Utilizing high-density genotyping data for wheat improvement

Alexander Coulton

A dissertation submitted to the University of Bristol in accordance with the requirements for the awards of degree of Doctor of Philosophy in the Faculty of Life Sciences

Under the supervision of Professor Keith J. Edwards

University of Bristol

Department of Life Sciences

March 2021

Abstract

Wheat's wide-ranging distribution, in addition to its vast levels of production and consumption make it an essential component of global food security. Ever-increasing population sizes necessitate an increase in global wheat yields to match. This thesis aims to contribute to this goal by addressing a broad range of seemingly disparate themes: evolution, recombination and segregation distortion. What unites these themes is their methodological underpinnings - the use of high-density genotyping arrays, which have undergone considerable development in the past decade.

The genetic diversity of wheat is limited by bottlenecks that have occurred in its evolutionary history, both through polyploidization and domestication. This limitation presents difficulties for future yield increases, potentially increasing wheat's susceptibility to pathogens. One area of interest is the rate of novel polymorphism formation over time. The results presented here indicate that this question will be difficult to answer using molecular clock methodology.

Another route to increasing wheat yield may be the manipulation of wheat recombination distribution, removing large areas of linkage drag in the central regions of chromosomes. Previous work in barley suggests that an increase in environmental temperature could shift recombination distribution inwards. The results presented here suggest that whilst this might be the case for some chromosomes in wheat, for the majority of chromosomes, recombination distribution is unaffected by changes in temperature.

Segregation distortion, a deviation from Mendelian ratios in progeny of a cross, is also investigated here, with a focus on current practices of detection in the literature. My results indicate that many studies have been using inappropriate methods for the detection of segregation distortion.

Also presented in this thesis are novel methods and tools for wheat research, such as the

AutoCloner gene-cloning pipeline, allowing researchers to efficiently clone large numbers of genes in previously unsequenced varieties.

Author's declaration

I declare that the work in this dissertation was carried out in accordance with the requirements of the University's Regulations and Code of Practice for Research Degree Programmes and that it has not been submitted for any other academic award. Except where indicated by specific reference in the text, the work is the candidate's own work. Work done in collaboration with, or with the assistance of, others, is indicated as such. Any views expressed in the dissertation are those of the author.

Acknowledgements

First and foremost I would like to thank my primary supervisor, Professor Keith Edwards for his unending support during my PhD studies, whether that be through technical advice, moral support, providing opportunities to advance my own scientific career through conferences and travel, or simply imparting wisdom acquired through his many years in science. I would also like to thank my secondary supervisors, Gary Barker and Sacha Allen for their wisdom and support, whether that be through general scientific advice, reading manuscripts, providing access to invaluable computing resources, or organising travel and teaching opportunities.

This thesis would also not be possible without the many other members of the lab that provided invaluable knowledge and help during the process - Amanda Burridge for her wealth of knowledge on molecular procedures and help in the lab, Paul Wilkinson for advice and guidance on bioinformatic procedures, Mark Winfield for helpful theoretical discussions in the office.

My family and friends have also been invaluable to me, in particular my father, mother and partner Swetha, who have always been ready to lend an ear when I'm in need of a chat. I'm also thankful to the other members of the Biological Sciences department, whether students or staff, who I've had the pleasure of sharing my time with. I'm always impressed by the friendliness and warmth of the community, as well as the fascinating array of research in the department.

Contents

List of Tables

List of Figures

Abbreviations

ASY1	Arabidopsis asynaptic 1
AxP	Apogee x Paragon
BAC	Bacterial artificial chromosome
BLAS	Basic linear algebra subprogram
CENH3	Centromere specific histone 3
CO	Crossover
DHJ	Double Holliday junction
DQC	Dish quality control
DSB	Double-stranded break
FAOSTAT	Food and Agriculture Organization Corporate Statistical Database
FDR	False discovery rate
FISH	Fluorescense in-situ hybridization
GC	Guanine-cytosine
HSPs	High-scoring segment pairs
IWGSC	International Wheat Genome Sequencing Consortium
KASP	Kompetitive allele-specific PCR
LOD	Logarithm of odds
MRD	Mean recombination distance
MSA	Multiple sequence alignment
MTP	Microtitre plate
NCO	Non-crossover
PCR	Polymerase chain reaction
RAPD	Random amplified polymorphic DNA

RFLP	Restriction fragment length polymorphism
RIL	Recombinant inbred line
Rsp	Responder
SC	Synaptonemal complex
SD	Segregation distortion
SDR	Segregation distortion region
SEI	Single-end invasion
SNP	Single nucleotide polymorphism
Sd	Segregation distorter
TM	Melting temperature
TMRCA	Time to most recent common ancestor
WISP	Wheat improvement strategic programme
ZYP1	zipper 1

1 General introduction

1.1 Wheat and the world

1.1.1 Economic importance of wheat

Roughly 9000 years ago in the fertile crescent, now modern-day Turkey, humans began the transition from searching for food to cultivating their own, sparking the first agricultural revolution. This period marked the inception of bread wheat, which is now a globally distributed crop with huge importance to the world economy. In 2018, global wheat production was 734 million tonnes (FAOSTAT), making it the third highest food crop in terms of production, eclipsed only by rice (782 million tonnes) and maize (1.14 billion tonnes). One of the upcoming challenges of the 21ˆst century is the divergence of human population growth and crop yield growth trajectories, which if not addressed will result in a global shortage of food supplies in the coming decades. The 2017 UN report of world population prospects projected a global population of 8.5 billion people by 2030, increasing to 9.7 billion by 2050. It is therefore important that governments and private companies invest into the advancement of wheat research, facilitating the development of tools, such as the recently released IWGSC chromosome-level genome sequence, that will significantly increase our knowledge of wheat, and hopefully allow us to produce varieties that are suited to meeting this increased demand.

1.1.2 Wheat breeding strategies

Although agricultural practices, including the active cultivation of plants for food, date to over 10,000 years ago, breeding based on the theoretical foundation of genetics is very recent. Mendel published his hybridization studies on peas in 1866 [@allenMendelModernGenetics2003], which were not regarded seriously until around 1900. This provided breeders with more concrete knowledge of inheritance, but even so, breeding was still limited to the selection of plants based on their phenotypic characteristics, and consequently was largely focussed on the selection of qualitative traits underpinned by

a small number of genes, such as plant height, which notably was the driving factor behind the massive yield increases of the green revolution. Accurate dissection of quantitative traits, as well as heterozygotes with recessive alleles remained elusive without knowledge of plant genotypes.

The transition from using phenotypic traits to inform plant breeding strategies to molecular markers that survey the genetic material was revolutionary, potentially saving breeders a significant amount of time. Genetic markers allow researchers or breeders to survey large numbers of plants for the trait of interest without having to grow plants to late stages of development. Before examining the current state of wheat breeding, it will be illuminating to review the history of genotyping as a means of providing context. Genetic markers have undergone significant development in the last few decades. The first genetic markers were restriction fragment length polymorphism (RFLP) markers, initially developed in humans [@botsteinConstructionGeneticLinkage1980] and then later for maize [@helentjarisRestrictionFragmentPolymorphisms1985] and wheat [@chaoRFLPbasedGeneticMaps1989]. These relied on fragmentation of the DNA via restriction enzymes, followed by gel electrophoresis and hybridization of probe sequences to the DNA, which defined polymorphisms. The invention of the polymerase chain reaction (PCR) then opened up several new techniques for producing genetic markers, such as random amplified polymorphic DNA (RAPD) [@williamsDNAPolymorphismsAmplified1990], DNA amplification fingerprinting [@caetano], amplified fragment length polymorphism (AFLP) [@barrettAFLPBasedGeneticDiversity1998], microsatellites [@conditAbundanceDNASequence1991; @devosApplicationTwoMicrosatellite1995]. These methods were restricted by their reliance on agarose or polyacrylamide gels, which limit the number of polymorphisms that can be analysed concurrently. Later on it became widely acknowledged that single nucleotide polymorphisms (SNPs), would become the standard [@henryHenryPlantGenotyping2001] for human and plant genotyping due to their abundance in the genomes of organisms, meaning that for any particular gene of interest, there would likely be a SNP within close proximity, allowing the inheritance of the gene to be tracked through generations.

Progress in SNP genotyping in wheat has lagged significantly behind humans and other crops due to the complexity and size of the wheat genome in comparison, which is hexaploid and comprises around 16 Gb. The current state of wheat genotyping has seen the emergence of a small number of dominant technologies for surveying SNPs. Kompetitive allele specific PCR (KASP) [@allenTranscriptspecificSinglenucleotidePolymorphism2011; @semagnSingleNucleotidePolymorphism2014] provides an efficient, cost-effective method for genotyping samples at a small number of SNP loci. This method is ideal for the situation in which the researcher already has a target gene of interest in the genome. In addition, SNP arrays have also been developed for simultaneous genotyping of large numbers of samples (90 400) and SNPs (35,000 820,000).

1.2 SNP Array development

SNP arrays are ultimately based on the hybridization of DNA fragments of interest to allele-specific probes that are immobilized on a surface. Arrays are typically restricted to biallelic SNPs for ease of detection. Two probes are present for each SNP to be detected, each of which contains a different base at the SNP position, along with identical flanking sequences. These probes have different fluorescent tags attached to them (for example red and green), such that after hybridization, non-hybridized probes will be removed and the genotype of the sample at that allele will be deduced through the colour of the fluorescent signal. A mixture of the two colours indicates that the sample was heterozygous at that locus. The first SNP array for wheat was described in 2013, and was reliant on the Illumina iSelect technology, comprising 9,000 SNPs [@cavanaghGenomewideComparativeDiversity2013] discovered through transcriptome sequencing of 26 bread wheat accessions. This was later expanded on to produce an array containing 90,000 SNPs [@wangCharacterizationPolyploidWheat2014] based on 19 bread wheat varieties. Subsequently, an array was developed with a much higher density of SNPs using Affymetrixs Axiom technology, containing 820,000 SNPs discovered via exome capture from 43 bread wheat varieties and relatives, enabling, amongst other typical array applications, the tracking of introgressions from wide crosses in

wheat. These SNPs were then filtered to produce a high-quality set of 35,000 optimized for elite UK bread wheat varieties, known as the Wheat Breeders array [@allenCharacterizationWheatBreeders2016]. Notably, all of these arrays were built around genic sequences, whether in their full genomic of transcriptomic forms. To address this, a further array was designed and released [@rimbertHighThroughputSNP2018] that contained 280,226 SNPs, many of which were intergenic, and also utilized the IWGSC RefSeq v1.0 chromosome-level genome assembly [@consortiumiwgscShiftingLimitsWheat2018] as a reference for alignment of sequencing reads.

1.3 Wheat Genome Development

The assembly of a chromosome-level genome for wheat has proven to be a significant challenge for the scientific community. The first attempt at sequencing a complete chromosome of wheat, chromosome 3B, which is two-times bigger than the entire rice genome, came in the form of long insert bacterial artificial chromosome (BAC) libraries [@pauxPhysicalMap1Gigabase2008]. The Chinese Spring variety of wheat was chosen on the basis of its previous use in cytogenetic studies as well as the availability of aneuploid lines for every chromosome [@endoDeletionStocksCommon1996]. This progressed to a full genome assembly with the rise of next generation sequencing [@brenchleyAnalysisBreadWheat2012], albeit a highly fragmented version with short contigs. Several further assemblies were then produced [@chapmanWholegenomeShotgunApproach2015; @clavijoImprovedAssemblyAnnotation2017; @iwgscChromosomebasedDraftSequence2014; @ziminFirstNearcompleteAssembly2017] before the canonical assembly, the International Wheat Genome Sequencing Consortium Reference Sequence v1.0 [@consortiumiwgscShiftingLimitsWheat2018] was released. As the wheat genome is highly repetitive, not only with duplicate copies of genes due to hexaploidy but also due to a large number of transposable elements, de novo assembly of the genome is very difficult. The authors of the IWGSC assembly therefore used a variety of methods to overcome these difficulties, including traditional next-generation sequencing, Hi-C [@lieberman], Bionano optical maps, radiation hybrid maps, genetic

mapping and BAC libraries with microtitre plate (MTP) sanger sequencing . Since the release of the IWGSC assembly, further progress has been made in sequencing other wheat varieties, with the 10 Wheat Genomes project [@walkowiakMultiple-WheatGenomes2020] providing whole genome assemblies of varieties Mace, Lancer, CDC Landmark, Julius, Norin61, ArinaLrFor, Jagger, Cadenza, Paragon, Robigus and Claire, although at the time of writing, the latter four assemblies are not chromosome-level, and instead are comprised of scaffolds.

1.4 Wheat Diversity and Evolution

1.4.1 Phylogenetic placement of wheat

Wheat, or more formally *Triticum aestivum* L., is an allohexaploid species contained within the Poaceae, or the grass family, which is a large clade containing all three major food crops (wheat, rice and maize) as well as many other common grasses. Wheat is said to have an AABBDD genome constitution, a nomenclature originally derived from hybridization experiments between polyploids and diploid grasses [@kiharaConjugationHomologousChromosomes1929; @lilienfeldKiharaGenomeAnalysisTriticum1951]. Hybridizations in which meiosis proceeded normally, e.g. without nondisjunction of chromosomes and resulting aneuploidy, were said to have homologous genomes [@lilienfeldKiharaGenomeAnalysisTriticum1951], and the parents were therefore assigned the same letter / genome designation. From the viewpoint of modern phylogenetics, based on an abundance of genomic sequence data, this concept could be simplified into the dates of divergence between genomes or sequences, however these designations are still useful in a plant breeding context, and have thus remained in widespread usage in the agricultural scientific literature.

1.4.2 Overview of close relatives of wheat and estimated time of origin

Contained within the Poaceae is the Pooideae subfamily, which is comprised of important food crops such as barley, oats, rye and wheat. *Triticum*, the wheat genus,

contains many species of varying ploidies and significance to the global economy, including hexaploid bread wheat (*Triticum aestivum* L.), used for making bread; tetraploid durum wheat, used for making pasta (*Triticum turgidum* conv. durum); tetraploid emmer wheat (*Triticum turgidum subsp. dicoccoides*), the wild progenitor of modern cultivated tetraploid wheats; hexaploid spelt (*Triticum spelta*), an important crop in the Bronze age [@salaminiGeneticsGeographyWild2002] which is now used as a health food; diploid einkorn wheat (*Triticum monococcum* L.), one of the first grasses to be actively cultivated by humans and *Triticum urartu*, the closest living relative of the A subgenome of bread wheat. Another genus that played an important role in wheat evolution is *Aegilops*, commonly known as goatgrasses, with evidence pointing to *Aegilops speltoides* being the closest living relative to the B subgenome of wheat [@petersenCa2SignallingPancreatitis2006], and *Aegilops tauschii* the closest relative to the D subgenome of wheat. A phylogenomic molecular clock analysis of bread wheat and its close diploid relatives [@marcussenAncientHybridizationsAncestral2014] suggests that the D genome lineage of wheat arose around 5.5 million years ago through homoploid hybridization, i.e. where the hybrid retains the same number of chromosomes as the parental species, of the B and A genome ancestors. Later, less than 0.8 million years ago, the A and B genome progenitors underwent allopolyploid hybridization to form the tetraploid AABB plant, which underwent a subsequent allopolyploid hybridization event with the D genome progenitor. Archaeological evidence suggests this final hybridization event occurred around 9000 years ago, with the occurrence of hexaploid free-threshing wheats at Cafer Höyük that have been radiocarbon-dated to ~8700 years ago [@nesbittWheatEvolutionIntegrating2001]. This order of hybridization is further supported by the relative lack of SNPs in the D genome of bread wheat compared to A and B genomes [@allenCharacterizationWheatBreeders2016], which is congruent with a more recent origin.

1.5 Phenotypic changes in wheat domestication

The process of domestication involves the accumulation of traits that are often detrimental to the selected organism in the wild, but beneficial to humans in terms of agriculture. The first of such traits to be introduced into cultivated wheat was a non-brittle rachis [@charmetWheatDomesticationLessons2011], which causes the spikelets to be retained on the spike rather than be dispersed into the environment, meaning yield is increased as much of the edible material of the plant, the kernel, is retained for harvesting. The gene conferring a non-brittle rachis, Br, is located on chromosome 2A [@pelegGeneticAnalysisWheat2011] and has monogenetic inheritance, and would thus be relatively straightforward to select and breed into a population of wheat plants. The second important gene involved in wheat domestication is the Tg or Tenacious glume, located on the short arm of chromosome 2D [@charmetWheatDomesticationLessons2011; @soodMajorThreshabilityGenes2009]. This controls ease at which the glume, the protective bract covering each spikelet, separates from the grain, thus plants with recessive tg are easier to thresh. Perhaps the most important gene involved in the domestication of wheat is the Q gene, an APETELATA2-like transcription factor located on chromosome 5A [@simonsMolecularCharacterizationMajor2006]. This gene is pleiotropic, producing effects on spike length, rachis fragility (and therefore threshability), glume tenacity, spike emergence time and plant height. Q is therefore a crucial development in cultivated wheat that significantly enhanced the efficiency with which farmers could harvest their grain. As we can see, one common factor uniting these genes is that they are all qualitative traits controlled for the most part by single loci. We can expect that since there was strong selection for these genes that wheat will have significantly reduced genetic diversity in the regions surrounding these genes in the genome.

1.6 Genetic mapping

1.6.1 Origin of genetic mapping

First conceived by Sturtevant in the fruit fly drosophila [@sturtevantLinearArrangementSix1913], genetic mapping is a method of deducing the linear order of genes or molecular markers on a chromosome via their segregation ratios in the progeny of a biparental cross. It is based on the principle that the number of recombination events that occur between any pair of loci on a single chromosome is proportional to the distance between those loci. Whilst in organisms with smaller genomes, such as the model plant Arabidopsis, genetic mapping has been almost completely superseded by more advanced techniques such as whole-genome sequencing, in many crops, including wheat, it remains an important technique for examining the genome, as whole-genome sequencing remains challenging and expensive due to the size of the genome. Genetic mapping has progressed over the years from using phenotypic markers of genes - yellow body, white eyes, vermillion wings, miniature wings and rudimentary wings were the traits used by Sturtevant - to using molecular markers such as SNPs [@allenCharacterizationWheatBreeders2016; @allenTranscriptspecificSinglenucleotidePolymorphism2011].

1.6.2 Producing genetic maps with SNPs

As genotyping technology has advanced, genetic maps have become increasingly large, now often including thousands of molecular markers per map [@allenCharacterizationWheatBreeders2016]. Only SNPs that are polymorphic between parents can be used for the creation of a map. In addition, an ideal SNP would be codominant between parents such that the heterozygotes are able to be distinguished, as this provides more information on recombination events. We can then assign letters to represent the genotype of a sample at any particular SNP, for example an “A” could represent a sample that is homozygous for the allele from the first parent of a biparental cross, likewise a “B” could represent a homozygote of the second parental allele, whilst a “H” could represent a heterozygote. Examination of the genotypes at any pair of molecular mark-

ers would then reveal whether a recombination event had occurred between them, as they would have different genotypes. There are some exceptions to this, for example when two recombination events occur between a pair of markers, or when both of the markers are heterozygous and both of the gametes that made up the zygote of the sample underwent recombination between these markers. Genetic mapping becomes feasible when we expand these principles to the population level: if a large proportion of individuals in the population have recombination events between two loci, the loci must be far from each other on the chromosome. This proportion is known as the recombination fraction r, which, in absence of sampling error, ranges from $0 < r < 0.5$. The upper limit is 0.5 due to double recombination events: as the distance between two loci increases, odd and even numbers of crossovers between them become equally likely [@xuPrinciplesStatisticalGenomics2013].

1.6.3 Computational stages of genetic mapping

After genotyping, genetic mapping involves three primary computational stages, which are the clustering of markers, ordering of markers, and calculation of genetic distance between markers. Of these stages, the ordering of markers is the most computationally intensive, with $n!/2$ potential orders for n markers [@mesterConstructingLargescaleGenetic2003]. With modern computational hardware it is almost impossible to examine every possible order in a reasonable amount of @wuEfficientAccurateConstruction2008] have been created to try and optimise this. A notable development is the MSTmap software, which uses an algorithm based on the minimum spanning tree of a graph to determine the order of markers on the map, which not only was shown to be significantly faster than previous methods [@wuEfficientAccurateConstruction2008], but was also released free-of-charge to the community as open source software, and has since been developed into an R package that simplifies its use even further [@taylorPackageASMapEfficient2017]. Other algorithms of comparable speed, such as MultiPoint [@mesterFastAccurateConstruction2015] are proprietary software and have cumbersome graphical interfaces that are very difficult to incorporate into automated bioinformatics pipelines.

After the inference of both marker clustering and ordering, the calculation of genetic distances between markers is trivial, with either the Haldane or Kosambi mapping functions most often used to convert recombination fractions between markers to units of centiMorgans, the latter accounting for crossover interference.

1.7 Experimental lines

There are several types of experimental line used commonly in wheat research, each with its own advantages and disadvantages depending on the research application. The simplest is an F1 line, which is a cross between two different parental varieties, and is heterozygous at every locus at which the parents differ. Since wheat is self-fertilizing, F1 plants can be selfed to produce F2 lines, which unlike the F1 are genetically different due to the difference in recombination position and frequency between meioses; this difference forms the basis for creation of a genetic map. With repetition of this selfing process for several generations, the plants eventually become completely homozygous at every locus, and are referred to as recombinant inbred lines (RIL). These lines can be more useful than F2 lines for genetic mapping, as the lack of heterozygosity allows increased detection of recombination events. In addition, they also have a stable genotype, and will not change significantly over generations, meaning they could form useful germplasm for further input to breeding programmes. The obvious downside is that they take much longer than F2 plants to produce. Doubled haploid lines allow researchers to expedite this process significantly, achieving complete homozygosity in a single generation by taking the pollen cells of F1 lines and subjecting them to chemicals, usually caffeine or colchicine, which disrupt meiosis and cause chromosome doubling. A potential downside of this method is that a selection pressure is applied favouring alleles that allow the pollen to thrive during tissue culture, which are not necessarily beneficial to agronomic and physiological traits in later stages of development. It has also been noted in the literature that the tissue culture process during doubled haploid production causes segregation distortion at some loci [@adamskiSegregationDistortionHomozygous2014; @sayedSegregationDistortionDoubled2002], which

is one of the points of investigation of this PhD. The final type of experimental line which I will mention here is the backcross population, which as the name implies is the hybridization of the progeny of a biparental cross with one of the parents that produced it. Backcrosses have been utilized, amongst other things, in the study of segregation distortion to discover segregation distortion loci and to examine whether the effect of distortion is caused in the female or male gametal development cycle [@kumarIdentificationCharacterizationSegregation2007].

1.8 Computing environment and the R programming language

The large amount of data produced from high-density SNP arrays necessitate a bioinformatics approach to processing and analysing the data. Since such a large proportion of the work done during the PhD was done in a computational context, it might be illuminating to discuss the choices made in terms of computational environment and hardware used. To support computations, I utilized various high-performance servers, including a computing cluster located in the Life Sciences building at Bristol, and most importantly a powerful server dubbed "Wilkins", with ~750 Gb RAM, ~60 simultaneous threads and over 20TB total hard drive space. For the majority of the data analysis I chose the R programming language, which offers significant advantages for this application over any other language. R is an object-oriented, interpreted language, allowing inspection and manipulation of objects in a dynamic manner without compilation of the script being worked on. R also offers significant advantages over Python, also popular in the field of data science, as it provides native support for tabular data in the form of its dataframe class, and vectorized computation is a central feature of the language, whereas Python requires an additional libraries to implement these features. The consequence of this is that computations over entire columns of tabular data are very simple to implement, not even requiring a for loop, and are quick to execute, as the underlying vector calculation is performed in a basic linear algebra subprogram (BLAS) written in a compiled language, such as C++. Finally, R has a wealth of packages that extend its functionality significantly, such as ggplot for plotting, qtl for qtl analysis, ASMap

for genetic mapping and Bioconductor for handling genetic sequence data.

1.9 Segregation Distortion

1.9.1 Mendelian genetics

Mendel's law of segregation states that when two F1 individuals, both heterozygous for the same gene, are crossed together, the gametes produced by these individuals will bear the two alleles in a 1:1 ratio, and furthermore, the progeny of the cross will have a 1:2:1 ratio of homozygotes for the first allele, heterozygotes, and homozygotes for the second allele. This is a fundamental law of genetics which many students of biology will recognise in the form of a Punnett square. With further studies into genetics catalysed by the highly efficient research organism Drosophila, later followed by the revolution in molecular genetics enabling detailed study of crop species, it was revealed that there are many cases in which this law is violated. This violation is referred to as segregation distortion (SD), or alternatively as meiotic drive. SD is common in wheat and many other crops, although the exact mechanisms causing it have yet to be fully elucidated.

1.9.2 Causes of segregation distortion and relation to meiotic drive

SD can be caused by simple sampling error, or alternatively by the action of a selection pressure at some stage of the developmental cycle, such as during meiosis, gametogenesis, fertilization or zygotic development [@gadishDifferentialZygoticAbortion1987; @rickDifferentialZygoticLethality1963; @xuPrinciplesStatisticalGenomics2013]. One much studied cause of segregation distortion is meiotic drive, initially defined as the manipulation of meiosis [@bucklerMeioticDriveChromosomal1999; @sandlerMeioticDriveEvolutionary1957], and later broadened to include all of gametogenesis [@lindholmEcologyEvolutionaryDynamics2016; @zimmeringMechanismsMeioticDrive1970], by selfish genetic elements such that their own frequency of transmission is increased, or conversely, the frequency of transmission of their alleles is decreased. The literature is not in complete consensus regarding the definitions of meiotic drive

and segregation distortion, with some authors simply indicating that the two terms are synonymous [@kozielskaSegregationDistortionEvolution2010]. In this thesis I will limit the definition of meiotic drive to the developmental stages preceding or up to the completion of gametogenesis, which seems most in line with the original proposal of the term [@sandlerMeioticDriveEvolutionary1957], differentiating meiotic drive from gamete competition. This could explain the incongruity in the literature, as all cases of meiotic drive are also cases of segregation distortion, but all cases of segregation distortion have the potential to be, but are not necessarily, cases of meiotic drive, requiring further investigation to determine. In any case, meiotic drive is a fascinating form of intra-genomic conflict that is almost anthropogenic in quality, and as such has provoked almost 50 years of study: "Mendelian inheritance is a marvellous device for making evolution by natural selection an efficient process. The Mendelian system works with maximum efficiency only if it is scrupulously fair to all genes. It is in constant danger however, of being upset by genes that subvert the meiotic process to their own advantage" [@crowGenesThatViolate1979].

1.9.3 Canonical example of segregation distortion

The canonical example of meiotic drive is the segregation-distorter system in Drosophila [@larracuenteSelfishSegregationDistorter2012; @sandlerMeioticDriveNatural1959]. This system is comprised primarily of two loci, Segregation distorter (Sd) and Responder (Rsp), both located on autosomal chromosome 2, with several modifier loci that also play a role by either enhancing or reducing the intensity of the driving locus. Sd alleles act to distort segregation, whereas their wild-type alleles, Sd+, do not. In addition, Rsp has both sensitive (Rsp~s) and insensitive alleles (Rsp~i). Sd is usually paired with insensitive Rsp~i, whereas wildtype Sd+ is usually paired with a sensitive Rsp~s allele. In heterozygotes containing Sd/Sd+ and $\mathrm{Rsp_{i/Rsp}s}$, gametes with Rsp~s will fail to develop, resulting in close to 100% transmission of the segregation-distorter allele. In addition, individuals with Sd and Rsp~s on the same chromosome will self-destruct. Collectively, these observations indicate an interaction between the two

loci that ultimately produces segregation distortion, however, despite the length of time that this system has been studied, an exact mechanism for how the two loci interact is not entirely clear, with several competing hypotheses involving nuclear transport and small RNAs respectively [@larracuenteSelfishSegregationDistorter2012].

1.9.4 Example of segregation distortion in wheat

Could segregation distortion in wheat be caused by meiotic drive systems with similar mechanisms to the segregation-distorter system of Drosophila? Loegering and Sears [@loegeringDistortedInheritanceStemRust1963] found evidence of a meiotic drive system in wheat in the form of the pollen killer gene Ki. Chinese spring was crossed to a substitution line, Timstein-6B, the latter composed of Chinese spring with chromosome 6B substituted for the corresponding chromosome in variety Timstein, shown in previous experiments to contain a resistance gene for stem rust. F1 samples of this cross were then reciprocally backcrossed into Chinese Spring. Whilst the cross in which Timstein-6B was the female parent yielded close to a 1:1 ratio of resistant to susceptible plants, the reciprocal cross showed strong distortion in favour of susceptible plants. Further investigation revealed that many of the microspores resulting from this cross began to degenerate early in development, suggesting that a meiotic drive gene, Ki, was highly linked to the gene conferring stem rust resistance, causing the extreme distortion in plants heterozygous for this gene. Although discovered over 50 years ago, this remains one of the best examples of meiotic drive operating in wheat, and illustrates how meiotic drive and segregation distortion can be detrimental to agronomic aims, in this case reducing the number of progeny with resistance to stem rust.

1.9.5 Segregation distortion and hybrid incompatibility

Segregation distortion has also been shown to play a role in hybrid incompatibility between species or subspecies [@johnsonHybridIncompatibilityGenes2010]. The Bateson-Dobzhansky-Muller model [@orrDobzhanskyBatesonGenetics1996] posits that negative

epistatic interactions between genes in hybrids act to reproductively isolate different species, and are the result of the gradual accumulation of mutations along different evolutionary trajectories in populations. One such system is the killer-protector system of rice [@yangKillerProtectorSystemRegulates2012], in which hybrids of indica and japonica subspecies are often sterile, or show strong segregation distortion. This is due to a gene that induces endoplasmic reticulum stress in indica, which is rescued by another gene. The rescuing gene is not present in japonica varieties, and as such the endoplasmic reticulum stress ultimately leads to premature programmed cell death and embryo-sac abortion.

1.9.6 Disparity of in methods of detection between studies

The increased genotyping capacity in recent years, driven by developments in technology, has led to an increase in the number of loci that supposedly exhibit segregation distortion, with examples found in cotton [@daiIdentificationCharacterizationSegregation2017], maize [@luChromosomalRegionsAssociated2002; @wangHighSegregationDistortion2012], potato [@manrique], chickpea [@castroSegregationDistortionLocus2011], barley [@liConstructionHighdensityComposite2010] and wheat [@allenCharacterizationWheatBreeders2016; @gardnerHighlyRecombinedHigh2016; @wingenWheatLandraceGenome2017]. There is however a disparity in the statistical methods used to detect segregation distortion, with some authors using a simple chi-square test with a alpha threshold of 0.05, others using multiple alpha thresholds, others using a false-discovery rate multiple-testing correction procedure, and some using the even stricter Bonferroni correction. One of the aims of this thesis will be to identify which of these methods is the most accurate for the detection of segregation distortion when many markers are being tested simultaneously, as is the case with modern high-density genotyping data from e.g. arrays or genotyping by sequencing methods.

1.9.7 Distinction between Bonferroni and FDR multiple testing correction procedures

Statistics is a method of classifying certain experimental results as "significant" if they pass a specified p-value threshold, which by convention is set to 0.05, although this choice was an arbitrary one made by the father of statistics, Ronald Fisher. This value indicates that a result could be significant by chance, e.g. through sampling error, and the probability of this false-positive result is 1/20. When multiple statistical tests are performed simultaneously, the probability that one of these tests will have a false-positive result increases by a factor, the number of tests being performed. The Bonferroni correction [@bonferroniCalcoloAssicurazioniSu1935] is a widely used procedure that accounts for the increasing probability of a false-positive result with increasing number of tests, otherwise known as the family-wise error rate, by dividing the significance threshold a, by the number of tests being performed. Although a plausible strategy for small datasets, the Bonferroni has its downsides; it is highly conservative when the number of hypotheses being tested is large, and has been criticised for its low statistical power in these cases [@nakagawaFarewellBonferroniProblems2004]. The advent of large biological datasets, initially in the form of microarray data to assess gene expression, demanded a new multiple testing correction procedure, as it would often be the case that thousands of hypotheses of differential gene expression, one for each gene under analysis, were being tested simultaneously [@storeyStatisticalSignificanceGenomewide2003]. The Benjamini-Hochberg false-discovery rate (FDR) correction [@benjaminiControllingFalseDiscovery1995] rose to meet this demand, with 63,089 citations on Google Scholar at the time of writing, by reconceptualizing the error to be corrected for away from the family-wise error rate, i.e. the probability of a false-positive, to instead focus on the FDR, which is the proportion of significant results that are incorrectly categorised as significant. It offers a test in which the correction scales to the number of hypotheses - for small numbers it offers a correction similar to the Bonferroni correction, whilst for large numbers it becomes more lenient. The FDR correction can be thought of geometrically as a plot of ranked p-values from smallest to largest. A line

is then drawn through the origin with slope a/m (m being the number of hypotheses tested), and all p-values beneath the line are retained as significant. Whether this is the most appropriate test for segregation distortion remains to be seen, and I will test this hypothesis in a later chapter.

1.10 Phylogenetic inference

Like many fields in biology, increasing availability of sequencing and genotyping data, as well as increasing computational power in recent years, have made viable the field of molecular phylogenetics - the study of evolutionary history through DNA molecules themselves rather than phenotypic characteristics. In this thesis I will utilize phylogenetic methods to examine the evolutionary history of wheat, particularly the landraces: locally adapted cultivars from diverse regions around the globe. As with the ordering of molecular markers in a genetic map, the number of possible rooted topologies for a bifurcating labelled tree rises dramatically with the number of taxa, with (2n - 3)!! possible trees in a tree with n taxa [@felsensteinNumberEvolutionaryTrees1978]. It is therefore impossible to examine every topology in a reasonable amount of time, and so algorithms must be used. Two principle methods of phylogenetic inference are generally used, which are maximum likelihood and Bayesian methods.

To construct a phylogeny, it is first necessary to estimate the amount of evolution that has occurred between two sequences (e.g. DNA or amino acid sequences). A naïve assumption would be to simply calculate the number of differences in homologous positions between the two sequences, and indeed this is the foundation of simple phylogenetic methods such as neighbour-joining [@saitouNeighborjoiningMethodNew1987]. However, this calculation ignores the possibility of multiple nucleotide substitutions, or multiple hits, at the same sequence site over time, resulting in a number of hidden mutations, and thus underestimate the total difference between the two sequences (Yang, 2014). To account for this, models of sequence evolution are commonly used in maximum likelihood and Bayesian inference, which take into account empirically determined properties of DNA evolution - for example the fact that transversion mutations occur

less frequently than transition mutations. Maximum likelihood methods calculate the likelihood for any particular tree, which is the probability of the data given then tree and the model of sequence evolution or P(D | T, M). The likelihood can then be calculated for many trees using an algorithm, and the tree that has the maximum likelihood is reported as the most likely phylogeny for those particular sequences and taxa.

1.11 QTL Analysis

Quantitative trait loci analysis, unlike qualitative traits, focusses on phenotypic traits that are continuously distributed in a population, and are underpinned by the combined action of many different genes, as well as the effects of environment. Many important agronomic and physiological traits in wheat are influenced by QTL, such as plant height, harvest index, thousand grain weight [@tshikundeAgronomicPhysiologicalTraits2019], recombination frequency, and recombination distribution [@jordanGeneticArchitectureGenomewide2018]; the latter two will be investigated in this thesis. With high-density molecular marker data and genetic mapping of the genome, it is often the case that some of the molecular markers will be linked to these underlying genes, visible in the data as a correlation between genotype and phenotype. In its simplest form, the identification of a QTL significantly associated with a trait of interest can be done through marker regression, which involves performing an ANOVA at each marker, using genotype categories as predictors and the trait of interest as a response. This method is improved upon by interval mapping, which takes into account missing genotype data at markers via maximum likelihood estimation. Furthermore, the standard interval mapping method is made more computationally efficient by the Haley-Knott regression, which provides a fast and accurate approximation of the results [@bromanGuideQTLMapping2009].

1.12 Recombination

For many millennia, one of the fundamental processes in agriculture has been the crossing of varieties with different phenotypic traits to produce new, hybrid varieties with a mixture of these traits. The molecular process underlying this is meiotic recombination, where chromosomes are shuffled to create new, hybrid chromosomes containing a mixture of alleles from both parents. In wheat and many other important grasses, the distribution of recombination events is limited to the distal ends of the chromosomes, meaning genes surrounding the centromere and in the pericentromeric regions stay in linkage disequilibrium and are not mixed during recombination. One of the foci of this thesis is to test whether environmental temperature, a factor known to influence recombination, alters the distribution and frequency of recombination events in wheat in a way that could be useful to breeders. Does an increased temperature act to induce recombination events in the pericentromeric regions, breaking up genes that were previously inaccessible to manipulation by breeders?

2 The Watkins landraces: their evolutionary history and congruity of genotyping data between platforms

2.1 Introduction

Archaeological evidence indicates that wheat was domesticated around 9000 years ago in the fertile crescent. This event involved the hybridization of a tetraploid progenitor containing what are now referred to as the A and B subgenomes of modern bread wheat with a wild diploid grass related to the modern-day *Aegilops tauschii*, now comprising the D genome of bread wheat.

Before the inception of modern commercial breeding practices, wheat as a species was composed of many locally adapted cultivars known as landraces [@jaradatWheatLandracesGenetic2012]. Recent research has focussed on the Watkins collection, which consists of landrace cultivars from a broad range of countries, originally collected in the 1930s [@wingenEstablishingWatkinsLandrace2014]. Modern elite bread wheat varieties suffer from a lack of genetic diversity, which makes them susceptible to evolving biotic stresses such as pathogenic fungi (e.g. karnal bunt, *Tilletia indica* [@reifWheatGeneticDiversity2005]), as well as changing environmental conditions such as climate. Research has shown that the Watkins collection is more genetically diverse compared to modern elite varieties [@winfieldHighDensityGenotyping2017; @wingenEstablishingWatkinsLandrace2014] and could therefore serve as a valuable source of novel alleles for wheat breeding programmes.

Much of the current focus within the wheat community is to improve elite wheats through incorporation of existing genetic variation. The Wheat Improvement Strategic Programme (WISP) [@mooreStrategicPrebreedingWheat2015] aims to utilize three primary sources of variation: landraces, synthetic wheats and introgression from wild relatives of wheat. There has however been little investigation to date into the rate

at which wheat accumulates novel polymorphisms. This is an interesting question both from a historical perspective as well as in future projections of wheat evolution – if the current pool of genetic diversity, encompassing landraces, synthetics and wild relative introgressions is exhausted, what length of time would it take for new beneficial mutations to accumulate in global germplasm? To answer this question, I first intend to use previously generated exome-capture data of lines from the Watkins collection [@gardinerHiddenVariationPolyploid2018] to determine the time to most recent common ancestor (TMRCA) of these lines. This will be done using a molecular clock analysis with calibration stemming from the known divergence point of hexaploid bread wheat *Triticum aestivum* and the ancestor of the D genome of wheat, *Aegilops tauschii*, which was around 9000 bp. After obtaining the TMRCA estimate, it will then be possible to form an estimate of how many novel alleles have accumulated in each variety over that period of time.

The concept of a molecular clock was first suggested by Zuckerkandl and Pauling [@zuckerkandlMoleculesDocumentsEvolutionary1965], who suggested that constancy in the rate of amino acid substitution between haemoglobin proteins could provide a mechanism with which to estimate the time of divergence between species. Knowing the rate of mutation between two molecules and the number of differences between those molecules, it is possible to calculate the time at which they diverged from each other. In mathematical form, this is represented by the equation $T = rL/2$, where T represents time, r represents the rate of mutation and L represents the combined branch lengths of the phylogeny leading to the common ancestor of both molecules. This concept was bolstered by the suggestion of the neutral theory of evolution, which posits that the majority of differences in nucleotide sequences are in selectively neutral regions [@kimuraEvolutionaryRateMolecular1968]. Consequently, nucleotide substitution in these regions might operate at a constant clock-like rate, rather than having a rate that shifts over time due to selection.

In addition to the molecular clock analysis, I will also be using the genotyping data generated for the Watkins lines to investigate congruity between data sources (i.e. be-

tween array and exome capture). There has been a recent trend towards the use of exome capture data within the wheat research community [@gardinerAnalysisRecombinationLandscape2019; @olohanModifiedSequenceCapture2018]. Exome capture has the potential to provide information on much more sequence variation than array genotyping data at the cost of speed and throughput volume. It would be of interest to compare these two datatypes to examine whether the increased resolution of exome capture data significantly effects downstream analyses – does it give increased insight that compensates for the increased cost? Whilst [@gardinerHiddenVariationPolyploid2018] made some comparisons of their dataset to the data of [@winfieldHighDensityGenotyping2017], namely noting that European accessions formed separate clusters to Asian and Middle Eastern accessions, they did not use the same method to cluster their SNP data, opting for a hierarchical clustering approach rather than a STRUCTURE-based analysis. STRUCTURE [@porras-hurtadoOverviewSTRUCTUREApplications2013] is a population genetics software package which aims to infer population structure using a Bayesian clustering approach in conjunction with Markov Chain Monte Carlo estimation. This differs to hierarchical clustering in that the posterior probabilities for a range of K values, or the number of clusters, must be evaluated. In contrast, the hierarchical clustering approach initially assigns each individual to its own cluster, then proceeds to join closely related clusters together. The papers also differ in that only one of them performs a phylogenetic analysis [@gardinerHiddenVariationPolyploid2018], only one of them performs a principle components analysis [@winfieldHighDensityGenotyping2017], and the sample sizes of Watkins lines differ in each. Here I perform a more direct and detailed comparison of array and exome capture datasets in both population genetics and phylogenetic contexts, ensuring the use of the same methodology in each case.

2.2 Methods

2.2.1 Phylogenetic analysis

Alignment of paired-end sequence reads and identification of SNPs in exome-capture data

Exome capture data for 104 Watkins lines was obtained from the Grassroots Genomics repository [@gardinerHiddenVariationPolyploid2018]. Non-bisulfite-treated paired-end reads were mapped to the IWGSC v1.0 genome assembly of wheat using BWA MEM. Processing of mapping results was performed with Samtools. Reads were filtered so that only mapped reads and unique reads, defined as reads with a MAPQ value higher than 10, were used. Duplicate reads were removed from the alignment. VCF files were generated using samtools mpileup to calculate genotype likelihoods in conjunction with bcftools call for SNP calling. Only homozygous SNPs with a VCF QUAL value higher than 20 and at least 20x coverage across all varieties were used.

Obtaining homologous sequences from *A. tauschii*

After VCF files were generated for each Watkins line, it was then necessary to determine the genotype values of *Aegilops tauschii* at orthologous positions to the Watkins SNPs. Mummer was initially used in an attempt to align the entirety of the A. tauschii genome sequence to the D genome of the IWGSC assembly, but this approach was found to be prohibitively slow due to the size of the sequences involved. In addition, I decided to forgo the use of other commonly used tools for determining orthology such as OrthoMCL, as these are typically designed for the identification of families of orthologues between many species, whereas here we were only dealing with two. My custom pipeline began with the extraction of D-genome subsequences of the IWGSC assembly based on positions that had 20x coverage across all Watkins varieties. A BLAST search of these subsequences was then performed against the A. tauschii genome assembly. BLAST does not return full length alignments of query against target, but instead returns a series of local alignments called high-scoring segment pairs (HSPs). The results of the initial BLAST search was therefore used to identify regions in which the homologues was most likely located. HSPs less than 7000 bp apart (a distance determined empirically) were grouped together and their average bitscore was calculated. The group of

HSPs with the highest bitscore was determined to be the homologous sequence, and both the lowest and highest base positions of HSPs within the group were used as a coordinate range for sequence extraction from the genome.

Supermatrix construction and inference of phylogeny using BEAST

These extracted *A. tauschii* sequences were then aligned to their corresponding IWGSC query sequences using MUSCLE. The genotypes and positions of SNPs were determined using R. Unknown genotypes (represented as "N") and insertions (represented as "-") were removed from both Chinese Spring and *A. tauschii* sequences. Alignments with more than a 40% difference were excluded as these were likely to be erroneous alignments between sequences that were not truly homologous. A supermatrix containing all of the multiple sequence alignments was then generated containing all 104 Watkins varieties as well as Chinese Spring and *A. tauschii*, which would be used as input to phylogenetic inference software. It is well known that phylogenies inferred using only SNPs without correction for acquisition bias can lead to overestimation of the divergence between taxa [@leacheShortTreeLong2015]; we therefore included both SNPs and invariant loci in the supermatrix. Insertions and deletions (indels) were not included in any of the sequences as the differences in source between sequences, namely the full genomic sequence for *A. tauschii* and exome capture data for Watkins varieties may have led to a bias in the length of indels towards *A. tauschii*. BEAST was used to generate the phylogeny and estimate divergence times of each node. A strict clock model was used with a calibration of 9,000 years for the divergence between *A. tauschii* and the rest of the varieties was used. More specifically, the prior used for this calibration was a normal distribution with a mean of 0.009 (measured in millions of years) and a standard deviation of 0.0001. This date was based on the occurrence of hexaploid free-threshing wheats at Cafer Höyük that have been radiocarbon-dated to ~8700 years ago [@nesbittWheatEvolutionIntegrating2001].

Bootstrap analysis to assess reliability of inferred subclades

Whilst the phylogeny generated with BEAST allowed the dating of particular nodes, it was also important to assess the reliability of the tree topology, as clades with low levels of support would affect the inference of dates of divergence. To do this, a maximum likelihood phylogeny was also generated with IQTREE using a HKY+F model of sequence evolution and 1000 bootstrap trees. Nodes within this tree that had low bootstrap support values could then be disregarded from the dating analysis, whilst nodes with higher bootstrap support would indicate that dating was more reliable.

Functional characterization of SNPs

Functional characterization of SNPs, such as whether they would result in a change in amino acid sequence (missense mutations) or if they were silent with regards to the amino acid sequence, was performed with Ensembl Variant Effect Predictor [@mclarenEnsemblVariantEffect2016]. The predicted effect of missense mutations on protein function was evaluated using SIFT [@vaserSIFTMissensePredictions2016]. This assigns each mutation a score from 0 to 1, based on how conserved the position is in homologous sequences, with lower scores representing mutations that are more likely to be deleterious to the organism (i.e. positions that are highly conserved in most homologues).

Estimation of the rate of novel polymorphisms in wheat

To estimate the rate at which novel polymorphisms occur during the evolution of wheat, it was first necessary to calculate the TMRCA for each Watkins variety and Chinese Spring. This was done by calculating the cophenetic distance between each variety and CS using the cophenetic.phylo() function of the APE package in R. This distance was then divided by two to correct for the inclusion of branches leading to both the Watkins variety and Chinese Spring in the distance value. This calculation returns the period of time over which the observed mutations, whether missense, synonymous or intronic, were estimated to have occurred. The rate of polymorphism can then be calculated by dividing the number of observed polymorphisms by the time over which they occurred, then dividing this by the breadth of sequence with the minimum coverage threshold

(20x) in the exome capture dataset to give the number of polymorphisms per year per bp that occurred in a particular Watkins variety.

2.2.2 Comparison of Exome and Array data

PCO Analysis

Array-based genotyping data of the Watkins lines from the Axiom 35k wheat breeder's array [@allenCharacterizationWheatBreeders2016] was obtained from CerealsDB [@wilkinsonCerealsDBExpansionResources2016]. To compare the effect of exome vs array data on population genetic analysis, two methods were used, PCO and STRUCTURE, as in [@winfieldHighDensityGenotyping2017]. For the PCO analysis, pairwise genetic dissimilarity was calculated between all combinations of Watkins varieties by dividing the total number of genotypes in common between two varieties by the total number of genotypes. Genotypes with missing values were not included in this calculation. This value was then subtracted from 1 to give the dissimilarity score for each pair. Principle coordinates were calculated using the cmdscale function in R.

STRUCTURE analysis

Further to this, a STRUCTURE analysis was also performed, giving information on the number of populations (K). STRUCTURE was automated using StrAuto [@chhatreStrAutoAutomationParallelization2017], which parallelizes STRUCTURE, running each iteration of K on a separate core for much faster computation. In addition, StrAuto runs StructureHarvester [@earlSTRUCTUREHARVESTERWebsite2012] as part of the pipeline, which calculates the uppermost bound for K using the Evanno method [@evannoDetectingNumberClusters2005]. STRUCTURE was run with each individual represented as a diploid to incorporate heterozygosity in the analysis. Values of K ranging from 1 to 10 were tested, with 5 repeats for each value. The ancestry model used was admixture, which assumes that each individual inherits fractions of its genetic composition from a combination of the K populations. A burnin length of

10000 was used for the Markov chain, and the Markov chain was then run for 10000 iterations. CLUMPAK [@kopelmanClumpakProgramIdentifying2015] was used to align STRUCTURE runs across multiple values of K, compensating for label-switching.

2.3 Results

2.3.1 Phylogenetic analysis

Coverage of exome-capture data

15.30 Mbp of the Chinese Spring sequence had at least 20x coverage in all Watkins samples. This equates to 1.28 % of the genomic sequences of the high-confidence gene set included in the IWGSC RefSeq v1.0 assembly, which is 1196.52 Mb in size. To assess the quality of the custom pipeline used to determine homologues between Chinese Spring and *Aegilops tauschii*, the number of mismatching sites between all sequences was calculated — 4% — indicating that the alignments were of high quality. The total number of SNPs found between Chinese Spring and at least one of the other varieties, including *Aegilops tauschii*, was 181043. The mean (± s.d.) number of SNPs between lines in the Watkins collection and Chinese Spring was 5962.06 ± 887.32, whereas the number of SNPs between *A. tauschii* and Chinese Spring was 131042.

Estimation of the TMRCA of Watkins lines

The time to the most recent common ancestor (TMRCA) for the clade containing the wheat varieties, including both Watkins lines and Chinese Spring, was estimated to be 859 years (figure 2.1). The smallest TMRCA for any node in the tree was 237 years (varieties USSR_1990753 and Syria_1190045). This is congruent with our expectation that the dates of divergence should be older than 1930, which is when the Watkins lines were originally collected [@wingenEstablishingWatkinsLandrace2014]. In addition to the general rate of evolution between Watkins varieties, we were also interested in examining the mutations that could potentially effect protein function, and therefore influence phenotype. We will limit this to the analysis of the D genome, as this is the

genome for which we have TMRCA estimates, allowing the association of total number of mutations to time. Of the 16867 genes in the D genome that were at least partially covered by the exome capture data, 6942 contained missense mutations in at least one of the Watkins varieties.

The functional analysis of SNPs between all Watkins varieties and Chinese Spring revealed a total of 9385 unique missense mutations. The mean (± s.d.) number of missense mutations per variety was 1747.89 ± 240.7. The total number of missense mutations shared by all Watkins varieties was 32. In a pairwise comparison of shared missense mutations between all Watkins varieties, the mean ± s.d. number shared was 640.75 ± 199.53. The maximum number of shared missense mutations between any two Watkins varieties was 1485, between Watkins 1190044 from Morocco and 1190045 from Syria. Interestingly, these varieties were located in distinct clades on the maximum likelihood phylogenetic tree, indicating that many of these shared polymorphisms were the result of convergent evolution. This is reinforced by the fact that the sister sample to the Watkins 1190045 sample was Watkins 1190753 from the USSR, with a 100% bootstrap support value in the maximum likelihood tree, and the number of shared missense mutations between Watkins 1190753 and 1190044 was only 434. The similar climate of the countries where these samples originated from (Morocco and Syria) indicates that some of the mutations could be involved in heat or drought tolerance. The minimum number of shared missense mutations between any two Watkins varieties was 328, from Watkins 1190040 from France and Watkins 1190700 from China. These two varieties were located in distant clades on the phylogeny.

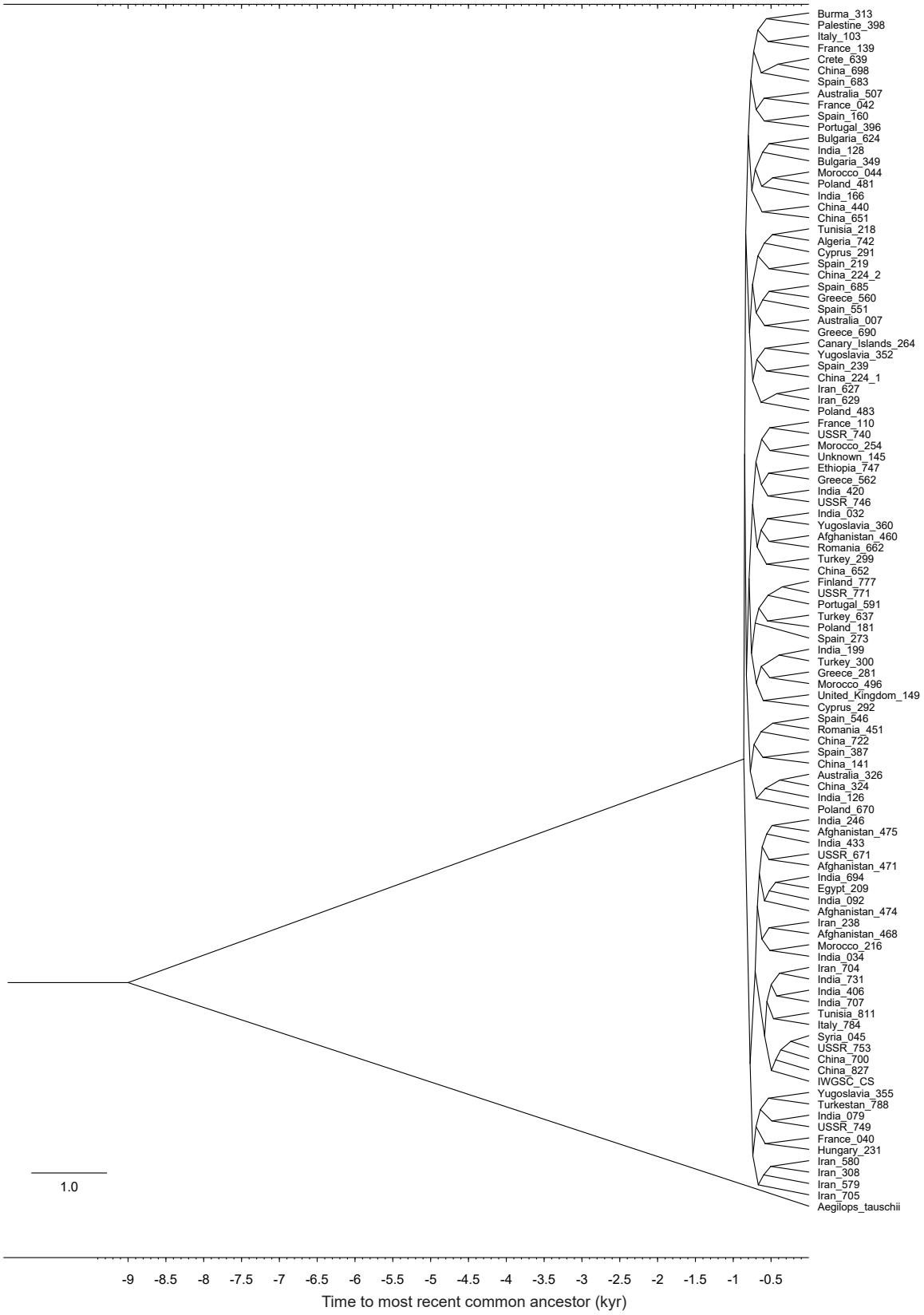

Burma_313
Palestine_398
Italy_103
France_139
Crete_639
China_698
Spain_683
Australia_507
France_042
Spain_160
Portugal_396
Bulgaria_624
India_128
Bulgaria_349
Morocco_044
Poland_481
India_166
China_440
China_651
Tunisia_218
Algeria_742
Cyprus_291
Spain_219
China_224_2
Spain_685
Greece_560
Spain_551
Australia_007
Greece_690
Canary_Islands_264
Yugoslavia_352
Spain_239
China_224_1
Iran_627
Iran_629
Poland_483
France_110
USSR_740
Morocco_254
Unknown_145
Ethiopia_747
Greece_562
India_420
USSR_746
India_032
Yugoslavia_360
Afghanistan_460
Romania_662
Turkey_299
China_652
Finland_777
USSR_771
Portugal_591
Turkey_637
Poland_181
Spain_273
India_199
Turkey_300
Greece_281
Morocco_496
United_Kingdom_149
Cyprus_292
Spain_546
Romania_451
China_722
Spain_387
China_141
Australia_326
China_324
India_126
Poland_670
India_246
Afghanistan_475
India_433
USSR_671
Afghanistan_471
India_694
Egypt_209
India_092
Afghanistan_474
Iran_238
Afghanistan_468
Morocco_216
India_034
Iran_704
India_731
India_406
India_707
Tunisia_811
Italy_784
Syria_045
USSR_753
China_700
China_827
IWGSC_CS
Yugoslavia_355
Turkestan_788
India_079
USSR_749
France_040
Hungary_231
Iran_580
Iran_308
Iran_579
Iran_705
Aegilops_tauschii
1.0
-9
-8.5
-8
-7.5
-7
-6.5
-6
-5.5
-5
-4.5
-4
-3.5
-3
-2.5
-2
-1.5
-1
-0.5
Time to most recent common ancestor (kyr)

Figure 2.1 Phylogeny of Watkins lines showing time until most recent common ancestor for each node. The phylogeny was dated using a strict molecular clock, which was calibrated using the estimated divergence of Aegilops tauschii and hexaploid wheat of 9000 years before present. Node labels indicate the location from which each line was collected, as well as the Watkins number of each line in the format Location_Number. Also included is Chinese Spring from the IWGSC reference sequence (labelled IWGSC_CS). Scale bar indicates number of base substitutions per site.

Functional SNP analysis

The majority of the missense mutations were categorised as deleterious by SIFT with mean ± s.d. SIFT score for all unique missense mutations among all varieties of 0.35 ± 0.36. This trend was consistent when examining missense mutations within varieties, with the minimum mean SIFT score among all the mutations within a single variety of 0.41 in 1190460, and the maximum of 0.46 in 1190224.

