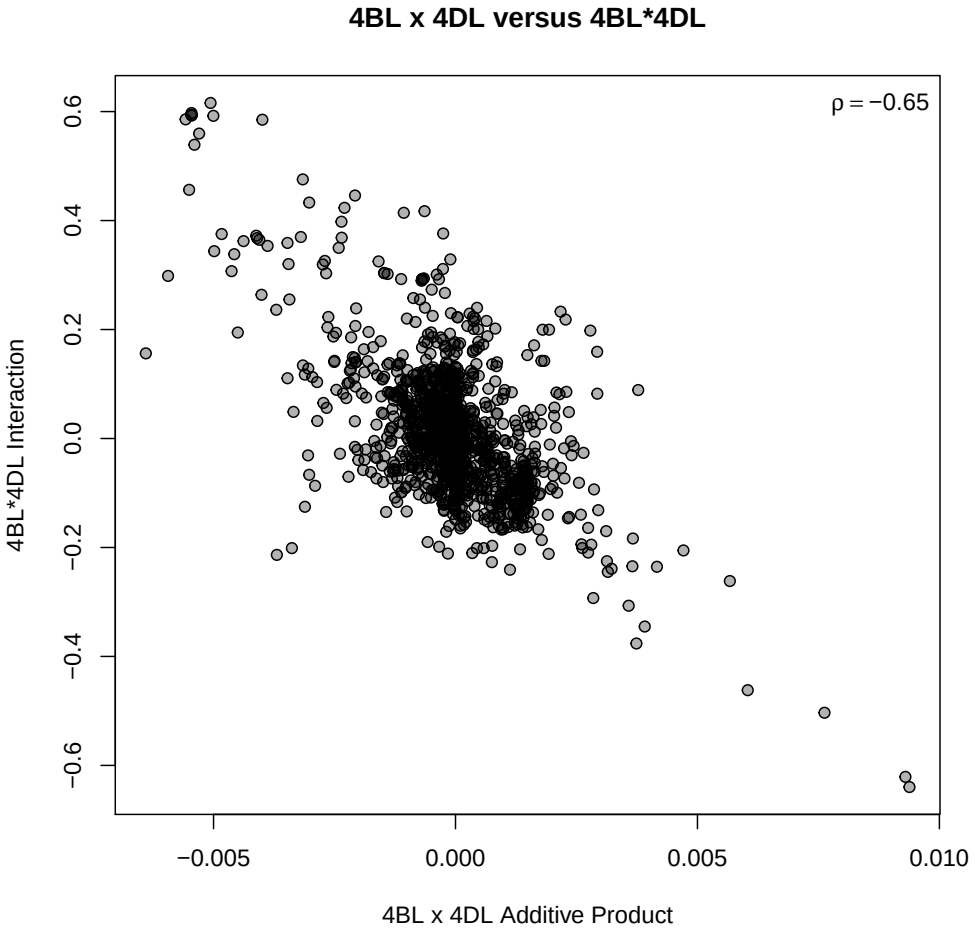


Figure 5.5: Interaction effect of chromosome 4BL by 4DL plotted against the product of the additive effects for 4BL and 4DL for PH. ρ indicates the Pearson correlation coefficient.

were detected for any of the traits involving chromosome arms 1AS, 1DL, 2AS, 2DL, 3DL, 4AS, 5AS, 5BL, 5DL, 6BL, 6DS, and 7BS at this threshold.

There were several chromosome arms that appeared to be interacting with multiple loci (Table 5.6). Of these, several clearly stand out (Figure 5.8). Chromosome arms 1AL, 2AL, 2DS, 4BS, 4DS, 4DL, 6AS and 7AL were involved in five or more interacting pairs for PH, with 2DS, 4DS and 4DL. The 4D chromosome in particu-

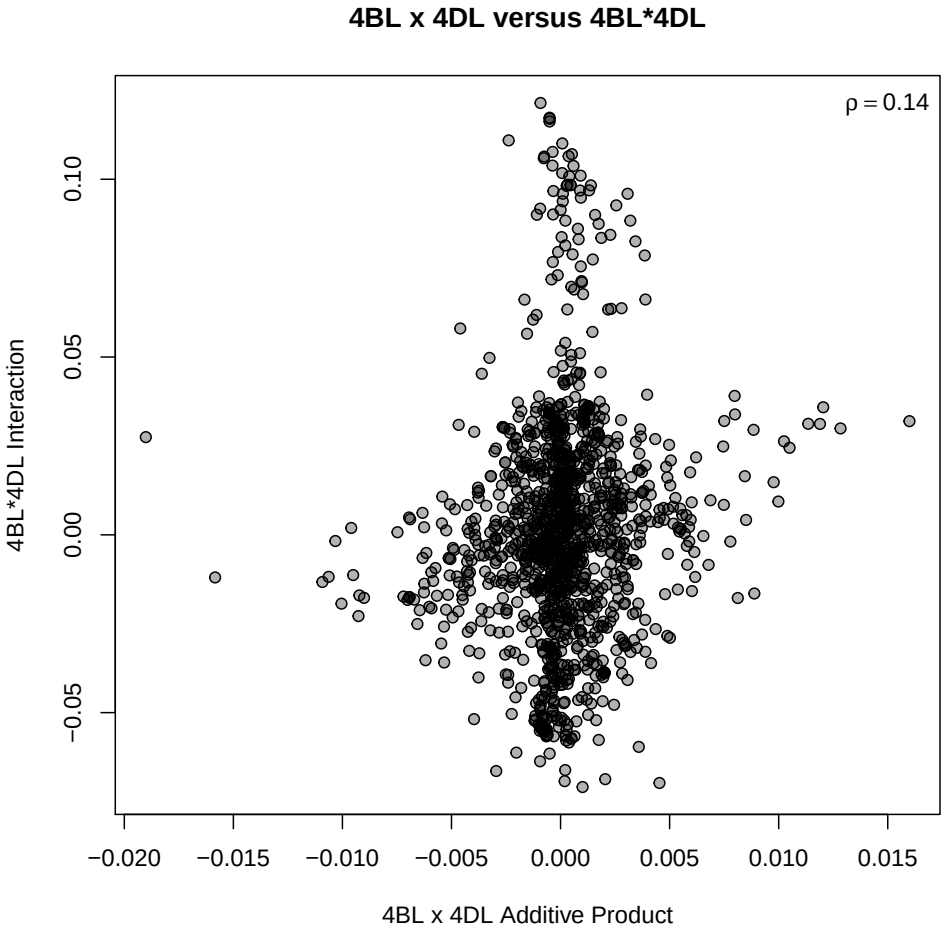


Figure 5.6: Interaction effect of chromosome 4BL by 4DL plotted against the product of the additive effects for 4BL and 4DL for TW. ρ indicates the Pearson correlation coefficient.

lar was involved in almost half (21) of the interacting arm pairs for this trait. 7DL was involved in all but three of the interacting pairs detected for the TW trait. Arm interactions for HD did not cluster to one or a few arms in the same way as PH and TW, but 6AS and 7BL were each involved in five interacting pairs for this trait.

Most correlations between the additive products and the epistatic effect were

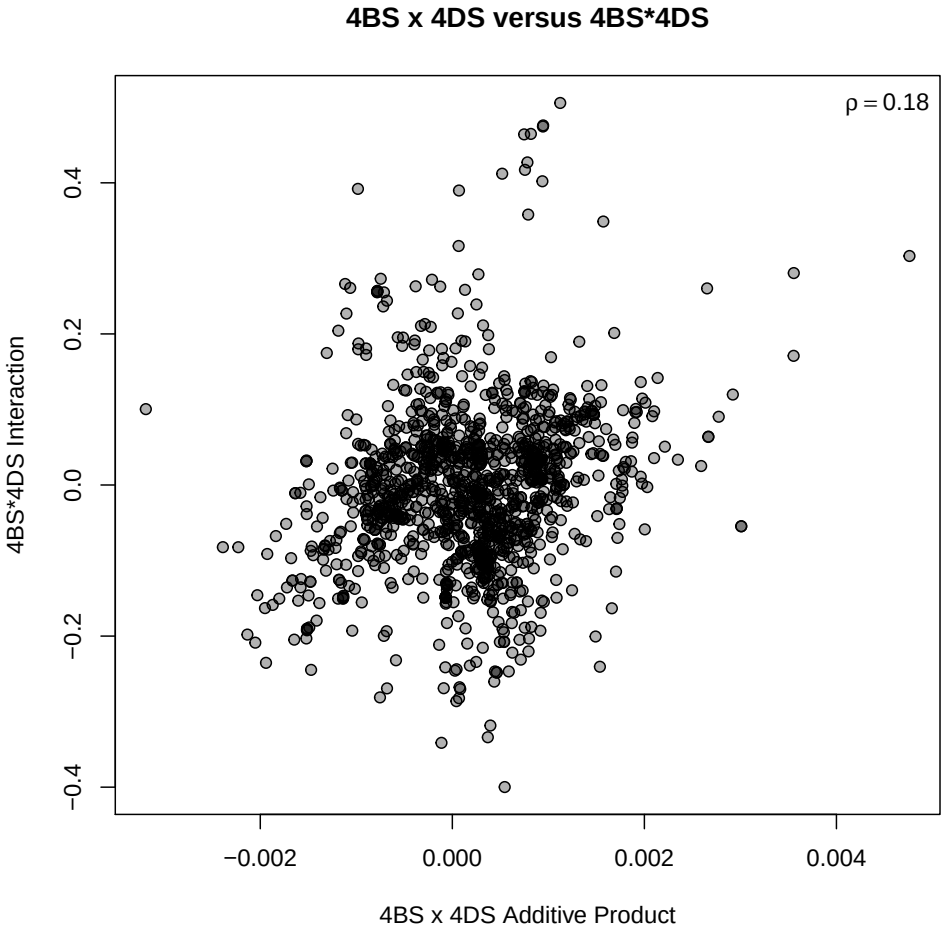


Figure 5.7: Interaction effect of chromosome 4BS by 4DS plotted against the product of the additive effects for 4BS and 4DS for PH. ρ indicates the Pearson correlation coefficient.

low in magnitude (i.e. < 0.3), particularly for the TW and HD traits. Notable exceptions include the 4BL/4DL pair for PH, which had a highly negative correlation, as previously noted. Pairs with moderate magnitude tended to also include the 4DL chromosome, but other pairs with moderate correlations between the product of their additive and interaction effects included the 1AL/2AL, 1AL/7AL, and 3AS/6AS arm pairs.

