University of Oxford

Genetic and genomic analysis of *Arabidopsis thaliana* with low-coverage next-generation sequencing data

Martha Imprialou, Balliol College

Supervisors: Professor Richard Mott and Professor Jotun Hein

DPhil thesis

Department of Statistics, University of Oxford

April 2015

This is my own work (except where otherwise indicated)

Candidate: Martha Imprialou

Signed: Martha Imprialou

Abstract

Next-generation sequencing technologies have transformed our understanding of genetic variation segregating in populations and its relationship with phenotypic traits. Sequencing large populations at low coverage, thus sampling only a fraction of the genome of each individual, may increase statistical power in genetic mapping [99] compared to genotyping arrays. This thesis explores several novel applications of low-coverage population-based sequencing, using data from 488 recombinant inbred lines from the MAGIC population of *Arabidopsis thaliana*, descended from 19 inbred founder accessions. Based on the full catalogue of genetic variation that is available in the 19 founders [36], I describe every MAGIC genome as a mosaic of founder haplotypes and analyse the accuracy of the mosaics by simulation. I then use the mosaics in three ways. First, I investigate structural variation using a novel method that treats anomalies in the alignment of sequencing reads, potentially representing signatures of structural variants (SVs), as quantitative traits. These can be mapped genetically to identify loci in which genetic variation correlates with signatures of SVs. The method can distinguish short- (e.g. indels) and long-range (e.g. translocations) SVs and has led to the discovery of a large number of SVs segregating in the MAGIC population, including thousands of long-range SVs. I show that SVs have a significant impact on silencing gene expression and that they explain a large fraction of the phenotypic variation in several physiological traits. Second, I use the mosaic structure of the MAGIC lines to map recombination events and analyse lineage-specific recombination in MAGIC. I

infer recombination hotspots and compared recombination in the MAGIC lines to the *Arabidopsis* genetic map. Finally, I detect bacterial endosymbionts hosted in MAGIC genomes from unmapped reads that have high sequence similarity with bacterial DNA and examine whether variation in the presence of endosymbionts can be explained by host genetic variation.

Acknowledgements

The time I spent at the University of Oxford working on this thesis has been one of the most fulfilling experiences of my life. This I owe, first and foremost, to my supervisor, Richard Mott, whom I want to thank for his encouragement, guidance and support. Richard is one of the brightest people I know and has contributed a lot of his knowledge, time and ideas to this thesis. His genuine interest and commitment to science motivated and inspired me, even during tough times, and have contributed to my decision to pursue a career in research.

I am grateful to Jotun Hein, who co-supervised my DPhil with Richard, for the advice and ideas he has contributed to this thesis and the funding opportunities he has offered me. Jotun also gave me many opportunities to teach and supervise students, which was an invaluable experience that I thoroughly enjoyed. I wish to thank Miltos Tsiantis for the collaboration in this project and for providing his valuable scientific knowledge on plant biology. I thank Jonathan Flint for his advice, support and his sense of humour.

Several other scientists have contributed to this work, either by performing experiments, providing data or offering insightful comments. In particular I must acknowledge Amarjit Bhomra, from WTCHG, Xiangchao Gan and Janne Lempe (Max Planck Institute of Plant Breeding Research, Cologne), Paula Kover (University of Bath), Oliver Stegle (European Bioinformatics Institute), Gunnar Raetsch and Andre Kahles (Memorial Sloan-Kettering Cancer Center), EJ Osborne and Richard Clark (University of Utah), Magnus Nordborg (Gregor Mendel Institute) and Ian Henderson (University of Cambridge).

For providing a great space for learning and scientific development I thank the Department of Statistics in Oxford, that also funded me, and the Wellcome Trust Centre for Human Genetics which offered me a space to work all these years (being, to the best of my knowledge, the only person working on plant genetics in the building). I want to thank my colleagues from the WTCHG, all the members of the Mott, and Flint labs, and specifically my friends from the centre Na, Sanja, Carme, Amelie, Loukas and Juan.

I would not have completed this work without the support and friendship of all the great people I have met in and around Oxford. I want to thank the extraordinary community of the Balliol

College MCR in which I have met some of my best friends, Gabi, Andy, Seb, Sara, Fiona, Hilary, Rahul and Elisabetta, and my dancing group, Oxford Lindy Hoppers, with whom I had some of the most amusing days (and nights) in Oxford. Hard times require furious dancing.

I want to thank my partner, Mike, for his support and encouragement as I was writing up this thesis.

Finally I want to thank my parents and my sister Marianna, for helping me all along the way in my studies and for their love and support in all my endeavours.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Genetic variation explains much of the observed phenotypic variation within a population and is the material from which natural selection drives evolution. Its main generators are mutation and recombination. Mutation randomly alters genomic sequence, either by substituting a single nucleotide, giving rise to single nucleotide polymorphisms (SNPs) or by altering chromosomal structure, giving rise to structural variants (SVs). Recombination generates new combinations of existing alleles, allowing two different DNA molecules to exchange genetic material and form a new one. Transposable elements, which are genic features that change their genomic position, are another source of genetic variation, affecting both genomic size and structure.

Cataloguing the genetic variation segregating in a population is important and useful. For example, associations between DNA variants and phenotypic traits may be used to discover genes that contribute to a phenotype. Genotype-phenotype associations can be detected by genome-wide association studies (GWAS), which examine the relationships of many genetic variants with a phenotypic trait. GWAS commonly use SNP arrays of millions of SNPs which genotype every individual in the population.

With the advent of next-generation (high-throughput) sequencing technologies it has become possible to concurrently sequence thousands of genomes at high precision. This has allowed sequencing of large populations and has uncovered many novel sequence variants. However, our knowledge of genomic polymorphism is incomplete because of variability in the size, copy-number

and order of loci within populations [17]; this variability is usually caused by structural variants, but centromeres, heterochromatic regions and transposable elements also contribute. Assembly of sequencing reads over such regions is challenging - for example, if a genome carries multiple copies of a locus, it may be difficult to assign reads to their correct place unambiguously. Sequencing at high coverage i.e. obtaining a large number of reads so that each locus is covered by multiple reads can help resolve some, but not all, ambiguities, as the read lengths are relatively small (typically between 35 and 150bp) and read-mapping algorithms are imperfect. Furthermore, statistical power is key to most GWAS, as complex traits are often functions of many common polymorphisms with small effects which can be detected only with very large population sizes. However, sequencing all individuals at high coverage for such a study can be expensive and impractical.

As an alternative, techniques that do not sequence the entire genomes of individuals have emerged. Restriction-site associated DNA marker (RAD) sequencing for example, simultaneously sequences at high coverage target loci in multiple individuals. RAD sequencing is cost-effective and does not require a reference genome to align reads, but it is not applicable over highly diverged loci. Furthermore, target sequence information is not always available and can be hard to collect, so RAD-seq data may miss important genomic loci. Another approach is low-coverage whole genome sequencing, which is priced comparatively to a SNP array study, but has been shown to produce superior results, as it can increase statistical power and reduce false positive rates [99]. Depending on the population, the depth of coverage required varies, but is typically in the range $c = 0.1 - 2$x i.e. each position in each individual is covered by an average number of c reads, although coverage will vary randomly as well as for systematic reasons. Consequently, only a fraction of the genome is observed in each individual. Missing genotypes can be inferred by comparing the data with known haplotypes, a procedure known as genomic imputation [87]. With the advancement of sequencing technologies the cost of sequencing is continuously dropping in comparison to SNP arrays, so the paradigm of low-coverage sequencing may soon allow the sequencing of very large populations, potentially increasing statistical power for GWAS [34].

1.1 Thesis description

This thesis explores genetic variation in a population of the plant *Arabidopsis thaliana* and proposes novel uses of low-coverage sequencing in genomic analysis. Several aspects of genetic variation analysis using low coverage sequencing data have been considered. The main focus is the identification and analysis of structural variants, particularly long-range, which are harder to detect. Genomic imputation from low-coverage sequence is also discussed. Recombination maps of *A. thaliana* using imputed data are also analysed and compared to the *Arabidopsis* genetic map. Finally, a method for mapping the differential presence of cohabitating bacteria and viruses in a population is introduced.

1.1.1 Structural genetic variation

Structural variants (SVs) alter the structure of chromosomes by deleting, duplicating, transposing or reversing the orientation of DNA sequence (Figure 1.1). Their size is highly variable, ranging from a few bases to entire chromosomes. They may only affect the local sequence (short-range structural variation) or juxtapose remote loci (long-range). Since SVs can disrupt individual genes or groups of genes, they can have significant phenotypic effects [53, 25, 101, 103, 77, 118].

Early studies detected rare and very large ($>$ 3Mb) SVs by microscopy [53, 25], which however accounted for a very small fraction of the total structural variation. More recently, SVs were detected using the technology of array comparative genome hybridisation (aCGH) which used arrays of long oligonucleotides enabling the detection of common SVs in populations, down to a resolution of about 100kb. However, aCGH-based studies were highly inconsistent with each other as results were highly dependent on the array used. This resulted in large discrepancies between studies and to low replication rates with different arrays [1, 111]. The development of next-generation sequencing addressed many of these issues and rendered much more accurate SV predictions, thus allowing one to estimate the extent of structurally variant genomic loci. In humans, several studies on SVs using next-generation sequencing [3, 93, 134] have predicted different sets of SVs, while the total number of distinct SVs recorded in the Database of Genomic Variants Archive from all submitted studies includes over $200,000$ short-range SVs in the human genome [85]. In the mouse, a single study predicted $700,000$ [137] short-range SVs in thirteen classical and four wild-type strains, while

in *Arabidopsis thaliana* over 175,000 have been identified in 80 geographically diverse strains [17].

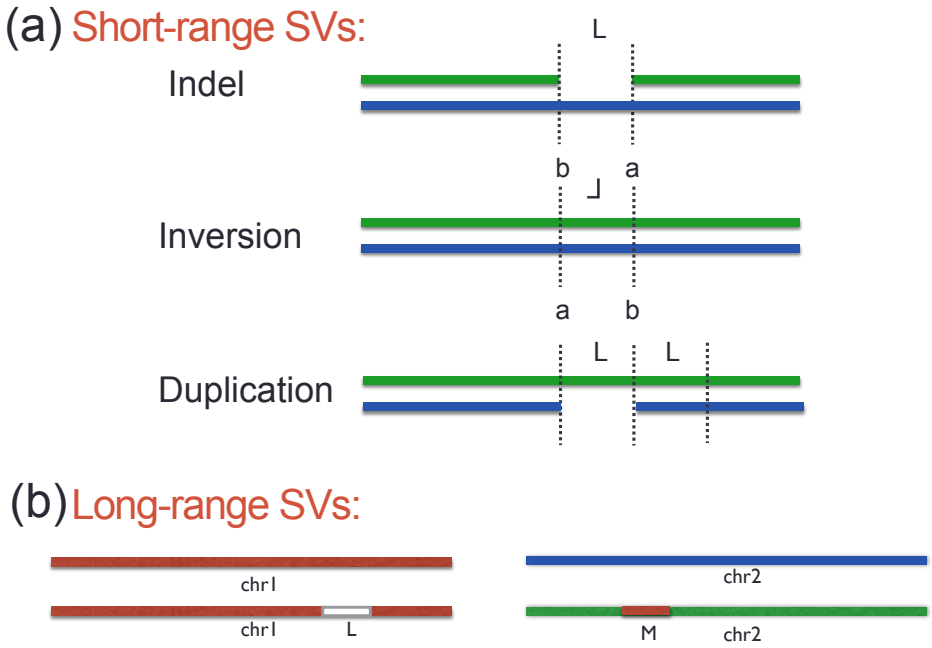


Figure 1.1: Sketch of structural variants (SVs). (a): Examples of short-range structural variants. The blue line represents a reference genome without an SV and the green line a genome carrying an SV at locus L. Locus L can be deleted, inverted or duplicated with respect to the reference. (b): An example of an interchromosomal translocation, in which locus L from chromosome 1 is moved to locus M of chromosome 2. Thus, there is a deletion in chromosome 1 and a symmetric insertion in chromosome 2 in one of the genomes. Again, the reference genome is coloured in blue, while the genome carrying the translocation is coloured in green. Chromosome 1 is coloured in red in both genomes, as we can assume without loss of generality that it is genetically identical in both genomes.

However, the extent of long-range structural variation segregating in natural populations is unclear; apart from certain long-range structural variants that are large enough to be detected by microscopy, or copy-number variants detected by comparative genome hybridisation (CGH), we do not yet know the frequency of small-scale long-range variants (of the order of $1-100$ kilobases (kb) in length), or what their phenotypic impact is. Nonetheless it is plausible that they would affect the chromatin landscape, the gene expression and complex traits.

Short-range structural variants are currently detected by anomalies in the alignment of reads to a reference genome: for example, gaps in the alignment of reads are associated with deletions and increases in read coverage with copy-number gains. This approach cannot be perfect even

for short-range SVs: indeed, computational methods for the prediction of SVs are known to have high false negative rates [49, 133], which partly explains the large discrepancies between studies. One reason is that structurally variant loci are usually surrounded by sequence features that are difficult to align to the reference genome, such as transposable elements or repetitive sequences [61] while others are complex, combining different signatures [137]. Moreover, structural variants in a population may be diverged causing different signatures of read anomalies at each individual which are harder to resolve. In the case of long-range rearrangements such problems are more prominent, thus it is even harder to distinguish a real event from sequencing errors. For example, to search for an intrachromosomal translocation breakpoint, in theory all that is needed is a read-pair that spans the breakpoint; one read will map to one locus and the other to a different chromosome. In practice, the great majority of these events will be false positives. Consequently we require additional means to resolve true variants from noise.

There are alternative experimental methods for identifying long-range structural variants in individuals. In addition to microscopy, Optical Mapping [97] provides approximate long-range information based on restriction sites but is not yet a high-throughput method. Standard short-read sequencing of individuals at very high coverage and with libraries with a range of insert sizes makes de-novo assembly more tractable, although it is not yet possible to assemble large chromosomes completely. The Illumina Moleculo technology simulates long-read sequencing by a form of multiplexed sequencing of barcoded localised reads. Over the next few years longer-read sequencing technologies (eg PAC-BIO [19], Oxford Nanopore MINION [119]) are likely going to establish themselves, and will help with the identification of structural variants. However, the timescale before they become mature, reliable, cost-effective and applicable to population-scale surveys is uncertain. It would also be preferable on grounds of cost to avoid the wholesale resequencing of previously-sequenced genomes if possible. More importantly, if our goal is to understand complex traits, we require methods that identify common long-range structural variants that segregate in populations, rather than being unique to an individual.

In this thesis I present a novel computational method which can identify both short and long-range SVs segregating in populations, and can be applied to low-coverage data. It differs from

classical methods in that it combines signatures of SVs in sequencing data with the genetic background of the local sequence. Effectively, the method transforms the problem of SVs detection into a genetic mapping one, thus instead of trying to predict SVs *de novo* in each genome separately, it uses information from the entire population. Therefore it can identify inherited SVs but not rare somatically acquired ones which are unique to each individual. The method is not specific to *Arabidopsis* and could be extended to any species and any population.

I have used this approach to generate a catalogue of structural variants in 488 recombinant inbred lines of the *Arabidopsis thaliana* MAGIC population [67]. These lines are descended from 19 inbred founder accessions, which have been sequenced and annotated at high precision [36]. Each MAGIC genome is an (almost) homozygous mosaic of the 19 founders, so the haplotype space at each locus is identical to the space of founder haplotypes. I determined these mosaics after obtaining low-coverage Illumina sequence of the MAGIC lines and using Hidden Markov Models. The low-coverage sequence data and the genetic mosaics have been then reused to find $4,898$ short-range and $1,604$ long-range SVs in the MAGIC lines. The expression of genes within these SVs is dysregulated and some SVs account for complex trait heritability missed by orthodox genetic variation.

1.1.2 Recombination

The second part of the thesis deals with recombination, making use of the population history of the MAGIC lines. Most recombination events are concentrated in recombination hotspots [96], but hotspot location varies greatly within subpopulations [48]. Studying variation in recombination hotspots may therefore explain some of the differences between populations. My approach is to study lineage-specific recombination in order to observe potential differences in hotspot formation. The MAGIC population design is well-suited for this purpose, as genome mosaics partially record the recombination history of each lineage, and recombinants can be mapped at high precision. I use the mosaics to compare the recombination landscape in MAGIC to the *A. thaliana* genetic map, determined from natural *A. thaliana* accessions. I report large genomic regions in which recombinants appeared to cluster - these were different in appearance from ordinary hotspots in size

and in that they were specific to individual lineages. With extensive quality control, I have shown that a large fraction of these were artifactual, caused predominantly by residual heterozygosity, but also by unmapped long-range SVs. A smaller fraction of those may be genuine although further data collection is required for a definitive answer.

1.1.3 Detection of endosymbionts from unmapped reads

The third part of the thesis studies variation within endosymbiont organisms living on tissues or inside the cells of *A. thaliana*. I show that many unmapped reads from sequencing data derive from the genome of bacterial or viral endosymbionts. There is differential enrichment of certain endosymbionts across lines and I explore the impact of genetics on this. I have mapped the number of unmapped reads corresponding to endosymbionts as a quantitative trait and report marginally significant associations, even with low coverage data. Interestingly, several traits are associated with the same locus, indicating that the same genes may control the enrichment of certain types of endosymbionts. I also show that low coverage sequencing could be sufficient for such an analysis, as on many occasions the data recapitulate the variance that would have been observed in high-coverage data.

1.2 Statement of collaboration

Parts of this work were completed in collaboration with colleagues from the Wellcome Trust Centre for Human Genetics and other institutions. In this report I have included only my personal contribution to these projects, unless otherwise mentioned. Below I present specific aspects of this project which were completed by or in collaboration with others.

Genome assembly of the 19 MAGIC founders was performed by Dr Xiangchao Gan, currently based at the Max Planck Institute for Plant Breeding Research, Cologne, using his IMR/DENOM pipeline [36]. MAGIC lines were grown in Bath (laboratory of Dr Paula Kover) and Oxford (laboratory of Professor Nick Hardberd) and were sequenced by the Oxford Genomics Centre.

In Chapter 3 the mosaic reconstruction algorithm based on Viterbi path was originally designed and implemented by Professor Richard Mott while I have extended the algorithm for heterozygous

genomes and implemented the Forward-Backward reconstruction algorithm.

In Chapter 4, gene expression data for 200 MAGIC lines were provided by Dr Richard Clark, Utah USA and Dr Gunnar Raetsch, New York USA. Furthermore, experimental confirmation of long-range SVs by PCR was completed by Amarjit Bhomra, WTCHG and Dr Janne Lempe, Max Planck Institute for Plant Breeding Research, Cologne.

Chapter 2

The MAGIC Arabidopsis Thaliana population

2.1 *Arabidopsis thaliana*

Arabidopsis thaliana is a small flowering plant that has been extensively used as model organism in plant biology and genetics [5]. Being native in almost all Eurasia, it can be found in a wide range of natural habitats, displaying great genetic and phenotypic variation. *A. thaliana* has a small genome, comprising ∼120Mb in 5 nuclear chromosomes encoding $\sim 33,000$ genes, and a short life cycle of about 6 weeks, allowing the growth and maintenance of large population stocks. Its ability to form naturally inbred strains by self-pollination, called accessions, enables the construction of stable inbred populations. Extensive genetic and phenotypic maps of the plant are available (`http://www.arabidopsis.org`), providing excellent datasets for research.

Figure 2.1: Variation in morphology and leaf-shape of *Arabidopsis Thaliana* wild-type strains.

2.2 Populations descended from inbred founders

Current studies for mapping quantitative trait loci (QTL) use either synthetic or naturally occurring populations. Synthetic biparental populations such as F_2, backcrosses and advanced intercrosses (AILs) [28] are created by crossing two individuals that differ significantly for a specific quantitative trait. This design offers high statistical power: all QTLs are biallelic so the power to detect them depends on the variance explained, which is proportional to $p(1-p)$, with p the allele frequency. This is maximised in F_2-type crosses, where allele frequencies are close to $p = 0.5$. However, each individual in a typical F_2 will has only one recombinant per chromosome so the mapping resolution

is poor. Naturally occurring populations, on the other hand, have the opposite properties: they offer fine resolution due to huge numbers of accumulated recombinants, but can suffer from low power as minor alleles may have extremely low frequency. Moreover, natural population studies are confounded by extensive population structure, which inflates false positive rates [6].

A good compromise between these extremes can be achieved with synthetic mosaic populations descended from multiple ancestors [13]. These are characterised by higher genetic diversity than F_2 and by a higher number of visible recombinants per genome. Additionally, the controlled set of founder strains sets a lower bound to allele frequencies, offering intermediate power between biparental and natural populations (Figure 2.2). Examples of mosaic population designs include:

- the Collaborative Cross (CC) mice [127, 52], in which every animal originates from an 8-way breeding funnel. At the start of the breeding eight founder strains are crossed to form four F_2, subsequently two F_4, and finally one F_8 cross, in which all founder haplotypes are represented. The end genomes are maintained by inbreeding, so their chromosomes are fixed.

- Heterogeneous Stocks (HS) [128] also originate from an 8-way funnel breeding. During maintenance HS genomes cross with each other for about 50 generations to build up recombinants and produce fine-grained mosaics. HS populations are currently available in mice [127] and rats [2, 27].

- the Diversity Outcross (DO) mice [120] in which the CC genomes are maintained by outcrossing with wild-type mice. The resulting population preserves the segregating alleles of the collaborative cross, but with high heterozygosity levels, offering higher genetic and phenotypic variation.

Regardless of population design strategy, all individuals in a mosaic population comprise genomes that are mosaics of the founder genomes (Figure 2.3).

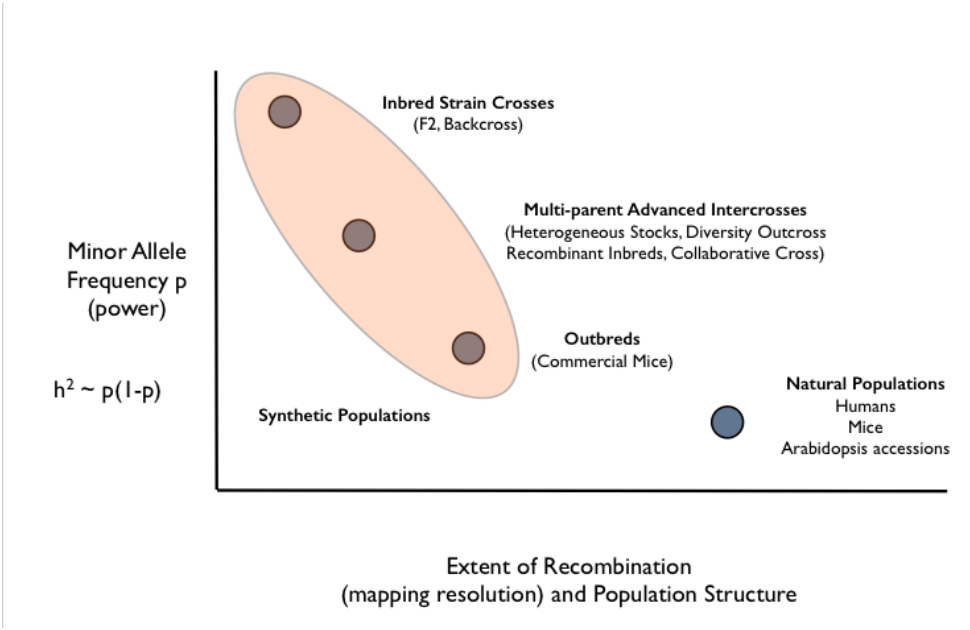


Figure 2.2: Sketch comparing different types of synthetic populations with natural populations with respect to statistical power and mapping resolution.

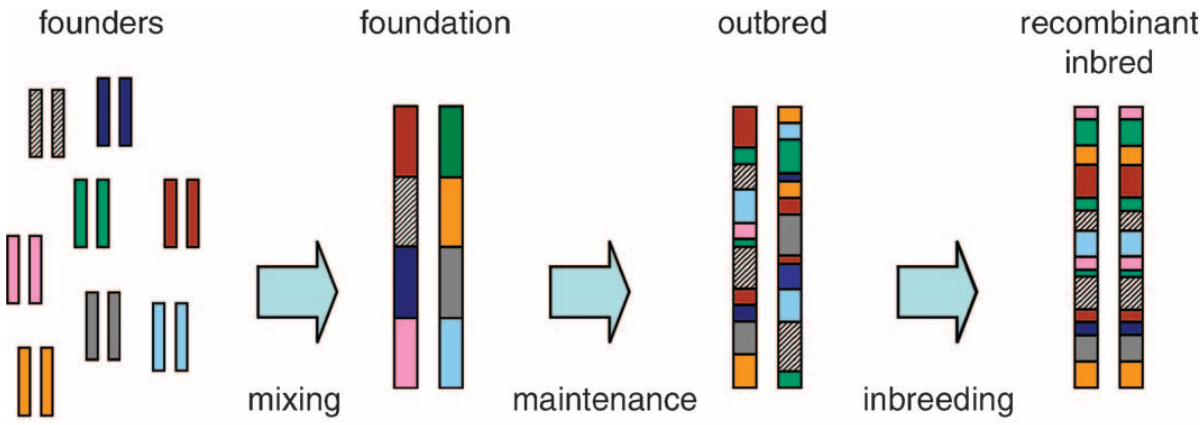


Figure 2.3: Stages of a RIL breeding programme. In the first stage, several homozygous founder strains are selected and are subsequently mixed into genomes with multiple founder haplotypes. In the maintenance stage, further intercrossing increases the number of recombinants per genome. Finally, several generations of inbreeding results in a population of homozygotes (figure taken from [127])

.

2.2.1 The MAGIC population of RILs in *A. thaliana*

The Multiparent Advanced Generation Inter-Cross (MAGIC) population of *A. thaliana* [67], consists of about 700 recombinant inbred lines (RILs) descended from 19 founder accessions. The 19 founders include commonly used accessions such as the *A. thaliana* reference genome Col-0 (TAIR10), as well as geographically diverse accessions, to achieve high genetic diversity (Table 2.1). There are 342 independent MAGIC lineages; each descended from different combinations of eight distinct F_2 plants sampled from a full F_1 diallele cross of the 19 founders. Each MAGIC line is descended from two further generations of outcrossing (F_3 to F_4), as shown in Figure 2.4. Up to 3 F_4 lines from each family were selected and selfed for 6 generations to render homozygous genomes. Lines from the same family are referred to as "cousins" and they are expected to share about 25% of their genomes. However, this does not introduce population structure; we have estimated the mean haplotypic overlap for two arbitrary MAGIC lines to be 5.7%.

Compared to the general mosaic scheme, shown in Figure 2.3, the MAGIC population design includes only the mixing and inbreeding stages, but not the maintenance (outbreeding) stage. Figure 2.4 illustrates the breeding strategy for MAGIC. On average 11 founders are represented in the genome of each line.

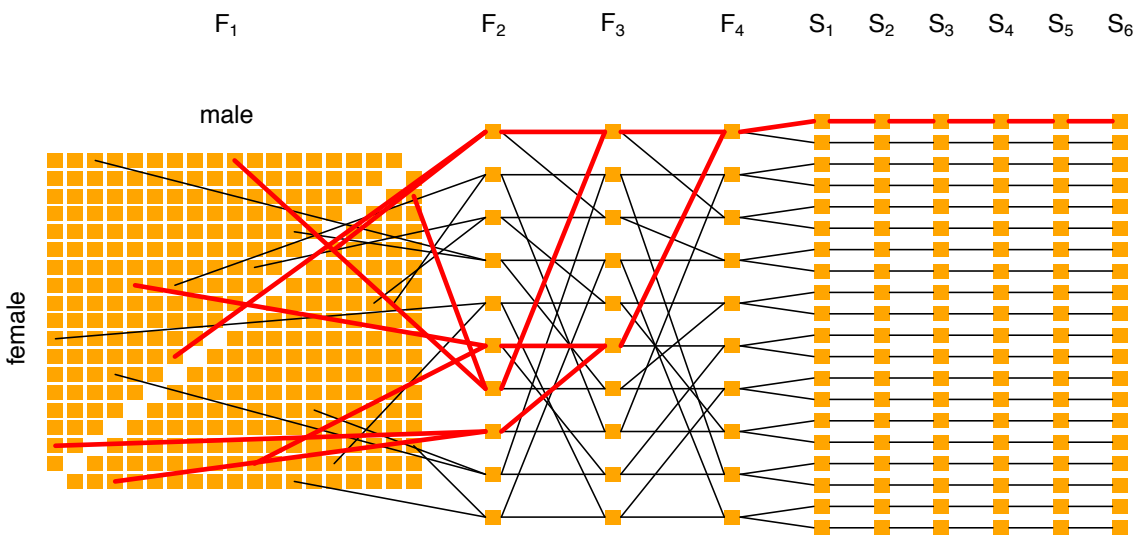


Figure 2.4: The MAGIC breeding scheme. The left-hand 19x19 matrix represents the F_1 diallele cross between the 19 MAGIC founder accessions; each orange square represents one of $19 \times 18 = 342$ F_1 plants, taking into account both cross directions. The columns labeled F_2, F_3, F_4 show 10 examples from the 2nd, 3rd and 4th generations of crossing (in reality there are 342 individuals in each generation). The F_2 generation is formed by crossing pairs of F_1 plants, each of which is used at least once as a parent to one of the 342 in the full design. The parents of each exemplar plant are indicated by the black lines connecting two plants from generation F_N with one plant in $F_{(N+1)}$. Parents of each $F_{(N+1)}$ plant are selected at random from generation F_N. At the 4th generation up to three seeds are taken from each F_4 plant to initiate cousin MAGIC lines, which are selfed for six generations ($S_1 - S_6$). The ancestral lineage of the top-most MAGIC line is traced back using the thicker, red lines, showing its breeding funnel.

The MAGIC lines were genotyped at 1260 SNPs and used for QTL mapping [67]. With respect to population structure, this study showed that on average the SNP sharing between founders, and hence between MAGIC lines, is about 70%, but haplotype sharing is much lower, estimated at about 7% (with the exception of cousins). Linkage disequilibrium (LD) between SNPs decays within 10kb in the founders, while there is no LD between remote loci. Therefore, the genome mosaics were inferred probabilistically using Hidden Markov Models without ambiguity (Figure 2.5). Furthermore, the population was used to map QTLs, using both binary and continuous traits. The binary traits tested were presence of the ERECTA mutation (affecting leaf morphology and shape) and glabrousness, both of which are known to be private to a single accession, Ler-0 and Ws-0 respectively. Flowering time was used as the representative complex trait, with the gene

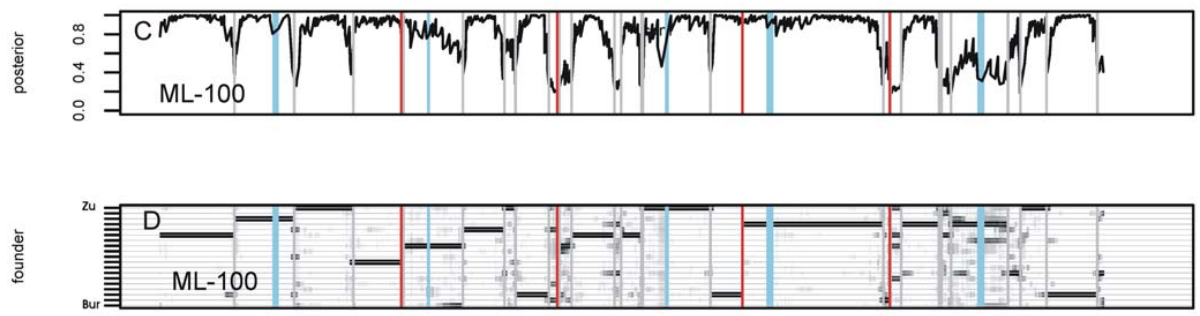


Figure 2.5: Genetic mosaic reconstruction of a MAGIC line based on 1260 SNPs [67]. The mosaics were generated by computing posterior probabilities $P_{s,L}$ of carrying founder s at locus L. In both figures the horizontal axis show genomic position, while vertical red lines mark chromosome boundaries and blue lines mark the centromeres. (a) Maximum posterior probability across all founder haplotypes L at a locus s. Vertical grey lines show likely recombination breakpoints - a sharp decay in the maximum posterior followed by a sharp increase indicates a change in the identity of the most probable founder. (b) The maximum posterior table. The vertical axis encodes the 19 possible founder assignments s, while the shade at each position (s, L) encodes the value of $P_{s,L}$, with white corresponding to $P_{s,L} = 0$ and black to $P_s^{(L)} = 1$. (Figure taken from [67])

FRIGIDA being the most likely causal gene. Different QTL mapping strategies gave concordant QTLs for these phenotypes, suggesting that population structure does not affect the predictions.

The founder accessions were resequenced at high coverage (27-60x) and reassembled by iterative mapping and *de novo* assembly, allowing the identification of over 500k Single Nucleotide Polymorphisms (SNPs) and 22k indels per accession [36]. Of the 3.07 million SNPs that are segregating in the 19 founders, 45.2% are private to single accessions. Therefore, a nearly complete catalogue of sequence variants for the population is available, which was used for the re-annotation of the genomes.

488 MAGIC lines were sequenced at low $(0.1 - 0.5$x) coverage with Illumina paired-end reads of length 51bp, comprising the data set that is used in this thesis. The MAGIC lines were grown at Bath (lab of Dr Paula Kover) or Oxford (lab of Professor Nick Harberd) in greenhouses or growth chambers respectively. Leaves were harvested for DNA extraction. DNA preps were performed at the John Innes Centre, in 96 well plates using the DNeasy 96 Plant Kit and DNeasy 96 Protocol (www.quiagen.com). Sequencing was performed by the Oxford Genomics Centre. The Illumina reads were mapped to the *A. thaliana*reference genome (TAIR10) using stampy ver-

AIMS stock center	**Accession**	**Origin**
CS6643	Bur-0	Ireland
CS6660	Can-0	Canary Isles
CS6673	Col-0	USA
CS6674	Ct-1	Italy
CS6688	Edi-0	Scotland
CS6736	Hi-0	Netherlands
CS6762	Kn-0	Lithuania
CS20	Ler-0	Poland (formerly Germany)
CS1380	Mt-0	Libya
CS6805	No-0	Germany
CS6824	Oy-0	Norway
CS6839	Po-0	Germany
CS6850	Rsch-4	Russia
CS6857	Sf-2	Spain
CS6874	Tsu-0	Japan
CS6889	Wil-2	Russia
CS6891	Ws-0	Russia
CS6897	Wu-0	Germany
CS6902	Zu-0	Germany

Table 2.1: The 19 founder accessions of the MAGIC lines. The columns show the stock centre numbers, the accessions names and the place of origin.

sion v1.0.20 [83]. Alignments were stored in a separate bam file for each MAGIC line. Previous sequencing for the 18 MAGIC line progenitors had produced a catalogue of $3,316,270$ segregating SNPs [36] (http://mus.well.ox.ac.uk/19genomes). I ran GATK v2.626 [80] on the segregating SNPs to call variants for the 19 founders, setting the following read filters: Allele Balance, BaseQualityRankSumTest, Clipping RankSumTest, Coverage, DepthPerAlleleBySample, FisherStrand,

GCContent, HaplotypeScore, LowMQ, MappingQualityRankSumTest, MappingQualityZero, MappingQualityZeroBySample, RMSMappingQuality, ReadPosRankSumTest. I filtered out SNPs that were triallelic, within transposons, or heterozygous for any founders.

2.3 Next-generation sequencing

Next generation sequencing [88] allows the rapid sequencing of genomes in parallel. Its efficiency lies in the ability to produce a large number of short ($25 - 500$bp) reads quickly, so that any locus can be covered by many reads to resolve ambiguities. The most widespread next-generation sequencing method is *paired-end sequencing*, in which DNA is fragmented in short pieces of about $200 - 400$bp and both ends of each fragment are sequenced to produce a pair of reads. Each read in the pair originates from opposite DNA strands. The fragment size is pre-determined, therefore the *insert size* i.e. the distance between reads in a pair, is known. This facilitates genome assembly, as ambiguities in the alignment, caused mainly by structural variants - see also Chapter 4, can be resolved by looking at the alignments of both mates. An example of normal read-pair alignments from chromosome 1 of the MAGIC founder Ler-0, is shown in Figure 2.6.