Estimation of rate of novel polymorphism accumulation in wheat

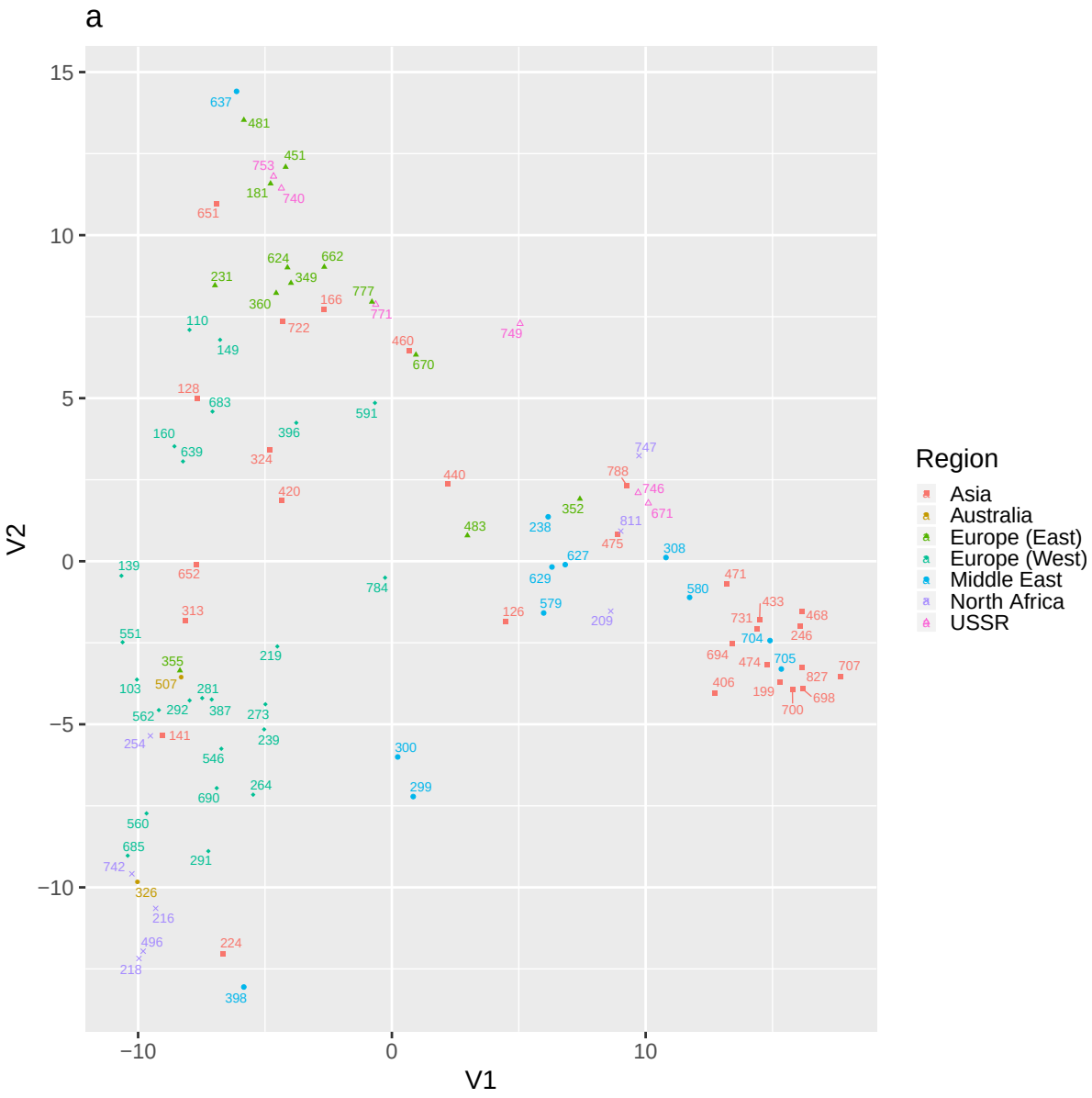
The rate of novel polymorphisms that could affect protein function was estimated by dividing the number of missense mutations between each Watkins variety and Chinese Spring by the respective TMRCA values for each (see methods). The mean ± s.d. value of this calculation among all Watkins varieties was 2.19 ± 0.3. This can then be divided by the breadth of the exome that was successfully sequenced in the exome-capture data (3.78 Mb), giving an estimated rate of 0.58 novel missense mutations per Mb per year. Alternatively, if the TMRCA value from the molecular clock analysis is an underestimation of the true value, we can still provide an estimate of the range of values that the number of missense mutations might take by using a range of values for the TMRCA of the Watkins lines based on what is known about the global dissemination of wheat. The minimum value for the TMRCA can be assumed to be 8000 years, since the inception of hexaploid wheat occurred roughly 9000 years ago. Performing the calculation with this new TMRCA estimate gives 1747.89 / 8000 / 3.78 = 0.06 novel missense mutations per Mb per year. On the other hand, using a younger estimate for

the TMRCA (nonetheless older than the molecular clock estimate) of 4000 years gives a value of 1747.89 / 4000 / 3.78 = 0.12 missense mutations per Mb per year.

2.3.2 Comparison of Exome and Array data

PCO analysis

The PCO plots revealed remarkable similarity between the exome capture and array datasets (figure 2.2). Both plots show the same broad pattern of clusters, with the Asian and Middle Eastern varieties separated from the European, Australian, USSR-originating varieties along the x-axis, and the y-axis separating western European and North African lines away from Eastern European lines. More specific patterns are also preserved between datasets, such as the positioning of varieties 300 and 299, which in both plots occupy their own space in between Middle Eastern and Western European clusters around 0 on the x-axis, as well as varieties 440 and 749, which lie in between Asian and Eastern European clusters in both plots. There are some varieties which differ in positioning between plots, such the variety 753 from the USSR, which in the array data clusters together with Eastern European lines at the top of the y-axis, whereas in the exome data clusters with Asian lines along the far right of the x-axis, as well as variety 326 from Australia, which in the array data clusters with Western European lines at the bottom of the y-axis, whilst in exome capture data clusters with Eastern European lines at the top of the y-axis.


a
15
10
5
0
−5
−10
V2
−10
0
10
V1
Region
Asia
Australia
Europe (East)
Europe (West)
Middle East
North Africa
USSR

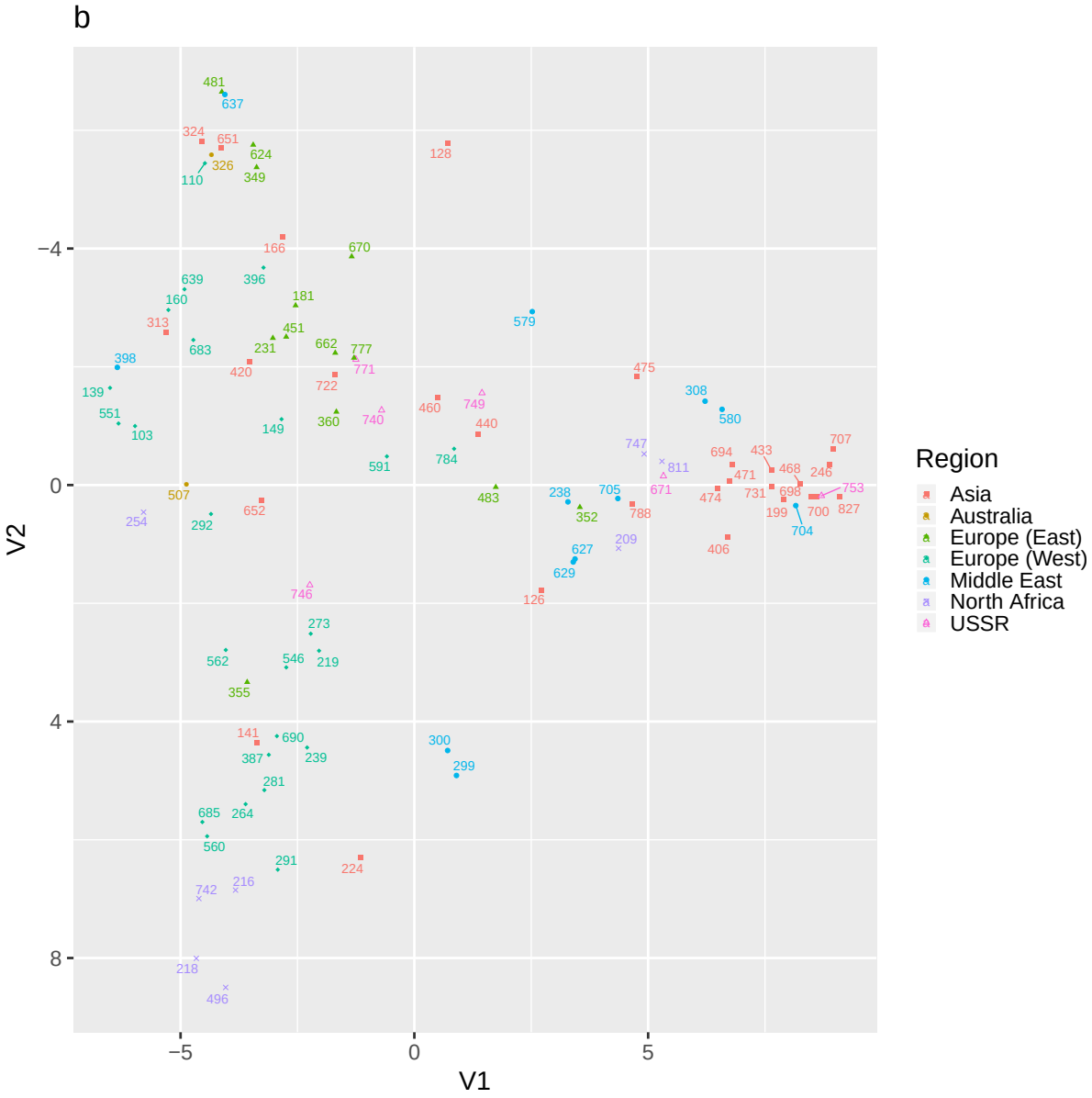


Figure 2.2 PCO plots comparing array (a) to exome capture data (b). Individual points are labelled with their Watkins variety numeric identifiers. The y-axis on the right plot has been inverted for visual ease of comparison. Points are coloured and shaped by region. Represented are varieties from Asia, Australia, Europe (east), Europe (west), Middle East, North Africa and the USSR. The plots are remarkably similar considering the use of datasets from different labs using different methods to generate them. For example, in each plot, varieties 300 and 299 show the same configuration in relation to the remainder of the varieties, emerging around 0 on the x-axis.

Pairwise differences by variety in genotype values between data types

Pairwise differences in genotype values between varieties were highly correlated between

data types, as shown in figure 2.3 (Pearson test, t = 123.94, df = 4369, p < 2-15). A linear regression of exome pairwise differences as a function of array pairwise differences revealed that 77.86 % of variation in the exome data was explained by the array data (R^2 = 0.7786).

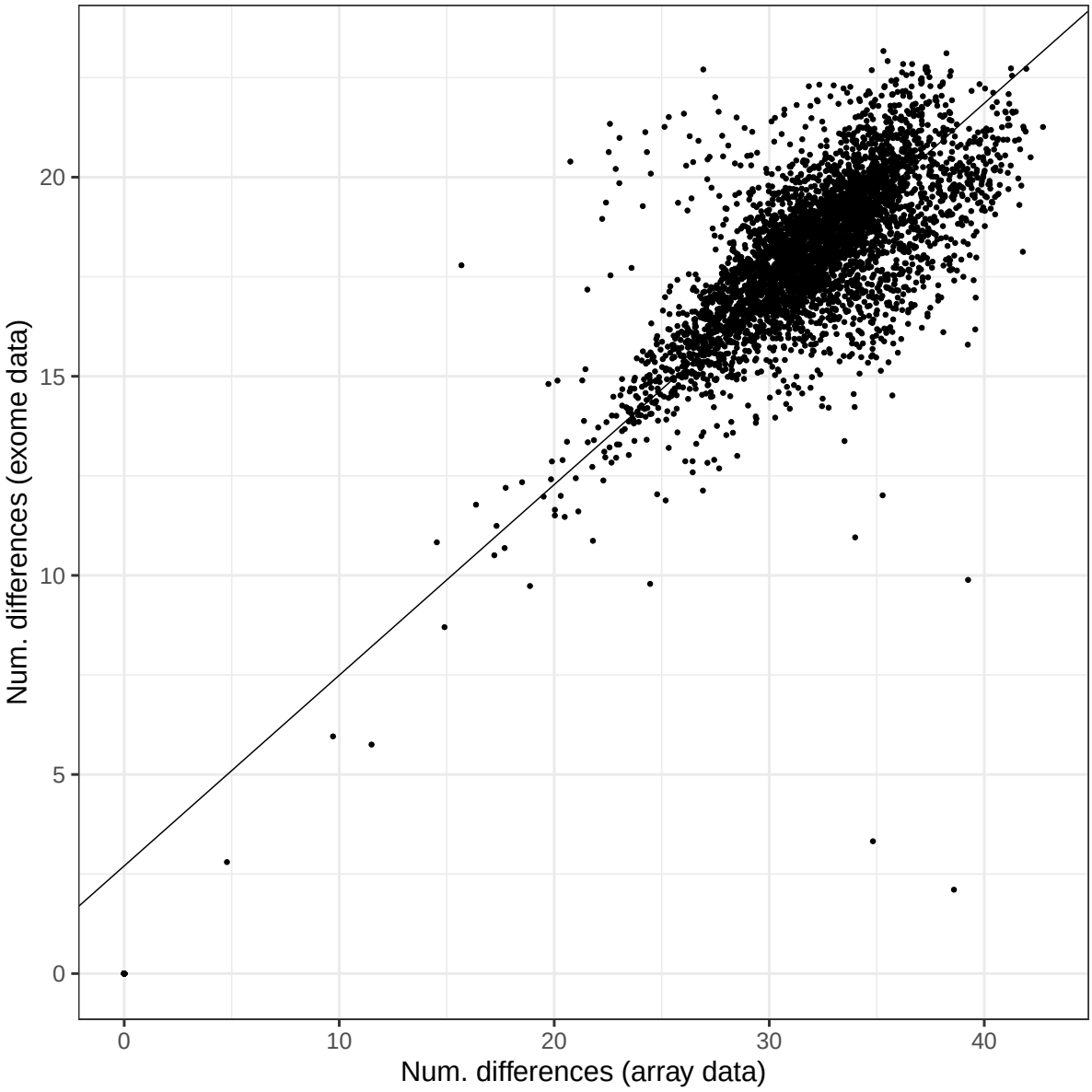


Figure 2.3 Scatterplot showing pairwise distances between varieties between exome capture data and array data. Also shown is a regression line.

Comparison of probe distribution

Probe distribution was also highly similar between the array and exome capture data (figure 2.4). Both datasets contained peaks in numbers of probes at the start of chromosome 1A. Although distribution was similar, exome capture data contained many more probes than the array data (168440 and 35144 respectively).

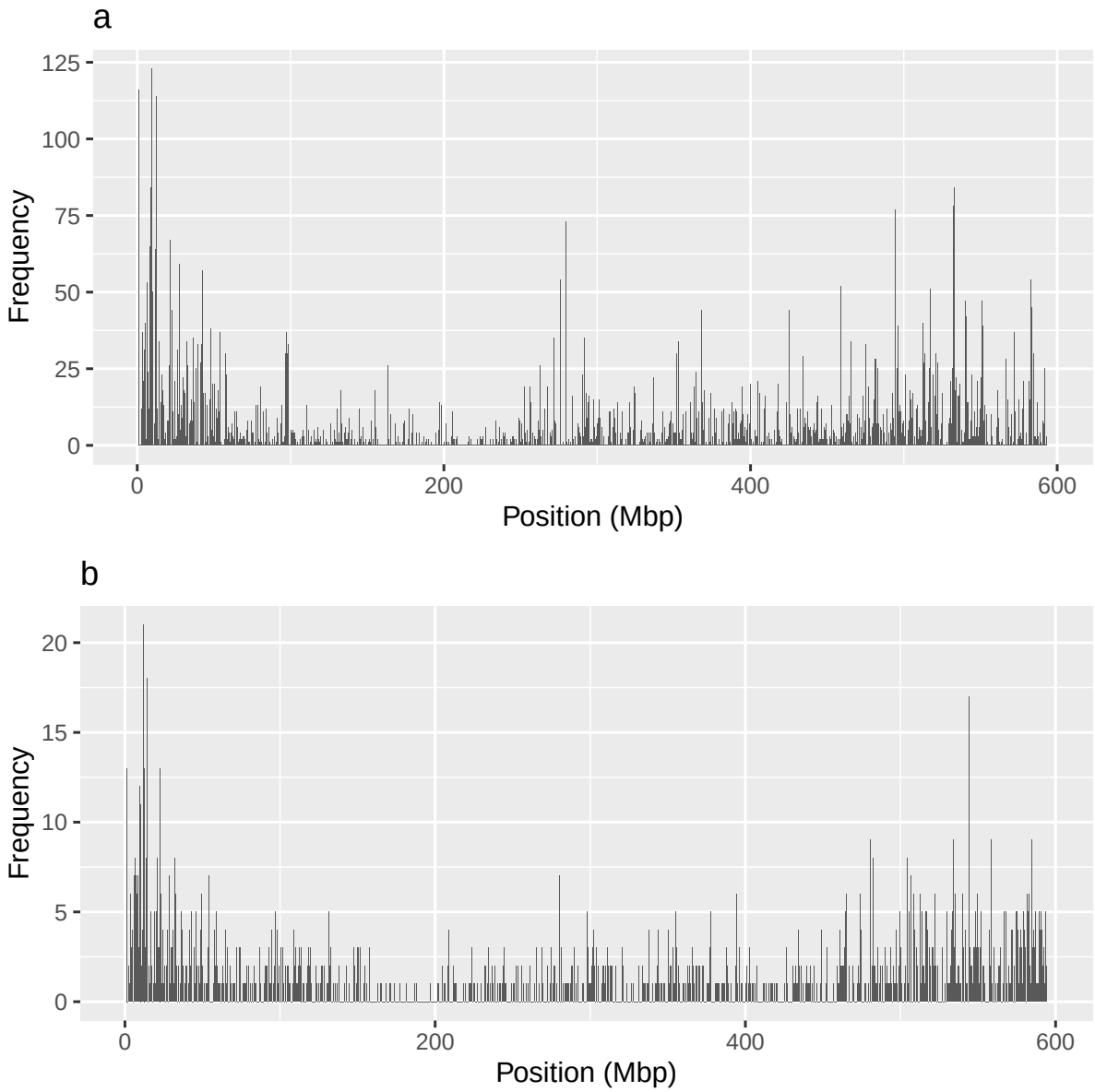


Figure 2.4 Comparison of probe distribution for chromosome 1A for exome capture data (top) and array data (bottom).

STRUCTURE analysis

The STRUCTURE analysis also revealed a high degree of similarity between exome capture and array datasets, as shown in figure 2.5. In both datasets, around half of the Asian varieties have high degree of membership to the third cluster (K3), as do many of the Middle Eastern, North African and USSR-originating lines. The other two clusters, K1 and K2, largely correspond to Eastern and Western European lines

respectively, but also contain varieties from the other regions. Differences between exome capture and array datasets include Watkins line 753, which in array data has full membership to K1, whereas in exome capture data has full membership to K3, as well as the Australian Watkins line 326, which in array data mostly corresponds to K2, whereas in exome capture data mostly corresponds to K1.These two lines are the same lines that were highlighted as being different between datasets in the PCO analysis, which indicates that the STRUCTURE and PCO methods are largely congruent in their results. The Evanno method [@evannoDetectingNumberClusters2005] for determining the uppermost bound for K in the data revealed that for the array data, the upper bound was 7, whilst for the exome data, the upper bound was 5.

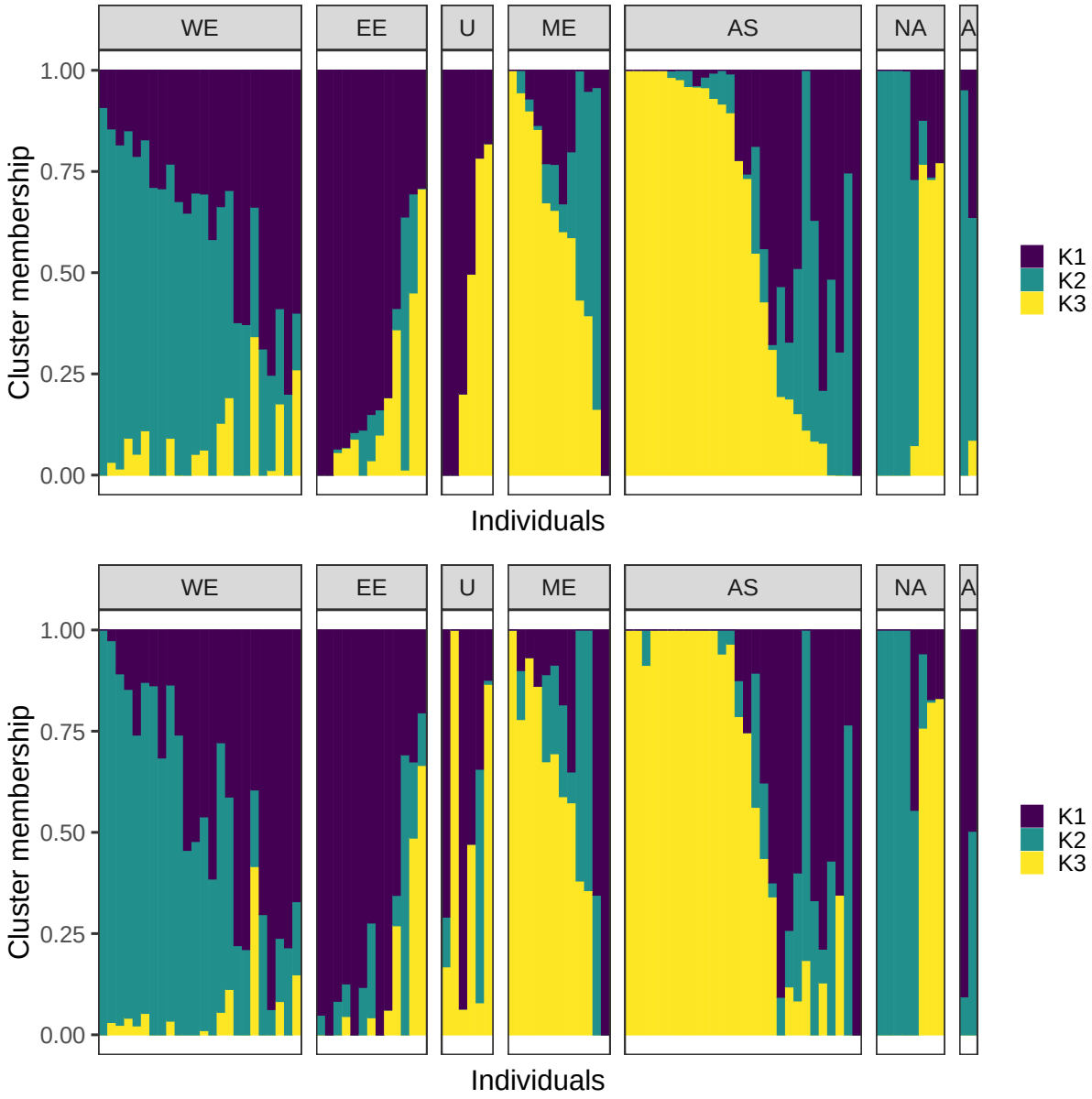


Figure 2.5 Structure plots for K = 3, showing array (top panel) and exome capture (bottom panel) data. WE = Western Europe, EE = Eastern Europe, U = USSR, ME = Middle East, AS = Asia, NA = North Africa, A = Australia.

2.4 Discussion

The analyses in this chapter provide novel results covering the evolutionary history of the Watkins landraces in terms of their divergence times. In addition, a direct comparison of two distinct genotyping platforms, high-density arrays and exome-capture

systems, is performed using genotyping data from Watkins lines in a population genetics context. With regards to the evolutionary analysis, the initial stage was to infer a maximum likelihood tree to examine the reliability of the tree topology. In this tree, many of the subclades within the Watkins clade had low bootstrap support values, indicating that the fine-scale topology of the tree is unreliable. This mirrors the findings of Gardiner et al. [-@gardinerHiddenVariationPolyploid2018]. The implication of this is that the TMRCA values for these subclades in the maximum-clade consensus tree produced using BEAST are not very useful, as they do not relate to the true topology of the phylogeny. Nonetheless, we can still use the TMRCA value for the Watkins clade as a whole as a useful benchmark of the origin of these varieties.

The results of the TMRCA analysis itself are perplexing, in that they don't conform to our expectations given what is currently known about the evolutionary history of wheat varieties. The tree inferred by BEAST along with the calibration of the molecular clock based on the suspected divergence point of wheat and *Aegilops tauschii* (~ 9000 ybp), produced an estimation of 860 ybp for the date of the most recent common ancestor of all Watkins lines. Taking this at face value, the late divergence of Watkins lines compared to the divergence of hexaploid bread wheat from the D genome ancestor (860 ybp compared to 9000 ybp) indicates that wheat underwent significant genetic changes in those preceding 8200 years. This seems unlikely and is probably an underestimation of the true TMRCA of the Watkins clade, as it is widely believed that the spread of germplasm leading to global landraces occurred from 8000 to 2300 years ago [@balfourierWorldwidePhylogeographyHistory2019]. Indeed archaeological evidence in the form of radiocarbon dated wheat grains has shown that wheat was cultivated in China from at least 4000-5000 years ago [@bonjeanWorldWheatBook2001]. Whether these were the ancestors of the lines that went on to become Chinese Watkins lines is unknown, but it seems unlikely that these lines, adapted to local environmental conditions, would have been replaced by landraces from elsewhere in the last 500 years, as suggested by the molecular clock analysis.

What could have caused this dramatic difference between the results and our expecta-

tions? The inference of the TMRCA may be influenced by different rates of evolution between *Aegilops tauschii* and *Triticum aestivum*. Since *T. aestivum* is a hexaploid organism, we could hypothesize that the individual sub genomes evolve faster than the diploid genome of *A. tauschii*, as the increased redundancy provided by homeologous copies of each gene reduces the impact of detrimental mutations in individual genes, causing a reduction in stabilizing selection. This is sometimes referred to as mutational robustness [@vandepeerEvolutionarySignificanceAncient2009]. The calibration of the molecular clock was performed by examining the number of differences in genotype between Chinese Spring and *A. tauschii*, and then equating these differences to a time of ~ 9000 years, which is thought to be when hexaploid bread wheat originated.

Following the hypothesis that these two organisms evolve at a different rate, the differences in genotype then are the result of a mixture of two evolutionary rates, the first, slower rate, which we will name r~1, operating along the branch of the phylogeny leading to *A. tauschii*, and the second, faster rate, r~2, operating along the branch leading to *T. aestivum*. As Watkins lines are wheat varieties, they should evolve under the faster rate given the mutational robustness hypothesis, and so any differences in genotype between Watkins lines are the result of r~2 operating along each branch leading to each respective variety. This would lead to an inflation in the difference in genotypes between any pair of Watkins lines when compared to *A. tauschii* for the same time period, and should therefore artificially increase the TMRCA values between pairs of Watkins lines, as the calibration of the phylogeny was based on a combination of rates r~1 and r~2, rather than r~2 and r~2 (an equal rate of evolution in both species). Whilst the mutational robustness hypothesis is logical, it actually predicts that the inferred TMRCA should be older than the true date, which is the opposite of what is observed here, suggesting that *A. tauschii* actually evolved at a faster rate than the landraces of wheat. Why this is the case is difficult to say, but it is possible that this is due to the constraints on wheat landrace genetic diversity imposed by farming practices, whereas *A. tauschii* is a wild plant free from the constraints of artificial selection. This may have allowed it to accumulate more mutations than wheat landraces, outweighing the mutational robustness that wheat varieties gained as a polyploids.

In addition to the molecular clock analysis, this chapter focuses on exome capture and array based genotyping data - how do they compare to each other in a population genetics context? Does one provide additional information that is missing from the other? The large hexaploid genome of wheat presents many opportunities for off-target hybridization events in arrays and in exome capture, as both use DNA probes to target specific regions of the genome. Does this affect e.g. array-based data more than exome-capture data? Another factor to consider is the lack of a complete, chromosome-level genome assembly at the time of producing one of the more popular wheat arrays, the 35k Wheat Breeder's array [@allenCharacterizationWheatBreeders2016], and so it is likely that many of the probe sequences are actually chimeric, containing sequences from a mixture of wheat's three sub genomes. This means that it is common for probe sequences to have lowest e-value BLAST hits (or highest scoring, best hits) to sequences in a sub-genome that is not the target of probe hybridization itself in the molecular reaction.

The population genetics analysis as a whole, including both STRUCTURE and PCO analysis, show that these factors are not overly impactful on array-based data, as the array and exome capture results were highly similar, both in their broad scale patterns, e.g. the separation of varieties by region along the axes of the PCO plots (figure 2.2), as well as in the minute details, such as the placement of Watkins varieties 300 and 299 (figure 2.2). It is clear then that both array and exome capture data contain much of the same information in a population genetics context even though the array data contained fewer SNPs. This is because a large number of SNPs in the exome capture data are likely to exhibit a high degree of linkage disequilibrium, whereas the SNPs on the array were curated to contain a high degree of genetic information across varieties by selecting those with the highest polymorphic information score.

3 Detecting a shift in recombination distribution using high-density genotyping data

The results described here regarding the effects of temperature on recombination distribution and frequency have been published in the following manuscript:

Coulton, A., Burridge, A., Edwards, K. 2020. Examining the effects of temperature on recombination in wheat. Frontiers in Plant Science

Genotyping was performed by Burridge, A. Temperature treatments were performed by Edwards, K.

3.1 Introduction

Meiosis is a specialised type of cell division that leads to the production of haploid gametes. A key feature of meiosis is the process of recombination, where parental genetic material is shuffled together to create chimeric chromosomes, somewhat akin to shuffling a pack of playing cards. Recombination is crucial to the evolution of species, facilitating the spread of beneficial combinations of alleles whilst allowing unfavourable ones to be reduced in the population [@wilkinsEvolutionMeiosisMitosis2009]. In addition to its role in the formation of natural populations, meiosis is also exploited in agriculture, where breeders cross different varieties of crops or animals together to produce offspring with a mixture of both parental phenotypic traits [@acquaahPrinciplesPlantGenetics2007]. This principle has previously been exploited to produce dramatic yield increases in staple food crops, such as during the green revolution, in which short-stemmed Japanese wheat varieties were crossed with high-yielding American varieties [@russellProgressPlantBreeding1985]. This resulted in plants that were less susceptible to lodging [@rajikumaraLodgingCerealsReview2008].

The process of recombination occurs during prophase I of meiosis, preceded by DNA replication in S phase. In addition to the shuffling of parental DNA, recombination is also essential to the mechanics of meiosis, both serving to bind homologs together

into bivalents and also ensure their correct segregation [@zicklerRecombinationPairingSynapsis2015]. Bound homologs impart tension on the meiotic spindle during their alignment along the equator of the cell, triggering chromosome disjunction at anaphase [@zicklerRecombinationPairingSynapsis2015]. Initially, each pair of sister chromatids is linked together by a linear protein axis from which DNA extends in loops, known as the leptotene subphase of prophase I. During this phase, telomeres cluster together at the nuclear envelope to form a structure known as the bouquet. The respective axes of each pair of homologs will go on to become the lateral elements of the synaptonemal complex (SC), a tripartite, ladder-like structure, visible cytologically, that enables the formation of bivalents through synapsis of homologous chromosomes, formed during the zygotene subphase. The lateral elements are joined together along their lengths by transverse filaments. Once the synaptonemal complex is in place, non-sister chromatids are within close proximity to each other within the cell, allowing the process of recombination to begin. Recombination is initiated through the formation of double-stranded breaks (DSBs) in the DNA via the topoisomerase-like protein SPO11 [@nealeClarifyingMechanicsDNA2006]. Individual DNA strands are then resected, generating overhanging 3‘ tails. Strand invasion of the non-sister chromatid can then occur, forming a Holliday junction, or a formation in which non-sister chromatid DNA strands are physically crossed over each other. At this stage, various recombination-intermediate molecules are possible, each associated with a different outcome. Double Holliday junctions, formations in which the non-sister chromatid DNA strands cross each other twice, result in crossover (CO) events, visible cytologically as chiasmata, producing hybrid DNA molecules composed of long stretches of DNA from each progenitor molecule / parent (illustrated in figure 3.1). In contrast, when only a single Holliday junction is formed, no crossover occurs, the recombination-intermediate instead being resolved with only a short stretch of DNA exchanged between the non-sister chromatids, known as a gene conversion. The ratio of DSBs to CO events is high, with only a minority of DSBs resulting a CO. In addition, when a CO is formed it prevents further COs occurring in its vicinity, a phenomenon known as crossover interference. Also possible is a non-crossover (NCO) event, where neither a gene conversion or a crossover event results

from the DSB [@filippoMechanismEukaryoticHomologous2008].

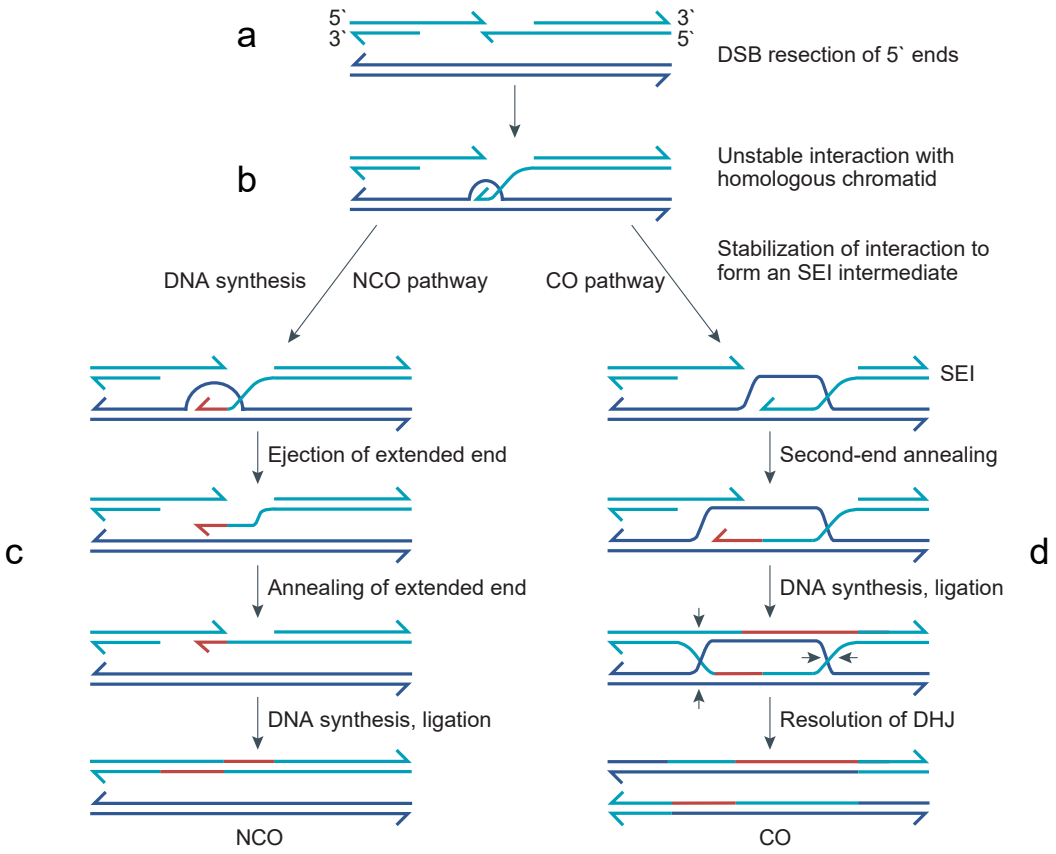


Figure 3.1 Diagram illustrating the process of recombination between non-sister chromatids during meiosis, adapted from [@marstonMeiosisCellcycleControls2004]. DNA synthesis is denoted by the colour red, whilst 3 prime ends are denoted by half-arrows. (a) Spo-11 creates a double stranded break in the DNA, which is then resected at the 5 prime ends to reveal short overhanging 3 prime tails. (b) One of the overhanging tails begins the process of strand invasion, forming an unstable interaction that can take one of two paths, the first leading to a crossover (CO) event (right) and the second leading to a noncrossover (NCO) event. (c) In the path leading the non-CO event, DNA synthesis begins on the invading strand but the strand is ejected from the non-sister chromatid. Following this, DNA synthesis and ligation of the invading strand to its original location occurs. (d) In contrast to this, in the CO pathway, the interaction between non-sister chromatids is stabilised to form a single-end invasion (SEI) intermediate, which contains a single Holliday junction. DNA synthesis then occurs from the 3 prime tails, after which they are annealed to their original location, but with the strands in the formation of a double Holliday junction (DHJ). The DHJ proceeds to be nicked at the positions indicated by arrows and recombinant DNA molecules are formed.

Whilst recombination has previously been harnessed to great effect by breeders, its utility is limited in important staple food crops such as wheat [@saintenacDetailedRecombinationStudies2009; @chouletStructuralFunctionalPartitioning2014], maize [@heGenomicFeaturesShaping2017] and barley [@higginsFactorsUnderlyingRestricted2014]. The distribution of recombination events in these crops is strongly skewed towards the distal ends of the chromosomes, with little to no crossovers occurring in the region surrounding the centromere, known as the pericentromeric region [@zelkowskiDiversityDeterminantsMeiotic2019]. This contrasts with *Arabidopsis*, in which the distribution of COs is much more uniform, with only the centromeric region showing highly reduced numbers of COs [@girautGenomeWideCrossoverDistribution2011; @choiArabidopsisMeioticCrossover2013]. With the advent of a chromosome-level genome assembly for wheat @consortiumiwgscShiftingLimitsWheat2018, it is now known that genes are distributed somewhat evenly along the chromosomes, with many potentially important genes being found in the pericentromeric region. The limitation in recombination distribution therefore creates a problem for breeders, as it is not possible to break up large central linkage blocks. A central area of current research in crops is to examine the cause of this skewed recombination distribution and, following this, to try and implement measures to induce recombination in the pericentromeric region [@higginsSpatiotemporalAsymmetryMeiotic2012; @phillipsEffectTemperatureMale2015].

Much of the salient research on recombination distribution in cereals has been performed in barley [@higginsSpatiotemporalAsymmetryMeiotic2012a; @phillipsEffectTemperatureMale2015], *Hordeum vulgare*, a diploid member of the pooideae subfamily of poaceae, the grasses. The two primary avenues with which to investigate this are cytologically, making extensive use of fluorescent staining in various forms [@higginsSpatiotemporalAsymmetryMeiotic2012a], as well as genetically via SNP genotyping of mapping populations [@phillipsEffectTemperatureMale2015]. Cytological analysis includes immunofluorescence and fluorescence in situ hybridization (FISH), the former making use of fluorescent antibodies targeted at homologues of proteins that have been shown to be part of the meiotic apparatus in *Arabidopsis thaliana*, and the latter

utilizing labelled DNA probes to target specific regions of the DNA. Notable among these antibodies are *Arabidopsis asynaptic 1* (ASY1), associated with the linear axis formation, and therefore the early stages of synaptonemal complex formation during the G2 phase of meiosis [@higginsSpatiotemporalAsymmetryMeiotic2012a], as well as the *Arabidopsis* SC transverse filament protein zipper 1 (ZYP1), which allows the monitoring of synapsis. The research in [@higginsSpatiotemporalAsymmetryMeiotic2012a] is of special interest as it presents data that forms the beginning of one of the only mechanistic explanations of the distal distribution of recombination events in crops. By using a telomere FISH-labelled probe in conjunction with ASY1 and ZYP1 antibodies it was found that early synaptonemal complex formation, and hence chromosome synapsis, exhibits a spatiotemporal bias towards the distal ends of the chromosomes. As meiosis proceeds, ASY1 and ZYP1 fluorescent signals extend from the subtelomeric regions to the interstitial regions of the chromosomes. In addition, an antibody for the MutL homolog MLH1, which is involved in the resolution of double Holliday junctions and therefore marks CO sites, was used in conjunction with ZYP1. The authors suggest that at a temperature of 22°C, MLH1 primarily occurred in the distal / subtelomeric regions of the chromosomes. They did not perform a stain that utilized the telomeric FISH-labelled probe in conjunction with the MLH1 antibody, so this distal localization of MLH1 must be inferred from previous experiments that show colocalization of ZYP1 and the telomere probe in early prophase.

Overall, these experiments suggest that the distal distribution of recombination events is primarily due to the spatiotemporal bias of chromosome pairing and synapsis to the distal regions of the chromosomes, facilitating preferential crossover formation these regions. The authors further hypothesized that crossover interference [@bromanCrossoverInterferenceMouse2002], the phenomenon in which one CO prevents further COs from forming in its vicinity, could be one of the factors that prevents further CO formation in the pericentromeric region [@higginsSpatiotemporalAsymmetryMeiotic2012]. This is not a recent hypothesis, being suggested as early as 1993 in a study utilizing C band polymorphism between varieties to assess recombination distribution [@lukaszewskiPhysicalDistributionRecombination1993]. The immediate question that

follows from this is: why does chromosome pairing and synapsis occur first in the distal regions of the chromosomes? Interestingly, in a study of rye chromosome 1R [@lukaszewskiUnexpectedBehaviorInverted2008], which usually has a distal distribution of recombination events, an inversion of the long arm resulted in chiasmata proximal to the centromere, indicating that chiasma formation is not dependent on physical position along the chromosome, but rather some property of individual chromosome segments themselves, whether that be genetic, epigenetic, or due to the chromatin composition of segments.

Factors shown to be influence meiotic recombination range from internal, genetic factors, such as the FANCM gene, which limits meiotic crossovers in Arabidopsis @crismaniFANCMLimitsMeiotic2012, to external stresses such as soil magnesium content [@reyMagnesiumIncreasesHomoeologous2018] and environmental temperature [@jainEffectHighTemperature1957; @loidlEffectsElevatedTemperature1989]. Temperature has recently been examined more thoroughly in barley with particular focus on its effect on the distribution of COs [@higginsSpatiotemporalAsymmetryMeiotic2012; @phillipsEffectTemperatureMale2015]. Cytological analysis of meiocytes revealed a small but significant reduction in mean chiasma frequency between meiocytes grown at 22°C and 30°C, as well as significantly more interstitial chiasmata for chromosome 5H when grown at 30°C compared to those grown at 22°C [@higginsSpatiotemporalAsymmetryMeiotic2012]. Further research attempted to expand on these results by utilizing SNP genotyping of a mapping population in addition to cytological analysis [@phillipsEffectTemperatureMale2015]. This method has the potential to provide a more precise evaluation of how large the shift in distribution of recombination events in response to temperature is, expanding on the categorical assignment of recombination events to either "distal" or "interstitial" positions in analysis of cytological data, as well as revealing the genes that are effected by a shift. Unfortunately, the authors did not perform a statistical analysis of the distribution of events in their SNP data, which was not made available for public use. In addition, they did not identify individual chromosomes in their cytological data [@phillipsEffectTemperatureMale2015]. It is therefore difficult to compare the two studies for consistency in results; these limitations of the literature

provide the incentive to explore this topic further.

Existing research in wheat suggests that high temperatures prevent normal meiotic progression and therefore reduces the fertility of plants, with nullisomic lines identifying 5D as one of the chromosomes effecting temperature sensitivity [@draegerShortPeriodsHigh2017]. However, there has yet to be a detailed analysis of the effects of temperature on recombination distribution in wheat. Here we have utilized a high-density SNP genotyping array [@allenCharacterizationWheatBreeders2016] and four F2 mapping populations of an Apogee X Paragon (A x P) cross, each subjected to a different temperature during meiosis (10°C, 14°C, 26°C and 28°C respectively), to examine whether wheat behaves in the same way as barley, and to assess the utility of any shift in distribution to breeders.

Recombination is known to be influenced by external factors such as temperature in many organisms, but are there internal factors also involved, such as particular loci in the genome? It is possible to extract the recombination frequency of each genotyped individual in a mapping population by examining changes in the phase of parental genotypes between clustered, ordered markers in a genetic map. This can then be used as a phenotype in QTL analysis, allowing potential loci involved in the determination of recombination frequency to be identified. Several recombination frequency QTL have already been identified in this manner [@gardinerAnalysisRecombinationLandscape2019; @wingenWheatLandraceGenome2017]. In addition, it is also possible to extract the distribution of recombination events as a phenotype. So far this has been attempted by dividing the chromosomes into sections comprising the distal (1/3 of chromosome arm away from the centromere) and pericentromeric (2/3 of chromosome arms nearer the centromere) regions and counting the number of events in each [@jordanGeneticArchitectureGenomewide2018]. In this study, we analyse several mapping populations, including Chinese Spring X Paragon, Apogee X Paragon, Avalon X Cadenza, Opata X Synthetic, Rialto X Savannah, Paragon X Watkins 49 and Paragon X Watkins 94 to try and identify QTL involved in both frequency and distribution of recombination events. For our recombination distribution phenotype we use a different method to

the one already mentioned, in which we assign markers positions in the physical wheat genome assembly [@consortiumiwgscShiftingLimitsWheat2018]. We then take the average distance of recombination events from the centre (50% mark) of the chromosome and use this as the phenotype for each individual for each chromosome. We hope that this will allow a more precise estimation of the genetic effects of parental alleles on recombination distribution.

Although putative QTL have been identified effecting recombination frequency in wheat, their phenotypic effect is often limited, with a QTL in a Chinese Spring X Paragon population only explaining 6.037% of phenotypic variance [@gardinerAnalysisRecombinationLandscape2019]. This presents the question: how much genetic variation is their between meiotic genes in wheat, and is there potential to modify recombination distribution and frequency simply through coordinated breeding of particular varieties? Here we perform a phylogenetic analysis of four wheat varieties and a close relative of wheat, barley, including two gene sets from the IWGSC RefSeq v1.0 genome assembly. The first contains homologues of genes that have been functionally characterised to be involved in recombination in *Arabidopsis* [@alabdullahCoexpressionNetworkHexaploid2019], whilst the second contains a random selection of 5000 genes. We construct a phylogeny based on a superalignment of these genes to test the hypothesis that meiotic genes are generally more conserved than a random selection of genes due to stabilising selection.

3.2 Materials and Methods

3.2.1 Plant cultivation and temperature treatments

F1 seed from a Paragon X Apogee cross were obtained from Dr Peter Jack at RAGT Seeds Ltd. These were randomly separated into four populations and grown initially in uniform conditions. Plants were grown in pots filled with peat-based soil and kept in a glasshouse at 15-25 °C with 16-h light, 8-h dark. Plants were deemed to be undergoing meiosis when the base of the stem showed visible swelling due to the growth of the developing head within the flag leaf sheath, often referred to as the 'booting' stage of

development [@barberDecimalGrowthStages2015]. At this point, they were transferred to temperature-controlled cabinets at 10 °C, 14 °C, 26 °C & 28 °C respectively for around 3 weeks. They were then transferred back to the glasshouse to avoid effects of temperature on pollen tube development. Seeds were then harvested from each of the populations. Leaf-tissue was harvested from F2 plants 12-14 days post-sowing, when the plants were at an early seedling stage. The sizes of the F2 populations were 80, 75, 70 and 78 individuals for temperature treatments of 10°C, 14°C, 26°C and 28°C respectively. DNA was extracted following the protocol in [@pallottaMarkerAssisted-Wheat2003] with minor modifications.

3.2.2 Sample genotyping

DNA concentration was assessed using a Qubit 2.0 Fluorometer and was the normalized to 23 ng / µl ready for analysis with the Axiom® Wheat Breeder's array [@allenCharacterizationWheatBreeders2016]. Sample preparation for array genotyping was performed with the Beckman Coulter Biomek FX. Samples were then genotyped using the Axiom® 35K Wheat Breeders array in conjunction with the GeneTitan® using standard Affymetrix protocols (Axiom® 2.0 Assay for 384 samples P/N 703154 Rev. 2).

Axiom Analysis Suite (version 3.1.51.0) was used to assign genotype calls. Of the 6536 polymorphic SNPs between the two parental varieties present on the array, 2504 codominant SNPs of highest quality were selected through visual inspection of each cluster plot. Only markers with a clear delineation between genotyping clusters representing homozygotes for the Apogee allele, heterozygotes and homozygotes for the Paragon allele were used, and borderline markers were recoded as no-calls as a precaution against genotyping errors that could affect the recombination analysis. A custom R script was used to assign genotype calls to parental varieties, such that an "A" genotype represented an allele from Apogee, whilst a "B" genotype represented an allele from Paragon (supplementary file 1).

3.2.3 Genetic map construction

Data from all four populations was amalgamated and used for initial genetic map construction using MultiPoint Complete (version 4.1). Markers exhibiting large segregation distortion ($2 > 20$), low informativity (LOD < 7), and large amount of missing data were removed from dataset before proceeding with genetic map construction. MultiPoint first performs binning of markers that have the same genotype across all individuals. These bins or "skeleton markers" are then clustered based on an initial threshold recombination fraction, in this case 0.2, which was iteratively increased up to a value of 0.34. After clustering, the markers were ordered in MultiPoint using a guided evolutionary strategy optimization algorithm [@mesterFastAccurateConstruction2015] in conjunction with a jackknife resampling strategy to remove any markers that caused unstable regions in the marker order. Genetic distances of markers were estimated from recombination fractions using the Kosambi mapping function.

Only the skeleton markers in this initial genetic map were retained, as these are the only informative markers for evaluating recombination events. The cluster and ordering information from this initial genetic map was applied to the genotyping data from each of the four populations. Assignment of chromosomes to linkage groups was performed both by comparison of linkage groups to previous marker assignments based on nullisomic lines @winfieldHighdensitySNPGenotyping2016, as well as BLASTN searches of probe sequences to the IWGSC RefSeq v1.0 assembly @consortiumiwgscShiftingLimitsWheat2018, hereafter referred to as the IWGSC assembly. In all cases, only BLAST hits with an e-value smaller than 10-19 were used. Four further genetic maps were then generated, one for each population, retaining the same clustering and ordering information from the first map, using the quickEst function in the R package ASMap (version 1.0-2) [@taylorPackageASMapEfficient2017]. This allowed comparison of recombination distribution between temperature treatments using centimorgan values of markers. To verify the quality of the genetic maps, we performed a comparison with the Apogee X Paragon F5 map produced by Allen et al., [-@allenCharacterizationWheatBreeders2016]. As these maps both involve the same parental varieties, they should show close resem-

blance in their clustering of markers, and should also approach colinearity in their ordering of markers, allowing for small perturbations that may be caused by genotyping error or missing data [@hackettEffectsGenotypingErrors2003; @wuEfficientAccurateConstruction2008].

3.2.4 Detection and processing of recombination events

Recombination events were detected by a change in genotype between consecutive markers in the genetic map. Transitions from a homozygous allele for one parent to a homozygous allele for the second parent were scored as two recombination events, as these would have required recombination to occur in this position in both of the gametes that formed the zygote. With the inference of recombination events from SNP data, genotyping error has the potential to erroneously inflate the number of events observed. The magnitude of this effect depends on both where the genotyping error occurs and what the erroneous assignment is, either homozygous or heterozygous. For instance, let us consider three markers used to genotype a single individual, M~1, M~2 and M~3. If the genotypes of the markers are A, A, A, (i.e. homozygous for the allele from the first parent) and M2 erroneously changes to B (i.e. homozygous for the allele from the second parent), in an F2 individual, this would indicate that four recombination events have occurred. In order to compensate for potentially erroneous genotypes, genotypes conforming to this three marker scenario where M1 and M3 are within 30 Mb of each other were recoded as no-calls.

3.2.5 Statistical method of detecting differences in recombination distribution between temperature treatments

To test statistically whether there was a shift in distribution of recombination events between treatments, it was necessary to take the mean distance of all recombination events within an individual from the centromere of the chromosome, which we will refer to henceforth as the mean recombination distance (MRD). Positions of centromeres

were based on previously published ChIP-seq data for centromere specific histone 3 (CENH3) @consortiumiwgscShiftingLimitsWheat2018, whilst the position of each recombination event was taken as the midpoint between the two markers exhibiting the genotype change. Recombination events cannot be tested individually as events within a sample are not independent. For example, if two recombination events occur on the same chromosome during a single meiosis, the position of the second event is likely to be influenced by the position of the first due to crossover interference. In SNP data derived from mapping populations, individuals are a product of two meioses, one for each gamete that contributed to the formation of the individual. It is not possible to assign recombination events to one or the other meiosis, and it is therefore not possible to determine which events may have been influenced by crossover interference. For example, if we observe two events at 2% physical distance either side of the centre, and a third event 10% physical distance from one of the telomeres, we know that it is very unlikely that the two central events occurred in the same meiosis, but we do not know which of them comes from the same meiosis as the distal event. Using individual recombination events then in our statistical analysis could cause conflation of the effects of the treatment, e.g. external temperature during meiosis, on recombination distribution with the effects of crossover interference.

In addition, since wheat chromosomes are not metacentric, with a mean ± s.d. centromere position of 40.5 ± 6.14 % of the physical sequence for each chromosome @consortiumiwgscShiftingLimitsWheat2018, we partitioned our measurement of MRD between chromosome arms. Previous work suggests that crossover interference in wheat is strongest at distances smaller than 10 cM, which in addition to the distal distribution of recombination events in wheat, should preclude strong inter-arm effects of crossover interference. MRD was measured in units of percentage physical distance along chromosome arms.

The analysis of recombination distribution was performed using physical distances of markers along the IWGSC assembly as determined by BLAST. This allows us to relate shifts in distribution to genomic features such as genes. Wheat has a recombination

distribution that is biased towards the telomeric regions of the chromosomes. This being the case, it was necessary to ensure that genetic map representations of the chromosomes had sufficient marker coverage to detect potential shifts away from the telomeric regions towards the centromeres. To do this, we defined anchor points along the chromosome at 0, 25, 50, 75 and 100 (% physical distance). We then identified the nearest marker to each of these anchor points. If the distance between any of the anchor points and their nearest markers was greater than 25, the chromosome was excluded from the analysis. So for example, if a chromosome consisted of markers at 40, 50, 60, 70, and 80%, it would be excluded as the nearest marker to the first anchor point (0), is 40, which is more than 25% away, precluding the analysis of recombination events at that end of the chromosome. In addition to ensuring adequate coverage, it was also necessary to ensure that the order of markers in the genetic maps and the physical map were concordant, as a discordant order could bias our analysis of recombination distribution. To do this, the longest increasing subsequence of physical positions of markers was taken for each chromosome, and all markers not included in this sequence were removed.

3.2.6 Examination of the influence of genotyping error

The data was also examined for other potential factors that could cause differences in MRD between temperature treatments. Genotyping error can be assessed with standard quality control variables produced by Axiom Analysis Suite, such as dish QC, which is based on the contrast between probe hybridization signals at non-polymorphic genome locations, and QC call rate, which measures the call rate of a subset of probes for a particular sample [@affymetrixAxiomGenotypingSolution2015]. To assess whether genotyping error influenced recombination distribution, linear regressions were performed against these two variables using MRD as the predictor.

3.2.7 Examination of the influence of sample size

Another possible explanation for any observed shifts in MRD is that the sample sizes of the mapping populations used were too small. The distribution of recombination events may appear to be skewed towards the centromeres simply due to sampling error, i.e. if we happened to produce individuals that had more centromere-proximal events and were not observing the true value of MRD in the population. To test whether this was the case, we utilized simulated genotyping data from PedigreeSim [@voorripsSimulationMeiosisDiploid2012]. This performs a detailed approximation of meiosis including the formation of bivalents, chiasmata, and meiotic divisions for a specified population structure. 200 genotyping datasets were produced, each containing an F2 mapping population of 500 individuals. These had a marker distribution based on chromosome 3B from the Apogee X Paragon F2 genetic map produced here. Further datasets with sample sizes of 250, 100 and 30 individuals were then created by randomly subsetting the original 200 datasets. For each sample size, 100000 random combinations of four datasets were generated. We then performed a Kruskal-Wallice test on MRD across the whole chromosome, for every combination, for every sample size, examining whether there was an increasing number of significant tests as sample size decreased, which would indicate that sample size has an influence on MRD.

3.2.8 QTL analysis

QTL analysis was performed using the R package rQTL. Marker LOD scores were calculated using the extended Haley-Knott method, and the LOD threshold for significance was calculated though permutation of phenotypes relative to the genotype data, using 1000 permutations. LOD scores were restricted to the markers themselves, with no imputation of genotypes between markers. To account for the testing of multiple phenotypes, a second significance threshold was also calculated which was the initial threshold after Bonferroni correction at $p < 0.05$. Where multiple QTLs were detected for the same phenotype in this initial single QTL model, QTLs were combined into a multi-QTL model to examine the combined effect on phenotype as well as any po-

tential QTL interactions. The recombination frequency phenotype was calculated by examining the genetically clustered and ordered genotyping data for changes in genotype between consecutive markers. For example, consider three markers, M1, M2, and M3, with genotypes A, A and B respectively. We can infer that a recombination event has occurred between markers two and three. If this is an F2 population, this would be considered as two recombination events as recombination would need to occur in both gametes for a transition from being homozygous for the allele from the first parent to homozygous for the allele for the second parent. Likewise, a transition from heterozygous to homozygous would be considered as one recombination event in an F2 population. For the recombination frequency phenotype, genotype data was did not undergo the same filtering as for the recombination distribution phenotype (i.e. taking the subset of markers consisting of the longest increasing subsequence of physical positions relative to the physical genome assembly), as in this case it is not necessary to have strict concordance with the IWGSC RefSeq v1.0 physical genome assembly.

3.2.9 Phylogenetic analysis

The aim of the phylogenetic analysis was to determine whether genes involved in meiosis are generally more conserved among wheat varieties than most other genes in the genome. Genes potentially involved in meiosis were taken from [@alabdullahCoexpressionNetworkHexaploid2019], who searched for orthologues of 103 functionally characterized meiotic genes in model plant species. A random sample of 4600 genes from the genome was used as a comparison gene set. Homologues of these genes were then obtained from genomes of other wheat varieties, as well as the barley genome to provide an idea of the evolution of these genes in a more distantly related grass species. Genome sequences of wheat varieties Robigus, Paragon, Cadenza and Claire were obtained from the Grassroot genomics web page (http://www.earlham.ac.uk/grassroots-genomics); the genome for Chinese Spring was the IWGSC RefSeq v1.0 [@consortiumiwgscShiftingLimitsWheat2018]; the barley genome was obtained from [@mascherChromosomeConformationCapture2017]. To obtain homologues from these genomes, a strategy

using BLAST was chosen over bioinformatics packages formally used to map CDS sequences to genomic sequences, e.g. Exonerate [@slaterAutomatedGenerationHeuristics2005a], GMAP [@wuGMAPGenomicMapping2005] and genBlastG [@sheGenBlastGUsingBLAST2011], as these programs were found to be prohibitively slow in producing alignments.

First, BLAST searches of genomic sequences for meiotic genes extracted from the IWGSC assembly against genomic sequences from other samples were performed with default settings, with the -culling_limit parameter set to 10. BLAST does not return full length alignments of query against target, but instead returns a series of local alignments called high-scoring segment pairs (HSPs). The results of the initial BLAST search was therefore used to identify regions in which the homologues was most likely located. HSPs less than 7000 bp apart (a distance determined empirically) were grouped together and their average bitscore was calculated. The group of HSPs with the highest bitscore was determined to be the homologous sequence, and both the lowest and highest base positions of HSPs within the group were used as a coordinate range for sequence extraction from the genome. Once all sequences were extracted, a multiple alignment of genomic sequences was produced using MUSCLE [@edgarMUSCLEMultipleSequence2004], and the exonic regions of the Chinese spring sequence was determined using the GFF3 file containing exon coordinates for each gene. These exonic regions were assumed to be the exonic regions for all sequences. A new multiple alignment was then produced containing only exonic sequences.

To assess the amount of evolution that had occurred for each gene among samples, IQTREE was used to produce gene trees. IQTREE was used with default settings, and therefore for each gene tree inference, an AIC test was performed to identify the best model of sequence evolution.

3.2.10 Gene distribution analysis

In the analysis of recombination with respect to gene distribution, genes were taken from the high-confidence annotation of the IWGSC assembly v1.0 [@consortiumiwgscShift-

ingLimitsWheat2018]. Centromeres were marked according to ChIP-Seq data [@consortiumiwgscShiftingLimitsWheat2018]. Where more than one region was specified, the region that correlated with the highest peak of transcriptome element families Cereba and Quinta was taken as the position of the centromere.

3.3 Results

3.3.1 Gene distribution analysis

Whilst the distribution of recombination events is mainly towards the distal ends of the chromosomes (figure 3.2), the distribution of genes is much more even along the chromosomes (figure 3.3). There are also many more genes than probes on the Axiom 35k array. This effect differs among sub-genomes, with the D genome having fewer probes compared to genes than the other genomes (figure 3.3).