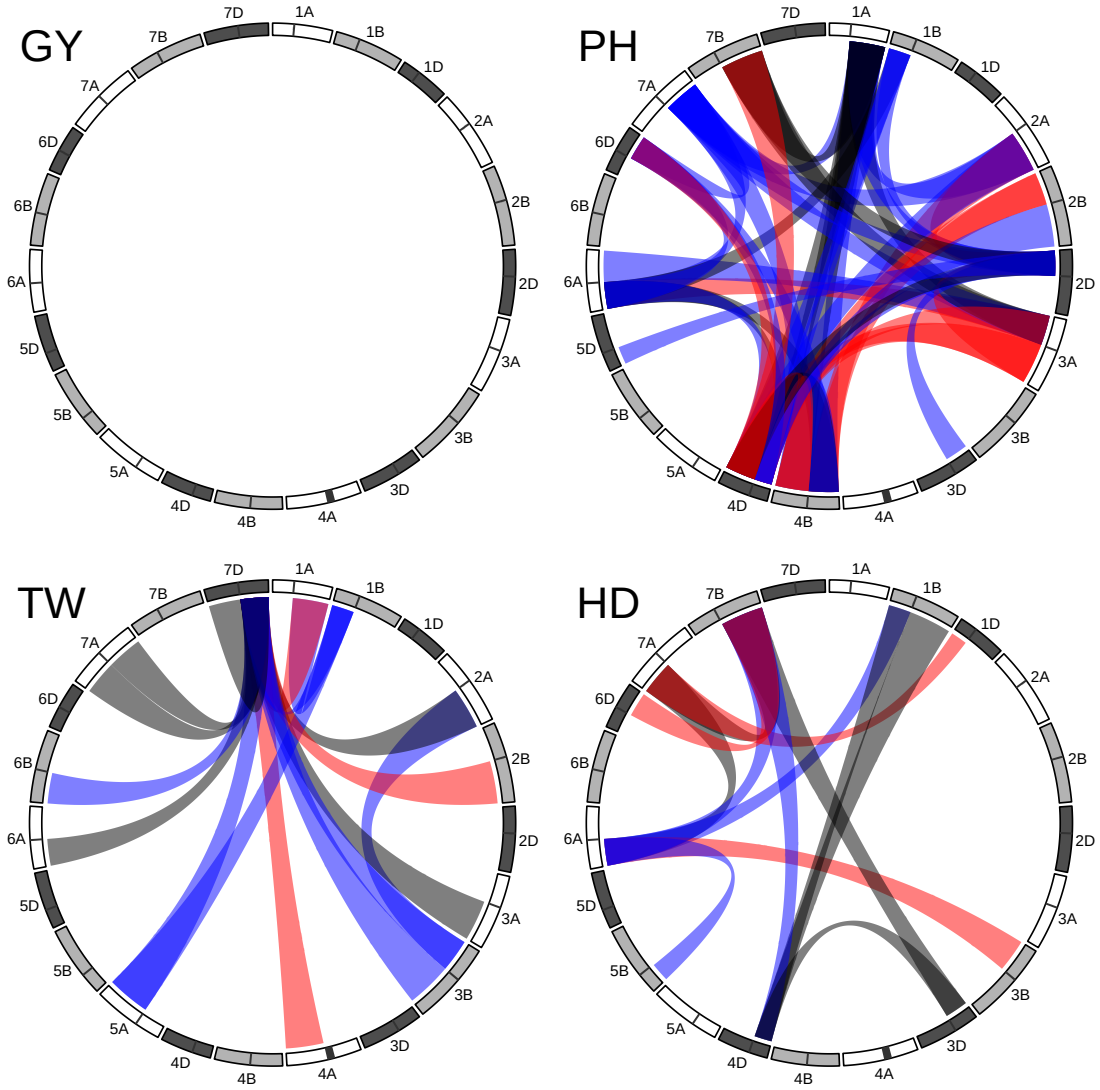


Figure 5.8: Chromosome arm interactions significant at a Bonferroni correction of $0.05/861 = 5.8 \times 10^{-5}$. Blue and red bridges indicate interactions with a significant positive or negative correlation between the product of the additive effects and their interaction effect, respectively. Black bridges indicate significant interactions that did not have a significant correlation between additive products and the interaction effect.

Table 5.4: Table of significant chromosome arm interactions for all four traits. The proportion of genetic variance attributed to each arm and their corresponding interaction are shown with statistical significance from a nested likelihood ratio test.

Trait	arm_i	$\text{arm}_{i'}$	$(h^2_{\text{arm}_i}, h^2_{\text{arm}_{i'}})^a$	$h^2_{\text{arm}_i \times \text{arm}_{i'}}$	ρ^b
PH	1AL	2AL	$(0.039, 0.06)^{****a}$	0.09^{****}	0.37^{***}
PH	1AL	3AS	$(0.065, 0.011)^{**}$	0.076^{****}	-0.01
PH	1AL	7AL	$(0.063, 0)^{*}$	0.067^{****}	0.35^{***}
PH	1AL	4BS	$(0.042, 0.024)^{**}$	0.073^{****}	0.08^{**}
PH	1AL	4BL	$(0.066, 0.012)^{***}$	0.086^{****}	-0.05
PH	1AL	7BL	$(0.091, 0.043)^{***}$	0.059^{****}	0.01
PH	1AL	2DS	$(0.079, 0.094)^{****}$	0.065^{****}	0.18^{***}
PH	1AL	4DS	$(0.048, 0.037)^{****}$	0.064^{****}	-0.04
PH	1AL	4DL	$(0.07, 0.002)^{**}$	0.048^{****}	0.05
PH	2AL	7AL	$(0.04, 0)^{**}$	0.056^{****}	0.27^{***}
PH	2AL	4BS	$(0.06, 0.032)^{***}$	0.09^{****}	-0.17^{***}
PH	2AL	2DS	$(0.036, 0.097)^{****}$	0.061^{****}	-0.17^{***}
PH	2AL	4DL	$(0.039, 0)^{***}$	0.069^{****}	0.23^{***}
PH	3AS	6AS	$(0.015, 0.069)^{**}$	0.084^{****}	-0.22^{***}
PH	3AS	6AL	$(0, 0.086)^{**}$	0.07^{****}	0.35^{***}
PH	3AS	7AL	$(0.0081, 0)$	0.06^{****}	0.27^{***}
PH	3AS	7BL	$(0.00084, 0.038)^{*}$	0.068^{****}	0.01
PH	3AS	4DS	$(0.02, 0.033)^{****}$	0.057^{****}	-0.09^{**}
PH	3AL	2DS	$(0.0042, 0.068)^{**}$	0.042^{****}	-0.11^{***}
PH	3AL	4DS	$(0.0075, 0.033)^{****}$	0.05^{****}	-0.17^{***}
PH	3AL	4DL	$(0.0015, 0.0097)$	0.023^{****}	-0.07^{**}
PH	6AS	7AL	$(0.083, 0)^{**}$	0.069^{****}	0.4^{***}
PH	6AS	1BS	$(0.078, 0.011)^{***}$	0.1^{****}	0.02
PH	6AS	4BS	$(0.054, 0)^{***}$	0.13^{****}	-0.01
PH	6AS	4DS	$(0.16, 0.042)^{****}$	0.052^{****}	0.03
PH	6AS	4DL	$(0.087, 0)^{**}$	0.073^{****}	0.23^{***}
PH	6AS	6DL	$(0.069, 0.0077)^{**}$	0.084^{****}	0.09^{**}
PH	7AL	4BL	$(0, 0.029)^{*}$	0.058^{****}	0.29^{***}
PH	7AL	2DS	$(0.00064, 0.064)^{**}$	0.054^{****}	0.11^{***}
PH	7AL	6DL	$(0, 0.016)$	0.051^{****}	0.31^{***}
PH	1BS	2DS	$(0, 0.11)^{***}$	0.062^{****}	0.22^{***}
PH	1BS	4DS	$(0, 0.035)^{****}$	0.042^{****}	0.26^{***}
PH	1BS	4DL	$(0.023, 0.0091)$	0.041^{****}	0.3^{***}
PH	2BS	4BS	$(0.027, 0.053)^{*}$	0.084^{****}	-0.11^{***}
PH	2BS	4DS	$(0.038, 0.032)^{****a}$	0.069^{****}	-0.12^{***}
PH	2BL	4DS	$(0.2, 0.029)^{****}$	0.052^{****}	0.13^{***}
PH	4BS	4DS	$(0.0021, 0.031)^{****}$	0.049^{****}	0.18^{***}
PH	4BS	4DL	$(0, 0)$	0.056^{****}	0

[a] h^2 represents the proportion of the chromosome arm additive or interaction variance component estimates to the total genetic variance.
[b] ρ indicates the correlation between the product of the additive arm effects and their interaction effect with correlation coefficients significantly different from zero indicated by stars. If only one additive effect had a non-zero variance, the correlation coefficient shown is the correlation between the additive effect with the non-zero variance and the interaction effect. *, **, and *** correspond to p-values < 0.05, 0.01, and a Bonferroni correction of $0.05/42 = 0.0012$, respectively.

Table 5.5: Continuation of Table 5.4 of significant chromosome arm interactions.