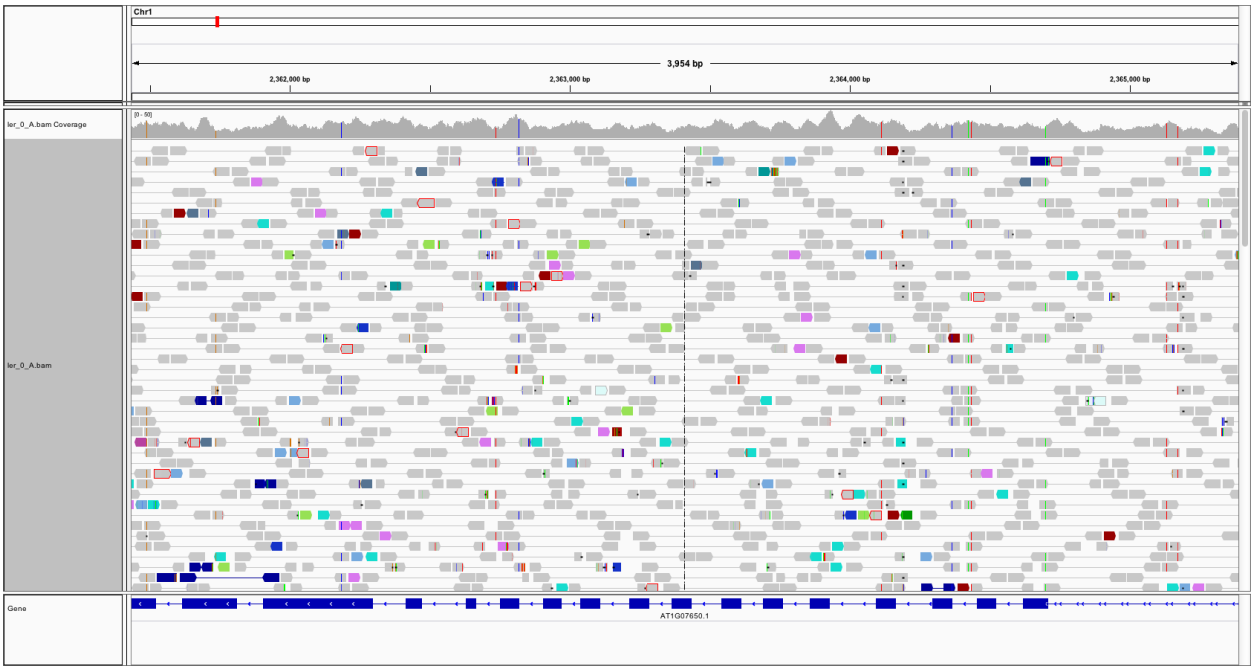


Figure 2.6: Normal paired-end reads from Ler-0, aligned to the reference genome Col-0 (TAIR10) from chromosome 1, visualised in iGV[123] . The top panel shows the total read coverage of the region, below the read alignments are displayed and the final panel shows annotated genes. Paired reads are connected with a line. Normal reads (i.e. reads with normal insert size and good mapping quality) are grey and coloured reads are anomalous. While some anomalous reads are present they are scattered throughout the region and every position is predominantly covered by normal reads. Therefore, the anomalous reads present in the region are probably random error.

2.3.1 Genome assembly

Genome assembly is the procedure of reconstructing a genome from short reads. The most widely used approach is to map reads against a reference genome and then identify differences. A common read mapping strategy, employed by software such as MAQ, MOSAIK, Smalt and Stampy[76, 72, 102, 83] is to use hash tables with small k-mers as the unique keys, matched to genomic loci in which these k-mers are present. Before alignment, each read is matched to the hash table to detect potential alignment positions. This is faster than aligning a read against the whole genome, as only specific genomic locations are considered.

Other read mappers, such as BWA, Bowtie and SOAP2 [136, 70, 78], use the Burrows-Wheeler transform (BWT), a technique used for reversible sequence compression based on character permutation. BWT compresses the sequence by placing the same characters next to each other in

the compressed sequence, along with information on the steps needed to decompress the sequence. The transform is used to compress selected fragments of the reference sequence that are then used as entries in arrays containing FM-indices, which are data structures designed to efficiently detect overlaps of BWT sequences [32]. Overall, use of BWT speeds up performance, but may sacrifice the alignment accuracy, so additional steps are employed for ambiguous alignments. For example, if a read cannot be aligned consecutively, as in the case of structural variants, most read-mappers use a read realignment strategy. For example, BWA addresses insertions and deletions by splitting unmapped reads in two or more fragments which are then separately aligned to the reference. Stampy [83] uses information about the alignment of the mate and invokes a separate gapped alignment algorithm. Reads that map to multiple positions get a low mapping quality score and may be filtered out.

A different approach, employed by software such as SOAPdenovo2, ALLPATHS, Velvet [84, 14, 141] is to assemble reads *de novo*. Instead of using a reference genome *de novo* assemblers align reads against each other and merge them into larger sequence contigs. *de novo* assembly is much slower than read mapping, but can be more informative with respect to genomic rearrangements or insertions. On the other hand, ambiguous regions such as repeats or copy number variations are hard to assemble *de novo* and read mappers perform better in those cases. To benefit from both approaches, assembly strategies that combine read mapping and *de novo* assembly have been implemented, such as IMR/DENOM, used on the 19 MAGIC founders [36].

2.3.2 SNP calling

After mapping reads to the reference genome it is possible to identify SNPs using a variant calling algorithm, such as GATK [89] and Samtools [74]. GATK uses a Bayesian framework that considers a column of all aligned reads over a site and computes the probability of observing a genotype given the data. The model can be extended to deal with noise and ambiguities in read alignments, for example setting a maximum allowed read coverage, requiring balanced allele calls in heterozygous genotypes, or computing haplotype scores assuming a maximum of two segregating haplotypes.

Samtools [74] variant caller is based on mapping quality scores of reads produced by read map-

ping software. A quality score is assigned to each base, which is the minimum between the overall read mapping quality and the individual base sequencing quality. Samtools considers only the two most frequent base calls and computes the posterior probability of the site being homozygous or heterozygous. It uses bayesian priors based on whether the site is a known variant site or not and based on concordance with the reference genotype at the site.

Chapter 3

Reconstruction of genome mosaics in the MAGIC *Arabidopsis* population

3.1 Introduction

Low-coverage sequencing of a population, in particular where average coverage is < 1, will genotype each individual at a different and random set of SNPs. However any genetics study will compare all genomes using a common set of SNPs. In a synthetic population such as MAGIC, the population history is known and can be used to infer missing genotypes. The MAGIC genomes are mosaics of the 19 founders of the population, whose variation is known at high precision [36], so by comparing the observed genotypes in each individual with the catalogue of variants in the founders we can reconstruct the MAGIC genetic mosaics and impute their complete genotypes.

In this chapter I describe two algorithms used for mosaic reconstruction, based on dynamic programming and Hidden Markov Models (HMMs). The first reconstruction algorithm (best single-sequence reconstruction) uses dynamic programming to retrieve the maximum-scoring assignment of founder haplotypes to a given set of genotypes, and is similar to the Viterbi path of an HMM [94]. The algorithm makes the assumption that genomes are fully homozygous and hence behave as haploids. There is also a diploid extension of the algorithm which allows both homozygous and heterozygous haplotypes. The second algorithm is a standard implementation of the Forward-

Backward algorithm for Hidden Markov Models, which computes posterior probabilities of all haplotype assignments at each position. The algorithm was developed primarily to resolve ambiguities in mosaic predictions over the (few) regions with multiple possible assignments, for example in regions where multiple founders are genetically similar.

Both algorithms are evaluated by simulation, which also serves to select optimal parameters. Simulations show that both algorithms, in haploid and diploid mode can accurately infer the mosaics at the majority of loci, in both high and low-coverage data. The imputation accuracy with low-coverage data is only slightly compromised compared to high-coverage, while the gains in computational efficiency are great. I also show that the mosaic reconstruction algorithms can be used to infer recombination breakpoints, with recombinants mapped to the intervals between the closest pairs of genotypes with different haplotype assignments. Simulations show that the haploid algorithm can predict recombination events at high precision.

The last section of the chapter presents and analyses the 488 MAGIC genetic mosaics, reconstructed using the haploid and diploid version of the best-single reconstruction algorithm.

3.2 Variant calling in MAGIC

To infer the mosaic structure of the MAGIC lines, we obtained next-generation sequencing data from 488 MAGIC lines from 284 lineages at mean read coverage 0.36x using Illumina 52bp paired-end reads. We aligned the reads to the reference genome (Col-0 TAIR10) using Stampy [83]. Using the IMR/DENOM variant-caller [36] and GATK [89], we called single nucleotide polymorphisms (SNPs) at the 3.07 million sites known to be segregating in the 19 founder genomes [36], and which are distributed almost uniformly across the genome (Figure 3.1). GATK was used to genotype variants at the sites previously discovered by IMR/DENOM.

The two variant callers gave substantially different heterozygosity estimates. IMR/DENOM called 21.8% of SNP sites as heterozygous in at least one founder; this number was raised to 40.4% by GATK. Heterozygosity estimated by the two programs in each founder genome is shown in Table 3.1 and in Figure 3.2. The average heterozygosity y of a founder was typically estimated at $1-2\%$ by IMR/DENOM and at $4-5\%$ by GATK. The exceptions were the founders Po-0 (8.9% or 16.3%) and

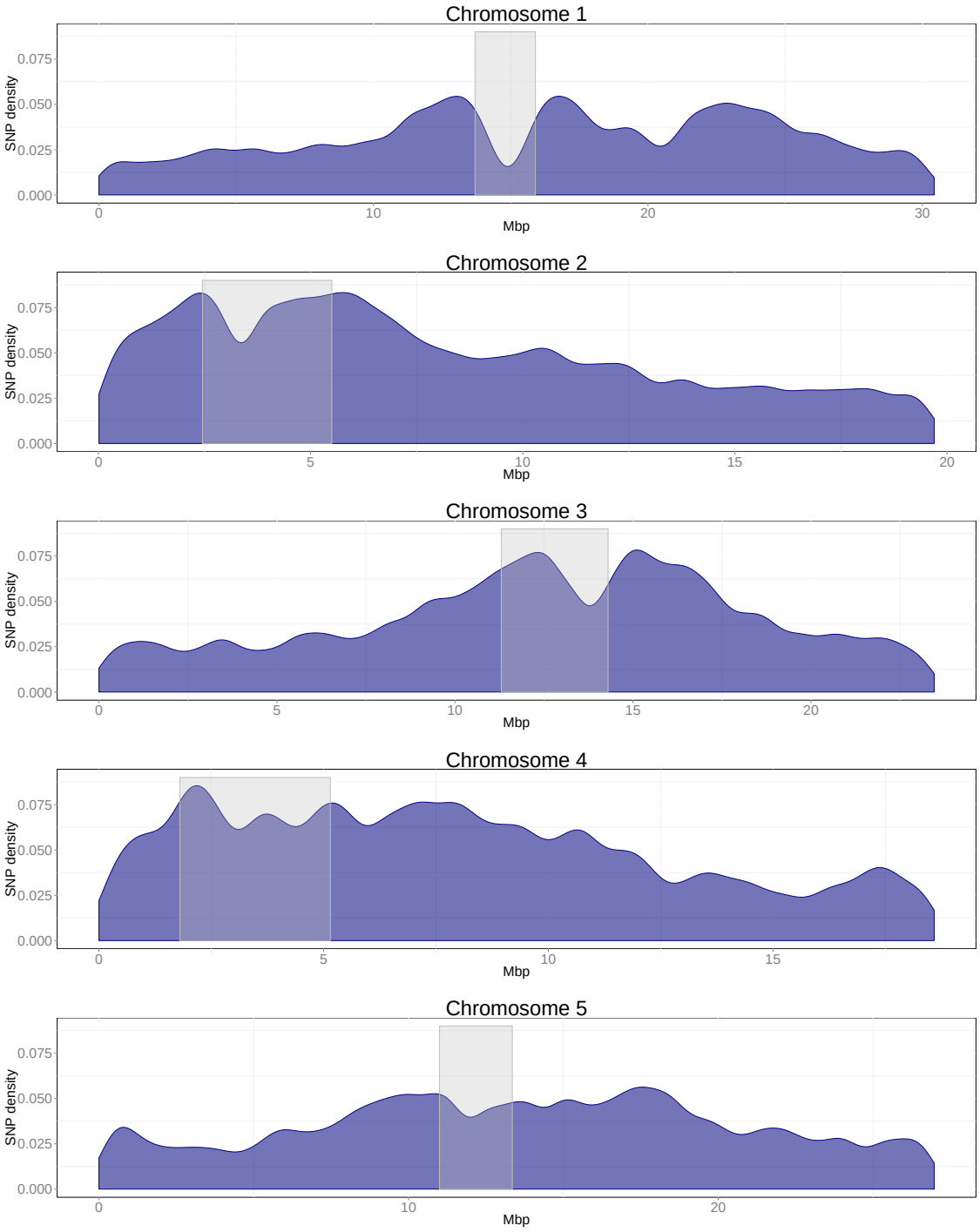


Figure 3.1: Density of SNPs tagged in the 19 MAGIC founders, per chromosome. The grey blocks mark the centromeres. Density plots are designed using density() function in R, which uses kernel density estimation so that the area under the curves is equal to 1.

Hi-0 (5% or 13.7%). Heterozygosity levels can affect mosaic reconstruction because they are related to the number of haplotypes present in the data: if one of the 19 founders was heterozygous, then the number of haplotypes present in MAGIC is 20, not 19. In the MAGIC mosaic reconstruction, we can bypass this problem by filtering out heterozygous SNPs in the founders. However, in case of undetected heterozygosity the number of haplotypes detectable in MAGIC would technically increase, but the extra haplotypes would be "hidden" in the founder data as the corresponding genotypes would be missing. In this sense, the more conservative GATK heterozygosity estimates are favourable; even if heterozygosity is overestimated, downsampling to a smaller set of SNPs and so decreasing mapping resolution is a smaller problem than dealing with unknown haplotypes.

Throughout this thesis I present results based on the GATK variant calls. Chapter 5 is an exception where I analyse the initial genome mosaics of MAGIC, generated with the variant calls of IMR/DENOM and discuss the implications of having "hidden" haplotypes in mosaic reconstruction.

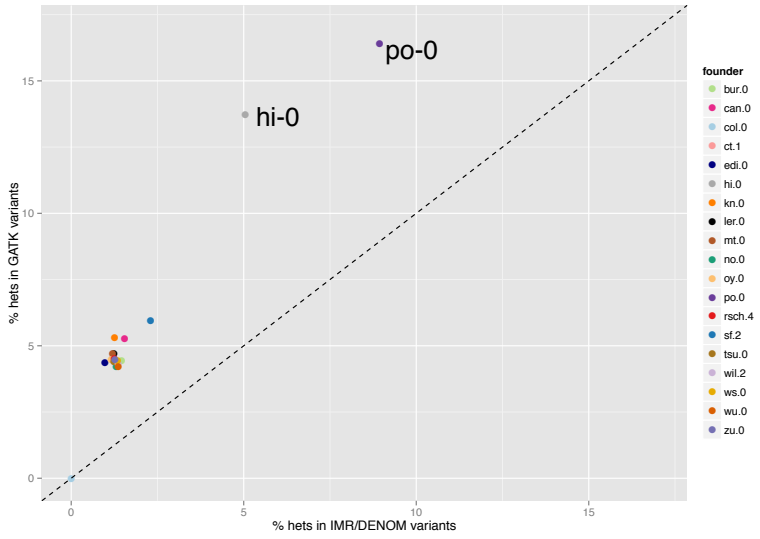


Figure 3.2: Comparison of heterozygosity levels in the alleles called in the 19 founders (among 2,487,199 SNPs), estimated by GATK and by IMR/DENOM. Each point corresponds to a founder genome; the x-coordinates show the fraction of heterozygous alleles in the IMR/DENOM set and the y-axis this fraction in GATK.

Of the total 3.07M SNPs, 157k SNPs within known transposons or repeated regions were filtered out, as well as 1.72 million SNPs for which at least one of the founders was heterozygous or were

Founder	**Hets - IMR/DENOM**	**Hets - GATK**
Bur-0	36267 (1.5%)	110511 (4.4%)
Can-0	38499 (1.5%)	130640 5.3%
Col-0	0 (0%)	0 (0%)
Ct-1	30203 (1.2%)	113006 (4.5%)
Edi-0	24460 (0.9%)	108360 (4.4%)
Hi-0	125514 (5.0%)	341827 (13.7%)
Kn-0	31402 (1.3%)	132242 (5.3%)
Ler-0	30729 (1.2%)	116513 (5.3%)
Mt-0	29997 (1.2%)	116921 (4.7%)
No-0	32599 (1.3%)	104823 (4.2%)
Oy-0	28893 (1.2%)	111544 (4.2%)
Po-0	222247 (8.9%)	407568 (16.3%)
Rsch-4	31013 (1.2%)	112718 (4.5%)
Sf-2	57171 (2.3%)	147846 (5.9%)
Tsu-0	30814 (1.2%)	109858 (4.4%)
Wil-2	34035 (1.4%)	107398 (4.3%)
Ws-0	33402 (1.3%)	110833 (4.5%)
Wu-0	33942 (1.4%)	110833 (4.4%)
Zu-0	31519 (1.3%)	111782 (5.0%)

Table 3.1: Number of heterozygous alleles called in the 19 founders (among $2,487,199$ SNPs) estimated with the IMR/DENOM variant-caller and with GATK

not biallelic. This left on average 301 thousand informative genotyped sites per line (interquartile range 275 - 446 thousand).

For mosaic reconstruction I used a dynamic programming algorithm which retrieves the highest-scoring sequence of founder assignments [94] and also implemented the Forward-Backward algorithm for Hidden Markov Models, which computes posterior probabilities $P_{s,L}$, i.e. the probability

of having founder haplotype s at locus L. Both algorithms have two modes: haploid and diploid. In haploid mode, the algorithms assume homozygous genomes so they consider 19 possible haplotype assignments (hidden states). In diploid mode, no assumption about zygosity is made (the algorithm considers both homozygous and heterozygous assignments).

On average 1 in 5 variants is different between any given pair of founders [36]. Moreover, each genome has $300,000$ typed SNPs so a SNP is typed every 400bp. Hence, we expect to be able to assign a distinct haplotype to each locus within $5 \times 400 = 2$kb on average (i.e. we expect the algorithm to be able to distinguish the correct founder assignment between any pair of founders in a region longer than 2kb), so 2kb is the expected error in the prediction of mosaic breakpoints, i.e. recombination events. The only exception is the pair of founders Oy-0 and Po-0, which are related and can be identical over Mb-sized genomic regions [36].

3.3 Best single-sequence mosaic reconstruction

The data used for reconstruction were the catalogue of variant sites in the founders and a sequence of genotypes at a fraction of these sites, which were the allele calls in the genome that is being reconstructed. I describe the problem of reconstruction as the optimisation problem of finding the founder assignment that maximises the matches between the genotypes of the sequenced genome and the called haplotypes.

The haploid version of the algorithm considers 19 haplotype assignments. The algorithm iterates through the allele calls of each genome at the known variant sites and recursively searches for the maximum-scoring path for each possible founder assignment. Assuming a sequence of loci $\{1, \ldots, N\}$ and a set of founders S, at each locus L the algorithm tries to find the founder assignment s that maximises the objective function:

$$\mathsf{Score}(L, s) = M_L(s) + \max(\mathsf{Score}(L-1, t)) - C_{s,t} \tag{3.1}$$

In (3.1), $M_L(s)$ is a function which has a positive value when allele at L matches founder s, otherwise it is zero. $C_{s,t}$ is a cost function which is subtracted from the total path score after a

change of state, to prevent the algorithm from randomly switching states. If a change of state occurs i.e. $s \neq t$ then $C_{s,t} = c, c > 0$ otherwise it is zero. The value of c is user-defined and depends on the number of genotypes available (and hence the sequencing coverage) and on the distance in generations from haplotypes to reconstructed genomes, which account for different amounts of recombination. For the MAGIC data, I estimated optimal values for c in haploid and diploid mode by simulation (Section 3.3.1). After the algorithm computes the objective function at each locus it iterates through the scores to recover the maximum-scoring founder assignment at each locus. The algorithm is summarised in Algorithm 1.

The diploid version of the algorithm has a larger state space, as it involves 190 hidden states comprising the 19 homozygous states and their 171 pairwise combinations. With low-coverage data there are few heterozygous calls in a genome, in the rare event of two reads overlapping, one from the maternal and the other from paternal chromosome. For read coverage 0.3x, for example, if we model the distribution of reads covering a site as a Poisson process with rate $\lambda = 0.3$, the probability of two or more reads overlapping a locus is $P(X \geq 2) \simeq 3.4\%$ and only in half (1.7%) of them the two reads will come from different chromosomes. Thus, most of the allele calls are based on a single read so the algorithm has to predict heretozygous haplotypes from frequent oscillations between two founder states in a genomic region; depending on the cost function, it may be cheaper to stay in the same heterozygous state than frequently change homozygous states.

There are some differences in the definition of parameters in the diploid algorithm. At rare heterozygous genotypes, if genotype $\langle u, v \rangle, (u \neq v)$ matches both founder haplotypes of state $\langle s, t \rangle$ then the score assigned to the match function $M_L(\langle s, t \rangle) = 2$, otherwise $M_L(\langle s, t \rangle) = 0$. Thus, the few heterozygous alleles guide the algorithm into selecting the correct heterozygous state. For homozygous allele calls $\langle u \rangle$, if allele a matches one of s or t then for a heterozygous haplotype $M_L(\langle s, t \rangle) = 1, (s \neq t)$ and for a homozygous haplotype $M_L(s) = 1.05$. I give a slightly increased cost in homozygous states, otherwise the algorithm would favour heterozygous states in which at least one strand always matches the allele. Also, the cost function prefers single recombinant transitions, i.e. either between homozygous states or between states that share one haplotype.

With high-coverage data, all heterozygous alleles are observed, so we do not need to give a

Data:

N = number of sites,

S = list of available founder states

Matching function: $M_L(s)$

Transition Cost: $C_{s,t}$

Output: Q - vector with q_L the maximum scoring founder at L

Initialisation:

for $s \in S$ **do**
$\mathsf{Score}(1, s) = M_1(s)$

Recursion:

foreach locus L **do**

foreach founder $s \in S$ **do**
$\mathsf{Score}(L, s) = M_L(s) + \underset{t \in S}{\mathrm{Max}}(\mathsf{Score}(L-1, t) - C_{s,t})$
$\mathsf{Path}(L, s) = \underset{t \in S}{\mathrm{Arg\,max}}(\mathsf{Score}(L-1, t) - C_{s,t})$

Recover best sequence:

$Q_N = \underset{s \in S}{\mathrm{Arg\,max}}\, \mathsf{Score}(N, s)$

foreach locus L **do**
$q_{L-1} = \mathsf{Path}(L, q_L)$

Algorithm 1: The highest-scoring sequence mosaic reconstruction algorithm.

positive score to states when only one strand matches the allele. Doing so, would in fact lead the algorithm to sometimes incorrectly prefer homozygous states, because of their higher matching score $M_L(s)$. Instead, with high-coverage data we require both strands of the state to match the allele at all times, as in homozygous reconstruction. The only difference is that in heterozygous allele calls of the form $\langle u, v\rangle, u \neq v$ I give to all matching states $\langle s, t\rangle$ twice the score of any other match, $M_L(\langle s, t\rangle) = 2$, as this narrows down the selection of states.

The parameters used by the haploid and diploid modes of the algorithm are summarised in

Table 3.2 and the parameter values for the haploid and diploid reconstruction are in 3.3.

Parameter	**Description**
$a(L)$	Allele of current genome at locus L
$g_L(s)$	Genotype of founder s at SNP L
$M_L(s)$	Value added to the score if $a(L)$ and $g_L(s)$ are consistent
$C_{s,t}$	Cost of transition from founder s to founder t

Table 3.2: Description of parameters the reconstruction algorithm.

Haploid mode value	**Diploid mode value**
$M_L(s) = \begin{cases} 1 & \text{if } a(L) = g_L(s) \\ 0 & \text{otherwise} \end{cases}$	Low-coverage data: If $a(L) = \langle u, v \rangle$ and $u \neq v$: $M_L(\langle s,t \rangle) = \begin{cases} 2 & \text{if } u = g_L(s) \text{ and } v = g_L(t) \\ 0 & \text{otherwise} \end{cases}$ If $a(i) = \langle u \rangle$: $M_L(\langle s,t \rangle) = \begin{cases} 1.05 & \text{if } s = t \text{ and } a_L(s) = u \\ 1 & \text{if } s \neq t \text{ and } a_L(s) = u \text{ or } a_L(t) = u \\ 0 & \text{otherwise} \end{cases}$ High-coverage data: $M_L(\langle s,t \rangle) = \begin{cases} 2 & \text{if } u = g_L(s) \text{ and } v = g_L(t) \text{ and } u \neq v \\ 1 & \text{if } u = g_L(s) \text{ and } v = g_L(t) \text{ and } u = v \\ 0 & \text{otherwise} \end{cases}$
$C_{s,t} = \begin{cases} c & \text{if } s \neq t \\ 0 & \text{if } s = t \end{cases}$	$C(\langle s,t \rangle, \langle s',t' \rangle) = \begin{cases} 3c & \text{if } s \neq t, s' \neq t', s \neq s', t \neq t' \\ 0 & \text{if } s = s', t = t' \\ c & \text{otherwise} \end{cases}$

Table 3.3: Parameter values for haploid and diploid mode of the most-likely sequence reconstruction algorithm.

3.3.1 Evaluation of the algorithm by simulation

I evaluated the prediction accuracy of the algorithm by running it on simulated allele calls from predetermined mosaics. I use the word segment to describe an unrecombined genomic locus imputed with the same founder haplotype in one line and the word breakpoint to describe boundaries between two segments - mosaic breakpoints should correspond to recombination events. The statistics computed for the evaluation are:

1. total false positive mosaic segments, i.e. number of mosaic segments falsely inserted by the algorithm
2. total false negative segments, i.e. number of segments missed by the algorithm
3. median mosaic breakpoint position error
4. total false haplotype segments, i.e. number of segments with correct breakpoints (within 100kb) but incorrectly annotated
5. mean genotype error, i.e. fraction of loci in which the haplotype prediction was not concordant with the called SNP allele
6. mean haplotype error, i.e. fraction of loci that were not annotated with the correct haplotype

Apart from the soundness of algorithms, the simulations help compare low and high-coverage data in terms of imputation accuracy and optimise the c parameter used in the cost function $C_{s,t}$ (see also Tables 3.2, 3.3).

I simulated the genome mosaics of two sets of 50 *A. thaliana* genomes, one fully homozygous and the other partly homozygous and partly heterozygous, each designed to test the haploid and diploid version of the algorithm. In the homozygous set, "SIM.HOMO", each genome had 25 to 35 uniformly distributed haplotype segments. The heterozygous set, "SIM.HET", was based on SIM.HOMO: the mosaic breakpoints were the same, but some mosaic segments were heterozygous. Selection of segment zygosity was random with the requirement that neighbouring segments had to be either both homozygous or share one haplotype (for example, a homozygous mosaic segment with haplotype ⟨Col-0⟩ can be adjacent to another homozygous segment e.g. ⟨Ler-0⟩ or to a heterozygous

⟨Col-0, Ler-0⟩ segment, but not to a heterozygous ⟨Bur-0, Ler-0⟩). This is because change of haplotypes in both strands would imply two coincident recombinants, an extremely unlikely event.

For both simulated sets, I generated allele calls for the genomes corresponding to low and high-coverage sequencing data, with 300,000 and 1.291M (i.e. the full set) SNPs, respectively. For heterozygous genomes in the low-coverage set there were about 2% heterozygous allele calls per genome, as expected. In the high-coverage set all heterozygous alleles were observed.

Single-read sequencing data using the Illumina platform have about 1% of error in allele base calls [98] and this error rate should be present in the real sequencing data. To account for this, at 1% of sites the called allele was selected at random, so we expect 0.75% of total genotype error in the reconstructed mosaics (since a quarter of the time the correct genotype is substituted at random). The simulation datasets are summarised in Table 3.4.

N	**Simulated mosaic**	**Simulated genomes**	**Total breaks**	**Hets**	**Allele Calls**	**% hets observed**
1	SIM.HOMO	50	1,471	no	300k	0
2	SIM.HOMO	50	1,471	no	1.291M	0
3	SIM.HET	50	1,471	yes	300k	2%
4	SIM.HET	50	1,471	yes	1.291M	100%

Table 3.4: Simulation datasets. Each row corresponds to a simulation dataset. The columns are: **N**: index of simulation set, **Simulated mosaic**: simulated mosaic on which the simulation was used, **Simulated Genomes**: number of simulated genomes in each set, **Total breakpoints**: Total number of mosaic breakpoints in each set, **Hets**: boolean indicator of the presence of heterozygous haplotypes in the mosaic, **Allele calls**: number of allele calls in each of the simulated genomes, **% hets observed**: if heterozygous haplotypes are present in the mosaic, fraction of heterozygous allele calls observed

Evaluation of the haploid best-single reconstruction algorithm

I ran the haploid reconstruction dynamic programming algorithm on the low and high-coverage SIM.HOMO allele calls, for several values of the c parameter, namely for $c \in \{1, 2, 3, 4, 5, 10, 20, 35, 50, 75\}$, and computed the evaluation statistics described in Section 3.3.1.

The results of simulations for haploid reconstruction are summarised in Table 3.5 and Figure 3.3 shows an example of a simulated mosaic and its corresponding predicted mosaic by the recon-

struction algorithm. If $c > 2$, no false positive breakpoints were made, i.e. the algorithm detected a breakpoint only where there was one. There was a small number of false negatives, affecting small segments ($<$ 200kb) that did not contain enough segregating SNPs to be detectable. As expected, the number of false negatives was higher in the low-coverage dataset than in the high-coverage (for example, 1.97% vs 0.88% for $c = 5$). The haplotype error was small in both datasets, and entirely due to the (known) sequence similarity between founders Po-0 and Oy-0, causing Po-0 segments to be tagged as Oy-0. The genotype error was extremely low and close to the expected value of 0.75%. Finally, the median breakpoint prediction error was elevated in the low-coverage dataset and was close to the expected error of 2kb. In both low and high-coverage datasets, lower values of the cost parameters c increase breakpoint prediction accuracy with optimal value $c = 2$; when $c < 2$ false positives were introduced (Figure 3.4). Higher c values result in more false negatives, because the algorithm requires more allele mismatches to call a change of state, so it is likely to miss smaller segments.

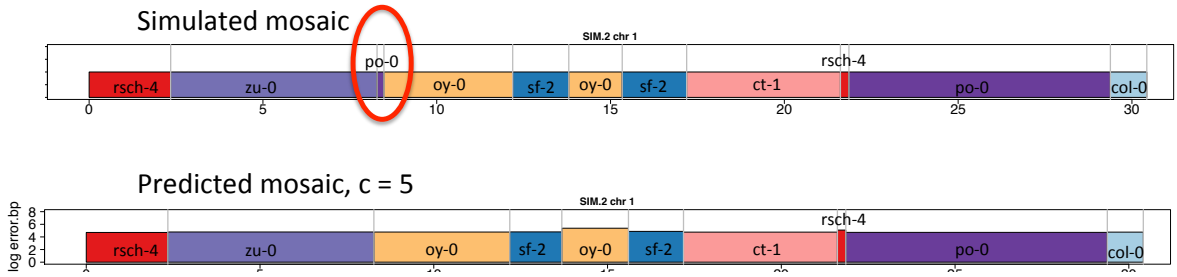
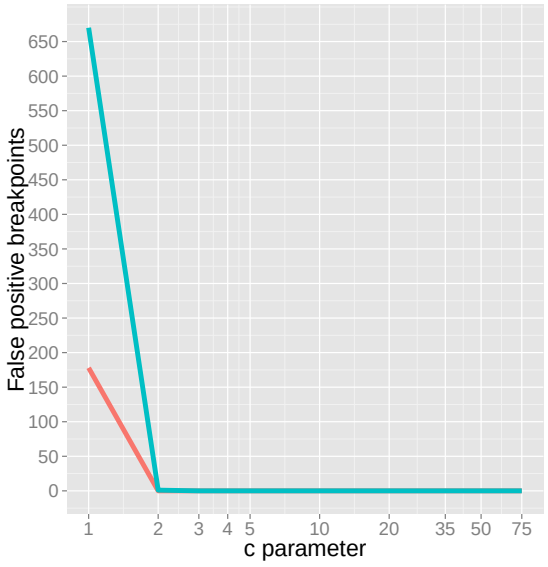


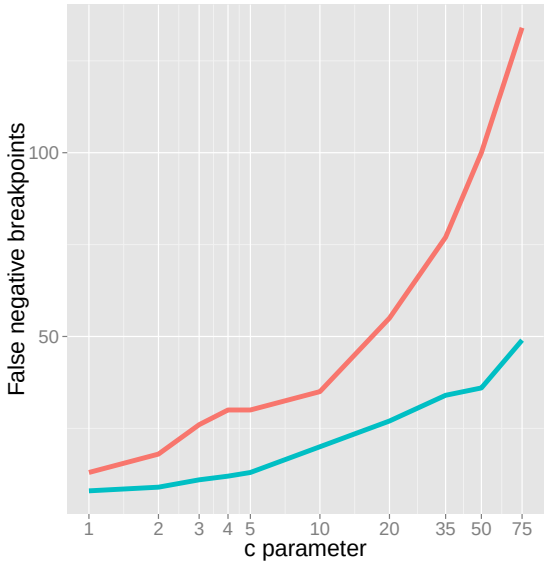
Figure 3.3: Simulated haploid mosaic (simulated genome SIM.2, chr 1) and corresponding predicted mosaic by haploid reconstruction, with cost parameter $c = 5$. The x-axis shows genomic position (in Mb) and each coloured segment corresponds to a genomic interval descended from a specific founder haplotype. In the simulated mosaic the y-axis is constant while in the predicted mosaic the y-axis shows the $-log10(E)$, with E the genotype error divided by the total length of the segment. There is one false negative, circled in red.

Simulated set	c	**Predicted breaks**	**False positives**	**False negatives**	**False haplotypes**	**Break error (bp)**	**Genotype error**	**Haplotype error**
	1	1653	178	13	36	2634	1965 (0.7%)	4310 (1.4%)
	2	1454	0	18	36	2555	1966 (0.7%)	4703 (1.6%)
SIM.HOMO	3	1446	0	26	32	2540	1966 (0.7%)	4758 (1.6%)
(300k SNPs)	4	1442	0	30	32	2527	1966 (0.7%)	4761 (1.6%)
	5	1442	0	30	32	2527	1966 (0.7%)	4761 (1.6%)
	10	1437	0	35	32	2519	1967 (0.7%)	4812 (1.7%)
	20	1417	0	55	32	2519	1973 (0.7%)	4989 (1.7%)
	35	1395	0	77	29	2497	1984 (0.7%)	5094 (1.7%)
	50	1372	0	100	29	2512.5	2004 (0.7%)	5320 (1.8%)
	75	1338	0	134	29	2569.5	2047 (0.7%)	5707 (1.9%)
	1	2209	670	8	28	1154	8381 (0.6%)	11303 (0.9%)
	2	1464	1	9	21	1084	8384 (0.6%)	10562 (0.8%)
SIM.HOMO	3	1461	0	11	19	1084	8384 (0.7%)	10571 (0.8%)
(1.291M)	4	1460	0	12	19	1084	8384 (0.7%)	10571 (0.8%)
	5	1458	0	13	19	1084	8385 (0.7%)	10610 (0.8%)
	10	1451	0	20	18	1084	8386 (0.7%)	13085 (1%)
	20	1444	0	27	18	1084	8388 (0.7%)	13138 (1%)
	35	1437	0	34	18	1084	8392 (0.7%)	13244 (1%)
	50	1435	0	36	18	1084	8393 (0.7%)	13375 (1%)
	75	1422	0	49	20	1084	8410 (0.7%)	13964 (1%)

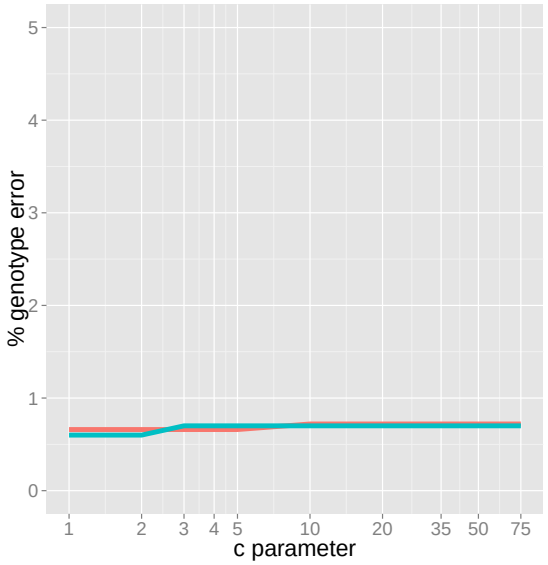
Table 3.5: Results of the evaluation of homozygous reconstruction algorithm, using simulated genome mosaics and simulated allele calls. The columns are: **Simulated mosaic**: simulated mosaic on which the simulation was used, c **parameter**: cost parameter used in each run of the algorithm **Predicted breaks**: total number of breakpoints predicted by each run of the algorithm in all genomes **False positives**: total number of breakpoints falsely inserted by the algorithm **False negatives**: total number of unpredicted breakpoints **Incorrect haplotypes**: total number of segments annotated with incorrect haplotype **Break error (bp)**: median distance between the actual breakpoint position in the simulated set and its position predicted by the algorithm **Genotype error**: mean number of sites per line where genotype (called allele) disagrees with the predicted haplotype, **Haplotype error**: average number of sites per line with incorrect haplotype prediction.