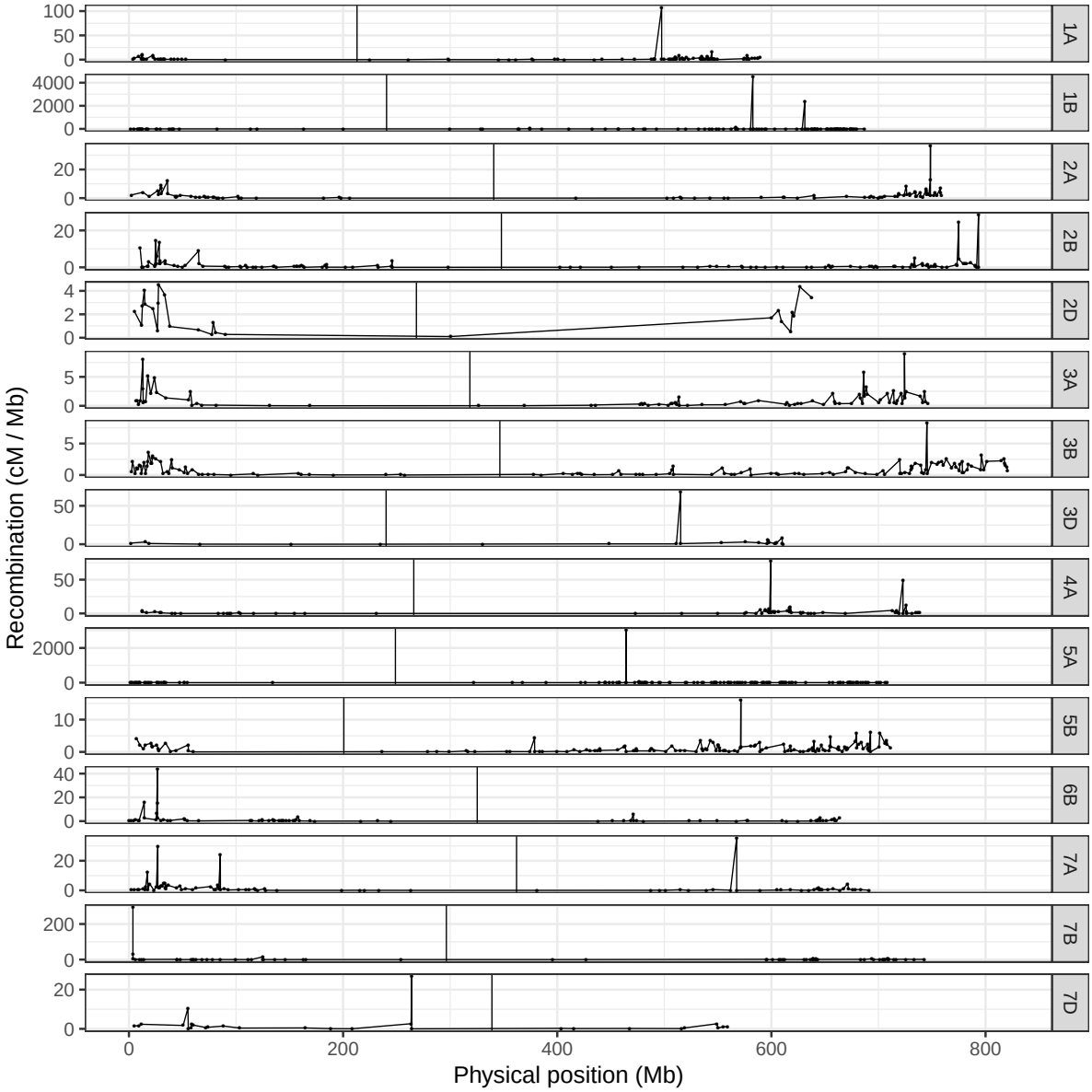


Figure 3.2 Recombination maps for a Chinese Spring X Paragon F5 mapping population. Positions of centromeres are marked by vertical black lines, as determined by ChIP-Seq data [@consortiumiwgscShiftingLimitsWheat2018]. Recombination is mostly absent in regions surrounding the centromeres. Marker positions are denoted by circles. Chromosomes have been selected based on overall marker coverage, whilst markers have been filtered, selecting the longest increasing subsequence of markers that are concordant between genetic and physical maps.

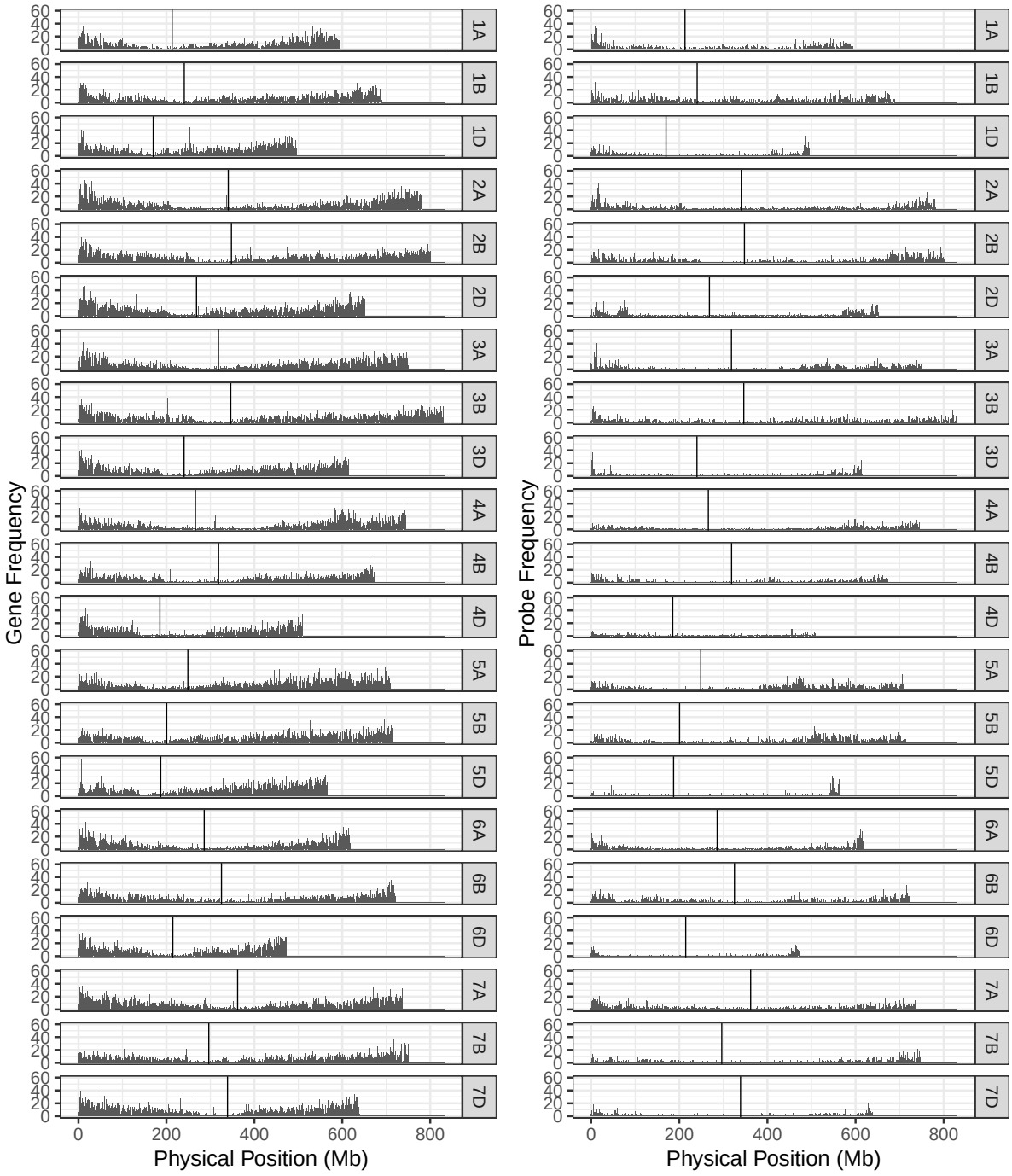


Figure 3.3 Comparison of gene and probe distribution in wheat. The left column shows the distribution of high-confidence genes along each of the chromosomes of the IWGSC RefSeq v1.0 wheat genome assembly, whilst the right column shows the distribution of Axiom probes from the wheat breeder's 35k array along the genome.

3.3.2 Apogee X Paragon Results

Processing the data for concordance between genetic and physical maps

To analyse potential changes in recombination distribution between temperature treatments, it was first necessary to select only the chromosomes that had a sufficient marker distribution to detect an inward shift in recombination distribution. We found that chromosomes 1A, 2A, 2D, 3A, 3B, 4A, 5A, 6B, 7A and 7B met our criteria for marker distribution (figure 3.4). These chromosomes contained a total of 444 markers in the initial genetic map prior to any filtering. 870 of these markers had valid BLAST hits to the IWGSC assembly and could therefore be assigned physical positions (figure 3.5a); the remaining markers were discarded. Markers were then filtered further: any markers that had a discordant order between the genetic and physical maps were removed (figure 3.4b), leaving 442 markers, or 44.2 ± 12.15 (mean ± s.d.) markers per chromosome for further analysis (supplementary file 2). We will refer to this map as the filtered genetic map. Overall marker distribution was generally preserved during this stage, although the number of markers immediately adjacent to the centromere for chromosomes 2A and 6B was noticeably reduced (figure 3.4b). The mean ± s.d. distance between markers (Mb) was 8.9 ± 26.4, 17.33 ± 48.5, 24.39 ± 62.64, 20.58 ± 42.2, 15.78 ± 44.34, 10.71 ± 33.35, 13.4 ± 22.58, 20.69 ± 46.89 and 17.3 ± 39.21 for chromosomes 1A, 2A, 2D, 3A, 3B, 4A, 5A, 6B, 7A and 7B respectively. Marker density was generally highest at the distal ends of the chromosomes with a drop in the number of markers in regions surrounding the centromeres (figure 3.4).

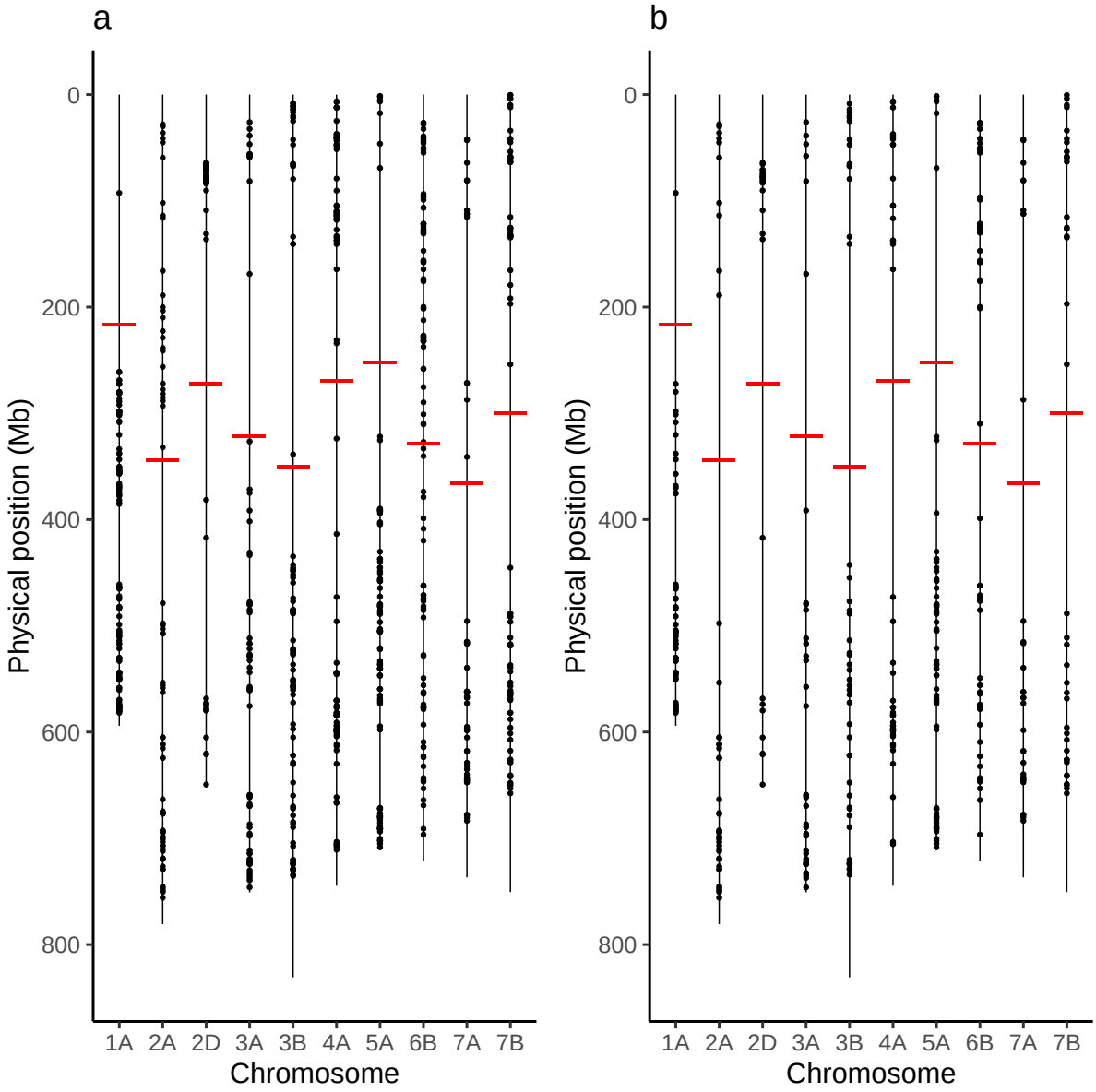


Figure 3.4 Marker distribution for chromosomes that passed our filtering criteria. (a) Marker distribution before removal of markers with discordant order between genetic and physical maps via the longest increasing subsequence. (b) Marker distribution after removal. Vertical lines represent the entirety of the length of each chromosome, taken from the IWGSC assembly, whilst points represent the positions of markers. Horizontal red lines mark the position of the centromere on each chromosome.

Performing comparisons with previously generated F5 Apogee x Paragon map

Of the 2503 markers selected from Axiom Analysis Suite for the F2 A x P genotyping data produced here, 1563 were present in the final F5 A x P genetic map produced in Allen et al., [-@allenCharacterizationWheatBreeders2016]. For comparison to the F5 A x P map, the filtered F2 map was used due to its importance in subsequent analyses such as assessing MRD. Clustering between genetic maps was highly consistent; none of the linkage groups from the F2 map contained markers that were present in a linkage group that was identified as a different chromosome in the F5 map (table 3.1). To test the concordance of marker distribution and order between maps, linear regressions for each chromosome were performed predicting the genetic position (cM) of a marker in the F5 map based on the position of the marker in the F2 map. All were highly significant ($p <$ 10-16 for all chromosomes after Bonferroni correction), with over 96% of the variation in the F5 position being explained by the F2 position for every chromosome tested (figure 3.5). Chromosomes 1A, 2D, 3B, 4A, 5A and 7B had perfectly colinear marker bins between maps. In the six cases where marker order did differ between maps, in chromosomes 2A, 3A, 6B and 7A, it did so in adjacent pairs of markers contained within very small centimorgan windows: the mean $\pm$ s.d. distance (cM) between these inverted pairs of markers was 0.33 $\pm$ 0.25 cM. This was expected; it is well known that small genetic distances between markers can influence map ordering algorithms [@hackettEffectsGenotypingErrors2003].

Table 3.1 Comparison of clustering of markers between the F2 Apogee X Paragon genetic map generated here and the F5 map generated previously. The first column indicates the linkage group / chromosome from the F2 genetic map, whereas the subsequent "LG" columns indicate linkage groups that share markers with this F2 linkage group. Columns labelled "Num. markers" indicate the number of markers shared between linkage groups. The final column indicates the number of markers in the F2 linkage group that were not present in the F5 genetic map.

A x P F2 LG	A x P F5 LG 1	A x P F5 Num Markers 1	A x P F5 LG 2	A x P F5 Num Markers 2	A x P F5 LG 3	A x P F5 Num Markers 3	Not present
1A	1A	44	-	-	-	-	12
2A	2A	24	-	-	-	-	19
2D	2D	12	-	-	-	-	13
3A	3A	28	-	-	-	-	8
3B	3B	35	-	-	-	-	12
4A	4A	34	-	-	-	-	12
5A	5A	34	5A2	11	-	-	22
6B	6B	39	-	-	-	-	12
7A	7A	19	7A3	2	-	-	11
7B	7B	27	-	-	-	-	12

Examining the effect of temperature on recombination distribution

To test whether there was a significant difference in the distribution of recombination events between temperature treatments, the mean recombination distance (MRD; see methods) was calculated for each individual in each treatment. We first examined MRD for all chromosomes at once to see if there was a genome-wide effect of temperature on recombination distribution. In the Apogee X Paragon F2 populations, a Kruskal-Wallice test of the MRD in the long arms of all chromosomes reveals a highly significant difference between all four temperature treatments (10°C, 14°C, 26°C and 28°C) (^{2} = 25.63, d.f. = 3, $p < 0.0001$). Likewise, this test was highly significant for the short arms (^{2} = 12.13, d.f. = 3, $p < 0.007$). Temperatures 10°C and 14°C showed evidence of a more distal distribution of recombination events compared to 26°C and 28°C, with mean ± s.d. MRD values (units are percentage of chromosome arm) of 73.7 ± 15.2, 74.34 ± 15.56, 70.19 ± 18.75 and 70 ± 18.03 respectively for long chromosome arms (figure 3.6), and values of 81.48 ± 13.6, 81.19 ± 14.24, 77.87 ± 17.52 and 77.99 ± 16.69 for short chromosome arms (figure 3.6).

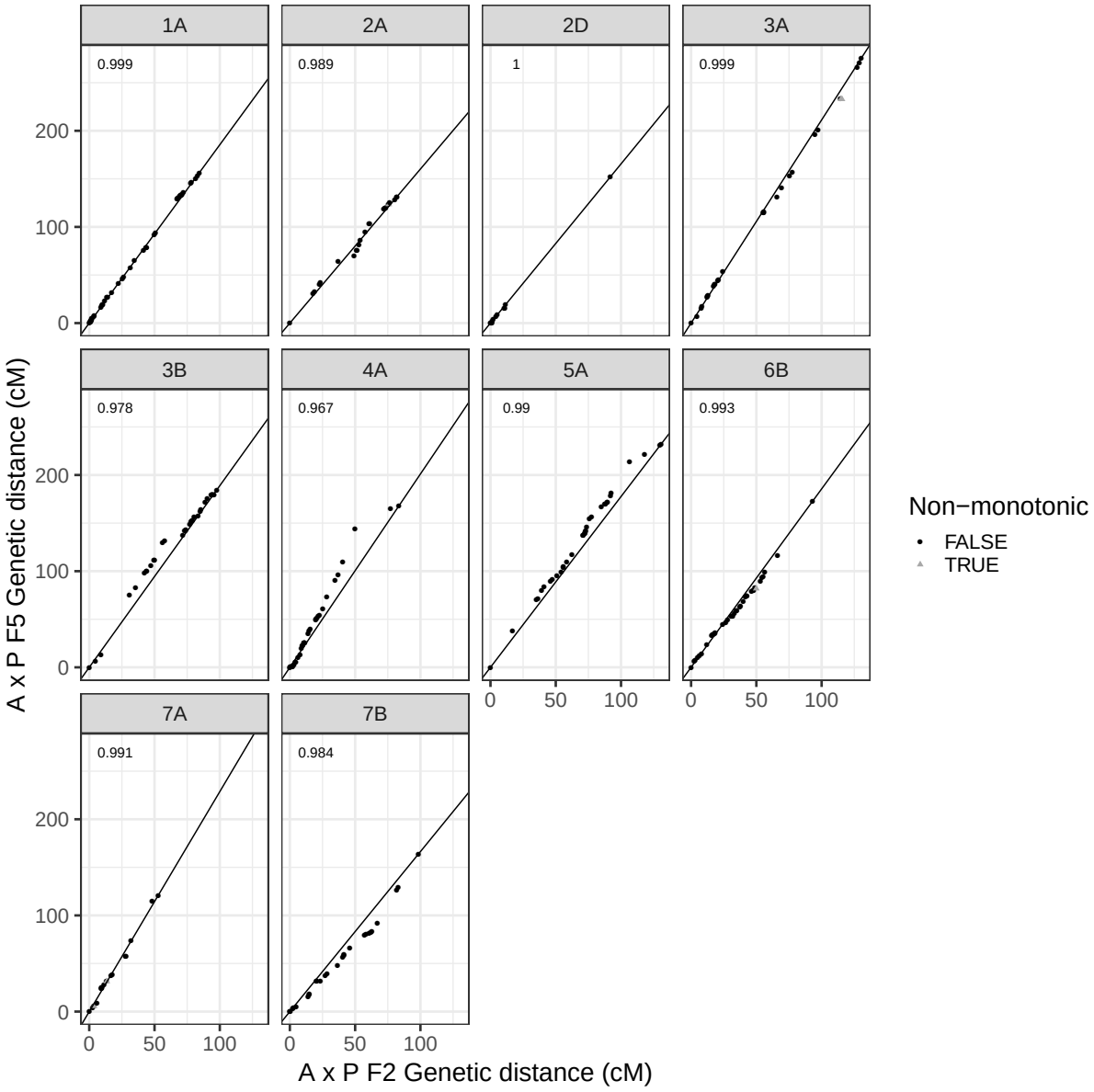


Figure 3.5 Comparison of marker order and distribution between the filtered F2 Apogee X Paragon genetic map and the F5 genetic map produced by Allen et al. (2016). Points represent markers and their genetic positions (cM) in the respective maps. Some of the markers present in the F2 map are not present in the F5 map due to a difference in marker selection procedure between studies. Centimorgan values have therefore been normalized such that map comparisons start at zero whilst retaining inter-marker distances. Deviations from the diagonal line represent differences in the recombination distribution between maps; perfect adherence to the line represents complete coherence between maps in both marker order and marker distribution. Markers that have an inverted order between maps (markers deviating from monotonicity) are represented as grey triangles, whereas markers that are consistent in order represented as black circles. R2 values of linear regressions of the F5 position as a function of the F2 position are shown in the upper left corner of each plot. Chromosomes are labeled in grey panels above each plot.

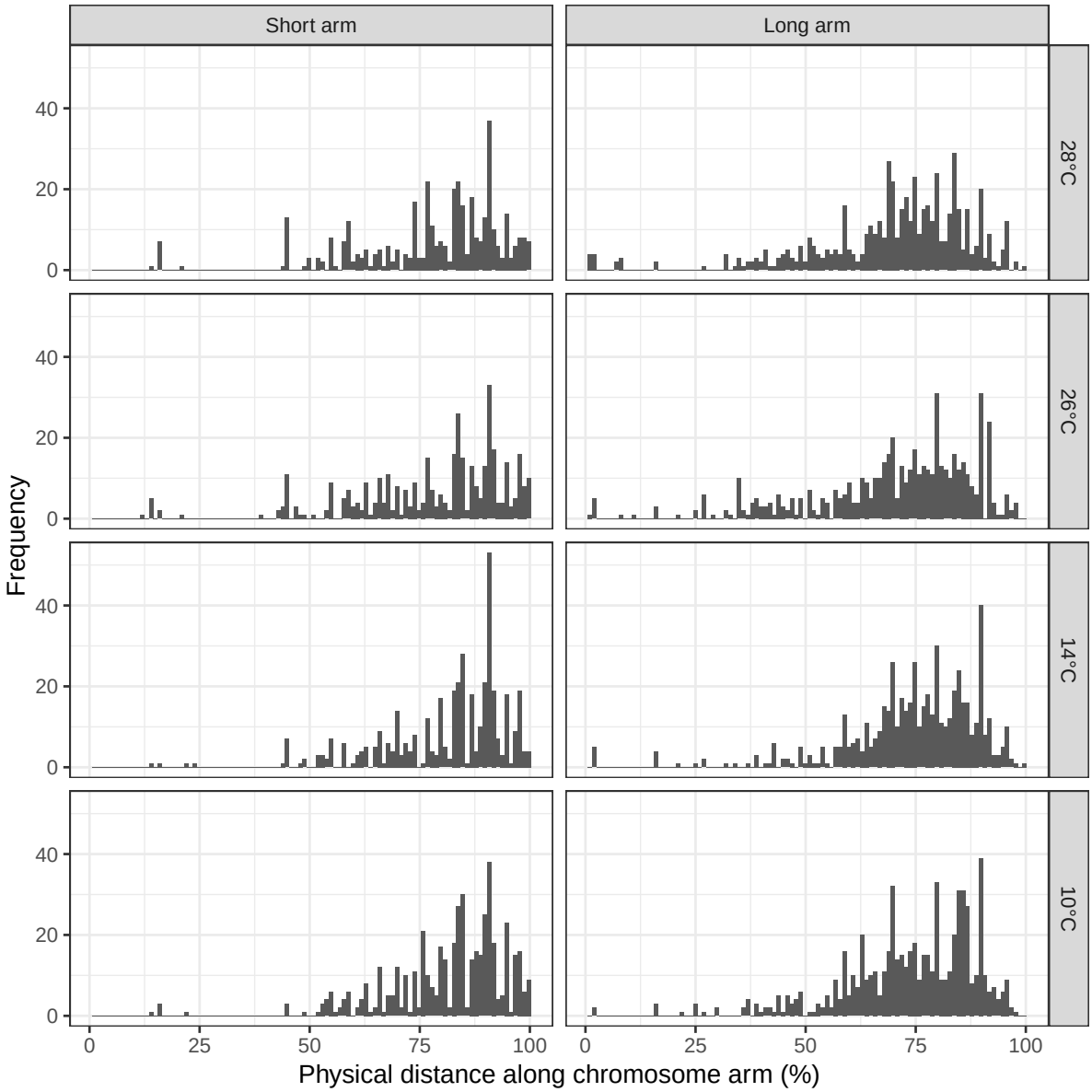


Figure 3.6 Recombination distribution of Apogee X Paragon F2 populations for temperature treatments of 10°C, 14°C, 26°C and 28°C respectively at meiosis. Recombination is measured as the mean distance of recombination events from the centromere of the chromosome in each individual plant to avoid conflation of crossover interference with temperature treatment.

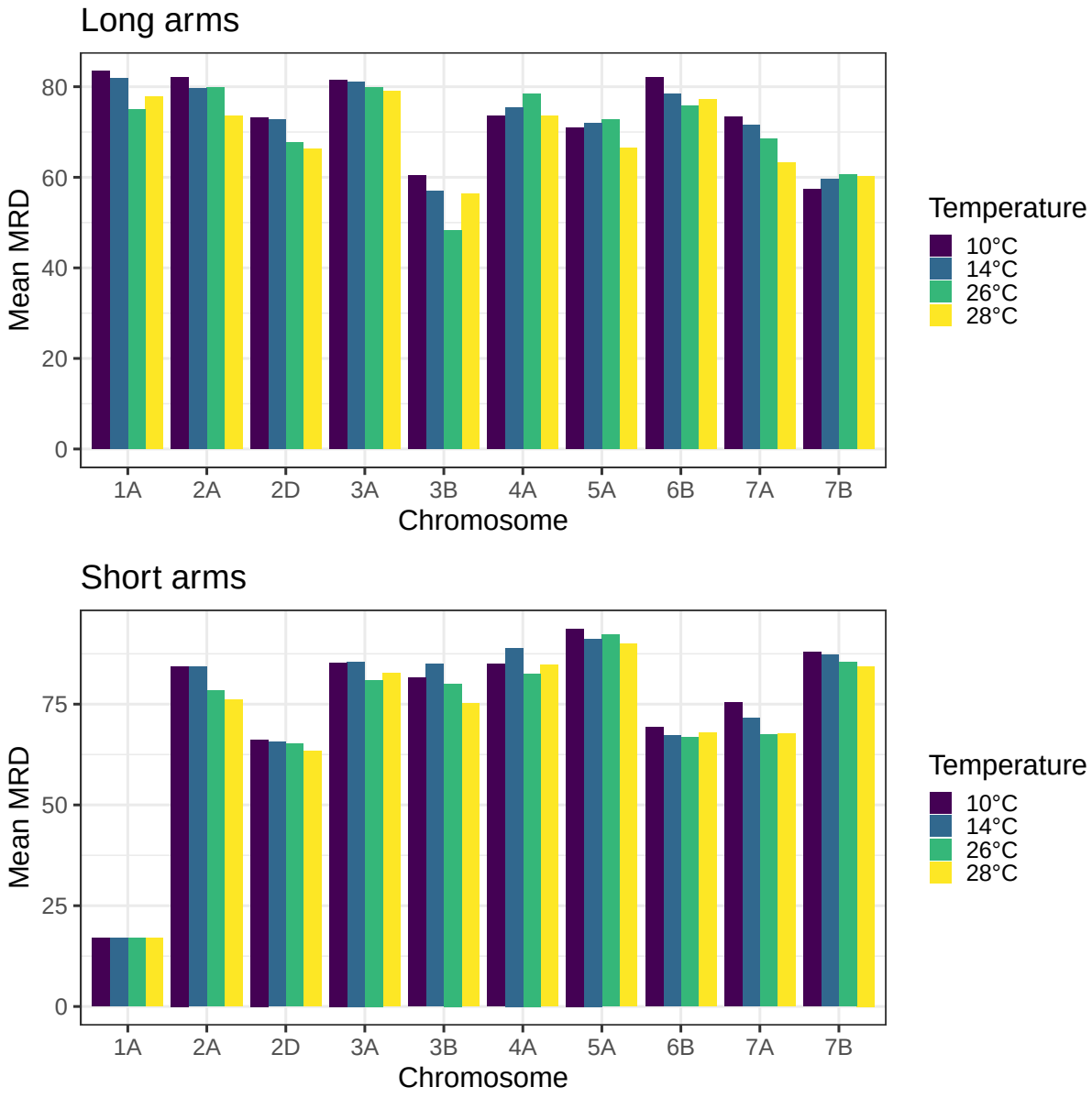


Figure 3.7 Mean MRD for each chromosome for long and short arms.

To examine whether any one treatment was exerting a strong influence on the Kruskal-Wallice test of MRD, individual treatments were removed before performing the test again. MRD remained significantly different between temperature treatments when any of the individual treatments were removed before the test ($p < 0.005$ for all temperatures in the long arm; $p < 0.05$ for all temperatures in the short arm). We then removed pairs of treatments before performing the test again to examine whether high and low temperature treatments were clustered in their effect on the test (table 3.2). For the long chromosome arms, the test only became insignificant when either both low temperature

treatments (10°C and 14°C) were removed (2 = 0.5, d.f. = 1, p = 0.48) or when both high temperature treatments (26°C and 28°C) treatments were removed (2 = 0.75, d.f. = 1, p = 0.37). This was also the case for the short chromosome arms (table 3.2).

Table 3.2 Examining the effect of removal of pairs of temperature treatments before performing the Kruskal-Wallice test on differences in mean recombination distance (MRD). Significant p-values are highlighted in bold. Bonferroni corrections were performed within chromosome arms.

Treatments removed	p-value (long arm)	chisq (long arm)	Bonferroni corrected p-value (long arm)	p-value (short arm)	chisq (short arm)	Bonferroni corrected p-value (short arm)
10°C, 14°C	0.4473	0.5776	1	0.7668	0.08797	1
10°C, 26°C	**0.00001**	19.46	**0.00006**	**0.00627**	7.471	**0.03762**
10°C, 28°C	**0.00094**	10.95	**0.00562**	**0.02589**	4.963	0.1553
14°C, 26°C	**0.0003**	13.06	**0.00181**	**0.00884**	6.854	0.05307
14°C, 28°C	**0.00974**	6.682	0.05842	**0.02821**	4.815	0.1692
26°C, 28°C	0.3716	0.7984	1	0.7525	0.09944	1

In addition to the genome-level analysis of recombination distribution, we also tested for differences in MRD between temperature treatments for individual chromosomes, which differed in their response to changes in temperature during meiosis. The long arms of chromosomes 1A, 3B, as well as the short arms of chromosomes 2A and 7A showed significant differences in MRD between all temperature treatments as determined by a Bonferroni-corrected Kruskal-Wallice test (table 3.3). The long arm of chromosome 1A was most significant, followed by the short arm of 2A (table 3.3).

Table 3.3 Results of Kruskal-Wallice test for difference in MRD between temperature treatments for individual chromosomes. P-values have undergone a Bonferroni correction for multiple testing within chromosome arms. The short arm of 1A is not included as the marker distribution was not sufficient. Significant p-values ($p < 0.05$) are highlighted in bold.

Chromosome	p-value (long arm)	chisq (long arm)	p-value (short arm)	chisq (short arm)	d.f.
1A	**0.00003**	28.61			3
2A	0.2165	9.664	**0.00032**	23.3	3
2D	1	0.5984	1	1.099	3
3A	1	4.971	0.8884	6.281	3
3B	**0.04108**	13.26	0.9845	6.046	3
4A	1	6.199	0.334	8.477	3
5A	0.3848	8.397	1	2.912	3
6B	1	5.256	1	0.5728	3
7A	0.3354	8.701	**0.00485**	17.57	3
7B	1	1.95	1	1.74	3

To investigate these chromosome-level differences in recombination distribution between temperature treatments further, we compared the centimorgan distribution of markers between genetic maps (figures 3.7 - 3.16). Examination of the chromosome 7B, which was the least-significant chromosome in our MRD analysis, showed a reduction

in the number of recombination events between temperatures 10°C and 14°C, with little change in the distributions of events between 14°C, 26°C and 28° treatments (figure 3.16). Chromosome 3B on the other hand had a similar distribution of recombination events between temperatures 10°C and 14°C, before expanding in central regions of the genetic map between temperatures 14°C and 26°C (figure 3.11). These differences highlight the fact that temperature does not act equally on all chromosomes during meiosis.

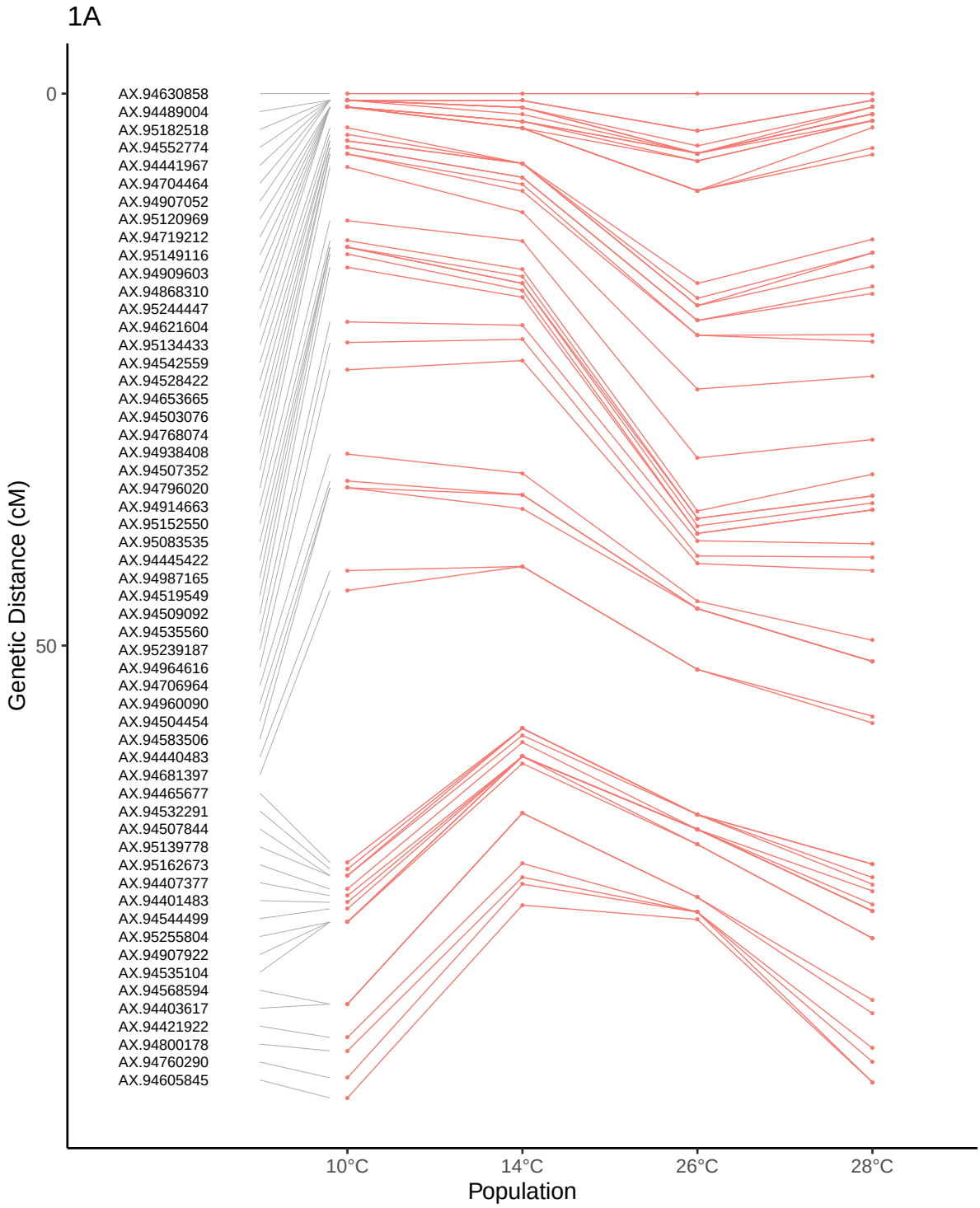


Figure 3.8 Comparisons of genetic maps for four different temperature treatments in Apogee X Paragon F2 populations for chromosome 1A.

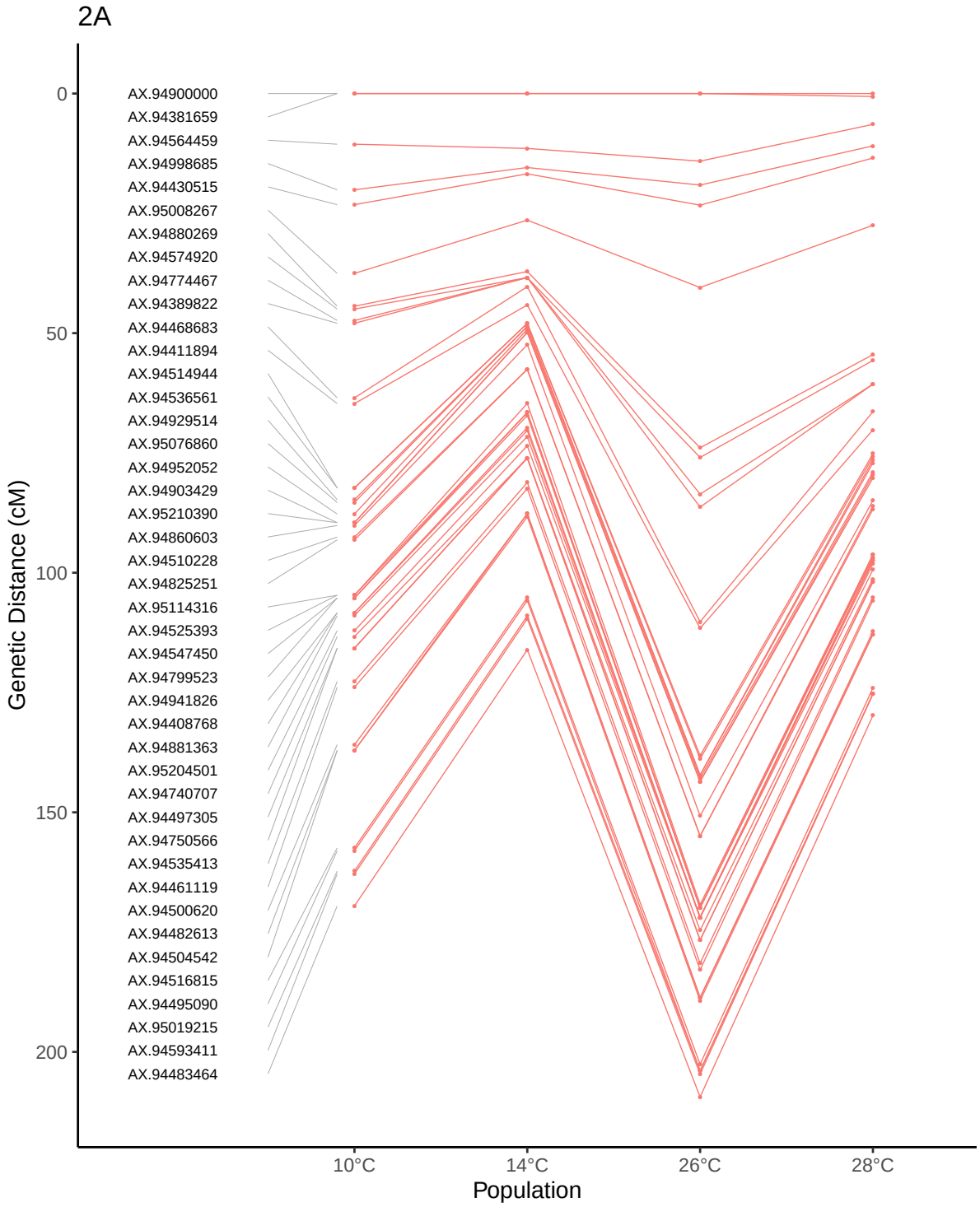


Figure 3.9 Comparisons of genetic maps for four different temperature treatments in Apogee X Paragon F2 populations for chromosome 2A.

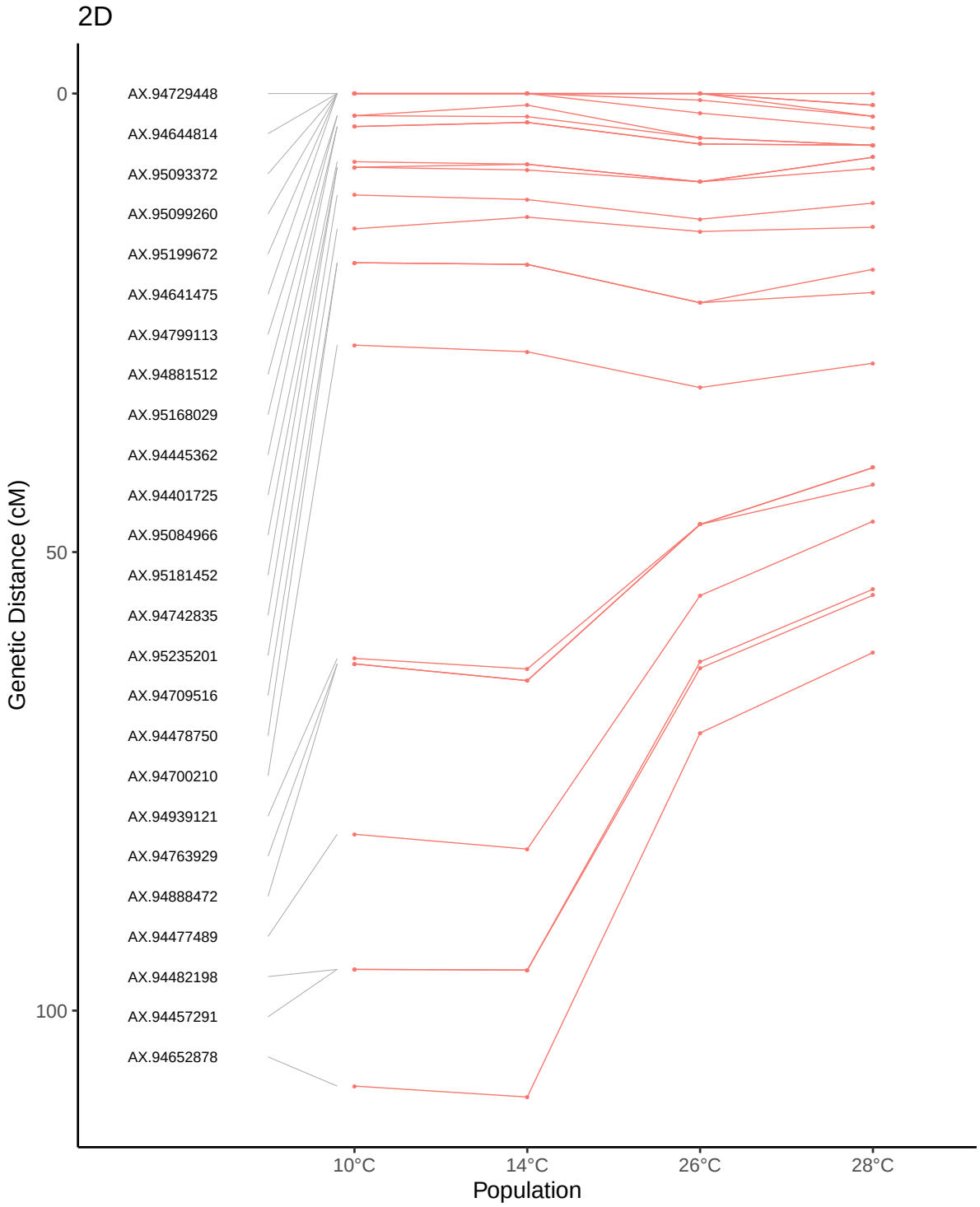


Figure 3.10 Comparisons of genetic maps for four different temperature treatments in Apogee X Paragon F2 populations for chromosome 2D.

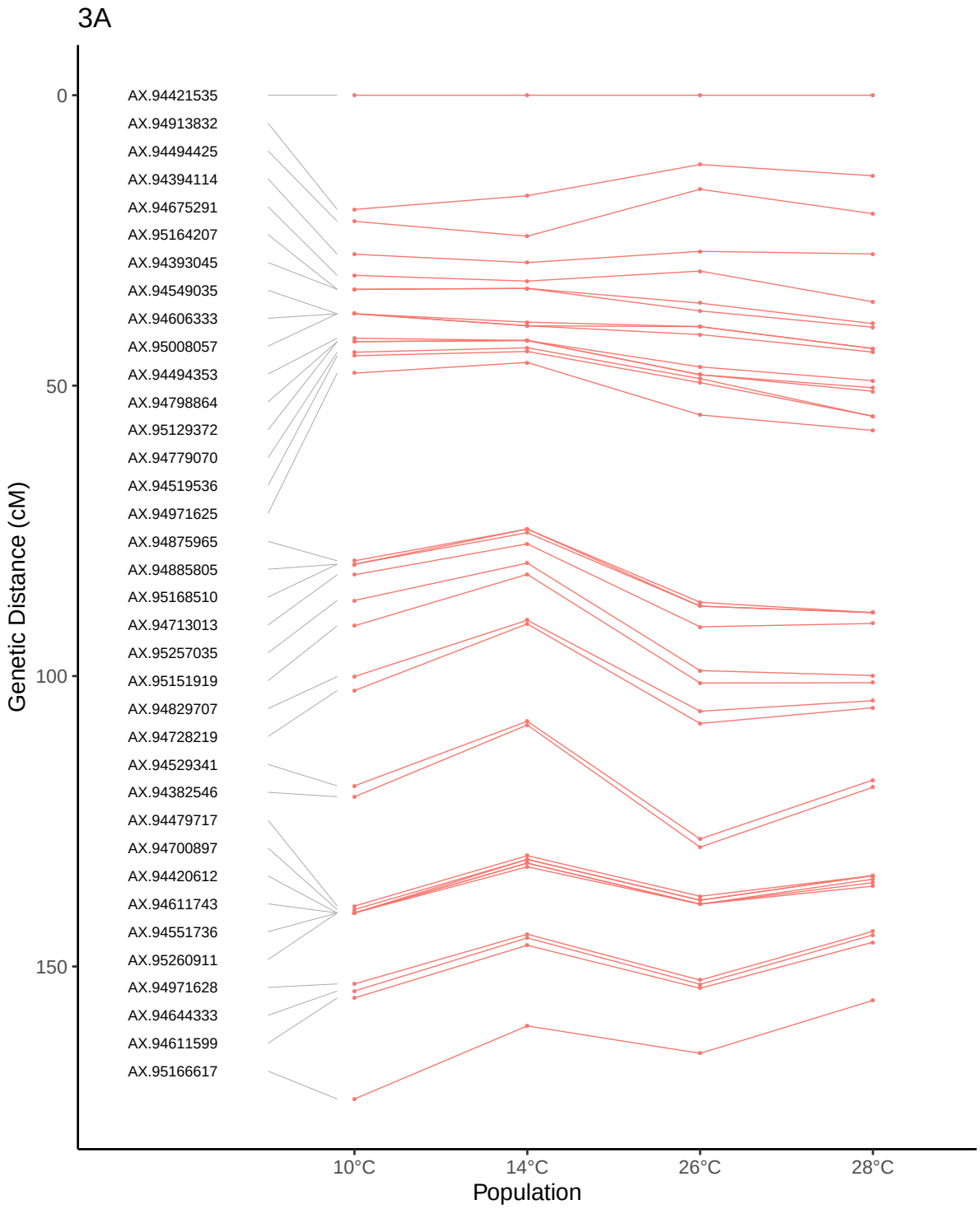


Figure 3.11 Comparisons of genetic maps for four different temperature treatments in Apogee X Paragon F2 populations for chromosome 3A.

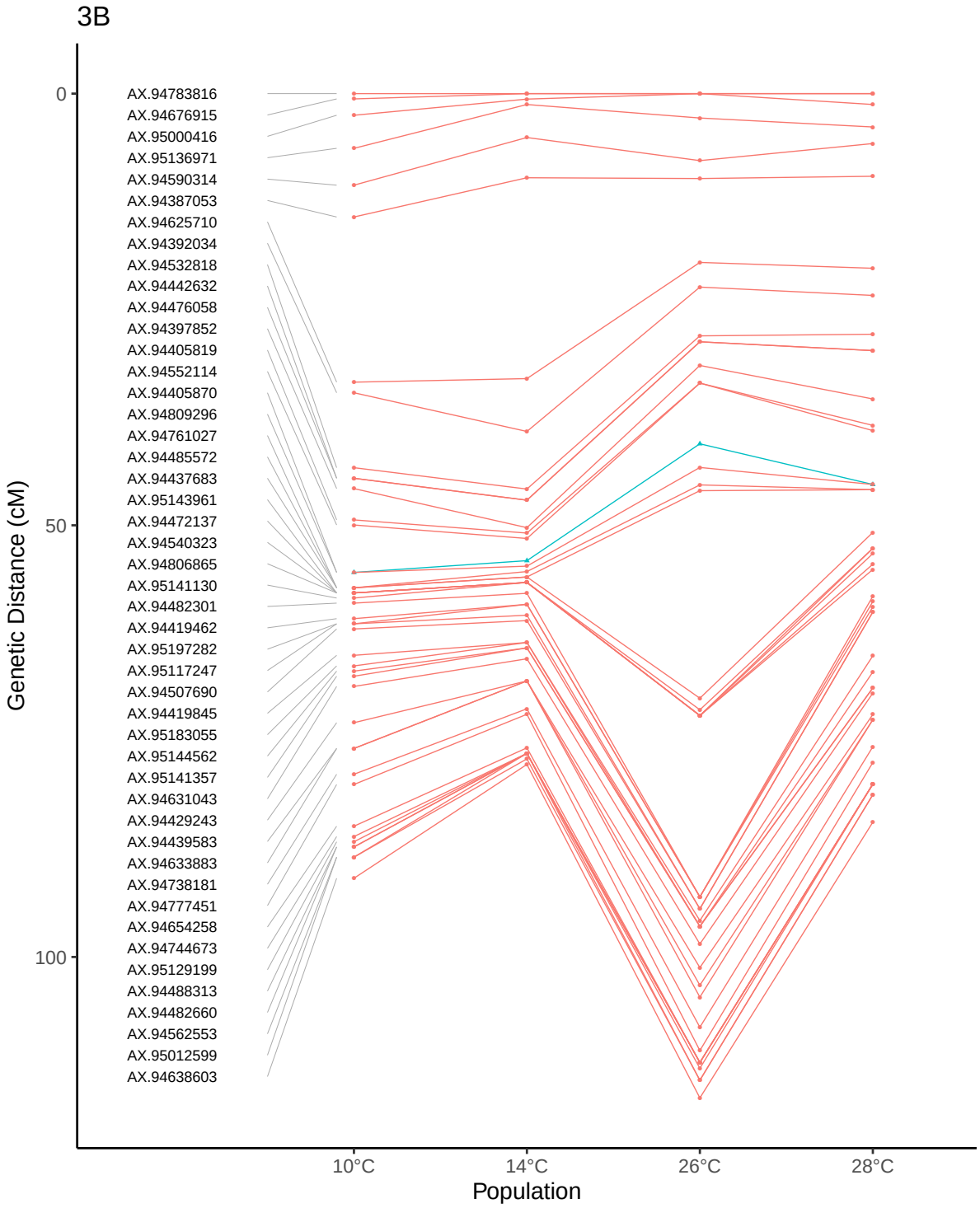


Figure 3.12 Comparisons of genetic maps for four different temperature treatments in Apogee X Paragon F2 populations for chromosome 3B. Markers closest to the centromere (the location of which are taken from (Consortium (IWGSC) et al., 2018) are highlighted as blue triangles.

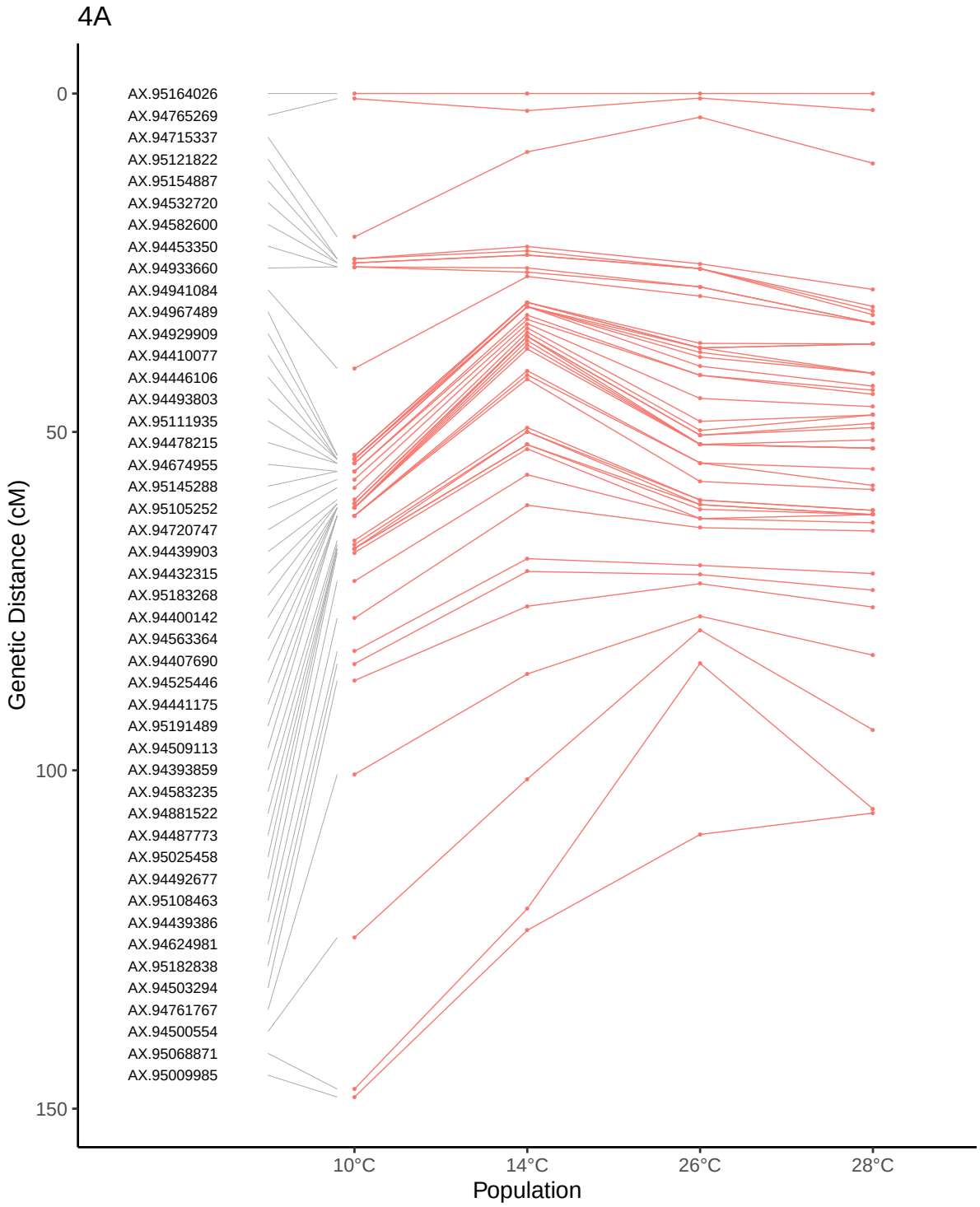


Figure 3.13 Comparisons of genetic maps for four different temperature treatments in Apogee X Paragon F2 populations for chromosome 4A.

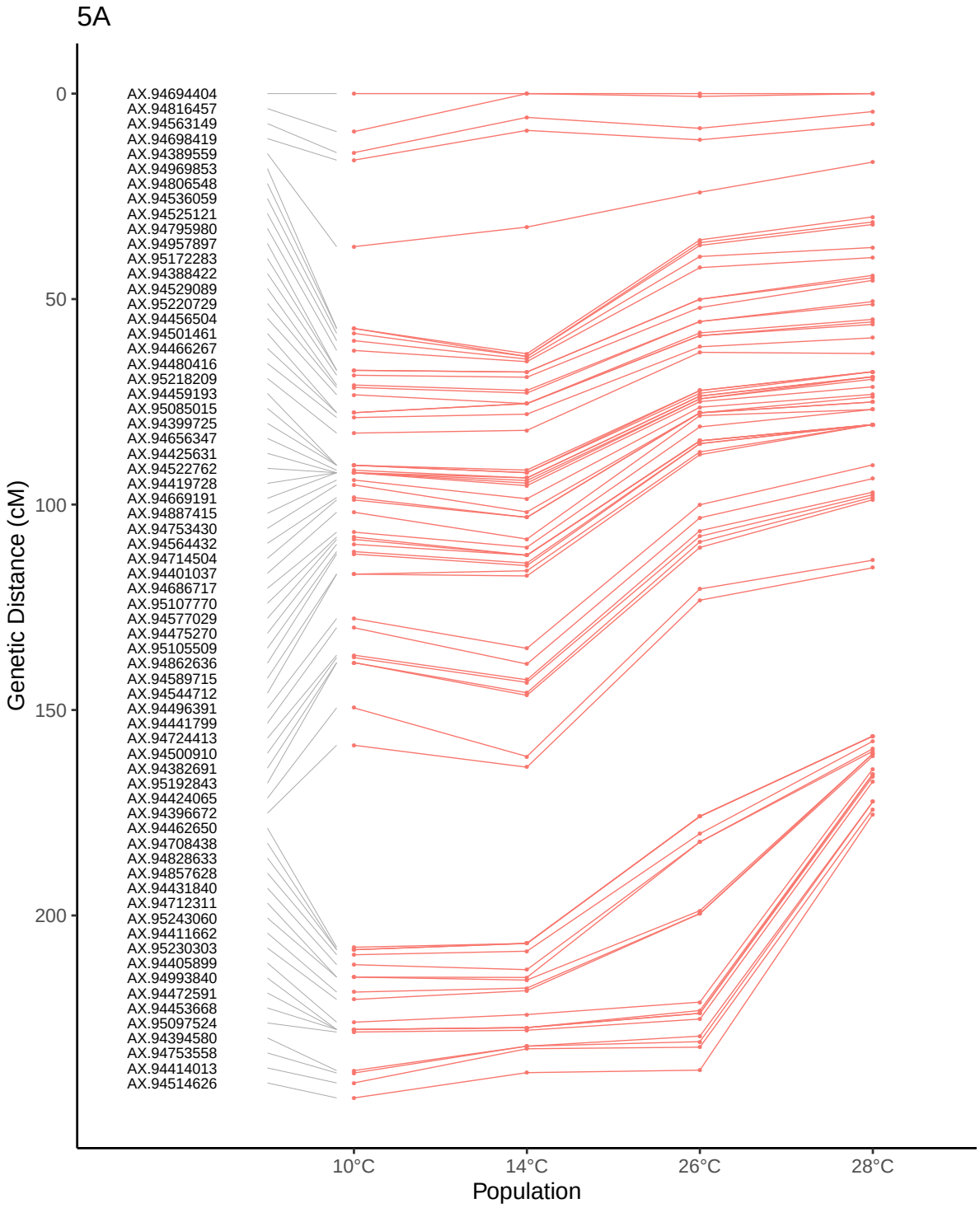


Figure 3.14 Comparisons of genetic maps for four different temperature treatments in Apogee X Paragon F2 populations for chromosome 5A.