Trait	arm_i	$\text{arm}_{i'}$	$(h^2_{\text{arm}_i}, h^2_{\text{arm}_{i'}})^a$	$h^2_{\text{arm}_i \times \text{arm}_{i'}}$	ρ^b
PH	4BS	6DL	$(0, 0.021)$	0.11^{****}	0.07^{**}
PH	4BL	4DS	$(0.063, 0.029)^{****}$	0.051^{****}	-0.52^{***}
PH	4BL	4DL	$(0.048, 0.0033)^{*}$	0.058^{****}	-0.65^{***}
PH	7BL	2DS	$(0.03, 0.09)^{***}$	0.071^{****}	0
PH	7BL	4DL	$(0.031, 0.00081)^{*}$	0.094^{****}	-0.12^{***}
PH	2DS	3DS	$(0.058, 0.047)^{****}$	0.034^{****}	0.07^{**}
PH	2DS	4DS	$(0.13, 0.038)^{****}$	0.046^{****}	0.07^{**}
PH	2DS	4DL	$(0.11, 0.0057)^{***}$	0.031^{****}	0.04
PH	2DS	5DS	$(0.08, 0.0031)^{**}$	0.031^{****}	0.11^{***}
PH	4DS	6DL	$(0.031, 0.0023)^{****}$	0.036^{****}	0.08^{**}
PH	4DL	6DL	$(0.0018, 0.011)$	0.026^{****}	-0.38^{***}
TW	1AL	1BS	$(0.015, 0)$	0.13^{****}	0.23^{***}
TW	1AL	7DL	$(0.015, 0.061)$	0.097^{****}	-0.13^{***}
TW	2AL	3BS	$(0.013, 0)$	0.13^{****}	0.19^{***}
TW	2AL	7DL	$(0.032, 0.063)$	0.11^{****}	0
TW	3AL	7DL	$(0.019, 0.072)$	0.085^{****}	0.01
TW	4AL	7DL	$(0.054, 0.053)^{*}$	0.11^{****}	-0.11^{***}
TW	5AL	1BS	$(0.031, 0)$	0.15^{****}	0.26^{***}
TW	5AL	7DL	$(0.026, 0.026)^{*}$	0.14^{****}	0.06^{*}
TW	6AS	7DL	$(0, 0.058)$	0.11^{****}	0.04
TW	7AS	7DL	$(0.007, 0.059)$	0.09^{****}	0.01
TW	7AL	7DL	$(0.019, 0.03)$	0.14^{****}	-0.04
TW	1BS	7DL	$(0, 0.018)$	0.18^{****}	0.18^{***}
TW	2BL	7DL	$(0.0042, 0.063)$	0.11^{****}	-0.21^{***}
TW	3BS	7DL	$(0, 0.071)$	0.12^{****}	0.11^{***}
TW	3BL	7DL	$(0.0026, 0.054)$	0.16^{****}	0.11^{***}
TW	6BS	7DL	$(0, 0.056)$	0.093^{****}	0.08^{**}
TW	7DS	7DL	$(0, 0.057)$	0.088^{****}	-0.02
HD	6AS	7AS	$(0.0042, 0.008)$	0.098^{****}	-0.03
HD	6AS	1BS	$(0.007, 0)$	0.058^{****}	0.11^{***}
HD	6AS	3BS	$(0.012, 0.064)^{***}$	0.082^{****}	-0.12^{***}
HD	6AS	5BS	$(0.0036, 0.00018)$	0.051^{****}	0.11^{***}
HD	6AS	7BL	$(0.0076, 0.033)^{***}$	0.075^{****}	0.06^{*}
HD	7AS	7BL	$(0.0089, 0.037)^{****}$	0.073^{****}	-0.05
HD	7AS	1DS	$(0.013, 0)$	0.053^{****}	-0.09^{**}
HD	1BS	4DS	$(0, 0.0019)$	0.044^{****}	0
HD	1BL	4DS	$(0, 0.0038)$	0.035^{****}	0.02
HD	7BL	3DS	$(0.036, 0.0015)^{***}$	0.061^{****}	0.03
HD	7BL	4DS	$(0.031, 0.0028)^{***}$	0.053^{****}	0.05^{*}
HD	7BL	6DL	$(0.035, 0.0098)^{***}$	0.056^{****}	-0.15^{***}
HD	3DS	4DS	$(0, 0.003)$	0.029^{****}	0

[a] h^2 represents the proportion of the chromosome arm additive or interaction variance component estimates to the total genetic variance.
[b] ρ indicates the correlation between the product of the additive arm effects and their interaction effect with correlation coefficients significantly different from zero indicated by stars. If only one additive effect had a non-zero variance, the correlation coefficient shown is the correlation between the additive effect with the non-zero variance and the interaction effect. *, **, and *** correspond to p-values < 0.05, 0.01, and a Bonferroni correction of $0.05/42 = 0.0012$, respectively.

Table 5.6: Counts of significant homeologous chromosome arm interactions by arm and traits.

Chomosome Arm	GY	PH	TW	HD	Total
1AL	0	9	2	0	11
1BL	0	0	0	1	1
1BS	0	4	3	2	9
1DS	0	0	0	1	1
2AL	0	5	2	0	7
2BS	0	2	0	0	2
2BL	0	1	1	0	2
2DS	0	10	0	0	10
3AS	0	6	0	0	6
3AL	0	3	1	0	4
3BS	0	0	2	1	3
3BL	0	0	1	0	1
3DS	0	1	0	2	3
4AL	0	0	1	0	1
4BS	0	7	0	0	7
4BL	0	4	0	0	4
4DS	0	11	0	4	15
4DL	0	10	0	0	10
5AL	0	0	2	0	2
5BS	0	0	0	1	1
5DS	0	1	0	0	1
6AS	0	7	1	5	13
6AL	0	1	0	0	1
6BS	0	0	1	0	1
6DL	0	5	0	1	6
7AS	0	0	1	3	4
7AL	0	7	1	0	8
7BL	0	4	0	5	9
7DS	0	0	1	0	1
7DL	0	0	14	0	14
Total	0	98	34	26	158

5.4 Discussion

5.4.1 Centromere positions

While my assigned position for the 3B centromere position is an estimate, most of the chromosome estimates on other chromosomes were close to the known centromere position. The centromere position estimate reported here should be sufficient to assign most of the 3B markers to the correct chromosome arm for the subsequent analyses.

5.4.2 Model fit and p-value distribution

The distribution of p-values from the likelihood ratio test should be uniform if no true interactions exist. If interactions are important, then we would expect to see a skewed distribution with many small p-values. However, the p-values were often calculated to be one because the variance components were estimated on the parameter boundary (i.e. zero), resulting in the U-shaped distribution. When variance parameters are estimated on the parameter boundary, the p-value becomes one simply due to the fact that the variance component is zero. This is likely due to a lack of sufficient population size to distinguish and resolve multiple small variance components. Perhaps another explanation may be provided by the use of the the average information algorithm to fit the mixed model, which may lose a small portion of information by avoiding the calculation of the second derivative of the likelihood function. While other algorithms exist for solving REML problems, the computational burden of resolving multiple variance components with dense covariance structures may be restrictive. Further investigation is necessary to

determine how large a population need be to resolve multiple genetic variance parameters with magnitudes of 1% or less of the total variance.

5.4.3 Homeologous arm tests and additive interaction effect relationships

Most of the homeologous chromosome arm interactions detected across all traits involved homeologs 4 and 7. The less than additive trend observed for the 4BL/4DL pair may suggest a significant degree of gene functional redundancy between these two arms. Despite having a weak positive additive genetic trait correlation between PH and TW (Table 2.3 in Chapter 2), the 4BL/4DL pair had a weak, yet positive relationship for TW. This provides evidence that the pattern is not simply a genetic artifact and may indicate differential gene function for these two traits.

A negative correlation for PH was not observed for the 4BS/4DS chromosome arm pair, as might be expected from previous results for the *Rht-1* genes that reside on those chromosome arms. This casts some doubt on the usefulness of these correlations to infer the direction of the epistatic effect. It is unclear if the missing double-dwarf genotype is contributing to this positive correlation in the CNLM population. The relationship between the product of the additive effects and the interaction was thought to mirror the $\{-1, 1\}$ Additive $\times$ Additive epistatic model (Table 4.1) using a multi-locus approach, but it is unclear what is driving these trends.

For inbred allopolyploids, multi-subunit protein complexes can be comprised of genes from a single subgenome, or from multiple subgenomes. If functional copies of subunits exist on both genomes, the formation of subgenome hetero-complexes

may occur. Protein complexes comprised of evolutionarily divergent subunits may have increased or, more likely, decreased functionality. If heterocomplexes display decreased functionality, then we would expect the relationship between the additive and epistatic effects to be negative.

It is unlikely that all homeologous interactions are so large in effect that they are quickly fixed after the hybridization event. The distribution of epistatic effects is likely similar in shape to the distribution of additive effects. These distributions will change based on the complexity of the trait. If a trait is governed by relatively few loci, the relatively few epistatic interactions could have larger effects, and may be easier to detect. In contrast, a distribution of effects with many small non-zero effects may have many more non-zero epistatic effects, but are too small to detect.

5.4.4 All pairwise arm tests

PH appears to exhibit a higher degree of epistasis than either TW or HD. However, the number of interacting loci or chromosome arms detected was not directly related to the observed increase in genomic prediction accuracy (Chapter 2). HD had the largest percent increase in accuracy from the additive only model, yet had the fewest detectable interacting chromosome arms. GY showed no evidence of epistasis controlling the trait. This may be due to one of two explanations. The first and most obvious is that grain yield is not subject to epistatic gene action. This would mean that all genes contribute additively to the collection and allocation of resources to vegetative tissue, and then reallocation to the ear during flowering and grain fill. The second and more likely explanation is that GY is the culmination of essentially all the genes working in concert to produce the final outcome, and interactions with such small effects may simply be too small to detect

(Xu and Jia 2007; Wu, Chang, and Jing 2012).

While I corrected for population structure on both the additive and epistatic levels (i.e using additive and additive by additive genetic covariance terms), it is possible that residual structure is causing these observed relationships. The drastically different patterns in the arm pair test results for each trait suggests otherwise. If these interactions were due to population structure, we would expect to see similar patterns of significance across all traits. When I omitted the background epistatic effect, most of the 861 interactions were significant (results not shown). I deemed this to be due to chromosome arm epistatic relationship matrices modeling close relationships in the population regardless of which unit of chromatin was used to determine those relationships. However, it is possible that these interactions are far more prevalent than suggested here, and that correction for background epistatic effects is diluting true genetic signal.