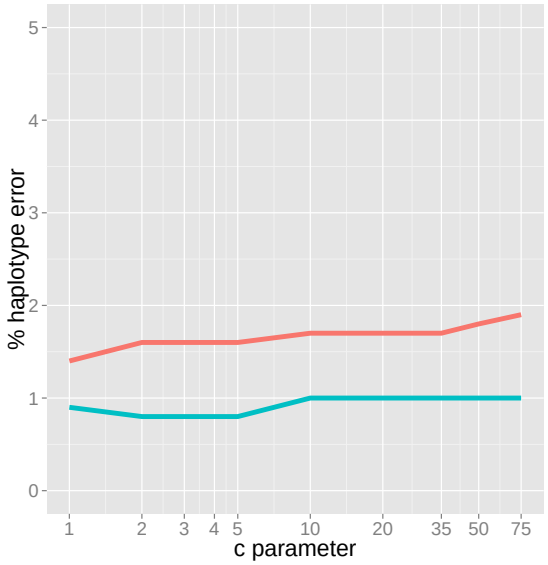
(a) False positive rate



(b) False negative rate



(c) Genotype error rate



(d) Haplotype error rate

Figure 3.4: Error rates per c (transition cost) value in haploid best single-sequence reconstruction algorithm, estimated by simulation. The x-axis shows the c values considered. The y-axis shows: in Figure 3.4a number of false positives, in Figure 3.4b number of false negatives, in Figure 3.4c the genotype error (percentage of called alleles mismatching the haplotype predicion) and in Figure 3.4d the haplotype error (percentage of sites annotated with incorrect haplotype). The two lines in each figure correspond to the two different simulated datasets for homozygous reconstruction: in salmon, the set corresponding to low-coverage data, with 300k SNPs per genome and in turqoise the set corresponding to high-coverage data, with 1.291M SNPs per genome. The two lines overlap for $c \geq 2$ in Figure 3.4a and for $c \geq 10$ in Figure 3.4c.

Evaluation of the diploid best-single reconstruction algorithm

The same statistics were used for the evaluation of the diploid reconstruction algorithm, and for $c \in \{5, 10, 20, 35, 50, 75\}$. Results of the evaluation are shown in Table 3.6. Some false positives were observed, which were predominantly very small segments inserted between large correctly-called states (Figure 3.5 shows an example). False positive segments are mostly inserted over loci with sequencing errors - hence the genotype error is lower than the expected 0.7%. False positives decrease as c increases but they are also negatively correlated with false negatives (Figures 3.6a, 3.6b). Haplotype error is minimised and equal to 5% at $c = 10$ with low-coverage data, and at $c = 50$ with high-coverage data (Figures 3.6c, 3.6d). The higher haplotype error compared to the homozygous data is mainly due to false positive segments and to Po-0 and Oy-0 haplotypes which are related and hard to distinguish - the diploid mosaics have twice as many haplotypes with at least one strand Po-0 and Oy-0 compared to the haploid, so they are affected more by their relatedness.

Simulated set	c	**Predicted breaks**	**False positives**	**False negatives**	**False haplotypes**	**Break error (bp)**	**Genotype error**	**Haplotype error**
SIM.HET (300k SNPs)	5	2101	648	68	158	18470	1138 (0.4%)	18721 (6.2%)
	10	1653	258	103	141	18890	1130 (0.4%)	15122 (5.0%)
	20	1502	188	173	139	18208	1150 (0.4%)	16194 (5.4%)
	35	1360	133	257	164	19150	1212 (0.4%)	20606 (6.9%)
	50	1237	94	345	182	21215	1293 (0.4%)	25249 (8.4%)
	75	1091	380	69	462	21652	1440 (0.5%)	34468 (11.4%)
SIM.HET (1.291M)	5	1585	148	48	90	1377	8734 (0.7%)	89944(7.0%)
	10	1544	119	55	84	1330	8741 (0.7%)	79048 (6.1%)
	20	1522	109	67	81	1314	8748 (0.7%)	71821 (5.6%)
	35	1509	108	79	84	1300	8761 (0.7%)	69183 (5.4%)
	50	1488	103	94	82	1307	8784 (0.7%)	65896 (5.1%)
	75	1471	101	108	85	1361	8819 (0.7%)	66363 (5.1%)

Table 3.6: Results of the evaluation of diploid reconstruction algorithm, using simulated genome mosaics and simulated allele calls. The columns are the same as in Table 3.5

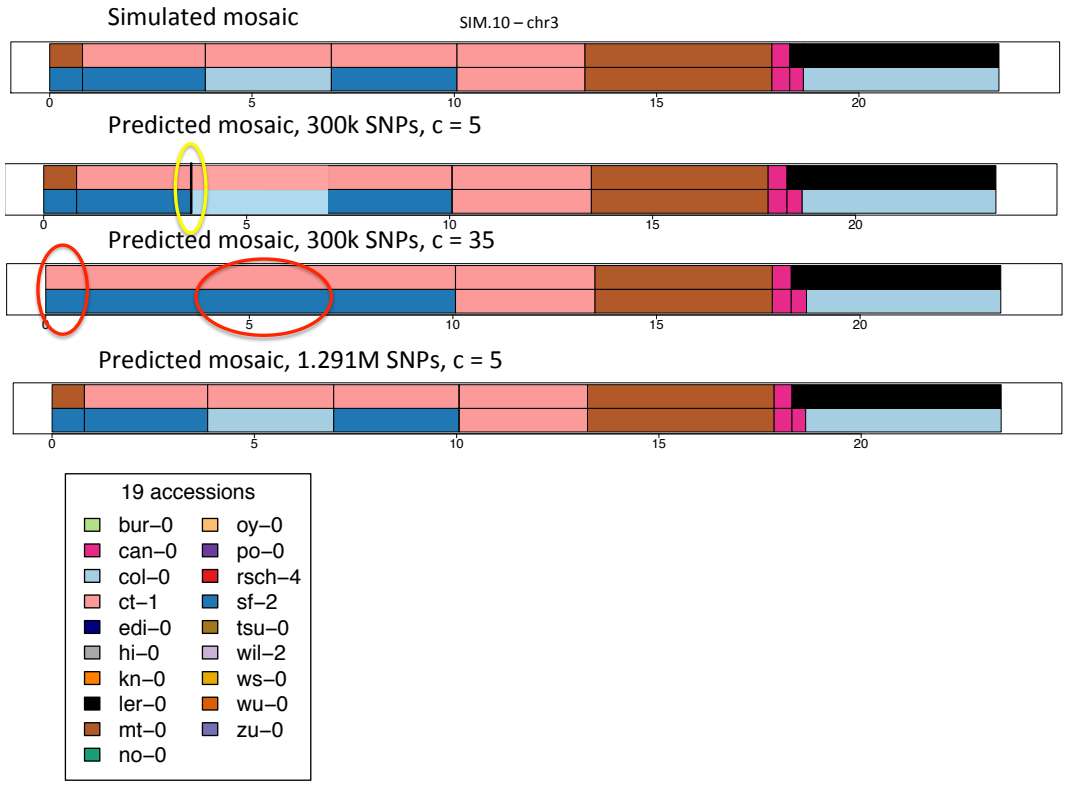
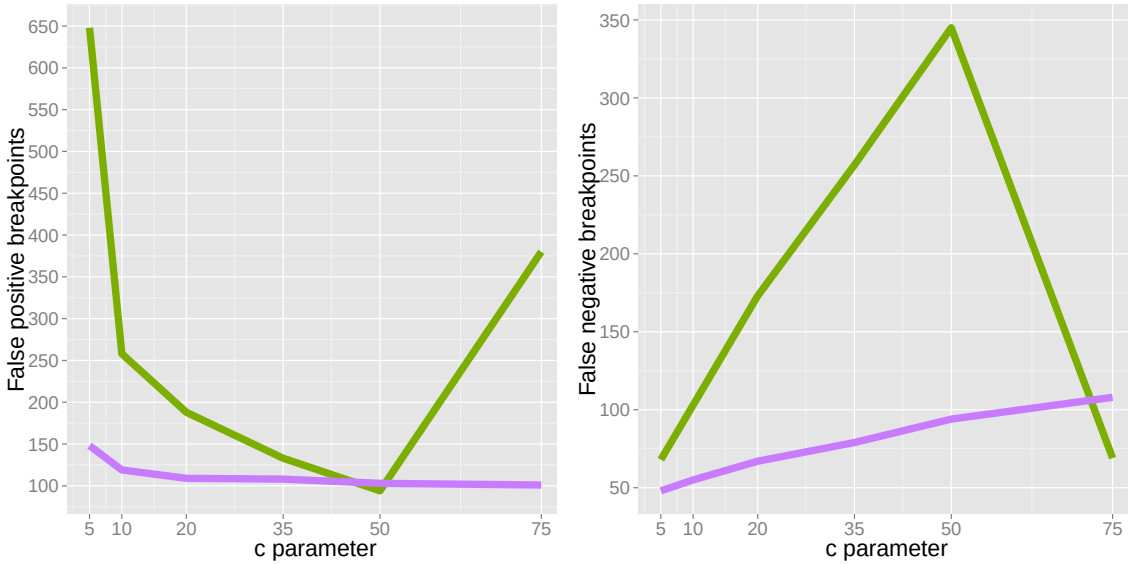


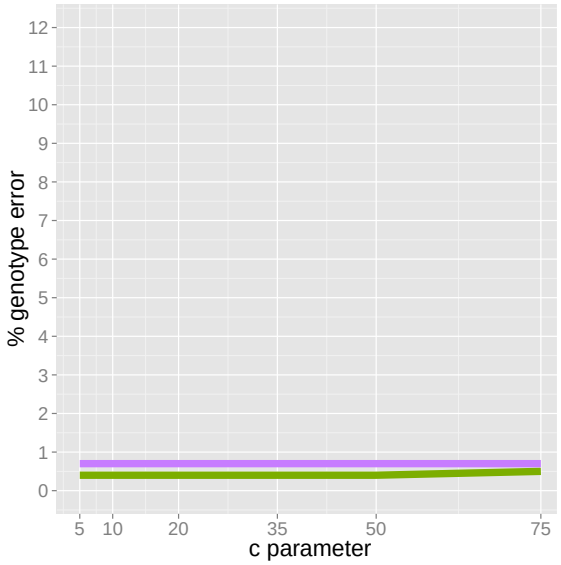
Figure 3.5: Simulated diploid mosaic and the corresponding mosaics predicted by diploid reconstruction with cost parameters $c = 5, 35$ and for $c = 5$ with low and high-coverage data from chromosome 3 of simulated genome SIM.10 . The x-axis shows genomic position (in Mb) and the y-axis is constant in all mosaics. The top-panel shows the simulated mosaic, while the three panels below show reconstructed mosaics with low-coverage data and $c \in \{5, 35\}$ and with high-coverage data and $c = 5$, The yellow circle marks a false positive: a very small segment is inserted on the first predicted mosaic - segment is too small to see the predicted haplotype. Red circles indicate false negatives. The legend shows the colour-coding of the 19 founders.

Overall, both the haploid and the diploid algorithm performed sufficiently well and they were able to accurately reconstruct the mosaics for the most part. The haploid algorithm is more accurate, mainly due to its smaller state space. In particular, the minimum haplotype error with low-coverage data was 1.4% for the haploid algorithm and 5% for the diploid. In both algorithms false positives were negatively correlated with false negatives - as c approached zero false positives were introduced, because the algorithm became sensitive to sequencing errors; however, with higher c values the chance that the algorithm would disregard small segments increased. In haploid mode

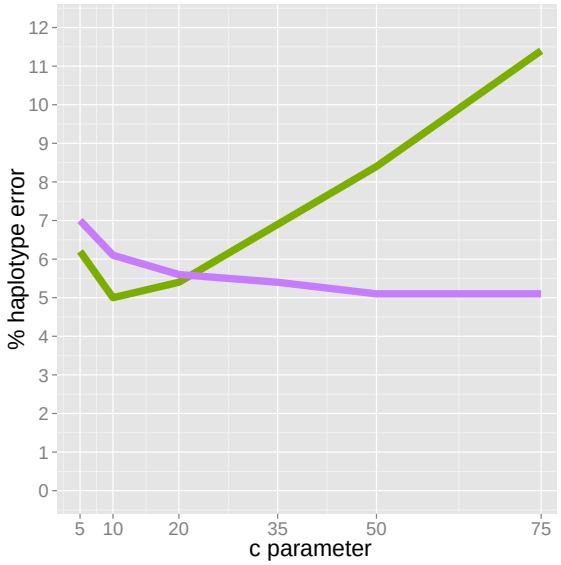


(a) False positive rate

(b) False negative rate



(c) Genotype error rate



(d) Haplotype error rate

Figure 3.6: Error rates per c in diploid best single-sequence reconstruction algorithm, estimated by simulation. The figure can be interpreted as Figure 3.4. The green line shows low-coverage simulated data and the purple high-coverage simulated data.

$c = 2$ is optimal in both low and high-coverage data with respect to the trade-off between false positive and false negative rates. With low-coverage data, the breakpoint position error is minimised

with $c = 10$ while false negatives only slightly increase, so this is the optimal value for recombination analysis. The diploid algorithm is generally more sensitive to changes of c and is optimal for $c = 10$ with low-coverage data (minimum haplotype error and minimum sum of false positive and false negative predictions) and for $c = 20$ (minimum false positive and false negative predictions) or $c = 50$ (minimum haplotype error) with high-coverage data.

3.4 Mosaic reconstruction using the Forward-Backward algorithm

The highest-scoring sequence reconstruction can fail in regions where multiple founder states are identical or very similar, as then the algorithm chooses one of the matching haplotypes at random. An alternative approach to recovering the optimal sequence is to average over all possible reconstructions for all founders s at each locus L using the Forward-Backward algorithm. The algorithm cannot strictly be used for haplotype annotation, as the most likely haplotype at each locus does not necessarily belong to the correct annotation. However, it can help determine the quality of predictions of the best single-sequence reconstructions and identify regions of uncertainty in the reconstructed genome. I have implemented the Forward-Backward algorithm in its classical stochastic form, as described in [104] and summarised in Algorithm 2.

The algorithm scans the genome twice: one from the first locus to the last (forward scan) and one from the last to the first (backward scan). Assuming a sequence of N loci, the forward scan computes the forward probabilities $P^F_{s,L}$ at each locus L and for each possible founder s, i.e. the probabilities of being in state s after observing the sequence of alleles at $\{1, 2, \ldots, L\}$. Similarly, the backward scan computes the backward probabilities $P^B_{s,L}$, which are the probabilities of observing sequence $\{L, \ldots, N-1, N\}$ starting with state s. At each locus the forward and backward probabilities $P^F_{s,L}$ and $P^B_{s,L}$ are standardised so that $\sum_{s \in S} P^F_{s,L} = 1$ and $\sum_{s \in S} P^B_{s,L} = 1$. This description is for the haploid version of the algorithm; the diploid is similar.

The emission matrix E encodes the probability of observing an allele at L given annotation s and has dimensions $N \times 10$, as there are 10 possible genotype calls (allowing heterozygous calls). In the haploid Forward-Backward algorithm, at any locus L each founder state s emits $g_L(s)$ with probability 99%, otherwise it randomly emits some other nucleotide with probability 1% (modelling

a sequencing error of 1%). In diploid Forward-Backward with low-coverage data, the emission probability is 0.99 for matching genotypes if only one allele of the genotype is observed, or if a homozygote is called and $1 - 10^{-5}$ otherwise, i.e. if the called genotype is of the form $\langle u, v \rangle, u \neq v$. With high-coverage data, the emission probabilities are identical to the haploid algorithm.

The transition matrix R carries the transition probabilities $r_{s,t}$ for moving from state s to state t. In the haploid version of the algorithm there are $S = 19$ potential transitions to states: one of which corresponds to staying to the same state and the remaining 18 reflect the change to any other state. I give high transition probability i.e. $1 - 10^{-5}$ to staying to the same state, while 10^{-5} is divided among the 18 remaining states. The same transition probabilities are used in the diploid algorithm, with the additional condition that transitions between heterozygous states involving two disjoint haplotypes are not allowed. Thus, for every state (haploid and diploid), there are 37 possible transitions, one to the same state and the remaining 36 to all diploid states (homozygous and heterozygous) in which one haplotype is shared with the current state. The differences in E and R between homozygous and heterozygous reconstruction are presented in Table 3.7.

3.4.1 Evaluation of the Forward-Backward algorithm by simulation

We evaluated the algorithm using the simulation datasets used in 3.3.1, using two statistics: the maximum posterior probability $P_{\mathsf{max}}(L)$ at a locus L and the posterior probability of the true haplotype assignment on that site $P_{\mathsf{true}}(L)$. $P_{\mathsf{max}}(L)$ tests whether the algorithm converges to a single state and can help estimate the fraction of the genome with ambiguous haplotype assignments. $P_{\mathsf{true}}(L)$ is estimated by comparing the initial simulated mosaic to the posteriors table P and indicates the predictive accuracy of the algorithm, i.e. how often the most likely assignment is the right one. Table 3.8 summarises the results of computing these statistics for different transition probability values in haploid reconstruction and Figures 3.7 and 3.8 show example reconstructions with the Forward-Backward algorithm with low-coverage genomes as well as the distribution of $P_{\mathsf{max}}(i)$ and P_{true}.

In both haploid and diploid mode, and in both low and high-coverage datasets, $P_{\mathsf{max}}(i)$ and P_{true} were close to 1 at the majority of sites. This means that the algorithm usually converges to a

Haploid mode	**Diploid mode**
$E(L,s) = \begin{cases} 0.99 & \text{if } a(L) = g_L(s) \\ 0.01 & \text{otherwise} \end{cases}$	Low-coverage data: If $a(L) = \langle u, v \rangle$ and $u \neq v$: $E(L,\langle s,t\rangle) = \begin{cases} \frac{1}{2}(1-10^{-5}) & \text{if } s \neq t, u = g_L(s) \text{ and } v = g_L(t) \\ \frac{1}{2}(10^{-5}) & \text{otherwise} \end{cases}$ If $a(L) = \langle u \rangle$: $E(L,\langle s,t\rangle) = \begin{cases} \frac{1}{2}0.99 & \text{if } u = g_L(s) \text{ or } u = g_L(t) \\ \frac{1}{2}0.01 & \text{otherwise} \end{cases}$ High-coverage data: $E(L,\langle s,t\rangle) = \begin{cases} 0.99 & \text{if } u = g_L(s) \text{ and } v = g_L(t) \\ 0.01 & \text{otherwise} \end{cases}$
$R_{s,t} = \begin{cases} 1-10^{-5} & \text{if s = t} \\ \frac{10^{-5}}{S-1} & \text{otherwise} \end{cases}$	$R_{\langle s,s'\rangle,\langle t,t'\rangle} = \begin{cases} 1-10^{-5} & \text{if } s = t \text{ and } s' = t' \\ \frac{10^{-5}}{2(S-1)} & \text{if } s \neq s', t = t' \text{ or } s = s', t' \neq t' \\ & \text{or } s = s', t = t', s \neq t \\ 0 & \text{otherwise} \end{cases}$

Table 3.7: Parameter values used for the emission and transmission matrices, E and R in the haploid and diploid implementations of the Forward-Backward algorithm for mosaic reconstruction.

Data:

N = number of sites,

S = list of available founder states

Emission matrix: $E(L, s)$ **Transition table:** $R_{s,t}$

Output: P - the posterior probability matrix, dimensions $N \times S$

Initialisation:

for $s \in S$ **do**

$P^F_{s,(1)} = E(1, s)$

$P^B_{s,(N)} = 1$

Recursion (forward and backward scan):

foreach $L \in N$ **do**

foreach $s \in S$ **do**

$P^F_{s,L} = E(L, s) \times \sum_{t \in S} P^F_{s,(L-1)} \times R_{s,t}$

$P^B_{s,L} = E(L, s) \times \sum_{t \in S} P^B_{s,(N-L+1)} \times R_{s,t}$

Compute posterior probabilitiy matrix P**:**

foreach $L \in N$ **do**

foreach $s \in S$ **do**

$P_{s,L} = P^F_{s,L} \times P^B_{s,L}$

Algorithm 2: The mosaic reconstruction algorithm.

single state with very high probability and that most of the time this is the true state. The haploid algorithm gives very similar answer to the best single-sequence reconstruction as in over $\sim 99\%$ of loci, $P_{\mathsf{true}} \geq 50\%$, i.e. the correct answer gets at least 50% posterior probability. In the diploid version, the same fraction is lower (about 90%) as there is more uncertainty and the algorithm converges to multiple states more often.

Only results with transition probabilities of the order of magnitude of 10^{-5} for a change of state are shown here, but different parameters have been tested, namely $10^{-2}, 10^{-3}, 10^{-4}$. Giving a higher transition probability to a change of state generally increases uncertainty in the data, so

	P_{max}				P_{true}			
Dataset	**Mean** (μ)	**Median** (m)	% **loci** $> 50\%$	% **loci** $> 90\%$	**Mean** (μ)	**Median** (m)	% **loci** $> 50\%$	% **loci** $> 90\%$
SIM.HOMO (300k SNPs)	98.9%	1	99.1%	97.6%	98.7%	1	98.4%	97.5%
SIM.HOMO (1.291k SNPs)	99.5%	1	99.7%	98.9%	99.4%	1	99.3%	98.8%
SIM.HET (300k SNPs)	95.7%	1	97.0%	90.1%	93.1%	1	92.2%	88.8%
SIM.HET (1.291k SNPs)	97.2%	1	97.9%	93.2%	87.5%	1	87.4%	85.6%

Table 3.8: Results of the evaluation of the Forward-Backward algorithm for mosaic reconstruction. Column **Dataset** shows set of simulations used for the evaluation and for P_{max} and P_{true} the columns are: **Mean** (μ) - P_{max}, P_{true} per site, averaged across all simulated genomes, **Median** (m) - median P_{max}, P_{true} per site **% loci** $> 50\%(90\%)$ - mean fraction of sites per genome in which P_{max} and P_{true} are over $50\%(90\%)$

the algorithm converges to the true state less often. Paradoxically, the diploid algorithm performed better with low-coverage than high-coverage data. This should not be the case, and is probably entirely due to parameter choice: the transition probabilities for change of state should be lower, so the algorithm changes state less often. Due to time constraints and because the error rates using my choice of parameters were low, I did not implement the Baum-Welch algorithm, which might have helped optimise parameter estimation in the diploid algorithm.

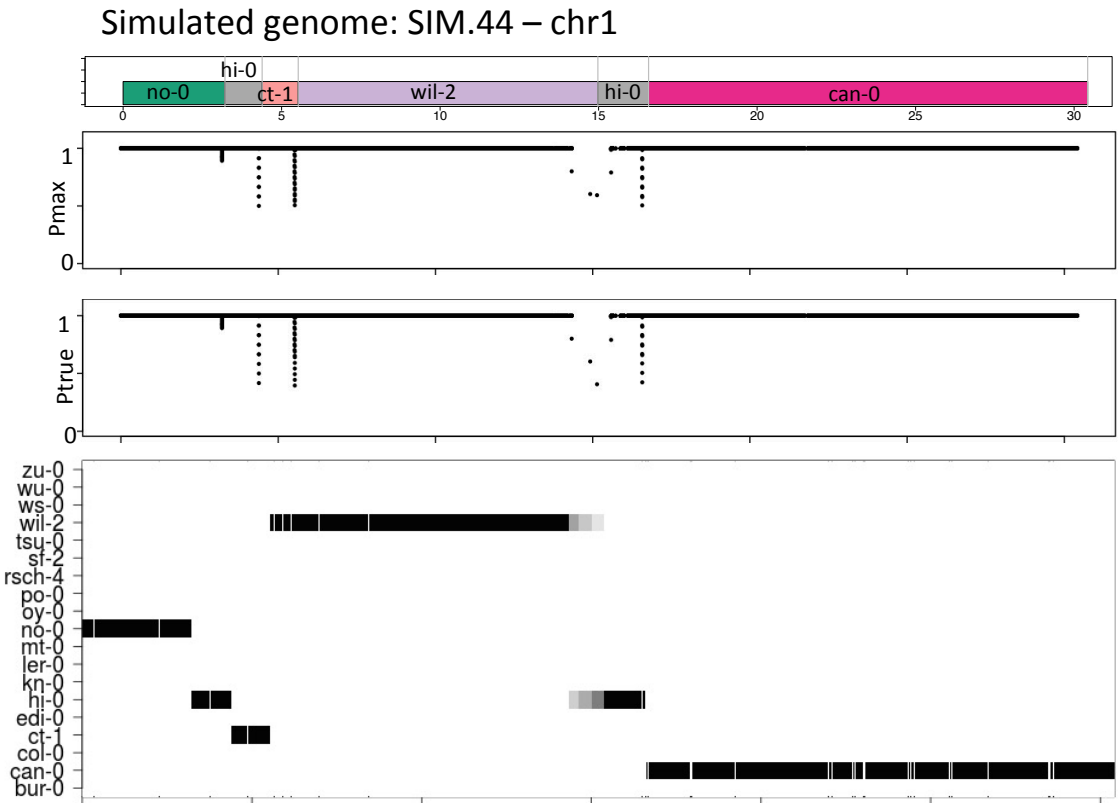


Figure 3.7: Maximum posterior probability (P_{max}), true state posterior P_{true} and visualisation of the posterior matrix of chromosome 1 of the simulated genome SIM.44 from the Forward-Backward algorithm in haploid mode. The x-axis is chromosomal position in Mb. The top panel shows the simulated mosaic and below the chromosome-wide distributions of the maximum posterior P_{max} and of the true haplotype posterior P_{true} are shown. The final panel shows the posterior matrix of the chromosome for the 19 founder states. The posterior probabilities are encoded with shade of grey, with white meaning $P = 0$ and black $P = 1$.

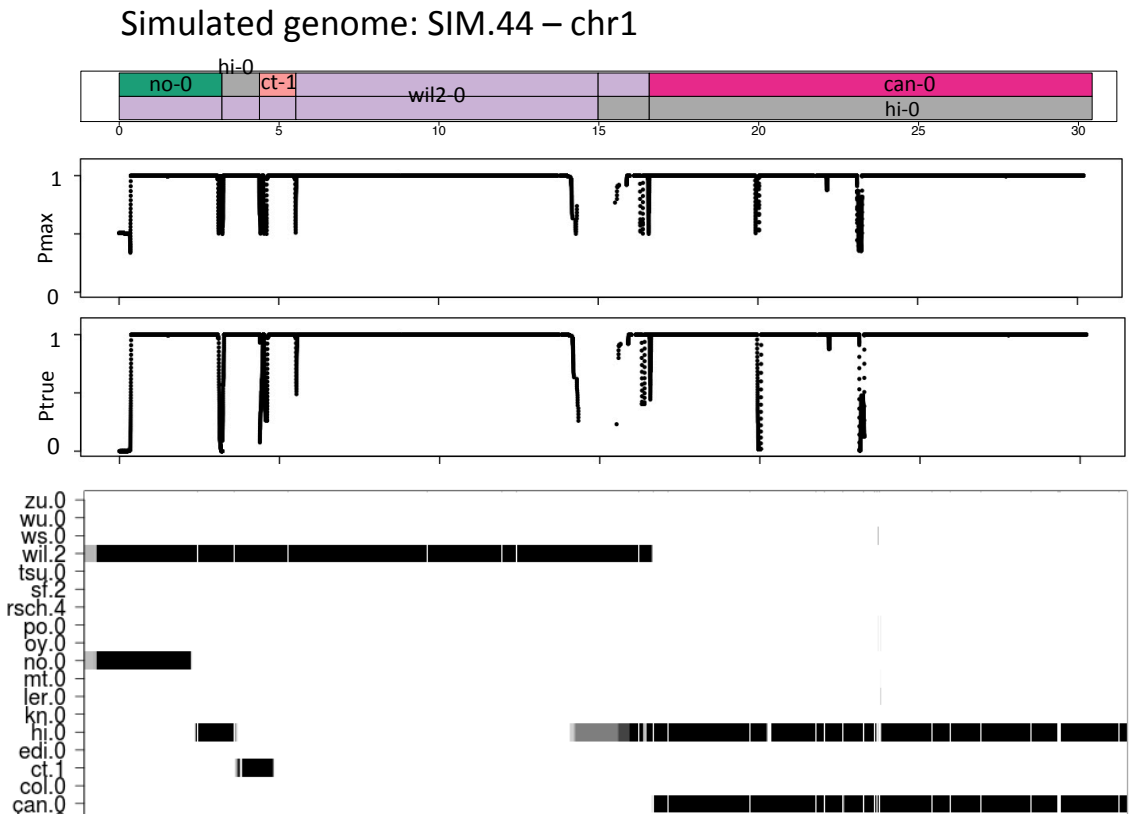


Figure 3.8: Maximum posterior probability (P_{max}), true state posterior P_{true} and visualisation of the posterior matrix of chromosome 1 of the simulated genome SIM.44 from the Forward-Backward algorithm in diploid mode. Here the 190 rows of the matrix have been marginalised into a 19-row matrix in which each row corresponds to a founder. The colour of a founder at one locus corresponds to the sum of posterior probabilities of all states with that founder. Thus if $P_{\langle \text{Col-0, Ler-0} \rangle, i} = 0.99$, $P_{\langle \text{Col-0} \rangle, i} = 0.005$ and $P_{\langle \text{Ler-0} \rangle, i} = 0.005$, in the table $P_{\langle \text{Col-0} \rangle, i} = 0.995$, $P_{\langle \text{Ler-0} \rangle, i} = 0.995$.

3.5 Computational efficiency of the algorithms

Both algorithms were implemented in C and are available from http://mus.well.ox.ac.uk/19genomes/magic.html. The algorithms differed substantially with respect to computational efficiency. Time (based on an AMD 1.4GHz processor with 2048kb of Cache) and memory used by one run of the algorithm for the reconstruction of a single genome is summarised in Table 3.9.

3.6 MAGIC genome mosaics

Based on the previous analysis, I present here the MAGIC genome mosaics using the haploid and diploid best single-sequence algorithm [94] with cost parameter $c = 10$. The results with the Forward-Backward algorithm were nearly identical and not shown.

3.6.1 Mosaics with the haploid algorithm

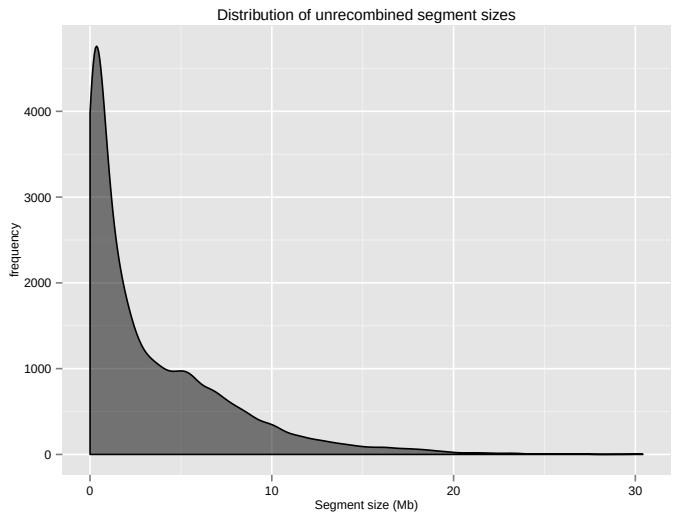
The haploid algorithm predicted a total of $16,700$ segments in 488 lines (i.e. $14,260$ recombination breakpoints). Each line has on average 34 segments, i.e. 29 observable recombination events per line, while the mean segment size is 3.48Mb. Segments are mostly uniformly distributed between founder haplotypes. Founder Po-0 is slightly underrepresented, which is probably due to its similarity to founder Oy-0 . The distribution of segment sizes is presented in Figure 3.9a while the distribution of founder haplotypes in the mosaics is shown in Figure 3.9b. An example mosaic, from line MAGIC.135 is shown in Figure 3.10.

3.6.2 Mosaics with the diploid algorithm

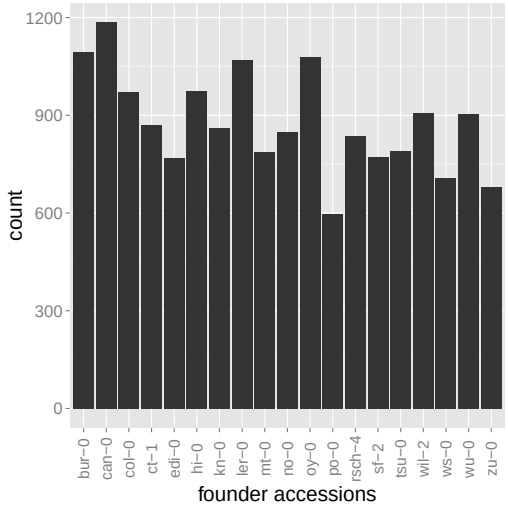
The results of the diploid algorithm are, in most lines, very similar to those of the haploid algorithm. In total $16,012$ segments (i.e. $13,572$ recombinants) are predicted, with average size 3.63Mb. In most lines the diploid algorithm mosaics are identical to the haploid ones, i.e. almost no heterozygosity. Median heterozygosity level per line, defined as fraction of the genome imputed as heterozygous is 0.1%. The very small levels of heterozygosity likely are false positives inserted over sequencing errors. However, 24 lines comprise of more than 5% heterozygous haplotypes, while

	300K SNPs (low coverage)		1.291K SNPs (high coverage)	
Algorithm	**Time (sec)**	**Mem (Gb)**	**Time (sec)**	**Mem (Gb)**
Haploid best-single recon	4.46	2.95	29.99	2.98
Diploid best-single recon	118.96	3.08	793.51	4.32
Haploid FB	9.37	2.96	33.82	2.98
Diploid FB	166.99	3.08	1195.71	4.32

Table 3.9: Computational performance of reconstruction algorithms with respect to time and memory used. Time refers to the main function call, without including reading in the data, which takes about 25 seconds (to read founder genotypes and allele calls for one sample). Memory refers to total heap memory, computed using valgrind.



(a) Frequency distribution of unrecombined haplotype segments in the mosaics (generated with the GATK variant calls). The x-axis shows the segment size in Mb and the y-axis the frequency.



(b) Distribution of founder haplotypes in the revised mosaics. x-axis lists the founder accessions, y-axis gives counts of numbers of segments assigned to the corresponding founder.

Figure 3.9: MAGIC mosaic statistics

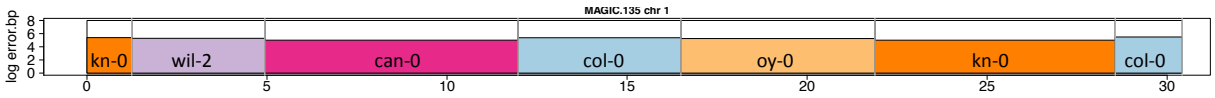


Figure 3.10: Genome mosaic of line MAGIC.135 chromosome 3 indicating a chromosome with seven haplotype segments. x-axis: genome position (Mb). Coloured segments indicate genomic regions descended from the corresponding founder. y-axis: negative log 10 discrepancy rate between called SNPs from high-coverage sequence data and alleles predicted from the imputed mosaics, i.e. the genotype error of Section divided by the size of each segment.

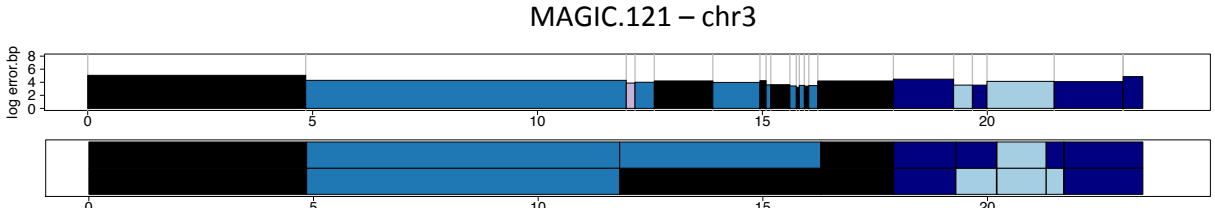


Figure 3.11: Mosaics of chromosome 3 reconstructed by haploid and diploid reconstruction. The breakpoint clustering the region from $12 - 17$Mb can be explained by heterozygosity, and is not present after allowing diploid states in reconstruction.

8 (MAGIC.113, MAGIC.13, MAGIC.149, MAGIC.26, MAGIC.329, MAGIC.367, MAGIC.426 and MAGIC.433) are over 50% heterozygous. The lines with very high heterozygosity levels possibly represent samples from earlier selfing generations that were accidentally sequenced. The remaining 18 have small heterozygous segments in some of their chromosomes. This residual heterozygosity could be present by chance after several generations of selfing or because of hybrid incompatibilities between specific pairs of SNPs which may prevent certain regions from fixing.

Heterozygous regions are reconstructed by the haploid algorithm as clusters of recombinants with two oscillating founder haplotype states, but can be reconstructed correctly by the diploid algorithm 3.11. However, because simulations have shown the haploid algorithm predictions to be free of false positives and to also predict recombination breakpoint more accurately, I chose to use the haploid mosaics throughout the analysis, despite the few areas of residual heterozygosity. In the computation of recombination hotspots of Chapter 6, residual heterozygosity was taken into consideration, so false breakpoints inserted by the algorithm over heterozygous regions were masked out.