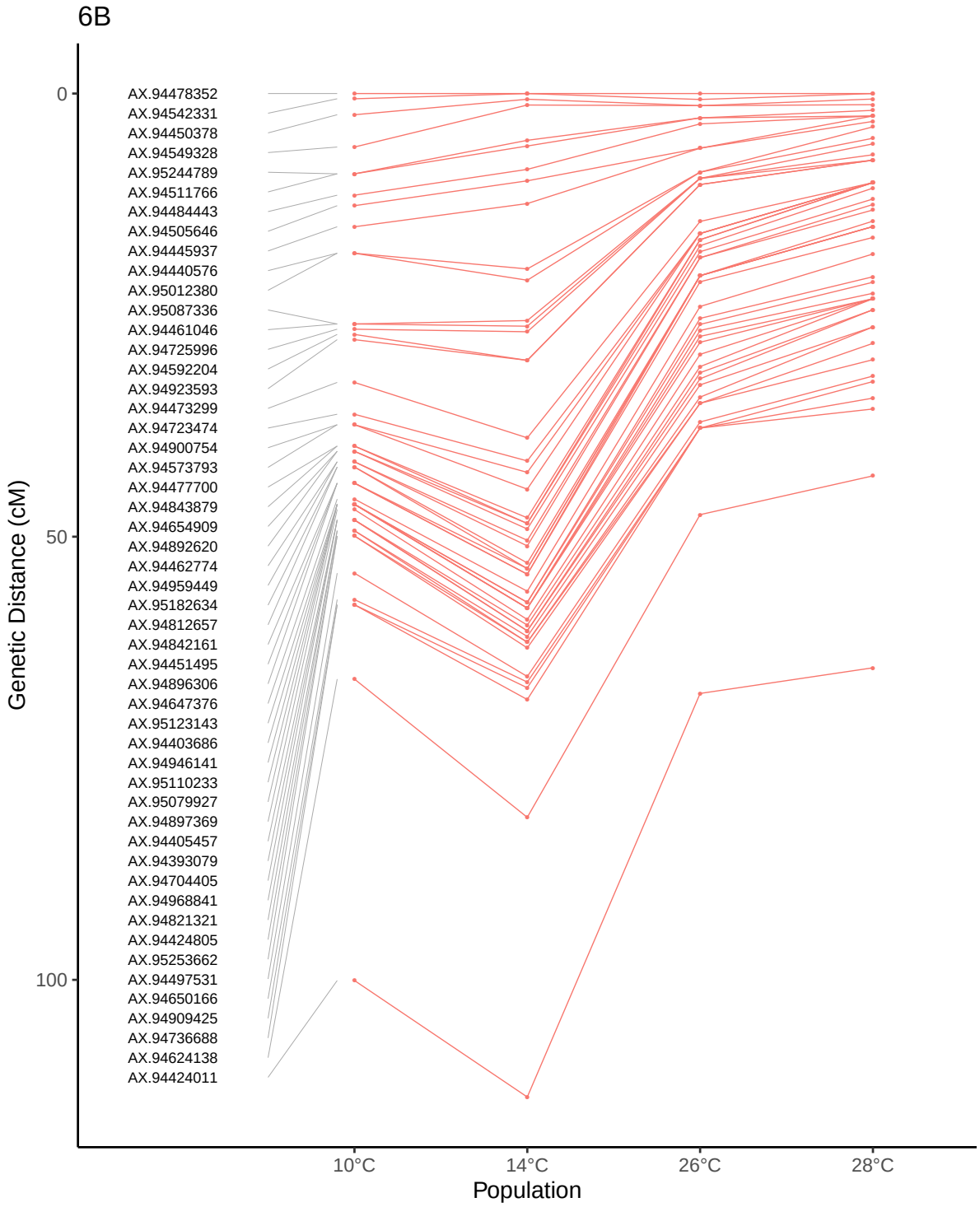


Figure 3.15 Comparisons of genetic maps for four different temperature treatments in Apogee X Paragon F2 populations for chromosome 6B.

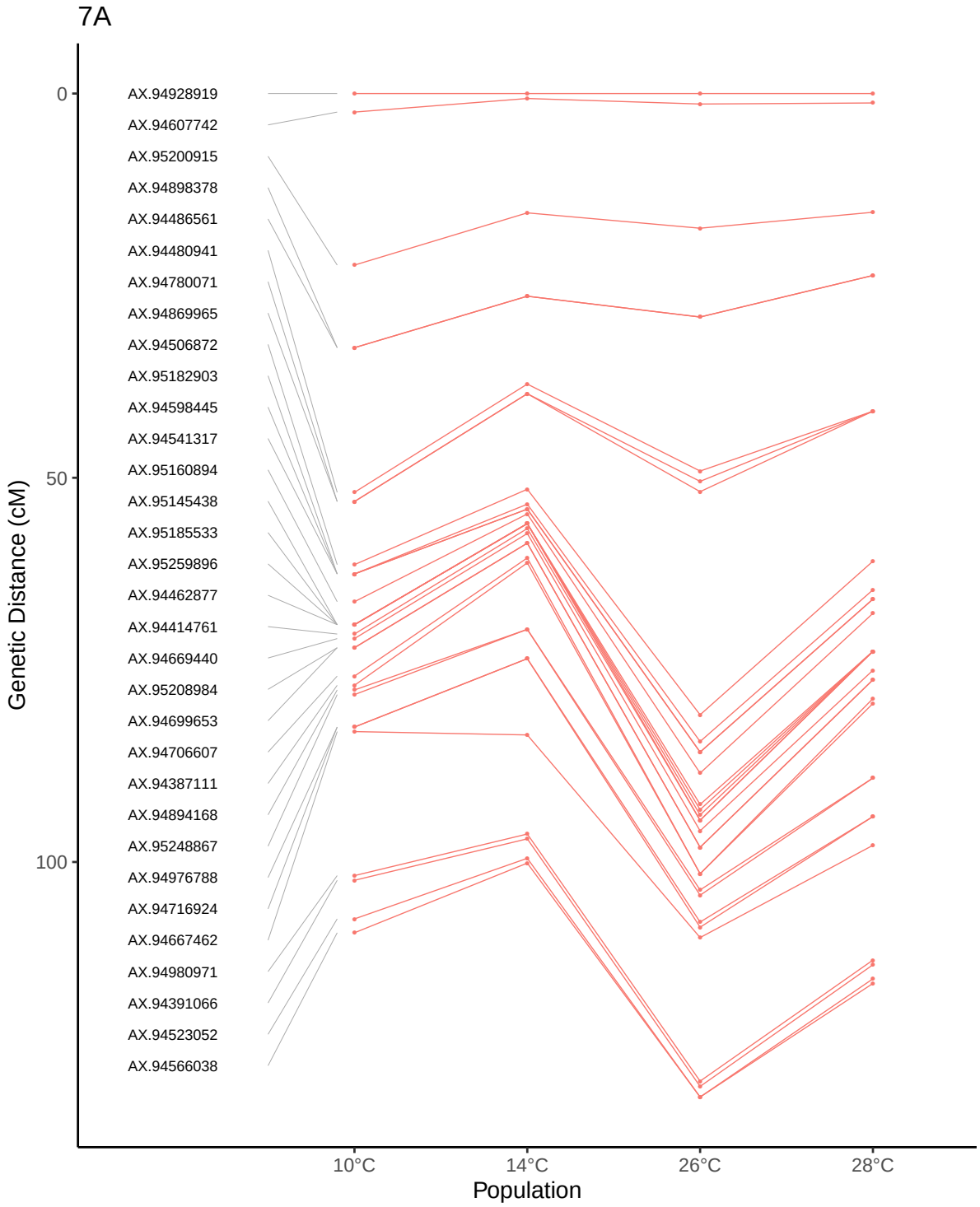


Figure 3.16 Comparisons of genetic maps for four different temperature treatments in Apogee X Paragon F2 populations for chromosome 7A.

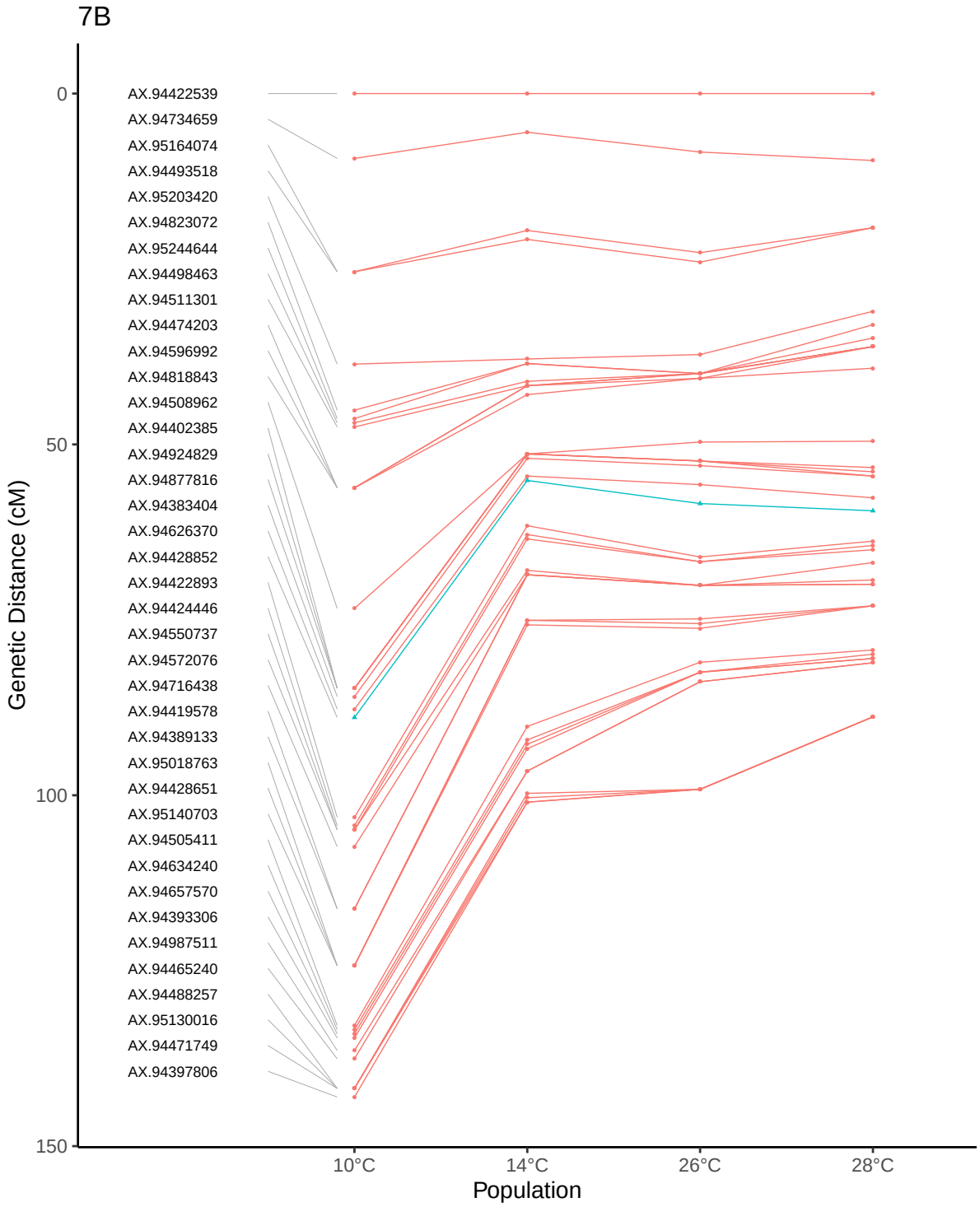


Figure 3.17 Comparisons of genetic maps for four different temperature treatments in Apogee X Paragon F2 populations for chromosome 7B. Markers closest to the centromere (the location of which are taken from (Consortium (IWGSC) et al., 2018) are highlighted as blue triangles.

Whilst we can detect differences MRD between temperature treatments for some chro-

mosomes, it is also important to assess the potential utility of this difference to wheat breeders in terms of its effect on linkage disequilibrium. To do this, we compared chromosome 1A recombination distributions for each treatment (figure 3.17, top four panels), to the distribution of genes along the chromosome (figure 3.17, bottom panel). Higher temperature treatments of 26°C and 28°C appear to induce recombination in regions closer to the centromere from 375 Mb to 461 Mb (markers AX-94621604 and AX-95134433), with a difference of 8.4 cM and 7.7 cM between markers in 26°C and 28°C respectively compared to 1.87 cM and 3.19 cM in 10°C and 14°C treatments. Despite these differences in recombination distribution between treatments, many genes remained highly linked regardless of temperature, for example from 280 to 350 Mb (figure 3.17). Recombination distributions for other chromosomes can be seen in figures 3.18 - 3.26.

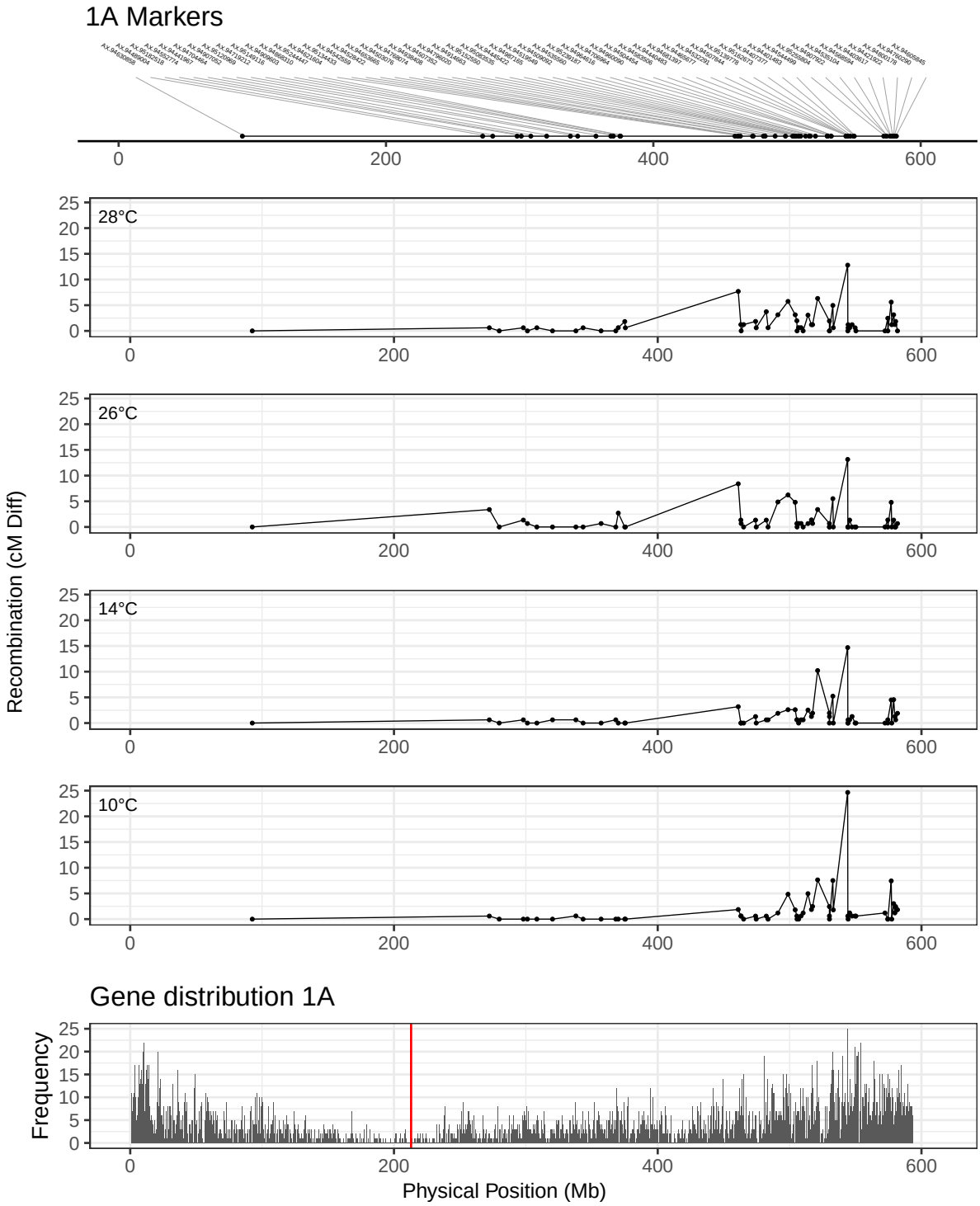


Figure 3.18 Recombination distribution amoung temperature treatments for chromosome 1A with high-confidence gene distribution according to the IWGSC assembly for comparison.

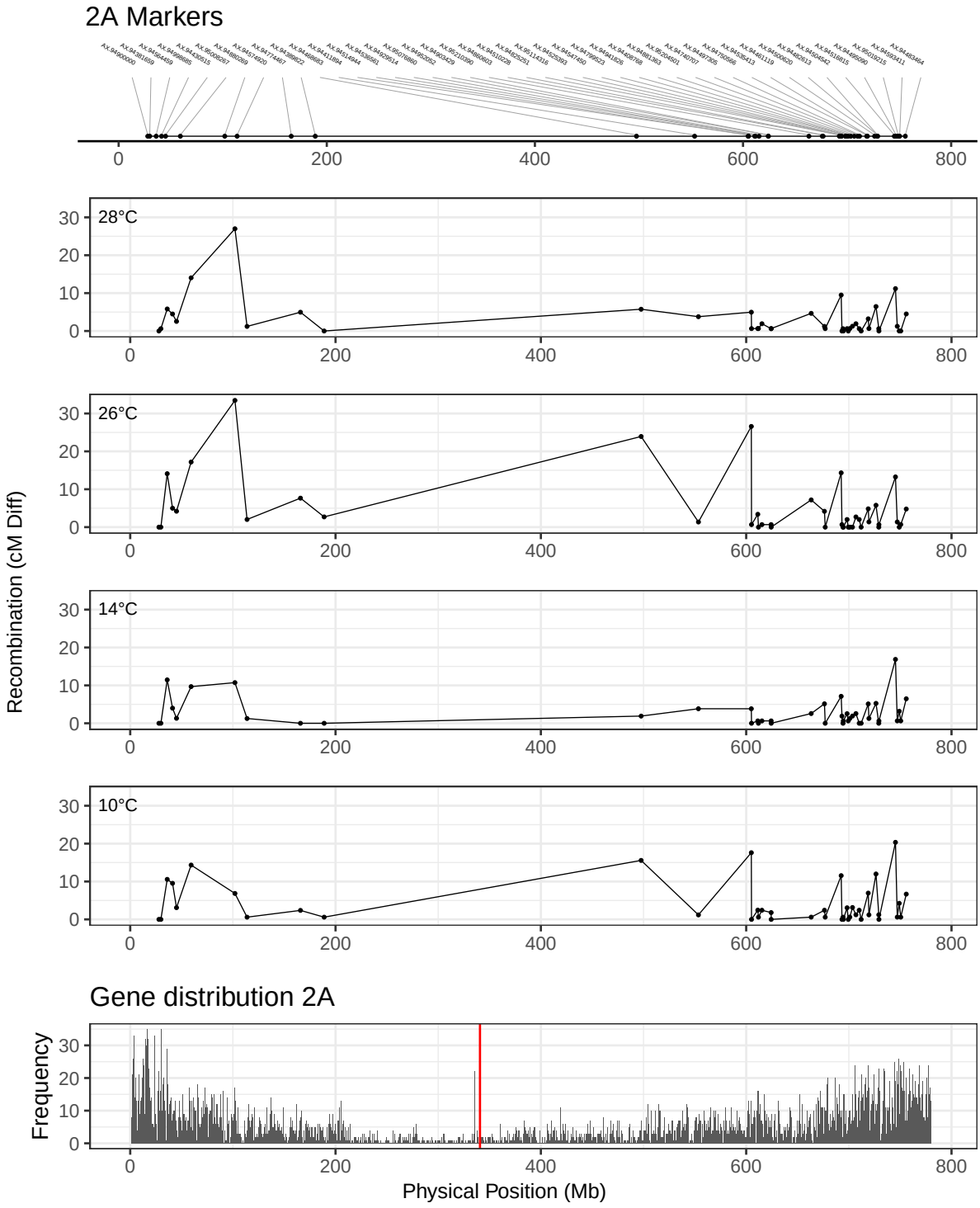


Figure 3.19 Recombination distribution amoung temperature treatments for chromosome 2A with high-confidence gene distribution according to the IWGSC assembly for comparison.

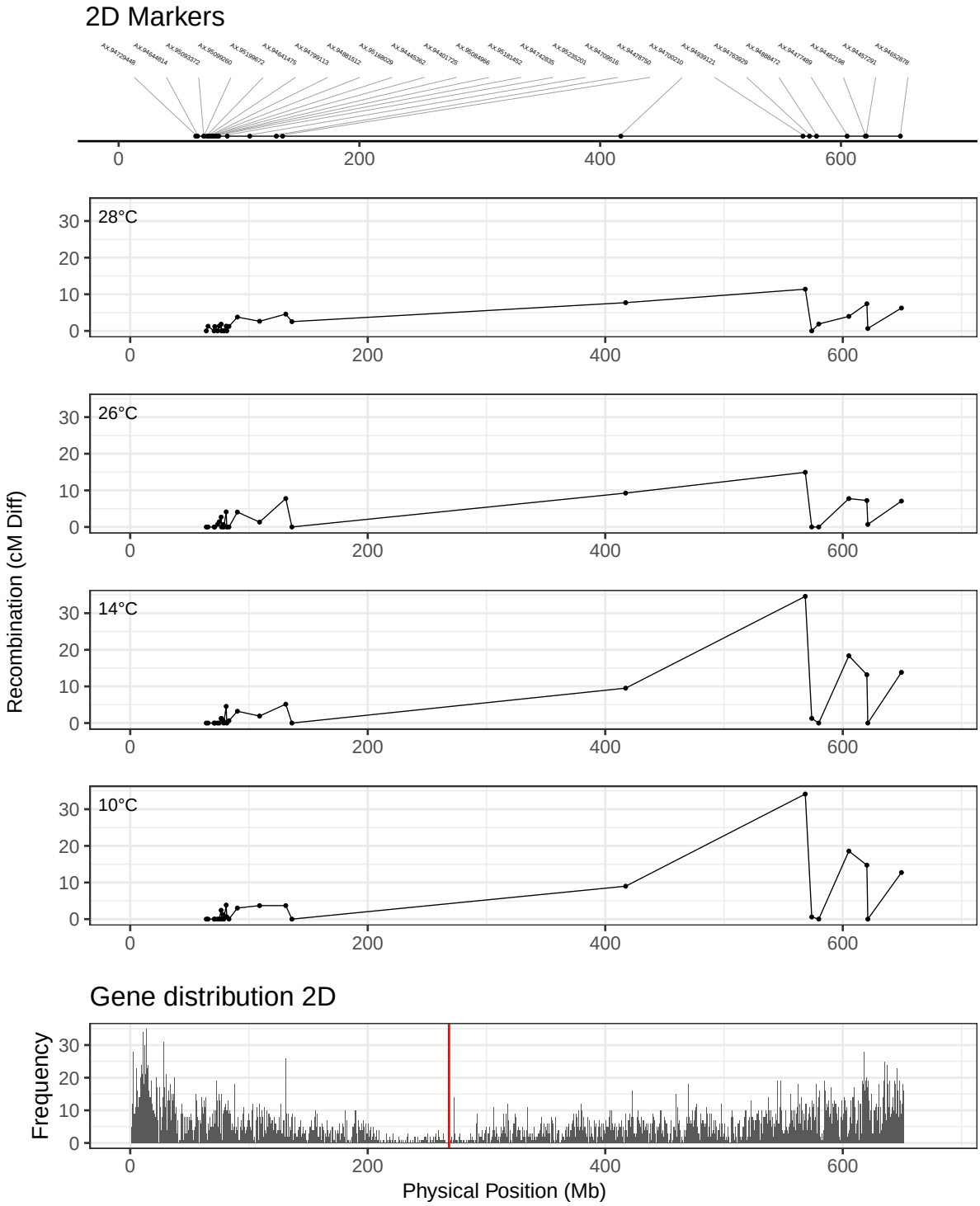


Figure 3.20 Recombination distribution amoung temperature treatments for chromosome 2D with high-confidence gene distribution according to the IWGSC assembly for comparison.

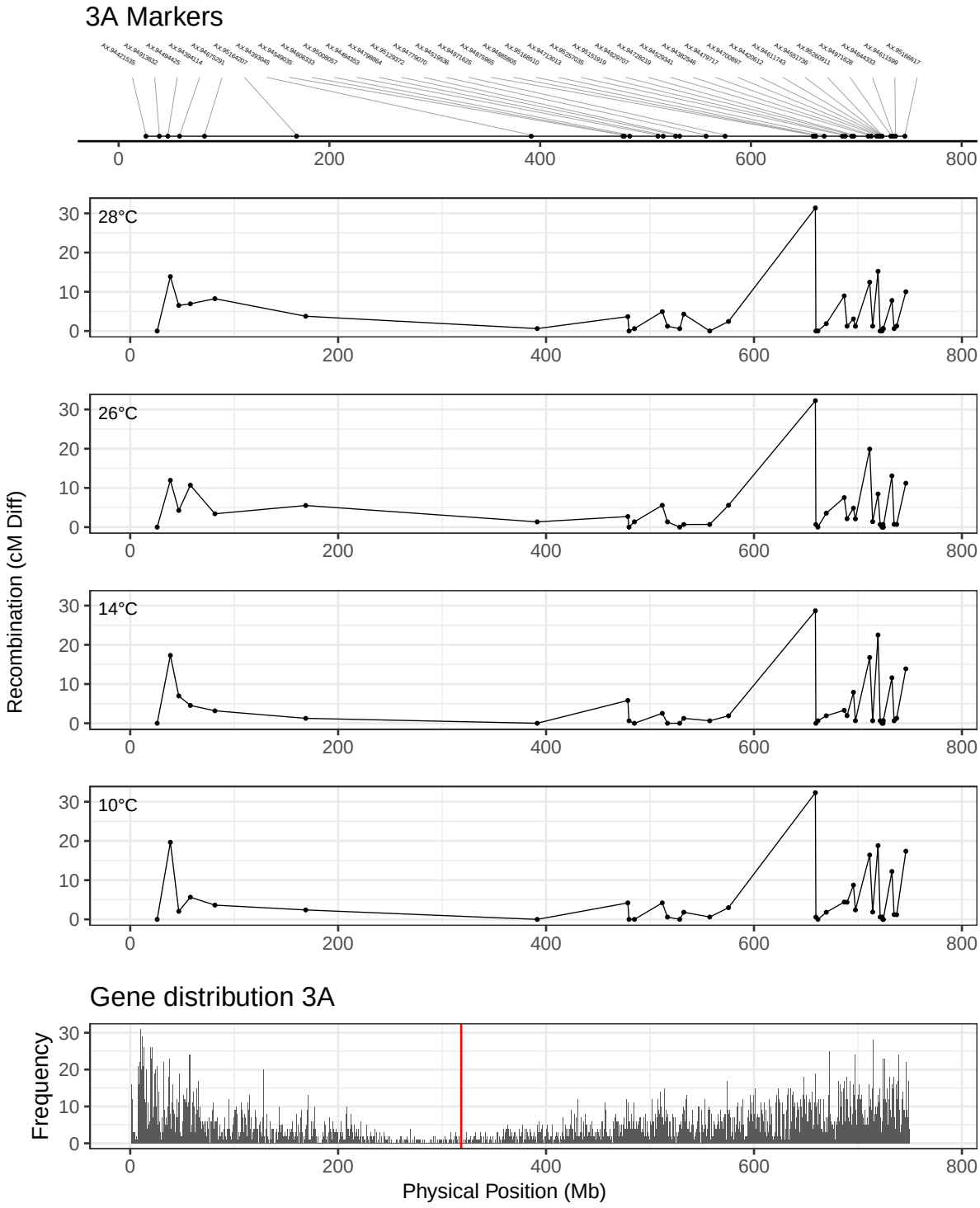


Figure 3.21 Recombination distribution amoung temperature treatments for chromosome 3A with high-confidence gene distribution according to the IWGSC assembly for comparison.

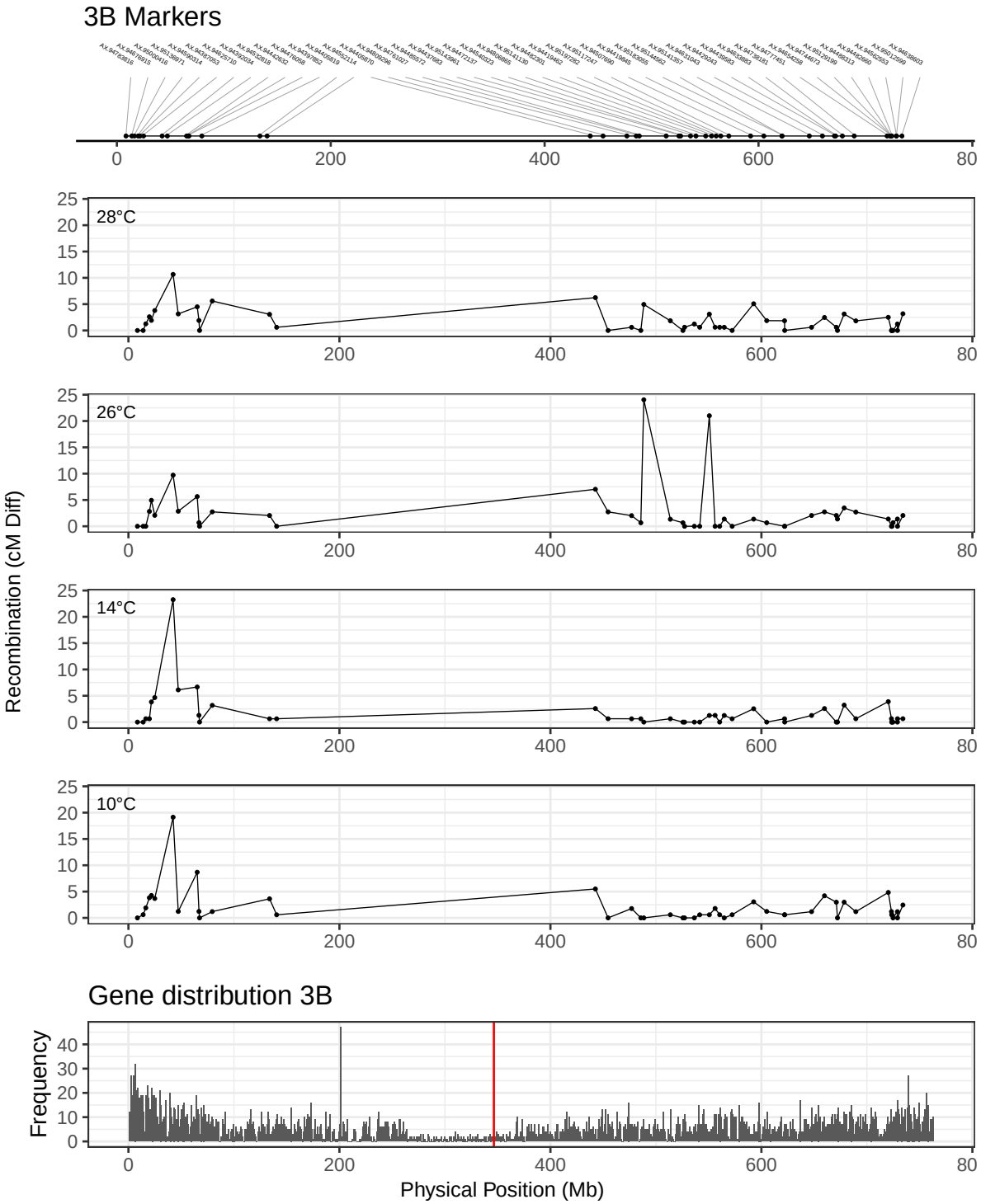


Figure 3.22 Recombination distribution amoung temperature treatments for chromosome 3B with high-confidence gene distribution according to the IWGSC assembly for comparison.

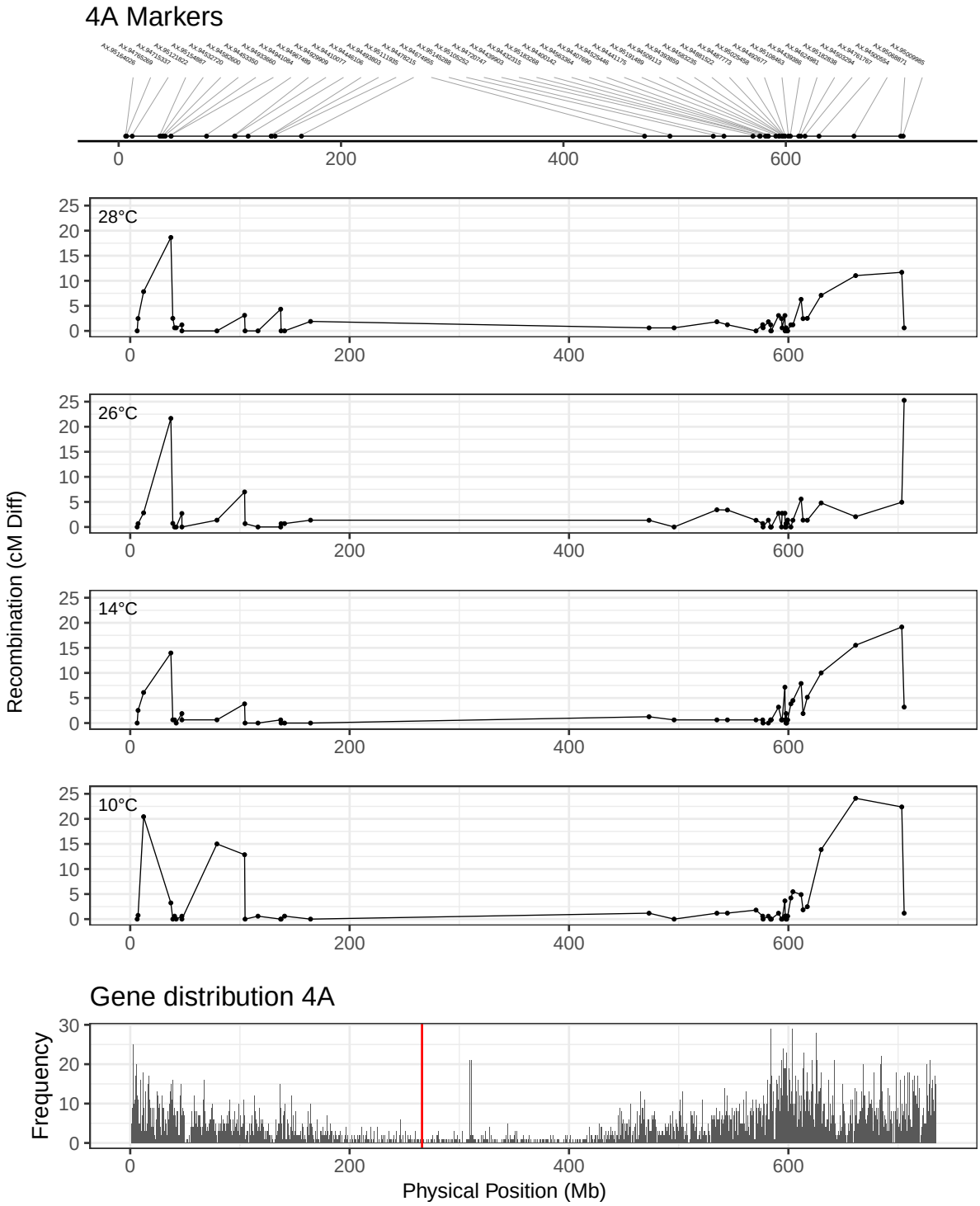


Figure 3.23 Recombination distribution amoung temperature treatments for chromosome 4A with high-confidence gene distribution according to the IWGSC assembly for comparison.

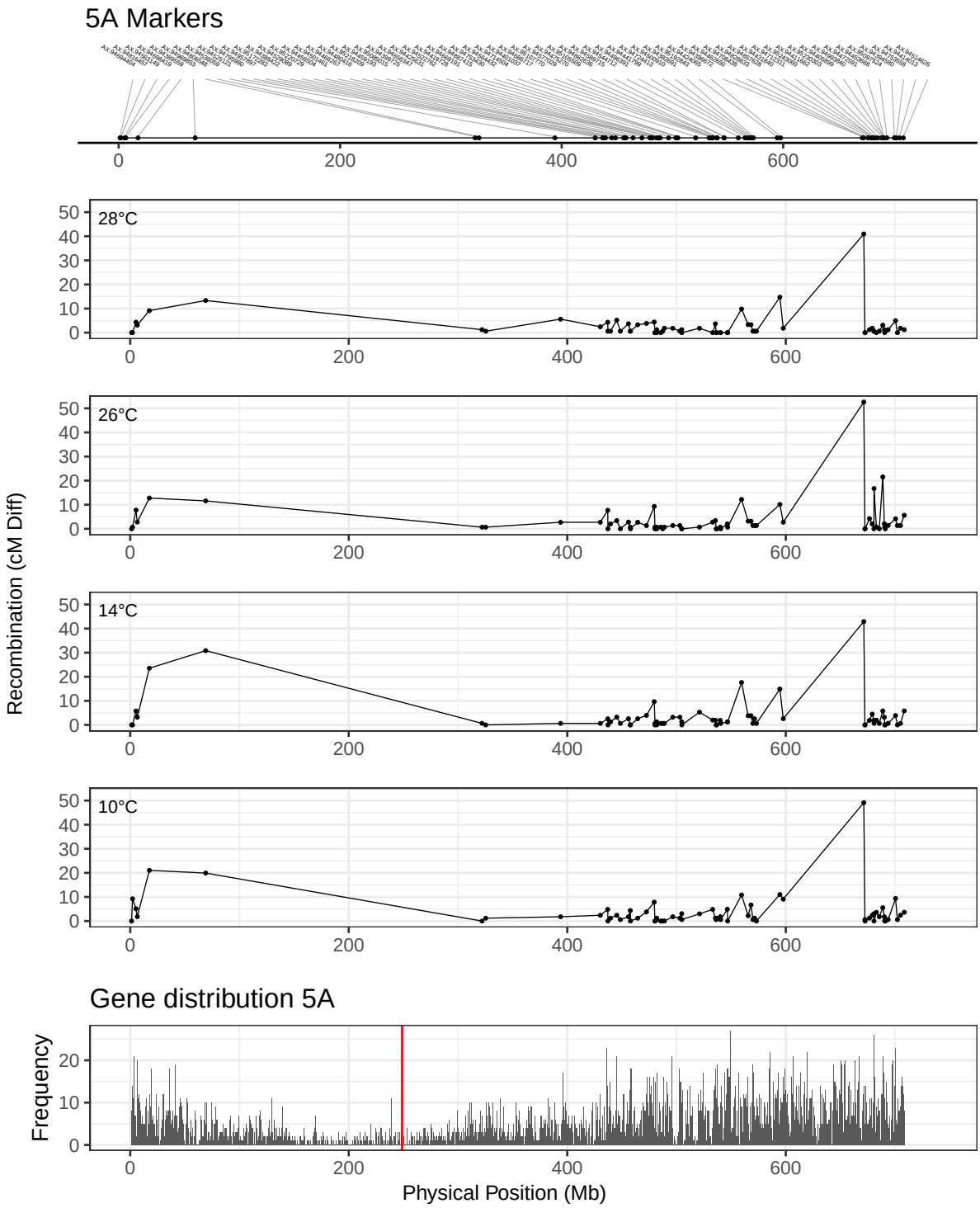


Figure 3.24 Recombination distribution amoung temperature treatments for chromosome 5A with high-confidence gene distribution according to the IWGSC assembly for comparison.

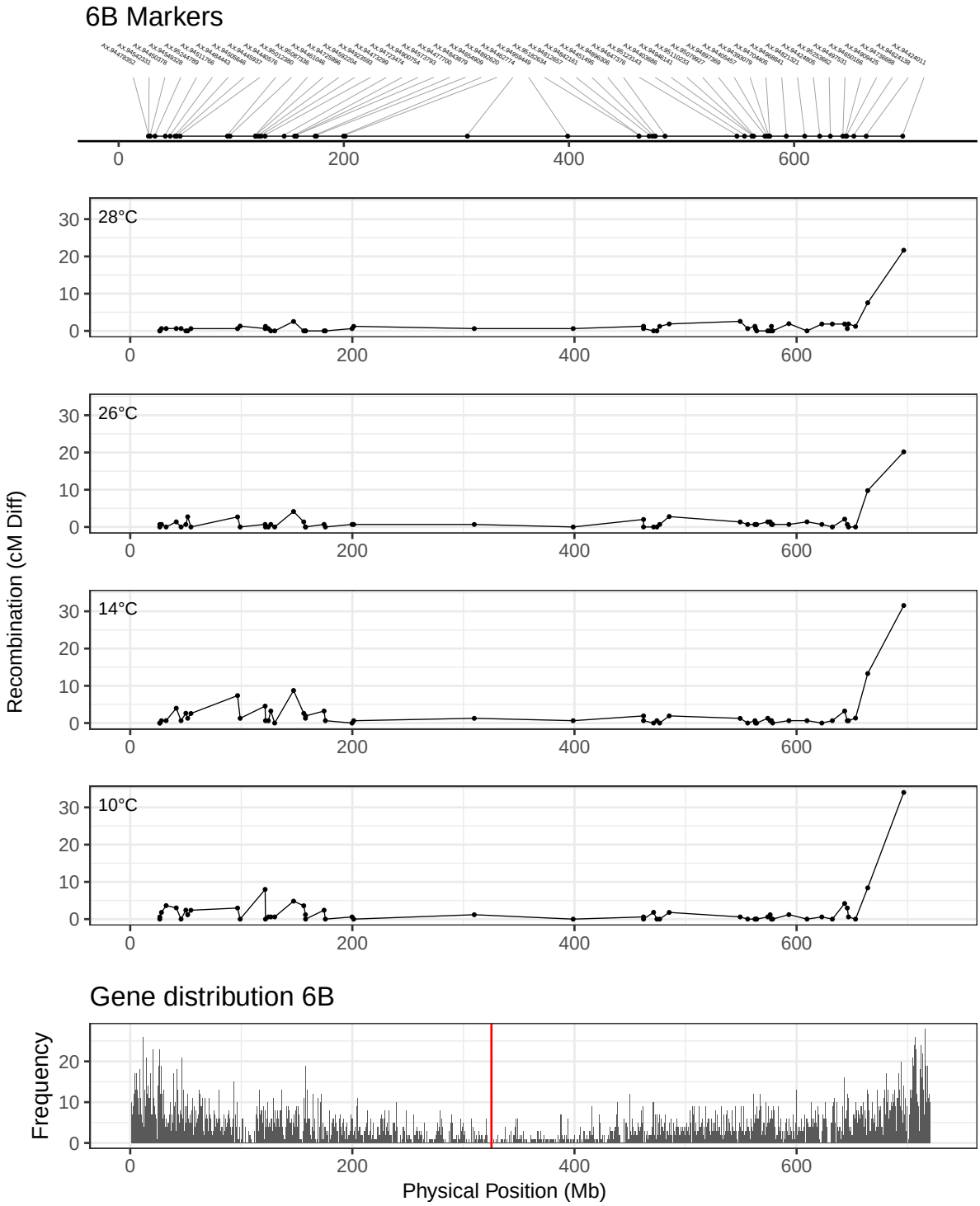


Figure 3.25 Recombination distribution amoung temperature treatments for chromosome 6B with high-confidence gene distribution according to the IWGSC assembly for comparison.

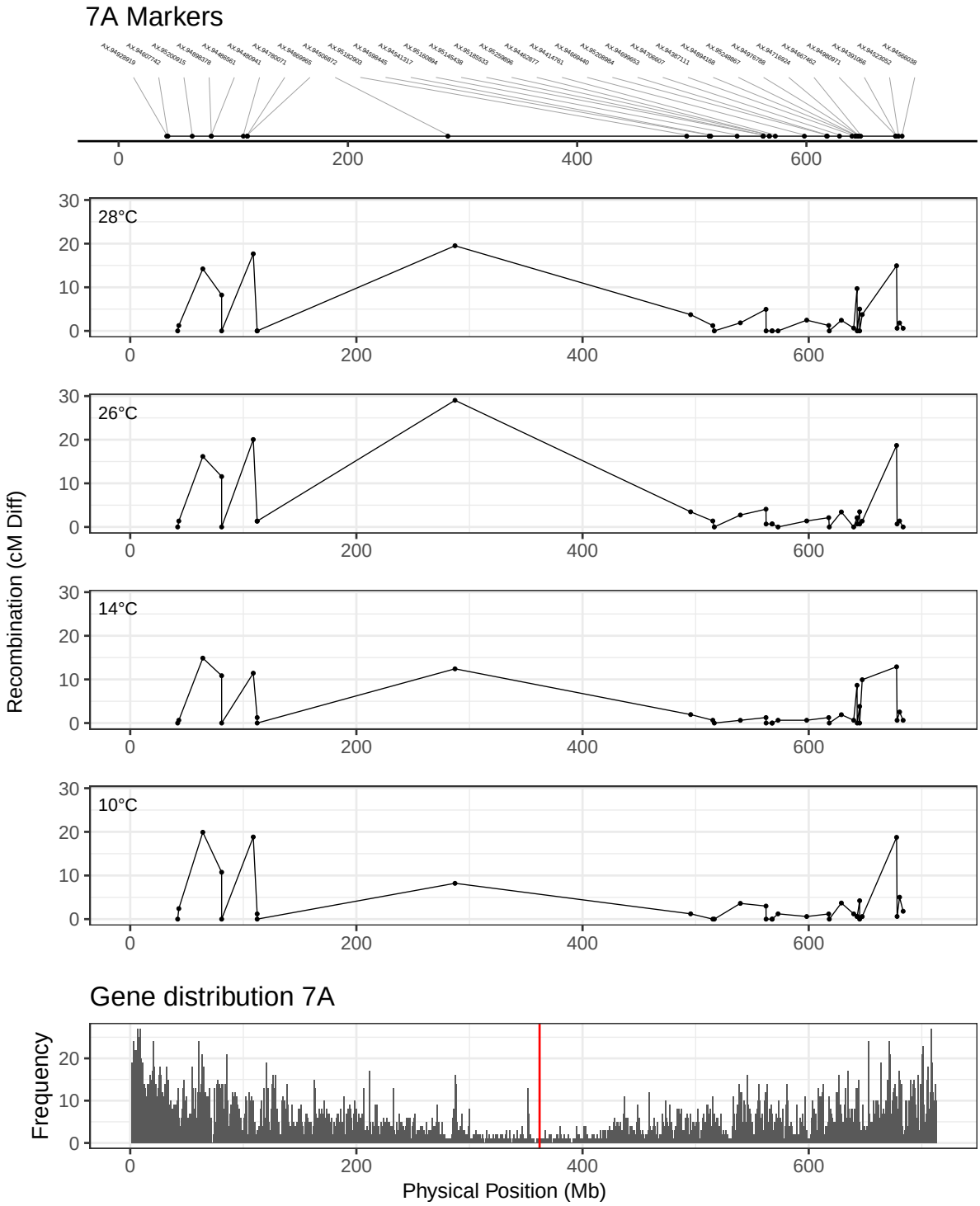


Figure 3.26 Recombination distribution amoung temperature treatments for chromosome 7A with high-confidence gene distribution according to the IWGSC assembly for comparison.

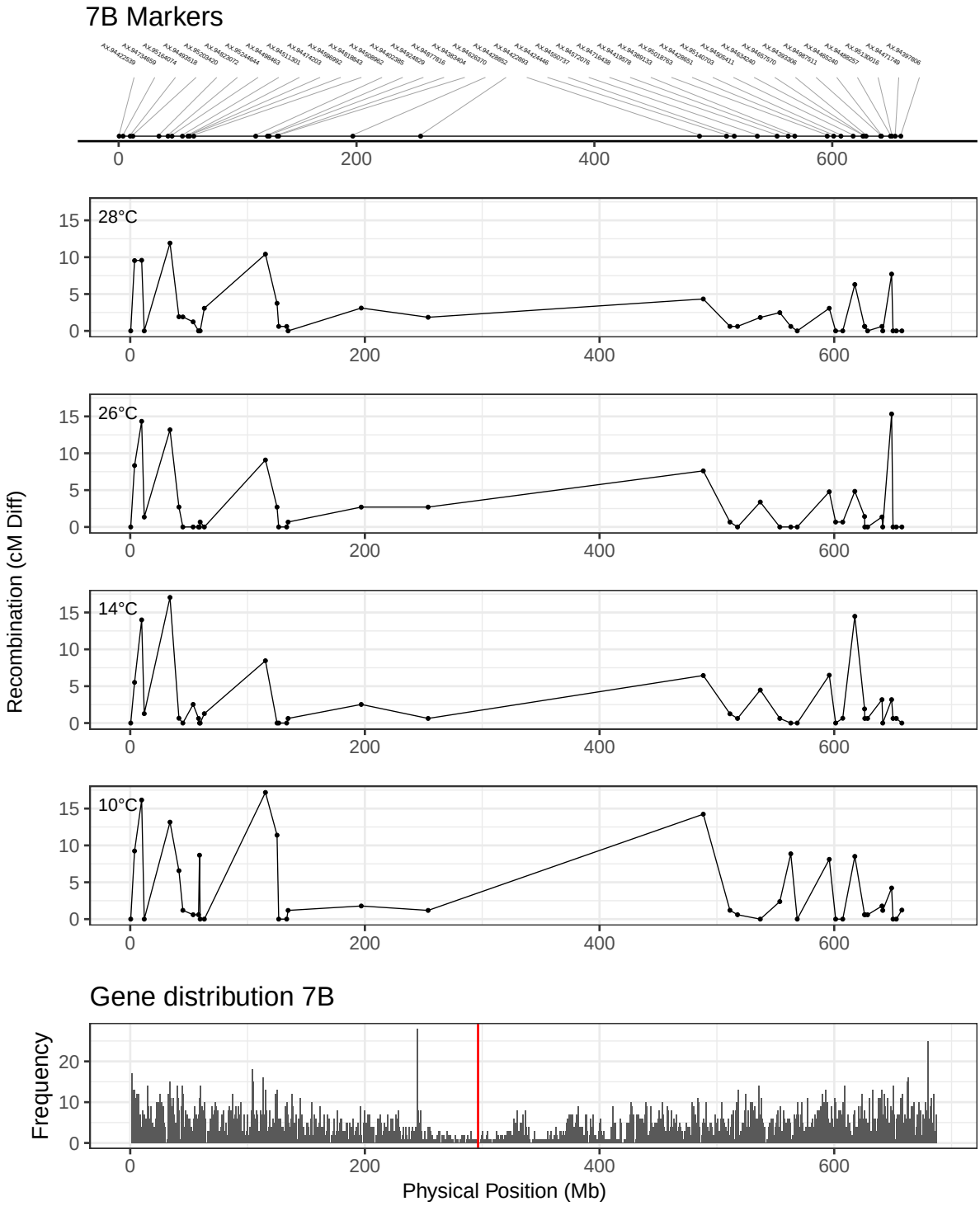


Figure 3.27 Recombination distribution amoung temperature treatments for chromosome 7B with high-confidence gene distribution according to the IWGSC assembly for comparison.

Another area of interest was the potential presence of temperature-dependent recombination hotspots, defined here as regions that contained recombination events in the two high temperature treatments, that also lacked recombination events in both lower temperature treatments. There was evidence of these hotspots on chromosomes 1A, 2A, 2D, 3A, 3B, 4A and 6B (table 3.4). The inter-marker areas of these hotspots contained a total of 868 genes. The hotspot with the largest disparity in number of recombination events between temperatures was found on chromosome 3B, spanning from 485.69 Mb to 488.53 Mb, with differences in cM values between flanking markers of 0, 0, 24.04 and 4.95 at temperatures of 10°C, 14°C, 26°C and 28°C respectively. Inspection of the genotype data revealed that these recombination events were not double events, and visual examination of the SNP cluster plots for both flanking markers revealed clearly delineated genotype clusters in both cases, indicating that genotyping error was unlikely to be the cause of this difference. Annotations for the 19 genes within these two makers include DNA-directed RNA polymerase III subunit RPC3, 50S ribosomal protein L15 (putative), Ubiquitin-conjugating enzyme E2, NAC domain protein and an Auxin-responsive protein. The second most notable hotspot was on chromosome 1A, where the 26°C and 28°C treatments have differences of 2.7 and 0.68 cM respectively between markers AX-94909603 and AX-94868310 at 370 Mb (figure 3.17).

Table 3.4 Regions containing potential temperature-dependent hotspots, defined as having recombination events in both high temperature treatments, whilst lacking recombination events in both low temperature treatments. Also shown are the number of high-confidence genes within the marker interval from the IWGSC assembly annotation. Negative distances from the centromere indicate that the hotspot occurs on the short arm of the chromosome. (continued below)

Chromosome	Marker before	Marker after	Physical position 1st marker	Physical position 2nd marker	Phys difference
3B	AX.94485572	AX.94437683	485.7	488.5	2.84
1A	AX.94909603	AX.94868310	368.2	370.1	1.9
2D	AX.95199672	AX.94641475	73.59	74.94	1.35
1A	AX.95255804	AX.94907922	572.4	574.5	2.13
3A	AX.94606333	AX.95008057	479.8	485	5.22
4A	AX.95111935	AX.94478215	140.7	164.4	23.69
3A	AX.95164207	AX.94393045	168.8	391.5	222.7
2D	AX.94482198	AX.94457291	620.3	621.1	0.8
6B	AX.94451495	AX.94896306	474.2	476.7	2.5
6B	AX.95123143	AX.94403686	549.2	556	6.76
1A	AX.94445422	AX.94987165	505.7	506.9	1.18
6B	AX.94946141	AX.95110233	562.5	563.1	0.58
2A	AX.94514944	AX.94536561	605	605.2	0.18

10°C cM difference	14°C cM difference	26°C cM difference	28°C cM difference	Number of genes	Distance from centromere
0	0	24.04	4.95	19	140.7
0	0	2.7	0.63	28	156.2
0	0	1.42	1.31	16	-194.2
0	0	1.36	2.47	37	360.4
0	0	1.36	0.6	40	164
0	0	1.36	1.89	117	-113.5
0	0	1.34	0.62	508	-38.24
0	0	0.7	0.66	20	352.2
0	0	0.69	1.23	18	150.2
0	0	0.68	0.62	44	227.3
0	0	0.68	0.64	11	293.3
0	0	0.67	0.62	3	237.5
0	0	0.67	0.64	7	264.4

Examining the effects of temperature on recombination frequency

Recombination frequency varied between temperature treatments, with total map lengths of 1377.43, 1213.2, 1279.8 and 1071.65 cM for 10°C, 14°C, 26°C and 28°C degree treatments respectively. Chromosome 5A had the highest number of recombination events in all temperature treatments (figure 3.27). An ANOVA of recombination frequency in individuals across all chromosomes between populations revealed significant differences between temperature treatments 10°C and 14°C ($p < 0.0001$), 10°C and 28°C ($p < 0.00001$) as well as 26°C and 28°C ($p < 0.001$) as determined by a Tukey post-hoc test (figure 3.28). The frequency of recombination events follows a U-shaped pattern between temperatures 10°C, 14°C and 26°C, before declining between 26°C and 28°C (figure 3.28). The mean ± s.d. number of recombination events across all individuals was 21.83 ± 5.32.

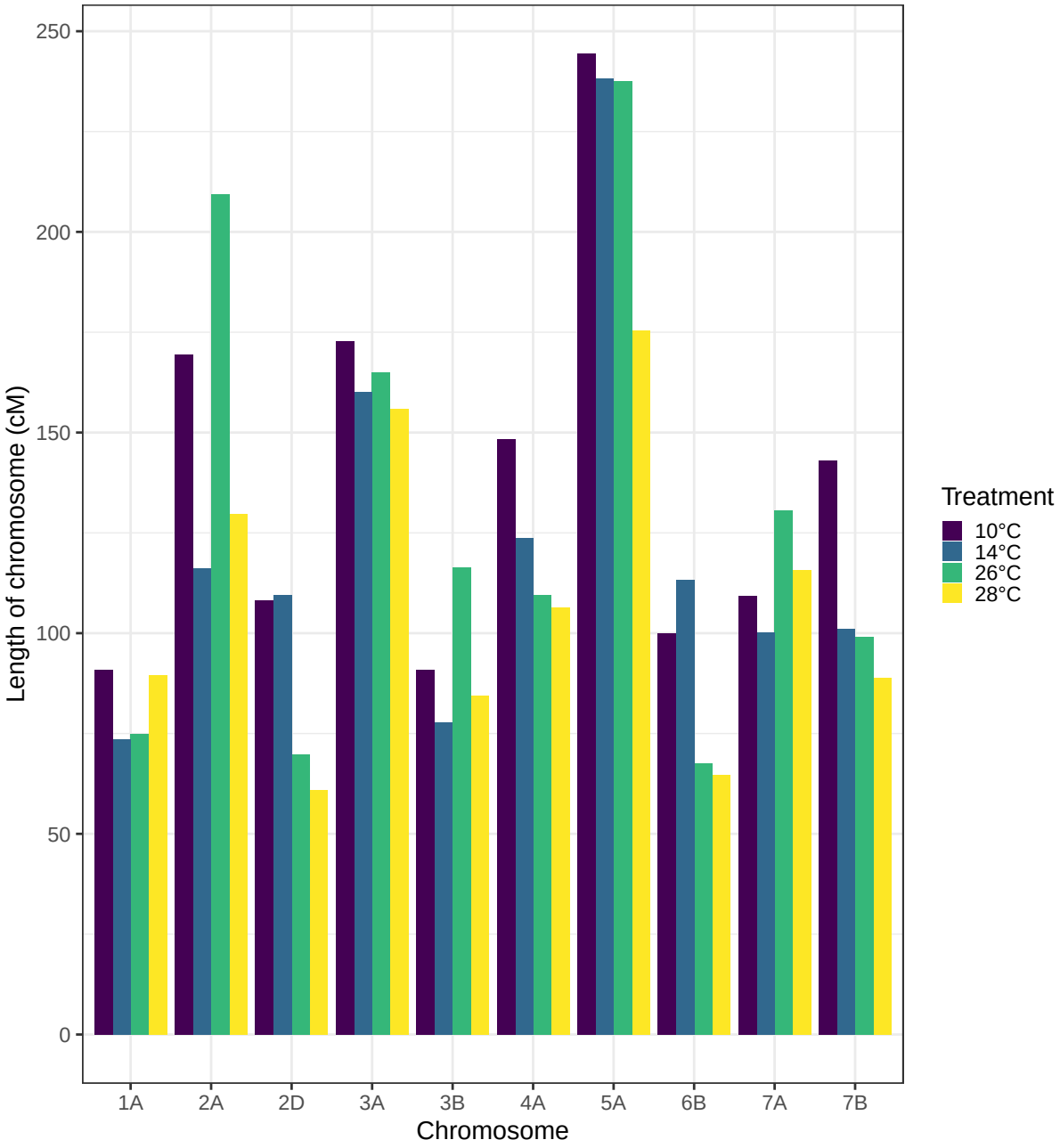


Figure 3.28 Apogee X Paragon genetic map lengths by chromosome across temperature treatments. Chromosome 5A has the largest map length and therefore the highest number of recombination events in all temperature treatments. In some chromosomes, higher temperature treatments have less recombination events overall, such as in chromosome 2D, 6B and 7B.