The prevalence of a few chromosome arms interacting with many other arms is of particular interest, due to the potential for one site to influence the expression of so many other sites. Jiang et al. (2017) observed a large proportion of the epistatic interactions affecting GY involved chromosomes 4A and 7D in a large population of hybrid wheat. While I did not detect a large number of interactions involving 4A, 7D was particularly important for TW. However, the interactions that they detected appear to be on the short arm of chromosome 7D, instead of the long arm as I observed. It may be that the signal detected for arms influencing multiple loci is due to the presence of functional and non-functional alleles at important upstream regulators, such as transcription factors. In this case, a non-functional transcription factor would cause the suppression of differential additive alleles. However, it appears that the loci detected to interact with many other loci

in this study are not the same as those of Jiang et al. (2017).

The detection of chromosome arm interactions not identified in the homeologous marker sets suggests that single marker sets may miss important interactions. It is unclear if these interactions would have been detected if I had tested all pairwise epistatic interactions between markers. While all possible tests can be conducted, this increases the multiple testing problem drastically and may result in the loss of ability to detect even the largest effect interactions. It is unclear how large the effect sizes of a single pair of interacting loci would need to be to show up in a variance component estimated from multiple loci. While this method may not work well for a single large effect interaction, it may work well for many small effect interactions as might be expected for homeologous interactions.

It should be noted that epistatic relationships formed from the Hadamard product of covariance matrices have the property of shrinking distant relationships while emphasizing close ones. For example, two lines with an additive covariance of 0.1 will have an epistatic covariance of 0.01, whereas two lines with an additive covariance of 0.9 will have an epistatic covariance of 0.81. It may be that there are several levels of relatedness that must be considered to properly account for genetic relatedness. The pedigree is an example of a covariance estimation procedure that emphasizes close relationships and deemphasizes more distant ones. Considering both pedigree and marker based covariance matrices has been shown to be more predictive than using either alone (De Los Campos et al. 2009; Crossa et al. 2010). Other multi-kernel methods, including Reproducing Kernel Hilbert Spaces (RKHS), can be used to model these various degrees of genetic relatedness (Campos, Gianola, and Rosa 2009; Crossa et al. 2010), but may have less genetic interpretability than the method presented here.

5.5 Conclusion

The interacting pairs presented here do not have the precision to make claims of interacting genes. Nor are these interactions necessarily targets for selection. They do, however, demonstrate that there appears to be global patterns of epistasis across the genome. Seemingly additive only traits have often been shown to be under a high degree of epistasis when careful investigation is used to elucidate the trait (Carlborg et al. 2006; Forsberg et al. 2017). Some have argued that essentially all genetic variation is subject to epistasis (Huang et al. 2012; Forsberg et al. 2017), where the rest of the genome must be functional to express additive differences in alleles.

This is evident when we consider the the complexity of the cell, where no genes truly work independently of one another. In order to create the complex structure of the cell, proteins may interact with other proteins, both alike and dislike to them, to form multi-subunit complexes. Therefore allelic variation alone should be sufficient to produce epistatic variation. It is merely our inability to separate this variation from "additive" variation under classic parameterizations that leads many to conclude that epistasis is not important (Hill, Goddard, and Visscher 2008; Huang et al. 2012; Huang and Mackay 2016; Forsberg et al. 2017).

Further research into this methodology might be used to identify meaningful haplotypes. Once interacting segments are identified, they can each be split into multiple pieces for further refinement of the method, while nominally increasing the number of tests performed. The low resolution epistasis mapping approach presented here emphasizes the power of using multiple genetic markers to test for interacting genomic regions, albeit at the cost of low precision.

CHAPTER 6

CONCLUSION

Biological organisms are perhaps the most complex systems known to exist. While the hard sciences can be described quite well using mathematical models, models for biological systems are drastically simplified from reality. A single organism has billions of base pairs, comprising tens of thousands of genes, operating in hundreds of pathways that must be regulated through time and space. It is unlikely we will ever build models that take into account all the factors that affect the growth and development of a single biological organism, much less a population or ecosystem. If we do, they will likely be black box approaches, and we certainly won't be able gather more than the most general inferences from them.

When it comes to phenotypic expression, the most simplistic models assume that a phenotype is the combination of two independent factors, genetics and the environment. Some models try to account for an interaction between the environment and genetic factors, but even these models assume that the genetic factor is strictly additive. The effect of genetic variants, or alleles, on the phenotype is typically assumed to be linear, with all genes operating independently of one another. Such models are deemed to be additive, with the phenotype being the sum of all the genetic effects. However, only a portion of the total genetic variability is due to "independent" gene factors. Other parameterizations take into account gene interactions at the same or different loci, which we refer to as dominance and epistasis, respectively.

Plant and animal breeding evolved during the neolithic revolution, and has become one of the defining characteristics of the human species. No organism on earth manipulates the genetics of other organisms in the way that humans have,

either intentionally or unintentionally. Additive models are particularly useful for changing a population through generations. This additive genetic variation is exploited during population improvement, where favorable alleles are selected to increase their frequency within a population.

The seminal paper by Hill (Hill, Goddard, and Visscher 2008) suggests that most genetic variation is additive in nature, regardless of the mechanism. However, recent work has suggested that epistasis is prevalent (Forsberg et al. 2017), and is important for maintaining long term selection (Carlborg et al. 2006; Paixão and Barton 2016). Much like the central limit theorem finds the summation of realized samples from non-normal distributions to be normally distributed, the effect of many small non-additive physiological processes only appears additive because they are summed in the expression of a complex trait.

I find some evidence of subfunctionalization of homeoallelic genes, particularly for heading date (HD). However, the results presented here do not point to high levels of epistasis between homeoallelic loci. The larger degree of allele pair fixation at these loci may indicate that the most important interactions have been fixed across homeologous regions in the Cornell soft winter wheat breeding population. While this dataset represents an ideal situation to evaluate the contribution of these interactions to the genetic variation in the breeding population, it is a poor dataset to maximize the likelihood for detection of these effects. The allele frequency distribution of this population is heavily weighted with low frequency alleles, limiting the detection power of epistasis in this population.

The TILLING population developed by Kasileva et al. (2017) consisting of 2,735 mutant lines each with thousands of genic mutations could be a useful resource for future investigation into homeoallelic gene interactions. Lines with com-

plementary loss of function homeologous genes could be used to develop bi-parental mapping populations to test the degree of subfunctionalization with high statistical power afforded by allele frequencies of 0.5.

Other genetic resources also exist that would maximize the likelihood of detecting these interactions. A segregating synthetic hexaploid wheat population was developed from a cross between a spring wheat variety, Opata, and a synthetic hexaploid, W7984, with durum A and B genomes coupled to an *Ae. tauschii* D genome (Sorrells et al. 2011). The population, consisting of roughly 200 doubled haploid and over 2,000 recombinant inbred lines, will have high statistical power to detect interactions between the common wheat homeologs and their durum and *Ae. tauschii* ancestors due to both optimal allele frequencies and high genetic differences. I envision a study using the Synthetic Opata population, where the doubled haploid lines are planted in two locations across two years. Plant tissue would be sampled from several organs through time including seedling leaf, flag leaf and ear. Then a direct connection between variation across homeologous loci and phenotypic expression can be drawn using homeoallele specific gene expression as an intermediate phenotype.

Another interesting area of future research will be in transvection across homeologous chromosomes. Transvection is the trans-regulation of genes across chromosomes. For example, a promoter of one allele can recruit transcription machinery that result in the expression of a second allele on another homologous chromosome. This phenomenon was first identified by by Lewis (Lewis 1954) who proposed that it be used as a tool for detection of chromosomal rearrangements. While there is a plethora of research on transvection (Henikoff and Comai 1998, reviewed by), a search for transvection across homeologs of allopolyploids turned up no results.

Prediction of unobserved beneficial homeoallelic epistatic regions may prove difficult, as it currently is in diploid hybrids. Additionally, directed selection for homeoallelic interactions across subgenomes will be challenging in autogamous allopolyploids due to the intensive labor involved in making crosses. However, these tools provide a way to screen large populations for beneficial homeologous interactions such that they may in turn be selected in breeding populations before intensive field trials are conducted, saving phenotyping costs and potentially reducing the time to variety release.

Treating the genome as consisting of purely additive gene action assumes that genes are independent machines, whose products sum to the final value of an individual. While convenient for selection, this is almost certainly not true when we consider the molecular mechanisms of biological organisms. Instead, genes work in concert to produce an observable phenotype. Allopolyploids have traditionally been treated as diploids in breeding programs because they undergo disomic inheritance. With modern DNA marker technology and ever increasing computational power, breeders of allopolyploids can further exploit the genetic complexity of their crops. We now have the technical ability to view and start to breed these organisms as the ancient immortal hybrids that they are.

CHAPTER 7

APPENDIX: PROPOSAL - APRIL 5TH 2015

The following proposal was written for the PLBRG 7160 course, Perspectives in Plant Breeding Strategies, instructed by Dr. Mark Sorrells in the spring of 2015. This proposal was inspired by the work of James Mac Key in the paper 'Significance of mating systems for chromosomes and gametes in polyploids' (1970), and subsequently inspired much of the work presented in this dissertation. It is presented here exactly as it was submitted as an assignment in April of 2015.

PLBR 7160 Genome specific kernels for genomic selection in allopolyploid species

Nicholas Santantonio, April 5^{th}, 2015

Introduction

As an autogamous disomic allohexaploid, some aspects of genetic improvement of common wheat, *Triticum aestivum*, can be a challenge. The last polyploidization event occured approximately 9,000 years ago through the hybridization of *Aegilops tauschii* and *T. dicoccoides* (Dubcovsky and Dvorak[2], 2007). Because this event is thought to have occurred just a few times, hexaploid wheat has relatively less polymorphism than other crop species. This is particularly true for the D genome, which comes from the *A. tauschii*, and has consistently been shown to have less polymorphism than either the A or B genomes (e.g. Wang et al.[5], 2014). On the other hand, the B genome has the most genetic diversity because the B genome donor, *A. speltoides* is a cross pollinated species, and therefore contributed more genetic variation during the hybridization with the A genome progenitor, *T. urartu*. The A genome is more polymorphic than the D genome, but less so than the B genome due to the autogamous nature of *T. urartu*.