3.7 Conclusion

Many algorithms that estimate haplotypes in genotype data, without prior knowledge of the population history, are currently available [71, 122], however in MAGIC the problem is simpler as the founder genomes are known to high precision and can be used for imputation. This chapter has analysed two algorithms for mosaic reconstruction in MAGIC. The first algorithm recovers the best annotation of founder haplotypes at each locus by dynamic programming and, thus, is similar to finding the Viterbi path in a Hidden Markov Model. Unlike Viterbi, the algorithm is not stochastic as it uses additive scores and a cost function for change of states instead of transition probabilities, but it can be perceived as a Viterbi run in a logarithmic scale. The second algorithm is the Forward-Backward Hidden Markov Model algorithm. Each algorithm can be run in two modes: haploid and diploid; the diploid version can deal with heterozygous genomes.

To evaluate the algorithms I simulated 50 genome mosaics as well as allele calls from low and high-coverage sequencing data. Both algorithms accurately predict the correct haplotype at the majority of loci, although the haploid algorithm is more precise. Nevertheless, the diploid algorithm performs sufficiently well, particularly when a genome comprises large haplotype segments, so it would be suitable for reconstructing haplotypes in other synthetic populations involving heterozygous data, such as F_2 populations. Because the diploid algorithm was shown to be sensitive to changes of the cost parameter, the optimal c values suggested here may need to be reconsidered with different data.

The evaluation results suggest that mosaic reconstruction is possible with low-coverage data for both haploid and diploid genomes. Simulations showed that the mosaics produced by both the Viterbi-based reconstruction and the Forward backward algorithm accurately predicted the underlying haplotypes with low-coverage data, with the exception of loci in which multiple founders are highly similar. High-coverage data slightly increase the prediction accuracy of the algorithm, but the improvement is negligible. Furthermore, use of low-coverage data is much more efficient computationally in time and space. Thus, low-coverage sequencing and imputation can efficiently and accurately determine the DNA sequence of every genome in the MAGIC population.

Chapter 4

Detection of Structural Variants by genetic mapping

4.1 Introduction

By Structural Variants (SVs) I mean alterations in chromosomal structure with respect to the reference genome, other than single nucleotide substitutions. Thus, under this definition, SVs can range in length from a few bases to several millions (Mb) and include insertions, deletions, copy number variations, inversions or translocations. I shall focus in this chapter on longer (kb-sized) SVs.

Despite the fact that structural genomic variation can have important phenotypic effects in eukaryotes, SVs that segregate in populations are currently under-reported. As I discuss below, methods that identify SVs based on signatures associated with short-read sequence data suffer from high false negative rates. This is particularly the case in respect to long-range SVs such as translocations, because their signatures are hard to distinguish from sequencing errors; translocations are for the most part ignored in studies of complex traits.

This chapter presents a general approach that detects SVs by analysing low-coverage sequence collected in many individuals in a population. I treat signatures of SVs, measured as counts of anomalously mapped reads, as quantitative traits that I map to identify loci containing genetic

variation correlated with read-mapping anomalies. The method can distinguish between short- and long-range SVs, and in the MAGIC lines I use it to identify 4,898 short-range and 1,604 long-range SVs at 1% false discovery rate. I confirm these results computationally using paired-end reads and *de novo* assembly contigs in the 19 founders, and also experimentally confirm by PCR in the case of 31 translocations and 8 inversions (success rate 83%). I estimate that 34.2% of the *Arabidopsis* genome is within 10kb of a structural variant in the 19 founders, with long-range SVs affecting 9.9% of the genome. Using gene expression data from 200 MAGIC lines, I show that SVs have a significant impact on silencing gene expression in genes that are disrupted by SV boundaries or that are transposed or inverted. Finally, I also show that some SVs lie within QTLs of physiological phenotypes and can explain a very large fraction of the phenotypic variation, consistent with, but not proof that, the SV being causal.

4.2 Structural Genetic Variation

Structural variants can contribute significantly to phenotypes, for example in conditions such as Down syndrome [53], mental retardation [25], autism [101] and schizophrenia [103]. Many cancers are related to somatically acquired rearrangements [77, 118]. Long-range SVs such as translocations, albeit less common can have very large phenotypic effects in many disease phenotypes [33, 41].

Early studies of structural variation used microscopy to identify inherited *de novo* structural variants responsible for severe disorders, mainly rare aneuploidies or large-scale (over 3Mb) chromosomal rearrangements that are visible under a microscope. Later studies used array Comparative Genomic Hybridisation (aCGH), which identified copy number variations by hybridisation of a genomic sequence to arrays of oligonucleotides ($>$ 50 bp) from a reference genome, with resolution of about 100kb. These studies revealed that structural variants are much more common than expected and that at least 3% of the human genome is copy number variant [26]. However, aCGH studies had severe reproducibility issues [111, 1] as use of different arrays led to inconsistent results. For example, estimates for the extent of copy number variations in mouse inbred strains using different CGH arrays ranged from 3% [16] to 10.7% [45].

4.2.1 Next-generation sequence studies of local structural variation in eukaryotes

Next generation sequencing technologies have greatly improved SV mapping resolution and prediction accuracy and, consequently, have helped discover a larger number of SVs [93]. Studies of local structural variation using paired-end sequencing data have been completed in several species. In humans, a study of 185 human genomes identified 22,025 deletions, 5,499 insertions and 501 tandem duplications, affecting 10,995 genes [93] and also identified 'SV hotspots', i.e. genomic loci with higher than expected concentration of SVs. The same study investigated the mechanism of SV formation and concluded that non-allelic homologous recombination sites, characterised by high sequence similarity, are associated with many SVs. Furthermore, the second phase of the 1000 genomes project, which included 1,092 genomes from 14 populations, reported 14,000 deletions and 1.4M biallelic indels [3]. A study on 96 deep-sequenced Asian Malay genomes identified 1.6M segregating indels smaller than 50bp and 34,000 larger deletions. Of the small indels only 0.1% were found to affect genes [134]. The Database of Genomic Variants Archive has catalogued 2.5M human SVs from both SNP array and next-generation sequencing studies covering 202,431 regions. Of these 1,149 are inversions and the rest are indels [85]. A technique that is slightly related to the method proposed here has been used to complete maps of the human genome from admixed samples of West African and European origin [37]. That study used SNP arrays to compute linkage disequilibrium between large unmapped contigs (assembled from individuals from different populations) and the rest of the reference genome, to attempt to place these contigs. In this way it was possible to identify novel euchromatic islands within heterochromatin.

In the mouse, a combination of computational and experimental analyses were used to map SVs in 17 murine laboratory strains, 13 classical and 4 wild-derived [137] (wild-derived strains are about four times as divergent, in terms of SNPs, from the C57BL/6J reference genome than are classical strains). The study combined computational SV predictions with experimental confirmation by PCR and manual inspection. In total they mapped 710k SVs larger than 100bp, and estimated another 49k smaller SVs, affecting 1.2% of the genome in classical strains and more (3.7%) in wild-derived strains. 98% of all these SVs were indels, 1.4% were copy number gains and inversions,

while the remaining 0.6% were complex, i.e. adjacent simple SVs of different types that abut each other. Most (54%) SVs were associated with transposons, while another 15% by tandem repeats and 2% by pseudogenes. They also investigated the impact of SVs on gene expression variation and found a modest contribution of up to 10% which significantly increased when the SV overlapped with over 50% of a gene: in those cases SVs explained over 25% of the variance in the gene's expression. False negative rates for detecting SVs were high, estimated at up to 24% for simple and 54% for complex structural variants.

In animals the germ line is distinct from the soma, so mitotic structural variants (i.e those arising in the non-reproductive tissues during the life of an individual) are not inherited. In contrast, in plants there is the potential for somatically acquired structural variants to be transmitted. Thus, in *A. thaliana*, high temperature stress increases the number of inherited copy number variants [30]. Moreover, genome size varies greatly between accessions, due to massive copy number variations in rDNA [81]. To date, no comprehensive catalogue of structural variants has been reported in *A. thaliana*, although sequence analysis of the 19 MAGIC founders [36] identified over $22,000$ indels longer than 100bp segregating in 19 accessions, and another study that sequenced 80 accessions [17] identified over $174,789$ deletions longer than 20bp and $1,059$ copy number variants longer than 1kb. Both studies are likely to have under-reported the true degree of structural variation in *Arabidopsis*. In general, surveys of structural variation in plants are less complete than in animals, despite the fact that copy number variants for specific genes have been linked to important traits such as flowering time, plant height and resistance [130]. A recent study discovered an 80kb translocation in *A. thaliana* and speculated that a large number of undetected translocations lay in regions of apparent heterozygosity [132].

4.2.2 Detection of Structural Variation using next-generation sequencing

Computational methods that identify structural variants from short-read sequence data look for signatures in the alignment of paired-end reads from an individual to the reference genome. Such signatures include reads with split alignments, anomalous insert sizes or increased read coverage (Figure 1.1). Figures 4.1 and 4.2 show examples of paired-end read alignments from the MAGIC

founder Ler-0 that display signatures of structural variants. Figure 4.1 shows read alignments with large insert sizes and zero read coverage that indicate a deletion, while Figure 4.2 shows alignments with a mixture of anomaly signatures, whose origins are hard to resolve.

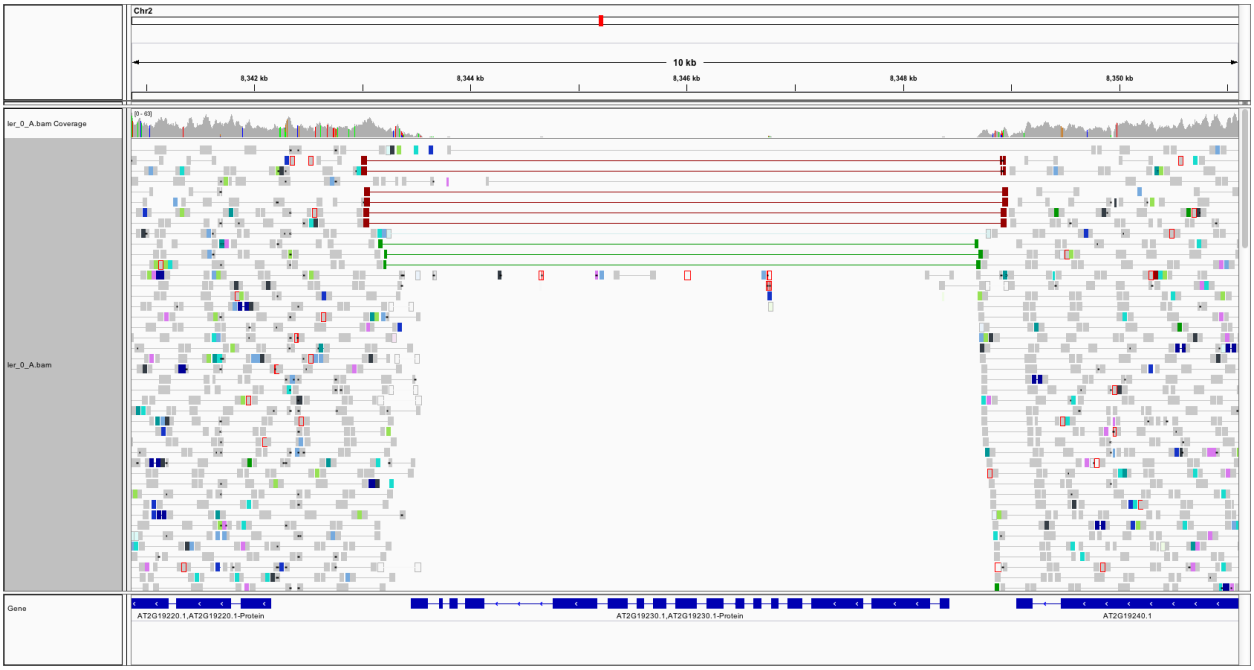


Figure 4.1: Read pair alignments from chromosome 2 of Ler-0 mapped to Col-0 (TAIR 10) over a deleted region. Read pairs flanking the deleted locus have a larger than normal insert size and are colour-coded in red. There are also three read-pairs with a strand error (i.e. both reads are mapped to the same strand), which are coloured in green; indicating that the sequence near one of the boundaries of the deletion may also be inverted. Figure is a screen-shot from IGV [123], displaying sequence data from [36].

One approach is to detect SVs from split read alignments [139, 92, 86]. Pindel [139] for example, searches for reads that span deletions, i.e. split alignments in which two fragments flank the deletion breakpoint. First, it selects all read pairs with only one read mapped to the genome - this read is used as an anchor point which can indicate the approximate location of its pair, which is unmapped. If two fragments of the unmapped read can be mapped separately, then they mark the breakpoints of a deletion. If the two fragments of the unmapped reads map to opposite strands, then they indicate an inversion. At a duplication, both fragments have the same orientation, but in a different order than expected (Figure 4.3). If the alignment of fragments fails, Pindel looks for insertions. Thus, the unmapped read is aligned against other unmapped reads in the opposite orientation: if a match

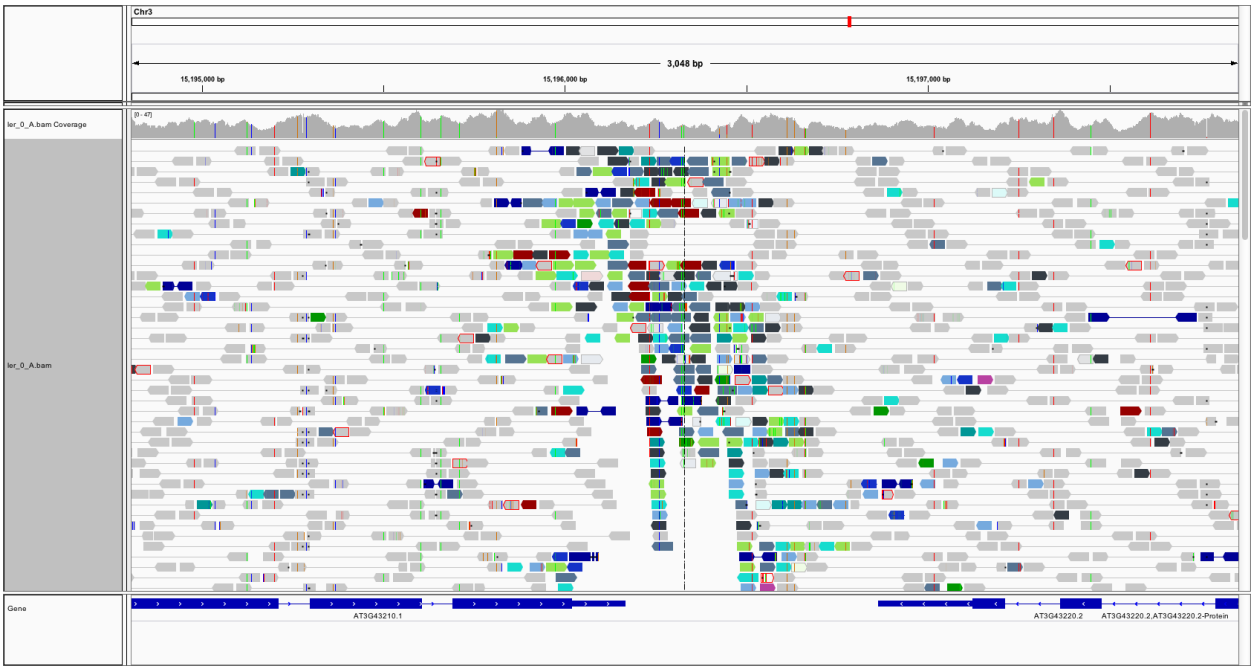


Figure 4.2: Paired-end alignments with a mixture of anomaly signatures. The region is taken from chromosome 3 of Ler-0 (from [36]) and contains reads with large insert size (red), reads with short insert size (dark blue), reads whose pairs are mapped to different chromosomes (colour-coded in black, navy, green, purple, turquoise and pale blue for chromosomes $1, 2, 4, 5$ and chloroplast respectively) and unpaired reads (grey with red outline).

is found and its pair is fully mapped to the reference, then the mapped read defines approximately the breakpoint of an insertion, otherwise the procedure is repeated until an unmapped read with a mapped pair is detected.

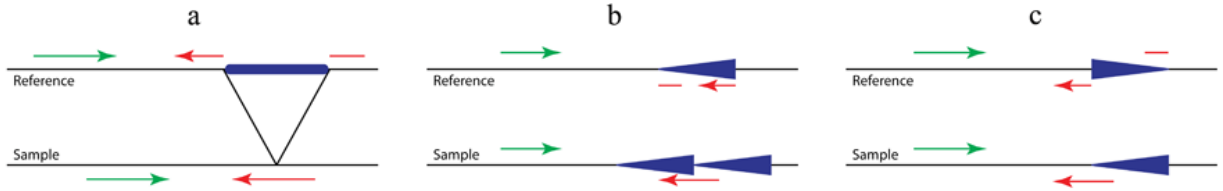


Figure 4.3: Detection of a deletion (a), tandem duplication (b) and inversion (c) by Pindel. The first row shows the alignment of reads to a reference and the second the true alignment, to the sample genome from which they were collected. The green read is the anchor point and the red is its unmapped pair, which is fragmented and re-aligned. Figure taken from `http://gmt.genome.wustl.edu/packages/pindel/index.html`

Another approach is to use insert sizes and read-pair orientation [21, 115, 65]. BreakDancer

[21] considers only read pairs with both pairs mapped to the reference and maps structural variants by looking at insert size or read orientation irregularities. Large insert sizes are associated with deletions, small insert sizes with insertions and wrong relative orientation (i.e. both reads in a pair map to the same strand) to inversions. Long-range rearrangements can also be detected by looking for reads mapping to remote genomic loci. Figure 4.4 summarises the main ideas.

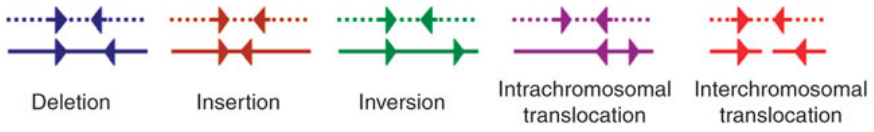


Figure 4.4: Structural variants identified by regularities in insert size and orientation of paired-end reads in BreakDancer (figure from [21])

Other algorithms predict SVs analysing read coverage fluctuations [114, 22]. The CnD algorithm [114] uses variations in read coverage to map copy-number variations, i.e. duplications, triplications and deletions. When copy-number gains are present in a sample with respect to the reference, reads from all copies are mapped to a single reference locus, increasing read coverage and, in inbred genomes, causing apparent heterozygosity. CnD models each genome as a Hidden Markov Model with hidden copy number states such as $1-$fold gain, $2-$fold gain, $1-$fold loss or deletion (loss of all copies). Each emit different distributions of read coverage and heterozygosity rates, estimated by simulation. For example, heterozygosity rates follow a Poisson distribution in normal or copy number loss states and a negative binomial in copy number gain states. CnD handles only local copy number gains, so it cannot map insertions, inversions or translocations and works only in homozygous inbred genomes.

The above algorithms have been used to map structural variants in several studies, and similar approaches are incorporated in many state-of-the-art read mapping [83, 75, 36] and variant calling [89, 74] programs. However, they cannot fully capture the landscape of structural variation in a population. Next-generation sequencing data have high levels of noise. SVs are only one source of read mapping irregularities, along with transposons, repeated sequence, centromeric regions or sequencing errors unrelated to structural variation. Thus, SV signatures cannot always be unambiguously attributed to SVs. Moreover, most read-mapping algorithms assign low mapping-quality scores to reads related to structural variants, because they appear unreliable, so potentially

relevant information is lost. As a result, even in studies that predict short-range structural variants a high number of false negatives is reported [139, 21]. This is especially true for complex structural variants displaying a mixture of adjacent simple SVs [137]. Finally, perhaps the main limitation of these algorithms is that they do not detect long-range structural variants, with the exception of BreakDancer, that nevertheless had very low validation scores for translocations. Therefore, in the case of translocations further experimental confirmation such as PCR sequencing is required to resolve true signal from noise.

4.3 Mapping structural variants segregating in populations as quantitative traits

Whilst SVs manifest themselves as anomalies in the alignment of next-generation sequencing reads, only a fraction of SVs can be completely resolved solely by analysing read-pair alignments. The alignments shown in Figure 4.2 for example, include read pairs mapping to all five *Arabidopsis* chromosomes. A likely cause of such alignments is translocations, but it is difficult to predict the chromosome to which the locus has been transposed from the alignments. Furthermore, noisy alignments such as that of Figure 4.2 are not unusual, and the presence of structural variation is not their only possible interpretation: alternatives include sequence repeats, poorly-assembled reference genome (particularly around centromeres), transposable element genes, contamination from other genomes, and poor quality sequence data.

In this chapter I present a novel approach for the identification of SVs which uses population genetics as additional information to resolve complex read-pair alignments. I treat signatures of SVs, measured by sequencing a population, as traits that can be mapped genetically as quantitative trait loci (QTLs). In this way I identify and distinguish between short- and long-range SVs , corresponding to cis and trans QTLs respectively. This work generates the first catalogue of long-range structural variation in a eukaryote.

I define the location of a QTL as the *sink* locus associated with trait variation at a *source* locus. The basic principle of the method is best illustrated with regard to translocations. A

translocation links two loci, the source and sink, potentially on different chromosomes. Suppose the translocation is ancient, arising once in an individual, when a DNA segment at the source locus A, was translocated to the sink locus B. The translocation is embedded within a specific haplotype b at the sink. Recombination and segregation over subsequent generations will then randomise the relationship between the haplotypes at the two loci. Only present-day individuals who carry the b haplotype at the sink carry the translocation, regardless of their haplotype at the source.

To find these events given a sequenced sample of individuals, we align the reads to the reference genome using a standard read mapper [75, 83] and look for the signature of a translocation at the source locus, for example an excess of unpaired reads mapping to the source. Only individuals that carry the b sink haplotype will show an excess of unpaired reads. Therefore we have transformed the problem of identifying a translocation to that of mapping a quantitative trait locus (QTL). Here the trait is the numerical measure of read mapping anomalies at the source locus, and the QTL position identifies the sink. Because the positions of the sources and sinks are not known *a priori*, I divide the genome into segments, construct a trait for each segment, based on reads with anomalous mappings within the segment, and then map each trait analogously to an expression QTL study; trans-QTLs correspond to long-range structural variants and cis-QTLs to local structural variants. See Figure 4.5.

This general idea extends to other types of SVs. Thus we can identify short and long-range inversions (translocations where the inserted locus is also inverted) by defining the number of reads aligning to the same strand as a trait. Similarly, short and long-range copy number variations (i.e. a translocation where the source locus is duplicated rather than deleted) result in elevated read coverage and in a mixture of read pairs some of which map properly to the reference genome at the source whilst others link the source to the sink. The method only identifies relatively common SVs, segregating in the population, and is blind to *de novo* events. However, by definition common SVs contribute more than rare *de novo* SVs to overall phenotypic variation.

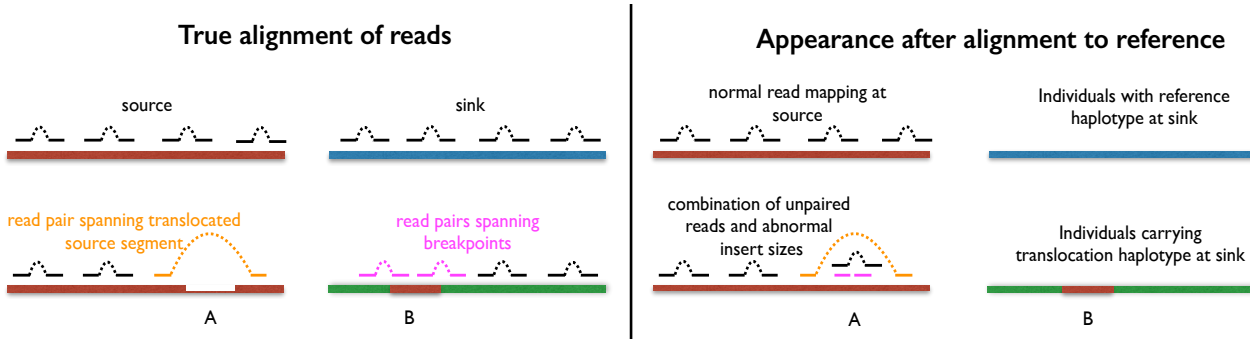


Figure 4.5: The effect of a translocation on short-read mapping. Chromosomes are horizontal bars and read pairs horizontal lines linked by dotted curves. Left panel: alignment of paired-end reads to chromosomes labelled "source" and "sink" when the true genomes are used. Right panel: appearance when alignment is to the reference instead. The upper pair of chromosomes in each panel shows a normal individual without a translocation; at the sink it carries a normal blue haplotype. The lower chromosome pair in the left panel shows an individual in which a source segment A is translocated to sink locus B, on a green b sink haplotype. The lower chromosome pair in the right panel shows an individual carrying the normal source haplotype and a green b sink haplotype with the translocation. When reads are aligned to the reference then, irrespective of the haplotype at the source locus, if the sink is carrying the green b haplotype, read anomalies occur at the source. By counting the number of anomalous reads at the source in each individual we construct a quantitative trait that is large if and only if the individual carries the green b sink haplotype.

4.4 SV signatures as quantitative traits

I divided the TAIR10 (Col-0) reference genome ($\simeq$ 120 Mb) into 11,915 abutting 10 kb segments. Within each segment I computed six measures of anomalously mapped reads that are signatures of SVs: high coverage, large insert size, unpaired reads, pairs mapping to the same strand, large insert size or unpaired reads and improperly paired reads (combining all of the above anomalies). A mapped read pair r is written using the notation:

$$r = \langle r_1 : (\mathsf{chr}_1, p_1, s_1), r_2 : (\mathsf{chr}_2, p_2, s_2) \rangle \tag{4.1}$$

where r_1, r_2 are the paired reads, $\mathsf{chr}_1, \mathsf{chr}_2$ are the chromosomes to which the reads map, p_1, p_2 are their start positions (in bp), and s_1, s_2 the DNA strands (± 1). I assume that the read length is constant throughout the sample. Let S be the vector of insert sizes of all read pairs in the sample and μ_S the median insert size for those pairs mapped to the same chromosome.

In what follows, I shall denote the set of reads that have a property X as R_X. The size of the set is $\rho_X = |R_X|$. Where X is a numerical value, I also define $\mu_X, \mathrm{m}_X, \sigma_X^2, \mathrm{IQR}_X$ as the mean, median, variance and inter-quartile range of X. I define a read pair r as normal if:

$$\mathsf{chr}_1 = \mathsf{chr}_2 \neq 0 \tag{4.2}$$

$$\mid p_2 - p_1 \mid \leq \mathrm{m}_S \pm 5\mathrm{IQR}_S \tag{4.3}$$

$$s_1 \neq s_2 \tag{4.4}$$

where S is the insert size of the pair. Equation 4.2 means that both reads in a normal pair are mapped to the same chromosome and none of the two is unmapped (unmapped reads are symbolically assigned to chromosome 0). Equation 4.3 specifies a normal range for the insert size which is up to 5 times the interquartile range larger than the median. Finally, Equation 4.4 requires that a pair should map to opposite DNA strands. The set of reads mapped within a 10-kb segment l with coordinates $[B_{\mathsf{chr}}, B_{\mathsf{start}}, B_{\mathsf{end}}]$ and length $B = B_{\mathsf{end}} - B_{\mathsf{start}} + 1$ is:

$$\begin{aligned} R_l = &\{r_1 \in R : \mathsf{chr}_1 = B_{\mathsf{chr}}, p_1 \geq B_{\mathsf{start}}, p_1 \leq B_{\mathsf{end}}\} \cap \\ &\{r_2 \in R : \mathsf{chr}_2 = B_{\mathsf{chr}}, p_2 \geq B_{\mathsf{start}}, p_2 \leq B_{\mathsf{end}}\} \end{aligned} \tag{4.5}$$

where R is the total number of reads in the genome. The expected read coverage of R_l is defined as:

$$\mathbb{E}[\rho_l] = \frac{|R| \times B}{L} \tag{4.6}$$

where L the length of the genome.

Based on these definitions I characterised six read anomaly measures. Their values were computed in segments with coordinates $[B_{\mathsf{chr}}, B_{\mathsf{start}}, B_{\mathsf{end}}]$. With a slight abuse of notation, they are defined as follows:

1. **High coverage:** Number ρ_{hc} of reads exceeding the expected read coverage of the locus

$$\rho_{hc} = \rho_l - 1.5\mathbb{E}[\rho_l] \tag{4.7}$$

2. **Unpaired reads**: Number of reads ρ_u whose pair is unmapped, defined as

$$R_u = \{r_1 \in R_l : \mathsf{chr}_2 = 0\} \cup \{r_2 \in R_l : \mathsf{chr}_1 = 0\} \tag{4.8}$$

$$\rho_u = |R_u| \tag{4.9}$$

3. **Number of read pairs ρ_s on the same strand :**

$$R_s = \{r_1 \in R_l : s_1 = s_2\} \cup \{r_2 \in R_l : s_2 = s_1\} \tag{4.10}$$

$$\rho_s = |R_s| \tag{4.11}$$

4. **Number of reads ρ_i with large insert size**: Read pairs whose insert size exceeds the normal range, or reads whose pair is mapped to a different chromosome

$$\begin{aligned} R_i =& \{r_1 \in R_l :\mid p_2 - p_1 \mid > \mathrm{m}_S \pm 5IQR(S)\} \cup \{r_1 \in R_l : \mathsf{chr}_1 \neq \mathsf{chr}_2 \neq 0\} \cup \\ & \{r_2 \in R_l :\mid p_2 - p_1 \mid > \mathrm{m}_S \pm 5IQR(S)\} \cup \{r_2 \in R_l : \mathsf{chr}_1 \neq \mathsf{chr}_2 \neq 0\} \end{aligned} \tag{4.12}$$

$$\rho_i = |R_i| \tag{4.13}$$

5. **Number of reads ρ_{ui} unpaired or with large insert size**

$$\rho_{ui} = \rho_u + \rho_i \tag{4.14}$$

6. **Number ρ_{uis} of improperly paired reads**

$$\rho_{uis} = \rho_u + \rho_i + \rho_s \tag{4.15}$$

The last two traits are combinations of other traits as often SVs have multiple anomaly signatures, so merging them is expected to increase power.

Each type of read pair mapping anomaly was measured in each of the 11,915 segments, so in total 71,490 traits were determined. The genome-wide distribution of trait variances for improperly paired reads ρ_{ius} is shown in Figure 4.6 and for the other five types of anomaly in Appendix A. The figure also marks loci at which sink QTLs are detected, so are probably structurally variant (the QTL mapping is described in Section 4.5). Many source loci, distributed throughout the genome but with a higher concentration in centromeres, have distinctly elevated trait variance. I show these loci are also more likely to have associated sink QTLs, indicating that read anomaly variance can indicate likely SV loci.

4.5 Mapping structural variants

I treated the set of traits analogously to a gene expression eQTL study, by performing a genome scan for each trait. Association was tested by fitting trait vectors to the imputed ancestral haplotype at each locus in the 488 MAGIC genome mosaics. In combination, the mosaics partitioned the genome into 16,700 founder haplotype segments, such that within each segment the ancestral state of all lines was unchanged, so the whole genome could be scanned using 16,700 haplotype tests each with 18 degrees of freedom (Figure 4.7). I define a quantitative trait as the number of anomalous reads y_{Ai} at source locus A in line i. At every locus L (from the 16,700 mosaic segments) I fit a linear model, described by:

$$y_{A,i} = \mu_A + \sum_{s \in S} X_{Li}(s)\beta_{AL}(s) + e_i \tag{4.16}$$

μ_A is the average trait value at the source, $X_{Li}(s)$ is a binary variable indicating whether line i carries haplotype s at L, $\beta_{AL}(s)$ is the founder effect and e_i the (Normally distributed) error. The founder effects $\beta_{AL}(s)$ are estimated by standard one-way Analysis of Variance (ANOVA). The haplotype test implicitly assigns an SV allele to each of the founders, based on the estimated haplotype effects $\hat{\beta}_{AL}(s)$, thereby predicting which founders carried the SV when the population

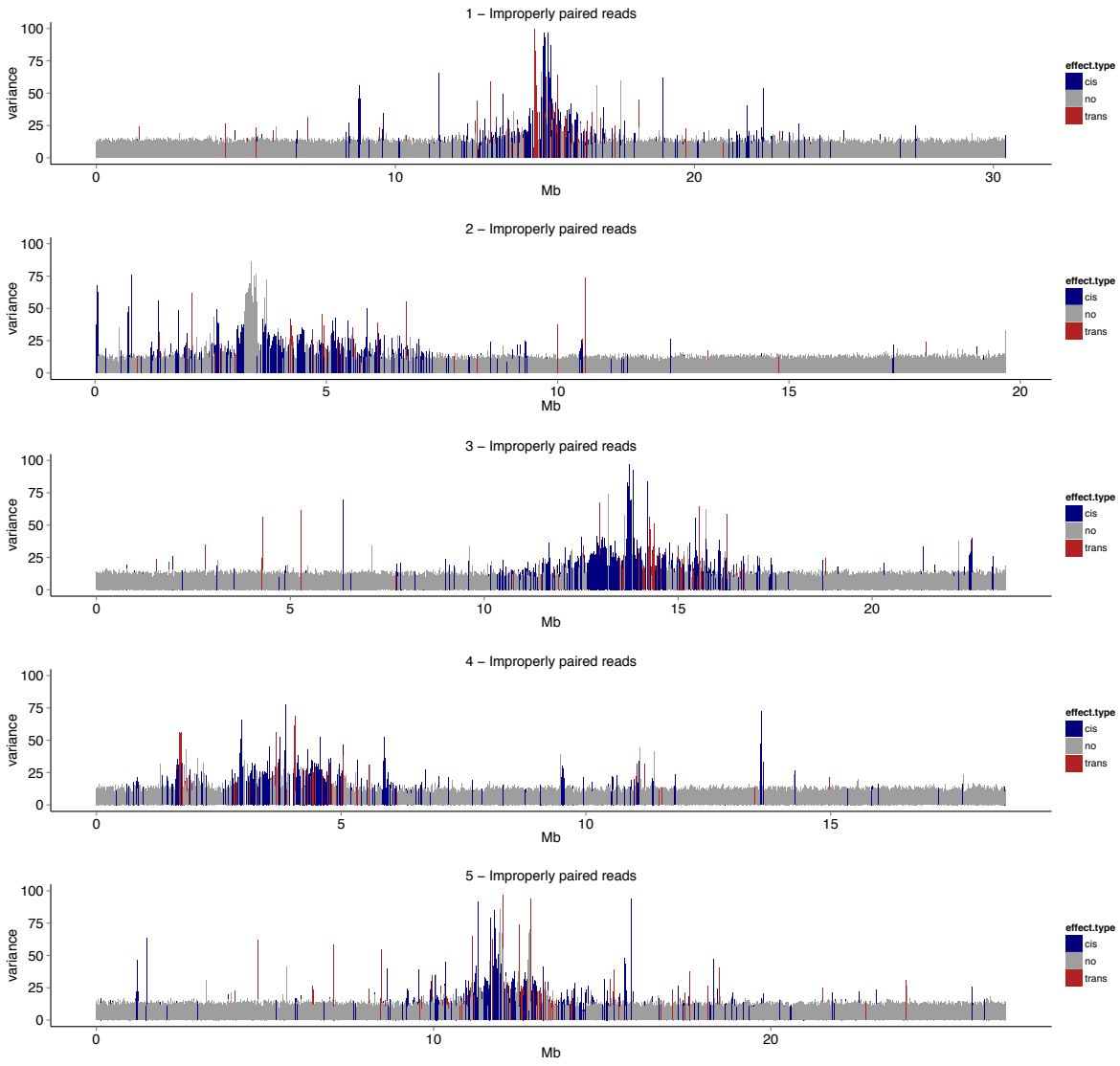


Figure 4.6: Genome-wide distribution of the phenotypic variance for the trait improperly paired reads. The x-axis shows genomic position and the y-axis variance of trait value (scaled by the mean). Each bar corresponds to a 10kb segment for which the trait was measured. Source segments that had a sink QTL for improperly paired reads are coloured in blue and red, for cis and trans-QTLs, respectively.

originated. Locus L is associated with the trait value at source A if founder effects differ amongst

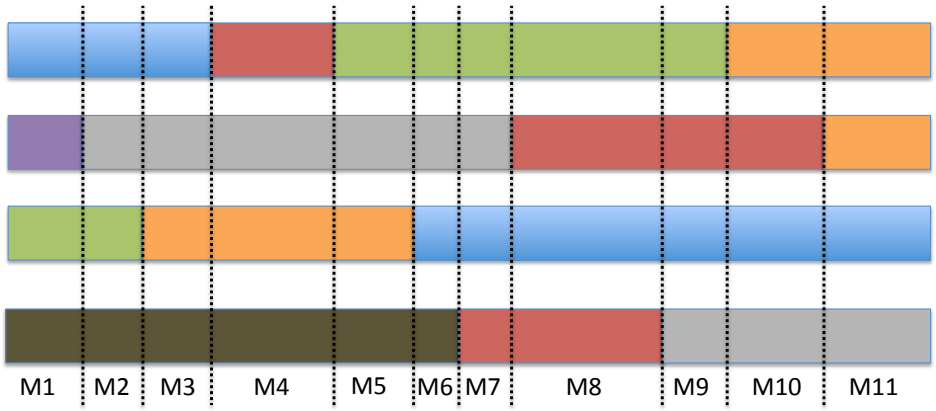


Figure 4.7: A sketch of the partitioning of the genome into intervals based on the reconstructed mosaics. Each horizontal coloured bar represents one genome mosaic: each colour corresponds to a different founder haplotype. The genome is split into loci $M1 \dots M11$, whose boundaries are defined by union of the positions of all mosaic breakpoints in the 488 sequenced MAGIC genomes.

the founders, i.e. if the following null hypothesis is rejected:

$$H_0(L) : \beta_{AL}(s) = 0 \, \forall \, s \tag{4.17}$$

4.6 Analysis of genome-wide significance from multiple genome scans

I call the location of a sink locus associated with an SV trait an SV QTL. The genome scan for the source locus A returns p-value P_{AL} for each of the $16,700$ scanned loci L indicating the association of founder haplotypes at L with the structural variant trait at A. To decide which traits had an association with a sink (i.e. an SV QTL) I set out to find a universal significance threshold for the study. For this I first selected as a candidate SV QTLs for each of the $71,490$ traits, the locus with maximum genome-wide negative logarithm of P_{AL}, such that:

$$\lambda_A = \max(-\log_{10}(P_{AL})). \tag{4.18}$$

referred to as the trait logP. I reasoned that if the genome-wide maximum logP λ_A was not significant after correction for multiple testing at the level of individual scans ($16,700$ tests) and for the number of traits ($71,490$) then it could be ignored. On the other hand if it was significant then there might be more than one such associated SV QTL per trait.