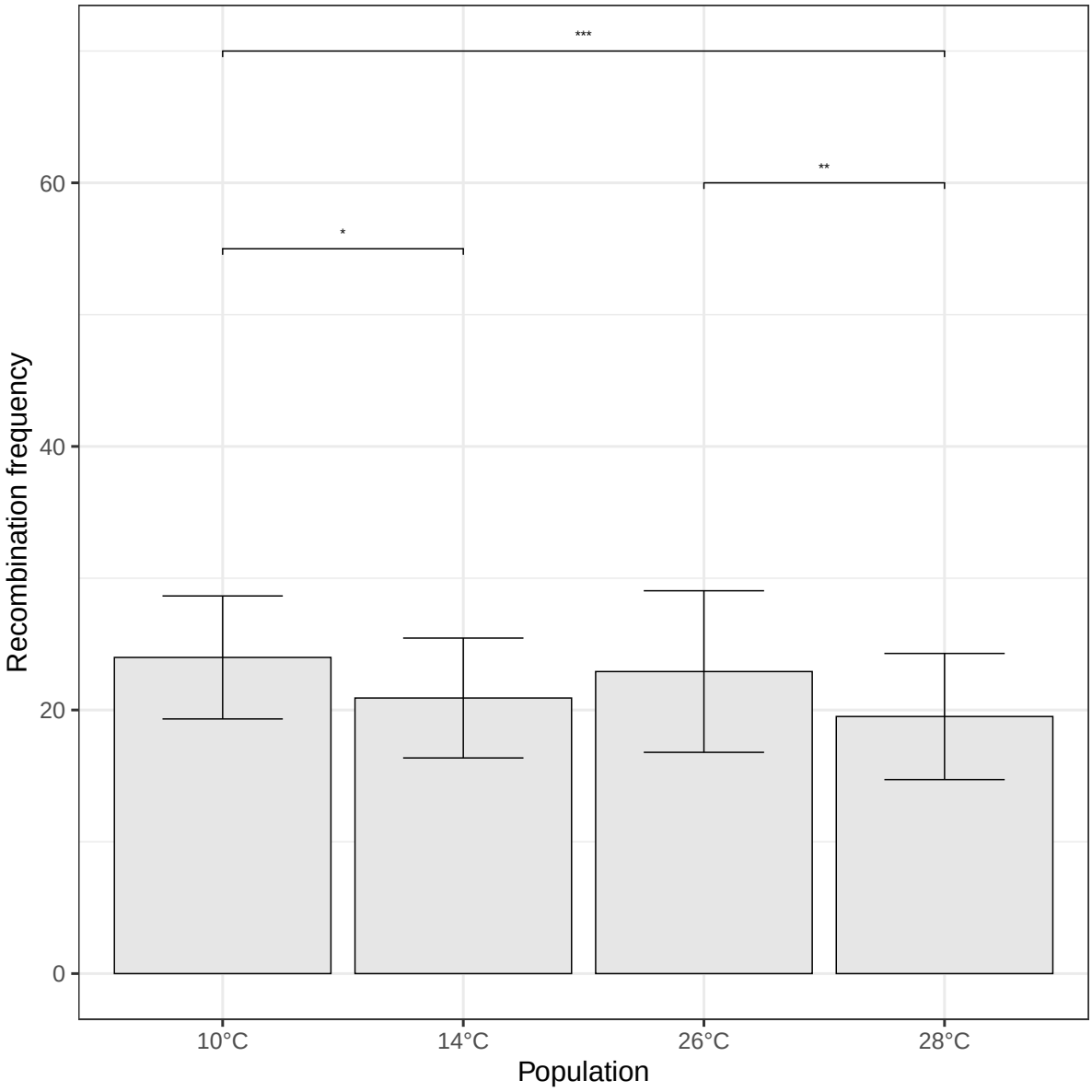


Figure 3.29 Mean recombination frequency across all chromosomes for each temperature treatment. Error bars represent ± s.d. from the mean. Significantly different populations are indicated by asterisks (* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$).

Investigating genotyping error as a possible cause of the observed effects

Genotyping error was found not to influence MRD. Linear regressions of standard sample quality control metrics (either dish quality control (DQC) or quality control call rate) as the explanatory variable and MRD as the dependent variable for each chromosome were all none-significant. R^2 values for each chromosome were all smaller than 0.026, meaning that less than 2.6% of the variation in MRD was explained by these variables in every case.

In addition, the simulation experiment indicated that sample size does not influence MRD either. For sample sizes of 500, 250, 100 and 30, 6.29%, 3.01%, 7.15% and 2.03% of the 100000 combinations of samples exhibited significant differences in MRD, as shown by a Kruskal-Wallice test. If sample size had an effect, we would expect these percentages to show a consistent trend, i.e. increasing or decreasing with sample size. It should also be noted that these percentages are close to the expected number of false-positives (5%) at this alpha threshold (0.05).

3.3.3 QTL Analysis

No significant QTL were found for recombination frequency in any of the populations analysed. Significant QTLs for MRD were found in the Opata X Synthetic population on chromosome 1A (phenotype MRD 5A), as well as the Paragon X Watkins 94 population on chromosome 1A (MRD 7A). The increasing effect came from Opata and Watkins 94 respectively for these QTLs, meaning that these parental genotypes yielded MRDs that were further away from the centre of the chromosome. The O x S QTL accounted for the highest phenotypic variance of any single QTL (24.906%). None of these QTL were significant after Bonferroni correction for the number of phenotypes tested.

Table 3.5 Table showing significant QTLs for late-filial generation (> F4) RIL populations and DH populations. Numbers have been rounded to three decimal places, so where p-values are listed as 0, represents a value smaller than 0.001. (continued below)

Population	Phenotype	Chromosome	Position (cM)	LOD	% Phenotypic Variance
A x P F5	MRD.6B	4A	28.6	2.833	4.781
O x S	MRD.5A	1A	73.97	3.615	24.93
CS x P	MRD.2D	6B	72.11	3.057	5.237
CS x P	MRD.2D	6B	72.49	3.279	5.626
CS x P	MRD.2D	6B	72.68	3.187	5.455
CS x P	MRD.2D	6B	74.05	3.053	5.234
CS x P	MRD.2D	6B	74.43	3.148	5.392
CS x P	MRD.2D	6B	74.62	3.346	5.722
CS x P	MRD.2D	6B	75	3.572	6.097
CS x P	MRD.2D	6B	75.19	3.671	6.264
A x C DH	MRD.3A	5A	10.15	3.045	12.07
A x C DH	MRD.7B	3A	77.41	3.553	14.82

Table continues below

p-value (chi-square)	p-value F	AA Mean	BB Mean	AA SE	BB SE	Position Previous Marker (cM)
0	0	42.03	38.98	0.581	0.589	28.44
0	0	38.49	31.12	1.116	1.244	72.17
0	0	41.13	37	0.737	0.801	71.73
0	0	41.18	36.89	0.732	0.804	72.11
0	0	41.17	36.95	0.736	0.8	72.49
0	0	41.08	36.96	0.73	0.811	72.68
0	0	41.08	36.88	0.725	0.818	74.05
0	0	41.13	36.79	0.721	0.82	74.43
0	0	41.22	36.75	0.724	0.811	74.62
0	0	41.26	36.73	0.726	0.807	75
0	0	42.16	45.48	0.609	0.603	8.463
0	0	39.65	43.87	0.679	0.732	66.42

LOD previous marker	Physical position of marker before (%)	Physical position of marker before (bp)	Position of marker after (cM)	LOD of next marker	Physical position of next marker (%)	Physical position of next marker (bp)
2.614	80.88	602245890	30.08	2.103	80.97	602911128
2.058	84.88	504243610	74.83	2.93	86.92	516373177
2.871	89.55	645661984	72.49	3.279	90.35	651418713
3.057	89.89	648086087	72.68	3.187	91.06	656511775
3.279	90.35	651418713	74.05	3.053	91.38	658818903
3.187	91.06	656511775	74.43	3.148	91.44	659233712
3.053	91.38	658818903	74.62	3.346	91.59	660381093
3.148	91.44	659233712	75	3.572	92.03	663531485
3.346	91.59	660381093	75.19	3.671	92.05	663669873
3.572	92.03	663531485	NA	NA	NA	NA
1.89	0.699	4959413	13.65	2.373	1.135	8059358
1.76	94.73	711296348	84.81	1.5	97.34	730849182

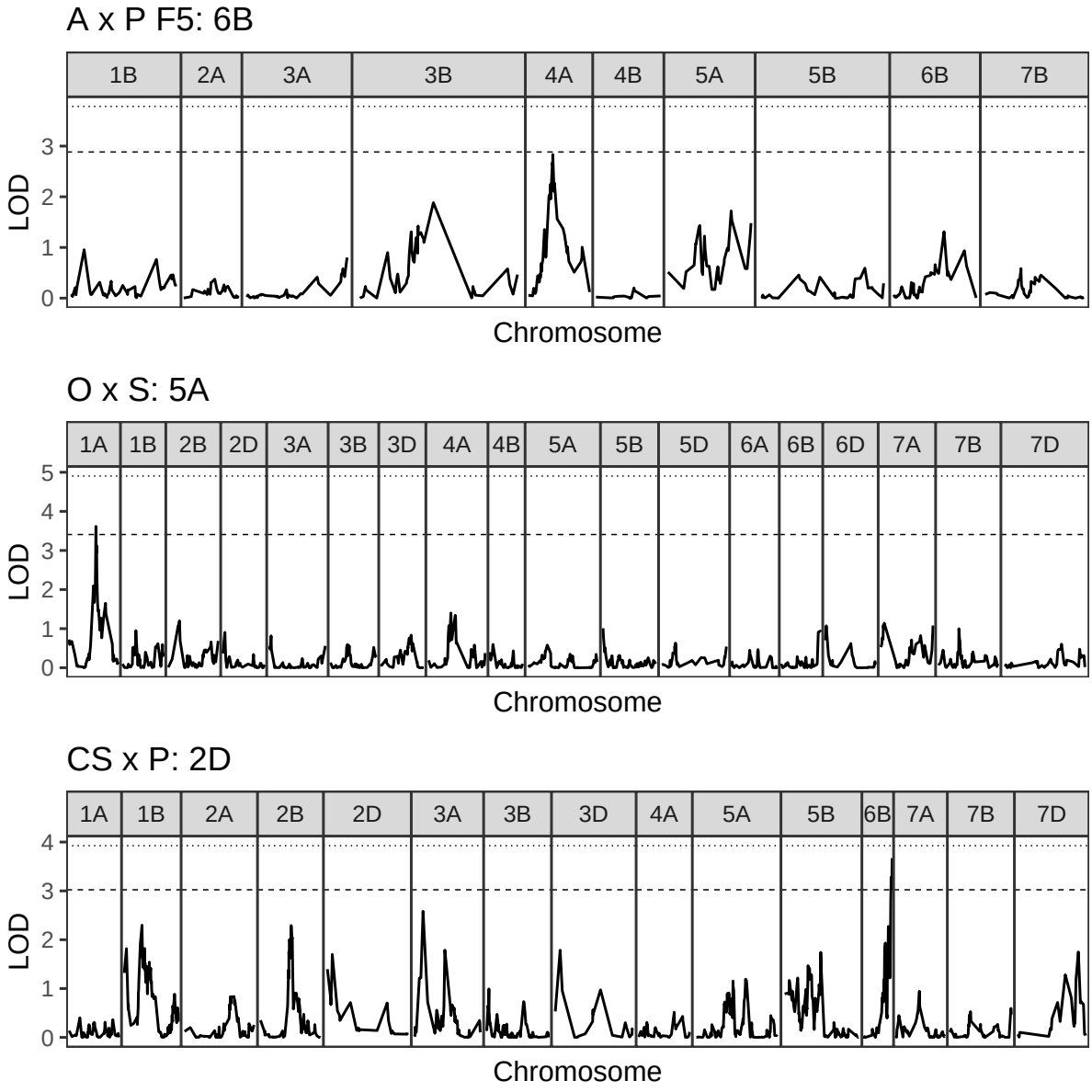


Figure 3.30 QTL Plots for late-filial generation (> F4) RIL populations and DH populations. Dashed line indicates a 0.05 significance threshold for the LOD value based on a permutation test, whilst the dotted line indicates a 0.05 significance threshold after Bonferroni correction for the number of phenotypes tested.

3.3.4 Phylogenetic Analysis

The mean ± s.d. length of gene trees for the meiotic gene set and the random gene set were 0.172 ± 0.221 and 0.227 ± 0.306 respectively. When barley was removed from the analysis, the mean and ± s.d. length of gene trees decreased in all cases to 0.011 ± 0.051

and 0.03 ± 0.17 for meiotic and random gene sets respectively. The tree length was not significantly different between both gene sets (t-test, t = -4.36, d.f. = 476.21, p = 0.00002), and this remained the case when barley was removed from the analysis (t-test, t = -5.01, d.f. = 1235.78, p = 0, Mann-Whitney U test W = 745407, p = 0.03069). Boxplots of tree lengths are shown in figure 3.29. In both phylogenies constructed from superalignments of the respective sets of genes, barley was by far the most distantly related compared to all of the wheat varieties (figure 3.30). For the tree constructed from the superalignment of the meiosis gene set, the patristic distances between pairs of nodes (table 3.5) were mostly shorter than the distances in the random gene set (table 3.6). Phylogenies without barley are shown in figure 3.31.

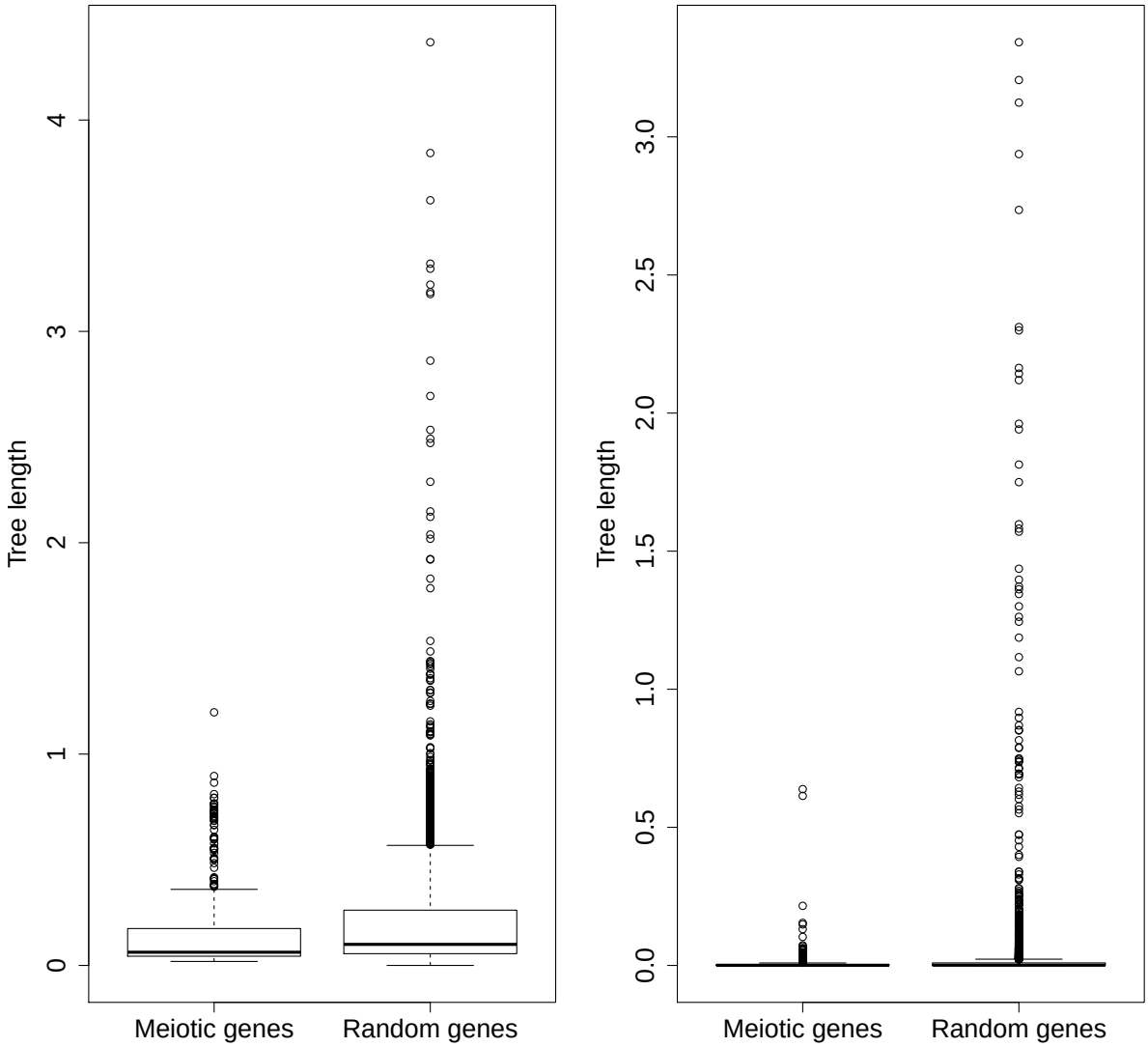


Figure 3.29 Box plots of lengths of gene trees, including wheat varieties Chinese Spring, Paragon, Cadenza, Claire and Robigus as well as Barley (Morex) for meiotic and random gene sets. The left panel includes barley, the right panel does not.

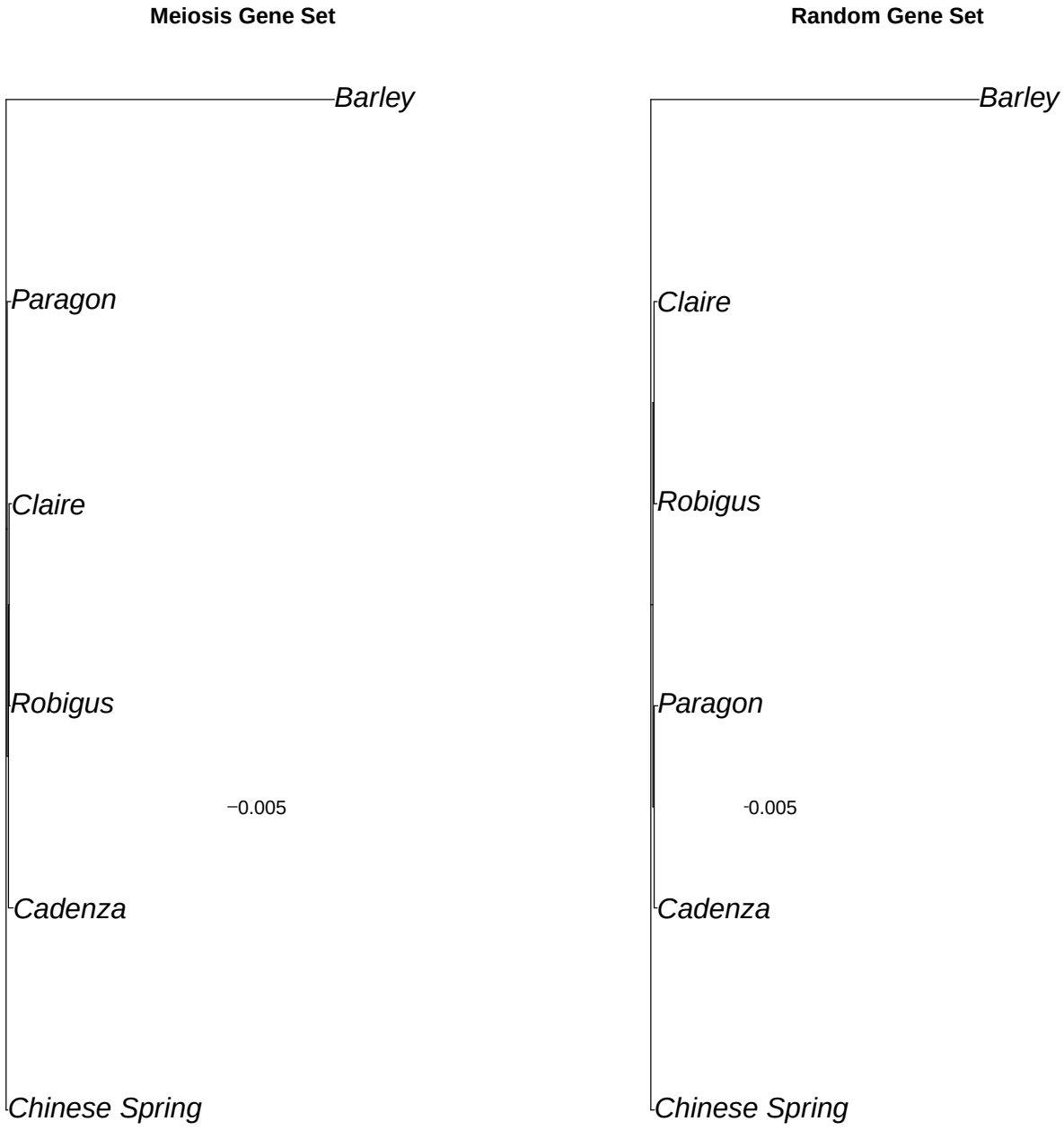


Figure 3.32 Phylogenies contructed from superalignments of all genes for both meiosis and random gene sets, including barley. Scale bars indicate number of substitutions per site.

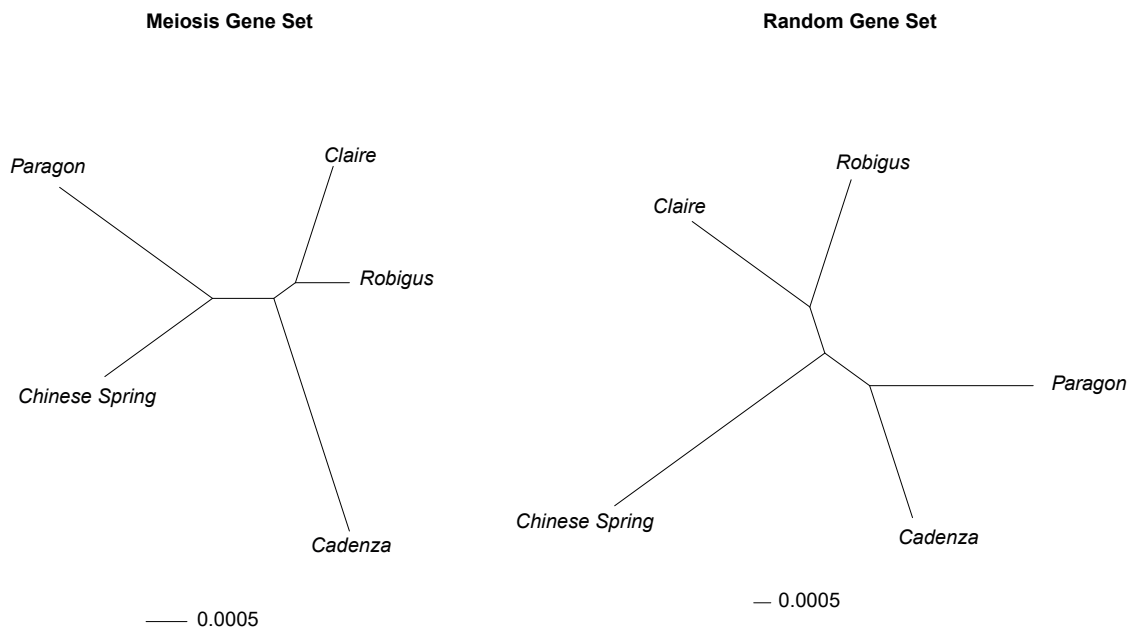


Figure 3.31 Phylogenies constructed from superalignments of all genes for both meiosis and random gene sets, without barley. Scale bars indicate number of substitutions per site.

Table 3.6 Patristic distances of nodes in the tree constructed from a superalignment of the meiosis gene set. Upper triangle has been removed due to redundancy.

	Chinese Spring	Cadenza	Robigus	Claire	Paragon	Barley
Chinese Spring						
Cadenza	0.004771					
Robigus	0.003206	0.003568				
Claire	0.003906	0.004269	0.001992			
Paragon	0.003563	0.005323	0.003758	0.004458		
Barley	0.1786	0.1815	0.1799	0.1806	0.1803	

Table 3.7 Patristic distances of nodes in the tree constructed from a superalignment of the random gene set. Upper triangle has been removed due to redundancy. Interestingly, the distances here are longer than in the meiosis gene set.

	Chinese Spring	Cadenza	Robigus	Claire	Paragon	Barley
Chinese Spring						
Cadenza	0.01274					
Robigus	0.01243	0.01079				
Claire	0.01272	0.01108	0.007815			
Paragon	0.01341	0.00848	0.01146	0.01175		
Barley	0.4259	0.4296	0.4293	0.4296	0.4303	

3.4 Discussion

The data presented here is the first detailed analysis of the effect of environmental temperature during meiosis on the distribution and frequency of recombination events in wheat. Our data, based on high-density SNP genotyping of four Apogee X Paragon mapping populations, each subjected to a different temperature during meiosis, reveal a clear effect of temperature on the distribution of recombination events. This effect is visible, although subtle, in figure 3.6, where higher temperature treatments appear to have a more distal distribution of events in both short and long arms of the chromosomes. In figure 3.6, there are some MRD values that appear to be within 25% of the centromere, which contrasts with evidence from previous studies [@saintenacDetailedRecombinationStudies2009; @chouletStructuralFunctionalPartitioning2014] that the distribution of recombination in wheat is limited to the distal ends of the chromosomes. These MRD values are most likely are artefacts of the method used here to measure recombination, where the position of recombination is assigned as the midpoint between two markers, in conjunction with the reduced marker density in centromeric regions after filtering (figure 2). This should not have any impact on the analysis of

the relative difference in recombination distribution between temperature treatments, as all treatments used the same genetic map with the same distribution of markers. The statistical analyses of MRD for individual chromosomes show that the effect of temperature on recombination distribution is limited to specific chromosomes/chromosome arms, suggesting that chromosomal structure may influence the susceptibility of chromosomes to changes in temperature. This mirrors results from barley [@higginsSpatiotemporalAsymmetryMeiotic2012; @phillipsEffectTemperatureMale2015].

A comparison of the Apogee X Paragon F2 genetic map produced here to the F5 genetic map of the same cross from Allen et al., [-@allenCharacterizationWheatBreeders2016] shows a high degree of similarity in the clustering (table 3.1), ordering (figure 3.5) and genetic distribution (figure 2) of markers, indicating that the genetic map is robust. Noticeable in the comparison of clusters however is that some of the chromosomes in the F5 map, such as 5A and 7A, are split into multiple linkage groups. This could explain the markers that are present in our map that are not present in the map of Allen et al., as these were most likely resolved as smaller linkage groups in the latter, which were then discarded. Our marker selection process was more stringent than in Allen et al., [-@allenCharacterizationWheatBreeders2016] as we only used codominant markers that were categorized as "Poly high resolution", the highest quality marker categorization in Axiom Analysis Suite @bassilDevelopmentPreliminaryEvaluation2015, whereas multiple categories of marker were used in the map of Allen et al. In addition, in this study the SNP cluster plots for each marker were visually inspected and only markers with a clear delineation between genotyping clusters representing homozygotes for the Apogee allele, heterozygotes and homozygotes for the Paragon allele were used.

Our confidence in the validity of the results are increased by the fact that differences in MRD between treatments only become non-significant when two treatments are removed, either both of the low temperature treatments (10°C and 14°C) or both of the high temperature treatments (26°C and 28°C). This effect was less pronounced in the short arms of chromosome, indicating that perhaps temperature have less of an effect on these regions. However, this could also simply be the result of lower marker density

in short chromosome arms. It is clear the observed shifts in recombination distribution are not due to genotyping error, as the results of linear regressions of standard quality control metrics against MRD were all non-significant. Furthermore, the results of the simulation experiment show that the differences in MRD observed in the Apogee X Paragon crosses were not the result of sampling error due to small population size.

In order to perform a statistical analysis on the differences in distribution of recombination events between temperature treatments, it was necessary to associate genetic maps to the physical wheat genome assembly. To do this we devised a method that utilizes the longest increasing subsequence of BLAST positions of marker sequences to the IWGSC assembly, allowing the measurement of MRD in each individual. This unfortunately comes with a caveat, in that markers that do not conform to this sequence must be removed, which reduces the density of markers in certain regions (figure 3.4). This is one of the primary limitations of using MRD and means that our ability to precisely localize certain recombination events is reduced. The reason for the reduction in marker density is likely due to structural differences between varieties Apogee and Paragon, used here to generate the genetic maps, and Chinese Spring, the only wheat variety currently to have a completely publicly-available, chromosome-level genome assembly. A reanalysis of the data could provide further information should a chromosome-level genome assembly of either Apogee or Paragon be released in the future. Despite these caveats, the data has sufficient marker density to inform our picture of recombination distribution in many chromosomal regions (figure 3.4). The process used in this study conforms closely to the process that a wheat breeder might implement in their development of new varieties, and so has direct bearing on the applications of temperature in wheat.

In addition to the effect of temperature on the distribution of recombination events, we also observed an effect on the frequency of recombination events. Recombination frequency followed a U-shaped response from 10° to 26°C (figure 3.28), which is consistent with results in Arabidopsis presented by Lloyd et al. [-@lloydPlasticityMeioticRecombination2018]. This is followed by a

dip at 28°C, which differs from the results in Arabidopsis. Lloyd et al. [-@lloydPlasticityMeioticRecombination2018] suggest that this effect is primarily produced by changes in class I interfering crossovers and speculate that recombination may be minimized in organisms that are already well-adapted to their environment and living in optimal conditions.

The number of recombination events that occurred in each gamete can be estimated by dividing the mean number of recombination events across all individuals by the number of chromosomes, then dividing that by two (as each zygote is composed of two gametes). This gives us an estimate of 1.09 recombination events per gamete. This is lower than published counts of chiasmata per bivalent in wheat, which are around 2.3 [@millerEffectIncreasedDosage1985]. There are several factors that contribute to this underestimation. Firstly, double recombination events that occur in between marker positions are not detectable using SNP data. In addition, in a sequence of heterozygous genotypes, if both gametes have a recombination event between the same two markers, the recombination event will not be detected. Finally, in some of the chromosomes analysed, there were a lack of markers at the very distal ends of the chromosomes (figure 3.5), and so recombination events that occurred in these regions could not be detected.

Our data indicate that increasing the temperature during meiosis could have some limited use to breeders in breaking up centromeric linkage blocks. We observed four chromosome arms with significant shifts in recombination distribution between temperature, and several putative temperature-dependent recombination hotspots (table 3.4). Included in the 3B temperature-dependent hotspot were putative genes influencing plant development, including a NAC domain protein [@puranikNACProteinsRegulation2012] as well as an auxin-responsive protein [@tealeAuxinActionSignalling2006]. Despite these points, there are still many genes that remain highly linked in regions closer to the centromere on many chromosomes. To achieve thorough mixture of these genes in progeny of crosses, it will be important for breeders to explore other avenues of manipulating recombination distribution. One potential option is to produce fusion

proteins linking SPO-11, the protein that initiates recombination through production of double-stranded breaks in the DNA, to other DNA targeting proteins, such as Zinc finger elements, Transcription activator-like elements or dead Cas9. Research in yeast using these methods revealed a 2.3 to 6.3-fold increase in COs near targeted regions [@sarnoProgrammingSitesMeiotic2017]. Applications of this technique to agriculture outside of research could however be limited due to increasingly strict legislation, such as the 2018 European Union ruling in case C-528/16 that organisms edited using directed mutagenesis methods such as CRIPR-Cas9 will be officially classified as genetically modified organisms.

If breeders do decide to utilize temperature as a means of altering recombination distribution, they will also have to consider the loss in fertility of wheat plants at higher temperatures of meiosis [@draegerShortPeriodsHigh2017]. These elevated temperatures would therefore need to be employed strategically and transiently in between generations that are grown at lower temperatures. It would be of interest to examine the epigenetic effects of elevated temperature during meiosis, such as whether there are any lasting changes inherited by progeny, and whether these changes are detrimental to plant growth.

Interestingly, the results of our QTL analysis yielded no significant QTLs for recombination frequency or MRD after Bonferroni correction for the number of phenotypes tested. This is in contrast to other studies such as [@gardinerAnalysisRecombinationLandscape2019], where a significant QTL, albeit accounting for only a small proportion of phenotypic variance (6.037%), was found in the same Chinese Spring X Paragon population analysed here. This difference between studies suggests that the processing performed in [@gardinerAnalysisRecombinationLandscape2019] in which markers were grouped into large bins might have effected the analysis, and indeed no difference was found when the authors attempted to validate their CO frequency QTL using Cadenza TILLING populations containing knockout mutations for the most likely gene candidate driving this QTL. There are two possible explanations for the lack of QTL detected, either our method is lacking in statistical power or that there is no genetic

component involved in either recombination frequency or distribution in the varieties tested. To assess the latter hypothesis, we performed a phylogenetic analysis of homologues of genes that have been functionally characterized to be involved in meiosis in *Arabidopsis*. On the whole, the meiotic gene set was more conserved than the random gene set. However, when barley was included in the analysis, we observed far more variation in meiotic genes between barley and wheat than between any two wheat varieties. This suggests that the conservation of wheat genes involved in meiosis is largely due to the fact that they diverged fairly recently. The difference between meiotic and random gene sets in the phylogenies where barley was excluded suggests that stabilising selection also plays a role.

The phylogenetic analysis provides insight into the evolutionary constraints (or lack thereof) operating on meiotic genes, and therefore the potential for genetic control of recombination in wheat. The fact that the meiotic genes are similarly distant between barley and wheat in both meiotic and random gene sets indicates that there is potential for genetics to influence recombination, as if there was strong stabilizing selection operating at these loci we would expect barley to be more closely related to wheat in the meiotic gene set compared to the random gene set. However, the closeness of the meiotic genes in wheat varieties suggests that there simply hasn't been enough time in the history of wheat for this deviation to occur, which could explain the lack of QTL influencing recombination. This would be congruent with evidence on the inception of all modern bread wheat at a relatively recent point in history around 10,000 years ago.

In conclusion, the work here shows that temperature has a subtle effect on the frequency and distribution of recombination events in wheat. The analysis of recombination distribution in comparison to gene distribution indicates that the utility of this effect may be limited, with large amounts of genes remaining under strong linkage, and thus inaccessible to manipulation by breeders. Future work involving cytological analysis of wheat meiocytes subjected to different temperatures would be of interest, as the principle of consilience asserts that confidence in a result increases when reached independently through different methods.

4 Segregation distortion: utilizing simulated genotyping data to evaluate statistical methods

All results described here have been published in the following manuscript:

Coulton, A., Przewieslik-Allen, A., Burridge, A., Shaw, D., Edwards, K., Barker, G., 2020. Segregation distortion: Utilizing simulated genotyping data to evaluate statistical methods. PLoS ONE

4.1 Abstract

Segregation distortion is the phenomenon in which genotypes deviate from expected Mendelian ratios in the progeny of a cross between two varieties or species. There is not currently a widely used consensus for the appropriate statistical test, or more specifically the multiple testing correction procedure, used to detect segregation distortion for high-density single-nucleotide polymorphism (SNP) data. Here we examine the efficacy of various multiple testing procedures, including chi-square test with no correction for multiple testing, false-discovery rate correction and Bonferroni correction using an in-silico simulation of a biparental mapping population. We find that the false discovery rate correction best approximates the traditional p-value threshold of 0.05 for high-density marker data. We also utilize this simulation to test the effect of segregation distortion on the genetic mapping process, specifically on the formation of linkage groups during marker clustering. Only extreme segregation distortion was found to effect genetic mapping. In addition, we utilize replicate empirical mapping populations of wheat varieties Avalon and Cadenza to assess how often segregation distortion conforms to the same pattern between closely related wheat varieties.

4.2 Introduction

Segregation distortion is the phenomenon in which alleles in the progeny of a cross between two varieties or species deviate from expected Mendelian ratios. In an F2 population originating from a biparental cross, the expected ratio of genotypes AA:AB:BB (progeny homozygous for the allele from the first parent, heterozygotes and progeny homozygous for the allele from the second parent) in absence of segregation distortion is 1:2:1. Segregation distortion is observed across a wide range of taxa, including animals such as Drosophila [@lyttleCheatersSometimesProsper1993; @phadnisSingleGeneCauses2009] and mice [@silverPeculiarJourneySelfish1993], as well as crop species, including cotton [@daiIdentificationCharacterizationSegregation2017], maize [@luChromosomalRegionsAssociated2002; @wangHighSegregationDistortion2012], potato [@manrique], chickpea [@castroSegregationDistortionLocus2011], barley [@liuProgressSegregationDistortion2010] and wheat [@allenCharacterizationWheatBreeders2016; @gardnerHighlyRecombinedHigh2016; @wingenWheatLandraceGenome2017]. The primary explanation of the cause of segregation distortion is a selection pressure operating against one of the parental alleles at some stage of the development cycle, whether at meiosis through meiotic drive [@lyttleCheatersSometimesProsper1993], through male gamete competition [@daiIdentificationCharacterizationSegregation2017], or at the level of the zygote. An example of this is the pollen killer gene in wheat [@loegeringDistortedInheritanceStemRust1963], for which there is an allele that causes pollen cells to degenerate until unviable, leading to an over representation of the non-deleterious allele.

Segregation distortion can be problematic for crop breeders, who wish to generate varieties with novel genotypic compositions that are better suited to meeting the various aims of modern agriculture, such as increased yields or improved resistance to biotic or abiotic stresses. Distorted segregation at a locus could skew most lines in a recombinant inbred line population (RIL) away from the desired genotype, requiring breeders to create larger numbers of lines to compensate. It would be useful if we could attribute regions of segregation distortion to causative loci in the genome, as this would allow

breeders to plan for the occurrence ahead of time. One important obstacle to this goal is another potential cause of segregation distortion: sampling error. With small RIL population sizes, it is possible that a specific parental allele is, by chance, sampled more often than its alternative in the progeny, leading to a distorted ratio of segregation. Conflating this for distortion caused by a selective pressure would be problematic, as the same pattern of distortion in the progeny would not be repeated if the RIL population was recreated. Planning for this distortion in breeding programmes would therefore be counterproductive. How much of the segregation distortion typically observed in RIL populations is due to chance?

Distinguishing distortion caused by selection from that caused by chance is difficult, because both have the potential to produce similar patterns of segregation in the progeny of a cross. What we can assume though is that if the selection strength is high enough, the intensity of distortion produced would be unlikely to have occurred by chance. This is complicated by the fact that the effects of chance change with population size, being more prevalent when population size is small, and eventually evening out as population size becomes larger. Our sole detection criterion for separating selection from chance as the cause of segregation distortion then is the amount of distortion as a function of the population size. This leads to an important question: at what stage do we say that enough distortion is present for the event to be caused by selection? If we make our detection criteria too lenient, then we increase the risk of type I errors (false positives), whilst stricter criteria increase the risk of type II errors (false negatives). What is the optimal place to draw this proverbial line in the sand when detecting segregation distortion?

The diversity of criteria used in the literature reflect the difficulty of answering this question. Some authors settle for a simple chi-square test with the minimum significance threshold of $p < 0.05$ [@luChromosomalRegionsAssociated2002; @allenCharacterizationWheatBreeders2016; @takumiSegregationDistortionCaused2013; @avniUltradenseGeneticMap2014; @liHighDensitySNPSSR2015], others report multiple significance thresholds [@manrique; @thompsonInheritanceGeneticMark-

ers1991; @pengMolecularGeneticMaps2000; @singhIntegratedMolecularLinkage2007; @adamskiSegregationDistortionHomozygous2014], whilst others use corrections for multiple testing, including false discovery rate (FDR) [@gardnerHighlyRecombined-High2016; @wingenWheatLandraceGenome2017; @seymourGeneticArchitectureRe-current2017] and the even stricter Bonferroni correction [@daiIdentificationCharac-terizationSegregation2017; @manninenAssociationsAnthercultureResponse2000]. This inconsistency has the further implication that many of the studies on segregation distortion are not comparable, which is problematic for the general advancement of our knowledge of segregation distortion. It interferes with our ability to assess hypotheses such as: (i) levels of segregation distortion differ between different species, (ii) segregation distortion increases with the genetic distance between parents.

To circumvent the conflation between selection and chance, it would be useful if we could observe the processes that lead to the final genotypic composition of a RIL population. Whilst this is infeasible to achieve with real organisms, it is possible in an in-silico simulation of a RIL population. Here we utilize PedigreeSim [@voorripsSimulationMeio-sisDiploid2012], which computationally models single-nucleotide polymorphism (SNP) genotype data from a RIL population, starting with recombination between homologues during meiosis, generation of gametes and fusion of gametes to form a zygote. This process can be repeated for the desired number of filial generations. The simulation allows us to control multiple parameters that could influence segregation distortion, such as the number of SNP markers used, the position of selection in the genome, the strength of selection in the genome, the distribution of SNP markers, and the size of the population. We can also examine the interaction between different parameters, such as population size and selection. With knowledge of the parameters that produced the final genotyping dataset, we can then attempt to identify the appropriate threshold to detect segregation distortion by examining the performance of various statistical tests. For example, when a selection pressure of strength X is applied at a locus, in what pro-portion of simulated populations is this locus identified as being significantly distorted for a given statistical test and population size?

In addition to the simulation experiments performed, we also wanted to investigate how much of the purported segregation distortion typically observed in real populations is the result of random chance rather than a consistent selection pressure. To examine this, we produced replicate populations of the same cross between varieties Avalon and Cadenza. These consisted of two F2 populations with Avalon as the female parent, and two F2 populations with Cadenza as the female parent, with each population containing around 96 lines. We were then able to compare replicate populations and test whether they showed any consistency in the regions exhibiting segregation distortion, which if they did would imply that the distortion was the result of a selection pressure rather than random chance. There is a trend in the literature of removing markers exhibiting segregation distortion before the construction of a genetic map [@allenCharacterizationWheatBreeders2016; @liuChromosomespecificSequencingReveals2016; @roorkiwalDevelopmentEvaluationHighdensity2018]. It has already been shown by a previous simulation that segregation distortion does not affect the order of a genetic map [@hackettEffectsGenotypingErrors2003]. High levels of segregation distortion can however effect the estimation of recombination frequency between a pair of markers [@lorieuxMaximumlikelihoodModelsMapping1995], which is used in the clustering stage of genetic map construction. Here we use our simulation to examine whether clustering of markers is significantly affected by segregation distortion in modern genetic mapping software such as MSTMap [@wuEfficientAccurateConstruction2008].

Finally, after identifying appropriate statistical tests for the detection of segregation distortion from these experiments, we perform a reanalysis of some existing genotyping datasets from populations of hexaploid and tetraploid wheat [@allenCharacterizationWheatBreeders2016; @avniUltradenseGeneticMap2014]. This allows us to highlight important regions of segregation distortion that could be the subject of further investigation, potentially leading to the identification of the genomic position and mechanism of a causative locus of segregation distortion in wheat.

4.3 Materials and Methods

4.3.1 Plant cultivation

For the replicate empirical mapping populations, we generated F2 plants using bread wheat (*Triticum aestivum* L.) varieties Avalon and Cadenza in reciprocal crosses. All plants were grown in uniform conditions at the same time using pots filled with peat-based soil and kept in a glasshouse at 15-25 °C with 16-h light, 8-h dark. Leaf-tissue was harvested from F2 plants two weeks after sowing. DNA was extracted following the protocol in [@pallottaMarkerAssistedWheat2003] with minor modifications.

4.3.2 Sample genotyping

DNA concentration was assessed using a Qubit 2.0 Fluorometer and was the normalized to 23 ng / µl ready for analysis with the Axiom® Wheat Breeder's array. Sample preparation for array genotyping was performed with the Beckman Coulter Biomek FX. Samples were then genotyped using the Axiom® 35K Wheat Breeders array in conjunction with the GeneTitan® using standard Affymetrix protocols (Axiom® 2.0 Assay for 384 samples P/N 703154 Rev. 2).

4.3.3 Genetic Map Construction

Axiom Analysis Suite (version 3.1.51.0) was used to assign genotype calls using the Axiom Best Practices Genotyping Workflow. There were 3044 SNPs polymorphic between the parental varieties, Avalon and Cadenza, that were deemed suitable for genetic mapping. These were designated as PolyHighResolution, which is the category assigned to markers that are clearly codominant, by Axiom Analysis Suite and had a minor allele frequency > 0.1. The minor allele frequency criterion served as a simple metric to remove markers with highly erroneous cluster plots from the analysis. Cluster plots of the probes that did not meet the minor allele frequency criterion were inspected by eye to ensure that no genuine cases of segregation distortion were omitted.

To create the genetic map, the genotyping data from the Cadenza X Avalon population was used. The ASMap package in R, an implementation of the MSTmap algorithm, was used for clustering, ordering and calculation of genetic distance between markers. Various values for the clustering parameter were tested during the creation of the genetic map. The final value used was 10-25, which returned several linkage groups that contained around 200 markers, which is in accordance with other genetic maps of wheat produced with the 35k Wheat Breeder's array [@allenCharacterizationWheatBreeders2016]. Chromosome assignment to linkage groups was based on information from nullisomic lines from CerealsDB [@wilkinsonCerealsDBExpansionResources2016] as well as a BLAST search of probe sequences against the IWGSC RefSeq v1.0 sequence [@consortiumiwgscShiftingLimitsWheat2018] (hereafter referred to as the IWGSC assembly). Markers were assigned physical locations based on a BLAST search of probe sequences against the IWGSC assembly. Any linkage groups that spanned less than 80% of the physical distance of the chromosome were removed from the analysis, as we were interested in observing patterns of segregation along the entire length of the chromosome. Linkage groups representing the following chromosomes were retained: 1A, 1B, 1D, 2A, 2B, 2D, 3A, 3B, 4A, 4B, 5A, 5B, 6A, 6B, 7A.

4.3.4 Simulation of genotyping data

Genotyping data from a single seed descent population were simulated using PedigreeSim [@voorripsSimulationMeiosisDiploid2012] in conjunction with a custom wrapper script written in R [@coultonPedigreeWrapper2019]. The R script provides the capability to apply a selection pressure of a specified strength on gametes of a parental genotype at a locus. For example, we could apply a negative selection pressure of strength 1/20 at marker 200 against gametes with a "B" genotype, meaning that these gametes would be 5% less viable than gametes with an "A" genotype at this locus. We would therefore expect this selection pressure to produce a 100:95 ratio of A:B gametes. PedigreeSim allows the input of markers at specified centimorgan positions, meaning that we were able produce simulations that had the marker distribution of wheat chro-

mosomes. For many of the simulations, we used the existing genetic map from the Cadenza X Avalon population to provide these marker positions so that the segregation distortion data were comparable to empirical populations of wheat. When performing simulations, we ran 1000 simulations for each unique set of parameter values unless otherwise stated.

To examine the effect of segregation distortion on genetic map construction, we simulated two chromosomes using the centimorgan positions from chromosomes 1A and 6B of the Cadenza X Avalon genetic map, which were chosen based on marker density. Before genetic map construction, the order of markers in the genotyping data was scrambled to ensure that this information was not being used by the mapping software. Firstly, we tested clustering when one selection pressure resulting in the highest level of distortion (0:0:1 ratio of AA:AB:BB genotypes) was applied to chromosome 1A at marker 200. We then tested clustering when each chromosome had a selection pressure applied at marker 30 and 200 of chromosomes 6B and 1A respectively in favour of the same parental allele. We also tested the effect of segregation distortion on map length using selection pressures of varying strengths at the positions previously mentioned.

4.3.5 Measurement of segregation distortion and p-value adjustment

To measure segregation distortion, we used a variety of methods. These include the magnitude of distortion, referred to here as M, which is defined as a/(a + b) where a and b represent the number of plants with homozygous A and B genotypes respectively at an arbitrary locus. M ranges from 0 to 1, with 0 meaning no A genotypes are present and 1 meaning no B genotypes are present. For F2 populations, we use a chi-square goodness-of-fit test to measure deviation from a 1:2:1 ratio of AA:AB:BB genotypes, whilst for F6 populations, we measured deviation from a 1:1 ratio of AA:BB genotypes. Adjusted p-values were produced using the p.adjust function in R with either the Benjamini-Hochberg procedure for false discovery rate (FDR) correction or the Bonferroni correction.

4.4 Results

4.4.1 Validation of Simulation

Simulated data closely resembled empirical data from the Cadenza X Avalon mapping population. The mean (± s.d.) number of crossover events per plant for chromosome 1A was 2.72 ± 3.31, and 2.59 ± 1.31 in empirical and simulated populations respectively, each population containing 96 individuals. There was no significant difference between the number of crossover events in individuals between empirical and simulated data as determined by a Mann-Whitney U test (p = 0.07). The mean (± s.d.) length of simulated genetic maps over 1000 simulations, using 96 individuals and the marker distribution from chromosome 1A of the Cadenza X Avalon genetic map, was 130.9 ± (7.3) centimorgans (cM), whilst the length of the empirical map was 130.48 cM. Simulated data closely conformed to the expected levels of heterozygosity for each filial generation (which should reduce by half for each generation in selfing organisms), with mean (± s.d.) values over 1000 simulations of 49.96 (± 1.73)%, 25 (± 1.42)%, 12.47 (± 0.97)%, 6.24 (± 0.7)%, 3.13 (± 0.48)% for F2, F3, F4, F5 and F6 generations respectively.

In the recombination frequency heatmaps (figure 4.1) of empirical and simulated data, the regions of low recombination between closely linked markers along the diagonal are largely preserved, whilst in the simulation, recombination frequency rises faster than in the empirical data with increasing distance between markers. This is to be expected, as recombination frequency (or a proxy measure, in this case the hamming distance for MSTmap) is used in the clustering stage of genetic map construction, meaning we do not expect to see pairs of markers above a particular recombination frequency threshold together in a single linkage group of the empirical data. Segregation of genotypes across markers in the simulated data are more autocorrelated than in empirical data, with values of 0.95 ± 0.03 (averaged over 1000 simulations with populations of 96 individuals and no selection) and 0.875 respectively (figure 4.2). This is expected as the empirical data contains both genotyping errors and missing data whilst the simulated data does

not.

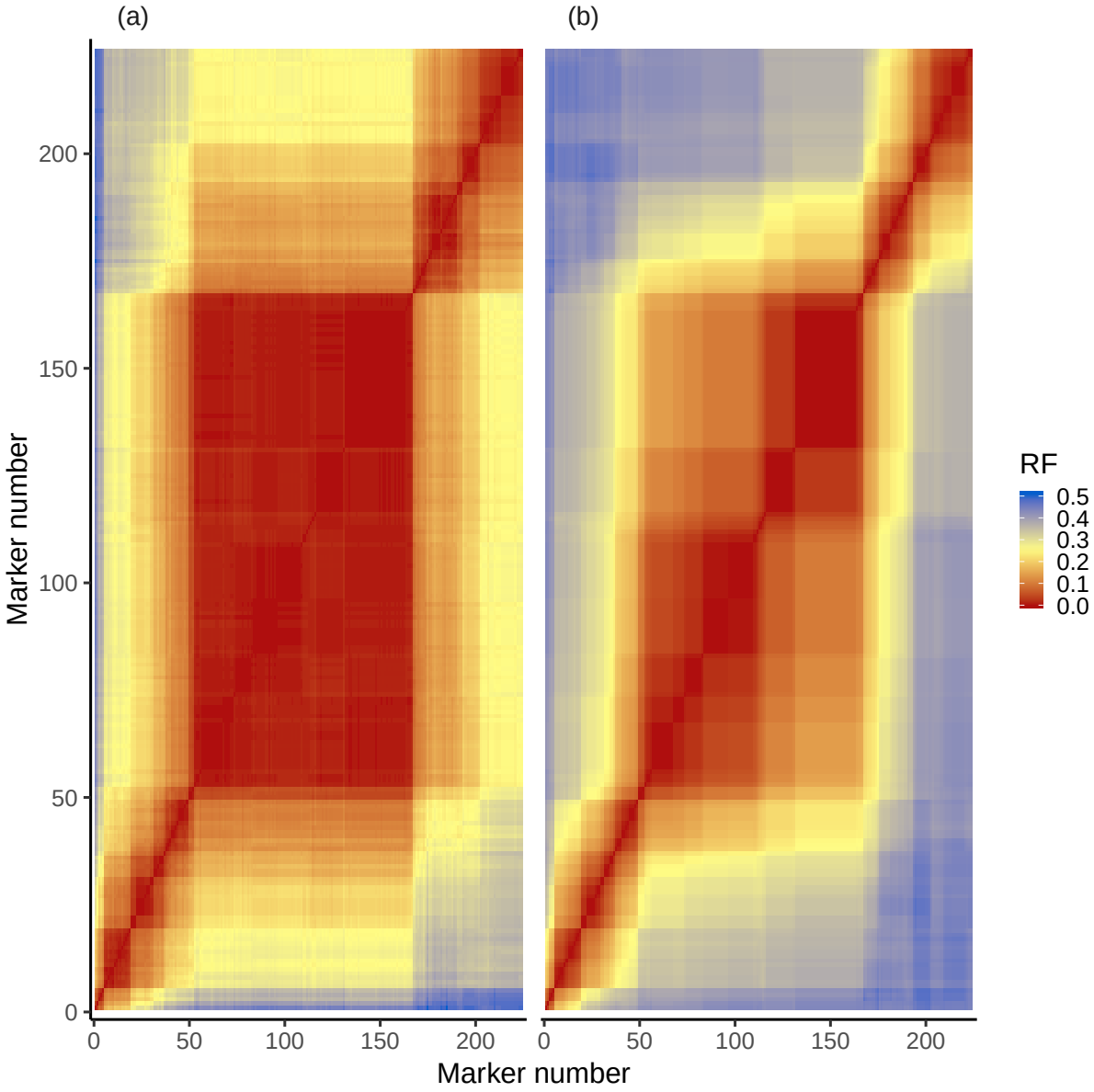


Figure 4.1 Comparison of recombination fraction heatmaps for both empirical (a) (Avalon X Cadenza 1A) and simulated data (b). The large central red block most likely represents the centromeric region of the chromosome, as wheat is known to have a lack of recombination in this area. The pattern of recombination cold spots (represented by red squares) is largely conserved between empirical and simulated data. The empirical data has low to medium levels of recombination between distant markers (represented by yellow regions), whilst the simulated data shows high levels of recombination (represented by blue regions)

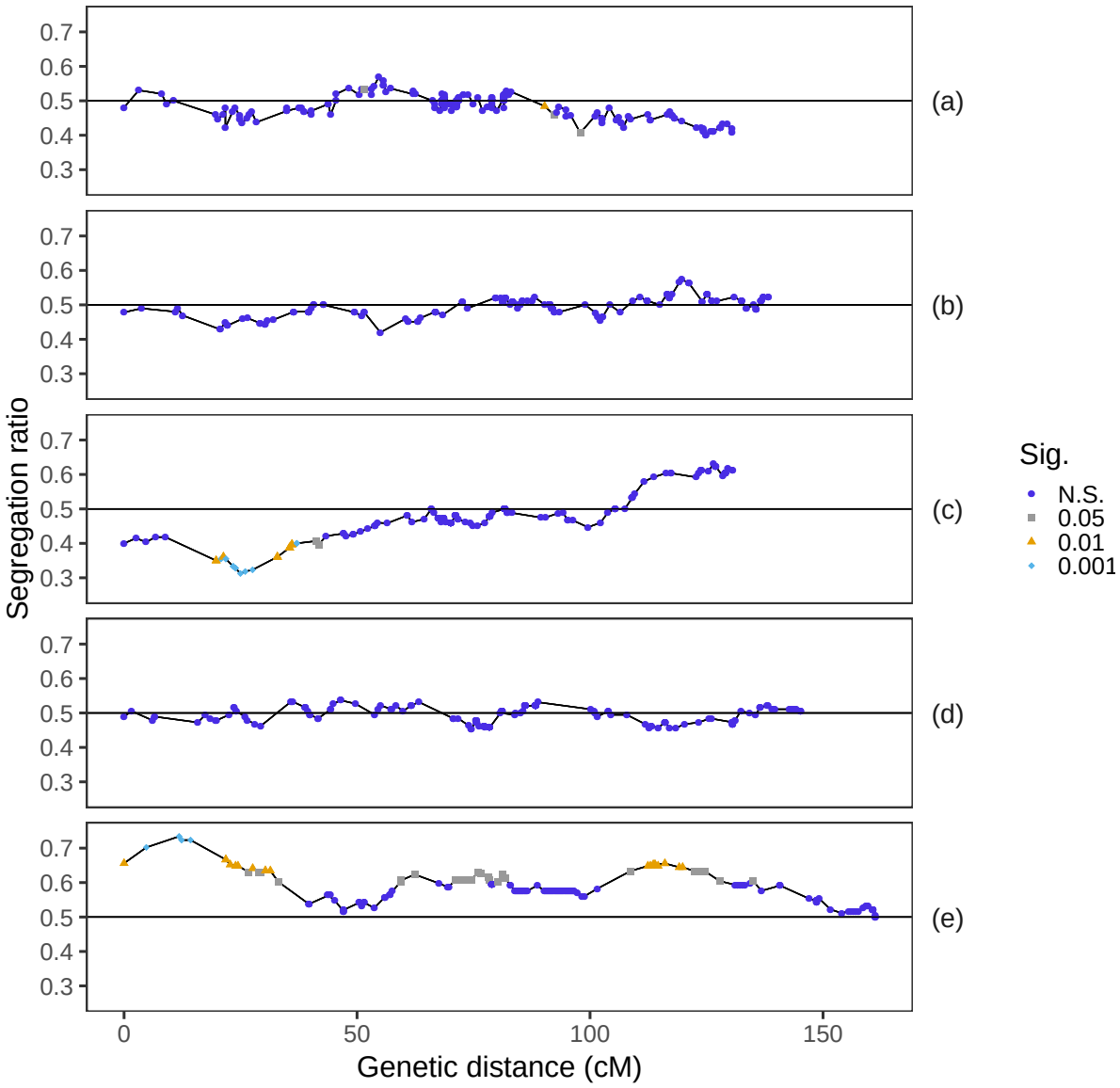


Figure 4.2 Comparison of empirical data (a) from chromosome 1A of a Cadenza X Avalon F2 mapping population with 96 individuals to simulated data (b - d). The simulations each have 96 individuals and were produced using the marker distribution from the empirical data. The y-axis shows the segregation ratio of homozygous genotypes, shown here as a proportion of the total number of homozygous genotypes ((a)/(a + b)). The black horizontal line indicates an even 1:1 ratio of homozygous genotypes. Included are simulations of both F2 (b, c) and F6 (d, e) single seed descent populations for comparison, as well as simulations exhibiting the least (b, d) and the most (c, e) amount of segregation distortion out of 1000 simulations. None of the simulations have any selection pressure applied, so these plots indicate the effect of sampling error on segregation. Sig. = significance threshold (chi-square test)

4.4.2 Simulation experiments

Initially, we ran simulations using the marker distribution for chromosome 1A of the Cadenza X Avalon cross for population sizes of 96, 300, 1000 and 10000, all with 224 markers and no selection applied. M decreased with increasing population size, whilst the proportion of simulations that contained markers exhibiting significant segregation distortion stayed relatively constant, as shown in Table 4.1. To test the effect of marker binning on the detection of segregation distortion, simulated genotyping datasets with a reduced marker set (93 markers) containing only skeleton markers were produced. Only the FDR and Bonferroni corrections showed any differences between marker sets (table 4.1). Increasing population sizes decrease the variance in segregation between simulations, but also make chi-square significance criteria more sensitive (figure 4.3). Filial generation did not influence the number of simulations that exhibited significant segregation distortion (table 4.1) according to a chi-square test, (comparison of F2 and F6 with population size 300, ?2 = 0.004, df = 1, p = 0.95).