An unfortunate consequence of phenotypic selection is that loci having the largest effects on phenotypic variance will be preferentially selected first. Loci with relatively small effects on phenotypic variance are less likely to be selected during the early cycles, and therefore less favorable alleles at these loci may become fixed through genetic drift. Because of the low genetic diversity in the D genome, it is likely that the D genome also contributes relatively small proportions of additive genetic variance compared to the A and B genomes. Therefore phenotypic selection is likely to act the most on loci in the B genome, and the least on loci in the D genome. Reduced genetic diversity is also associated with reduced polymorphisms that can be used as genetic markers. Because of this, each genome is not equally represented when using typical genomic selection techniques, such as GBLUP.

In the typical GBLUP senario, the number of markers in each genome weights the influence of that genome when calculating the additive genetic relationship between individuals. The D genome, which typically has very small numbers of markers compared to the B genome, is highly underrepresented when determining the relationship of lines. Because of this, genomic selection in wheat is highly biased toward variation in the A and B genomes, making it difficult to use for selection of useful variation within the D genome.

In either case, variation in the D genome is likely to be fixed by genetic drift through inbreeding and selection before the breeder can take advantage of it. Therefore, neither phenotypic selection nor typical GBLUP selection can take advantage of the relatively small portion of genetic variance in the D genome. However, a unique feature of genomic selection is that it can allow the breeder to make selections based on specific genomic regions, given that markers locations are known. This study proposes a solution to this problem in order to allow the breeder to take advantage of genome specific genetic variation in a manner that makes the most sense based on the trait of interest.

Marker Platforms

Because this method is concerned with the distribution of molecular markers across the three genomes of wheat, a brief discussion of marker platforms will be presented first. In the case of wheat, several SNP arrays are currently in use. While these arrays typically contain markers with "known" genetic positions based on biparental mapping populations, they are also subject to ascertainment bias because the arrays are designed to detect polymorphisms in the population from which they were designed. Genotyping by sequencing (GBS) markers, on the other hand, are not subject to ascertainment bias because SNP markers are called based on the population in which the genotyping is conducted. While GBS markers generally do not have known positions, they can be assigned to chromosome arms by BLASTing marker sequence tags to the Chinese Spring wheat 5× survey sequence. Because the sequencing was accomplished individually for each chromosome arm, unique alignments that have no gaps, and one or two mismatches for the SNP polymorphisms are probably reliable alignments and can be used to assign markers to a chromosome and genome.

GBS seems to be a more attractive marker platform for this exercise, however, GBS markers typically underrepresent the D genome even more drastically than SNP arrays, because no effort is made to exploit the somewhat rare polymorphisms of the D genome. In this light, it is unclear which marker platform would be preferable for the following proposed method. Should coverage be given higher priority than ascertainment bias? Answering this question would need a population(s) that has been genotyped with both platforms and compared for their relative efficacy under the following genome specific selection method. Therefore, I will assume that GBS markers are being used due to their recent popularity and relative flexibility.

Genome Specific Selection

A population must first be genotyped using GBS, and markers assigned a chromosome arm using a local BLAST as previously described. Markers without chromosome information can be assigned to a chromosome if they correlate to markers having chromosome information with a correlation coefficient greater than 0.7 or so, such that $0.7 \leq r < 1$. Any markers perfectly correlated to markers with chromosome information should be dropped, as they provide no additional information, and cause the marker matrix to lack full rank.

To estimate the relative importance of each genome to the trait of interest, the marker matrix can be separated into three matrices based on the genome origin of each marker. An additive relationship matrix can then be calculated for each of the genome specific marker matrices (denoted $\mathbf{A}, \mathbf{B}, \mathbf{D}$ for the A, B and D genomes respectively) using the method proposed by Van Raden [4] (2008):

$$\text{additive genetic relationship matrix, } \mathbf{K} = \frac{\mathbf{W}\mathbf{W}^T}{2\sum_i p_i(1-p_i)}$$

where $\mathbf{W}$ is the centered marker matrix, and p_j is the major allele frequency of the j^{th} marker.

A linear mixed model can then be fit with each genome specific relationship matrix corresponding to the covariance matrix for three separate kernels using the following model:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{a} + \mathbf{W}\mathbf{b} + \mathbf{V}\mathbf{d} + \mathbf{e}$$

with the assumptions,

$$\mathbf{a} \sim N(0, \sigma_a^2\mathbf{A}) \quad , \quad \mathbf{b} \sim N(0, \sigma_b^2\mathbf{B}) \quad \text{and} \quad \mathbf{d} \sim N(0, \sigma_d^2\mathbf{D})$$

where, $\boldsymbol{\beta}$ is a vector of fixed effects, $\mathbf{a}$, $\mathbf{b}$, and $\mathbf{d}$ are vectors of genome specific breeding values for the A, B and D genomes, respectively, and $\mathbf{X}$,$\mathbf{Z}$,$\mathbf{W}$, and $\mathbf{V}$ are incidence matrices. The relative amounts of additive genetic variation explained by each genome can subsequently be determined by estimating σ_a^2, σ_b^2, and σ_d^2 using software available for fitting multiple kernel linear mixed models that estimate variance components using Restricted Maximum Likelihood (REML) methods, such as the 'emmremlMulitKernel' function of the 'EMMREML' package (Akdemir and Okeke[1], 2014) in R (R Development Team[3], 2014).

Once the relative amounts of genetic variation explained by each genome is determined, weights can be given to each kernel based on the relative importance of the genome as determined by the researcher. Weighting each genome is accomplished by multiplying the relationship matrix by a constant between 0 and 1, such that the weight constants sum to 1, resulting in,

$$\mathbf{a} \sim N(0, p\sigma_a^2\mathbf{A}) \quad , \quad \mathbf{b} \sim N(0, q\sigma_b^2\mathbf{B}) \quad \text{and} \quad \mathbf{d} \sim N(0, (1-(p+q))\sigma_d^2\mathbf{D})$$
$$\text{for} \quad 0 \leq p \leq 1 \quad , \quad 0 \leq q \leq 1 \quad \text{and} \quad 0 \leq p+q \leq 1$$

One method would be to weight each kernel based on the relative importance of the genome to the genetic variability of the given trait; however, this would likely give results similar to the typical GBLUP if the proportion of markers in each genome is similar to the proportion of additive genetic variance contributed by each genome. A breeder might decide to put most of the weight on the D genome, such that the D genome has the highest priority and the B genome has the least priority during early stages of selection, and shift the weight back toward the A and B genomes in later cycles of selection. This method could also allow for several different selection schemes based on genomic regions of interest, perhaps using markers on genomic regions inherited from a wild parent in a separate kernel from the elite recurrent parent genomic regions.

Using this method, a plant breeder can take advantage of genetic variability in underrepresented or otherwise interesting genomic regions without being masked by large effect loci in over represented or otherwise less interesting regions.

References

[1] Akdemir, D. and Okeke, U. G. (2014). EMMREML: Fitting mixed models with known covariance structures.. R package version 1.0. URL http://CRAN.R-project.org/package=EMMREML

[2] Dubcovsky, J., and Dvorak, J. (2007). Genome plasticity a key factor in the success of polyploid wheat under domestication. Science, 316(5833), 1862-1866.

[3] R Core Team (2014). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.

[4] VanRaden, P. M. (2008). Efficient methods to compute genomic predictions. Journal of dairy science, 91(11), 4414-4423.

[5] Wang, S., Wong, D., Forrest, K., Allen, A., Chao, S., Huang, B. E., Maccaferri, M., Salvi, S., Milner, S., Cattivelli, L., Mastrangelo, A. M., Whan, A., Stephen, S., Barker, G., Wieseke, R., Plieske, J., International Wheat Genome Sequencing Consortium, Lillemo, M., Mather, D., Appels, R., Dolferus, R., Brown-Guedira, G., Korol, A., Akhunova, A. R., Feuillet, C., Salse, J., Morgante, M., Pozniak, C., Luo, M., Dvorak, J., Morell, M., Dubcovsky, J., Ganal, M., Tuberosa, R., Lawley, C., Mikoulitch, I., Cavanagh, C., Edwards, K. J., Hayden, M., and Akhunov, E. (2014). Characterization of polyploid wheat genomic diversity using a high-density 90 000 single nucleotide polymorphism array. Plant biotechnology journal, 12(6), 787-796.

BIBLIOGRAPHY

Abel, S, C Möllers, and HC Becker (2005). "Development of synthetic Brassica napus lines for the analysis of fixed heterosis in allopolyploid plants". In: *Euphytica* 146.1-2, pp. 157–163.

Adams, Keith L and Jonathan F Wendel (2005). "Polyploidy and genome evolution in plants". In: *Current opinion in plant biology* 8.2, pp. 135–141.

Adams, Keith L et al. (2003). "Genes duplicated by polyploidy show unequal contributions to the transcriptome and organ-specific reciprocal silencing". In: *Proceedings of the National Academy of sciences* 100.8, pp. 4649–4654.

Akdemir, Deniz and Jean-Luc Jannink (2015). "Locally epistatic genomic relationship matrices for genomic association and prediction". In: *Genetics* 199.3, pp. 857–871.

Akdemir, Deniz, Jean-Luc Jannink, and Julio Isidro-Sánchez (2017). "Locally epistatic models for genome-wide prediction and association by importance sampling". In: *Genetics Selection Evolution* 49.1, p. 74.

Akdemir, Deniz and Julio I Sánchez (2016). "Efficient breeding by genomic mating". In: *Frontiers in genetics* 7, p. 210.

Akhunova, Alina R et al. (2010). "Homoeolog-specific transcriptional bias in allopolyploid wheat". In: *BMC genomics* 11.1, p. 505.

Allard, Robert W and AD Bradshaw (1964). "Implications of genotype-environmental interactions in applied plant breeding". In: *Crop science* 4.5, pp. 503–508.

Álvarez-Castro, José M and Örjan Carlborg (2007). "A unified model for functional and statistical epistasis and its application in quantitative trait loci analysis". In: *Genetics* 176.2, pp. 1151–1167.