Most (57.9%) λ_A values were low (below 5) as expected, based on a crude Bonferroni correction for the number of tests per scan ($16,700 = 10^{4.22}$). Across the entire experiment the Bonferroni threshold, assuming all tests are independent (which is not the case - they are correlated due to linkage disequilibrium and because some traits are combinations of others), would be ($16,700 \times 71,490$) $= 10^{9.08}$. In fact, there was a large excess of high λ_A values, in the range $50 - 150$. However, high λ_A values do not necessarily indicate a true SV QTL. Examples are shown in Figure 4.8, which illustrates the distribution of same-strand reads in three representative traits. The figure also shows quantile-quantile (QQ) plots comparing the distribution of the negative logarithm of $P_{A,L}$ at all loci to the exponential distribution with mean 1, which is the theoretical distribution of independent log p-values under the null hypothesis. Finally, Manhattan plots showing the genomic positions of associations are shown. The trait a in Figure 4.8a is a randomly selected trait which most probably does not have a SV QTL. The trait distribution and the corresponding QQ and manhattan plots show that although a few MAGIC genomes have excessive numbers of anomalous reads at this source locus, no association of these anomalies with founder haplotypes was detected. For this trait $\lambda_a = 4.5$. The trait b in Figure 4.8b, in contrast, has one outlier with a much higher number of anomalous reads than the rest of the population. There is a small number of highly associated loci, with the maximum $\lambda_b = 57$; it is likely that the outlier drives this QTL. Finally, the trait c in Figure 4.8c is distributed similarly to Figure 4.8a, but in this case there is association ($\lambda_c = 15.2$) with a locus a few Mb away from the source. The peak of c in Figure 4.8i is much wider than that of b in Figure 4.8h, and is consistent with the mapping resolution of 200kb expected in the MAGIC population [67]. Although $\lambda_b > \lambda_c$ its SV QTL is probably artefactual, whilst the SV QTL of c in 4.8i appears genuine, and in fact corresponds to a known translocation [132].

Since λ_A alone is not a good indicator for the existence of an SV QTL, I tried permutation to determine genome-wide significance for each trait separately. I performed $N = 100$ phenotype

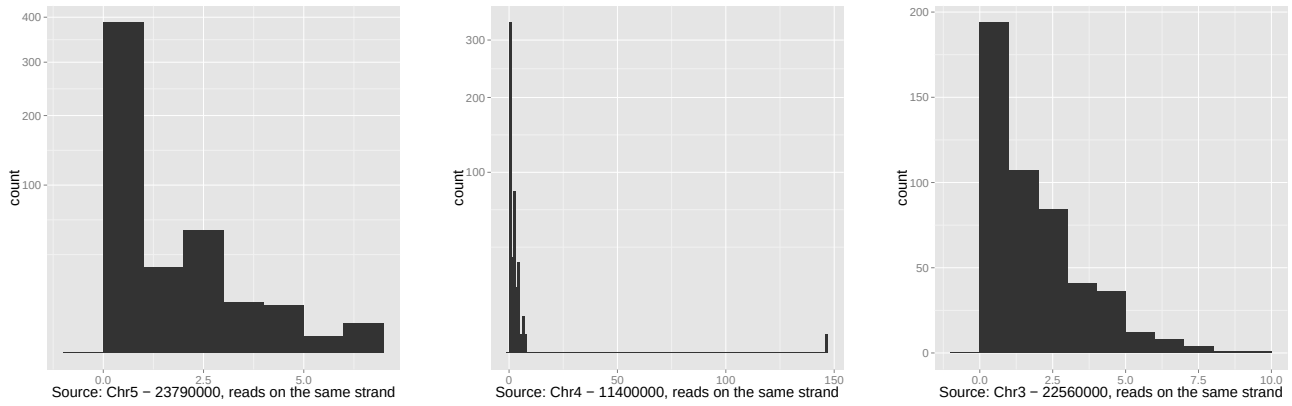
permutations for each trait and repeated the mapping for each one of them. From the permutations, I computed the permutation p-value π_A as:

$$\pi_A = \frac{\mid t \in T_A : \lambda_A(t) \geq \lambda_A \mid}{N} \tag{4.19}$$

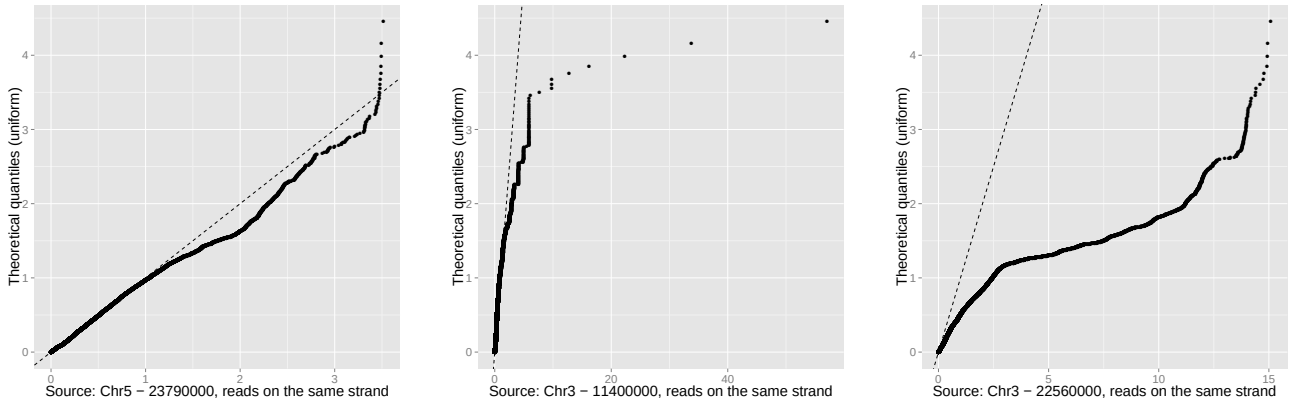
where T_A is the set of N permutations for the trait A, and $\lambda_A(t)$ is the maximal logP for the tth permutation. Thus π_A is the probability of observing a more extreme maximal association in the permutations than in the real scan. While λ_A is a negative logarithm of p-values following an exponential distribution under the null, π_A is a p-value following a uniform distribution under the null. The minimum non-zero value that π_A can take is $1/100 = 10^{-2}$ and given the size of the study, this does not cover the range of genome-wide p-values encountered. For this we would need at least $100,000$ permutations per trait, which is computationally very expensive. Instead, I fitted a generalised extreme value (GEV) distribution, to each set of 100 $\lambda_A(t)$ permuted values obtained for each trait, by maximum likelihood using the R package `evd`. I obtained a corrected genome wide log p-value γ_A based on the fitted GEV, which is defined in `evd` as:

$$\gamma_A = -\log(1 - \exp\left(1 + \hat{s}_A(\frac{\lambda_A - \hat{a}_A}{\hat{b}_A})^{(-\frac{1}{\hat{s}_A})}\right)) \tag{4.20}$$

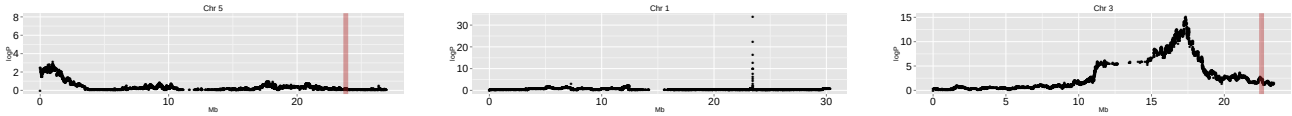
$\hat{a}_A, \hat{b}_A, \hat{s}_A$ are the MLEs of a_A, b_A, s_A. γ_A controls for the number of tests in an individual scan (16,700) and is effective in removing false positive SV QTLs caused by outliers: in the example b of Figures 4.8h $\gamma_b = 0.03$, while for the trait c in Figure 4.8i $\gamma_c = 6$.



(a) Source = Chr5, 23.79Mb (b) Source = Chr4, 11.40Mb (c) Source = Chr3, 22.56Mb



(d) QQplot: $x : -\log(P)$, y : Exponential (e) QQplot: $x : -\log(P)$, y :Exponential (f) QQplot: $x : -\log(P)$, y :Exponential

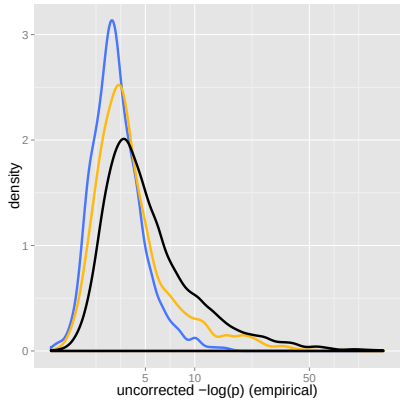


(g) $y : -\log(P)$, x : Mb (chr5) (h) $y : -\log(P)$, x : Mb (chr1) (i) $x : -\log(P)$, x Mb (chr3)

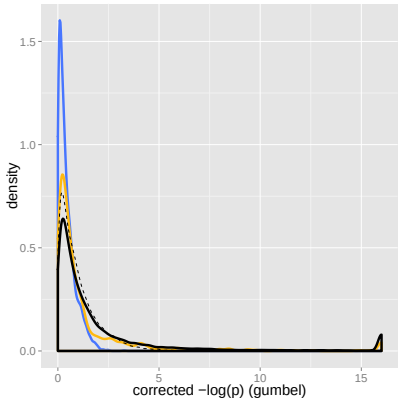
Figure 4.8: Distributions, QQ plots and Manhattan plots for three example traits, all indicating read pairs mapped to the same strand. The first row shows histograms of read anomaly traits, x-axis: Number of anomalous read pairs detected, y-axis: frequency distribution. In the second row, QQ plots compare the distributions of the $\log_{10}$ p-values estimated at each locus by the genome scan to the quantiles of the exponential distribution. The dashed line corresponds to the line $x = y$ to which the QQ plot should converge under the null. The third row shows the corresponding manhattan plots illustrating associations at each genomic position. Only the chromosome in which the genome-wide logP λ_A was maximum (i.e.candidate SV QTLs) is shown. For the first and the third example, the candidate QTLs are on the same chromosome as the source locus, which is marked by a red vertical bar. In the second example the source is in a different chromosome. Note that the horizontal scales in a-f differ, as do the vertical scales in g-i.

I thus obtained three genome-wide statistics for each trait: λ_A, the uncorrected genome-wide maximum logP, the p-values corrected by permutation π_A, and the logP corrected by permutation and GEV-fitting γ_A. I analysed the distribution of these statistics in 6 sets of $11,915$ traits of the same type. For a single genome scan the theoretical distribution of p-values at all tested loci under the null hypothesis is uniform. Making a similar argument for multiple genome scans, the distribution of corrected maximum genome-wide p-values for multiple independent scans under the null (i.e. no traits with genetic association) should also be uniform. Figure 4.9 shows the distributions of the statistics $\lambda_A, \pi_A, \gamma_A$ for the $11,915$ traits measuring read-pairs on the same strand (inversions), and illustrates how they compared to expectation as well as their pairwise relationships (in black). The distribution was far from uniform for all 3 statistics. (Figures 4.9a - 4.9f). λ_A has a long tail of very high p-values (up to 200, not shown in the plot), which π_A and γ_A compress to a smaller range. In the distribution of π_A there is an excess of traits with relatively low permutation p-value ($\pi_A < 0.1$). Consequently, the corresponding QQ-plots did not match the expected uniform distribution. Figures 4.9g show the effects of the transformation of λ_A to π_A and γ_A. As would be expected, traits with $\lambda_A < 4$ mapped to very low γ_A and $-\log(\pi_A)$. In contrast, certain traits with higher λ_A had equivalently high $\gamma_A, \log(\pi_A)$ values, while others were corrected to a much lower level of association. Finally $\exp(-\gamma_A)$ and π_A are well correlated over the range $0.1 - 1$, so γ_A is a good surrogate of π_A in the absence of sufficiently many permutations.

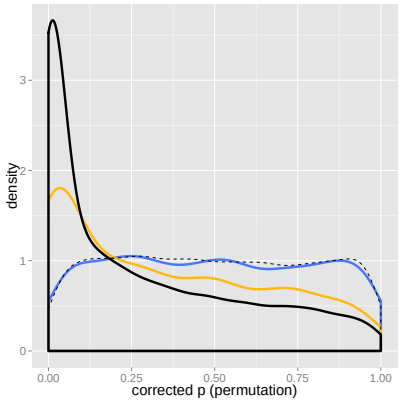
I investigated further the departure from the expected distribution of γ_A. First, I tested whether truly normally distributed traits would have uniformly distributed p-values by simulating 1000 independent normal traits with mean $m = 0$ variance $\sigma^2 = 1$ and repeating the genome scans and permutations (Figure 4.10). With normal data $\exp(-\gamma_A)$ and π_A were uniform as expected. Next, I tested whether non-uniformity was explained by very large differences in the statistical properties between traits (e.g. due to the presence of outliers), which could be corrected by quantile normalisation of the trait values prior to performing the genome scan. We repeated the analysis with a sample of 1000 quantile-normalised traits from the same strand pair counts, however the fit did not improve (Figure 4.9, in yellow). While quantile-normalisation helps in correcting outlier-driven traits (trait b from Figure 4.8b gets $\lambda_A = 3.3$ after quantile-normalisation), this does not



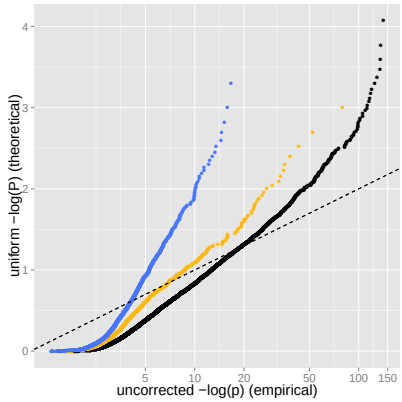
(a) Density of λ_A



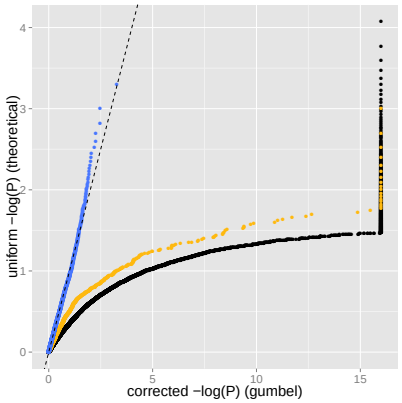
(b) Density of γ_A



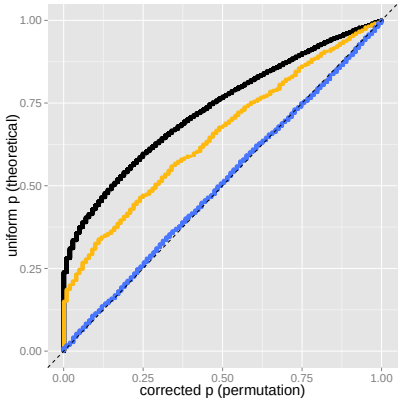
(c) Density of π_A



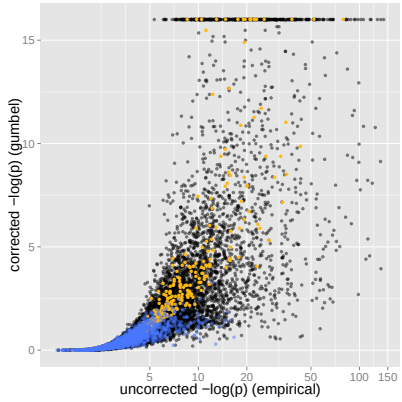
(d) QQplot: x: λ_A, y: Exponential



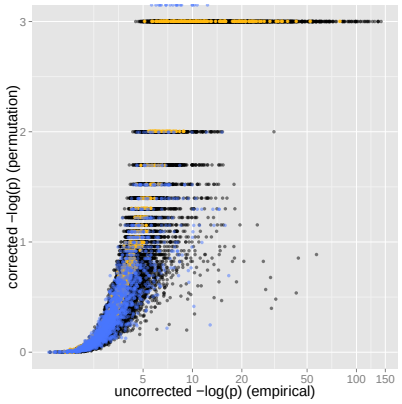
(e) QQplot: x: γ_A y: Exponential



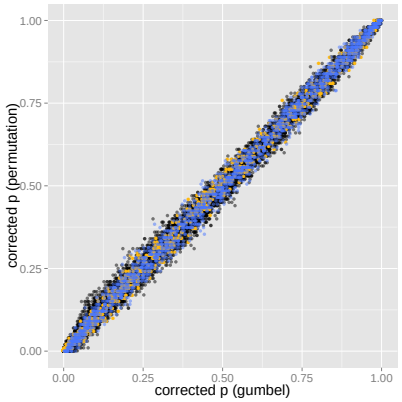
(f) QQplot: x: pi_A y: Uniform



(g) Scatterplot: x: λ_A (log) y: γ_A (log)



(h) Scatterplot: x: λ_A (log) y: γ_A (log)



(i) Scatterplot: x: γ_A y: π_A

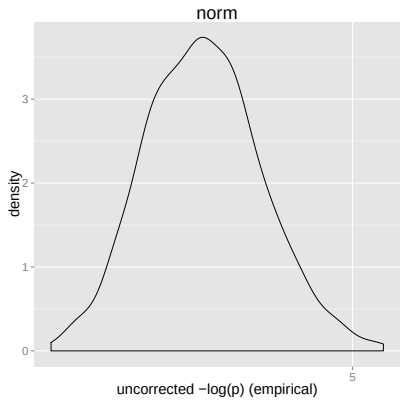
Figure 4.9: Summary statistics of trait p-values, λ_A, the negative logarithm of maximum genome-wide empirical p-value, γ_A, the corrected negative logarithm after fitting a GEV distribution to 100 permutation λ_A values, and π_A the permutation p-value. P-values estimated from raw trait values are shown in black, 1000 randomly selected quantile-normalised traits are shown in yellow and 1000 permuted trait distributions are shown in blue. Dashed lines indicate the theoretical null distributions. Figures 4.9a, 4.9b and 4.9c show p-value density, estimated using density function in R, so that the area under each curve is equal to one. Figures 4.9d, 4.9e and 4.9f are QQ plots comparing the distribution of the p-values to the quantiles of the theoretical (uniform or exponential) distribution. Figures 4.9g, 4.9h and 4.9i show pairwise correlation between λ_A, γ_A and π_A. Observations with $\lambda_A > 25$ are not shown. Because of lack of significant digits in p-values below 10^{-16} reported by R, observations with $\gamma_A > 16$ were set to 16. Similarly, because the lowest possible non-zero value of π_A was 10^{-2} (as there were 100 permutations), values with $\pi_A = 0$ were set to $P = 10^{-3}$.

improve the fit of γ_A to the exponential distribution. Besides, fitting the EVD and using γ_A has the same effect on outliers. Finally, I tested whether EVD is a suitable model for the distribution of trait p-values, by permuting 1000 traits from the same strand pair counts and repeating the mapping. The distribution of $\exp(-\gamma_A)$ and π_A for permuted traits were uniform (Figure 4.9, in blue), proving that under null γ_A and π_A are distributed as expected and hence they are appropriate corrections of λ_A.

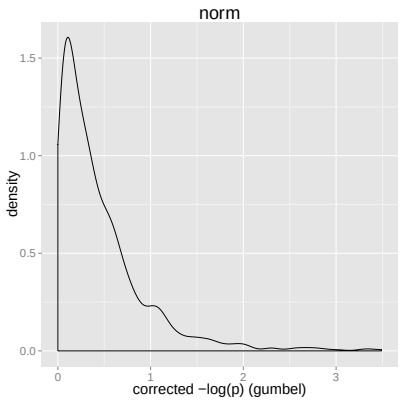
Inflation due to the presence of many true SV QTLs might also explain non-uniformity. In almost half (44.12%) of the traits the source locus and the candidate SV QTLs were on the same chromosome, and in about a third of cases (29.24%) they were less than 2Mb apart. Therefore, I investigated whether there were in fact a large number of local (cis) SV QTLs. These would be caused by sequence features in the local sequence of certain founders (including SVs, but also transposons or repeats) that could cause read mapping anomalies over certain haplotypes. To test, I omitted from the analysis all traits that had a candidate SV QTL on the same chromosome as the source locus and I also removed the top associations on other chromosomes, in which $\pi_A < 0.1$. In total 59.9% of all the traits were omitted. The fit to a uniform distribution considerably improved (Figure 4.11). The fit can be further improved by pruning more traits at the lower end of π_A (data not shown).

In conclusion, while read anomaly traits are generally not normally distributed, this is not the cause of the excess of associations. Instead, there are many traits with strong associations with nearby loci (cis effects). In particular, 27.6% of all traits have an associated SV QTLs in the same chromosome as the source locus, at $\pi_A < 0.1$, 82.1% of which lie within 2Mb from the source locus. These associations skew the distribution of λ_A away from the null distribution. Nevertheless, their genome-wide p-value is often not extreme so we cannot conclude that there is an SV underlying each one of these cis effects. After removing these cis-effects the distribution of corrected genome-wide p-values was closer to the null, thereby confirming the initial hypothesis that the distribution of genome-wide p-values under the null hypothesis is uniform.

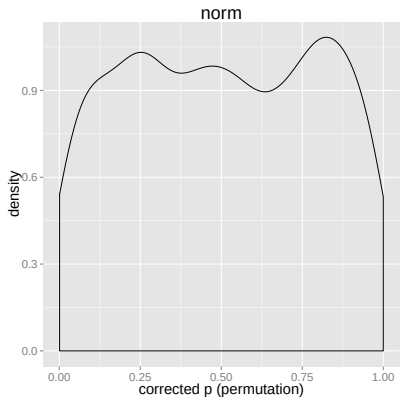
To determine study-wide SV QTLs, I used γ_A as the overall trait statistic. The significance threshold α for $\exp(-\gamma_A)$ was selected by minimising false positives using false discovery rates



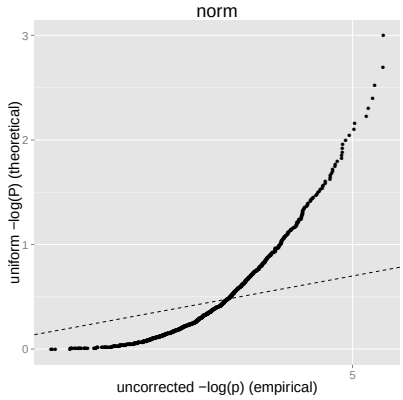
(a) Density of λ_A



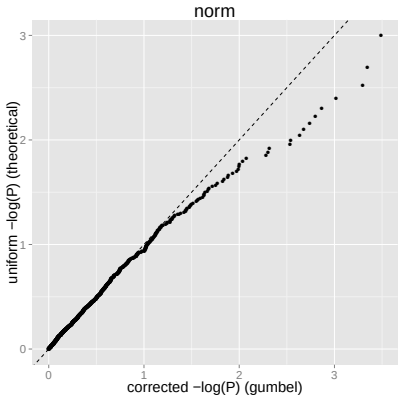
(b) Density of γ_A



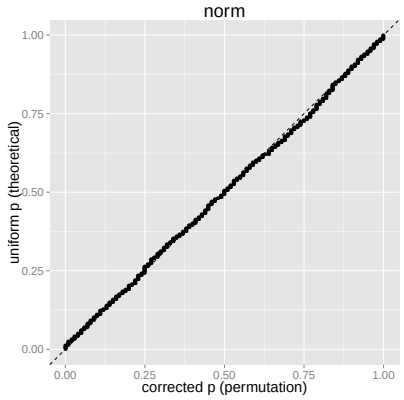
(c) Density of π_A



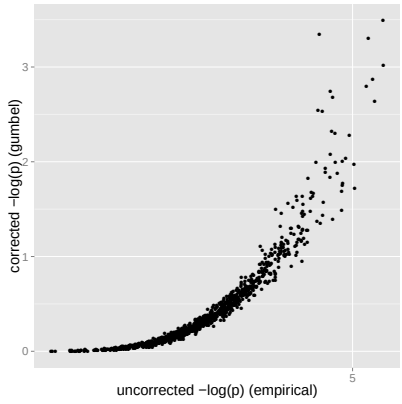
(d) QQ plot: x: λ_A, y: Exponential



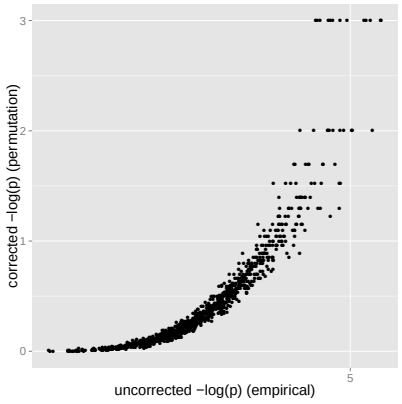
(e) QQ plot: x: γ_A y: Exponential



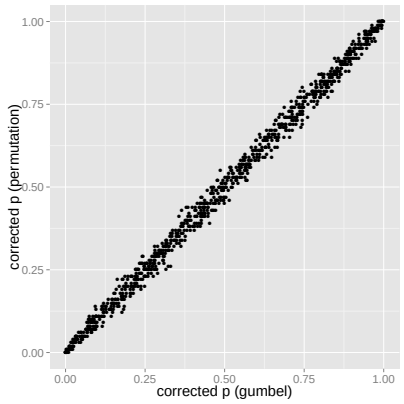
(f) QQ plot: x: π_A y: Uniform



(g) Scatterplot: x: λ_A (log) y: γ_A (log)

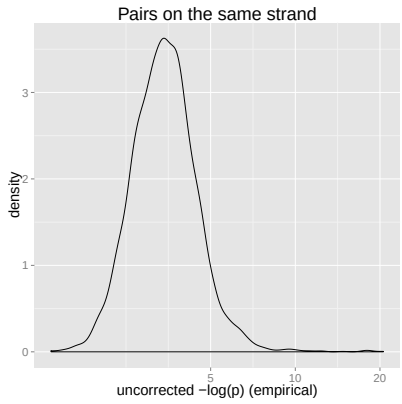


(h) Scatterplot: x: λ_A (log) y: γ_A (log)

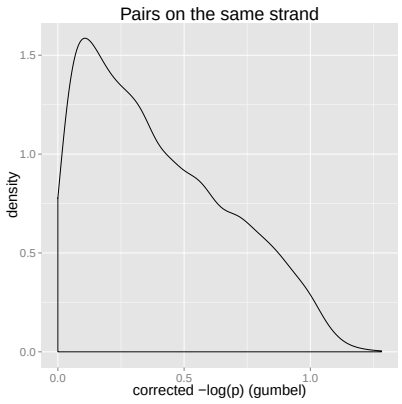


(i) Scatterplot: x: γ_A y: π_A

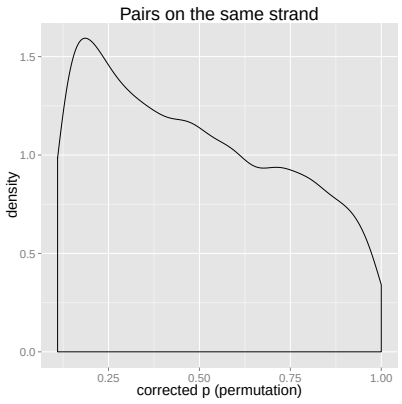
Figure 4.10: P-value density, QQ plots and correlations for simulated normal traits.



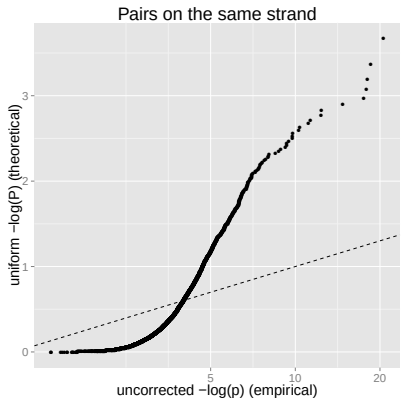
(a) Density of λ_A



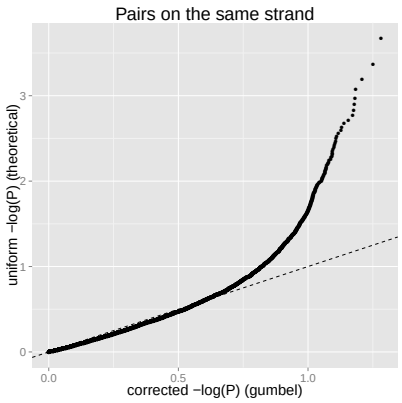
(b) Density of γ_A



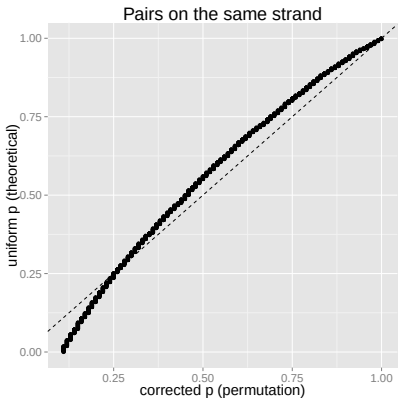
(c) Density of π_A



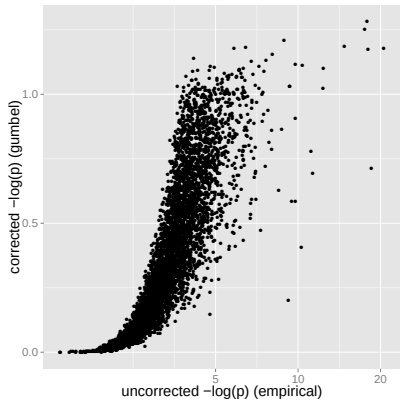
(d) QQplot: x: λ_A, y: Uniform



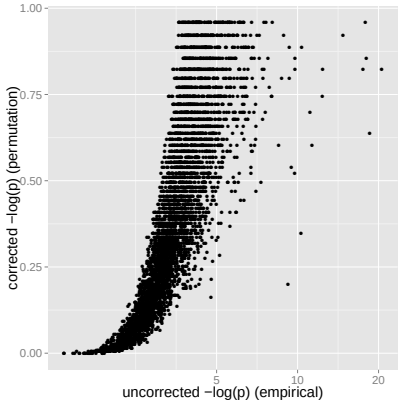
(e) QQplot: x: γ_A y: Uniform



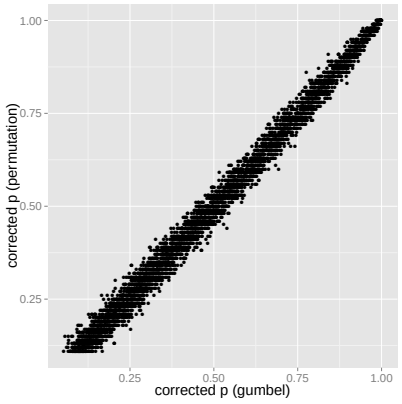
(f) QQplot: x: π_A y: Uniform



(g) Scatterplot: x: λ_A (log) y: γ_A (log)

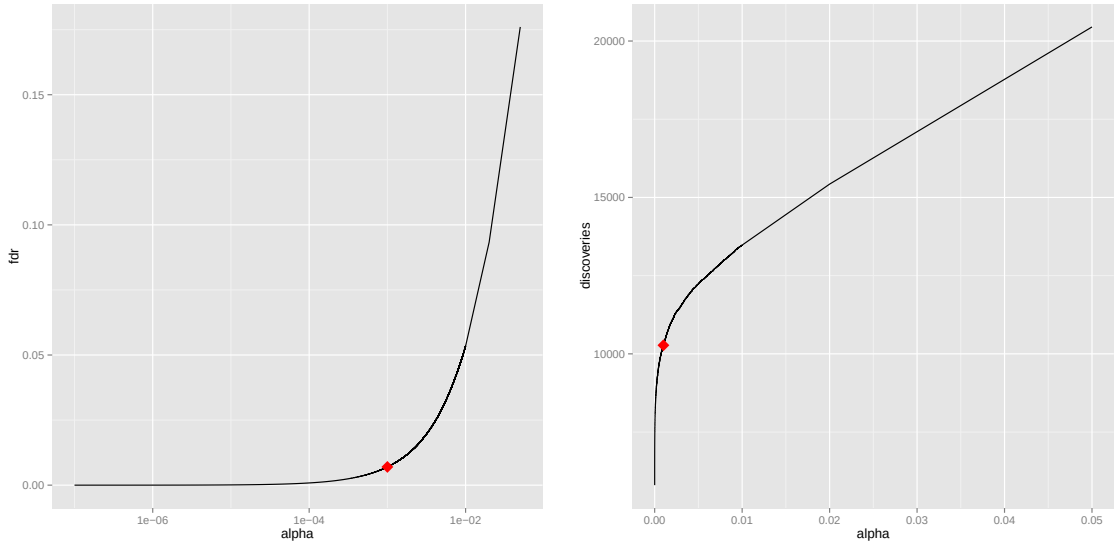


(h) Scatterplot: x: λ_A (log) y: γ_A (log)



(i) Scatterplot: x: γ_A y: π_A

Figure 4.11: P-value density, QQ plots and p-value correlations for pairs on the same strand after pruning candidate SV QTLs in which the source and the sink were on the same chromosome and after discarding associations in which $\pi_A < 0.1$. In total $4,694$ traits are shown.



(a) False discovery rate per α

(b) Discovered sink QTLs per α

Figure 4.12: False discovery rate and number of discovered SV QTLs per study-wide significance threshold α on $\exp(-\gamma_A)$. The red diamond shows the chosen level of significance $\alpha = 10^{-3}$, FDR $= 7 \times 10^{-3} < 10^{-2}$

(FDR). Given a significance threshold α, FDR was defined as the fraction of expected false discoveries over the total discoveries, thus:

$$\text{FDR} = \frac{N\alpha}{D(\alpha)} \tag{4.21}$$

where $D(\alpha)$ is the number of discoveries at level α and N is the total number of traits. For a single type of read mapping anomaly, FDR is zero with $\alpha < \frac{1}{11{,}915} < 10^{-4}$, but this would be too conservative. I thus selected as significance threshold $\alpha = 10^{-3}$ at which FDR $= 7 \times 10^{-3} < 1\%$. The FDR and number of discoveries for the range of α values are shown in Figure 4.12.