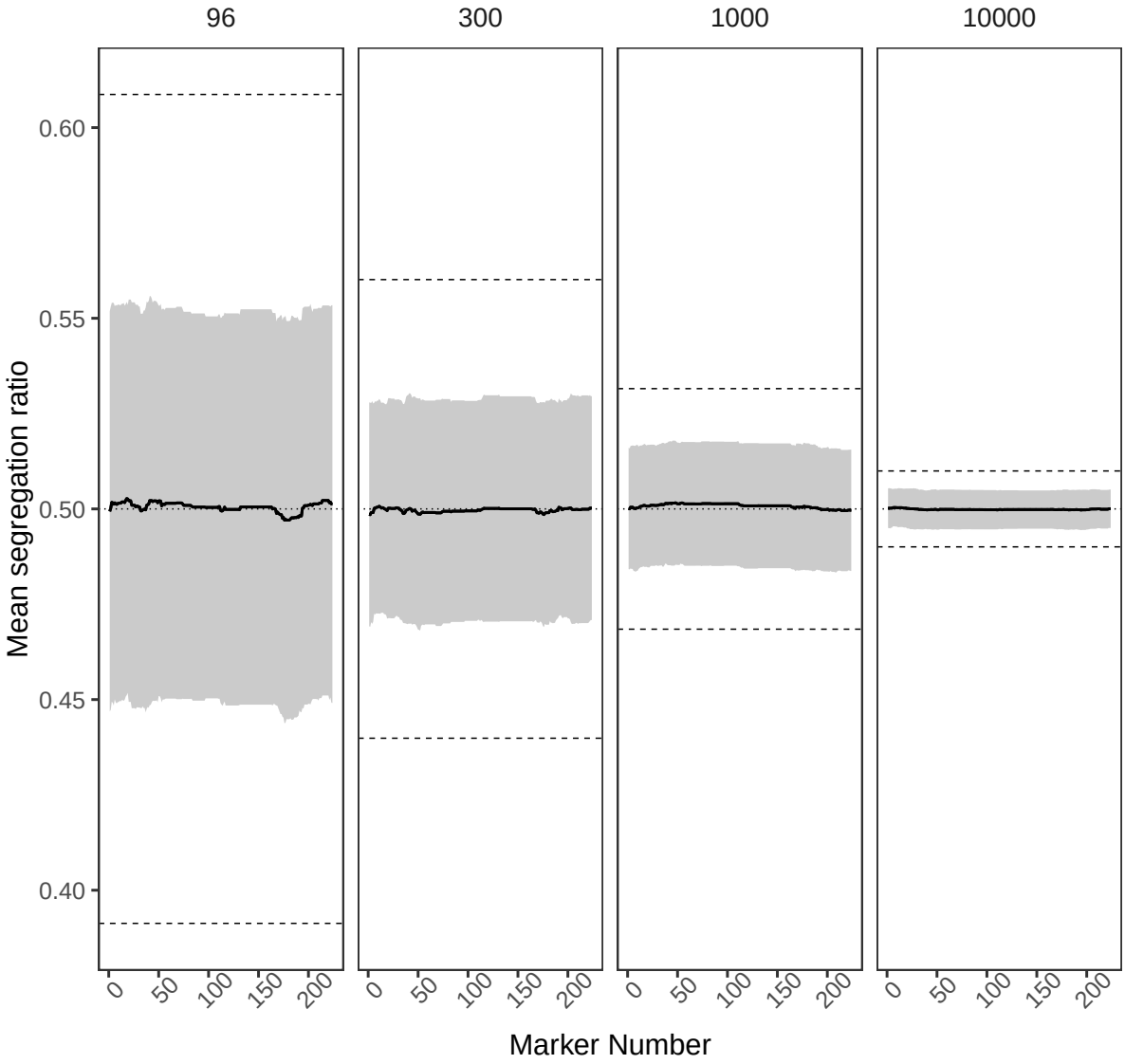


Figure 4.3 Effect of population size on the magnitude of distortion. Indicated in the header of each panel is the population size. Segregation ratio is calculated as (a)/(a + b), and each data point is the mean value over 1000 simulations. The simulations have no selection and use the marker distribution from chromosome 1A of the Cadenza x Avalon cross. The shaded area represents the mean segregation ratio value ± the standard deviation over 1000 simulations. The dashed lines mark the 5% significance threshold for a chi-square test, whilst the dotted line marks a 1:1 segregation ratio. The effect of sampling error on segregation ratio decreases as population size increases

Table 4.1 Measures of segregation distortion for simulations with 224 markers and marker distribution taken from chromosome 1A of a Cadenza X Avalon F2 cross. The last column indicates the mean value across all simulations of the magnitude of distortion at its highest value. Shown in the p-value columns are the number of simulations (out of 1000 performed) that contain significantly distorted markers. Marker set A refers to the full marker set of 224 markers, whilst marker set B refers to the skeleton marker set of 93 markers.

Population Size	Marker set	Filial Generation	P < 0.05	P < 0.01	P < 0.001	P < 0.05 (FDR Correction)	P < 0.05 (Bonferroni Correction)	Mean magnitude of peak distortion
96	A	2	561	162	16	18	4	0.1415
300	A	2	561	163	28	26	4	0.08122
1000	A	2	557	177	24	27	7	0.04366
10000	A	2	557	183	28	28	6	0.01377
96	A	6	602	179	27	30	7	0.1099
300	A	6	563	167	19	33	1	0.06153
1000	A	6	570	215	35	38	5	0.0343
10000	A	6	583	208	26	28	3	0.01085
96	B	2	561	162	16	22	8	0.1415
300	B	2	561	163	28	25	13	0.08122
1000	B	2	557	177	24	26	15	0.04366
10000	B	2	557	183	28	29	17	0.01377

In simulations containing a single marker for a population size of 1000, 5.2% exhibited significant segregation distortion, which is concordant with a chi-square alpha threshold of 0.05. Where two markers were placed near each other at centimorgan positions of 59 and 60, 5.4% of simulations contained markers with significant segregation distortion. Increasing the distance between these two markers by placing them at 20 and 60 centimorgans resulted in 9.2% of simulations containing markers with significant segregation distortion. The difference in number of simulations containing segregation distortion between these proximal and distal marker distributions was significant (?2 = 9.89, df = 1, p = 0.002).

To assess the effects of population size and selection strength on deviation from a 1:1 ratio of homozygous genotypes, and therefore segregation distortion, we ran a set of simulations in which both these parameters varied (figure 4.4). Population size ranged from 10 to 2000, whilst selection strength ranged from 1/20 to ½. As selection strength increases, the effect of population size on the deviation from 1:1 decrease. Simulations with population sizes less than 80 are very susceptible to distortion regardless of the selection strength.

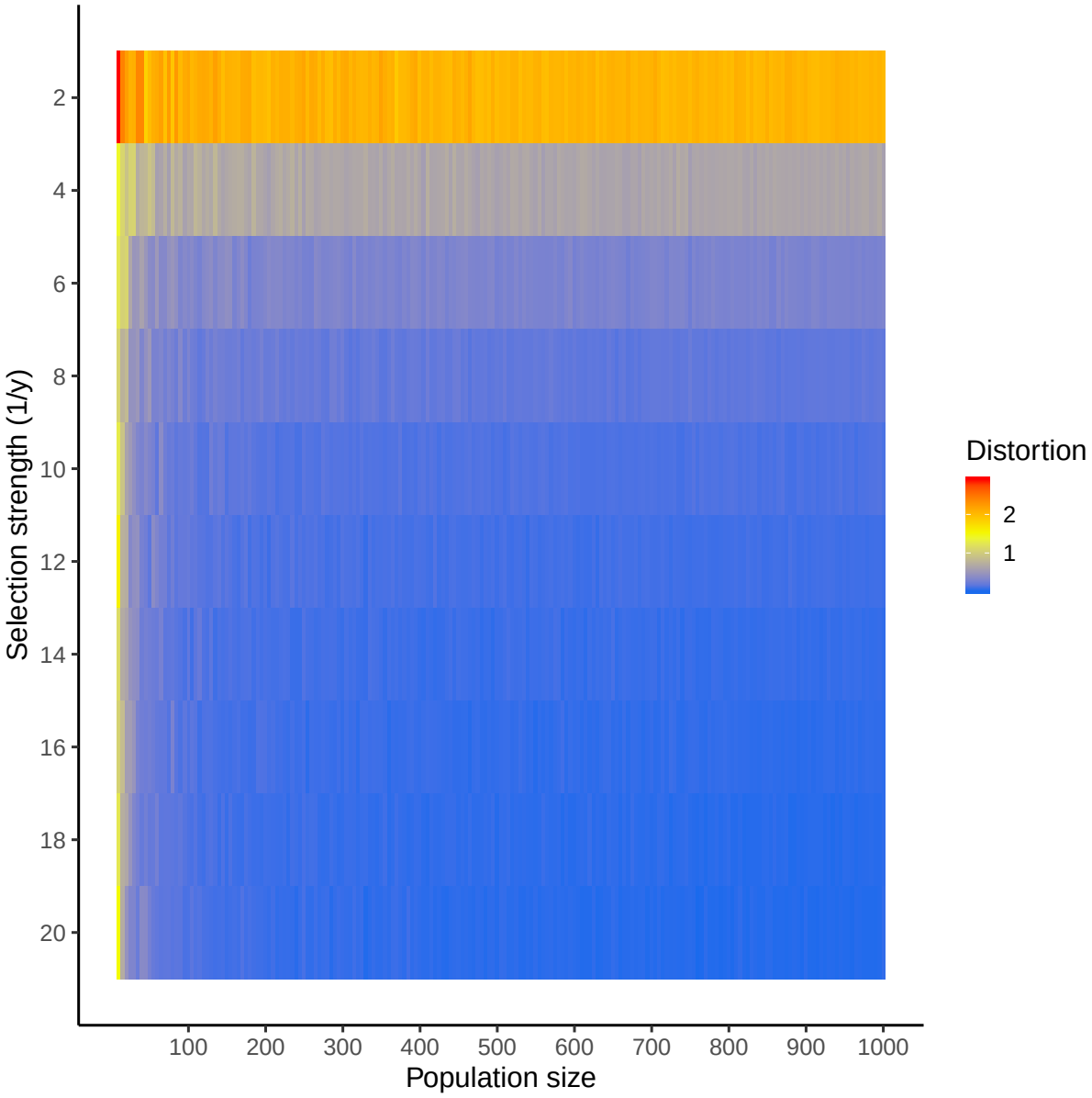


Figure 4.4 Heatmap of deviation from 1:1 segregation of homozygotes at various population sizes and selection strengths. Lower selection strengths are highly dependent on population size. As population size decreases, the influence of sampling error on segregation ratios increases, leading to high segregation distortion even in the case of weak selection. Each tile is an average value over 20 simulations with 20 markers at evenly spaced intervals, totalling 100 centimorgan. Deviation is calculated as Σ(y - 0.5)^2 where y is the ratio of homozygous genotypes (a)/(a + b) at an arbitrary locus; a is the number of homozygous genotypes from parent 1, b is the number of homozygous genotypes from parent 2 at an arbitrary locus

We examined the performance of various p-value thresholds and multiple testing procedures on the detection of segregation distortion at a range of selection strengths, with

a population of 1000 individuals per simulation (figure 4.5). 56.1% of simulations contained significantly distorted markers when a p-value threshold of 0.05 was used with no selection, compared to 3% at p-value thresholds of $p < 0.001$, $p < 0.05$ (FDR corrected) and $p < 0.05$ (Bonferroni corrected). As shown in Fig 4.5, $p < 0.001$ and $p < 0.05$ (FDR corrected) are almost equivalent for this distribution of markers. All the detection criteria reach saturation (100% of simulations having markers with significant segregation distortion) at a selection strength of 0.25. As expected, the Bonferroni test is strictest regardless of selection strength.

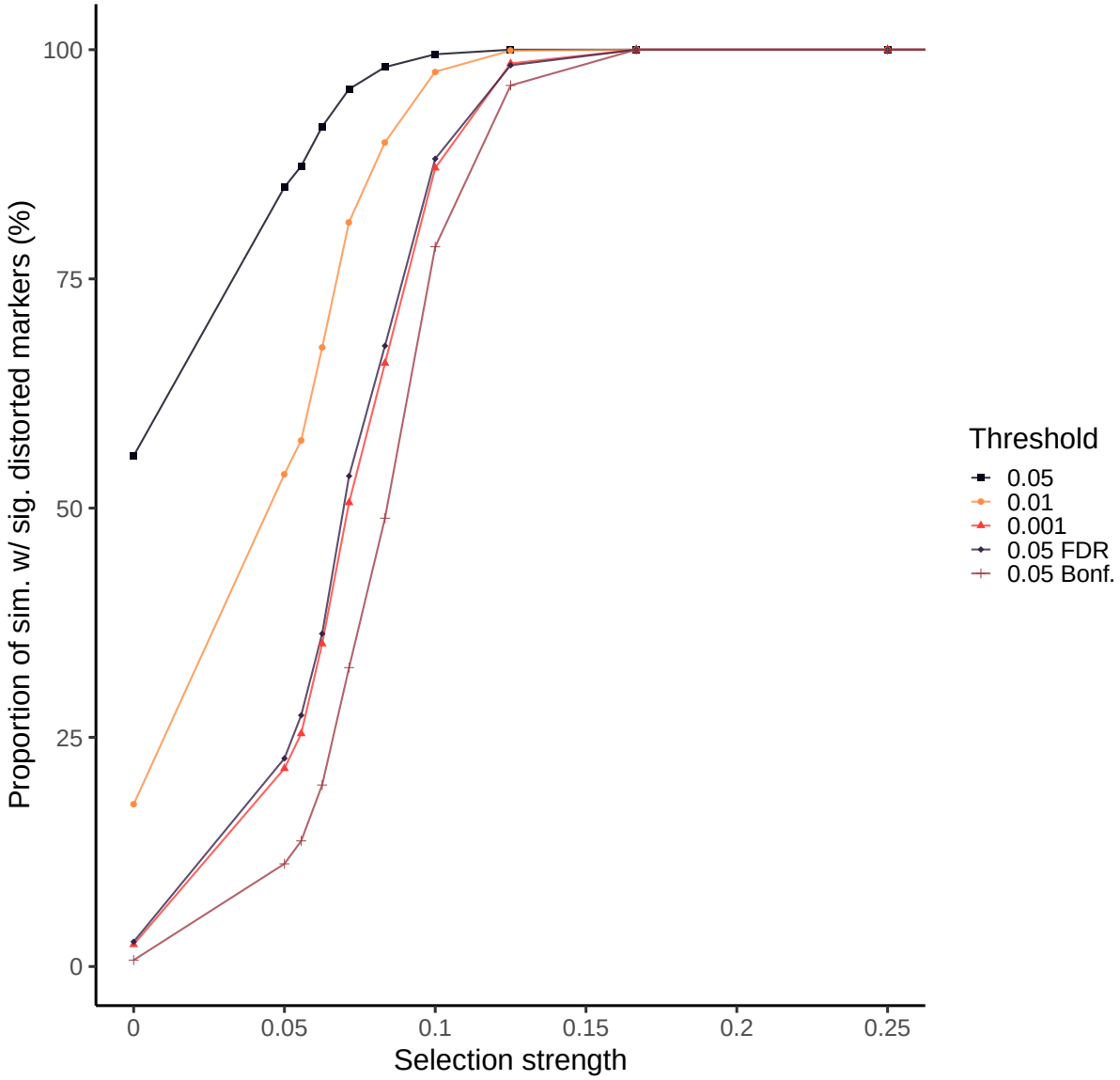


Figure 4.5 Proportion of 1000 simulations containing significantly distorted markers as a function of selection strength for various p-value threshold criteria. Simulations contain 1000 individuals and used the marker distribution of chromosome 1A from the Cadenza X Avalon F2 population. The position of selection was at locus 200 of 224 markers Sim = simulations, sig. = significant, pop. = population

In addition to the type of statistical test used, population size also influences the number of simulations exhibiting significant segregation distortion. As population size increases, so does the ability to reliably detect smaller selection strengths (figure 4.6). At a selection strength of 1/20, 60.3%, 66.2%, 85% and 100% of simulations contained markers exhibiting significant segregation distortion under a chi-square test with alpha threshold 0.05 for population sizes of 96, 300, 1000 and 10000 respectively (figure 4.7).

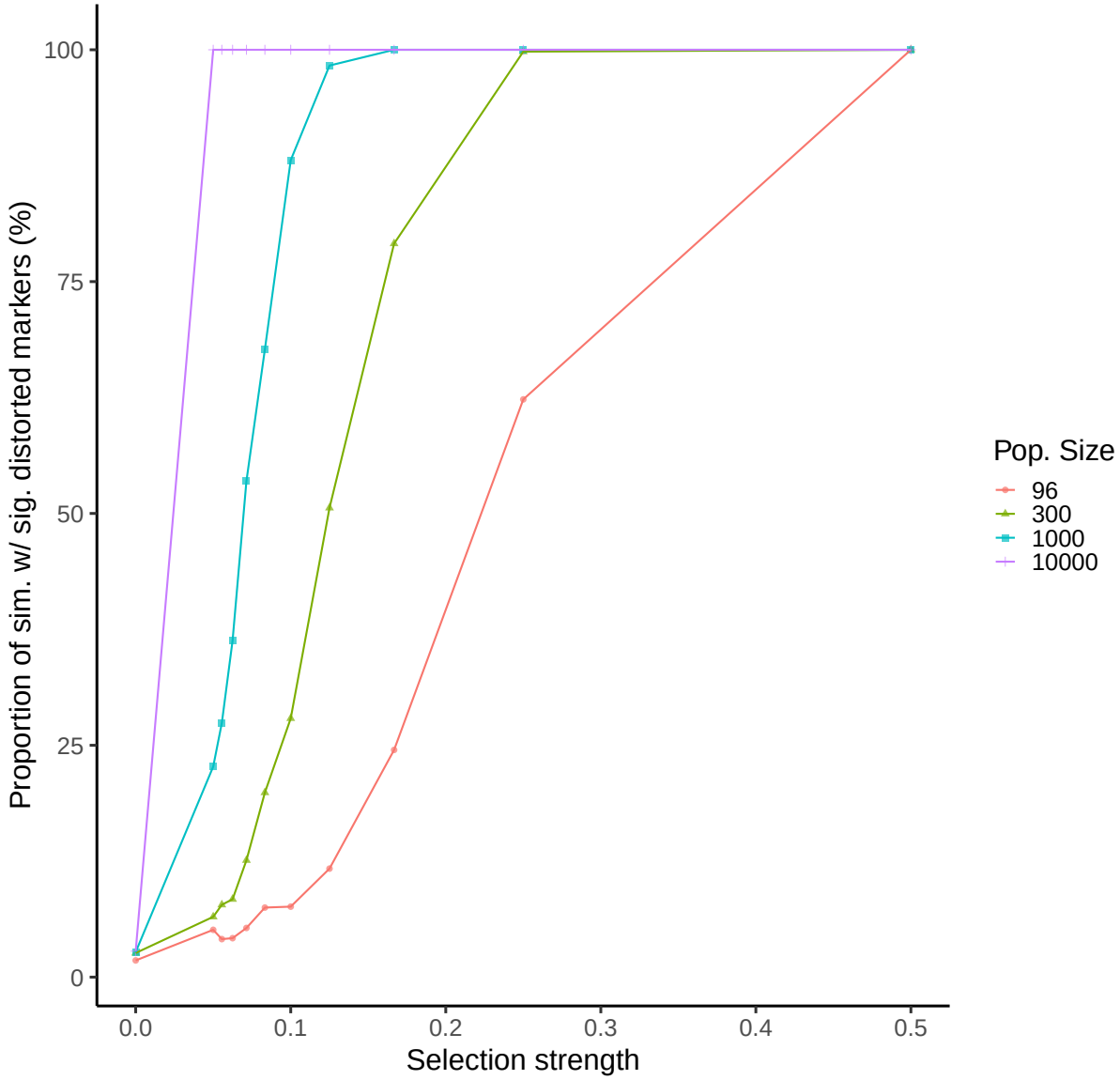


Figure 4.6 Effect of selection strength and population size on the number of simulations containing distorted markers (as determined by a chi-square test with significance threshold of 0.05 after correction for multiple testing with FDR). Sim = simulations, sig. = significant, pop. = population

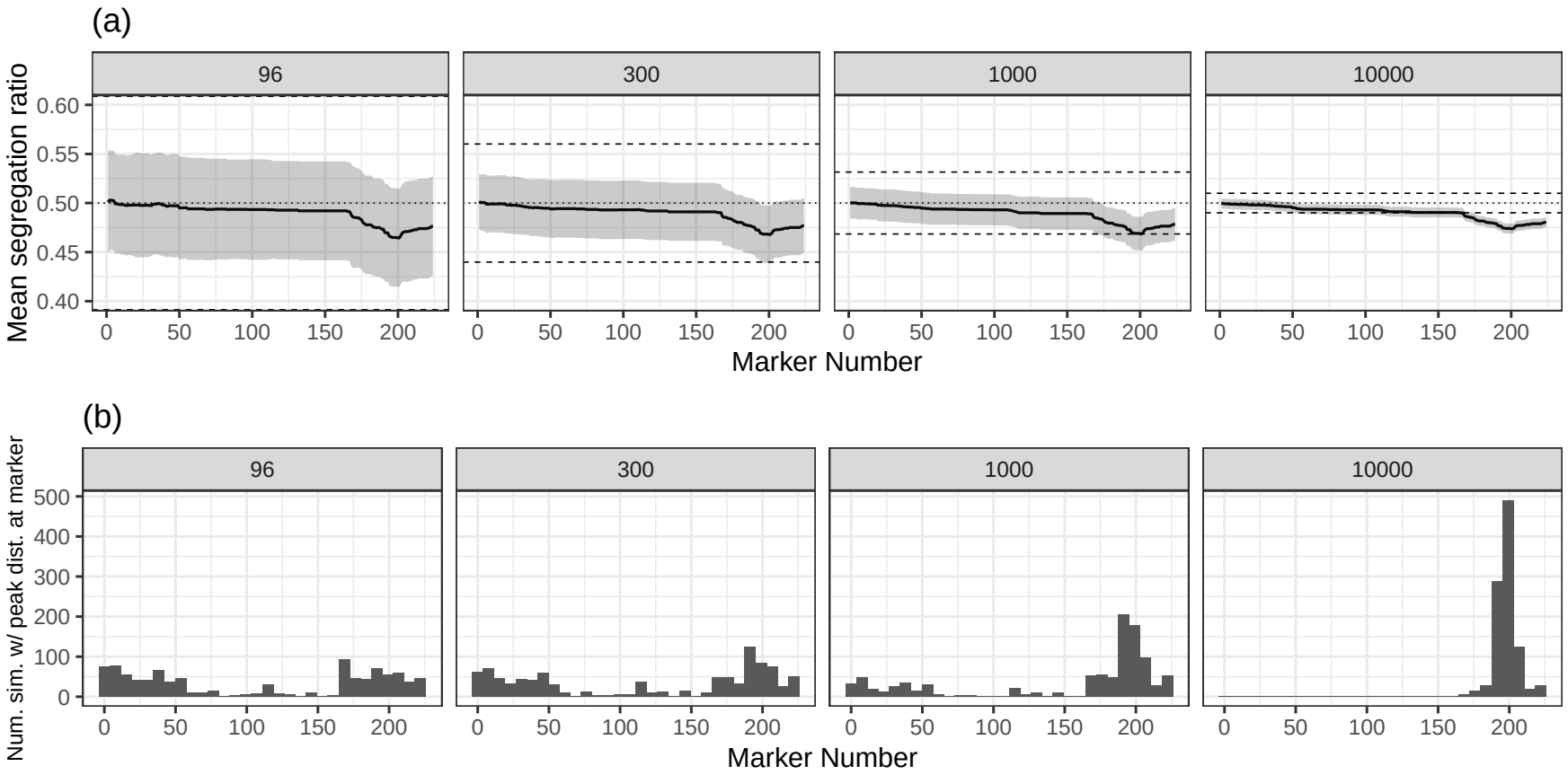


Figure 4.7 Simulation of an F5 RIL population with a selection pressure of strength 1/20 at locus 200. Indicated in the header of each panel is the population size. As the population size increases, the influence of sampling error on segregation of marker decreases, providing increased resolution of genuine selection events. (a) shows the mean magnitude of distortion ((a)/(a + b)) over 1000 simulations. The shaded area represents ± the standard deviation of the magnitude of distortion over 1000 simulations. The dashed lines mark the 5% significance threshold for a chi-square test, whilst the dotted line marks a 1:1 segregation ratio. (b) shows the number of simulations in which the peak of distortion occurs at the specified marker. As population size increases, so do the number of simulations in which the genuine selection event emerges as the peak of distortion. Num. = Number, sim. = simulations, dist. = distortion.

To test whether local recombination rate in the region of selection effected the detection of segregation distortion, we performed additional simulations with selection at marker 100 of chromosome 1A. This marker is located in a region of low recombination (figure 4.8), which contrasts with previous simulations where selection was at marker 200, located in a region of high recombination. The only statistical test that was affected by recombination rate in the region of selection was the FDR correction, which was consistently more powerful at all values of selection strength for population sizes 96, 300 and 1000 (figure 4.9).

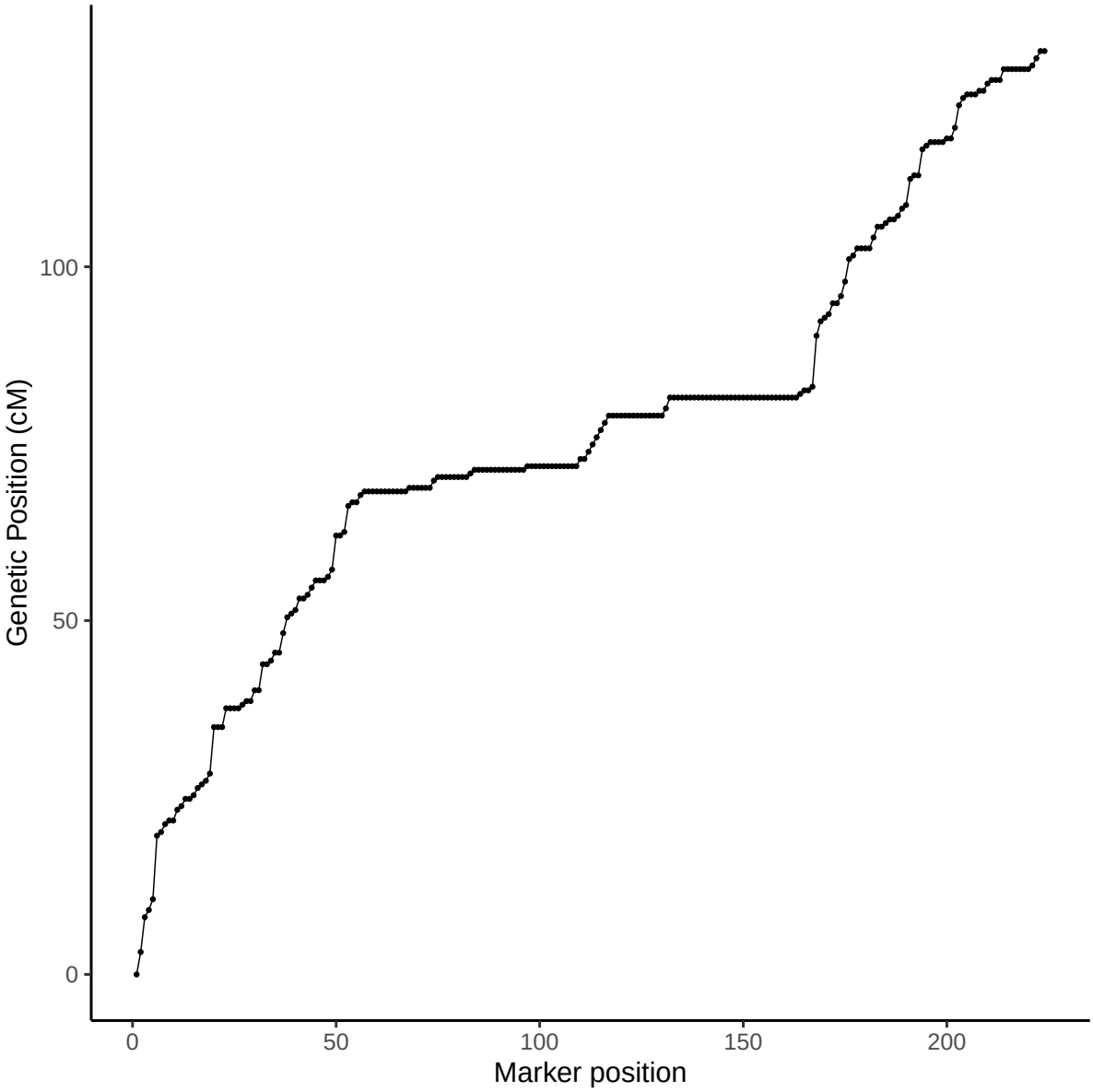


Figure 4.8 Recombination for chromosome 1A of the Avalon X Cadenza cross. The amount of recombination is represented by the slope of the line.

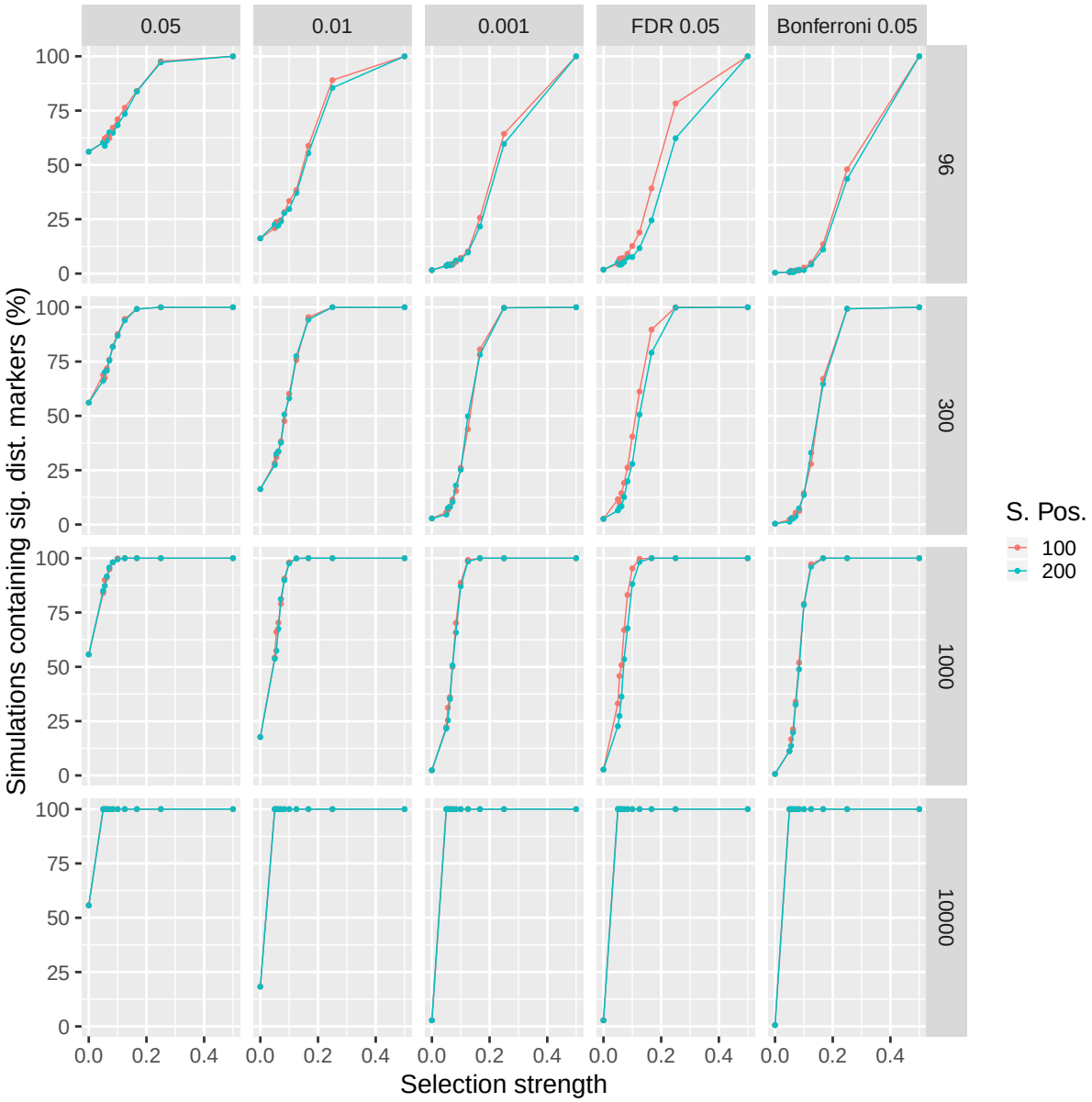


Figure 4.9 Examining the effect of selection position on the number of simulations containing significantly distorted markers. Position 100 is in a region of low recombination, whereas position 200 is in a region of high recombination. Columns are separated by the type of statistical test performed, whereas rows are separated by population size. The FDR correction has consistently more power when selection is in position 100. sig. = significant, dist. = distorted, S. = selection, Pos. = position.

These results contrast with the effect of local recombination rate on the detection of segregation distortion regions (SDRs), defined here as 2 or more consecutive markers exhibiting significant segregation distortion. The total number of SDRs generated among all 1000 simulations is generally higher when selection is positioned at marker 100 com-

pared to marker 200 (figure 4.10 (a)). An exception to this occurs at very high selection strengths (0.5), as these cause the entire chromosome to become one SDR at position 100, resulting in fewer, albeit larger in number of markers, SDRs overall. Changing the measure of SDRs to the number of simulations containing at least one SDR (figure 4.10 (b)). causes both selection positions to perform almost analogously when using no multiple testing correction and a detection threshold of $p < 0.001$. In comparison, the FDR correction for multiple testing results in a greater difference between positions.

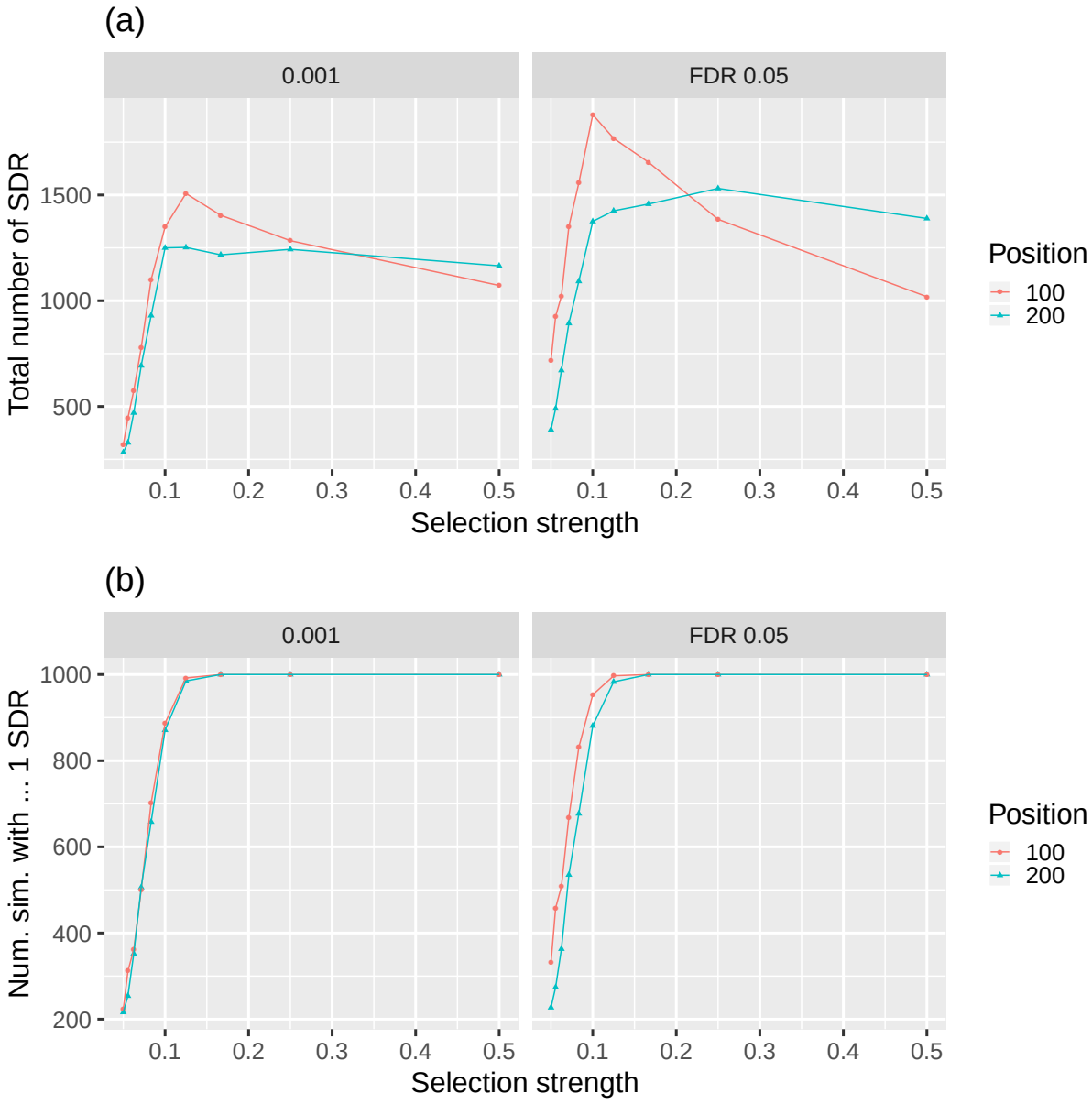


Figure 4.10 Comparison of the effect of selection position on formation of segregation distortion regions (SDRs). Selection position 100 is in a region of low recombination, whereas position 200 is a region of high recombination. (a) shows the total number of SDR among 1000 simulations as a function of selection strength, whereas (b) shows the number of simulations with at least 1 SDR. Shown in the panel titles are the thresholds / type of statistical tests used to detect segregation distortion. num. = number, sim. = simulations.

To examine the effect of selection strength on the position of maximum distortion, we performed simulations with a selection pressure at locus 200 of 224, using the marker distribution from Cadenza X Avalon 1A and a population size of 300 individuals. Selec-

tion strength ranged from 1/20 to ½. As expected, the number of markers exhibiting significant segregation distortion increased with the selection pressure, with mean values of 30.83 and 177.45 at selection pressures of 1/20 and ½ respectively. The percentage of simulations at which the peak of segregation distortion occurred within 10 markers of the applied selection pressure was52% for a selection pressure of 1/10, which gradually increased: 64.9%, 78.2%, 92.7% and 99.7% for pressures of 1/8, 1/6, 1/4, and ½ respectively. This was also affected by population size, with higher population sizes having an increasing number of simulations exhibiting peak segregation distortion at the selection locus when selection strength was fixed to 1/20 (figure 4.7 (b)).

4.4.3 Effect of segregation distortion on genetic mapping

In the simulations with a single selection pressure of strength 1 at locus 30 of chromosome 6B and no selection pressure on 1A, yielding a genotype ratio of 0:0:300 at the locus under selection, MSTMap was able to construct the genetic map with perfect clustering and ordering of marker bins using a clustering parameter of 10-43. The second simulation contained two selection pressures, one positioned at marker 30 of chromosome 6B and the other positioned at marker 200 of chromosome 1A, both favouring the same parental genotype. MSTMap was able to construct genetic map with perfect clustering and ordering of marker bins (using clustering parameter 10-45) up to a selection pressure of 1/1.11, which yielded genotype ratios of 1:31:268 (test of deviation from 1:2:1 ratio: ?2 = 664.07, df = 2, p < 10-15) and 2:32:266 (test of deviation from 1:2:1 ratio: ?2 = 650.29, df = 2, p < 10-15) for the markers under selection respectively. When the selection strength for this configuration was increased to 1/1.105, yielding genotypes ratios of 0:27:273 and 1:28:271 respectively, MSTMap was unable to cluster markers correctly for any of the clustering parameters tested, which ranged from 10-40 to 10-50. For example, using a clustering parameter of 10-45 yielded two linkage groups, the first consisting of markers 1 to 167 of chromosome 1A, the second consisting of a concatenation of 1A markers 168 to 223 and all the markers on 6B (figure 4.11).

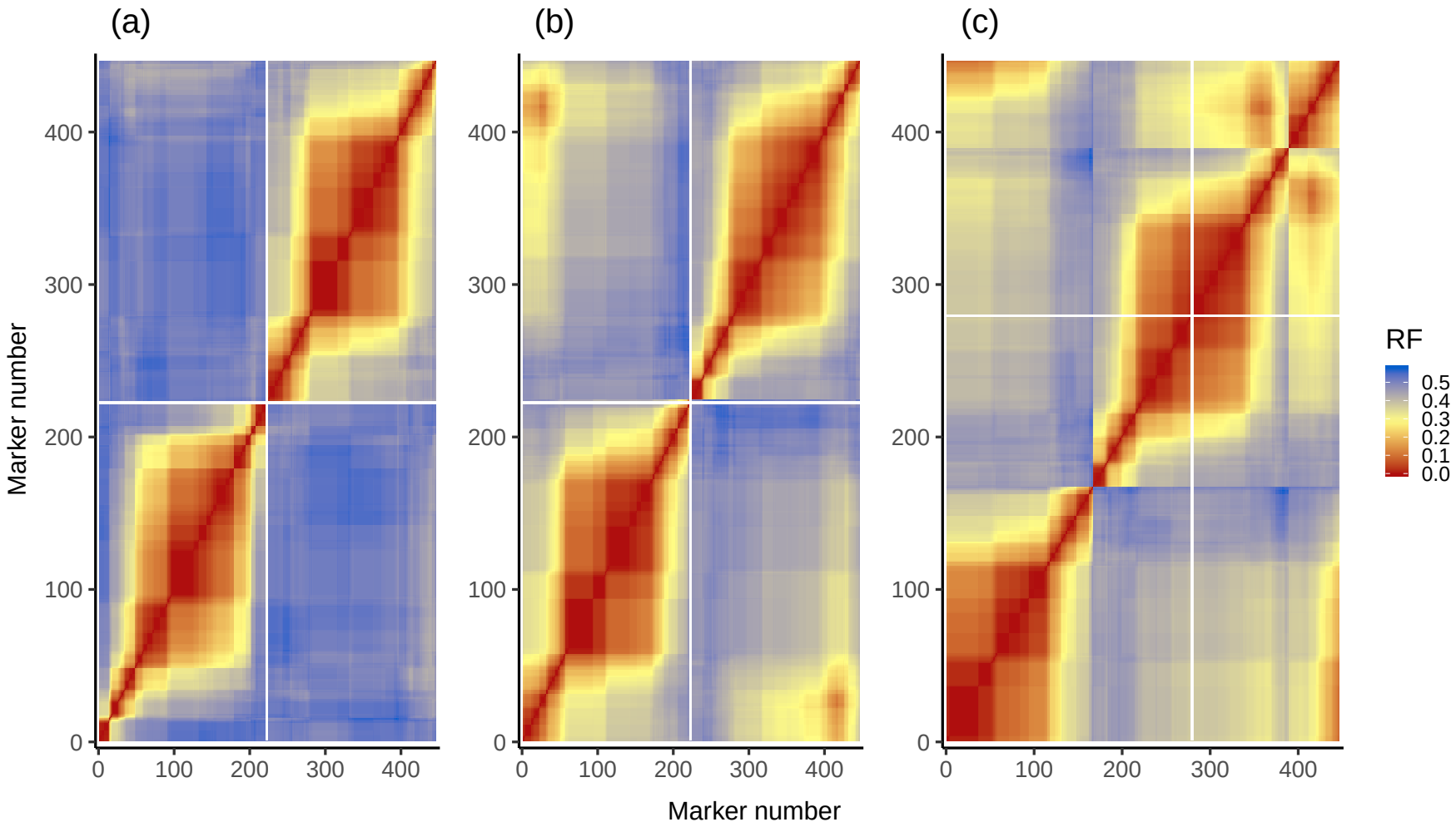


Figure 4.11 Effect of segregation distortion on genetic mapping. (a) When no selection is applied, two linkage groups representing simulated chromosome 1A and chromosome 6B are formed with perfect order of marker bins. (b) When a strong selection pressure of 1/1.11 is applied at locus 30 and 200 of 6B and 1A respectively, the recombination fractions between markers at these loci and surrounding markers are lowered, but not enough to disrupt clustering or ordering of the markers. (c) When a strong enough selection pressure of 1/1.105 is applied such that one of the loci has zero A genotypes, the recombination frequencies of markers under selection are low enough such that chromosomal fragments experiencing segregation distortion are concatenated into the same linkage group. White lines delineate linkage groups

Moving the position of one of the two selection pressures from a region of high recombination (marker 200 on chromosome 1A, (figure 4.8)) to a region of low recombination (marker 100 on chromosome 1A, (figure 4.8)), had little effect on genetic map construction, with MSTMap producing a map with perfect clustering and ordering of marker bins at a selection strength of 1/1.2, yielding genotype ratios of 2:50:248 and 2:55:243 at marker 30 of chromosome 6B and marker 100 of chromosome 1A respectively. MSTMap was unable to cluster markers correctly when selection strength was increased to 1/1.1, yielding genotype ratios of 0:29:271 and 0:31:269 for the respective markers under selection. When the position of selection on chromosome 6B was also moved to a region of low recombination (from marker 30 to marker 110, (figure 4.12)), MSTMap produced a map with perfect clustering and ordering of marker bins at a selection strength of 1/1.2, and failed to cluster markers correctly when the strength was increased to 1/1.1.

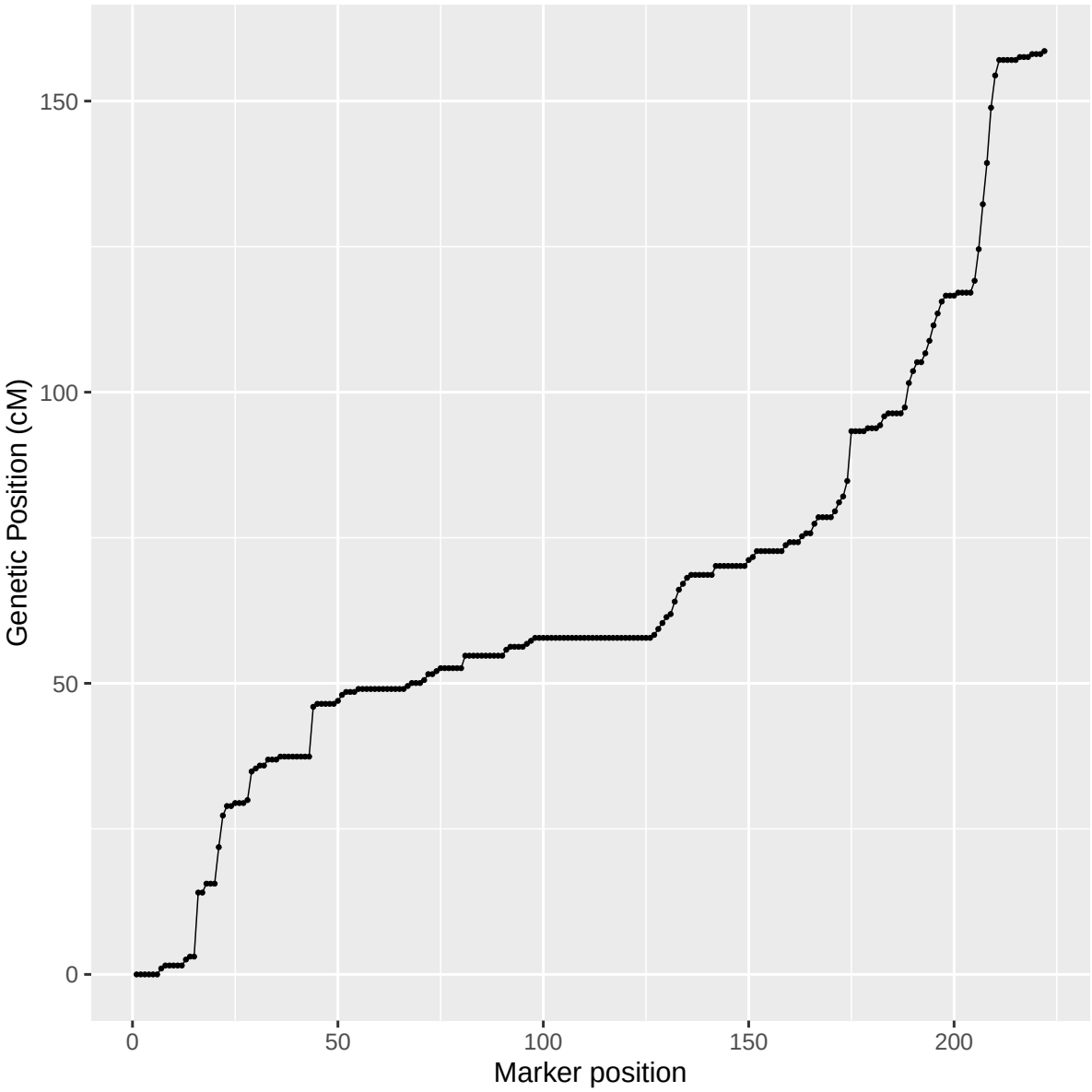


Figure 4.12 Recombination for chromosome 6B of an Avalon X Cadenza cross. The amount of recombination is indicated by the slope of the line.

Similarly to the results from the F2 population, in a simulation of an F8 SSD population with the same selection positions and a selection strength of 1/1.3, which yielded AA:BB genotype ratios of 28:272 (chi-square test of deviation from 1:1 ratio, ˆ2 = 198.45, df = 1, p < 10-15) and 31:269 (chi-square test of deviation from 1:1 ratio, ˆ2 = 188.81, df = 1, p < 10-15) for the respective markers under selection, MSTmap was able to produce perfect clustering and ordering of marker bins with a clustering parameter of 10-42.

Extreme segregation distortion caused a significant shortening of map length for simulated chromosome 6B (t-test, t = -27.57, df = 176.43, p < 10-15) by around 20 cM, with a selection pressure of 1 producing a map length of 140.14 ± 3.92 compared to 159.07 ± 5.64 with no selection applied. Less extreme selection pressures of 1/3, 1/5, 1/7 and 1/9 produced mean map lengths over 100 simulations of 153.04 ± 4.64, 155.08 ± 4.4, 155.39 ± 4.68 and 156.77 ± 3.95 respectively. Likewise, for simulated chromosome 1A, extreme distortion shortened the map length significantly (t-test, t = -5.17, df = 197.72, p < 10-6), but with a smaller effect size than for 6B, with selection pressure of 1 producing a mean map length of 127.39 ± 4.81 compared to 130.84 ± 4.64 cM with no selection.

4.4.4 Reanalysis of existing data

We reanalysed data from [@allenCharacterizationWheatBreeders2016; @avniUltra-denseGeneticMap2014], both of which used the minimum chi-square threshold of p < 0.05 to detect regions of segregation distortion (table 4.2). As expected, in both cases we observe a large reduction in the number of markers exhibiting significant segregation distortion when corrections for multiple testing are applied.

Table 4.2 Reanalysis of genotyping data from existing studies with corrections for multiple testing. Indicated in columns 3–5 are number of markers exhibiting significant segregation distortion with no correction for multiple testing, the FDR correction and the Bonferroni correction respectively.

Author	Mapping Population	P < 0.05	P < 0.05 (FDR)	P < 0.05 (Bonferroni)
Allen et al., 2016	Avalon X Cadenza	487	5	5
‘ ’	Savannah X Rialto	230	0	0
‘ ’	Opata X Synthetic	346	0	0
‘ ’	Apogee X Paragon	320	35	21
‘ ’	Chinese Spring X Paragon	774	0	0
Avni et al., 2014	Svevo X Zavitan	3789	1771	150

Markers that were still classified as significantly distorted in the Avalon X Cadenza population under Bonferroni correction were located on chromosomes 2A and 2D, whilst in the Apogee X Paragon population these were found on chromosomes 2D, 3B, 6A and 6B. Likewise, for the Svevo X Zavitan population, markers still significantly distorted under Bonferroni correction were found on chromosomes 2B and 3B.

4.4.5 Cadenza X Avalon Replicates

In the Cadenza X Avalon F2 replicates, there are 453 (14.88%) markers that exhibit significant segregation distortion ($p < 0.05$) in at least one of the replicates. Only 14 markers showed significant segregation distortion in both replicates. When both datasets were combined, 253 markers showed significant distortion. In the combined dataset, 173 of the 253 distorted markers were also distorted in one of the two original replicate datasets, whilst 80 were not. In the first and second replicates, 187 and 280

markers exhibited significant distortion respectively. In the first replicate, there were 22 SDRs, comprised of 161 markers in total. The mean ± sd length of the SDRs was 7.32 ± 5.57 markers. In the second replicate, there were 31 SDRs comprised of 238 markers in total. The mean ± s.d. length of the SDRs was 7.68 ± 10.46. Three of the SDRs on chromosomes 1D, 5B and 1A respectively overlapped between replicates; the lengths of the overlaps were 4, 2 and 2 markers respectively.

4.4.6 Avalon X Cadenza Replicates

In the Avalon X Cadenza F2 replicates, there are 510 (16.75%) markers that exhibit significant segregation distortion ($p < 0.05$) in at least one of the replicates. Only 38 markers showed significant segregation distortion in both replicates. When both datasets were combined, 173 markers showed significant distortion. In the combined dataset, 120 of the 173 distorted markers were also distorted in one of the two original replicate datasets, whilst 53 were not. In the first and second replicates, 193 and 355 markers exhibited significant distortion respectively. In the first replicate, there were 15 SDRs comprised of 155 markers total. The mean ± s.d. length of the segregation distortion regions was 10.33 ± 8.81. In the second replicate, there were 20 SDRs comprised of 328 markers total. The mean ± s.d. length of the segregation distortion regions was 16.4 ± 32.49. Six of the SDRs overlapped between replicates, these were all located on chromosome 6B and had widths of 8, 4, 8, 2, 4 and 4 markers respectively, (figure 4.13). The overlapping region did not have a skew towards the same parental genotype in each replicate.

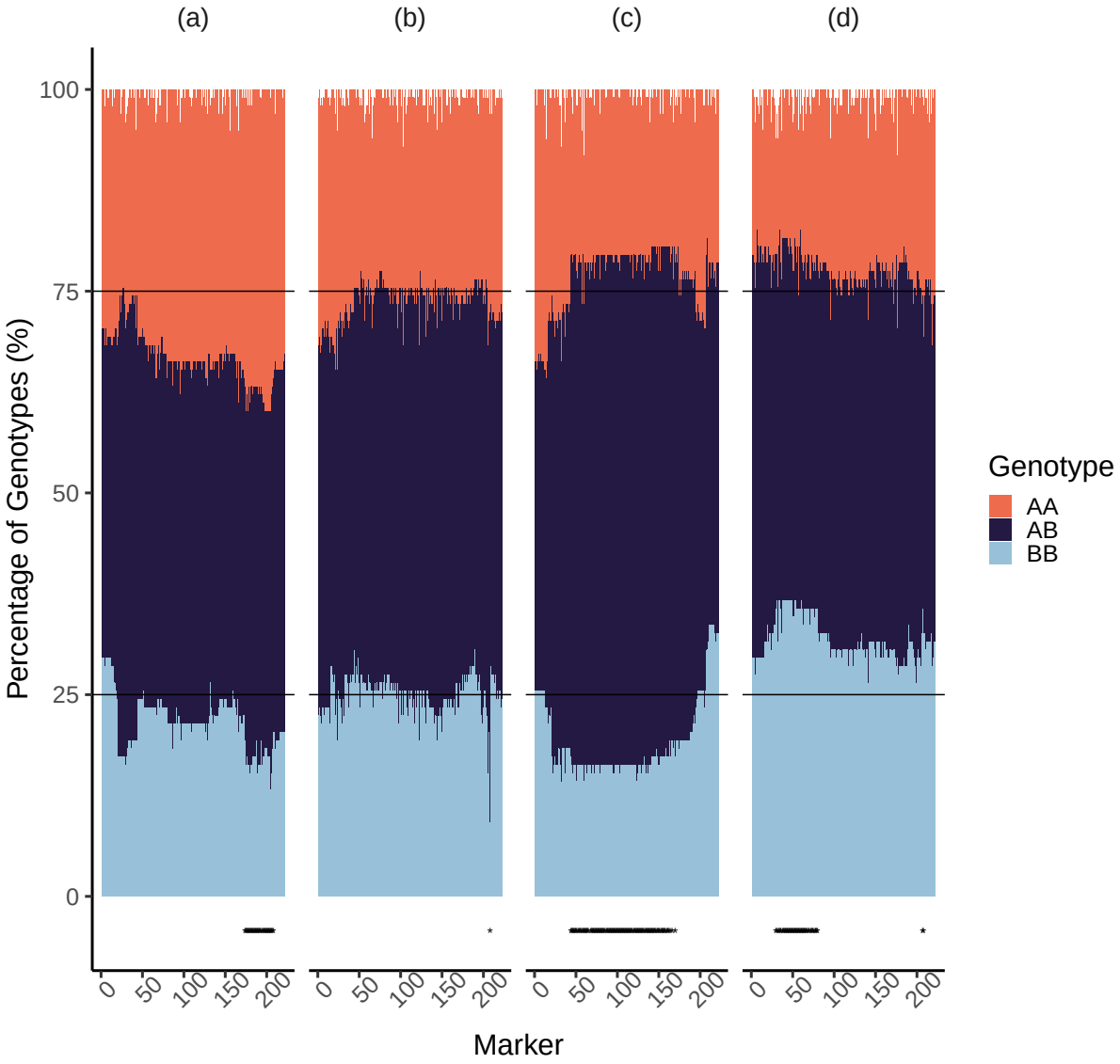


Figure 4.13 Comparison of segregation ratios on chromosome 6B for Cadenza X Avalon (a, b) and Avalon X Cadenza (c, d) replicates. Markers exhibiting significant segregation distortion as determined by a chi-square goodness of fit test for deviation from a 1:2:1 ratio of AA:AB:BB genotypes are highlighted by asterisks at the base of each plot. Black horizontal lines mark the expected transition from one genotype to the next under a 1:2:1 ratio. Markers are ordered on the x-axis as in the genetic map produced from the first replicate Cadenza X Avalon population

4.5 Discussion

Comparisons between the simulated data and the empirical data from the Cadenza X Avalon F2 mapping population show that the simulator is accurate in terms of map

length produced, number of recombination events per individual, degree of segregation distortion and recombination frequency heatmap. We can therefore be confident that the subsequent simulation experiments are an adequate representation of reality.

It is clear a priori that when we test for segregation distortion, the probability of a false-positive result increases with the number of markers, as multiple inferences are being made simultaneously. What complicates the calculation of how much more likely a false-positive result is with increasing number of markers, and therefore how strict our correction for multiple testing should be, is the marker distribution along the chromosome. All markers are ultimately linked together to varying degrees by the process of recombination, so not all the statistical tests performed are completely independent. Markers adjacent to each other at short genetic intervals along the chromosome segregate in a highly linked manner. This is confirmed by our simulation experiments with different distributions of two markers, one in which the markers are close together and one in which they are further apart, with the latter yielding almost double the amount of simulations containing markers that were significantly distorted. Interestingly, the chi-square test performs as expected when only one marker is used on the chromosome with no selection, with around 5% of simulations showing a false-positive result, which corresponds to the traditional alpha threshold of 0.05.