Assis, Raquel and Doris Bachtrog (2013). "Neofunctionalization of young duplicate genes in Drosophila". In: *Proceedings of the National Academy of Sciences* 110.43, pp. 17409–17414.

Bates, D et al. (2015a). *Fitting Linear Mixed-Effects Models using lme4, 1–51. Computation.*

Bates, Douglas et al. (2015b). "Parsimonious mixed models". In: *arXiv preprint arXiv:1506.04967.*

Bateson, William (2007). *Mendel's principles of heredity.* (Cambridge Univ. Press.

Beest, Mariska te et al. (2011). "The more the better? The role of polyploidy in facilitating plant invasions". In: *Annals of botany* 109.1, pp. 19–45.

Benjamini, Yoav and Yosef Hochberg (1995). "Controlling the false discovery rate: a practical and powerful approach to multiple testing". In: *Journal of the royal statistical society. Series B (Methodological)*, pp. 289–300.

Bernardo, Rex and Addie M Thompson (2016). "Germplasm Architecture Revealed through Chromosomal Effects for Quantitative Traits in Maize". In: *The Plant Genome* 9.2.

Bingham, ET et al. (1994). "Complementary gene interactions in alfalfa are greater in autotetraploids than diploids". In: *Crop Science* 34.4, pp. 823–829.

Birchler, James A et al. (2010). "Heterosis". In: *The Plant Cell* 22.7, pp. 2105–2112.

Blake, Nancy K et al. (1999). "Phylogenetic reconstruction based on low copy DNA sequence data in an allopolyploid: the B genome of wheat". In: *Genome* 42.2, pp. 351–360.

Blanc, Guillaume and Kenneth H Wolfe (2004). "Functional divergence of duplicated genes formed by polyploidy during Arabidopsis evolution". In: *The Plant Cell* 16.7, pp. 1679–1691.

Börner, A et al. (1996). "The relationships between the dwarfing genes of wheat and rye". In: *Euphytica* 89.1, pp. 69–75.

Bornkamp, Bjoern (2012). *txtplot: Text based plots*. R package version 1.0-3. URL: `https://CRAN.R-project.org/package=txtplot`.

Butler, David (2009). *asreml: asreml() fits the linear mixed model*. R package version 3.0. URL: `www.vsni.co.uk`.

Camacho, Christiam et al. (2009). "BLAST+: architecture and applications". In: *BMC bioinformatics* 10.1, p. 421.

Campos, Gustavo de los, Daniel Gianola, and Guilherme JM Rosa (2009). "Reproducing kernel Hilbert spaces regression: a general framework for genetic evaluation". In: *Journal of Animal Science* 87.6, pp. 1883–1887.

Campos, Gustavo de los and Paulino Pérez Rodriguez (2015). *BGLR: Bayesian Generalized Linear Regression*. R package version 1.0.4. URL: `http://CRAN.R-project.org/package=BGLR`.

Carlborg, Örjan et al. (2006). "Epistasis and the release of genetic variation during long-term selection". In: *Nature genetics* 38.4, p. 418.

Carlson, Christopher S et al. (2004). "Mapping complex disease loci in whole-genome association studies". In: *Nature* 429.6990, p. 446.

Chaudhary, Bhupendra et al. (2009). "Reciprocal silencing, transcriptional bias and functional divergence of homeologs in polyploid cotton (Gossypium)". In: *Genetics* 182.2, pp. 503–517.

Chen, Z Jeffrey (2010). "Molecular mechanisms of polyploidy and hybrid vigor". In: *Trends in plant science* 15.2, pp. 57–71.

— (2013). "Genomic and epigenetic insights into the molecular bases of heterosis". In: *Nature Reviews Genetics* 14.7, p. 471.

Chen, Z Jeffrey and Zhongfu Ni (2006). "Mechanisms of genomic rearrangements and gene expression changes in plant polyploids". In: *Bioessays* 28.3, pp. 240–252.

Cheverud, James M and Eric J Routman (1995). "Epistasis and its contribution to genetic variance components." In: *Genetics* 139.3, pp. 1455–1461.

Choulet, Frédéric et al. (2014). "Structural and functional partitioning of bread wheat chromosome 3B". In: *Science* 345.6194, p. 1249721.

Cockerham, C Clark and Zhao-Bang Zeng (1996). "Design III with marker loci". In: *Genetics* 143.3, pp. 1437–1456.

Comai, Luca (2000). "Genetic and epigenetic interactions in allopolyploid plants". In: *Plant molecular biology* 43.2-3, pp. 387–399.

Comai, Luca et al. (2003). "Do the different parental heteromes cause genomic shock in newly formed allopolyploids?" In: *Philosophical Transactions of the Royal Society of London B: Biological Sciences* 358.1434, pp. 1149–1155.

Cowman, Tyler and Mehmet Koyutürk (2017). "Prioritizing tests of epistasis through hierarchical representation of genomic redundancies". In: *Nucleic acids research* 45.14, e131–e131.

Crawford, Lorin et al. (2017). "Detecting epistasis with the marginal epistasis test in genetic mapping studies of quantitative traits". In: *PLoS genetics* 13.7, e1006869.

Crossa, José et al. (2010). "Prediction of genetic values of quantitative traits in plant breeding using pedigree and molecular markers". In: *Genetics* 186.2, pp. 713–724.

Dahl, David B. (2016). *xtable: Export Tables to LaTeX or HTML*. R package version 1.8-2. URL: `https://CRAN.R-project.org/package=xtable`.

De Los Campos, Gustavo et al. (2009). "Predicting quantitative traits with regression models for dense molecular markers and pedigree". In: *Genetics* 182.1, pp. 375–385.

Devos, KM et al. (1995). "Structural evolution of wheat chromosomes 4A, 5A, and 7B and its impact on recombination". In: *Theoretical and Applied Genetics* 91.2, pp. 282–288.

Doebley, John, Adrian Stec, and Charles Gustus (1995). "teosinte branched1 and the origin of maize: evidence for epistasis and the evolution of dominance." In: *Genetics* 141.1, pp. 333–346.

Doust, Andrew N et al. (2014). "Beyond the single gene: How epistasis and gene-by-environment effects influence crop domestication". In: *Proceedings of the National Academy of Sciences* 111.17, pp. 6178–6183.

Duarte, Jill M et al. (2005). "Expression pattern shifts following duplication indicative of subfunctionalization and neofunctionalization in regulatory genes of Arabidopsis". In: *Molecular biology and evolution* 23.2, pp. 469–478.

Dubcovsky, Jorge and Jan Dvořák (2007). "Genome plasticity a key factor in the success of polyploid wheat under domestication". In: *Science* 316.5833, pp. 1862–1866.

Duvick, Donald N (1999). "Heterosis: feeding people and protecting natural resources". In: *The genetics and exploitation of heterosis in crops*, pp. 19–29.

Dvořák, Jan et al. (1993). "The evolution of polyploid wheats: identification of the A genome donor species". In: *Genome* 36.1, pp. 21–31.

Eckart, Carl and Gale Young (1936). "The approximation of one matrix by another of lower rank". In: *Psychometrika* 1.3, pp. 211–218.

Ellstrand, Norman C and Kristina A Schierenbeck (2000). "Hybridization as a stimulus for the evolution of invasiveness in plants?" In: *Proceedings of the National Academy of Sciences* 97.13, pp. 7043–7050.

Elshire, Robert J et al. (2011). "A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species". In: *PloS one* 6.5, e19379.

Eshed, Yuval and Dani Zamir (1996). "Less-than-additive epistatic interactions of quantitative trait loci in tomato". In: *Genetics* 143.4, pp. 1807–1817.

Feldman, Moshe et al. (2012). "Genomic asymmetry in allopolyploid plants: wheat as a model". In: *Journal of experimental botany* 63.14, pp. 5045–5059.

Fisher, Ronald A (1919). "XV.The correlation between relatives on the supposition of Mendelian inheritance." In: *Earth and Environmental Science Transactions of the Royal Society of Edinburgh* 52.2, pp. 399–433.

Force, Allan et al. (1999). "Preservation of duplicate genes by complementary, degenerative mutations". In: *Genetics* 151.4, pp. 1531–1545.

Forsberg, Simon KG et al. (2017). "Accounting for genetic interactions improves modeling of individual quantitative trait phenotypes in yeast". In: *Nature genetics* 49.4, p. 497.

Garrick, DJ (2007). "Derivation of equivalent computing algorithms for genomic predictions and reliabilities of animal merit". In: *Journal of dairy science* 90.Suppl. 1, p. 376.

Gault, Christine, Karl Kremling, and Edward S Buckler (2018). "Tripsacum de novo transcriptome assemblies reveal parallel gene evolution with maize after ancient polyploidy". In: *bioRxiv*, p. 267682.

Gilmour, AR (1997). "ASREML for testing fixed effects and estimating multiple trait variance components". In: *Proceedings of the Association for the Advancement of Animal Breeding and Genetics*. Vol. 12, pp. 386–390.

Glaubitz, Jeffrey C et al. (2014). "TASSEL-GBS: a high capacity genotyping by sequencing analysis pipeline". In: *PLoS One* 9.2, e90346.

Goddard, ME and BJ Hayes (2007). "Genomic selection". In: *Journal of Animal breeding and Genetics* 124.6, pp. 323–330.

Gu, Zuguang et al. (2014). "circlize implements and enhances circular visualization in R". In: *Bioinformatics* 30.19, pp. 2811–2812.

Haldane, JBS (1933). "The part played by recurrent mutation in evolution". In: *American Naturalist*, pp. 5–19.

Hansen, Thomas F (2013). "Why epistasis is important for selection and adaptation". In: *Evolution* 67.12, pp. 3501–3511.

He, Dan, Zhanyong Wang, and Laxmi Parida (2015). "Data-driven encoding for quantitative genetic trait prediction". In: *BMC bioinformatics*. Vol. 16. Suppl 1. BioMed Central Ltd, S10.

Heffner, Elliot L, Mark E Sorrells, and Jean-Luc Jannink (2009). "Genomic selection for crop improvement". In: *Crop Science* 49.1, pp. 1–12.

Henderson, CR (1985). "Best linear unbiased prediction of nonadditive genetic merits in noninbred populations". In: *Journal of animal science* 60.1, pp. 111–117.