4.7 Structural variants in MAGIC

At FDR $< 10^{-2}$ I mapped $10,275$ SV QTLs. These were classified as cis if the sink is within 2Mb of its source, and trans otherwise. Table 4.1 summarises the results of mapping for each of the read-pair anomaly measures used. Different anomaly traits at the same source location had

coincident SV QTLs at 3,773 source loci. Duplicates probably correspond to the same SV so they were counted only once. Therefore, I mapped 6,502 distinct QTLs that correspond to unique SVs. Of these, 4,898 were cis and 1,604 (24.7%) trans. Figure 4.6 and Appendix A show the positions of the cis and trans QTLs. Overall, 4,073 out of 11,915(34.2%) 10kb segments were predicted to be structurally variant in at least one of the MAGIC founders. In 1,171 SV QTLs I could map exact breakpoints using *de novo* contig alignments (see below, Section 4.9.2) and hence estimate SV size. The mean SV size was 55kb and the maximum was 334kb. The distribution of SV sizes is shown in Figure 4.13.

trait type	**sink QTLss**	**unique**	**cis**	**trans**
High read coverage	184	165	112	72
Large insert size	2051	585	1677	374
Unpaired reads	2060	1998	1530	530
Pairs on same strand	1950	1887	1358	592
Unpaired + Large insert size	2033	431	1661	372
Improperly paired reads	1997	833	1617	380
Total	10275	5899	7955	2320

Table 4.1: Results of mapping per read pair anomaly measurement; rows are the read pair anomalies considered; columns: **sink QTLs**: total number of SV QTLs detected using each anomaly category (if the same QTLs was detected by multiple anomalies then it is counted multiple times in this column), **Unique**: number of SV QTLs detected only by a single anomaly category, **cis**: number of cis-QTLs for the anomaly, **trans**: number of trans SV QTLs for the anomaly.

I classified each SV based on the type of read anomaly trait that identified it. Thus, same-strand read pairs indicate inversions and high read coverage duplications. Abnormal insert sizes and unpaired reads (and their combinations) may indicate different SV-types in cis or in trans; cis SV QTLs were called indels if the distance from the source to the sink was below 1Mb, otherwise they were classified as short-range translocations. All other trans SV QTLs were classified as translocations. In total, there are 175 duplications, 1,976 inversions (1,373 short-range and 603 long-range) and 1,316 translocations (381 short-range and 935 long-range). Table 4.2 tabulates the number of SVs identified by SV type and SV QTL location (cis or trans) as well as independent

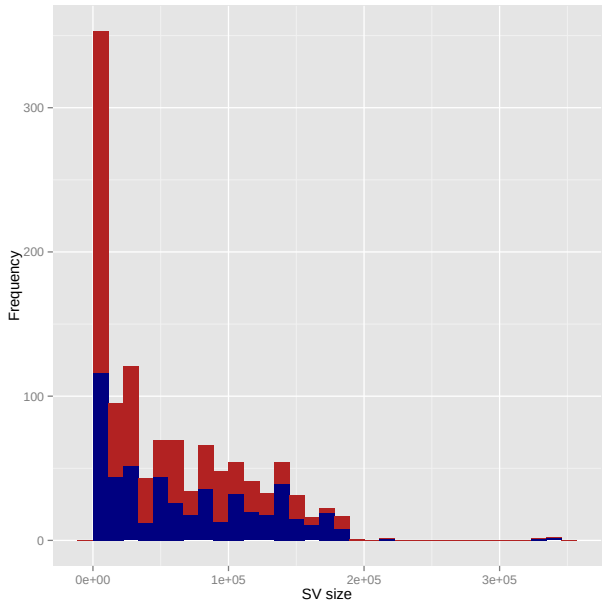


Figure 4.13: Distribution of SV size for the 1,171 SVs in which both breakpoints were mapped using *de novo* assembly contigs. The x-axis represents the size of SV in bp and the y-axis the frequencies. The blue bars correspond to short-range SVs (cis SV QTLs) while the red bars to long-range SVs (trans SV QTLs).

confirmation analysis by read-pair mappings or contig data (see 4.9). Figure 4.14 shows a circos plot summarising short and long-range structural variation in one of the founder haplotypes, Ler-0. The figure also contains these SVs that have been independently validated by *de novo* contigs and in vivo, by PCR (see Section 4.9). Circos plots for the remaining 17 founders are showing in Appendix D.

SV type	**N**	**cis**	**trans**	**R**	**C**
duplication	175	109	66	38	14
indel	3035	3035	0	1036	87
inversion	1976	1373	603	342	123
translocation	1316	381	935	138	196
Total	6502	4898	1604	1554	420

Table 4.2: Structural variants detected in the MAGIC population. The columns are: **N:** number of variants for each type, **cis:** variants in cis (source and sink lie within 2Mb from each other), **trans:** variants in trans, **R:** variants confirmed by paired reads, **C:** variants confirmed by *de novo* contigs.

4.8 SV QTL allele frequencies

The test of association between founder haplotype and SV trait implicitly assigns an SV allele to each of the founders (i.e. the estimate $\hat{\beta}_{AL}(s)$ for the founder s, in Equation (4.16)). However because these might be sensitive to outliers, especially at low haplotype frequencies, I chose to compute allele frequencies independently. Therefore I devised a procedure to predict the founder haplotypes carrying SVs at the origination of the MAGIC population and used them to estimate SV allele frequencies.

The procedure uses the fact that founder haplotypes carrying an SV allele at the sink will have elevated numbers of anomalous reads at the source than other haplotypes. For a given SV QTL, the founders' contributions to each SV QTL were arranged as a 19×19 table T whose cells (i, j) carried the sum of read anomalies (of a specific type) at the source for all lines carrying haplotype i at the sink and haplotype j at the source. At cis SV QTLs the source and sink are usually identical so usually only diagonal elements of T are non-zero. A founder is classified as carrying the SV allele if its corresponding row in the table generally had higher values than the rest of the table. If only sporadic cells of the row had high values then these might have been outliers so the founder was not associated with the SV.

Figure 4.15 shows tables T and corresponding Manhattan plots for example cis and trans sink SV QTLs.

To extract high-scoring rows, a score was computed for each row, indicating the contribution to the phenotype of each founder haplotype at the sink. For each cell in T, let t_{ij} be the sum of trait values for all genomes carrying haplotype i at the sink and haplotype j at the source respectively, and let $r_i = \sum_j t_{ij}$ be the sum of phenotype values for the row i. The rows r_i are then re-ordered so that $r_1 > r_2 > \ldots > r_{19}$. Under the null hypothesis, i.e. if no founder at the sink is associated with the trait, all row sums are equal $r_1 \approx r_2 \approx \ldots \approx r_{19}$. The null hypothesis is rejected if there is a set $\{r_1, \ldots r_k\}$ whose row sums are much larger than the rest of the table, such that $r_1 > \ldots > r_k > r_{k+1} \approx \ldots \approx r_{19}$. With $S = 19$ founder haplotypes there are 18 such possible sets.

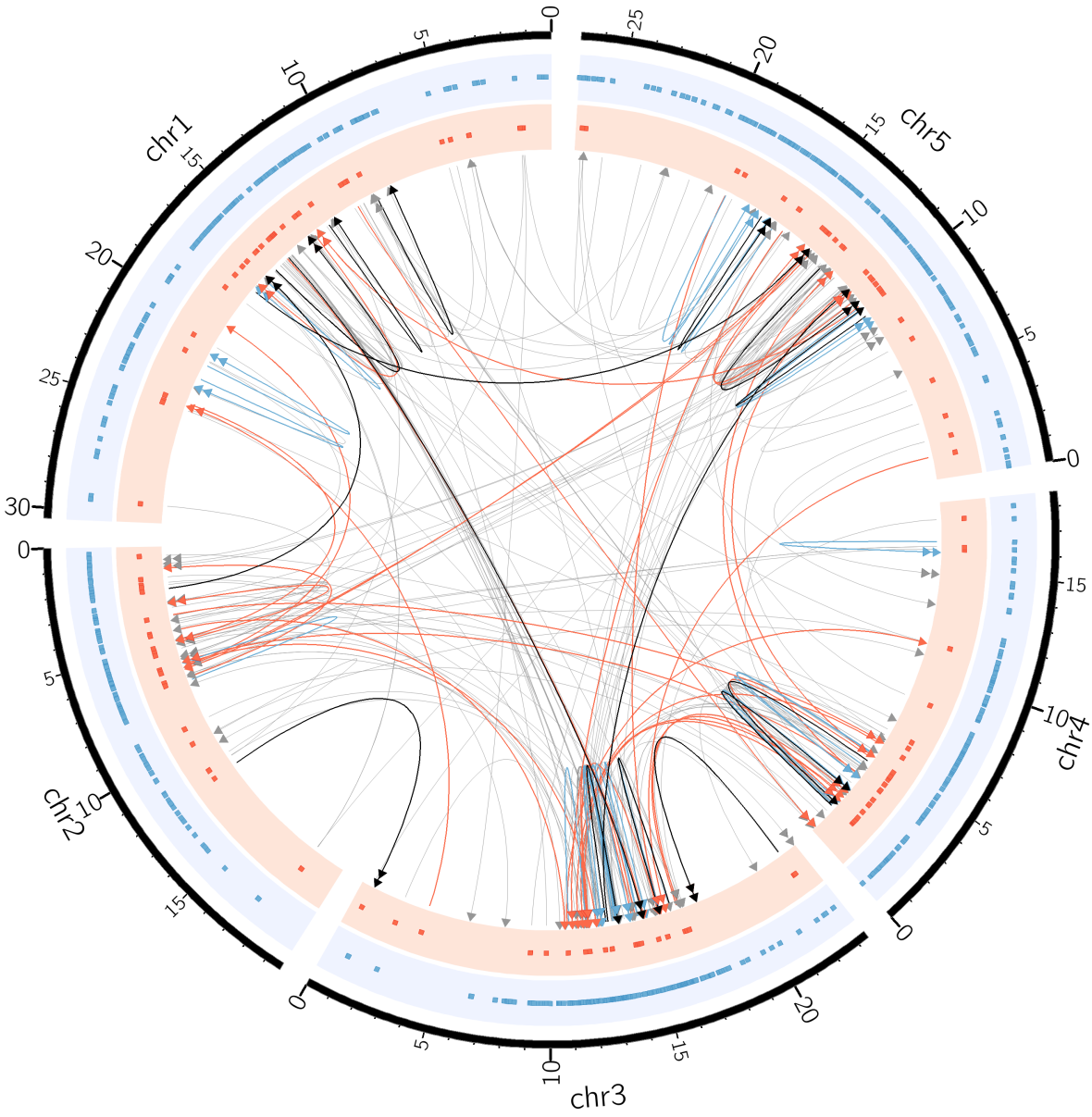
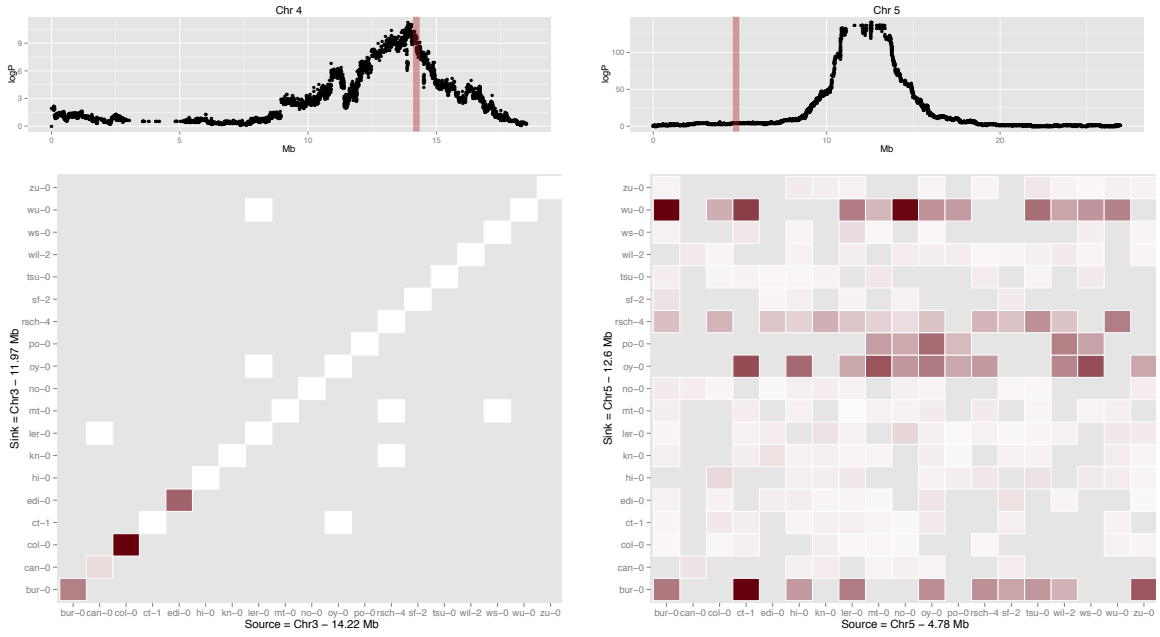


Figure 4.14: Circos plot displaying Structural Variants detected in the genome of the MAGIC founder Ler-0. The grey directed links show all SVs detected from SV QTLs with the arrows pointing towards the sink. Red and blue links indicate 37 trans and 30 cis SV QTLs confirmed by *de novo* contigs. The black links show 16 SVs confirmed in Ler-0 by PCR (7 cis, 9 trans). Double arrows in links indicate inversion sites. The dots in the red and blue tracks mark the sources of all SVs associated with the Ler-0 haplotype.

The z-score of each set k is defined as:

$$z_k = \frac{\sum_{(j=1)}^{k} r_j - \mathbb{E}(r_k)}{\sigma(r_k)} \tag{4.22}$$

The expected value $\mathbb{E}(r_k)$ and standard deviation $\sigma(r_k)$ are estimated by performing $N = 1000$ permutations of the elements of T. I then choose the set $k : \{r_1, \dots r_k\}$ which maximises z_k. Given $R(z_k)$, the vector of z_k estimated from the permutations then the permutation p-value $P(z_k)$ is



(a) Cis SV QTL - source: chr 3, 14.22Mb, sink: chr3, 11.97Mb, trait: high read coverage

(b) Trans SV QTL - source: chr5, 4.78Mb, sink: chr5, 12.60Mb, trait: high read coverage

Figure 4.15: Manhattan plots and founder contributions for the trait high read coverage in a cis (4.15a) and a trans(4.15b) SV QTL. In the manhattan plots the red line shows the source and the association peak the sink of the SV QTL. In the founder contributions tables rows and columns correspond to founder haplotypes at the sink and source, respectively. The colour hue at each cell is the trait value for each combination of founder haplotypes, darker colour means higher value. The values in Figure 4.15a range from 0 to 1000. The figure shows a duplication (confirmed by *de novo* contigs, see Section 4.9.2) present in 3 founders, namely Bur-0, Col-0 and Edi-0. In Figure 4.15b trait values (high read coverage) range from 0 to 600 and the figure is showing a trans QTL in chromosome 5, present in Bur-0, Oy-0, Po-0, Rsch-4 and Wu-0.

defined as:

$$P(z_k) = \frac{\mid r_{z_k} \in R_{z_k} : r_{z_k} \geq z_k \mid}{N} \tag{4.23}$$

i.e. the fraction of permuted tables that had a higher z_k score than the original table. Founders corresponding to $k : \{r_1, \ldots r_k\}$ are associated with the SV-carrying if $P(z_k) \leq 10^{-2}$. The test predicted founder haplotypes at $2,391$ QTLs. The mean SV allele frequency in the population, defined as the fraction of founders carrying an allele at that locus was $6/19 = 31\%$. SV alleles are more often minor than major (i.e. a minority of founders usually carry the SV allele). Only 387 (12%) are specific to a single founder (Figure 4.16). This is in contrast to the fraction of SNPs (45%) that are private to a single founder.

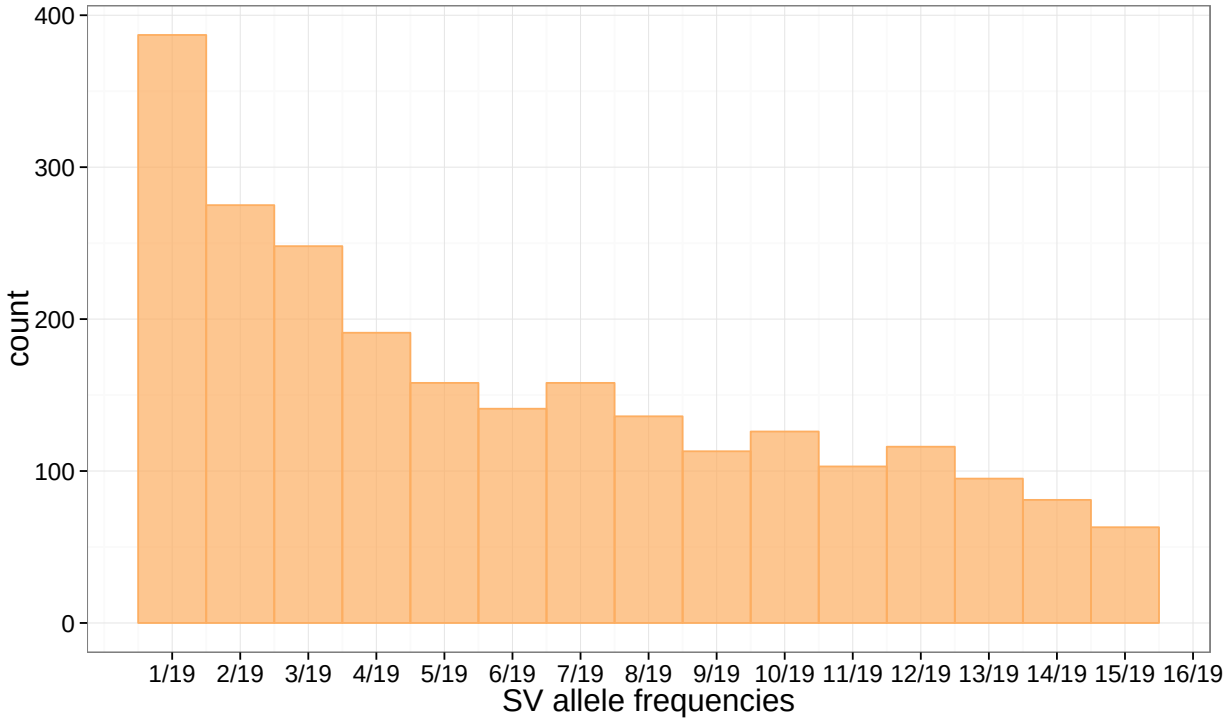


Figure 4.16: SV allele frequencies in MAGIC, defined as the fraction of founders carrying the SV allele.

4.9 Validation of SVs

Because the mapping resolution of QTLs in the MAGIC population is only about 200kb [67] I could not predict SV breakpoints from the QTL mapping alone. Therefore SVs were verified by read-pair data and by *de novo* assembly contig alignments. Both methods can help confirm SV predictions and pinpoint exact breakpoints, however lack of support from them does not definitively disprove an SV QTL, as they both have limitations in SV prediction. Thus, read-pair data contain a high fraction of noise, especially in the presence of repeats, while reads that correspond to SVs may be incorrectly aligned or unmapped. Validation from *de novo* contigs heavily relies on contig size and contigs may be misassembled over repeated or copy-number variant loci.

In total $1,898(29.2\%)$ of the SVs were validated by these methods. In addition, I realigned two manually assembled contigs from highly divergent regions of the founder genome Ler-0 [68] which also supported my SV predictions. 39 SVs were also validated experimentally, using polymerase chain reaction (PCR).

4.9.1 Validation with read pairs

I used the predictions of founders carrying an SV allele, available at $2,391$ SVs, to search for an excess of read pairs linking the source and sink loci in these founders compared to those not carrying the SV allele. In particular I looked for read pairs in which one read mapped within 10kb of the source and the other within 200kb of the sink (200kb is the QTL mapping resolution in MAGIC [67]). Note that the construction of the read anomaly traits did not use this information, thus all reads mapped to different chromosomes were counted as anomalous regardless of the position of the pair or whether it was mapped.

I first analysed the 19 founders high-coverage sequence data [36] restricting attention to the $2,391$ SV QTLs where founder alleles were predicted. I compared the number of such reads in the founders that carry the SV to the remaining founders using Fisher's exact tests (FET). In 759 out of $2,391$ (31.7%) SV QTLs I observed an excess of reads in the founder data connecting the source to the sink at FET $P < 0.05$.

I performed the same test for all $6,502$ source and sink pairs, using low-coverage reads from the

MAGIC lines. Where founder SV alleles were not predicted I compared the 100 lines with highest trait values to the rest of the population. At $P < 0.05$, there were $1,141 (17.5\%)$ source-sink pairs where MAGIC lines carrying an SV had more read-pairs connecting the source and the sink.

In total $1,554$ SV QTLs (23.9%) were supported by read pairs in the founders or in the MAGIC lines. Therefore I employed further validation methods to investigate the other SV QTLs.

4.9.2 Validation with *de novo* contigs

I used *de novo* assemblies of the 19 founder genomes [36] to identify SV breakpoints, based on the assumption that the SV QTLs correspond to SVs segregating among the founders. I used BLAT [60] to align $5,524,143$ short contigs (50bp-1kb long) from existing *de novo* assemblies of the 18 non-reference founder genomes to the reference Col-0 (TAIR10) to identify contigs split across the source and sink locus. I excluded genomic regions with annotated repeats or transposons. I also filtered out contiguous alignments that are unlikely to correspond to SVs i.e. alignments whose BLAT score (number of matches minus mismatches and gaps) exceeded $0.95l$, where l the length of the contig, because they matched the reference too closely. I also filtered out alignments that mapped to over 5 genomic loci as they probably correspond to unannotated repeats.

In the remaining alignments, two phenomena were observed. A minority of *split contigs* had two parts, one mapping to within 200kb of a source and the other within 500kb of the corresponding sink. I found $2,619$ contigs with alignments split into disjoint pieces over 420 QTLs sources and sinks. These split contigs suggest a simple cut-and-paste mechanism for the SV breakpoint. However, I found a much larger number of $460,656$ (8.3%) of *shared contigs*, whose alignments overlapped between source and sink regions; Such alignments may be present because of duplications, but also because of transposable elements, low-frequency repeats (high-frequency repeats were masked out) or Microhomology-Mediated Break-Induced Replication (MMBIR) loci [44] and Non-Allelic Homologous Recombination (NAHR), that are known to be overrepresented near SVs [92, 137]. In fact, some split alignments also had shared parts, but were long enough to pinpoint a breakpoint beyond which alignments were disjoint. Shared alignments are typically smaller (median size = 85bp, while split alignments have median size of 400bp) so it is possible that some of

them correspond to, or are nearby, SV breakpoints, but are not long enough to allow us pinpoint them.

To investigate whether split and shared alignments were enriched near SVs , I performed 100 circular genome permutations [15] of each of the $N = 6,502$ identified SV QTLs. Circular permutation of source and sink locations keep the distances between source and sink pairs constant while altering their locations in the genome. For each SV QTL i, if $a(i), b(i)$ are the original positions of the source and the sink respectively, then the permuted pair $a_k(i), b_k(i)$ is defined as:

$$\begin{aligned} a_k(i) &= (a(i) + \theta_k) \mod G_L \\ b_k(i) &= (b(i) + \theta_k) \mod G_L \\ \theta_k &\sim Unif(0, G_L) \end{aligned} \tag{4.24}$$

where $G_L = 119,146,348$ is the *A. thaliana* genome length. If $a(i), b(i)$ define a cis SV QTL but $a_k(i), b_k(i)$ are on different chromosomes, the permutation is repeated until they are placed in the same chromosome.

I then counted split and shared alignments between the permuted source-sink SV QTL locations and compared them to the original data using permutation values P_{split}, P_{shared} indicating the fraction of permutations that had more split or shared alignments. In the case of split alignments, none of the permutation sets exceeded the number of split alignments in the real data, thus $P_{split} < 0.01$ so split alignments are associated with SVs. Shared alignments are also associated with SVs at trans QTLs ($P_{shared,trans} < 0.01$), but not with cis $P_{shared,cis} < 0.2$, possibly because of the presence of local small CNVs.

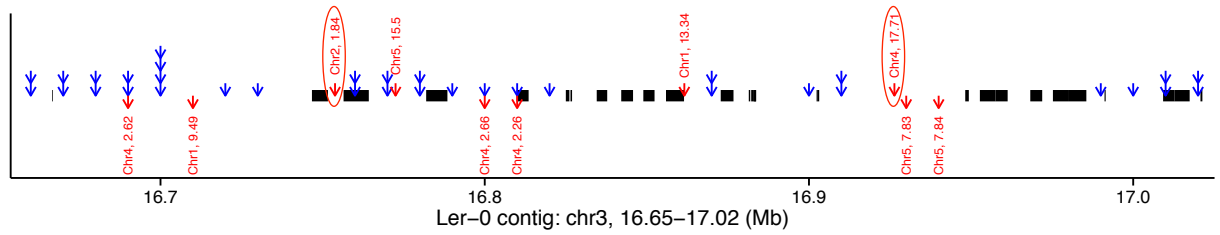
Split contigs mapped breakpoints at 420 SVs. At 367 trans QTLs where only one split contig defined a breakpoint, I used the closest shared alignment to define the second breakpoint (and hence minimise the predicted extent of the SV).

4.9.3 Validation using manually assembled Ler-0 contigs

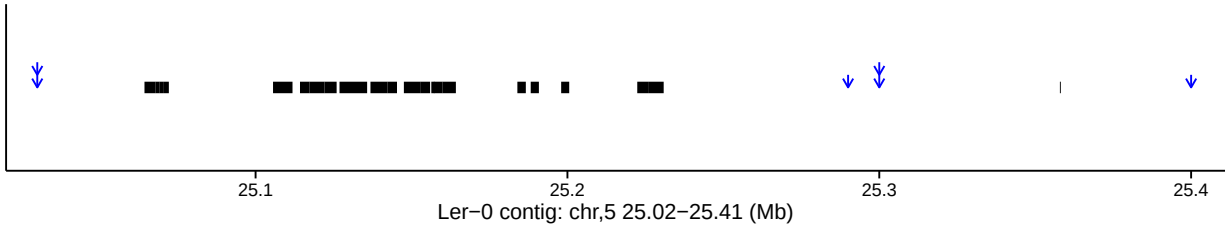
I analysed two manually assembled large contigs from Ler-0 [68] (TAIR10 coordinates: chr3:16.65 − 17.02Mb, chr5:25.06 − 25.23Mb) over regions where Ler-0 is highly divergent from the reference. In the chromosome 3 contig, manual assembly identified 83 indels, 31 larger than 100bp [68]. I mapped 36 cis SV QTLs (source and sink within the region) and 10 trans SV QTLs (6 with source and 4 with sink in the region). The region is 1.77-fold enriched for SVs compared to the rest of the genome (a region of the same size is expected to have 6 SVs). After realigning the contig to the reference using BLAT [60] cis SV QTLs coincided with the gaps in the contig alignments. Furthermore the alignments revealed two translocated segments, in chromosomes 2 and 4 respectively, whose positions are concordant with two of the trans QTLs mapped in the region (i.e. QTL sources at the positions of the segments in the reference and sinks at their corresponding loci in Ler-0). In the chromosome 5 contig, I found 6 related cis QTLs, corresponding to deletions in the start and end of the contig sequence. Figure 4.17 shows the alignments of both contigs, annotated with the locations of SV QTLs.

4.9.4 Validation by PCR

Finally, I validated a subset of SV breakpoints, detected by *de novo* contigs, by PCR. I designed PCR primer oligos from both split and shared (between the source and the sink) *de novo* contig alignments, using primer3 [110]. The first type of experiment I considered (type 1 experiment) tested the amplification of primer oligos that are in remote positions in the reference genome, but in close proximity in other founder haplotypes due to SVs. Thus the experiment should produce a product of length shorter than 1.5kb only in the founders carrying an SV whereas in the founders with the configuration of the reference it should not amplify. In split alignments corresponding to SVs, primer oligos testing a breakpoint were flanking it with both oligos overlapping the contig. In shared alignments, primer oligos were flanking the shared region and were not necessarily overlapping the contig (as it may have been shared in its entirety). Thus, if based on the shared contig alignments the region (source : $\langle \text{chr}_{\text{src}}, \text{start}_{\text{src}}, \text{end}_{\text{src}} \rangle, \langle \text{chr}_{\text{sink}}, \text{start}_{\text{sink}}, \text{end}_{\text{sink}} \rangle$) is duplicated, I designed primers such that the forward oligo was within $\langle \text{chr}_{src}, \text{start}_{\text{src}} - 1000, \text{start}_{\text{src}} \rangle$ while the



(a) Alignment of a manually assembled contig from Ler-0, chr3:16.65 − 17.02Mb to the reference, annotated with the positions of SV QTLs. The black lines show alignments to reference genome at the same position. Blue arrows show the sources of cis QTLs mapped within the region - stacked arrows indicate sources that had multiple sink QTLs for several anomalous read traits. Red arrows display trans QTLs with arrows starting from the source and pointing towards the sink. Gaps in the contig alignment are specific to Ler-0 and did not align to the reference, with the exception of two transposed segments that mapped to chromosomes 2 and 4 at positions concordant with the sources of two trans QTLs (circled).



(b) Alignment of a manually assembled contig from Ler-0, chr3:22.05 − 25.23Mb to the reference and corresponding SV QTLs. The figure can be interpreted as above.

Figure 4.17: Alignments of two manually assembled contigs and positions of corresponding cis and trans SV QTLs from highly divergent regions of Ler-0

reverse oligo was within $\langle \text{chr}_{\text{sink}}, \text{end}_{\text{sink}}, \text{end}_{\text{sink}} + 1000 \rangle$.

The second type of experiment I designed (type 2 experiment) is a control experiment, which is the reverse of type 1 in the sense that we expect a product in the reference and no amplification in the SV founders. In type 2 experiments, one of the primer oligos of the corresponding type 1 was fixed, while the other was coming from the region within 1.5kb from the reference genome.

Based on the above specification, I designed $20-30$bp primer oligo sequences from the reference genome after masking out repeats, transposons and known polymorphisms. Detecting specific primer sequence was challenging as breakpoint regions were often near repeats and transposons and so I had to relax criteria for the specification of primer oligos: the maximum allowed product

was 1.5kb, minimum annealing temperature was at 10^oC, maximum annealing temperature at 90^oC, gc-content of $10 - 90\%$ and self-complementarity of up to 8bp. With these specifications I could design type 1 experiments for 88 breakpoints, corresponding to 55 SV QTLs and type 2 experiments for 53 breakpoints in 29 SV QTLs. In 10 SV QTLs both breakpoints had corresponding primer pairs, while in the rest only one breakpoint could be validated. 23 SVs had polymorphic breakpoints, i.e. different type 1 experiments were required for different founders, in terms of oligo position or in terms of oligo orientation. Also, in 36 breakpoints type 1 both primer pairs aligned to the same strand in the reference genome, probably due to inversions. Because it was not possible to determine the orientation (forward or reverse) of the tested segment in the SV genomes, I designed and tested both possible orientations (i.e. both primers to forward strand and both primers to the reverse strand).

I selected 96 pairs to confirm a subset of 76 breakpoints testing 45 SVs: 15 translocations from 5 cis and 13 trans SV QTLs and 27 inversions from 8 cis and 19 trans SV QTLs (inversions from trans SV QTLs correspond to joint translocation and inversion events). For 7 of these SVs primers testing both breakpoints were available, while in the remaining 37 it was possible to test only a single breakpoint. The PCR was performed on DNA from the 19 founder genomes, provided by the University of Bath (lab of Dr Paula Kover). Before running the PCRs, we performed confirmatory tests on the 19 founder DNA, which validated private SNPs in each genome to verify that the DNA we received was the same as the sequenced founders. The DNA sample we received for the reference accession Col-0 was contaminated; it was partly heterozygous and contained untyped indels. All other 18 DNA samples passed verification tests. At the time this thesis was written, the experiments with the correct Col-0 DNA had not been completed. We report here results for the remaining 18 accessions and for the contaminated DNA sample. In most experiments the contaminated sample behaved as the reference, i.e. did not amplify for type 1 experiments and amplified for type 2 experiments. In some experiments it did not, which is probably is due to the contamination. In those cases, we considered an SV breakpoint to be validated only if there were at least three other founders in which the experiment worked as in the reference; and hence could be used as reference surrogates.

Most experiments had the expected outcome. The interpretation of results was not straightforward in a small number of cases, probably due to duplications - for example some lines produced a product for a subset of type 1 and type 2 experiments referring to the same SV. Furthermore, it appears that in some long-range SVs, some lines carry a translocation and others a translocation and inversion, as primers with different relative orientation to the reference produced products in different founder DNA samples. Overall, I considered a founder carrying an SV if at least one type 1 experiment produced a product in one founder and it did not in Col-0 or in three other founders. The full set of primer pairs used is in Appendix E and a summary of the results, showing confirmed SVs and founder haplotypes that were shown to carry it is presented in Table 4.3.

In total, I validated 37 (82%) SV QTLs, comprising 61 (77%) breakpoints, in 14 translocations (5 cis and 9 trans) and 23 inversions (6 cis and 17 trans). In 11 SVs the breakpoints confirmed were polymorphic, while in 5 translocations the orientation of the transposed locus differs between founders, thus some carry a translocation and the others a translocation and inversion.

4.10 Effects of SVs on physiological phenotypes

Large SVs are known to have important phenotypic effects in humans [53, 25, 101, 103]; however their effects on *A. thaliana* phenotypes have not been reported yet. I investigated whether SV traits were associated with 44 physiological phenotypes related to germination, bolting, leaf development and ionomics that had been previously measured in the MAGIC lines. I conjectured there could be two ways that SVs might be associated with phenotypes. First, there could be a direct correlation between an SV trait and a phenotype, without regard to any QTLs. Second, an SV trait could map to an SV QTL that overlaps with an existing phenotypic QTL. To look for direct correlations, I computed Pearson correlations and their corresponding p-values for each physiological phenotype with every SV trait in each of the six read anomaly types. I performed $11,915$ correlation tests for each anomaly type and for each locus A obtained a p-value p_A using the Fisher transformation of the correlation coefficient r_A. I selected significant correlations in which $\log P < 4$, which corresponds to $r > 0.17$ for a sample size of 488. I filtered out any correlations that were driven by up to three outliers, thereby testing whether the removal of the three most extreme samples would

N	Founders	source	sink	SV	QTL type	Inv founders
1	14	chr3, 16.28Mb	chr3, 15.65Mb	translocation,inversion	cis	1
2	10	chr1, 17.39Mb	chr5, 13.74Mb	translocation,inversion	trans	9
3	NA	chr3, 14.43Mb	chr3, 13.49Mb	inversion	cis	NA
4	NA	chr3, 2.09Mb	chr4, 2.77Mb	translocation	trans	NA
5	4	chr3, 11.9Mb	chr3, 0.63Mb	inversion	trans	NA
6	11	chr5, 12.63Mb	chr5, 10.04Mb	translocation,inversion	trans	9
7	13	chr2, 7.2Mb	chr3, 15.49Mb	translocation,inversion	trans	12
8	4	chr1, 9.07Mb	chr1, 24.07Mb	inversion	trans	NA
9	7	chr1, 12.24Mb	chr1, 11.8Mb	inversion	cis	NA
10	13	chr5, 14.88Mb	chr5, 16.33Mb	inversion	cis	NA
11	8	chr3, 14.66Mb	chr3, 14.32Mb	inversion	cis	NA
12	18	chr4,1.57Mb	chr4, 2.77Mb	inversion	cis	NA
13	NA	chr5, 10.2Mb	chr5, 15.07Mb	inversion	trans	NA
14	2	chr5, 6.29Mb	chr3, 15.89Mb	inversion	trans	NA
15	NA	chr3, 12.55Mb	chr4, 3.57Mb	inversion	trans	NA
16	15	chr4, 4.53Mb	chr4, 1.82Mb	translocation	trans	NA
17	2	chr1, 16.85Mb	chr4, 5.81Mb	translocation	trans	NA
18	1	chr3, 13.31Mb	chr3, 12.65Mb	translocation	cis	NA
19	4	chr4, 8.22Mb	chr5, 15.9Mb	inversion	trans	NA
20	1	chr3, 16.27Mb	chr3, 15.7Mb	inversion	cis	NA
21	15	chr3, 22.55Mb	chr3, 17.63Mb	inversion	trans	NA
22	NA	chr2, 9.18Mb	chr2, 2.59Mb	translocation	trans	NA
23	15	chr1, 14.94Mb	chr3, 14.76Mb	inversion	trans	NA
24	3	chr5, 18.83Mb	chr5, 17.03Mb	inversion	trans	NA
25	17	chr5, 9.2Mb	chr5, 9.63Mb	translocation	cis	NA
26	16	chr3, 12.97Mb	chr5, 10.8Mb	translocation	trans	NA
27	11	chr1, 17.69Mb	chr5, 15.95Mb	inversion	trans	NA
28	2	chr2, 1.84Mb	chr3, 16.75Mb	inversion	trans	NA
29	14	chr3, 12.86Mb	chr3, 13.48Mb	translocation	cis	NA
30	1	chr4, 3.72Mb	chr2, 6.00Mb	inversion	trans	NA
31	14	chr1, 13.35Mb	chr4, 7.43Mb	translocation	trans	NA
32	16	chr1, 15.86Mb	chr1, 13.47Mb	inversion	trans	NA
33	3	chr2, 2.16Mb	chr1, 16.42Mb	inversion	trans	NA
34	11	chr4, 3.67Mb	chr5, 12.08Mb	inversion	trans	NA
35	4	chr5, 14.82Mb	chr4, 7.05Mb	inversion	trans	NA
36	2	chr3, 14.89Mb	chr3, 16.23Mb	translocation	trans	NA
37	2	chr3, 14.25Mb	chr2, 4.85Mb	inversion	trans	NA
38	1	chr1, 16.89Mb	chr1, 12.86Mb	translocation	trans	NA
39	15	chr5, 16.69Mb	chr5, 16.25Mb	inversion	cis	NA
40	1	chr3, 3.64Mb	chr1, 24.28Mb	inversion	trans	NA
41	2	chr4, 2.78Mb	chr4, 1.96Mb	inversion	trans	NA
42	4	chr1, 9.38Mb	chr1, 8.5Mb	translocation,inversion	cis	1
43	NA	chr3, 14.51Mb	chr3, 13.49Mb	inversion	cis	NA
44	NA	chr3, 15.06Mb	chr2, 5.18Mb	translocation	trans	NA
45	NA	chr2, 2.91Mb	chr2, 5.53Mb	translocation	trans	NA

Table 4.3: SV QTLs validated by PCR. The columns are: **N**: experiment id, **Founders**: number of founders in which an SV was confirmed, **source, sink**: position of source and sink of the SV QTL, **type**: type of structural variant **QTL type**: cis or trans, **inv founders**: in cases where an inversion was detected in a subset of founders, the number of founders carrying an inversion

reduce correlation below the significance threshold. In total I found 549 SV traits associated with 40 phenotypes. Each physiological phenotype had on average 1.56 associated SV traits of the same read anomaly type (max = 21).