The FDR correction procedure reassuringly produced an alpha threshold that is only slightly stricter than the traditional 5% in the simulated chromosome 1A described earlier, as shown in table 1. The Bonferroni correction is only appropriate when statistical tests are completely independent from one another, which is not the case for highly linked markers. The Bonferroni test would therefore be appropriate if markers were on different chromosomes, or if they were located at large distances from each other on the same chromosomes. For high-density SNP data obtained from microarrays this is often not the case, and therefore the Bonferroni correction is often too strict, as shown by the results in table 1, where in the F2 population of 300 individuals only 4 simulations show significant segregation distortion, where we would expect around 50 if the test corresponded to the usual 0.05 alpha threshold. The fact that 56% of simu-

lations without any selection pressure show significant segregation distortion according to the chi-square test at the minimum p-value threshold ($p < 0.05$) should definitively rule out the use of chi-square without correction for multiple testing, or inclusion of lower thresholds, in future studies that utilize high-density genotyping data. The perfect multiple testing procedure for segregation distortion would be one that considers the distribution of markers on the chromosome such that the alpha threshold is adjusted depending on the degree of linkage between each marker. However, seeing as the FDR correction for multiple testing is only marginally more conservative than the traditional alpha threshold of 0.05, and taking into account the fact that the traditional alpha threshold was chosen arbitrarily [@wassersteinMovingWorld052019], the use of FDR as a new standard for the detection of segregation distortion seems appropriate.

One focus in the literature is the identification of segregation distortion loci. Our simulation experiments with a range of selection strengths show that resolution of selection events is increased with population size. This is because the effects of sampling error are neutralized as population size increases. Sampling error could lead to the erroneous conclusion that a segregation distortion locus is present by shifting the segregation of a marker away from expected Mendelian ratios. Conversely, it can make markers under true selection pressures appear as normally segregating. It can also skew the peak of segregation distortion away from a true selection locus at smaller selection strengths. To correctly identify the causative locus in this case then would require a wider search than is initially implied by the segregation data. These results emphasize the significance of sampling error in segregation distortion studies. In addition, our results show that local recombination rates in the region of selection have little influence on the detection of segregation distortion.

It has long been known that segregation distortion effects the estimation of recombination fraction between markers [@lorieuxMaximumlikelihoodModelsMapping1995]. However, there does not seem to be a practical guide in the literature that can assist researchers in knowing what degree of distortion will affect the mapping process. Our simulation experiments on the effect of segregation distortion on genetic mapping

show that only very extreme distortion effects the formation of linkage groups during the clustering of markers, as well as map length, meaning that markers experiencing moderate distortion can be retained in genetic maps. This conclusion persists regardless of selection position, whether in a region of high or low recombination. This result will be useful to future studies, as markers that would have previously been discarded will give us more information on potentially useful genomic regions of many crop species. In the re-analysis of data from Allen et al [@allenCharacterizationWheatBreeders2016] and Avni et al [@avniUltradenseGeneticMap2014], it is interesting to note that the latter had many more markers exhibiting segregation distortion both before and after corrections for multiple testing. The former study used varieties of bread wheat (*Triticum aestivum* L.) whilst the latter involved a cross between durum wheat (*Triticum turgidum* L. subsp. durum) and a wild relative of durum wheat, wild emmer (*Triticum turgidum* L. subsp. dicoccoides). It has been noted elsewhere in the literature that the degree of segregation distortion often increases with genetic distance of the parents [@liuProgressSegregationDistortion2010]. One hypothesis is that with increasing genetic distances, the fitness benefits conferred to the progeny of biparental crosses become increasingly different between parental alleles. If this is indeed the case, the description of a true segregation distortion locus in closely related crop varieties, including its mechanism of action, is a much more difficult task than in more distantly related crosses. Indeed, our best descriptions of segregation distortion loci are from crosses between rice (*Oryza sativa* L.) subspecies indica and japonica [@yangKillerProtectorSystemRegulates2012], as well as *Drosophila pseudoobscura* subspecies pseudoobscura and bogotana [@phadnisSingleGeneCauses2009]. To detect a true segregation distortion locus in closely related wheat varieties then would require population sizes large enough to detect much smaller selection strengths, as indicated in figure 4.6, as well as replicate populations to confirm the effect on segregation is due to selection. An exception to this statement may be in the production of doubled haploid mapping populations, where differences in amenability to doubled haploidy between closely related varieties has the potential to produce segregation distortion that is stronger than in an SSD population structure [@sayedSegregationDistortionDoubled2002].

When identifying segregation distortion in empirical populations, it is often convenient to assess segregation in terms of SDRs, as multiple consecutive markers exhibiting significant segregation distortion provide us with more confidence that the distortion observed is not due to erroneous genotype assignment. The fact that we only observed a few SDR overlaps that were distorted towards the same parent between replicates in our empirical populations shows that legitimate segregation distortion between varieties of wheat is rare. Our simulation experiments also confirm the intuitive deduction that the number of SDRs should increase when selection occurs within regions of low recombination. In conclusion, the results presented here emphasize the importance of using appropriate statistical methods when detecting segregation distortion. We must be sure that the observed distortion is due to a genuine selection pressure before we can commence further research into identifying the loci that are driving the distortion. We recommend that studies utilizing high-density genotyping data use an FDR correction for multiple testing when checking for segregation distortion, and that population size should be as high as possible to increase the chances of discovering genuine segregation distortion loci. Figure 4.6 serves as a guide for the appropriate population size to detect various selection strengths. For example, to reliably detect a selection strength of 1/10 at the 0.05 p-value threshold, a population size slightly bigger than 1000 individuals is required. As a result of our reanalysis of existing data based on these principles, we have discovered a candidate segregation distortion region on chromosome 3B of a cross between wheat varieties Apogee and Paragon that is likely to be caused by a genuine selection event. We hope that future studies of segregation distortion will also consider the findings presented here.

5 AutoCloner: automatic primer design for full-gene cloning in polyploids

All results described here have been published in the folllowing manuscript:

Coulton, A., Edwards, K., 2020. AutoCloner: automatic homologue-specific primer design for full-gene cloning in polyploids. BMC Bioinformatics

5.1 Introduction

Polymerase chain reaction (PCR) is a procedure that allows the amplification of small amounts of DNA into millions or billions of copies, originally conceived by Kary Mullis in the 1980s [@mullisUnusualOriginPolymerase1990]. There are four primary reagents required for the PCR. The first of these is the template sequence to be amplified, which is usually obtained using a simple DNA extraction procedure (e.g. [@edwardsSimpleRapidMethod1991]). In addition, two short oligonucleotide sequences, ranging from 15 - 30 bases in length, are also required. These must be complementary to the template sequence, as well as positioned such that they are flanking the template sequence. The first hybridizes at the start of the sequence to the sense strand of the DNA double helical structure, whilst the second hybridizes at the end of the sequence, to the anti-sense strand. For this reason these oligonucleotides are referred to as the forward primer and the reverse primer respectively. The third key reagent for PCR is a thermostable DNA polymerase, originally isolated from thermophilic bacterium *Thermus aquaticus*. Finally, PCR also requires a pool of free nucleotides that serve as base material for DNA synthesis.

The reaction itself consists of three stages: denaturation, annealing, and extension. The basis of change between these stages is a change in temperature of the reaction tube, done using a thermocycler machine. During the denaturation stage, temperature increases to around 94°C for 1 minute, which is enough to separate the individual strands of the DNA double helix, and thus making them free for hybridization to their respec-

tive primers [@lorenzPolymeraseChainReaction2012]. Following this, temperature is lowered to around 52°C for 30 seconds to allow annealing of primers to the template DNA. Finally, during the extension stage, temperature is raised to around 75°C, which is the optimal temperature for the thermostable DNA polymerase to synthesize the new DNA strands. These three stages are cycled through and repeated many times during the PCR, such that by the end of the reaction, the concentration of the target DNA outnumbers the concentration of the sample DNA by many orders of magnitude.

Cloning of genetic sequences via polymerase chain reaction (PCR) is a routine operation in biological research. In agricultural research specifically, this procedure facilitates the connection between varietal sequence differences and important phenotypic traits such as disease resistance, yield, and abiotic stress tolerance. This process is significantly complicated in polyploid crops due to the presence of multiple closely-related subgenomes, meaning that allele-specific primers must be used to prevent cloning of non-target sequences such as homeologues and paralogues. Although there already exists a tool for designing primers for use in Kompetitive allele specific PCR (KASP) assays in polyploids, PolyMarker [@ramirez], this only considers flanking sequences of 100 bases either side of a varietal SNP. This limitation means that it cannot used to clone entire genes, as the mean ± s.d. length of a high-confidence gene in the IWGSC RefSeq v1.0 [@consortiumiwgscShiftingLimitsWheat2018] wheat genome assembly is 3065 ± 3957 bases. There are currently no software packages to assist allele-specific primer design for the cloning of entire genes or other genomic sequences of interest, and indeed current practice within the community is to carry out this lengthy process manually [@wheat-training.comDesigningGenomeSpecific; @babbenEfficientApproachDevelopment2015; @babbenAssociationGeneticsStudies2018].

For example, consider the situation in which a researcher has a gene sequence from a single wheat variety and is interested in how this sequence differs between varieties. To assess this, they could design several pairs of allele-specific primers whose products overlap, covering the entire gene region, and then sequence these products after performing PCR. This primer design process involves several stages. First, the wheat

genome must be queried for closely related alleles to the sequence of interest. Once homologues have been identified and extracted, they must be arranged into a multiple sequence alignment. This alignment must then be scanned for SNPs to serve as the 3‘ bases of primers, which can then be designed using the appropriate primer-design software. In total, this is a lengthy process that would be significantly improved via the use of an automated tool. Here we present AutoCloner, (illustrated in figure 5.1) a fully automated allele-specific primer design pipeline that includes a simple web interface for users. Although developed in the context of wheat, AutoCloner can easily be configured to work with any species for which a genome assembly is available. It requires only a single input, the sequence of interest to clone.

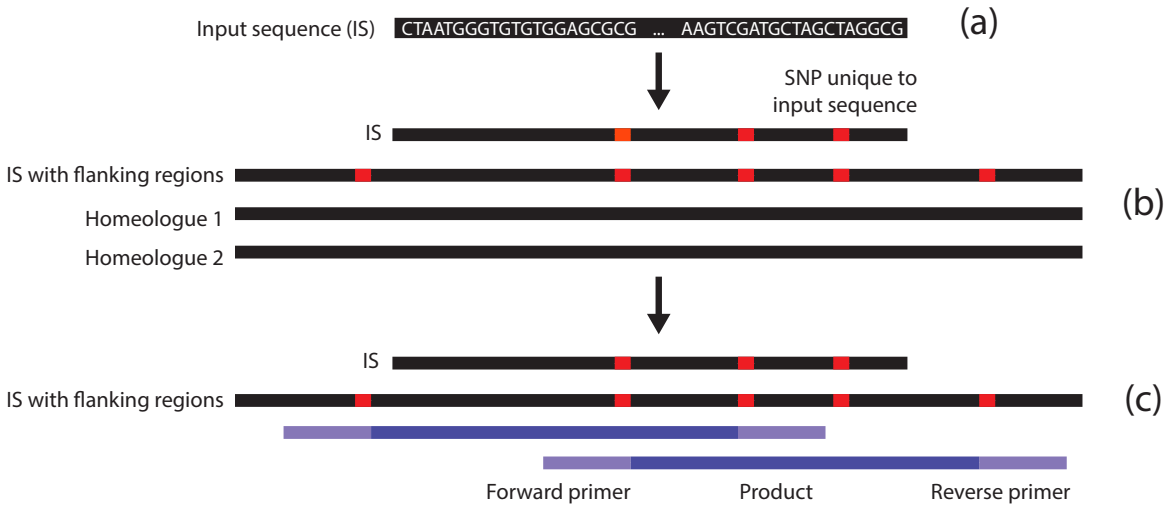


Figure 5.1 Schematic detailing AutoCloner, a homologue-specific primer design pipeline for polyploids. (a) The user inputs a DNA sequence which they wish to clone in a variety for which this sequence is unknown. (b) AutoCloner finds flanking regions and homologues through BLAST, generates a multiple sequence alignment with Muscle and scans the alignment for SNPs. (c) The best possible combination of primers, whose overlapping products span the entire input sequence, are returned by the pipeline via Primer3.

5.2 Materials and methods

5.2.1 Acquisition of homologous sequences via BLAST

AutoCloner first searches for homologues of the user input sequence by performing a BLASTN search of the input sequence against the latest IWGSC (International Wheat Genome Sequencing Consortium) RefSeq wheat genome assembly [@consortiumiwgsc-ShiftingLimitsWheat2018]. Alternatively, AutoCloner can use any genome that the user has specified in the configuration file, and could therefore be used for any species where homologous sequences with high similarity are common. The tabular output files of the BLAST search are parsed and used as a basis for sequence extraction from the genome assembly (for code see appendix A.1). BLAST breaks up query sequences into high-scoring pairs [@sheGenBlastGUsingBLAST2011], and as such it is necessary to examine groups of hits when using BLAST to extract homologues rather than individual

hits. Here a group of BLAST hits are defined as hits with the same query and subject sequence that are within 1000 bases of each other (figure 5.2). The group of BLAST hits that is most closely related to the input sequence is assumed to be the genomic representation of that sequence and is used to obtain the flanking regions of DNA. The next three, or alternatively the number specified by the user, best groups of hits are also used for sequence extraction and are assumed to be close homologues of the input sequence, providing their within-group average bitscore exceeds 200. This threshold means that only hits with a reasonable amount of sequence similarity are retained.

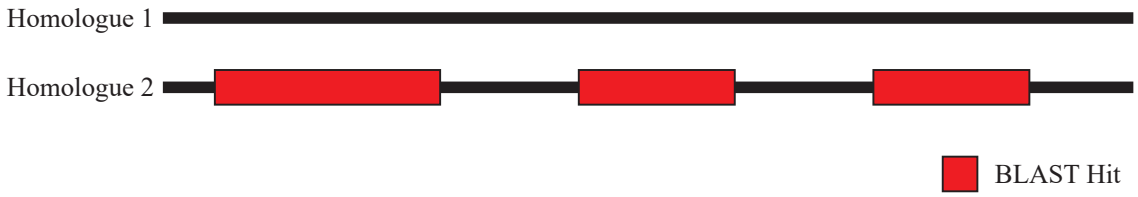


Figure 5.2 Schematic detailing BLAST parser. Although BLAST has returned three separate hits between these two homologues, these hits are actually within 1000 bases of each other, and so are grouped together, including the interstitial bases between hits, as one homologue by AutoCloner for sequence extraction.

Further input parameters to the pipeline include the amount of sequence to extract that flanks the input sequence, namely the start buffer (-s) and end buffer (-e) parameters, which default to 1000 nucleotides. The maximum product length and minimum product length parameters specify the maximum and minimum sizes of overlapping PCR products enclosed by the primers produced by the pipeline. To ensure memory is used efficiently during sequence extraction, AutoCloner makes fasta indices of any genome assemblies that are specified in the configuration file. AutoCloner also has the capability to include more than one genome if there are genome sequences available for more than one variety within the species. If this is the case, one sequence from each of the additional genomes is also extracted to increase the reliability of SNP identification in later stages, ensuring that varietal SNPs are not used as a basis for primer design.

5.2.2 Extraction of homologues and multiple sequence alignment

The extracted sequences must then be arranged into a multiple sequence alignment, which is used to identify SNPs between the sequence of interest and its homologues. AutoCloner uses Muscle [@edgarMUSCLEMultipleSequence2004], or alternatively Dialign [@alaitDIALIGNGOBICSMultiple2013], to achieve this. Dialign is useful when multiple homologues and partial homologues, i.e. sequences of different lengths, are present in the sequence set extracted from the BLAST results, as it allows specification of anchors that inform the alignment. A single-nucleotide mismatch at the 3‘ end of the primer significantly decreases the efficacy of Taq polymerase in the PCR reaction [@ayyadevaraDiscriminationPrimerNucleotide2000], and so the SNP locations can be evaluated as potential primer locations. It is unlikely that each of these locations will have the ideal sequence characteristics for a primer, such as adequate GC content, low probability of hairpin structures and an ideal melting temperature (TM). When performing this process manually, it is common for a researcher to evaluate multiple locations before finding one that is adequate for a primer. This is time consuming and can also produce sub-optimal results due to human error. It is for this reason that AutoCloner integrates Primer3 to evaluate each possible primer.

5.2.3 SNP Identification

Once the multiple sequence alignment has been produced, it can then be scanned for SNPs that could serve as potential locations for homologue-specific primers. Each SNP is evaluated based on various criteria, including whether the SNP provides complete chromosome specificity, and whether the SNP is in a suitable location. If more than one genome has been included in the sequence extraction process, varietal SNPs will also be identified. For example, if producing primers in wheat for the Apogee variety, the primary genome could be Chinese Spring [@consortiumiwgscShiftingLimitsWheat2018], whilst the secondary genome could be Paragon [@walkowiakMultipleWheatGenomes2020]. If the SNP is chromosome-specific and is present in both varieties, it is more likely to also be present in Apogee. A diagram of the SNP selection process

is shown in figure 5.3.

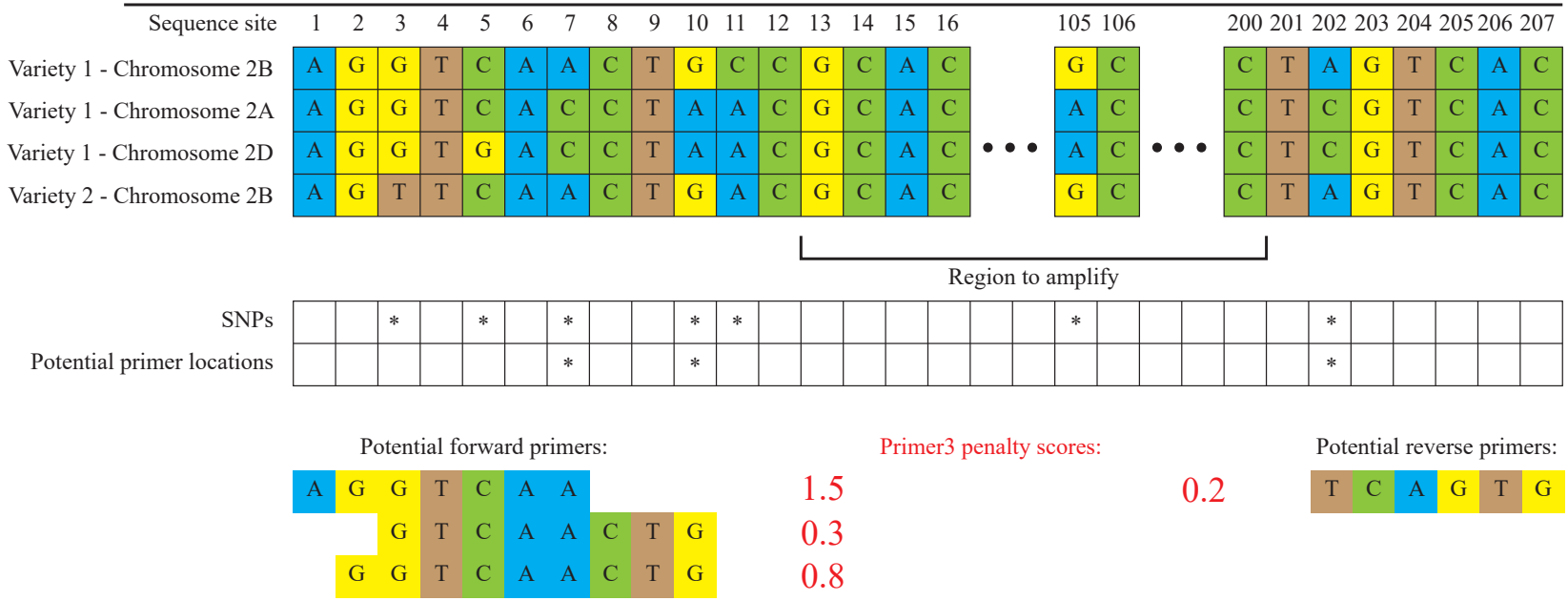


Figure 5.3 Detailed overview of the SNP categorisation and primer selection process in AutoCloner. Shown is a hypothetical multiple sequence alignment (MSA) containing the sequence to clone (first row), two homeologues (rows 2-3), and a homologue from a second variety (row 4). AutoCloner identifies SNPs in the MSA and identifies which might be suitable for potential primer locations. The SNP at sequence site 3 is a varietal SNP rather than a homeologous SNP and is therefore not suitable. The SNP at sequence site 5 only provides partial specificity and is also not suitable. Whilst the SNP at sequence site 11 provides specificity, however is not present in the second variety, meaning it could be unique to the first variety, and therefore not present in the variety under investigation. The SNP at sequence site 105 does not flank the desired region to clone and is therefore not suitable. Shown at the bottom of the figure are potential primers with 3' ends placed on SNPs deemed viable. Note that some of these include primers that are placed on the same SNP but are different lengths. Primer3 evaluates each of these primers, ultimately assigning each a penalty score. The primers with the lowest penalties are returned to the user. Note that the reverse primers are shown as the reverse complement of the sequence in the MSA.

5.2.4 Evaluation of potential primer combinations using Primer3

AutoCloner evaluates all possible combinations of primers at the SNP locations that fall within the user-specified minimum and maximum product size ranges by utilizing Primer3 [@untergasserPrimer3NewCapabilities2012]. The Primer3 output parameter PRIMER_PAIR_PENALTY is used to select the best primers. PRIMER_PAIR_PENALTY is a composite score calculated for each primer that corresponds to the overall quality of the primers: the lower the value the better. This is score is made up of several factors. Primer melting temperature, the temperature at which primers dissociate from the template DNA, is calculated using thermodynamic formulae that take into account factors such as the concentration of divalent cations in the solution [@koressaarEnhancementsModificationsPrimer2007]. This value is compared to the optimal melting temperature (set at 60°C), with large deviations resulting in higher penalties. Primer melting temperature is used to derive the primer annealing temperature, the temperature at which oligos bind to the template DNA. Sub- and super-optimal annealing temperatures have a negative effect on primer specificity to the target sequence [@rychlikOptimizationAnnealingTemperature]. Primer length is also taken into account, which optimally should be 20 bases. This is long enough to provide target specificity and short enough not to dramatically effect the efficiency of primer annealing [@dieffenbachGeneralConceptsPCR1993]. Primer GC content is closely linked to melting temperature, and should be around 60%. The probability of the formation of primer-dimers and hairpin structures (intermolecular and intramolecular secondary structures respectively) is another important factor that could negatively effect the reaction, and therefore the penalty score.

Using these principles, it is therefore possible to predict the likelihood of a successful PCR ahead of time. The PRIMER_PAIR_PENALTY metric allows AutoCloner to select optimal sets of forward and reverse primers. These sets are chosen such that they have overlapping products that cover the entire input sequence. These overlapping products allow the input sequence to be cloned and sequenced in its entirety. In addition to the primers intended for PCR, several within-product primers are also selected for

Sanger sequencing of large products. AutoCloner also allows the user to input their own multiple sequence alignment instead of a single sequence using the -a option; in this case the initial stages of the pipeline are skipped and the alignment is immediately scanned for SNPs. Note that if this is the case, AutoCloner expects the multiple alignment to 1) be in Fasta format with gaps indicated by "-" and 2) for the sequences to be in the following order: sequence to be cloned, same sequence but with flanking regions included, then any homologues.

5.2.5 Using Primer3 efficiently

One aspect that had to be considered when writing the AutoCloner pipeline was how to call Primer3 in a manner that would be most computationally efficient. This task is constrained by the options presented by the Primer3 developers that determine Primer3's mode of operation, namely the PRIMER_TASK configuration tag. The relevant values for PRIMER_TASK include "generic", which gives Primer3 free rein to pick and return only the best primer pairs it finds in the provided input sequence, unconstrained by position. This can be further modified to suit our needs (i.e. forcing the 3' ends of primers to be located at SNPs) by SEQUENCE_FORCE_LEFT_END and SEQUENCE_FORCE_RIGHT_END parameters. An additional possible value for PRIMER_TASK is the "pick_primer_list" option, which causes Primer3 to return a list of all possible primers in the input sequence.

Given these possibilities, a naive view would suggest that first the multiple sequence alignment should be scanned for SNPs, then a new instance of Primer3 should be called for each one of these SNPs using the "generic" value for PRIMER_TASK. This could be achieved using the SEQUENCE_FORCE_LEFT_END and SEQUENCE_FORCE_RIGHT_END parameters for forward and reverse primers respectively. This could save on the unnecessary computation of primers at invalid, non-SNP locations. This approach works when the number of SNPs is low, however was found to be poor in practice, as the inefficiency of calling Primer3 in separate instances significantly lengthened the computation when the number of SNPs, and

therefore potential primer locations, was high. A better approach is simply to force Primer3 to generate primers at every single base location of the input sequence using the "pick_primer_list" value for PRIMER_TASK, then perform filtering of these primers to include only those that are in valid locations. This is much faster in the case where the number of SNPs is high, and only slightly slower than running separate Primer3 instances when the number of SNPs is low. This is largely due to the fact that Primer3 is written in C / C++, compiled languages optimized for faster computation. Figure 5.4 shows the AutoCloner control flow.

5.2.6 Choosing sets of overlapping primers

Sanger sequencing and PCR both become less accurate after a certain number of bases. To clone large genes then, it is necessary to perform sequencing and PCR with smaller, overlapping subsections of the gene which can be pieced together later using sequence alignment software. To achieve this, AutoCloner uses the following while loop algorithm (for code see appendix A.2):

Variables:

- Minimum SNP coordinate (S_m)
- Maximum SNP coordinate (S_M)
- Maximum product size (P)
- Forward primer coordinate (S_F)
- Reverse primer coordinate (S_R)

Algorithm:

- Set $S_m = 0$
- Set S_M = Position of the start of the gene
- Set S_R = Position of the start of the gene
- While $S_R <$ Position of the end of the gene:
  - Find SNP within interval (S_m, S_M) with minimum primer penalty score. This is the position of the first forward primer (S_F).
  - Find SNP within interval (S_F, $S_F + P$) with minimum primer penalty score. This is the position of the first reverse primer (S_R).
    * While no SNPs within interval (S_F, $S_F + P$) and $P <$ sequence length, set $P = P + 10$
  - Set $S_m = S_F$
  - Set $S_M = S_R$

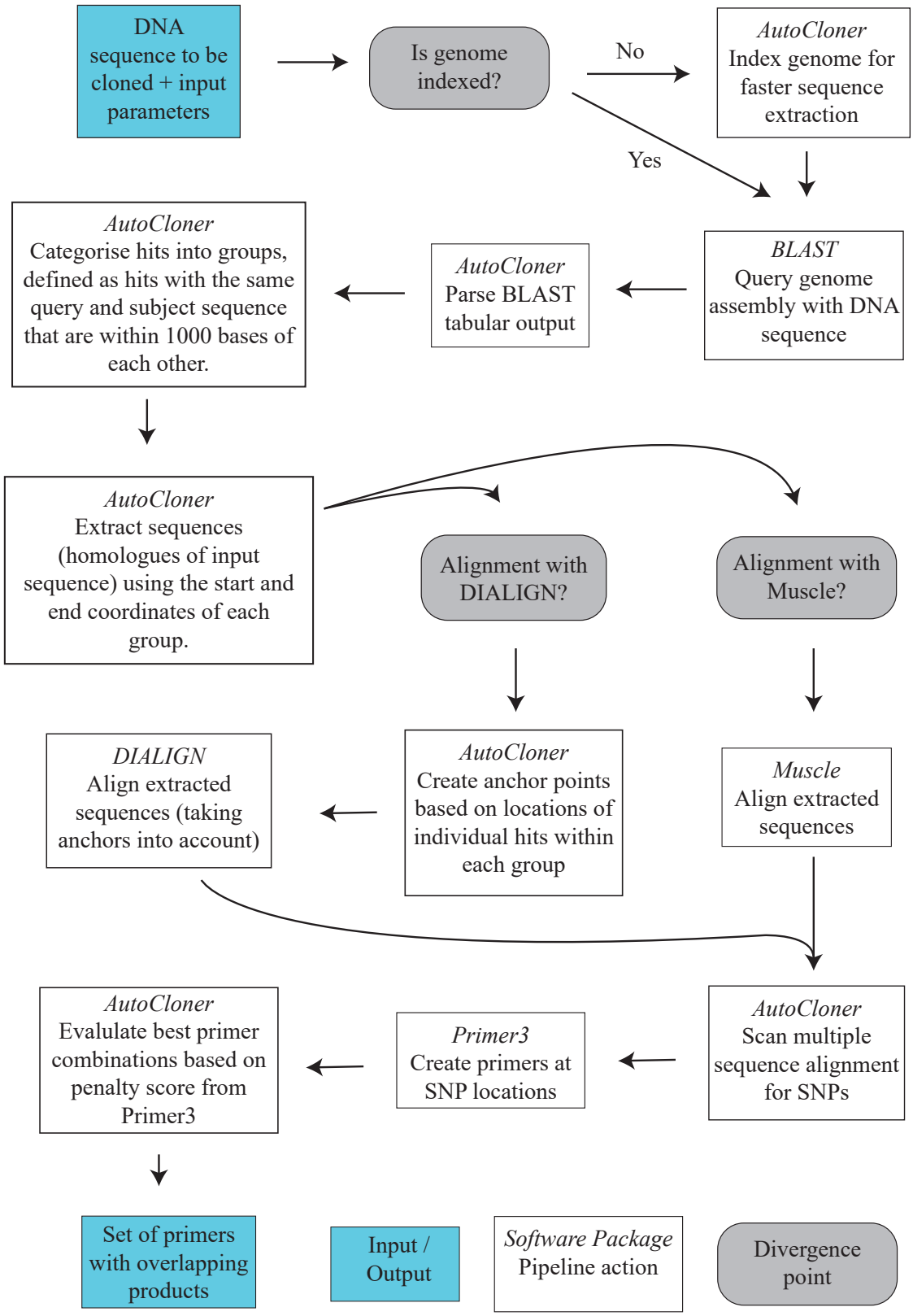


Figure 5.4 Flowchart showing the control flow of the AutoCloner pipeline. Names of software packages are shown in italics; box types are indicated in the legend.

5.2.7 Web interface

In order to maximise ease-of-use for wheat researchers, we designed a web interface for AutoCloner using the Django web development framework for Python. This site integrates into the popular wheat-resource website CerealsDB [@wilkinsonCerealsDB-ExpansionResources2016]. The web interface requires either a single sequence to be input by the user, or a multiple sequence alignment. The user can also specify all other pipeline parameters via the web interface, such as start and end buffers, minimum and maximum product sizes, and which alignment software to use. Once submitted, the sequence is entered into a queue for processing by AutoCloner. When the pipeline has finished processing the sequence, the user is forwarded to the Details page (figure 5.5) in which the details of their primers are displayed, with options to download the primer information as a CSV file as well as download the multiple sequence alignment in either fasta or clustal format. As well as showing optimal primer pairs chosen by the pipeline, a list of all possible primers (at all SNP locations) is also shown in case any of the primers do not work. The multiple sequence alignment is displayed directly on the website using MSA [@yachdavMSAViewerInteractiveJavaScript2016], a Javascript web component (figure 5.6). To maximise ease of use, the website was designed not to require any login details or personal information from the user. The website features two modes of running the pipeline. The first is a fully automatic mode in which all stages of the pipeline are run without user interaction. The second is a guided mode in which the user can inspect all found homologues and the multiple sequence alignment and choose to remove any unwanted homologues from the alignment before designing primers. This is useful as it gives the user more control over the primer design process, allowing them to debug errors in the pipeline when homologue classification is too lenient.

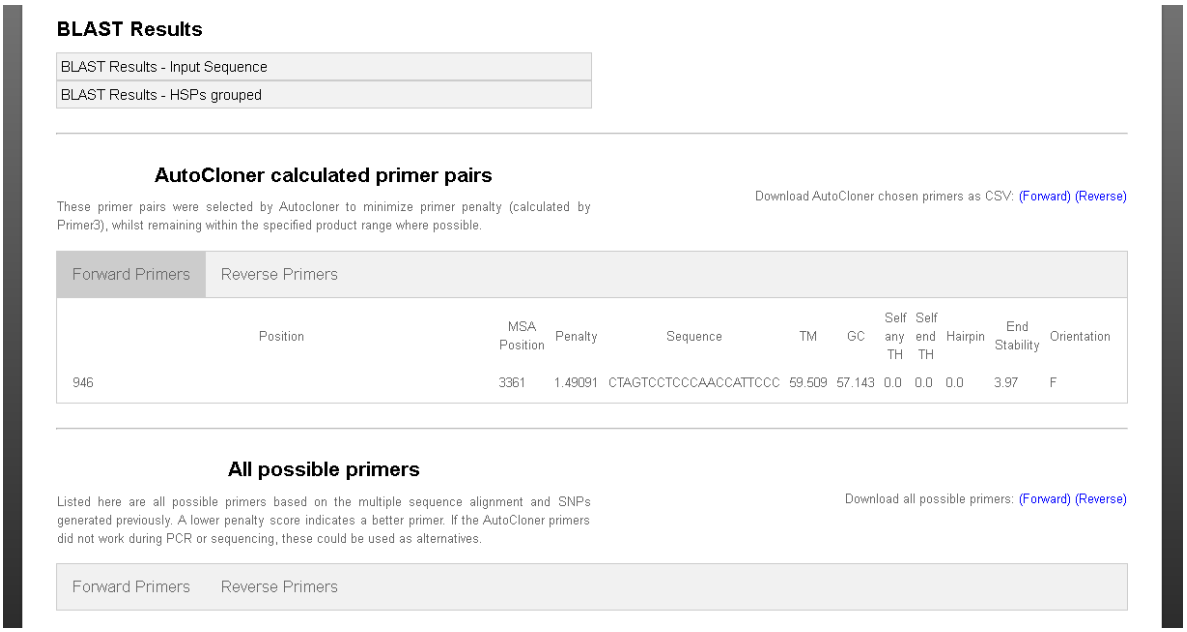


Figure 5.5 Picture of the AutoCloner website job details page. The user has the option to view the BLAST results themselves, as well as choose from all possible primers should the primers selected by AutoCloner not work.

Figure 5.6 Picture of the AutoCloner website job details page. The AutoCloner website presents the multiple sequence alignment, containing the input sequence, homologues, SNPs and primers, to the user for their own inspection.

5.3 Results and Discussion

5.3.1 Using AutoCloner to investigate segregation distortion

It is well established that a DNA mismatch at the 3‘ end of the primer significantly reduces the efficiency of Taq polymerase in a PCR, with previous research suggesting the magnitude of this effect could as much as a 100-fold decrease [@ayyadevaraDiscriminationPrimerNucleotide2000]. This principle serves as the basis for the design of homologue-specific primers. We used AutoCloner to investigate potential gene candidates underpinning a region of segregation distortion on chromosome 5A of a Chinese Spring X Paragon F5 mapping population [@allenCharacterizationWheatBreeders2016]. This region lacked distortion in an Apogee X Paragon F5 mapping population [@allenCharacterizationWheatBreeders2016]. Sequences for this region in both Chinese Spring and Paragon were available, whilst no sequencing data was available for Apogee. We therefore formulated the hypothesis that since there was a lack of distortion in the Apogee X Paragon population, any potential causative gene would have the similar or identical sequences between Apogee and Paragon, and different sequences between Chinese Spring and Paragon. Therefore, sequencing data for Apogee could highlight (or eliminate) genes worthy of further investigation.

5.3.2 Cloning TraesCS5A01G531300 in Apogee

The first gene we cloned was TraesCS5A01G531300, a 2.4 kb High Confidence gene from the IWGSC assembly [@consortiumiwgscShiftingLimitsWheat2018], with a BLAST search identifying 2 homeologues on chromosomes 4B and 4D (the result of a well-known translocation between 5A and 4A [@devosStructuralEvolutionWheat1995]), as well as a partial homologue on 5B, with homology beginning 325 bases into the gene and extending throughout the gene until ~10 bases downstream. Sequence identity, not including regions of outside of the HSPs identified by BLAST, between TraesCS5A01G531300 and each of the three homologues was 93.36%, 93.91% and 80.58% respectively, whereas GC content for TraesCS5A01G531300 and each of

the homologues was 49.67%, 49.51%, 49.67% and 49.4% respectively. These GC values are close the average GC percentage of all 110790 HC genes in the IWGSC assembly [@consortiumiwgscShiftingLimitsWheat2018], which is 51.43% ± 10 (mean ± s.d). AutoCloner returned four overlapping pairs of primers whose products covered the entire gene length (table 5.1). DNA was extracted following the protocol in [@edwardsSimpleRapidMethod1991]. All of the primers produced amplicons from the desired locus in the genome (figure 5.7), and the entire gene sequence was obtained via Sanger sequencing of these products. The resulting sequence was identical to the sequence in Chinese Spring apart from one non-synonymous single nucleotide variant.

Table 5.1 List of primers designed by AutoCloner to amplify and sequence TraesCS5A01G531300 in Apogee. Oligo names succeeded by an F are forward primers, whilst an R indicates a reverse primer. Both primers for PCR and primers for Sanger sequencing are included.

Oligo Name	Type	Sequence (5'->3')
T.300.577-1989.F	PCR	AGACTTCCTGAACACGGCACAA
T.300.577-1989.R	PCR	CTTCTTGATGGCGCGGCATATAT
T.300.1904-3138.F	PCR	TAGGTTGACGTCATCGAAGCAG
T.300.1904-3138.R	PCR	CGGGTGAGAAGCAAGGACTC
T.300.2632-3437.F	PCR	CCATGGTGAGGTTGAGGTCC
T.300.2632-3437.R	PCR	GGTGCAGCAAGAGTACGGAG
T.300.2888-4219.F	PCR	CGGACGATGACAACAGGGAG
T.300.2888-4219.R	PCR	CCCTGCTCTCTCTCTCTCTCT
T.300.577-1989.split2.F	Sanger	CTCGAACTCGCTATTGGGCT
T.300.577-1989.split3.F	Sanger	AGTCAAGGTACAATATGTGACTGA
T.300.1904-3138.split2.F	Sanger	CCGTGAAGTACCGAAACCCA
T.300.1904-3138.split3.F	Sanger	TAACGAACCTGGTGCCTTCG
T.300.2632-3437.split2.F	Sanger	TCAGGTCCTTGGCCAGTTTC
T.300.2888-4219.split2.F	Sanger	GTGGAGACATGGAGGAGCAC
T.300.2888-4219.split3.F	Sanger	ACCAACACTCAAGCAAAGGGA

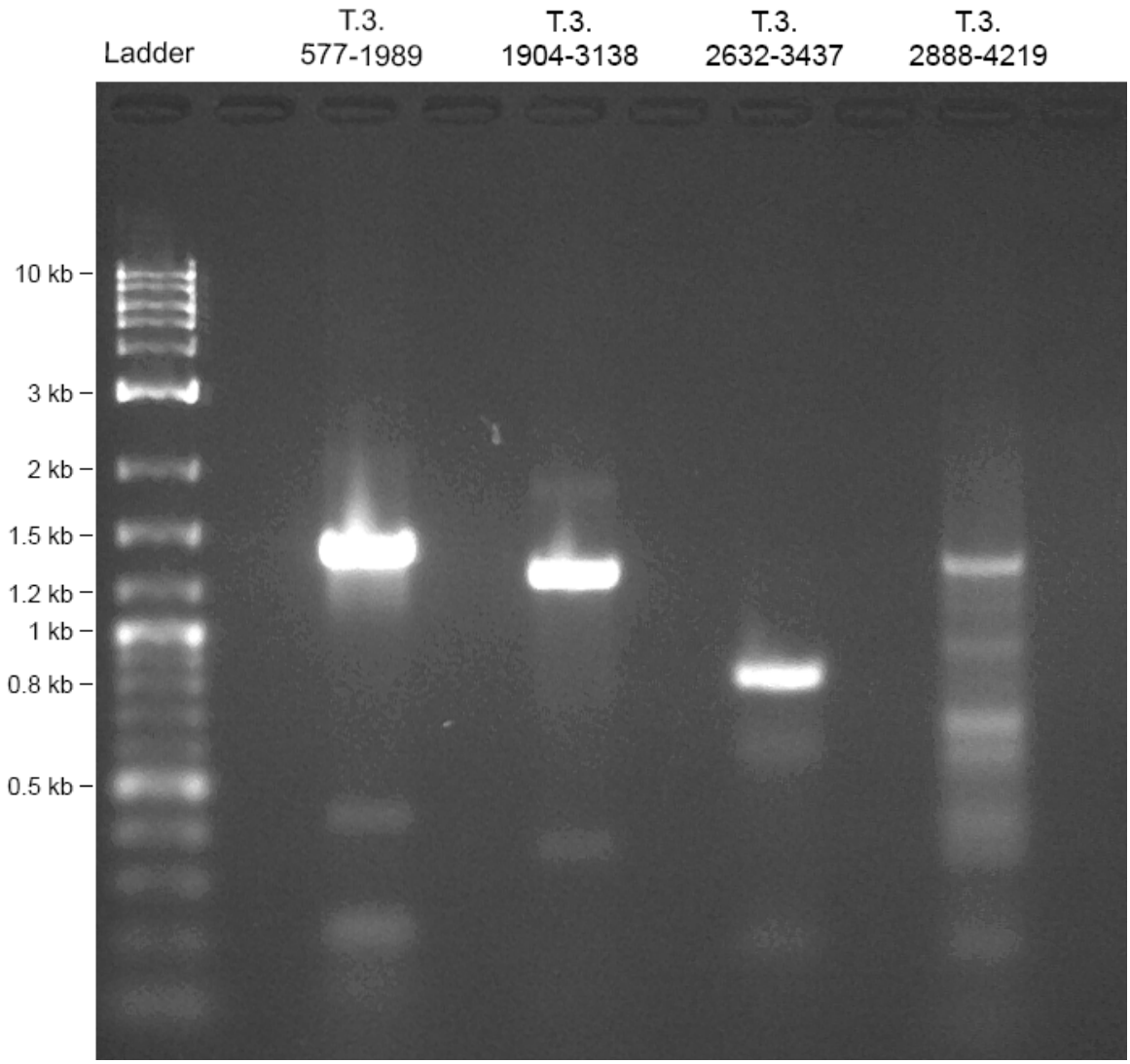


Figure 5.7 Agarose gels showing amplified PCR products for the TraesCS5A01G531300 gene using primers designed by AutoCloner. Details of the primers are shown in table 1. The DNA ladder used was the Quick-Load® Purple 2-log DNA ladder, manufactured by New England Biolabs, containing DNA fragments ranging from 0.1 kb to 10 kb in size. The expected product sizes for T.300.577-1989, T.300.1904-3138, T.300.2632-3437 and T.300.2888-4219 were 1412, 1234, 805 and 1331 bases respectively. A subsequent PCR (not shown) in which the annealing temperature was increased from 58 °C to 60 °C increased the specificity of the T.300.2888-4219 set of primers.

5.3.3 Cloning TraesCS5A01G531700.1 in Apogee

The second gene cloned in Apogee was TraesCS5A01G531700.1, a 3.6kb gene on chromosome 5A with homeologues on 4B and 4D, as well as a partial 2kb paralogue on

chromosome 5A, and smaller 600 bp homologous regions on 2D and 5D. We obtained a complete sequence of the coding region (introns defined by pairwise alignment to the IWGSC gene model) using four sets of overlapping primers produced by AutoCloner. All the amino acid substitutions present between the Chinese Spring and Paragon coding sequences of TraesCS5A01G531700.1 are also present between the Apogee and Paragon sequences. The Apogee sequence also contains some additional substitutions with Paragon between positions 35-69 (table 2).

Table 5.2 Amino acid substitutions between Paragon and Chinese Spring coding sequences of TraesCS5A01G531700.1. AA = Amino acid; "first variety" refers to the first variety listed in the corresponding "Comparison" column for each row.

Comparison	Position	AA in first variety	AA in second variety
Paragon & Chinese Spring	270	M	T
Paragon & Chinese Spring	292	Q	P
Paragon & Chinese Spring	552	I	V
Paragon & Chinese Spring	731	P	H
Paragon & Chinese Spring	760	A	V
Paragon & Apogee	35	L	F
Paragon & Apogee	64	F	V
Paragon & Apogee	66	E	K
Paragon & Apogee	69	E	A
Paragon & Apogee	270	M	T
Paragon & Apogee	292	Q	P
Paragon & Apogee	552	I	V
Paragon & Apogee	731	P	H
Paragon & Apogee	760	A	V

Comparison	Position	AA in first variety	AA in second variety
Chinese Spring & Apogee	35	L	F
Chinese Spring & Apogee	64	F	V
Chinese Spring & Apogee	66	E	K
Chinese Spring & Apogee	69	E	A

5.3.4 Cloning TraesCS5A01G530800 in Apogee

Also cloned was TraesCS5A01G530800, a 551 bp gene on chromosome 5A. The gene was cloned using a single set of flanking primers produced by AutoCloner, and the gene sequence was found to be identical between Apogee, Paragon and Chinese Spring. This sequence had homologues that encompassed the entire gene on chromosomes 5D, 4B and 4D, as well as 17 small sequences of around ~250 bases with high similarity to the flanking region downstream of the input sequence.

5.3.5 Running AutoCloner on 85,040 high-confidence genes

In addition to the sequences evaluated in the context of segregation distortion, AutoCloner was also run using all high-confidence gene sequences from the IWGSC assembly under 10,000 bases, amounting to a total of 85,040 genes. These alignments and primer sets are available to view on the AutoCloner website. For 30,186 of the genes, the top two homologous sequences identified by the pipeline were homeologues from the corresponding subgenomes and chromosomes (e.g. for an input sequence on 3A, sequences from 3B and 3D were most closely related). Also of interest was the composition of these alignments in terms of number of regions, not strictly limited to homeologues,

that contained high sequence identity to the input sequence. The mean ± s.d. number of these highly similar sequences, detected via BLAST HSPs, was 10.4 ± 7.47 per alignment. When limited to sequences that covered over 70 % of the input sequence, flanking regions of 1 kb upstream and downstream included, this number reduced to 2.4 ± 2.24 per alignment, indicating that the majority of genes only have a few close homologues that extend over large regions. The smaller regions with high sequence identity should not be problematic for allele-specific cloning if they do not fully encompass the PCR product. Even so, the AutoCloner web interface includes a "Guided Mode", allowing the user to manually inspect alignments and remove (or retain) sequences before SNP calling and primer design should these regions be of interest.

5.4 Conclusions

Whilst the scientific community has made incredible progress in producing genomic sequences for many different crop species, we are a long way from having a complete pangenome encompassing every single variety within each species. Until this time, cloning of genes will remain an important technique for assessing genetic variation, and AutoCloner makes this process significantly faster and easier than current methods.

6 General Discussion and Conclusions

Presented in this thesis are analyses of wheat evolution (Chapter 2), recombination (Chapter 3) and segregation distortion (Chapter 4). These seemingly disparate topics are linked by their methodological underpinnings, namely the combined use of mapping populations and high-density genotyping arrays to investigate genetic features at a genomic scale. Interest in array development for wheat has seen a significant increase in recent years, resulting firstly in the development of the lower density 9k iSelect [@cavanaghGenomewideComparativeDiversity2013], before progressing on to more advanced, denser arrays with 90k [@wangCharacterizationPolyploidWheat2014], 820k [@winfieldHighdensitySNPGenotyping2016] 35k [@allenCharacterizationWheatBreeders2016] and 280k [@rimbertHighThroughputSNP2018] SNPs respectively. These new genotyping technologies have in turn catalysed the production of a wealth of new genotyping data for wheat, making much of this thesis possible.

The findings presented here should help direct future research themes, prompting investigation into other avenues of influencing wheat recombination (in addition to temperature), as well as correct misconceptions among some researchers regarding the detection of segregation distortion, emphasizing the importance of correction for multiple testing. Additionally, the comparison of exome-capture and array-based genotyping data in Chapter 2 should add to general confidence around array usage, with high concordance between datasets from disparate technologies. Finally, the methods and tools generated during this thesis, including AutoCloner (Chapter 5), should be useful to other students of crop genetics, making common research processes such a full-gene cloning more efficient.

6.1 A lack of known divergence points between the inception of wheat and the proliferation of modern landraces makes dating via molecular clock difficult

One of the central foci of contemporary wheat research is the species' limited genetic diversity compared to wild relatives. This is largely the result of the various genetic bottlenecks that have occurred during the history of wheat, including several initial polyploidization events, e.g. the hybridization of tetraploid and diploid ancestors, as well as strong selection pressures during domestication, imbuing wheat with desirable agronomic traits such as a non-brittle rachis and free-threshing characteristics [@dubcovskyGenomePlasticityKey2007]. This limited diversity has the potential to stifle future yield increases as well as increase susceptibility to pathogens. Related to this limited genetic diversity is the question of novel allele accumulation: at what rate do novel polymorphisms accumulate in wheat over time? Answering this question could allow us to predict future levels of genetic diversity, potentially informing breeding practices. Whilst there are many studies aiming to assess the overall levels of genetic diversity in wheat [@allenTranscriptspecificSinglenucleotidePolymorphism2011; @laikaitaoIdentificationCharacterizationMore2014; @heExomeSequencingHighlights2019; @pontTracingAncestryModern2019], few have attempted to assess changes in genetic diversity over time.

In chapter 2, scientific interest in the Watkins lines has been leveraged to try and estimate the rate of novel allele accumulation in wheat. The Watkins lines are an ideal collection to examine this question as they are comprised of landraces, locally grown cultivars that have not been subjected to modern breeding practices designed to artificially induce novel variation, such as X-ray mutagenesis. In absence of a series of time-separated wheat samples stretching back to wheat's origin, I've attempted to answer this question using one of the only well-estimated divergences in wheat's history, the hybridization of tetraploid and diploid ancestors to make up hexaploid wheat. It has been shown through both hybridization experiments and genome analysis that this

diploid progenitor was most closely related to modern day goat grass, *Aegilops tauschii*. In theory then, if we compare the D genome of the Watkins wheat varieties to *A. tauschii* and use the date of their divergence, estimated to be ~ 9000 years ago based on archaeological data, to calibrate a molecular clock, we can estimate the rate of novel polymorphism.

Whilst the logical nature of this method is enticing, in reality, the assumption of an equal evolutionary rate between the D genome of wheat and *A. tauschii* is too simplistic, resulting in an estimate of TMRCA for the Watkins lines that seems too young based on what we know of historic wheat trade. Further insight into this question will need to arise through different methodologies, perhaps involving longitudinal study design with time-series genotyping data.

6.2 Population genetics inferences in wheat are consistent between different sources of genotyping data

Many of the conclusions generated in this thesis stem from SNP genotyping data generated from arrays, either in their 35k [@allenCharacterizationWheatBreeders2016] or 820k [@winfieldHighDensityGenotyping2017] formats. This data is produced initially through complex procedures in the lab, including DNA extraction, amplification and hybridization, followed by extensive bioinformatics pipelines that calculate quality control metrics both at the level of the sample and at the level of the probe, subsequent clustering of probe groups based on their fluorescent signals and assignment of genotypes informed by probe-specific priors. Also performed is extensive filtering of probes depending on the analysis, such as the exclusion of those with monomorphic genotypes across all samples in a recombination analysis, and more generally the exclusion of probes with signs of off-target hybridization, an important consideration in wheat due to its hexaploid nature, and therefore presence of homeologues within the genome. This long, complicated pipeline increases the opportunity for error compared to simpler procedures such as a PCR, and indeed previous authors have assessed the reliability of microarrays

[@jaksikMicroarrayExperimentsFactors2015], finding problems with the assignment of probes to genes [@daiEvolvingGeneTranscript2005], errors in evaluation of background signal [@krollModelingBackgroundIntensity2008], and effects of distinct probe features on data processing algorithms [@wuModelBasedBackgroundAdjustment2004].

With these considerations in mind, Chapter 2 includes an investigation into the reliability of array data by performing common population genetics analyses, PCO and STRUCTURE, with datasets generated using different genotyping methods, both array and exome-capture. These genotyping methods have been performed independently by different groups on the same set of lines, the Watkins collection. We would therefore expect that if the genotyping method did have an influence on the results that the PCO and STRUCTURE plots would show clear differences, whether that be in the placement of individual samples, or in the overall clustering of samples. This however did not turn out to be the case, with PCO and STRUCTURE plots appearing highly similar to each other between datasets. Both datasets produced similar broad patterns of clusters by region, with Asian and Middle Eastern varieties separated from European, Australian and USSR-originating varieties along the x-axis (figure 2.2), as well as similar placement of individual lines such as the positioning of Watkins lines 299 and 300 between Western European and Middle Eastern clusters (figure 2.2). Overall, the results of this analysis are highly reassuring, showing the arrays produce reliable, reproducible results.

6.3 The influence of temperature on recombination distribution in wheat is limited

Wheat, along with many other staple food crops such as barley and maize, has recombination events that are distributed in an inverted bell curve along the chromosomes, the nadir typically occurring in the region surrounding the centromere [@zelkowskiDiversityDeterminantsMeiotic2019]. In contrast, recent advances in wheat genome knowledge, including the development of multiple chromosome-level genome assemblies [@consortiu-

miwgscShiftingLimitsWheat2018; @walkowiakMultipleWheatGenomes2020], highlight the comparatively even distribution of genes along the chromosomes. These contrasting distributions present a problem for breeders: how is it possible to manipulate genes in the central regions of the chromosomes without bringing along large amounts of unwanted genetic material in the process, i.e. to avoid linkage drag? To achieve this, the distribution of recombination events would need to be modified.

There are multiple avenues with which this could be achieved. Environmental effects, such as soil magnesium content [@reyMagnesiumIncreasesHomoeologous2018], and temperature [@jainEffectHighTemperature1957; @loidlEffectsElevatedTemperature1989], have long been known to influence recombination. Recent research on barley using immunofluorescent staining techniques suggests that temperature could be effective in shifting the distribution of recombination inwards in cereals [@higginsAnalyzingMeiosisBarley2013; @higginsSpatiotemporalAsymmetryMeiotic2012], however research in wheat specifically remains limited. The primary aim of chapter 3 was to examine this effect in wheat.

As with much of the thesis, the analysis of recombination was achieved through the utilization of high-density array genotyping data. After treating F2 Apogee X Paragon lines with four respective temperatures during meiosis, their progeny were genotyped and polymorphic markers were clustered and ordered to create a genetic map. In this way, recombination events can be observed via the transition between parental genotypes in neighbouring markers on the map. The results revealed that recombination distribution on the majority of chromosomes was not affected by changes in environmental temperature, with only long arms of 1A and 3B, as well as the short arms of 2A and 7A showing significant changes in distribution (table 3.3). Closer examination of chromosome 1A showed that higher temperature treatments had a slight inward shift in the distribution of recombination events, although much of the region immediately surrounding the centromere remained highly linked regardless of temperature (figure 3.18). In addition to changes in distribution, recombination frequency was also shown to be significantly affected by environmental temperature, with a U-shaped curve observed

from temperatures 10°C to 26°C before a dip at 28°C (figure 3.29). Future work in the area will involve a similar investigation using immunofluorescent techniques, as has been done in barley [@higginsAnalyzingMeiosisBarley2013], to compliment the methodology used here.

Whilst environmental factors have been shown to play an important role in the distribution and frequency of recombination events in many species, also important are internal, genetic factors, such as the FANCM gene, which limits meiotic crossovers in *Arabidopsis* [@crismaniFANCMLimitsMeiotic2012]. In light of this, in chapter 3 I conducted a search for novel genetic factors influencing recombination frequency and distribution using a QTL analysis of previously published genetic maps [@allenCharacterizationWheatBreeders2016]. Whilst this search highlighted some potential candidates initially, after statistical correction for the number of phenotypes tested, none of the QTL appeared to be significant. Additionally, I attempted to assess the maliability of genes involved in meiosis. Are meiotic genes subject to harsher stabilizing selection than the rest of the genome due to their important function within wheat, or is there room for novel polymorphisms in these genes, and consequently potential for the manipulation of recombination through genetic modification? The results of this analysis showed that appear to be slightly more conserved in wheat compared to a random sample of genes, when we examine homologues of these genes in barley, they have accumulated many new polymorphisms since the divergence of barley and wheat. This indicates that there is room for genetic modification of meiotic genes without disastrous consequences for the organism, and that perhaps this could be a future route to the manipulation of recombination.

6.4 Misconceptions around the analysis of segregation distortion are common in scientific literature

Mendel's law of segregation states that gametes have an equal chance of inheriting either of the two parental alleles for a particular gene. Exceptions to this law occur

when selection operates in favour of one parental allele during the developmental cycle, causing deviation from the Mendelian ratio of offspring, known as segregation distortion. This phenomenon is common in wheat and many other species, and is often highlighted in studies of high-density genotyping data [@allenCharacterizationWheatBreeders2016; @gardnerHighlyRecombinedHigh2016]. Initially, my study of segregation distortion was aimed at finding potential causative genes underlying these regions in wheat. However, during my evaluation of the literature I noticed inconsistencies between authors in their methods of detecting segregation distortion.

Segregation distortion data consists of counts of categories, specifically genotypes, amongst samples in a breeding population. The statistical test used to assess this is therefore a chi-squared test. In the age of single-gene genetics assessed through phenotype, this test alone would be entirely sufficient to examine segregation distortion. However, the development of high-density arrays means that thousands to hundreds of thousands of molecular markers are now genotyped simultaneously. A chi-squared test must be performed for each marker, and therefore we are now faced with the problem of correcting for multiple testing. This is further complicated by the fact that markers are subject to different degrees of genetic linkage depending on their position relative to each other on the chromosome. Markers that are closer together exhibit strong genetic linkage, i.e. a small chance of an inter-marker recombination event, and therefore often show the same pattern of segregation, and as such are not independent in a statistical sense. My analysis into the literature showed that there was no consensus among authors on the most appropriate multiple-testing procedure to account for this - Chapter 4 aims to elucidate.

The primary means used to investigate this was simulated genotyping data. This allowed a selection pressure of known strength to be applied to a particular marker, followed by an examination of how effective detection of segregation distortion was with various statistical procedures. The results indicate that the false-discovery rate (FDR) procedure is most effective, offering a balance between no correction for multiple testing, and the more extreme control of familywise error rate offered by the Bonferroni

procedure. This is ideal for data in which events are linked to varying degrees rather than fully independent, such as markers on a chromosome. When I reanalysed datasets from published studies that had not used multiple-testing correction procedures in their analyses, I found that much of the reported segregation distortion was caused by sampling bias rather than a genuine selection pressures in the developmental cycle (table 4.2). These results were reinforced by an examination of replicate empirical populations, which had few regions of segregation distortion in common.

Another interesting outcome of my investigation into segregation distortion were the results on the effects of segregation distortion on genetic mapping. It is common for authors to remove markers exhibiting segregation distortion before the genetic mapping process [@allenCharacterizationWheatBreeders2016; @liuChromosomespecificSequencingReveals2016; @roorkiwalDevelopmentEvaluationHighdensity2018], as it is thought that they may interfere with the clustering or ordering processes of mapping. I found that contrary to this belief, segregation distortion only effects genetic mapping at extreme selection pressures, meaning that many markers that would have previously been removed before the mapping process can now be retained.

6.5 Wheat research can be made more efficient by the development of novel bioinformatics pipelines

PhD theses typically aim to investigate novel questions, making original contributions to the field of study. In the process of doing this, it is often the case that existing methods, whether laboratory procedures or bioinformatics pipelines, are not fit for purpose. In my search for a causative gene underlying regions of segregation distortion in wheat, I devised an experiment that would eliminate candidates based on sequence differences between varieties. This required the sequencing of a large number of genes that were not available in public sequence repositories.

The process of gene cloning in wheat is complicated by wheat's hexaploid genome, meaning that PCR primers must be orientated such that their 3' tail lands on a SNP unique

to the homologue of interest, preventing similar sequences from other subgenomes from also being amplified. This process involves various stages, including extraction of the sequence of interest and all homologues from an existing wheat genome assembly, multiple sequence alignment of these homologues, scanning of the alignment for SNPs, evaluation of all possible primer locations to optimize PCR-specific variables, and finally choosing optimal combinations of forward and reverse primers for gene cloning. Previous practice within the wheat community was to carry out each of these stages manually, a very time-consuming process when applied to a large number of genes. To make this more efficient, I developed AutoCloner, a fully-automated gene-cloning bioinformatics pipeline with a convenient web interface for use by the research community. I utilized AutoCloner to successfully sequence various genes from the Apogee wheat variety, shedding light on the potential causes of segregation distortion in a region of chromosome 5A.