Henikoff, Steven and Luca Comai (1998). "Trans-sensing effects: the ups and downs of being together". In: *Cell* 93.3, pp. 329–332.

Herskowitz, Ira (1987). "Functional inactivation of genes by dominant negative mutations". In: *Nature* 329.6136, pp. 219–222.

Heslot, Nicolas, Jean-Luc Jannink, and Mark E Sorrells (2015). "Perspectives for genomic selection applications and research in plants". In: *Crop Science* 55.1, pp. 1–12.

Hill, William G, Michael E Goddard, and Peter M Visscher (2008). "Data and theory point to mainly additive genetic variance for complex traits". In: *PLoS Genet* 4.2, e1000008.

Huang, Wen and Trudy FC Mackay (2016). "The genetic architecture of quantitative traits cannot be inferred from variance component analysis". In: *PLoS genetics* 12.11, e1006421.

Huang, Wen et al. (2012). "Epistasis dominates the genetic architecture of Drosophila quantitative traits". In: *Proceedings of the National Academy of Sciences* 109.39, pp. 15553–15559.

International Wheat Genome Sequencing Consortium (2014). "A chromosome-based draft sequence of the hexaploid bread wheat (Triticum aestivum) genome". In: *Science* 345.6194, p. 1251788.

IWGSC, International Wheat Genome Sequencing Consortium (2018, accepted). "Shifting the limits in wheat research and breeding using a fully annotated reference genome by the International Wheat Genome Sequencing Consortium (IWGSC)". In: *Science*.

Jannink, Jean-Luc, Aaron J Lorenz, and Hiroyoshi Iwata (2010). "Genomic selection in plant breeding: from theory to practice". In: *Briefings in functional genomics* 9.2, pp. 166–177.

Jannink, Jean-Luc et al. (2009). "Overview of QTL detection in plants and tests for synergistic epistatic interactions". In: *Genetica* 136.2, p. 225.

Jiang, Yong and Jochen C Reif (2015). "Modeling epistasis in genomic selection". In: *Genetics* 201.2, pp. 759–768.

Jiang, Yong, Renate H Schmidt, and Jochen C Reif (2018). "Haplotype-Based Genome-Wide Prediction Models Exploit Local Epistatic Interactions Among Markers". In: *G3: Genes, Genomes, Genetics* 8.5, pp. 1687–1699.

Jiang, Yong et al. (2017). "A quantitative genetic framework highlights the role of epistatic effects for grain-yield heterosis in bread wheat". In: *Nature genetics* 49.12, p. 1741.

Kashkush, Khalil, Moshe Feldman, and Avraham A Levy (2002). "Gene loss, silencing and activation in a newly synthesized wheat allotetraploid". In: *Genetics* 160.4, pp. 1651–1659.

— (2003). "Transcriptional activation of retrotransposons alters the expression of adjacent genes in wheat". In: *Nature genetics* 33.1, p. 102.

Keshet, Ilana, Judy Lieman-Hurwitz, and Howard Cedar (1986). "DNA methylation affects the formation of active chromatin". In: *Cell* 44.4, pp. 535–543.

Krasileva, Ksenia V et al. (2017). "Uncovering hidden variation in polyploid wheat". In: *Proceedings of the National Academy of Sciences* 114.6, E913–E921.

Kusterer, Barbara et al. (2007). "Analysis of a triple testcross design with recombinant inbred lines reveals a significant role of epistasis in heterosis for biomass-related traits in Arabidopsis". In: *Genetics* 175.4, pp. 2009–2017.

Lamkey, Kendall R, Bruce J Schnicker, and Albrecht E Melchinger (1995). "Epistasis in an elite maize hybrid and choice of generation for inbred line development". In: *Crop science* 35.5, pp. 1272–1281.

Law, CN, J Sutka, and AJ Worland (1978). "A genetic study of day-length response in wheat". In: *Heredity* 41.2, p. 185.

Lee, Hyeon-Se and Z Jeffrey Chen (2001). "Protein-coding genes are epigenetically regulated in Arabidopsis polyploids". In: *Proceedings of the National Academy of Sciences* 98.12, pp. 6753–6758.

Lee, Joshua A, C Clark Cockerham, and FH Smith (1968). "The inheritance of gossypol level in Gossypium I. additive, dominance, epistatic, and maternal

effects associated with seed gossypol in two varieties of Gossypium Hirsutum L". In: *Genetics* 59.2, pp. 285–298.

Lewis, EB (1954). "The theory and application of a new method of detecting chromosomal rearrangements in Drosophila melanogaster". In: *The American Naturalist* 88.841, pp. 225–239.

Lewontin, RC (1964). "The interaction of selection and linkage. I. General considerations; heterotic models". In: *Genetics* 49.1, pp. 49–67.

Li, Heng and Richard Durbin (2009). "Fast and accurate short read alignment with Burrows–Wheeler transform". In: *Bioinformatics* 25.14, pp. 1754–1760.

Li, Lanzhi et al. (2008). "Dominance, over-dominance and epistasis condition the heterosis in two heterotic rice hybrids". In: *Genetics*.

Liaw, Andy and Matthew Wiener (2002). "Classification and Regression by randomForest". In: *R News* 2.3, pp. 18–22. URL: `http://CRAN.R-project.org/doc/Rnews/`.

Lin, DY and D Zeng (2006). "Likelihood-based inference on haplotype effects in genetic association studies". In: *Journal of the American Statistical Association* 101.473, pp. 89–104.

Liu, Shao-Lun, Gregory J Baute, and Keith L Adams (2011). "Organ and cell type–specific complementary expression patterns and regulatory neofunctionalization between duplicated genes in Arabidopsis thaliana". In: *Genome biology and evolution* 3, pp. 1419–1436.

Liu, Zhenlan and Keith L Adams (2007). "Expression partitioning between genes duplicated by polyploidy under abiotic stress and during organ development". In: *Current Biology* 17.19, pp. 1669–1674.

Liu, Zhenshan et al. (2015). "Temporal transcriptome profiling reveals expression partitioning of homeologous genes contributing to heat and drought acclimation in wheat (Triticum aestivum L.)" In: *BMC plant biology* 15.1, p. 152.

Lukens, Lewis N and John Doebley (1999). "Epistatic and environmental interactions for quantitative trait loci involved in maize evolution". In: *Genetics Research* 74.3, pp. 291–302.

Lynch, Michael and John S Conery (2000). "The evolutionary fate and consequences of duplicate genes". In: *Science* 290.5494, pp. 1151–1155.

— (2003). "The origins of genome complexity". In: *science* 302.5649, pp. 1401–1404.

Lynch, Michael and Allan Force (2000). "The probability of duplicate gene preservation by subfunctionalization". In: *Genetics* 154.1, pp. 459–473.

Mac Key, James (1970). "Significance of mating systems for chromosomes and gametes in polyploids". In: *Hereditas* 66.2, pp. 165–176.

Malmberg, Russell L et al. (2005). "Epistasis for fitness-related quantitative traits in Arabidopsis thaliana grown in the field and in the greenhouse". In: *Genetics* 171.4, pp. 2013–2027.

Marcussen, Thomas et al. (2014). "Ancient hybridizations among the ancestral genomes of bread wheat". In: *Science* 345.6194, p. 1250092.

Martini, Johannes WR et al. (2016). "Epistasis and covariance: how gene interaction translates into genomic relationship". In: *Theoretical and Applied Genetics* 129.5, pp. 963–976.

Martini, Johannes WR et al. (2017). "Genomic prediction with epistasis models: on the marker-coding-dependent performance of the extended GBLUP and properties of the categorical epistasis model (CE)". In: *BMC bioinformatics* 18.1, p. 3.

McClelland, Michael, Michael Nelson, and Eberhard Raschke (1994). "Effect of site-specific modification on restriction endonucleases and DNA modification methyltransferases". In: *Nucleic acids research* 22.17, pp. 3640–3659.

McClintock, Barbara (1984). "The significance of responses of the genome to challenge". In: *Science* 266.4676, pp. 792–801.

Melchinger, AE, HH Geiger, and FW Schnell (1986). "Epistasis in maize (Zea mays L.)" In: *Theoretical and applied genetics* 72.2, pp. 231–239.

Meuwissen, THE, Benjamin J Hayes, and ME Goddard (2001). "Prediction of total genetic value using genome-wide dense marker maps". In: *Genetics* 157.4, pp. 1819–1829.

Microsoft, R Core Team (2017). *Microsoft R Open*. Microsoft. Redmond, Washington. URL: `https://mran.microsoft.com/`.

Muller, HJ (1925). "Why polyploidy is rarer in animals than in plants". In: *The American Naturalist* 59.663, pp. 346–353.

Mutti, Jasdeep S, Ramanjot K Bhullar, and Kulvinder S Gill (2017). "Evolution of gene expression balance among homeologs of natural polyploids". In: *G3: Genes, Genomes, Genetics* 7.4, pp. 1225–1237.

Nagamine, Yoshitaka et al. (2012). "Localising loci underlying complex trait variation using regional genomic relationship mapping". In: *PloS one* 7.10, e46501.

Nagasaki, Hideki et al. (2005). "Species-specific variation of alternative splicing and transcriptional initiation in six eukaryotes". In: *Gene* 364, pp. 53–62.

Nejati-Javaremi, A, C Smith, and JP Gibson (1997). "Effect of total allelic relationship on accuracy of evaluation and response to selection." In: *Journal of animal science* 75.7, pp. 1738–1745.

Neyman, Jerzy and Egon S Pearson (1933). "IX. On the problem of the most efficient tests of statistical hypotheses". In: *Phil. Trans. R. Soc. Lond. A* 231.694-706, pp. 289–337.

Ohno, Susumu (1970). *Evolution by Gene Duplication*. New York: Springer.

Ohta, Tomoko (1987). "Simulating evolution by gene duplication". In: *Genetics* 115.1, pp. 207–213.

Okeke, Uche Godfrey et al. (2018). "Regional Heritability Mapping Provides Insights into Dry Matter Content in African White and Yellow Cassava Populations". In: *The plant genome* 11.1.