I measured the overall effect of SVs on each physiological phenotype, using a heritability-like measure, h^2_{SV}, defined as the fraction of the phenotypic variance that could be attributed to SVs. I estimated h^2_{SV} as the proportion of variance explained by the SV traits that were deemed to be associated with the phenotype, using linear models. Let $\mathbf{y}$ be the vector of phenotypic values for a specific physiological trait across the MAGIC lines. Let the k significantly associated SV traits (of the same type) be $\mathbf{X_1}, \ldots \mathbf{X_k}$, represented by the matrix $\mathbf{X}$. Then the phenotype is modelled as

$$\mathbf{y} = \mathbf{X}\alpha + \mathbf{e} \tag{4.25}$$

I estimated the k parameters α by least squares using the `glm()` function in R and computed the residual sum of squares RSS. The heritability h^2_{SV} is defined as:

$$h^2_{SV} = 1 - \frac{\text{RSS}}{\text{TSS}} \tag{4.26}$$

where TSS is the total sum of squares (i.e. the variance of $\mathbf{y}$). I also computed the individual effect sizes of all traits contributing to the heritability, by fitting simple linear regression models using each of the traits as the single independent variable. Table 4.4 summarises the phenotypes in which large effect sizes were detected (above 10%). Based on this analysis, SV traits can explain up to 33% of the total phenotypic variance of a single trait.

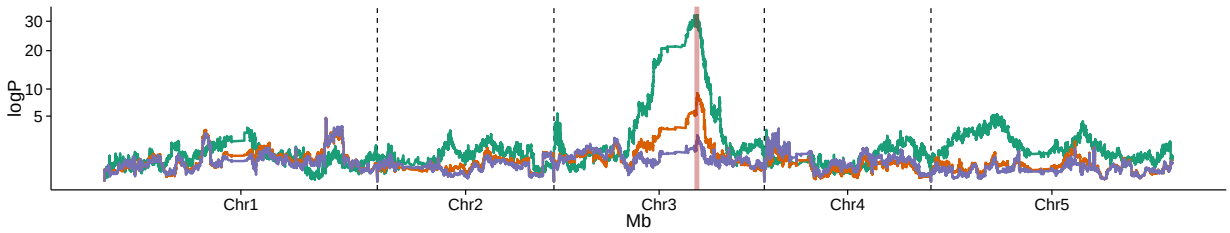
Phenotype	**SV trait type**	w	**N**	**SV QTL**	**max contrib.**
days.to.germ	same strand pairs	26.30%	13	4	7.30%
days.to.germ	unpaired reads	17.83%	6	3	8.15%
bolting	same strand pairs	11.10%	4	2	4.59%
Resistance	large insert size	18.33%	5	5	17.60%
Resistance	improperly paired reads	16.40%	5	5	15.73%
Resistance	unpaired reads	19.57%	3	2	8.72%
Resistance	unpaired or large insert size reads	17.07%	5	5	16.86%
Resistance	same strand pairs	24.71%	5	2	9.04%
ttl_branch.BATH	unpaired reads	11.01%	3	1	4.97%
ttl_branch.BATH	same strand pairs	13.42%	4	1	3.81%
fieldFT.pl	unpaired reads	16.06%	3	0	7.17%
fieldFT.pl	same strand pairs	12.21%	2	2	8.48%
ft.mean.h	unpaired reads	16.61%	4	2	6.93%
ft.mean.h	same strand pairs	10.52%	2	1	6.32%
fieldRD.pl	unpaired reads	17.01%	3	1	6.21%
Htbranches.CV	unpaired reads	24.71%	21	5	9.91%
Htbranches.CV	same strand pairs	29.85%	16	2	8.01%
leaves.day.28.given.days.to.germ	unpaired reads	10.63%	6	2	5.94%
RosetteLeafNumber.ShortDay	unpaired reads	22.13%	6	1	11.27%
RosetteLeafNumber.LongDay	unpaired reads	14.18%	5	3	2.72%
RosetteLeafNumber.LongDay	same strand pairs	23.92%	11	6	3.18%
Na	unpaired reads	22.95%	6	2	8.05%
Na	same strand pairs	23.42%	6	2	8.78%
Mn	unpaired reads	22.49%	7	1	8.81%
Mn	same strand pairs	33.25%	17	10	7.74%
Mg	same strand pairs	16.64%	3	3	7.65%
Ca	same strand pairs	21.53%	7	4	7.34%
Ni	same strand pairs	18.08%	4	0	7.37%
Zn	unpaired reads	29.50%	13	3	12.54%
Zn	same strand pairs	34.70%	13	4	13.81%
As	unpaired reads	12.50%	3	0	6.53%
As	same strand pairs	18.32%	3	0	11.26%
Rb	same strand pairs	13.97%	3	2	6.48%
Mo	same strand pairs	12.81%	3	2	5.78%

Table 4.4: Physiological phenotypes with large ($> 10\%$) SV effects. The columns are: **Phenotype** - physiological phenotype, **SV trait type** - type of SV traits (indicating a type of read pair anomaly), w - total phenotypic variance explained by SVs, **N** - number of associated traits, **SV QTL** - number of highly associated SV traits that have a sink QTL, **max contrib.** - the maximum contribution of a single trait of that type to the phenotype

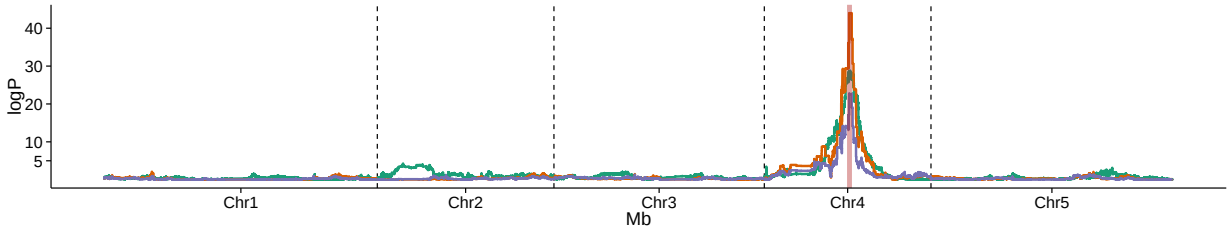
I then asked if any physiological QTLs could be explained by SV traits. I mapped QTLs for the phenotype residuals after first regressing out all the associated SV traits/each associated SV separately, and found that some of the associated effects might be causal. For example, for the physiological phenotype germination time, regressing out a single cis-SV that explained 8.13% of the variance, which lay under a germination QTL, ablates the QTL (Figure 4.18a). The source and sink of this SV are at $15.94, 15.93$Mb respectively, within $\sim$ 20kb from the zinc induced facilitator-like ZIFL 2 gene (at 15.66Mb), which was recently shown to control caesium (Cs(+)) and potassium (K(+)) channels [107]. Dysregulation and loss of function of ZIFL2 leads to excess supply in (K(+)), which has been shown to decrease seed germination [39].

Similarly, for resistance in the fungus *Albugo laibachii*, there are 5 cis SV QTLs, (mapped for almost all read anomaly traits), that explain up to 24.71% of the total phenotypic variance (Figure 4.18b). Regressing out two of them, both lying directly under a resistance QTL on chromosome 4 at the region between $8.5 - 10.5$Mb, leads to substantial drop of the association for resistance. The two most highly associated SVs are at 9.50Mb and 10.44Mb. The region under the QTL is rich in resistance genes that could be causal. The gene at the peak of the resistance QTL, which is also exactly at the sink SV QTL of one of the two SVs is RPP4 [129]. In Asian soybean CNVs of RPP4 have been detected and shown to affect susceptibility to pathogens [91]. The other highly associated SV is at 10.44 Mb, and has 3 putative candidate resistance genes within $30kb$ from the sink, namely: EDR2 [121], AT4G19050 and AT4G19060.

Finally, of the 549 traits 253(46.1%) did not have a sink QTL. In general, the traits with sink QTLs have a higher effect size (55% of the top associated SV traits have sink QTLs), but there are examples of unmapped SV traits with large effect sizes. For example, 7.3% of variation in germination time is explained by an SV trait on chromosome 1 with no sink, while 16.6% of variation in field flowering time is shared between 3 unmapped SV traits.



(a) Effects of SVs on germination. The figure shows the manhattan plots of three genome scans (x-axis: genomic position and y-axis: genome-wide logP). The orange line shows the genome scan of germination (days to germination) - the peak is at chr3, 15.94Mb. The green line shows the genome scan of a single SV trait, measuring same strand reads, which was found to explain 8.13% of the total germination variance. Its source is at 15.93Mb (marked by a red line) and sink at 15.94Mb, at the same position as the germination QTL. The purple line is the genome scan of the residuals of germination, after regressing out the SV effects, in which the QTL is lost.



(b) Effects of SVs on pathogen resistance. The figure can be interpreted as Figure 4.18a. The peak of association for resistance is at chr4, 9.50Mb. Two SV traits have been regressed out, (sources: chr3, 9.50Mb and 10.44Mb) together they explain 24.7% of the phenotypic variance. The SV trait genome scan shown is at 9.50Mb and its sink at 9.49Mb.

Figure 4.18: Effects of SVs on QTLs for physiological phenotypes

4.11 Effects of SVs on gene expression

SVs are known to affect gene expression, however the amount of variation of gene expression down to SVs is not fully known. A recent study that used data from the 1000 genomes project found that most variation in gene expression can be explained by CNVs and has a very strong impact in genetics [42]. In the mouse a small number of SVs disrupting exonic sequence with large effects has also been detected [137].

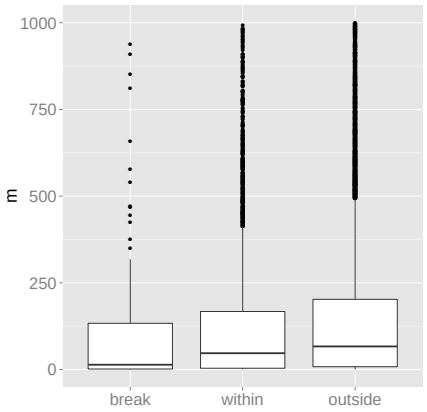
I investigated the effects of SVs on gene expression in MAGIC lines, using Illumina RNAseq leaf transcriptomes from 200 MAGIC lines. Expression levels for each gene were extracted from read counts normalised by fitting a negative binomial model in DeSeq [4]. Out of a total of $33,602$

genes in *Arabidopsis*, 21,747 were expressed. Due to mapping resolution issues, I only considered SVs with accurate breakpoints (from Section 4.9.2). In particular, 119 genes spanned the SV breakpoints and 6,909 laid inside them. The former are probably disrupted by SVs and the latter are not, but their copy-number or position in the genome is variable.

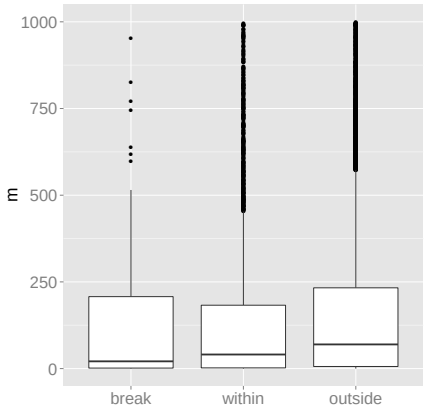
I used both raw read counts and filtered read counts. Filtered data excluded reads overlapping annotated introns, reads fully aligned to a region where multiple annotated genes overlap and reads that did not start within annotated exons. Applying the filters affected SV-related genes more than other genes as 18% of the transcripts spanning breakpoints and 14% of transcripts within breakpoints appeared unexpressed throughout all samples, but only 5% of all other genes were affected in the same way. From this I reasoned that some of these filters might be related to SVs so applying them might reduce SV-related signal, so I decided to also use the raw normalised read counts for comparison. From both datasets I removed genes expressed in fewer than 2% of samples.

Transcripts were divided in three categories: disrupted by breakpoints, within breakpoints (hence moved or with variable copy-number) and outside SVs. Though there was little difference in the mean expression of transcripts based on category (SV-related genes tend to have lower gene expression which is not statistically significant, Figures 4.19a, 4.19b), the gene expression variance, scaled by the mean, was elevated in genes spanning breakpoints (t-test comparing with genes outside SVs: $P < 10^{-2}$ for raw and filtered counts) and in genes within breakpoints (t-test comparing with genes outside SVs: $P < 10^{-13}$) as shown in Figures 4.19c, 4.19d). This difference in variance is mostly due to the larger fraction of silenced transcripts (i.e. with zero expression) for genes lying on SV boundaries or within SVs (t-test: $P < 10^{-4}, P < 10^{-30}$ for raw counts and $P < 10^{-3}, P < 10^{-52}$ for filtered) (Figure 4.19e, 4.19f). Genes on breakpoints are slightly more likely to be silenced than genes within breakpoints (t-test : $P < 0.02$). Table 4.5 summarises all the t-test p-values comparing between categories of transcripts relative to SVs.

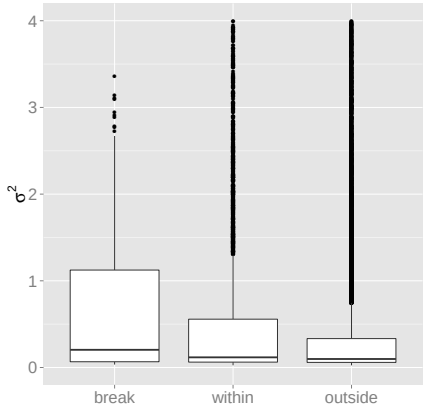
All these observations suggest the SVs dysregulate gene expression, even when the gene sequence itself is undisturbed.



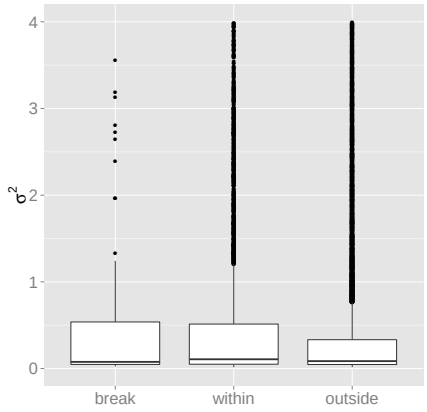
(a) mean m, raw normalised counts



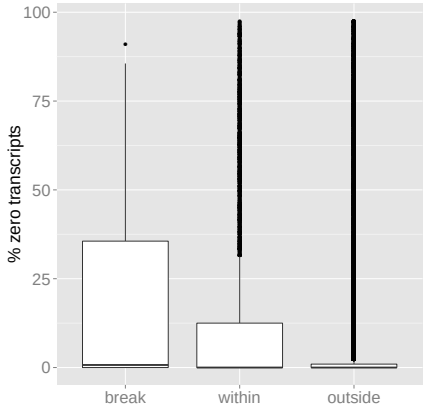
(b) mean m, filtered normalised counts



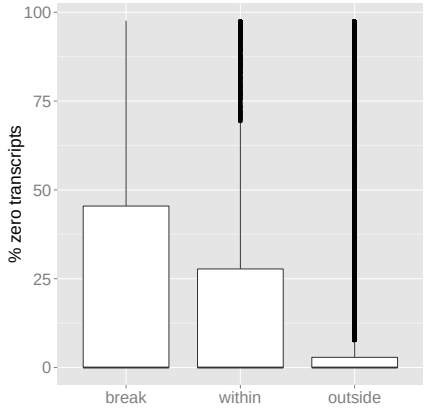
(c) variance σ^2, raw normalised counts



(d) variance σ^2, filtered normalised counts



(e) % zero transcripts, raw normalised counts



(f) % zero transcripts, filtered normalised counts

Figure 4.19: Comparison of gene expression transcripts for genes spanning SV breakpoints, genes within SVs and all other genes

	Bps vs Out		Bps vs In		In vs Out	
test	**R**	**F**	**R**	**F**	**R**	**F**
mean	0.51	0.64	0.59	0.71	0.19	0.12
var	$4 \cdot 10^{-3}$	$9 \cdot 10^{-3}$	0.13	$9 \cdot 10^{-2}$	$4 \cdot 10^{-33}$	$4 \cdot 10^{-13}$
% zero	$2 \cdot 10^{-5}$	$1.2 \cdot 10{-}2$	$1.4 \cdot 10^{-2}$	0.12	$6 \cdot 10^{-30}$	$2 \cdot 10^{-52}$

Table 4.5: P-values of one-sided t-tests comparing gene expression transcripts spanning SV breakpoints (**Bps**), within SV breakpoints (**In**) and outside SVs (**Out**). Columns labelled **R** show tests using raw counts, while **F** show filtered counts. The rows indicate comparisons of mean, variance and fraction of null gene expression per gene across all samples in each of the three SV-related categories. The alternative hypotheses were that SV-related genes had lower mean gene expression, higher variance and higher fraction of silenced genes.

4.12 Discussion

This chapter introduced a method for the identification of common structural variants segregating in a population, using population-based low-coverage sequencing data. Using genetic variation that underlies structurally variant regions I have jointly called structural variants in a population and also mapped long-range structural variants derived from remote locations (translocations, duplications), which are difficult to identify from next-generation sequencing data alone. The method allows one to readily distinguish between local and long-range SVs, and determine whether an SV is inverted.

Whilst I focused here on a population of recombinant inbred lines, the method is applicable with certain modifications to any sequenced population, including outbred species such as humans. I exploited the known ancestry of the MAGIC lines to generate haplotype mosaics and as a source of contigs to confirm breakpoints. In other species haplotype maps and careful *de novo* assemblies of a small number of individuals could be used similarly. High trait variances are good indicators of the presence of SVs (Figure 4.6); therefore instead of a whole genome scan, one could map target loci with high variance in anomalous read alignments without many false negative predictions, making the problem much more tractable in large genomes. The method could also be used to facilitate read-mapping software to correctly align ambiguous read alignments.

Despite the usual paradigm of using medium or high-coverage sequencing for SV mapping [142,

66, 21] we have seen that low-coverage sequencing data are suitable for structural variation mapping. The main benefits of this approach is that it can effectively resolve long-range SVs while offering the possibility of sequencing larger population samples, and get population-wide estimates for structural variation. Nevertheless, its predictions rely on mapping resolution, so high-coverage sequencing data are needed in order to pinpoint exact SV breakpoints. Novel sequencing technologies based around very long reads [18, 54] are also capable of resolving structural variants in individuals with high precision in breakpoints. However they are expensive and slow compared to low-coverage short-read sequencing and are therefore limited to small numbers of genomes. Similarly, complete *de-novo* assembly of reference genomes using a combination of short-read libraries with a range of insert sizes remains a challenge [38], although the incorporation of population data can aid this process [20]. By combining these technologies, i.e. sequencing a few representative individuals with expensive long reads, and the population as a whole with cheaper short reads, it will be possible to impute SVs, particularly translocations, into the larger population analogously to the way that reference haplotype panels such as 1000 genomes are used to impute SNPs and small indels.

In the *Arabidopsis* MAGIC population, we have found that structural variations are surprisingly common, affecting about a fifth of all genes. About a quarter of SVs are translocations. The *Arabidopsis* genome is thought to have emerged from successive chromosomal fissions and fusions, and it has been conjectured to be excessively rearranged [131, 35, 132]. In MAGIC, 45% of SNPs are private to a single accession [36], but the fraction of private SVs is much lower, about 12%. Therefore, it seems plausible that some SVs were formed before SNPs and thus may be related to the divergence and speciation of *A. thaliana*. Creating synteny maps of rearranged *de novo* contigs from *A. thaliana* with related species, such as *Arabidopsis lyrata* and *Cardamine hirsuta* would help resolve which SVs correspond to the ancestral configuration of loci.

Typically, detected SVs are small, of the order of 50kb. We did not find any very large (megabase-sized) common SVs; the largest being 334kb. Thus most SVs contain about 10 genes, implying the sequences of most genes and their promotors in SVs are not directly altered by the SV, although the genomic context tens of kb away is changed. This change of context appears to dysregulate gene expression, often causing silencing and increasing the variability of expression.

Presumably this is a consequence of alterations to local DNA-DNA interactions, which could be tested by chromosome conformation capture experiments. SVs are therefore strong candidates for QTLs that cause variability in phenotypes [109].

Furthermore, some SVs lie with QTLs for physiological traits, and could be causal variants since they explain most of the variation at the QTL. If so this lends support to the observation that many QTLs cannot be resolved to a single causal SNP but instead are an irreducible set of linked causal variants (a haplotype effect) [27, 124]. We have also shown that certain read anomaly traits cannot be mapped as SVs but nevertheless explain significant fractions of phenotypic variation. Some of these may be interpreted as polymorphic loci absent from the reference genome, thereby representing a source of missing heritability [37].

Chapter 5

Recombination and clusters of mosaic breakpoints in the MAGIC lines

As we saw in Chapter 3 recombination events can be mapped to within 2.5kb in MAGIC, so it is possible to reconstruct ancestral mosaics with high accuracy. The mosaics can be viewed as a partial recombination history of the population, and so can be used to draw inferences about recombination in *A. thaliana*. This chapter analyses the mosaics generated in MAGIC using the original variant calls in the 19 MAGIC founders [36], which were used for recombination analysis. More than half of the lines had a much larger number of mosaic breakpoints than expected, which formed clusters within each genome in different positions and with different haplotype combinations. My hypothesis was that they could either be artefacts caused by structural variation, introgression of non-MAGIC genomes in the population during breeding or residual heterozygosity or that they could be unusual signatures of recombination.

Clusters of recombination breakpoints are consistent with some of the relevant literature in *A. thaliana*; for example, recombination in Arabidopsis has a disease-resistance component and disease resistance gene families such as the R-genes are shown to substantially increase local recombination rates and induce double strand crossovers after each generation [23]. Furthermore, very high gene conversion rates have been reported [138], as well as non-interfering crossover regions in which recombinants tend to form close to each other [8]. We sequenced 9 cluster genomes at high

coverage and performed extensive data analysis which compared properties of cluster breakpoints (with respect to introgression, sequencing quality, number of SVs and heterozygosity) to normal breakpoints and to random genomic regions in order to test their validity. Many of these tests validated the clusters, in the sense that they did not detect any differences within and outside clustered breakpoints, however some showed evidence of artefactuality. A small number of clusters were linked to structural variants. More importantly, heterozygosity was elevated in some of the high coverage genomes and explained some of the clusters. Furthermore, running the diploid reconstruction algorithm from Section 3.3 over high and low-coverage data reconstructed some of the clusters as heterozygous. I resolved some of these observations by using revised sets of sequence variants, which revealed previously undetected heterozygosity in the 19 founders, which also gave rise to a large proportion of the clusters. The lesson learned from this analysis is that mosaic reconstruction is complex and can be affected by multiple sources of error, caused mainly by unexpected genetic features. Nevertheless, some of the clusters are still present in the revised data and need to be investigated further.

5.1 Recombination and genomic instability

Genetic recombination is the process by which a double-stranded DNA molecule breaks and joins with its sister chromatid. In chromosomal crossover, a pair of homologous parental chromosomes exchange DNA to form a new chromosome. Other recombination events include gene conversions i.e. double-strand breaks in which a DNA segment replaces its homologous sequence so that they become identical, and non-homologous recombinations in which a locus is deleted or translocated to a non-homologous region. Throughout this thesis, the term recombination refers to homologous events (crossovers or gene conversions).

Although recombination events can occur anywhere in the genome, recombination is not a random process. Most recombination events are concentrated in *recombination hotspots*, defined as short loci (typically $1 - 2$kb long) within the genome with significantly higher recombination rate [79, 55]. The existence of recombination hotspots implies that recombination is, to some extent, associated with other biological processes.

In humans,a 13-base pair sequence motif attracts more than 40% of chromosomal crossover events [96]. This motif is a binding site for the zinc-finger protein PRDM9 [95, 90]. The same protein has been also identified as a determinant for recombination in mice [9]. Moreover, variation in the genetic maps (and therefore, hotspot positions) of human sub-populations of African and European descent can be explained by variation in PRDM9 alleles and binding sites across these populations [48].

Recombination events per meiosis have been observed to be more evenly spaced within a chromosome than would have happened at random. This tendency of crossovers to preferentially occur far from each other is known as *crossover interference* [47]. Human and mouse recombination displays strong positive interference [51, 11, 12]. Sex-specific crossover pathways have been observed in the mouse [100], accounting for the differences in recombination rates between sexes.

5.1.1 Recombination in *A. thaliana*

Whilst plant recombination hotspots and motifs certainly exist [23, 50], no functional equivalent of PRDM9 has been found. Multiple pathways have been shown to affect recombination, each with distinct features [69, 8, 140, 23, 7]. Firstly, negative interference pathways have been detected, implying that crossovers may not always avoid close proximity. In particular, it has been shown that a fraction of meiotic crossovers in three *A. thaliana* chromosomes include both positive and negative interference pathways [69]. Moreover, recombination hotspots in which crossovers occur preferentially without interference have been identified, while the number of non-interfering crossovers is higher in male meiosis [8]. Epigenetic changes such as DNA methylation and histone modification also affect recombination rates [140], especially within the heavily methylated pericentromeric regions, increasing crossover frequencies. Environmental and pathogen stress are also recombinogenic. Chromatin modifications are associated with hotspots and with gene families that regulate response to disease (R-genes, defensin genes) [23]. Furthermore, transgenic resistance gene insertions influence recombination frequencies [7], while environmental changes induced by pathogen attacks also increase recombination [82]. Finally, because the germ line of plants is not separate from the soma, mitotic recombination events can be transmitted [40].

The first large-scale study on reccombination in *A. thaliana* used ancestral recombination rates inferred by linkage disequilibrium (LD) in 19 natural accessions to identify 260 hotspots [62]. This study showed that gene content is low within recombination hotspots and LD decays faster near intergenic regions, implying recombination is more likely to occur in intergenic regions rather than within genes.

Another survey analysed recombination in *A. thaliana* in 1300 *A.thaliana* genomes, and confirmed that intergenic regions are recombinogenic, showing further that genes in repeat families are enriched in the regions with the highest recombination rate [50].

5.2 Lineage-specific recombination

It is generally thought that recombination outside hotspots makes little overall impact, since each such event is random and therefore contributes little to the genetic map which is based on recurrent behaviour. Consequently, these sporadic recombinants are invisible to LD-based studies of natural populations predicated on the existence of genetic maps [24, 50].

Mosaic populations offer an alternative approach to recombination analysis: a genetic mosaic can be viewed as a partial recombination history of the population. In the MAGIC lines, each mosaic partially[1] encapsulates the recombination history of a line and as shown in Chapter 3 we can map fine-scale recombination events in MAGIC, using low-coverage sequencing data. Hence we can analyse both events that occured within and outside recombination hotspots. Furthermore, recombination hotspots in natural populations could represent a single ancient event that a large fraction of the population inherits; in a mosaic population all recombination events observable are unique (as each genome is a separate lineage) and recent. Therefore, recombination hotspots in a mosaic population indicate genomic regions that are currently active with respect to recombination.

However, it is important to distinguish true recombinations from sequencing artefacts, that can occur in mosaics over highly repetitive, divergent, rearranged or structurally variant loci. For example, reads mapping to a translocated genomic segment will be aligned to the incorrect location

[1]The mosaic of a MAGIC line captures a partial only recombination history because after the selfing stage a genome carries only half of the recombinants that occurred during intercrossing. Also, recombinantion events between loci with both strands carrying the same haplotype are invisible.

Type	**Number**	**Median segment size (Mb)**	**Mean segment size (Mb)**
Non-cluster	16,314	2.065	3.409
Cluster	5,012	0.136	0.214
All	21,326	1.085	2.658

Table 5.1: Nonrecombinant segment size statistics for 476 MAGIC lines and for regions inside and outside clusters. A segment is a genomic interval between two successive mosaic breakpoints.

in the reference, possibly causing false recombination breakpoints and apparent heterozygosity [132].

5.3 Genome mosaics inferred by IMR/DENOM

In Section 3.6 I presented mosaics inferred using GATK [89] allele calls. I initially generated mosaics based on the 3.07M SNPs reported in [36] called using the IMR/DENOM pipeline variant caller. After excluding heterozygous or triallelic SNPs as well as SNPs within known transposons or repeats, the set comprised 2.6M SNPs in total. On average there were 325k SNPs for each individual genome. This is a larger set of SNPs than the one used in the rest of the thesis. As we saw in Section 3.2 the two sets differ in their estimation of heterozygous SNPs: after filtering out heterozygous and triallelic SNPs, there are 1.3M SNPs left ($300k$ per line on average) in the GATK set, and 2.6M SNPs (325k per line on average) in the IMR/DENOM. Also the original set of SNPs includes only 476 genomes instead of 488, due to a sample renaming problem (12 genomes had duplicated names and were discarded from the set).

For the most part, the IMR/DENOM reconstructed genome mosaics were consistent with the *Arabidopsis* genetic map. Namely, they comprised 21,326 unrecombined blocks of founder haplotypes with a median length of 2.065Mb (Table 5.1). In total, 18,946 recombination breakpoints ($21,326 - (5 \times 476)$) were predicted. For example, Figure 5.1a shows the reconstruction of MAGIC.492 chromosome 3 from low-coverage sequence. Cluster definition, properties and verification are examined in detail in the remainder of this chapter.

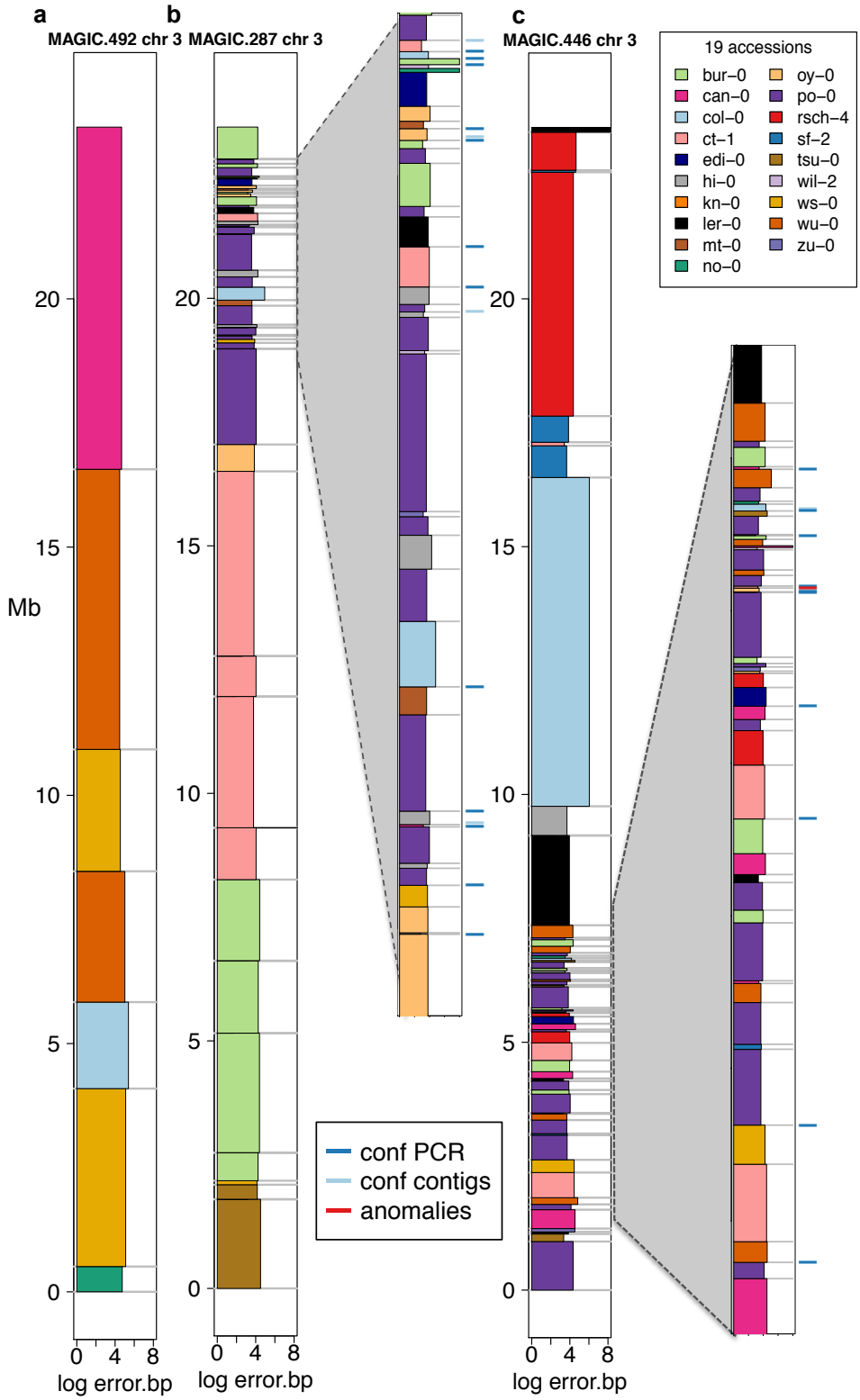
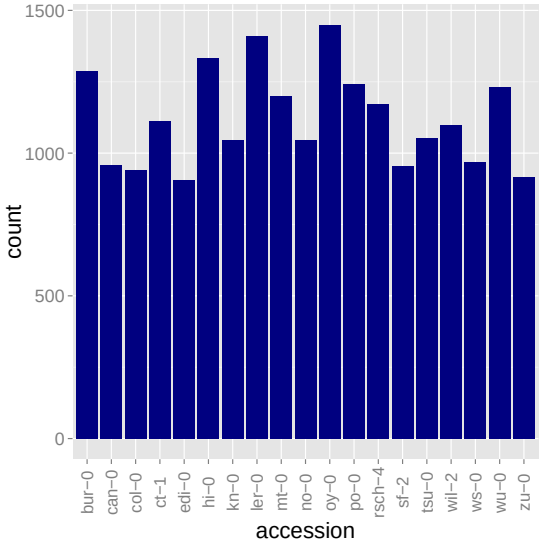
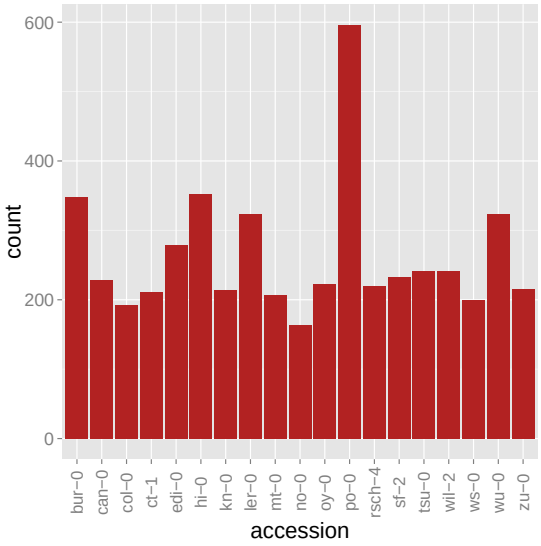


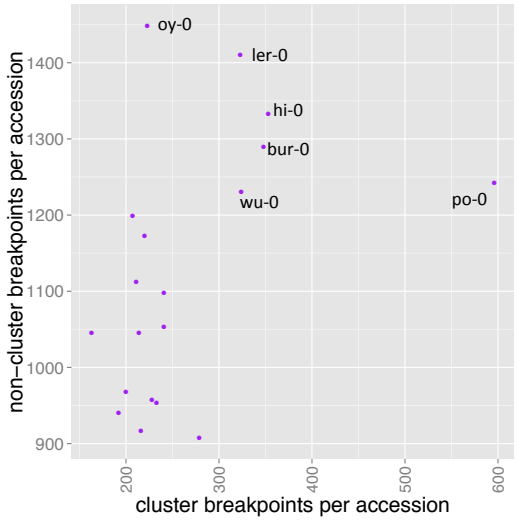
Figure 5.1: MAGIC genome mosaics. (a-c) Examples of mosaics. y-axis: genome position (Mb). Coloured vertical segments indicate genomic regions descended from the corresponding founder as in Legend. x-axis: negative log10 of the discrepancy rate between called SNPs from high coverage sequence data and alleles predicted from the imputed mosaics (this is equivalent to the genotype error which was used to evaluate genetic mosaics in Section 3.3.1, but the error is averaged across the length of the segment instead of the number of variant sites). (a) MAGIC.492 chromosome 3, indicating a normal chromosome with seven haplotype segments. (b, c) MAGIC.287 and MAGIC.446 chromosome 3, showing examples of clusters, which are also shown magnified (grey regions). Lines next to breakpoints mark validation results (Dark blue: breakpoints validated by capillary sequencing; pale blue: breakpoints validated by de-novo contigs; red: disproved breakpoints).