6.6 Final remarks

Whilst I hope that others find the research presented in this thesis useful, one of the primary outcomes of this PhD is my own personal growth as a scientist and an independent thinker. Research rarely conforms to the course set out in initial plans, and it is during these deviations that we uncover the questions of real interest, and begin to design solutions to elucidate them.

Most salient of the findings presented here are the results of the recombination analysis, which highlight the future importance of gene-editing technologies such as CRISPR-Cas9 in addressing linkage drag in wheat. If these technologies can be recognized as safe both by science and by legislation, they could play a crucial role in future yield improvement. In addition, the work on segregation distortion should educate researchers on the most appropriate methods for detection of distortion, ensuring only regions of true distortion are investigated. I also hope that other researchers wishing to clone genes in wheat find AutoCloner as useful as I did when seeking to clone a large number of genes in unsequenced varieties.

7 Appendix A - AutoCloner R Code

7.1 A.1 AutoCloner BLAST Scaffold parser

The primary purposes of the following script are to perform two actions. Firstly, the BLAST tabular output is parsed to identify which groups of hits correspond to homologues of the input sequence, primarily achieved by the parse.scaffold.blast function. These homologues are then extracted from the genome of interest by the extract.sequence function. Additionally, if Dialign is to be used for the multiple sequence alignment, an anchor file is setup based on BLAST coordinates to inform Dialign of the relative positions of the sequences in the alignment.

```
# BLAST PARSER FOR SCAFFOLDS
write("blast.scaffold.parser.rscript.R", p("jobs/", opt$sequence.name, "/pipeline.checkpoint.txt"))
gene.name <- opt$sequence.name
fa.path1 <- opt$fasta.path
input_sequence <- readDNAStringSet(p("jobs/", gene.name, "/seq/extended/seqs/input_seq.fa"))

#    ----------------------------------------------------------------------------
#   DEFINE FUNCTIONS                                                        ####

sort.blastdf <- function(blastdf) {
  # first clusters BLAST hits by chromosome, then sorts by the start location of each hit within clusters
  # blastdf - a dataframe containing BLAST output in tabular format
  sorted <- newdf(colnames(blastdf), no.rows = T)
  blastdf$orientation <- "F"
  rev.coords <- which(blastdf$sstart > blastdf$send)
```

```
16    blastdf$orientation[rev.coords] <- "R"
17
18    rev.starts <- blastdf$sstart[rev.coords]
19    rev.ends <- blastdf$send[rev.coords]
20
21    blastdf$sstart[rev.coords] <- rev.ends
22    blastdf$send[rev.coords] <- rev.starts
23
24    # do some sorting
25    for (i in unique(blastdf[, 2])) {
26      temp <- blastdf[blastdf[, 2] == i, ]
27      temp <- temp[sort(as.numeric(temp[, 9]), index.return = T)$ix, ]
28      sorted <- rbind(sorted, temp)
29    }
30    return(sorted)
31  }
32
33  parse.scaffold.blast <- function(blastdf1, dist.threshold) {
34    # parses a BLAST dataframe of a short query sequence against a genome ↩
        assembly
35    # composed of scaffolds or chromosomes. If the assemblie is ↩
        chromosomal, the parser will split
36    # the chromosome up into groups of hits where hits are more than ↩
        dist.threshold bp apart.
37    # returns a dataframe containing the best groups of hits (average ↩
        bitscore higher than 200, individual hits no more than ↩
        dist.threshold bp apart)
38    # args:
39    # blastdf1 - a BLAST dataframe imported using read.blast()
40    # dist.threshold - Integer; the maximum number of bases between two ↩
        hits for them to be considered part of the same group
41
42    blastdf_orig <- blastdf1
43    blastdf1 <- sort.blastdf(blastdf1)
44
45    unique.groups <- convert.to.character.data.frame(unique(blastdf1[, ↩
```

```
    1:2]))

potential.homeologues <- newdf(c("query", "scaffold", "start", "end", "length", "rev.comp", "avg.bitscore"), no.rows = T)

count1 <- make.counter()

for (i in 1:nrow(unique.groups)) {
  temp.df <- filter(blastdf1, qseqid == unique.groups[i, 1], sseqid == unique.groups[i, 2])

  split.numeric.vectorv2 <- function(sstart, send, threshold) {
    sstart <- sstart[-1]
    send <- send[-length(send)]

    g <- data.frame(send, sstart)
    g.diffs <- abs(g$sstart - g$send)

    cons1 <- function(x) {
      # get consecutive integer ranges / integer runs
      diffs <- c(1, diff(x))
      start_indexes <- c(1, which(diffs > 1))
      end_indexes <- c(start_indexes - 1, length(x))
      g <- data.frame(x[start_indexes], x[end_indexes])
      colnames(g) <- c("start", "end")
      g
    }

    group.coords <- which(g.diffs < threshold)
    if (length(group.coords) > 0) {
      groups1 <- cons1(group.coords)
      groups2 <- list()
      for (i in 1:nrow(groups1)) {
        groups2 <- c(groups2, list(c(groups1[i, 1]:groups1[i, 2], groups1[i, 2] + 1)))
      }
```

```
78
79          all.rows <- 1:(length(sstart) + 1)
80          all.rows <- all.rows[-which(all.rows %in% unlist(groups2))]
81          all.groups <- lapply(all.rows, function(x) x)
82          all.groups <- c(all.groups, groups2)
83        } else {
84          all.rows <- 1:(length(sstart) + 1)
85          all.groups <- lapply(all.rows, function(x) x)
86        }
87
88        all.groups
89      }
90
91      split.temp.df <- function(temp.df, orientation1) {
92        temp.df.corrected <- temp.df[which(temp.df$orientation == ↩
              orientation1), ]
93        # determine whether there is more than one locus involved in this ↩
              group of hits
94        correct.groups <- split.numeric.vectorv2(temp.df.corrected$sstart,↩
               temp.df.corrected$send, dist.threshold)
95        for (x in 1:length(correct.groups)) {
96          temp.df.corrected$sseqid[correct.groups[[x]]] <- paste0(↩
                temp.df.corrected$sseqid[correct.groups[[x]]], ".!!$", ↩
                orientation1, x)
97        }
98
99        temp.df.corrected
100     }
101
102     rev.orientation.coords <- which(temp.df$orientation == "R")
103     forward.orientation.coords <- which(temp.df$orientation == "F")
104     if (length(rev.orientation.coords) > 0) {
105       temp.df[rev.orientation.coords, ] <- split.temp.df(temp.df, "R")
106     }
107
108     if (length(forward.orientation.coords) > 0) {
```

```
109       temp.df[forward.orientation.coords, ] <- split.temp.df(temp.df, "F↩
              ")
110     }
111     # if two groups of hits are present in the same scaffold / ↩
            chromosome,
112     # these hits will no longer be ordered by bitscore due to ↩
            sort.blastdf()
113     # at the start of the script. Here we calculate mean bitscore for ↩
            these newly identified
114     # groups of hits, and sort groups of hits by this in descending ↩
            order.
115
116     temp.df.unique.scaffolds <- unique(temp.df$sseqid)
117     mean.bitscores <- unlist(lapply(temp.df.unique.scaffolds, function(x↩
            ) {
118       temp.df.filtered <- filter(temp.df, sseqid == x)
119       mean(as.numeric(temp.df.filtered$bitscore))
120     }))
121
122     transformation.coords <- sort(mean.bitscores, decreasing = T, ↩
            index.return = T)$ix
123
124     transformation.coords2 <- unlist(lapply(temp.df.unique.scaffolds[↩
            transformation.coords], function(x) {
125       which(x == temp.df$sseqid)
126     }))
127
128     temp.df <- temp.df[transformation.coords2, ]
129
130     check.group.orientation <- split(temp.df, factor(temp.df$sseqid, ↩
            levels = unique(temp.df$sseqid)))
131
132     check.group.orientation <- lapply(check.group.orientation, function(↩
            x) {
133       # if HSPs are near to each other but in different orientations, ↩
              separate them into different groups
```

```
    same.orientation <- (length(unique(x$orientation)) == 1)
    if (same.orientation == F) {
      group1 <- x[which(x$sstart < x$send), ]
      group1$sseqid <- paste0(group1$sseqid, "_1")
      group2 <- x[which(!x$sstart < x$send), ]
      group2$sseqid <- paste0(group2$sseqid, "_2")
      x <- bind_rows(group1, group2)
    }
    x
  })

  temp.df <- bind_rows(check.group.orientation)
  blastdf1[which(blastdf1$qseqid == unique.groups[i, 1] & blastdf1$↩
      sseqid == unique.groups[i, 2]), ] <- temp.df

  for (i2 in 1:length(unique(temp.df$sseqid))) {
    # CONCATENATE GROUPS OF HITS TOGETHER INTO potential.homeologues ↩
        DATAFRAME
    temp.df2 <- filter(temp.df, sseqid == unique(temp.df$sseqid)[i2])
    temp.df2$qseqid <- as.character(temp.df2$qseqid)
    temp.df2$sseqid <- as.character(temp.df2$sseqid)

    group.orientation <- temp.df2$orientation[1]

    # populate potential.homeologues dataframe where average bitscore ↩
        is higher than 200
    if (mean(temp.df2$bitscore) > 200) {
      # if this is in normal orientation, do x...
      potential.homeologues <- add_row(potential.homeologues)
      potential.homeologues$query[nrow(potential.homeologues)] <- ↩
          temp.df2[1, 1]
      potential.homeologues$scaffold[nrow(potential.homeologues)] <- ↩
          temp.df2[1, 2]
      potential.homeologues$start[nrow(potential.homeologues)] <- min(↩
          temp.df2$sstart)
      potential.homeologues$end[nrow(potential.homeologues)] <- max(↩
```

```
            temp.df2$send)
164         potential.homeologues$avg.bitscore[nrow(potential.homeologues)] ↩
              <- mean(temp.df2$bitscore)
165         potential.homeologues$max.bitscore[nrow(potential.homeologues)] ↩
              <- max(temp.df2$bitscore)
166         potential.homeologues$avg.percent.identical[nrow(↩
              potential.homeologues)] <- mean(temp.df2$↩
              percentage.identical)
167         if (group.orientation == "F") {
168           potential.homeologues$rev.comp[nrow(potential.homeologues)] <-↩
                F
169         } else {
170           potential.homeologues$rev.comp[nrow(potential.homeologues)] <-↩
                T
171         }
172
173         potential.homeologues$query.start[nrow(potential.homeologues)] ↩
              <- min(temp.df2$qstart)
174         potential.homeologues$query.end[nrow(potential.homeologues)] <- ↩
              max(temp.df2$qend)
175         potential.homeologues$num_hsp[nrow(potential.homeologues)] <- ↩
              nrow(temp.df2)
176       }
177     }
178   }
179
180
181
182   potential.homeologues$length <- as.numeric(potential.homeologues$end) ↩
        - as.numeric(potential.homeologues$start)
183   potential.homeologues$groupid <- potential.homeologues$scaffold
184   potential.homeologues$scaffold <- multi.str.split(↩
        potential.homeologues$scaffold, "\\.\\!\\!\\$", 1)
185   potential.homeologues$homo_length <- potential.homeologues$query.end -↩
         potential.homeologues$query.start
186
```

```
187   # try and identify the matching genomic sequence to the input sequence - avg.bitscore sometimes fails here
188   # e.g. if there is a small exon seperated by an intron from the main sequence, it will bring the avg.bitscore down
189   identi.coord <- which.max((potential.homeologues$homo_length / length(input_sequence[[1]])) * (potential.homeologues$avg.percent.identical / 100))
190   g <- 1:nrow(potential.homeologues)
191   g <- g[-identi.coord]
192   potential.homeologues <- potential.homeologues[c(identi.coord, g), ]
193
194   if (length(input_sequence[[1]]) > 1500) {
195     # this will remove all blast hits for small sequences. need an if statement
196     coord_to_rm <- which(potential.homeologues$length < 500)
197     if (length(coord_to_rm) != nrow(potential.homeologues)) {
198       if (length(coord_to_rm) > 0) potential.homeologues <- potential.homeologues[-coord_to_rm, ]
199     }
200   }
201
202   existing.homeologue.files <- grep("potential_homeologues", list.files(p("jobs/", gene.name, "/blast.results/")))
203   if (length(existing.homeologue.files) == 0) {
204     write.csv(potential.homeologues, p("jobs/", gene.name, "/blast.results/potential_homeologues1.csv"), row.names = F)
205   } else {
206     write.csv(potential.homeologues, p("jobs/", gene.name, "/blast.results/potential_homeologues", (length(existing.homeologue.files) + 1), ".csv"), row.names = F)
207   }
208
209
210
211   list(potential.homeologues, blastdf1)
212 }
```

```
213
214
215
216 extract.sequence <- function(genome1, blast.df.parsed, row.coords, ↩
      start.buffer, end.buffer) {
217   # extracts a related sequence from a genome assembly
218   # args:
219   # genome1 - DNAStringSet object containing the genome assembly of ↩
        interest
220   # blast.df.parsed - dataframe produced by parse.scaffold.blast()
221   # row.coords - Numeric vector; the row coordinates of the sequences to↩
         used in blast.df.parsed
222   # start.buffer - Integer; how much extra sequence before the start ↩
        indicated in blast.df.parsed to extract
223   # end.buffer - Integer; how much extra sequence after the end ↩
        indicated in blast.df.parsed to extract
224
225   # parse the original scaffold name from blastdf1.parsed (remove the ↩
        appended position in kb)
226   original.scaf.names <- multi.str.split(blast.df.parsed$scaffold, ".$!"↩
        , 1)
227
228   genome2 <- genome1[match(original.scaf.names[row.coords], names(↩
        genome1))]
229
230   for (i in 1:length(row.coords)) {
231     # check if start.buffer reaches before the start of the scaffold, ↩
          and likewise if end.buffer extends bigger than the total length
232     extract.start <- (as.numeric(blast.df.parsed$start[i]) - ↩
          start.buffer)
233     if (extract.start < 1) extract.start <- 1
234     extract.end <- (as.numeric(blast.df.parsed$end[i]) + end.buffer)
235     if (extract.end > length(genome2[[i]])) extract.end <- length(↩
          genome2[[i]])
236     print("extract.start")
237     print(extract.start)
```

```
238     genome2[[i]] <- genome2[[i]][extract.start:extract.end]
239     if (blast.df.parsed$rev.comp[i] == T) genome2[[i]] <- ↩
            reverseComplement(genome2[[i]])
240   }
241
242   # append the start coordinate (in kb) of the blast hit to the name of ↩
        the sequence for identification later
243   names(genome2) <- paste(blast.df.parsed$scaffold[row.coords], ".$!", ↩
        round((as.numeric(blast.df.parsed$start[row.coords]) / 1000)), sep↩
         = "")
244   print("genome2")
245   print(genome2)
246   genome2
247 }
248
249 #    ↩
      ------------------------------------------------------------------------↩

250 #    BEGIN PROCESSING ↩
                                                                        ####
251
252
253 number.genomes <- max(na.omit(unique(as.numeric(multi.str.split(↩
      config.variables, "_", 2)))))
254
255 main.processing <- function() {
256   extract.sequence.w.flanking.regions <- function(genome.number) {
257     # read the configuration file
258
259     config.variables <- multi.str.split(config.file, "=", 1)
260     # begin by parsing the config file for the genome name, fasta file ↩
          path and blastdb path
261     config.settings <- config.file[grep(genome.number, config.variables)↩
          ]
262
263     config.settings.temp <- config.settings[[1]]
```

```
264     genome.name <- strsplit(config.settings.temp[1], "=")
265     genome.name <- genome.name[[1]][2]
266
267     config.settings.temp <- config.settings[[2]]
268     fa.path <- strsplit(config.settings.temp[1], "=")
269     fa.path <- fa.path[[1]][2]
270
271     config.settings.temp <- config.settings[[3]]
272     blastdb.path <- strsplit(config.settings.temp[1], "=")
273     blastdb.path <- blastdb.path[[1]][2]
274
275     blast.files <- list.files(p("jobs/", gene.name, "/blast.results"), ↩
            pattern = paste0(genome.number, ".*?.blast"))
276
277     blastdf1 <- tryCatch(read.blast(p("jobs/", gene.name, "/↩
            blast.results/", blast.files[1])), error = function(e) {
278       write("No BLAST hits for this sequence", p("jobs/", gene.name, "/↩
              primers/error.txt"))
279       stop("No BLAST hits for this sequence")
280     })
281
282     blastdf1$sstart.mb <- blastdf1$sstart / 1000000
283     blastdf1$send.mb <- blastdf1$send / 1000000
284     blastdf_orig <- read.blast(p("jobs/", gene.name, "/blast.results/", ↩
            blast.files[1]))
285
286     # read the fasta index for this particular genome
287     fasta.index1 <- read.csv(p("./fasta.indexes/", genome.name, ".fa.idx↩
            "), stringsAsFactors = F, header = T)
288
289     print("opt$cds.max.intron.size")
290     print(opt$cds.max.intron.size)
291
292     blastdf1.parsed_orig <- parse.scaffold.blast(blastdf1, opt$↩
            cds.max.intron.size)[[1]]
293     blastdf1.parsed <- parse.scaffold.blast(blastdf1, opt$↩
```

```
        cds.max.intron.size)[[1]]
294
295     original.scaf.names <- multi.str.split(blastdf1.parsed$scaffold, ".$↩
            !", 1)
296
297     fasta.index1$offset <- as.numeric(fasta.index1$offset)
298
299     genome.assembly.subset.genomic.match <<- readDNAStringSet(↩
            fasta.index1[match(original.scaf.names[1], multi.str.split(↩
            fasta.index1$desc, " ", 1)), ])
300     names(genome.assembly.subset.genomic.match) <- multi.str.split(names↩
            (genome.assembly.subset.genomic.match), " ", 1)
301     template_sequence_genomic <- extract.sequence(↩
            genome.assembly.subset.genomic.match, blastdf1.parsed[1, ], 1, ↩
            opt$start.buffer, opt$end.buffer)
302
303     opt$fasta.path <<- p("jobs/", gene.name, "/seq/extended/seqs/↩
            input_w_flanking.fa")
304     query.fa.path <- p("jobs/", gene.name, "/seq/extended/seqs/↩
            input_w_flanking.fa")
305     writeXStringSet(template_sequence_genomic, p("jobs/", gene.name, "/↩
            seq/extended/seqs/input_w_flanking.fa"))
306
307     input_sequence <<- readDNAStringSet(p("jobs/", gene.name, "/seq/↩
            extended/seqs/input_seq.fa"))
308
309     # SECOND BLAST WITH FLANKING REGIONS OF INPUT SEQUENCE INCLUDED
310     source("scripts/perform.blast.rscript.R")
311   }
312
313   extract.homologues <- function(genome.number, number.to.extract, ↩
        setup.dialign.anchors) {
314     if (missing(setup.dialign.anchors)) setup.dialign.anchors <- T
315     if (missing(number.to.extract)) number.to.extract <- "all"
316
317     config.variables <- multi.str.split(config.file, "=", 1)
```

```
318     config.settings <- config.file[grep(genome.number, config.variables)]
319
320     config.settings.temp <- config.settings[[1]]
321     genome.name <- strsplit(config.settings.temp[1], "=")
322     genome.name <- genome.name[[1]][2]
323
324     # read the fasta index for this particular genome
325     fasta.index1 <- read.csv(p("./fasta.indexes/", genome.name, ".fa.idx"), stringsAsFactors = F, header = T)
326
327     #### ITERATION TWO ####
328     blast.files <- list.files(p("jobs/", gene.name, "/blast.results"), pattern = paste0(genome.number, ".*?.blast"))
329
330     blastdf0 <- read.blast(p("jobs/", gene.name, "/blast.results/", blast.files[grep(paste0(genome.number, ".*?w_flanking"), blast.files)]))
331
332     # SEQUENCE EXTRACTION WITH FULL TEMPLATE (INCLUDING FLANKING REGIONS)
333     blastdf1.parsed <- parse.scaffold.blast(blastdf0, opt$cds.max.intron.size)
334
335     fasta.index1$offset <- as.numeric(fasta.index1$offset)
336     genome.assembly.subset.genomic.match <<- readDNAStringSet(fasta.index1[match(unique(blastdf1.parsed[[1]]$scaffold), multi.str.split(fasta.index1$desc, " ", 1)), ])
337     names(genome.assembly.subset.genomic.match) <- multi.str.split(names(genome.assembly.subset.genomic.match), " ", 1)
338     sequences <- DNAStringSet()
339     if (number.to.extract == "all") number.to.extract <- nrow(blastdf1.parsed[[1]])
340
341     # SEQUENCE EXTRACTION
342     for (i in 1:number.to.extract) {
```

```
343       # MASKING OF INTER-HSP DISTANCES WITHIN THE SAME GROUP WITH Ns
344       rev.comp <- blastdf1.parsed[[1]][i, ]$rev.comp
345       rchr <- blastdf1.parsed[[1]][i, ]$scaffold
346       temp.df <- blastdf1.parsed[[2]][which(blastdf1.parsed[[2]]$sseqid ↩
              == blastdf1.parsed[[1]]$groupid[i]), ]
347       chr <- blastdf1.parsed[[1]]$scaffold[i]
348       # remove any _ concatenations that distinguished groups in ↩
              orientation check
349       # chr = strsplit(chr, "_")
350       # chr = chr[[1]][1]
351
352       if (nrow(temp.df) == 1) {
353         # if only 1 HSP, just add it to the list of sequences
354         if (rev.comp == F) sequences <- c(sequences, DNAStringSet(↩
                genome.assembly.subset.genomic.match[[chr]][temp.df$sstart↩
                [1]:temp.df$send[1]]))
355         if (rev.comp == T) sequences <- c(sequences, DNAStringSet(↩
                reverseComplement(genome.assembly.subset.genomic.match[[chr↩
                ]][temp.df$sstart[1]:temp.df$send[1]])))
356       } else {
357         if (opt$mask.inter.hsp.distances == F | i == 1) {
358           # extract sequences without masking inter HSP distances with ↩
                  Ns if this option is false
359           if (rev.comp == F) sequences <- c(sequences, DNAStringSet(↩
                  genome.assembly.subset.genomic.match[[chr]][min(temp.df$↩
                  sstart):max(temp.df$send)]))
360           if (rev.comp == T) sequences <- c(sequences, DNAStringSet(↩
                  reverseComplement(genome.assembly.subset.genomic.match[[↩
                  chr]][min(temp.df$sstart):max(temp.df$send)])))
361         } else {
362           temp.df <- temp.df[sort(temp.df$sstart, index.return = T)$ix, ↩
                  ]
363
364           if (rev.comp == T) temp.df <- temp.df[sort(temp.df$sstart, ↩
                  index.return = T, decreasing = T)$ix, ]
365
```

```
366         group.subsequences <- DNAStringSet()
367         subseq.differences <- as.numeric()
368         for (x in 1:nrow(temp.df)) {
369           if (all(temp.df$sstart < temp.df$send)) {
370             if (nrow(temp.df) == 0) browser()
371             if (rev.comp == F) group.subsequences <- c(↩
                    group.subsequences, DNAStringSet(↩
                    genome.assembly.subset.genomic.match[[chr]][temp.df$↩
                    sstart[x]:temp.df$send[x]]))
372             if (rev.comp == T) group.subsequences <- c(↩
                    group.subsequences, DNAStringSet(reverseComplement(↩
                    genome.assembly.subset.genomic.match[[chr]][temp.df$↩
                    sstart[x]:temp.df$send[x]])))
373           }
374
375           if (x != nrow(temp.df)) {
376             # calculate how far apart the HSPs are
377             subseq.differences <- c(subseq.differences, abs(temp.df$↩
                    send[x] - temp.df$sstart[(x + 1)]))
378           }
379         }
380
381         # combine HSPs with interleaving regions masked by Ns
382         subseq.gaps <- lapply(subseq.differences, function(x) ↩
              DNAStringSet(DNAString(paste(rep("N", 100), collapse = "")↩
              )))
383         subsequences.w.gaps <- DNAStringSet()
384         for (x in 1:length(group.subsequences)) {
385           subsequences.w.gaps <- c(subsequences.w.gaps, ↩
                group.subsequences[x])
386           if (x != length(group.subsequences)) subsequences.w.gaps <- ↩
                c(subsequences.w.gaps, subseq.gaps[[x]])
387         }
388
389         masked.subsequence <- DNAStringSet(DNAString(do.call(paste0, ↩
              lapply(subsequences.w.gaps, as.character))))
```

```
390           sequences <- c(sequences, masked.subsequence)
391         }
392       }
393     }
394
395     if (nrow(blastdf1.parsed[[1]]) <= 1 & genome.number == 1) {
396       write("No homologues found", p("jobs/", gene.name, "/error.txt"))
397       stop("No homologues found")
398     }
399     # else if(nrow(blastdf1.parsed[[1]] < 1)){
400     # write('No homologues found', p("jobs/", gene.name, "/error.txt"))
401     # stop('No homologues found')
402     # }
403
404     names(sequences) <- paste0(blastdf1.parsed[[1]]$scaffold, "_", ↩
          blastdf1.parsed[[1]]$query.start)[1:number.to.extract]
405
406     # sequences = c(DNAStringSet(input_sequence), sequences)
407
408     if (setup.dialign.anchors == T) {
409       # SETUP ANCHOR POINTS FOR DIALIGN
410       coord.query.start <- c(1, sort(blastdf1.parsed[[1]]$query.start[2:↩
              number.to.extract], index.return = T)$ix + 1) # order the ↩
              sequences by query start position
411       blastdf1.parsed[[1]] <- blastdf1.parsed[[1]][coord.query.start, ]
412
413       sequences <- sequences[coord.query.start]
414
415       dialign.df1 <- blastdf1.parsed[[1]]
416       dialign.df1$dialign1 <- 1 # position of the first sequence to be ↩
              anchored
417       dialign.df1$dialign2 <- 1:nrow(dialign.df1) # position of the ↩
              second sequence to be anchored
418       dialign.df1$dialign3 <- dialign.df1$query.start # beginning ↩
              position of the anchor point in sequence 1
419       dialign.df1$dialign4 <- 1 # beginning position of the anchor point↩
```

```
            in sequence 2
      dialign.df1$dialign5 <- 5 # length of anchor
      dialign.df1$dialign6 <- 20 # anchor priority
      dialign.df1 <- dialign.df1[-1, ]

      dialign.df1 <- dialign.df1[, grep("dialign", colnames(dialign.df1))]
      new_dialign_anchors <- dialign.df1

      new_dialign_anchors[, 1] <- new_dialign_anchors[, 1]
      new_dialign_anchors[, 2] <- new_dialign_anchors[, 2]
    } else {
      new_dialign_anchors <- "No anchors"
    }

    # system(p("scripts/run.dialign.sh jobs/", gene.name, "/"))
    return(list(sequences, new_dialign_anchors))
  }


  sequences <- lapply(1:number.genomes, function(genome.number) {
    if (genome.number == 1) {
      extract.sequence.w.flanking.regions(genome.number)
      return(extract.homologues(genome.number))
    } else {
      return(extract.homologues(genome.number, number.to.extract = 1, setup.dialign.anchors = F))
    }
  })

  if (number.genomes > 1) {
    # combine genome sequences from the other genomes
    other.genome.seq <- lapply(sequences[2:number.genomes], function(x) {
      x[[1]]
    })
```

```
452
453     other.genome.seq <- do.call(c, other.genome.seq)
454
455     # all.seq = c(input_sequence, sequences[[1]][[1]][1], ↩
            other.genome.seq, sequences[[1]][[1]][2:length(sequences↩
            [[1]][[1]])])
456
457     all.seq <- c(input_sequence, other.genome.seq)
458
459     dialign_anc <- sequences[[1]][[2]]
460     dialign_anc$dialign1 <- dialign_anc$dialign1 + 1
461     dialign_anc$dialign2 <- dialign_anc$dialign2 + length(↩
            other.genome.seq) + 1
462   } else {
463     all.seq <- c(input_sequence, sequences[[1]][[1]])
464     dialign_anc <- sequences[[1]][[2]]
465     dialign_anc$dialign1 <- dialign_anc$dialign1 + 1
466     dialign_anc$dialign2 <- dialign_anc$dialign2 + 1
467   }
468
469
470   write.table(dialign_anc, p("jobs/", gene.name, "/seq/extended/seqs/↩
          all.anc"), quote = F, sep = " ", col.names = F, row.names = F)
471   writeXStringSet(all.seq, p("jobs/", gene.name, "/seq/extended/seqs/↩
          all.fa"))
472 }
473
474
475 main.processing()
```

7.2 A.2 AutoCloner primer selection script

The following listing is an excerpt from the primer selection script of AutoCloner, showcasing two functions, grab.homeologous.snps_new (line 1) and find.best.primers (line 111). The first scans the multiple sequence alignment and returns coordinates of SNPs between the various homologues. The latter is used for primer selection, acting as an interface to Primer3, both through generation of input files and parsing of output files, as well as containing the algorithm that generates sets of overlapping PCR products through selection of optimal primers.

```
1  grab.homeologous.snps_new <- function(input.row, template.row, ↩
       homologue.rows, multiple.alignment,
2                                         perform.masking, mask.bin.size, ↩
                                             mask.threshold, ↩
                                             allow.hyphens.in.mask, ↩
                                             allow.hyphens.for.snp.detection↩
                                             ) {
3    # gets homeologous snps when there is only one genome
4    # takes a DNAMultipleAlignment object and returns a numeric vector of ↩
         the column coordinates containing homeologous SNPs
5    # args:
6    # input.row - Integer; the row of the sequence inputted by the user (↩
         usually 1)
7    # template.row - Integer; the row of the sequence to design primers ↩
         from (usually 2)
8    # homologue.rows - numeric vector containing the row coordinates of ↩
         the homologous sequences (either paralogous or homeologous)
9    # multiple.alignment - DNAMultipleAlignment class
10   # perform.masking - Boolean, indicates whether sequences of low ↩
         similarity to template should be masked
11   # allow.hyphens.for.snp.detection - Boolean, indicates whether hyphens↩
          in homologues (not the template sequence) will be considered SNPs
12   if (missing(perform.masking)) perform.masking <- opt$perform.masking
13   if (missing(mask.bin.size)) mask.bin.size <- opt$mask.bin.size
```

```
14  if (missing(mask.threshold)) mask.threshold <- opt$mask.threshold
15  if (missing(allow.hyphens.in.mask)) allow.hyphens.in.mask <- opt$↩
        allow.hyphens.in.mask
16  if (missing(allow.hyphens.for.snp.detection)) ↩
        allow.hyphens.for.snp.detection <- opt$↩
        allow.hyphens.for.snp.detection
17
18  # NB. Insertions "-" in the template sequence cannot be allowed when ↩
        classifying SNPs, as these
19  # will be subsequently removed by ↩
        get.coordinates.after.removing.hyphens(), meaning that the
20  # program maps the primers to the wrong locations in the final ↩
        multiple sequence alignment output.
21
22  mult.align.mat1 <- convert.to.character.data.frame(as.data.frame(↩
        as.matrix(multiple.alignment)))
23
24  perform.masking.function <- function(msa.matrix1, mask.bin.size, ↩
        mask.threshold, allow.hyphens.in.mask) {
25    # mask.threshold - integer, underneath what percentage of similarity↩
          should masking be performed?
26    #             e.g. 40 - when bins have less than 40% nucleotides in ↩
          common, mask them
27    if (missing(mask.bin.size)) mask.bin.size <- 10
28    if (missing(mask.threshold)) mask.threshold <- 40
29    if (missing(allow.hyphens.in.mask)) allow.hyphens.in.mask <- F
30
31    msa.matrix1.orig <- msa.matrix1
32    start.sequence.bins <- seq(1, ncol(msa.matrix1), mask.bin.size)
33    end.sequence.bins <- c((start.sequence.bins[2:length(↩
          start.sequence.bins)] - 1), ncol(msa.matrix1))
34
35    # performing masking of regions with low sequence identity (in bins ↩
          of 10)
36    bin.dfs1 <- lapply((2 + number.genomes):nrow(msa.matrix1), function(↩
          z) { # lapply across rows
```

```
37      bin.similarities <- unlist(Map(function(x, y) {
38        template.bin.seq <- msa.matrix1[2, x:y]
39        target.bin.seq <- msa.matrix1[z, x:y]
40
41        if (allow.hyphens.in.mask == F) {
42          if ((length(which(template.bin.seq == "-")) > 3) | length(which(target.bin.seq == "-")) > 3) {
43            return(10)
44          } # don't include bins with hyphens in masking
45        }
46
47        length(which(msa.matrix1[2, x:y] == msa.matrix1[z, x:y]))
48      }, start.sequence.bins, end.sequence.bins))
49
50      bin.df1 <- data.frame(start.sequence.bins, end.sequence.bins, bin.similarities)
51      colnames(bin.df1) <- c("sbin", "ebin", "nsim")
52
53      # mask bins with less than 4 nucleotides in common with the template sequence
54      mask.threshold2 <- (mask.bin.size / 100) * mask.threshold
55      bin.df1 <- bin.df1[which(bin.df1$nsim < mask.threshold2), ]
56
57      unlist(Map(function(x, y) {
58        seq(x, y, 1)
59      }, bin.df1$sbin, bin.df1$ebin))
60    })
61
62    print("Performing sequence masking")
63    # mask regions with low sequence identity
64    for (i in 1:length(bin.dfs1)) {
65      msa.matrix1[(i + 2), bin.dfs1[[i]]] <- "U"
66    }
67
68    msa.matrix1
69  }
```

```
70
71  if (perform.masking == T) mult.align.mat1 <- perform.masking.function(↩
        mult.align.mat1, mask.bin.size, mask.threshold, ↩
        allow.hyphens.in.mask)
72
73  counter1 <- 1
74  g <- lapply(mult.align.mat1, function(x) {
75    if (x[2] == "-") {
76      return(0)
77    } # return 0 if template sequence has an insertion
78
79    if (allow.hyphens.for.snp.detection == F) {
80      if ("-" %in% x[3:length(x)]) {
81        return(0)
82      }
83    }
84
85    if (number.genomes == 1) {
86      if (x[template.row] %in% x[homologue.rows]) {
87        snp <- 0
88      } else {
89        snp <- 1
90      }
91    } else {
92      if (length(unique(x[template.row:(template.row + number.genomes - ↩
            1)])) == 1) { # if all varieties have the same base
93        if (x[template.row] %in% x[homologue.rows]) {
94          snp <- 0
95        } else {
96          snp <- 1
97        }
98      } else {
99        snp <- 0 # no SNP if varietal genomes have different bases
100     }
101   }
102   counter1 <<- counter1 + 1
```

```
103     snp
104   })
105
106   unlist(g)
107 }
108
109
110
111 find.best.primers <- function(multiple.alignment, ↩
      template.sequence.row.number, snp.coords.after.filter, ↩
      start.coord.after.filter, end.coord.after.filter, ↩
      product.size.range, span.whole.gene, start.buffer, homologous.snps, ↩
      coords) {
112   # Automatically obtains primer sequences
113   # args:
114   # multiple.alignment - a DNAMultipleAlignment object
115   # template.sequence.row.number - Integer; the multiple alignment row ↩
        of the sequence to use as a template in primer3
116   # snp.coords.after.filter - Numeric vector; obtained using ↩
        grab.homeologous.snps() and then ↩
        get.coordinates.after.removing.hyphens()
117   # start.coord.after.filter - Integer; position of the first base of ↩
        the start codon after removing hyphens
118   # end.coord.after.filter - Integer; position of the final base in the ↩
        coding sequence after removing hyphens
119   # product.size.range - a numeric vector with two elements, the first ↩
        being the minimum product size, the second the maximum
120   # span.whole.gene - Boolean; should the product size span the entire ↩
        gene with only one set of primers?
121   # start.buffer - Integer; how many bases before the start of the gene ↩
        should be allowed in the product?
122
123   if (missing(span.whole.gene)) span.whole.gene <- F
124   if (missing(start.buffer)) start.buffer <- start.coord.after.filter
125
126   list.best.primer.start.coords <- as.numeric()
```

```
127 list.best.primer.end.coords <- as.numeric()
128 mult.align2 <- conv.mult.align.dnastringset(multiple.alignment)
129
130 generate.primer3.input.files <- function(template.sequence2, p3.seqid,↩
        product.size.min, product.size.max, left.end.coord,
131                                          right.end.coord, f.or.r) {
132   # args:
133   # template.sequence2 - a DNAString object without inserts ("-"s)
134   # p3.seqid - character string indicating name of sequence (used both↩
        inside the primer3 input file and in the title of the primer3 ↩
       input file)
135   # f.or.r - "F" for forward primer, "R" for reverse primer
136
137   # primer3 variables:
138   p3.template <- as.character(template.sequence2)
139   p3.product.size.range <- "100-10000"
140
141   # note here that line breaks "\n" have to be added in manually as
142   # writeLines automatically adds a line break to the end of every ↩
       line,
143   # whilst primer3_core will not accept a file in which the last line
144   # has a line break on it
145   if (f.or.r == "F") {
146     primer3.input <- c(
147       p("SEQUENCE_ID=", p3.seqid, "\n"),
148       p("SEQUENCE_TEMPLATE=", p3.template, "\n"),
149       p("PRIMER_PRODUCT_SIZE_RANGE=", p3.product.size.range, "\n"),
150       p("SEQUENCE_FORCE_LEFT_END=", left.end.coord, "\n"),
151       "="
152     )
153   } else if (f.or.r == "R") {
154     primer3.input <- c(
155       p("SEQUENCE_ID=", p3.seqid, "\n"),
156       p("SEQUENCE_TEMPLATE=", p3.template, "\n"),
157       p("PRIMER_PRODUCT_SIZE_RANGE=", p3.product.size.range, "\n"),
158       p("SEQUENCE_FORCE_RIGHT_END=", right.end.coord, "\n"),
```

```
159         "="
160       )
161     } else {
162       primer3.input <- c(
163         p("SEQUENCE_ID=", p3.seqid, "\n"),
164         p("SEQUENCE_TEMPLATE=", p3.template, "\n"),
165         p("PRIMER_PRODUCT_SIZE_RANGE=", p3.product.size.range, "\n"),
166         p("SEQUENCE_FORCE_RIGHT_END=", right.end.coord, "\n"),
167         p("SEQUENCE_FORCE_LEFT_END=", left.end.coord, "\n"),
168         "="
169       )
170     }
171     if (f.or.r == "F" | f.or.r == "R") {
172       output.filepath <- file(p(project.path, "jobs/", gene.name, "/↩
              primers/input/primer3.", p3.seqid, ".", f.or.r, ".txt"), "wb")
173     } else {
174       output.filepath <- file(p(project.path, "jobs/", gene.name, "/↩
              primers/input/primer3.", p3.seqid, ".txt"), "wb")
175     }
176     writeLines(primer3.input, output.filepath, sep = "")
177     close(output.filepath)
178     "Done"
179   }
180
181   # make primers for all possible SNPs, then evaluate
182   template.sequence <- mult.align2[[template.sequence.row.number]]
183
184   template.sequence2 <- remove.inserts(template.sequence)
185   generate.all.primer.penalites <- function(x) {
186     print("Generating primer3 files")
187     print("Running primer3")
188
189     primer3.forward.input <- c(
190       p("SEQUENCE_ID=AllForwardPrimers"),
191       p("SEQUENCE_TEMPLATE=", template.sequence2, ""),
192       "PRIMER_TASK=pick_primer_list",
```

```
193       "PRIMER_PICK_RIGHT_PRIMER=0",
194       p("PRIMER_NUM_RETURN=", (length(template.sequence2) * 6)),
195       "="
196     )
197
198     writeLines(primer3.forward.input, p("jobs/", gene.name, "/primers/↩
          input/forwardprimers.txt"))
199
200     primer3.reverse.input <- c(
201       p("SEQUENCE_ID=AllReversePrimers"),
202       p("SEQUENCE_TEMPLATE=", template.sequence2, ""),
203       "PRIMER_TASK=pick_primer_list",
204       "PRIMER_PICK_LEFT_PRIMER=0",
205       p("PRIMER_NUM_RETURN=", (length(template.sequence2) * 6)),
206       "="
207     )
208
209     writeLines(primer3.reverse.input, p("jobs/", gene.name, "/primers/↩
          input/reverseprimers.txt"))
210
211     system(p(project.path, "jobs/", gene.name, "/primers/run.primer3.sh ↩
          ", "jobs/", gene.name))
212     print("Finished")
213
214     # PARSE LEFT PRIMERS
215
216     primer3.forward.output <- readLines(p("jobs/", gene.name, "/primers/↩
          output/forwardprimers.txt.output.txt"))
217     pos1 <- primer3.forward.output[grep("PRIMER_LEFT_[0-9]*=", ↩
          primer3.forward.output)]
218
219     pos2 <- multi.str.split(pos1, "=", 2)
220     pos3.1 <- multi.str.split(pos2, ",", 1)
221     pos3.2 <- multi.str.split(pos2, ",", 2)
222     pos4 <- as.numeric(pos3.1) + (as.numeric(pos3.2)) - 1 # translate ↩
          primer3 coordinate to SNP coordinate (add length to starting pos↩
```

```
        - 1)
223
224     PRIMER_LEFT_X_PENALTY <- as.numeric(multi.str.split(↩
            primer3.forward.output[grep("PRIMER_LEFT_[0-9]*_PENALTY=", ↩
            primer3.forward.output)], "=", 2))
225     PRIMER_LEFT_X_SEQUENCE <- multi.str.split(primer3.forward.output[↩
            grep("PRIMER_LEFT_[0-9]*_SEQUENCE=", primer3.forward.output)], "↩
            =", 2)
226     PRIMER_LEFT_X <- pos2
227     PRIMER_LEFT_X_TM <- multi.str.split(primer3.forward.output[grep("↩
            PRIMER_LEFT_[0-9]*_TM=", primer3.forward.output)], "=", 2)
228     PRIMER_LEFT_X_GC_PERCENT <- multi.str.split(primer3.forward.output[↩
            grep("PRIMER_LEFT_[0-9]*_GC_PERCENT", primer3.forward.output)], ↩
            "=", 2)
229     PRIMER_LEFT_X_SELF_ANY_TH <- multi.str.split(primer3.forward.output[↩
            grep("PRIMER_LEFT_[0-9]*_SELF_ANY_TH", primer3.forward.output)],↩
             "=", 2)
230     PRIMER_LEFT_X_SELF_END_TH <- multi.str.split(primer3.forward.output[↩
            grep("PRIMER_LEFT_[0-9]*_SELF_END_TH", primer3.forward.output)],↩
             "=", 2)
231     PRIMER_LEFT_X_HAIRPIN_TH <- multi.str.split(primer3.forward.output[↩
            grep("PRIMER_LEFT_[0-9]*_HAIRPIN_TH", primer3.forward.output)], ↩
            "=", 2)
232     PRIMER_LEFT_X_END_STABILITY <- multi.str.split(↩
            primer3.forward.output[grep("PRIMER_LEFT_[0-9]*_END_STABILITY", ↩
            primer3.forward.output)], "=", 2)
233
234     left.parsed <- data.frame("name1", pos4, PRIMER_LEFT_X_PENALTY, ↩
            PRIMER_LEFT_X_SEQUENCE, PRIMER_LEFT_X, PRIMER_LEFT_X_TM, ↩
            PRIMER_LEFT_X_GC_PERCENT, PRIMER_LEFT_X_SELF_ANY_TH, ↩
            PRIMER_LEFT_X_SELF_END_TH, PRIMER_LEFT_X_HAIRPIN_TH, ↩
            PRIMER_LEFT_X_END_STABILITY)
235     colnames(left.parsed)[1:2] <- c("p.name", "pos")
236     left.parsed <- left.parsed[sort(left.parsed$pos, index.return = T)$↩
            ix, ]
237     left.parsed <- left.parsed[which(left.parsed$pos %in% ↩
```

```
    snp.coords.after.filter), ]
nrow(left.parsed)
left.parsed2 <- split(left.parsed, left.parsed$pos)
left.parsed2 <- lapply(left.parsed2, function(x) {
  x[which.min(as.numeric(x$PRIMER_LEFT_X_PENALTY)), ]
})

left.parsed3 <- bind_rows(left.parsed2)

left.parsed3$MSA.pos <- which(homologous.snps == 1)[which(
    snp.coords.after.filter %in% left.parsed3$pos)]
left.parsed3 <- left.parsed3[, c(1, 2, 12, 3:11)]
colnames(left.parsed3)[4] <- "pen"
colnames(left.parsed3)[5:ncol(left.parsed3)] <- gsub("X", "0", 
    colnames(left.parsed3)[5:ncol(left.parsed3)])

# PARSE RIGHT PRIMERS

primer3.reverse.output <- readLines(p("jobs/", gene.name, "/primers/
    output/reverseprimers.txt.output.txt"))
pos1 <- primer3.reverse.output[grep("PRIMER_RIGHT_[0-9]*=", 
    primer3.reverse.output)]

pos2 <- multi.str.split(pos1, "=", 2)
pos3.1 <- multi.str.split(pos2, ",", 1)
pos3.2 <- multi.str.split(pos2, ",", 2)
pos4 <- (as.numeric(pos3.1) - as.numeric(pos3.2)) + 1 # translate 
    primer3 coordinate to SNP coordinate

PRIMER_RIGHT_X_PENALTY <- as.numeric(multi.str.split(
    primer3.reverse.output[grep("PRIMER_RIGHT_[0-9]*_PENALTY=", 
    primer3.reverse.output)], "=", 2))
PRIMER_RIGHT_X_SEQUENCE <- multi.str.split(primer3.reverse.output[
    grep("PRIMER_RIGHT_[0-9]*_SEQUENCE=", primer3.reverse.output)], 
    "=", 2)
PRIMER_RIGHT_X <- pos2
```

```
264     PRIMER_RIGHT_X_TM <- multi.str.split(primer3.reverse.output[grep("↩
            PRIMER_RIGHT_[0-9]*_TM=", primer3.reverse.output)], "=", 2)
265     PRIMER_RIGHT_X_GC_PERCENT <- multi.str.split(primer3.reverse.output[↩
            grep("PRIMER_RIGHT_[0-9]*_GC_PERCENT", primer3.reverse.output)],↩
             "=", 2)
266     PRIMER_RIGHT_X_SELF_ANY_TH <- multi.str.split(primer3.reverse.output↩
            [grep("PRIMER_RIGHT_[0-9]*_SELF_ANY_TH", primer3.reverse.output)↩
            ], "=", 2)
267     PRIMER_RIGHT_X_SELF_END_TH <- multi.str.split(primer3.reverse.output↩
            [grep("PRIMER_RIGHT_[0-9]*_SELF_END_TH", primer3.reverse.output)↩
            ], "=", 2)
268     PRIMER_RIGHT_X_HAIRPIN_TH <- multi.str.split(primer3.reverse.output[↩
            grep("PRIMER_RIGHT_[0-9]*_HAIRPIN_TH", primer3.reverse.output)],↩
             "=", 2)
269     PRIMER_RIGHT_X_END_STABILITY <- multi.str.split(↩
            primer3.reverse.output[grep("PRIMER_RIGHT_[0-9]*_END_STABILITY",↩
             primer3.reverse.output)], "=", 2)
270
271     right.parsed <- data.frame("name1", pos4, PRIMER_RIGHT_X_PENALTY, ↩
            PRIMER_RIGHT_X_SEQUENCE, PRIMER_RIGHT_X, PRIMER_RIGHT_X_TM, ↩
            PRIMER_RIGHT_X_GC_PERCENT, PRIMER_RIGHT_X_SELF_ANY_TH, ↩
            PRIMER_RIGHT_X_SELF_END_TH, PRIMER_RIGHT_X_HAIRPIN_TH, ↩
            PRIMER_RIGHT_X_END_STABILITY)
272     colnames(right.parsed)[1:2] <- c("p.name", "pos")
273     right.parsed <- right.parsed[sort(right.parsed$pos, index.return = T↩
            )$ix, ]
274     right.parsed <- right.parsed[which(right.parsed$pos %in% ↩
            snp.coords.after.filter), ]
275     nrow(right.parsed)
276     right.parsed2 <- split(right.parsed, right.parsed$pos)
277     right.parsed2 <- lapply(right.parsed2, function(x) {
278       x[which.min(as.numeric(x$PRIMER_RIGHT_X_PENALTY)), ]
279     })
280
281     right.parsed3 <- bind_rows(right.parsed2)
282
```

```
283     right.parsed3$MSA.pos <- which(homologous.snps == 1)[which(↩
            snp.coords.after.filter %in% right.parsed3$pos)]
284     right.parsed3 <- right.parsed3[, c(1, 2, 12, 3:11)]
285     colnames(right.parsed3)[4] <- "pen"
286     colnames(right.parsed3)[5:ncol(right.parsed3)] <- gsub("X", "0", ↩
            colnames(right.parsed3)[5:ncol(right.parsed3)])
287
288     ########################################
289     if (!file.exists(p(project.path, "jobs/", gene.name, "/primers/↩
            penalties"))) dir.create(p(project.path, "jobs/", gene.name, "/↩
            primers/penalties"))
290     write.csv(left.parsed3, p(project.path, "jobs/", gene.name, "/↩
            primers/penalties/forward.all.pen.csv"), row.names = F)
291     write.csv(right.parsed3, p(project.path, "jobs/", gene.name, "/↩
            primers/penalties/reverse.all.pen.csv"), row.names = F)
292
293     return(list(left.parsed3, right.parsed3))
294   }
295
296   generate.best.primer.set <- function(forward.all.pen, reverse.all.pen,↩
          forward.coord.used, reverse.coord.used, iteration) {
297     if (missing(forward.coord.used)) forward.coord.used <- as.numeric()
298     if (missing(reverse.coord.used)) reverse.coord.used <- as.numeric()
299     if (missing(iteration)) iteration <- 1
300
301     if (length(forward.coord.used) > 0) forward.all.pen <- ↩
            forward.all.pen[-which(forward.all.pen$pos %in% ↩
            forward.coord.used), ]
302     if (length(reverse.coord.used) > 0) reverse.all.pen <- ↩
            reverse.all.pen[-which(reverse.all.pen$pos %in% ↩
            reverse.coord.used), ]
303
304     best.primer.file.end.coord <- as.numeric(start.coord.after.filter)
305
306     forward.primer.positions <- list()
307     reverse.primer.positions <- list()
```

```
308
309     minimum.snp.coord <- 0
310     maximum.snp.coord <- start.coord.after.filter
311
312     # run loop to get pairs of primers
313
314     while (best.primer.file.end.coord < end.coord.after.filter) {
315       valid.fwd.coords <- which(forward.all.pen$pos < (maximum.snp.coord↩
              - 1) & forward.all.pen$pos > minimum.snp.coord)
316
317       while (length(valid.fwd.coords) == 0) {
318         maximum.snp.coord <- maximum.snp.coord + 10
319         valid.fwd.coords <- which(forward.all.pen$pos < (↩
              maximum.snp.coord - 1) & forward.all.pen$pos > ↩
              minimum.snp.coord)
320         print("No SNPs found for forward primer, expanding start buffer"↩
              )
321
322         # add stop condition
323         if (maximum.snp.coord > max(coords$snp.coords)) {
324           if (iteration == 1) {
325             write("No valid forward primer coordinates", p("jobs/", ↩
                  gene.name, "/primers/error.txt"))
326             return(as.numeric())
327           } else {
328             return(as.numeric())
329           }
330         }
331       }
332
333       f.primer.candidates <- forward.all.pen[which(forward.all.pen$pos <↩
              (maximum.snp.coord - 1) & forward.all.pen$pos > ↩
             minimum.snp.coord), ]
334       f.primer.candidates <- f.primer.candidates[which(↩
             f.primer.candidates$pen == min(f.primer.candidates$pen)), ]
335
```

```
336       # debugging
337       valid.rev.coords <- which(reverse.all.pen$pos > (↩
              f.primer.candidates$pos + product.size.range[1]) & ↩
              reverse.all.pen$pos > best.primer.file.end.coord & ↩
              reverse.all.pen$pos < (f.primer.candidates$pos + ↩
              product.size.range[2]))
338
339       while (length(valid.rev.coords) == 0) {
340         product.size.range[2] <- product.size.range[2] + 10
341         valid.rev.coords <- which(reverse.all.pen$pos > (↩
                f.primer.candidates$pos + product.size.range[1]) & ↩
                reverse.all.pen$pos > best.primer.file.end.coord & ↩
                reverse.all.pen$pos < (f.primer.candidates$pos + ↩
                product.size.range[2]))
342         print("No SNPs found for reverse primer, expanding maximum ↩
                product size")
343
344         # add stop conditions
345         if (product.size.range[2] > length(template.sequence)) {
346           if (iteration == 1) {
347             write("No valid reverse primer coordinates", p("jobs/", ↩
                    gene.name, "/primers/error.txt"))
348             return(as.numeric())
349           } else {
350             return(as.numeric())
351           }
352         }
353       }
354
355       r.primer.candidates <- reverse.all.pen[which(reverse.all.pen$pos >↩
              (f.primer.candidates$pos + product.size.range[1]) & ↩
              reverse.all.pen$pos > best.primer.file.end.coord & ↩
              reverse.all.pen$pos < (f.primer.candidates$pos + ↩
              product.size.range[2])), ]
356       r.primer.candidates <- r.primer.candidates[which(↩
              r.primer.candidates$pen == min(r.primer.candidates$pen)), ]
```

```
357
358       forward.primer.positions <- c(forward.primer.positions, list(↩
              f.primer.candidates))
359       reverse.primer.positions <- c(reverse.primer.positions, list(↩
              r.primer.candidates))
360
361       best.primer.file.start.coord <- as.numeric(f.primer.candidates$pos↩
              )
362       best.primer.file.end.coord <- as.numeric(r.primer.candidates$pos)
363
364       minimum.snp.coord <- best.primer.file.start.coord
365       maximum.snp.coord <- best.primer.file.end.coord # The name of this↩
               variable originates from the first iteration. Of cause this ↩
              is not truly the start.coord.after.filter on subsequent ↩
              iterations
366     }
367
368     library(dplyr)
369
370     forward.primer.positions <- lapply(forward.primer.positions, ↩
          function(x) x[1, ])
371     forward.primer.positions <- bind_rows(forward.primer.positions)
372     forward.primer.positions$orient <- "F"
373
374     reverse.primer.positions <- lapply(reverse.primer.positions, ↩
          function(x) x[1, ])
375     reverse.primer.positions <- bind_rows(reverse.primer.positions)
376     reverse.primer.positions$orient <- "R"
377
378     p3.input.files.to.rm <- list.files(p(project.path, "jobs/", ↩
          gene.name, "/primers/input/"), full.name = T)
379     p3.output.files.to.rm <- list.files(p(project.path, "jobs/", ↩
          gene.name, "/primers/output/"), full.name = T)
380
381     lapply(p3.input.files.to.rm, file.remove)
382     lapply(p3.output.files.to.rm, file.remove)
```

```
383
384     Map(function(f.primer1, r.primer1) {
385       generate.primer3.input.files(template.sequence2, p((f.primer1 + 1)↩
            , "-", (r.primer1 + 1)), 100, 750, f.primer1, r.primer1, "B")
386     }, forward.primer.positions$pos, reverse.primer.positions$pos)
387
388     system(p(project.path, "jobs/", gene.name, "/primers/run.primer3.sh ↩
          ", "jobs/", gene.name))
389
390     output.files <- list.files(p(project.path, "jobs/", gene.name, "/↩
          primers/output/"))
391
392     output.files <- output.files[sort(as.numeric(multi.str.split(↩
          multi.str.split(output.files, "-", 1), "\\.", 2)), index.return ↩
          = T)$ix]
393     output.files.numbered <- paste0(1:length(output.files), output.files↩
          )
394
395     list.best.primer.start.coords <- c(list.best.primer.start.coords, ↩
          forward.primer.positions$pos)
396     list.best.primer.end.coords <- c(list.best.primer.end.coords, ↩
          reverse.primer.positions$pos)
397
398     # see if any sets have already made, if so make new set directory
399     i <- max(as.numeric(multi.str.split(list.files(p(project.path, "jobs↩
          /", gene.name, "/primers/best.primers/")), "set", 2))) + 1
400     if (is.na(i) | i == -Inf) i <- 1
401
402     if (!dir.exists(p(project.path, "jobs/", gene.name, "/primers/↩
          best.primers/set", i))) {
403       dir.create(p(project.path, "jobs/", gene.name, "/primers/↩
            best.primers/set", i))
404     }
405
406     Map(function(x, best.primer.file, fpos1, rpos1) {
407       file.copy(p(project.path, "jobs/", gene.name, "/primers/output/", ↩
```

```
            x), p(project.path, "jobs/", gene.name, "/primers/best.primers/set", i, "/", best.primer.file))
408       multiple.alignment.coord1 <- which(homologous.snps == 1)[which(fpos1$pos == coords$snp.coords)]
409       multiple.alignment.coord2 <- which(homologous.snps == 1)[which(rpos1$pos == coords$snp.coords)]
410       # add multiple sequence alignment coordinates to the best primer3 output file
411       system(p("echo multiple.alignment.forward.coord=", multiple.alignment.coord1, " >> ", project.path, "jobs/", gene.name, "/primers/best.primers/set", i, "/", best.primer.file))
412       system(p("echo multiple.alignment.reverse.coord=", multiple.alignment.coord2, " >> ", project.path, "jobs/", gene.name, "/primers/best.primers/set", i, "/", best.primer.file))
413       system(p("echo SEQUENCE_TEMPLATE_REV_COMP=", as.character(reverseComplement(template.sequence2)), " >> ", project.path, "jobs/", gene.name, "/primers/best.primers/set", i, "/", best.primer.file))
414       system(p("echo gene.start.coord=", start.coord.after.filter, " >> ", project.path, "jobs/", gene.name, "/primers/best.primers/set", i, "/", best.primer.file))
415       system(p("echo gene.end.coord=", end.coord.after.filter, " >> ", project.path, "jobs/", gene.name, "/primers/best.primers/set", i, "/", best.primer.file))
416     }, output.files, output.files.numbered, split(forward.primer.positions, forward.primer.positions$pos), split(reverse.primer.positions, reverse.primer.positions$pos))
417
418     # convert primer3 coordinates (coordinates from the sequence without
419     # insertions) to coordinates in the multiple sequence alignment
420     return(list(list.best.primer.start.coords, list.best.primer.end.coords))
421   }
422
```

```
423   print("Performing best primer selection")
424   # if(1 == 2){ #toggle to skip caching
425   if (file.exists(p(project.path, "jobs/", gene.name, "/primers/↩
          penalties/forward.all.pen.csv")) & file.exists(p(project.path, "↩
          jobs/", gene.name, "/primers/penalties/reverse.all.pen.csv"))) {
426     forward.all.pen <- read.csv(p(project.path, "jobs/", gene.name, "/↩
            primers/penalties/forward.all.pen.csv"), stringsAsFactors = F, ↩
            header = T)
427     forward.all.pen <- forward.all.pen[, 1:4]
428     reverse.all.pen <- read.csv(p(project.path, "jobs/", gene.name, "/↩
            primers/penalties/reverse.all.pen.csv"), stringsAsFactors = F, ↩
            header = T)
429     reverse.all.pen <- reverse.all.pen[, 1:4]
430   } else {
431     penalties1 <- generate.all.primer.penalites(1)
432     forward.all.pen <- penalties1[[1]][, 1:4]
433     reverse.all.pen <- penalties1[[2]][, 1:4]
434   }
435   used.coords1 <- generate.best.primer.set(forward.all.pen, ↩
          reverse.all.pen)
436   if (length(used.coords1) > 0) generate.best.primer.set(↩
          forward.all.pen, reverse.all.pen, used.coords1[[1]], used.coords1↩
          [[2]], 2)
437 }
```

8 Bibliography