Orr, H Allen (1990). "" Why Polyploidy is Rarer in Animals Than in Plants" Revisited". In: *The American Naturalist* 136.6, pp. 759–770.

Osborn, Thomas C et al. (2003). "Understanding mechanisms of novel gene expression in polyploids". In: *Trends in genetics* 19.3, pp. 141–147.

Ozkan, Hakan, Avraham A Levy, and Moshe Feldman (2001). "Allopolyploidy-induced rapid genome evolution in the wheat (Aegilops–Triticum) group". In: *The Plant Cell* 13.8, pp. 1735–1747.

Paixão, Tiago and Nicholas H Barton (2016). "The effect of gene interactions on the long-term response to selection". In: *Proceedings of the National Academy of Sciences* 113.16, pp. 4422–4427.

Patterson, Nick, Alkes L Price, and David Reich (2006). "Population structure and eigenanalysis". In: *PLoS genetics* 2.12, e190.

Paux, Etienne et al. (2008). "A physical map of the 1-gigabase bread wheat chromosome 3B". In: *science* 322.5898, pp. 101–104.

Pearce, Stephen et al. (2011). "Molecular characterization of Rht-1 dwarfing genes in hexaploid wheat". In: *Plant physiology* 157.4, pp. 1820–1831.

Peng, Jinrong et al. (1999). "Green revolutiongenes encode mutant gibberellin response modulators". In: *Nature* 400.6741, p. 256.

Pfeifer, Matthias et al. (2014). "Genome interplay in the grain transcriptome of hexaploid bread wheat". In: *Science* 345.6194, p. 1250091.

Phillips, Patrick C (2008). "Epistasisthe essential role of gene interactions in the structure and evolution of genetic systems". In: *Nature Reviews Genetics* 9.11, p. 855.

Poland, Jesse (Mar. 6, 2018). personal communication.

Poland, Jesse A et al. (2012). "Development of high-density genetic maps for barley and wheat using a novel two-enzyme genotyping-by-sequencing approach". In: *PloS one* 7.2, e32253.

Price, Alkes L et al. (2006). "Principal components analysis corrects for stratification in genome-wide association studies". In: *Nature genetics* 38.8, pp. 904–909.

Pumphrey, Michael et al. (2009). "Nonadditive expression of homoeologous genes is established upon polyploidization in hexaploid wheat". In: *Genetics* 181.3, pp. 1147–1157.

R Core Team (2015). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria. URL: `http://www.R-project.org/`.

Rastogi, Shruti and David A Liberles (2005). "Subfunctionalization of duplicated genes as a transition state to neofunctionalization". In: *BMC evolutionary biology* 5.1, p. 28.

Resende, Rafael Tassinari et al. (2017). "Regional heritability mapping and genome-wide association identify loci for complex growth, wood and disease resistance traits in Eucalyptus". In: *New Phytologist* 213.3, pp. 1287–1300.

Riggio, Valentina and Ricardo Pong-Wong (2014). "Regional Heritability Mapping to identify loci underlying genetic variation of complex traits". In: *BMC proceedings*. Vol. 8. 5. BioMed Central, S3.

Ritchie, Marylyn D (2011). "Using biological knowledge to uncover the mystery in the search for epistasis in genome-wide association studies". In: *Annals of human genetics* 75.1, pp. 172–182.

Rutkoski, Jessica E et al. (2013). "Imputation of unordered markers and the impact on genomic selection accuracy". In: *G3: Genes, Genomes, Genetics* 3.3, pp. 427–439.

Salamini, Francesco et al. (2002). "Genetics and geography of wild cereal domestication in the Near East". In: *Nature Reviews Genetics* 3.6, p. 429.

Scarth, R and CN Law (1983). "The location of the photoperiod gene, Ppd2 and an additional genetic factor for ear-emergence time on chromosome 2B of wheat". In: *Heredity* 51.3, p. 607.

Segovia-Lerma, A et al. (2004). "Population-based diallel analyses among nine historically recognized alfalfa germplasms". In: *Theoretical and applied genetics* 109.8, pp. 1568–1575.

Segre, Daniel et al. (2005). "Modular epistasis in yeast metabolism". In: *Nature genetics* 37.1, p. 77.

Shen, Guojing et al. (2014). "Dominance and epistasis are the main contributors to heterosis for plant height in rice". In: *Plant Science* 215, pp. 11–18.

Shirali, Masoud et al. (2016). "Regional heritability mapping method helps explain missing heritability of blood lipid traits in isolated populations". In: *Heredity* 116.3, p. 333.

Soltis, Douglas E et al. (2009). "Polyploidy and angiosperm diversification". In: *American journal of botany* 96.1, pp. 336–348.

Soltis, Pamela S and Douglas E Soltis (2009). "The role of hybridization in plant speciation". In: *Annual review of plant biology* 60, pp. 561–588.

Sorrells, Mark E et al. (2011). "Reconstruction of the Synthetic W7984× Opata M85 wheat reference population". In: *Genome* 54.11, pp. 875–882.

Stekhoven, Daniel J and Peter Bühlmann (2011). "MissForestnon-parametric missing value imputation for mixed-type data". In: *Bioinformatics* 28.1, pp. 112–118.

Stoltzfus, Arlin (1999). "On the possibility of constructive neutral evolution". In: *Journal of Molecular Evolution* 49.2, pp. 169–181.

Strandén, Ismo and DJ Garrick (2009). "Derivation of equivalent computing algorithms for genomic predictions and reliabilities of animal merit". In: *Journal of dairy science* 92.6, pp. 2971–2975.

Stuber, Charles W et al. (1992). "Identification of genetic factors contributing to heterosis in a hybrid from two elite maize inbred lines using molecular markers." In: *Genetics* 132.3, pp. 823–839.

Stuber, CW and RH Moll (1971). "Epistasis in maize (Zea mays L.). II: Comparison of selected with unselected populations". In: *Genetics* 67.1, pp. 137–149.

Tantau, Till (Apr. 8, 2018). *The TikZ and PGF Packages. Manual for version 3.0.1.* URL: `http://sourceforge.net/projects/pgf/`.

Technow, Frank et al. (2012). "Genomic prediction of hybrid performance in maize with models incorporating dominance and population specific marker effects". In: *Theoretical and Applied Genetics* 125.6, pp. 1181–1194.

VanRaden, PM (2008). "Efficient methods to compute genomic predictions". In: *Journal of dairy science* 91.11, pp. 4414–4423.

Veitia, Reiner A (2007). "Exploring the molecular etiology of dominant-negative mutations". In: *The Plant Cell* 19.12, pp. 3843–3851.

Vitezica, Zulma G, Luis Varona, and Andres Legarra (2013). "On the additive and dominant variance and covariance of individuals within the genomic selection scope". In: *Genetics* 195.4, pp. 1223–1230.

Vitezica, Zulma G et al. (2017). "Orthogonal Estimates of Variances for Additive, Dominance, and Epistatic Effects in Populations". In: *Genetics* 206.3, pp. 1297–1307.

Wagner, Andreas (2005). "Distributed robustness versus redundancy as causes of mutational robustness". In: *Bioessays* 27.2, pp. 176–188.

Walsh, J Bruce (1995). "How often do duplicated genes evolve new functions?" In: *Genetics* 139.1, pp. 421–428.

Wang, Jianlin et al. (2004). "Stochastic and epigenetic changes of gene expression in Arabidopsis polyploids". In: *Genetics* 167.4, pp. 1961–1973.

Wang, Jirui et al. (2013). "Aegilops tauschii single nucleotide polymorphisms shed light on the origins of wheat D-genome genetic diversity and pinpoint the geographic origin of hexaploid wheat". In: *New phytologist* 198.3, pp. 925–937.

Welsh J.R. Keim D.L., Pirasteh B. and Richards R.D. (1973). "Genetic control of photoperiod response in wheat". In: *Proceedings of the fourth International Wheat Genetics Symposium*. Ed. by Sears E.R. and Sears L.E.S. University of Missouri, Columbia, Mo., pp. 879–884.

Wendel, Jonathan F (2000). "Genome evolution in polyploids". In: *Plant molecular evolution*. Springer, pp. 225–249.

Wolf, Duane P and R Hallauer (1997). "Triple testcross analysis to detect epistasis in maize". In: *Crop science* 37.3, pp. 763–770.

Wolfe, Marnin D et al. (2016). "Marker-based estimates reveal significant non-additive effects in clonally propagated cassava (Manihot esculenta): implica-

tions for the prediction of total genetic value and the selection of varieties". In: *G3: Genes, Genomes, Genetics*, g3–116.

Wood, Andrew R et al. (2014). "Another explanation for apparent epistasis". In: *Nature* 514.7520, E3.

Wu, Xianshan, Xiaoping Chang, and Ruilian Jing (2012). "Genetic insight into yield-associated traits of wheat grown in multiple rain-fed environments". In: *PloS one* 7.2, e31249.

Xu, Shizhong and Zhenyu Jia (2007). "Genomewide analysis of epistatic effects for quantitative traits in barley". In: *Genetics* 175.4, pp. 1955–1963.

Yu, SB et al. (1997). "Importance of epistasis as the genetic basis of heterosis in an elite rice hybrid". In: *Proceedings of the National Academy of Sciences* 94.17, pp. 9226–9231.

Zeng, Zhao-Bang and C Clark Cockerham (1993). "Mutation models and quantitative genetic variation." In: *Genetics* 133.3, pp. 729–736.

Zeng, Zhao-Bang, Tao Wang, and Wei Zou (2005). "Modeling quantitative trait loci and interpretation of models". In: *Genetics* 169.3, pp. 1711–1725.

Zhang, Li et al. (2012). "Modeling haplotype-haplotype interactions in case-control genetic association studies". In: *Frontiers in genetics* 3, p. 2.

Zhang, Yumei et al. (2016). "Expression partitioning of homeologs and tandem duplications contribute to salt tolerance in wheat (Triticum aestivum L.)" In: *Scientific reports* 6, p. 21476.

Zhang, Zhiwu et al. (2010). "Mixed linear model approach adapted for genome-wide association studies". In: *Nature genetics* 42.4, pp. 355–360.