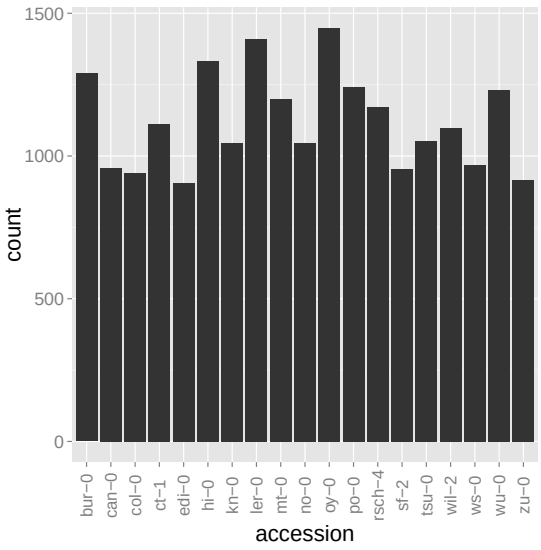
(a) Non-cluster segments



(b) Cluster segments



(c) Cluster (x) and non-cluster (y) breakpoints assigned to accessions



(d) All haplotype segments

Figure 5.2: Distribution of founders assigned to mosaic haplotype segments within and outside clusters

5.4 Lineage-specific clusters of breakpoints

Clusters of mosaic breakpoints, defined as islands of consecutive short haplotype segments (Figure 5.1b,c), were detected in 280 lines (58.82%). Of the total 18,946 breakpoints predicted, 5,465 (28.8%) lied in 536 clusters.

The clusters were detected by a recursive algorithm based on the Smith-Waterman (SW) algorithm, described in [58] and [94]. The algorithm can detect islands of genetic features with distinct properties within the genome. In particular, the algorithm was tuned to find dense clusters of 4 or more breakpoints by assigning positive scores only to segments shorter than 500kb, while it gave all other segments negative scores. Positive scoring islands were then detected by dynamic programming. The islands could include segments larger than 500kb if they were surrounded by sufficiently many shorter segments (Algorithm 3).

The average number of breakpoints per cluster was 9.35. Occasionally clusters were very large, affecting entire chromosomes. The distribution of clusters per line is plotted in Figure 5.3.

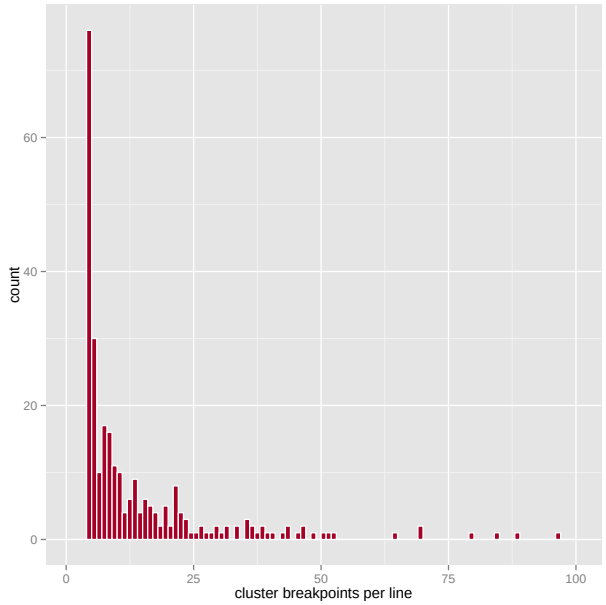


Figure 5.3: Frequency distribution of number of cluster breakpoints per line (for those lines with clusters).

About a quarter (118 lines, 24.79%) of the MAGIC genomes had multiple clusters, often on different chromosomes. However, most MAGIC lines had completely normal rates of recombination

Data: N - haplotype segments in chromosome

$l(i)$ - length of i-th segment in bp

Initialisation:

$$X(i) = 500,000 - l(i)$$

$$S(0) = 0$$

$$B(0) = 0$$

Recursion:

$$S(i) = \begin{cases} S(i-1) + X(i), & \text{if } S(i-1) + X(i) > 0 \\ 0, & \text{otherwise} \end{cases}$$

$$B(i) = \begin{cases} B(i-1), \text{if } S(i) > 0 \\ i, \text{otherwise} \end{cases}$$

Algorithm 3: A Smith-Waterman type recursion that detects islands of consecutive short haplotype segments (breakpoint clusters) in the mosaics.

except perhaps for one or two clusters spanning ∼2Mb. Furthermore, clusters occured genome-wide, as on average eight lines had a cluster spanning any randomly chosen location (Figure 5.4), and 1.68% of the genome was covered by a cluster in a given MAGIC line.

The clusters had distinct properties compared to classical recombination breakpoints in the mosaics. First, while the median segment size in the mosaics was 2.07Mb the median segment size within a cluster was only 136kb (Figure 5.6a), while the average total length of a cluster was 2.004Mb. Second, most of the cluster breakpoints appeared unique (i.e. specific to individual lineages): some of the clusters comprised two or three alternating haplotype segments, while others had a more complicated structure and involved a large number of haplotypes. The difference between the two types of cluster can be seen in Figure 5.5. There are relatively few cluster hotspots (as distinct from classical recombination hotspots). Lastly, the numbers of cluster and non-cluster breakpoints within a line are uncorrelated (Figure 5.6b).

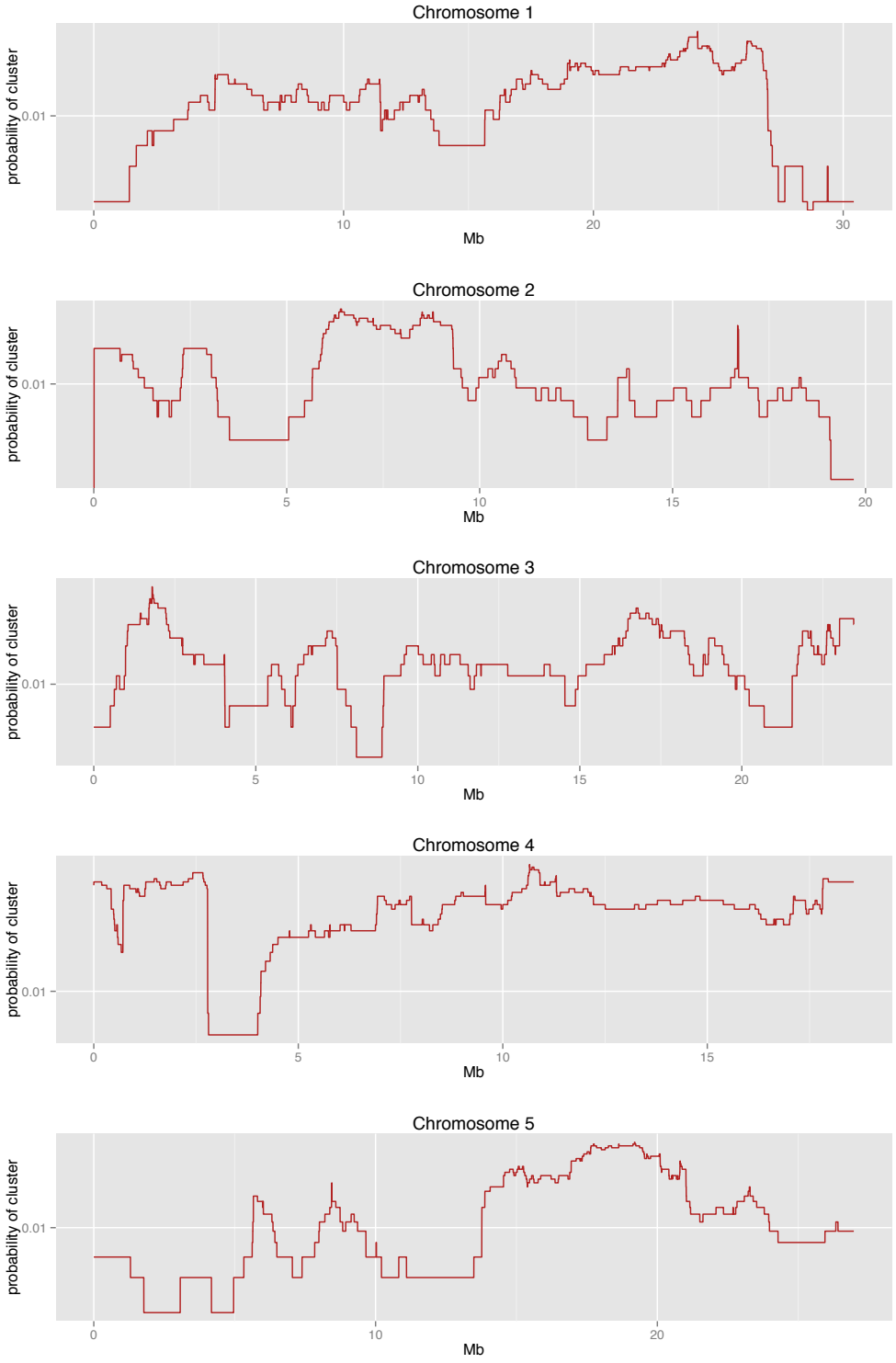


Figure 5.4: Spatial distribution of the probability that any genomic position is covered by a cluster. It is the result of averaging cluster occurrences over any one genomic location. Y-axis is in logarithmic scale.

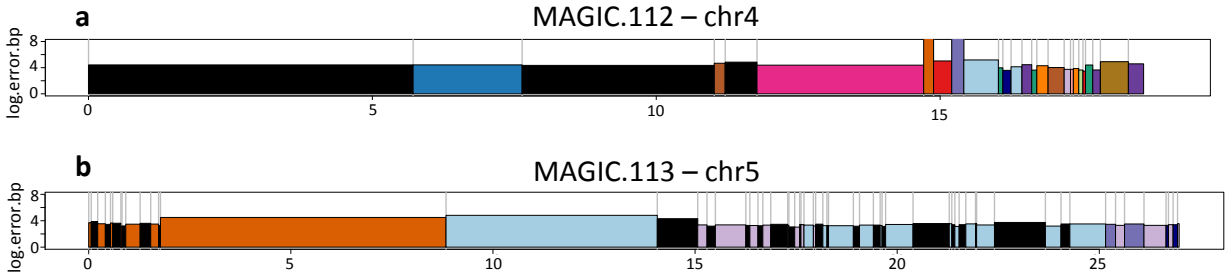
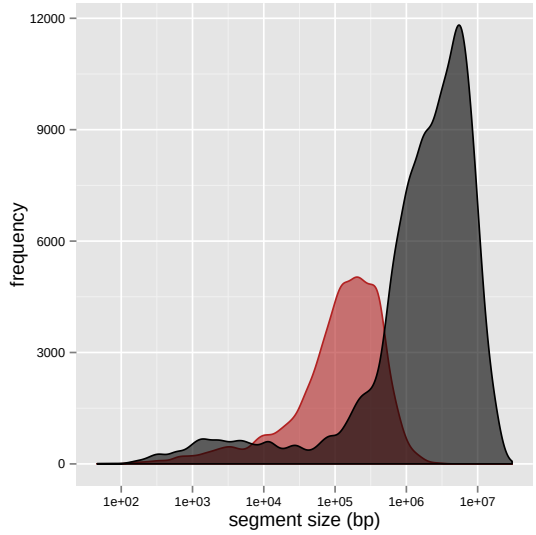
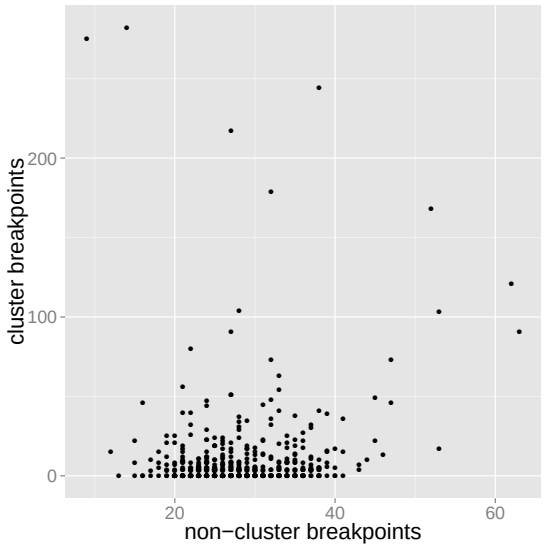


Figure 5.5: Examples of cluster mosaic with different characteristics. The cluster in (a), at the end of chromosome 4 of line MAGIC.122 carries many different haplotype segments, while in the cluster in (b) the cluster comprises of a small number of alternating haplotypes.



(a) Distribution of sizes of nonrecombinant founder haplotype segments for non-cluster (black) and cluster (red) segments. y-axis: frequency, x-axis: sizes in bp (log-scale).



(b) Scatterplot comparing the numbers of cluster (y-axis) and non-cluster (x-axis) breakpoints per line. Each point represents one MAGIC line.

Figure 5.6: Comparison of cluster and non-cluster haplotype segments

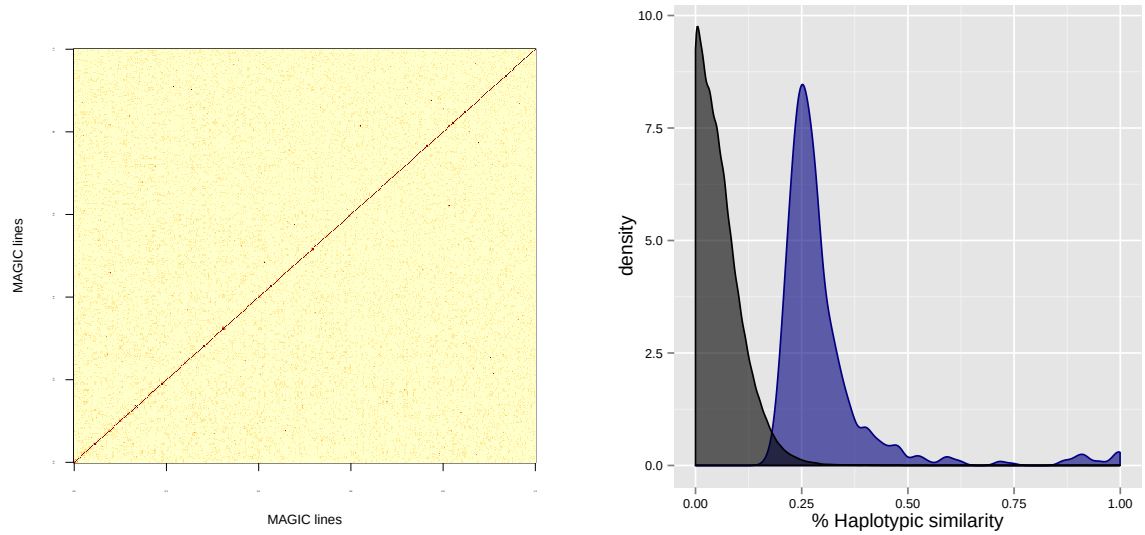
5.5 Cluster origins

The shared lineage structure of the cousin MAGIC lines can be used (Figure 2.4) to infer when the clusters arose. During the breeding of MAGIC, the inter-crossing (funnel-breeding) stage produced F4 families. From each family up to 3 MAGIC lines are generated by selfing. These "cousin" lines are expected to share about 25% of their genome.

Due to multiple sample renaming, historical breeding information on lines sharing the same funnel (i.e. cousins) was unreliable. However, it was possible to determine cousin lines by computing the pairwise haplotypic similarity between all MAGIC lines. I sampled the predicted haplotype every 1kb across the population and performed all pairwise comparisons. A heatmap illustrating the pairwise haplotypic similarity per MAGIC line is presented in Figure 5.7a. As expected, there is little population structure in MAGIC, since the average haplotype similarity of two arbitrary lines is 5.7% (median similarity = 4.99%). Any pair of MAGIC lines that shared 20% of their genome or above was considered a putative cousin pair. I found 1653 (1.46%) such pairs; 427 lines shared over 20% of their genome with more than two other lines. Therefore, for each line I selected the three lines with the highest haplotypic similarity, as cousin candidates, and formed a list of genomes with high haplotypic similarity. Any pair on this list will be hereafter referred to as a cousin pair, although some of these pairs may not actually originate from the same F_4 family. The mean similarity between cousins list was 30.36% (median 26.65%). The distribution of pairwise haplotypic similarity in cousins and non-cousins is shown in Figure 5.7b.

I also searched for local haplotypic similarity between all pairs of co-localised clusters, regardless of the genome-wide haplotypic similarity of the MAGIC lines containing the clusters. A cluster, or a part of a cluster, was defined to be shared between two lines if the clusters overlapped and the local haplotypic similarity exceeded 80%. If a cluster was shared between lines that were also cousins, then it might be a genuine event occurring during inter-crossing. If it was shared by more than four lines or by lines that are not cousins then it could be a genomic rearrangement or the result of introgression of haplotypes other than the 19 founders in the population (see Section 5.6).

If the cluster was unique, then I examined the possibility that it occurred during selfing. Clusters induced by selfing should only involve two alternating haplotypes at any locus, as only two specific



(a) Pairwise haplotypic similarity in the MAGIC lines. Red corresponds to 100% similarity and white no similarity.

(b) Density of pairwise haplotypic similarity between all MAGIC lines (black) and between cousin lines (blue).

Figure 5.7: Pairwise haplotypic similarity in MAGIC

DNA strands recombine with each other. A cousin of such a line that did not contain the cluster should comprise unrecombined segments of one of the two cluster haplotypes. Thus, if the local haplotypic similarity of a cluster with a cousin was about 50% then the cluster likely occurred during selfing. Contrary, if a cluster is composed of alternating segments but does not bear any local similarity with its cousins it probably did not originate during selfing.

There are 19 (3.5%) cases of a cluster shared between cousins (i.e. the cluster arose before selfing), and 196 (36.6%) cases of unique clusters, (i.e. it arose during selfing, but a cousin exists with the same sequence context and without the cluster). In 271 (50.5%) of clusters the origin could not be determined, either because no cousin was available or because the sequence context was different. The remaining 50 (9.3%) clusters were each present in more than 3 MAGIC lines so they are artifactual (because of rearrangements or introgression). These data suggest that clusters more often occur during selfing, rather than intercrossing, and are usually unique in a single MAGIC lineage.

5.6 Validation of cluster breakpoints

As discussed in Chapter 4, mapping short reads to a reference genome can cause problems with variant calling near repeats or where the sequenced focal genome is diverged from the reference. For example, reads originating from a translocated genomic segment will align to the incorrect location in the reference, possibly causing false recombination breakpoints and apparent heterozygosity [132]. I investigated whether cluster breakpoints are genuine recombinants or computational artifacts due to:

i. Introgression of other *Arabidopsis* genomes, during crossing of the population. If genetic material of a random *Arabidopsis* accession was present in MAGIC genomes and was reconstructed in terms of the 19 founders it would produce a very dense mosaic showing its relatedness to the other 19 founders.

ii. Structural variants. Over a translocation, reads from the transposed segment would be misaligned to the position of the reference causing apparent heterozygosity and introducing errors in variant calling [132].

iii. Heterozygosity, representing either genuine heterozygosity or genome duplication, which would also introduce errors in variant calling.

To test these alternatives, I used higher coverage sequence data (12x) for nine lines containing clusters, listed in Table 5.2, sufficient to call the great majority of variants in each genome. I then interrogated the high-coverage genomes for evidence that cluster breakpoints differed from non-cluster breakpoints or from randomly-chosen loci. The majority of each genome is free from clusters and therefore serves as an internal control. Therefore, all quality metrics were evaluated in the $1,132$ cluster-breakpoint regions from the 9 MAGIC lines sequenced at high coverage, and compared with 420 non-cluster breakpoint regions and 900 randomly selected regions, 100 from each line. Each metric was evaluated in fixed-size windows around each region; window size varied between metrics, in order to obtain sufficient statistical power. I estimated a normality range for every metric and identified regions in each class that were abnormal with respect to that range.

Line	Total breakpoints	Cluster breakpoints	Clusters
MAGIC.105	282	244	9
MAGIC.149	154	91	10
MAGIC.175	94	49	6
MAGIC.287	120	73	6
MAGIC.329	183	121	7
MAGIC.338	132	104	10
MAGIC.426	156	103	6
MAGIC.433	220	168	8
MAGIC.446	211	179	8

Table 5.2: Mosaic breakpoint statistics of the 9 MAGIC lines resequenced at $12-13$x coverage. Shown are the total numbers of breakpoints, the numbers of cluster breakpoints and the numbers of clusters.

I performed Fisher's Exact Tests (FET) for each line, comparing the number of abnormal regions among cluster breakpoint regions against non-cluster breakpoint regions and random regions.

I also used *de novo* assemblies and capillary sequencing to confirm breakpoints as well as independent SNP genotypes to confirm the accuracy of the genome mosaics away from breakpoints.

5.6.1 Introgression

I first tested whether introgression by *A. thaliana* accessions other than the 19 parents had occurred. This was unlikely a priori as no other accessions were grown in the same greenhouse during the construction of the MAGIC lines. Nonetheless, I searched for novel SNPs absent from the catalogue of variants segregating in the founders [36].

I called SNP sites in the 9 high-coverage MAGIC lines using Samtools [74]. On average $48,020$ novel sites per line were called, i.e. 1.7% of the total called per line, while the remaining 98.3% were already in the catalogue of sequence variants. 55.13% of novel calls were within known copy number variants from [36] and were discarded; therefore only 0.94% of the SNP calls in the nine high-coverage genomes were classified as novel. Novel SNPs in regions of interest were counted

in windows of 10kb. A region was considered normal if the number of novel SNPs was less than $N_\mu + 3N_{IQR}$, where N_μ is the median number of novel SNPs in a 10kb window throughout each genome, and N_{IQR} is the interquartile range. To compare counts of novel SNPs near cluster breakpoints with non-cluster breakpoints and random regions (that did not contain breakpoints) I used a FET for the following null hypotheses:

Hypothesis 1a: The fraction of regions with an excess of novel SNPs is the same around cluster breakpoints and non-cluster breakpoints.

Hypothesis 1b: The fraction of regions with an excess of novel SNPs is the same around cluster breakpoints and random regions.

Genome-wide, the novel SNPs were evenly distributed between cluster and non-cluster regions (Table 5.4, Figure 5.13). Cluster breakpoints versus random regions differed significantly in two of the nine lines. This is probably due to the elevated mutation rates observed near recombination events [43] and, given the very low absolute numbers of novel SNPs, does not support introgression. Therefore, the novel SNPs observed are most likely due to the longer read lengths in the resequenced data (100bp compared to 32bp and 51 bp used in [36]) making more of the genome accessible to variant calling.

In addition, I made a pileup [74] of all the reads in the 476 low-coverage genomes, and called SNPs *de novo*. $3,388,770$ variant sites were detected, 89.46% of which were present in the catalogue of variants [36], while 10.54% were novel. About a quarter of these novel SNPs (26.93%) were rare with Minor Allele Frequency (MAF) $< 3\%$ and were filtered out, as they probably correspond to individual sequencing errors. The distribution of novel SNPs across the genome is shown in Figure 5.8; the majority (82%) are inside or within 1Mb of a centromere.

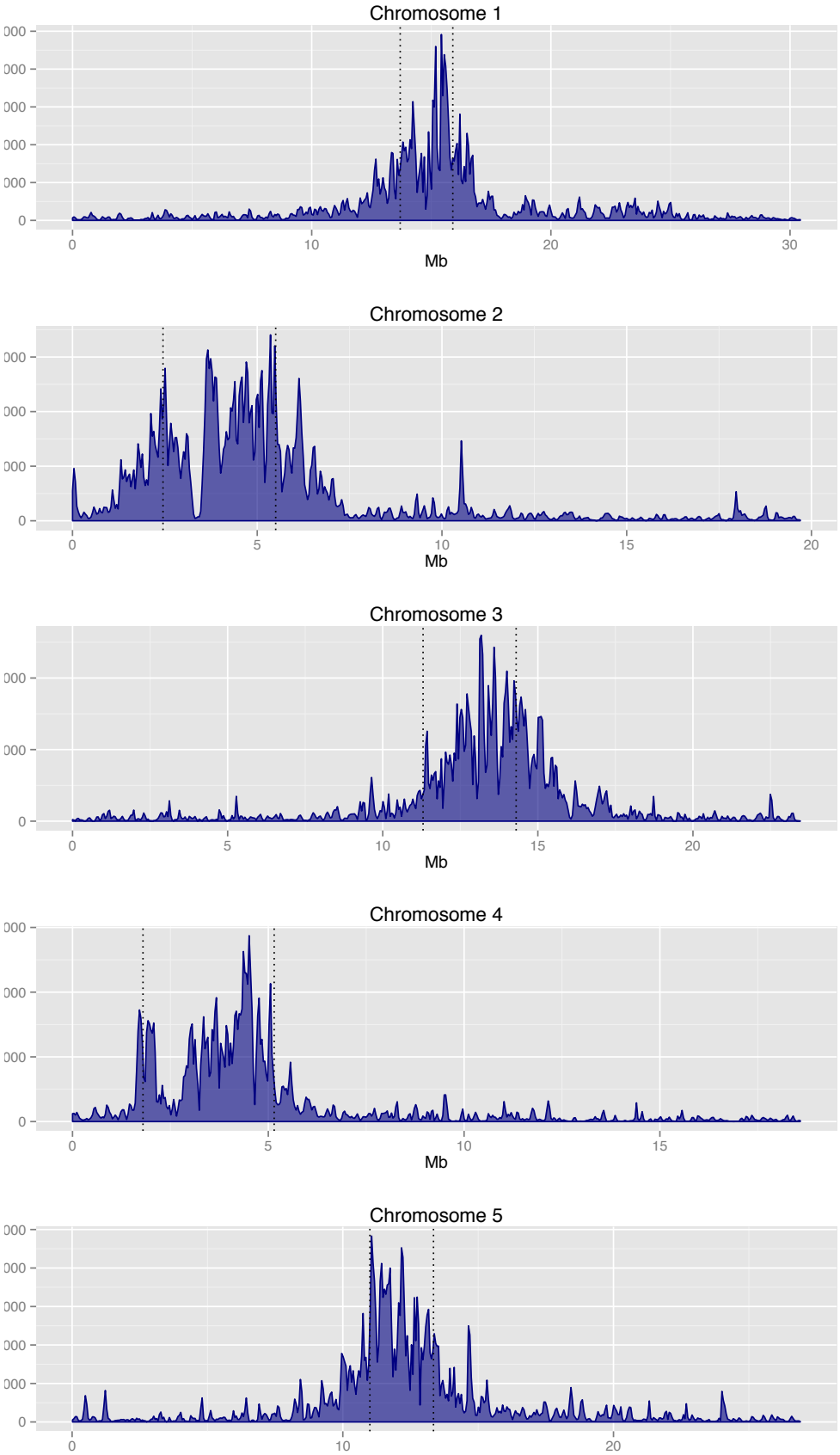


Figure 5.8: Spatial distribution of novel SNPs detected in a pileup of all reads obtained from low-coverage sequencing of 476 MAGIC lines (approximately 175x coverage). Dashed lines mark the centromeres.

5.6.2 Rearrangements because of structural variants

To search for structural variants, I used *de novo* assemblies [84] of the nine high-coverage genomes and mapped the resulting contigs to the reference. High-coverage genomes were reassembled using IMR/DENOM [36] (http://mus.well.ox.ac.uk/19genomes), and contigs were aligned to the reference using BLAT [59]. I identified contigs that mapped contiguously over cluster breakpoints. The contigs were too long to align to the reference using Stampy [83] that I used for short read alignment. BLAT and Stampy do not call variants consistently, because of different treatment of structural variants and repeats, so in order to make consistent variant calls, I extracted two 300bp sequences from each contig, flanking each breakpoint. Using Stampy and Samtools mpileup [74] I realigned the 300bp sequences to the reference using Stampy and called variants which I compared to the short-read data. A contig supports a breakpoint if it spans it and is long enough to both cover and confirm diagnostic SNPs for the two flanking haplotypes. Since only 8.32% of contigs were longer than the median distance of 950bp between diagnostic SNPs, one should expect to be able to validate a similarly small fraction of breakpoints. In fact, a higher fraction (169 out of 1132 (14.93%)) of cluster breakpoints in the nine lines were confirmed by a contig.

I also identified *de novo* contigs that did not map contiguously to the reference, but instead split across two multiple locations, providing evidence for rearrangements. I ignored contigs that mapped within transposons and repeat regions, and contigs that could be mapped to loci with distance lower to 50kb, as they were probably split because of indels. A breakpoint was deemed anomalous if a split *de novo* contig mapped to within 10kb of it.

High read coverage may also indicate presence of structural variants, such as duplications. Read coverage in cluster, non-cluster breakpoints and random controls was measured in 2kb windows. The read coverage of a region was considered normal if it was within $[C_\mu \pm 1.5C_{IQR}]$, with C_μ being the median coverage of the genome, and C_{IQR} the interquartile range.

Similarly, improperly-paired paired-end reads spanning breakpoints (i.e. aligned to the same strand, unpaired, or the incorrect distance apart) indicate potential structural variants. I computed the fraction of properly-paired reads in 1kb windows. A read was called properly-paired if it had a pair mapping to the same chromosome, and the insert size is less than $R_\mu + 3R_{IQR}$, with R_μ being

the median insert size of all reads in the genome (typically about 220bp). A region was considered normal if the fraction of properly-paired read pairs exceeded 90%. Strand error was defined as the fraction of read pairs mapping to the same strand in a 1kb window.

To assess presence of rearrangements near cluster breakpoints, I performed FET testing the following null hypotheses for each high coverage genome:

Hypothesis 2a: The fraction of regions with read coverage outside the range $[C_\mu \pm 1.5C_{IQR}]$ is the same in cluster breakpoints and non-cluster breakpoints.

Hypothesis 2b: The fraction of regions with read coverage outside the range $[C_\mu \pm 1.5C_{IQR}]$ is the same in cluster breakpoints and random controls.

Hypothesis 3a: The fraction of regions with less than 90% properly-paired read pairs is the same in cluster breakpoints and non-cluster breakpoints.

Hypothesis 3b: The fraction of regions with less than 90% properly-paired read pairs is the same in cluster breakpoints and random controls.

Hypothesis 4a: The fraction of regions with more than 10% read pairs mapped to the same strand is the same in cluster breakpoints and non-cluster breakpoints.

Hypothesis 4b: The fraction of regions with more than 10% read pairs mapped to the same strand is the same in cluster breakpoints and random controls.

Hypothesis 5a: The fraction of regions with a split *de novo* contig within 10kb is the same in cluster breakpoints and non-cluster breakpoints.

Hypothesis 5b: The fraction of regions with a split *de novo* contig within 10kb is the same in cluster breakpoints and random controls.

There is evidence for structural rearrangements in 206 (18.9%) of the cluster breakpoints, based on anomalies in the fraction of properly-paired paired-end reads, the read strand error, or the presence of split de novo contigs. This was not significantly different from non-cluster breakpoints but was significantly different from control regions (Tables 5.4, 5.5, Figure 5.13). From this analysis

I identified 377 potential structural variants that might explain a small number of clusters in the nine high-coverage genomes. Only 89 cluster breakpoints (7.86%) were anomalous with respect to high read coverage, a similar rate to that observed in non-cluster breakpoint regions and random loci (Tables 5.5, 5.4).

To check further if translocations in the founders might explain any of the clusters, I searched for segments that recurred in the low-coverage mosaics because they might be signatures of translocations. I found $2,573$ segments (12% of $21,326$) that were each repeated in at least 10 MAGIC lines with the same predicted founder, and whose start and end coordinates were the same to within 10kb. A translocation private to one of the 19 founders would be present in $\frac{476}{19} = 25$ lines on average. These represent 152 unique events and $1,731$ (67%) of them are in clusters.

This analysis of rearrangements preceded, and initially motivated the structural variation mapping in Chapter 4. In particular, to test whether these recurrent mosaic segments were related to translocations, I constructed binary traits for each of the 152 unique events, indicating whether each sample carried a segment. I mapped the traits using the genome scan presented in Section 4.5. All these traits (regardless of association with translocations) are expected to have a narrow QTL overlapping the tested segment, as all lines with a positive trait value have the same haplotype at that locus. A trait was classified as associated with translocations if the mapping rendered a region larger than 1Mb in which all loci inside it were associated with the trait at $\log P > 10$. In total, 39 traits had a cis (peak within 2Mb from the shared segment) and 14 had a trans QTL and so may correspond to SVs. These correspond to 491 (2.3%) mosaic segments An example of a trans QTL mapped in this way is shown in Figure 5.9.

All shared segments with a QTL also overlap with the sources of SV QTLs mapped with anomalous reads from Chapter 4. However, presence of translocations does not consistently generate false short mosaic segments, so mapping shared segments as traits can reveal a small number of SVs, but is not suitable for full mapping of SV QTLs. For example, from PCR validation we know that the translocation and inversion of Figure 4.8i is present in 12 founders, so the majority of the MAGIC lines should carry it. Nevertheless, only 23 carry an associated short segment whose haplotype is Ler-0 and whose genetic background at the sink is either Ler-0 or No-0 (which are

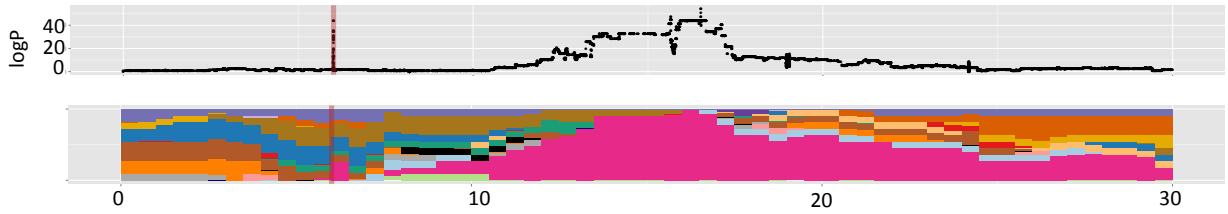


Figure 5.9: Genetic mapping of a recurrent mosaic segment, possibly caused by a structural variant. The segment is at chromosome 1, 6.009Mb and its founder haplotype is Can-0. 11 lines carry the segment, 8 of which in a cluster. The top panel of the figure shows results of genetic mapping of a binary trait, which was one in the 11 lines carrying the segment and zero in the rest. The x-axis shows genomic position, the y-axis genome-wide log P of each locus and the red vertical line marks the position of the segment. The bottom panel shows the (stacked) mosaic haplotypes for the 11 lines carrying the segment. A single haplotype is driving the QTL (Can-0, colour-coded in pink).

among the 12 founders carrying the rearrangement). Therefore mapping recurring short haplotype segments can only reveal a small subset of the SVs that segregate in MAGIC.

Returning to the analysis of recombination breakpoint clusters, this analysis showed that overall the fraction of breakpoints (cluster or non-cluster) that could be explained as recurrent translocations or SVs is small (2.3%). Thus, although rearrangements seem to affect the mosaics to some extent, they cannot explain the great majority of cluster breakpoints.

5.6.3 Validation by Sanger Sequencing

To further validate cluster breakpoints, I designed PCR primers and obtained capillary sequence data at 29 predicted cluster breakpoints, from two of the nine high-coverage genomes MAGIC.287, MAGIC.446. Regions for capillary sequencing were selected based on the following criteria:

1. the breakpoint interval (i.e. the interval between two diagnostic SNPs for each founder haplotype flanking the breakpoint) was smaller than 1kb, and therefore could be sequenced by a capillary read, and

2. a specific pair of primer sequences was available

The 26 out of 29 breakpoints selected for validation appeared in two clusters: the first cluster on the right arm of chromosome 3 of line MAGIC.287, covering the region from 18.97 - 22.81Mb and containing 36 breakpoints, 14 of which could be sequenced with respect to the above criteria,

and the second on the left arm of chromosome 3 of line MAGIC.446, spanning the region from 0.98 7.36Mb and containing 49 breakpoints, 12 of which could be sequenced. We also sequenced and genotyped 3 more cluster breakpoints, coming from different clusters in chromosomes 4 and 5 of MAGIC.446.

The selected regions and the primer sequences can be found in Appendix F, Table F.1. I searched for specific primer pairs flanking each cluster breakpoint using Primer3 [110], with parameters:

- Product size ranging between 500 and 1000bp
- Melting temperature between 55 and 65^{o}C
- Minimum primer size 20bp
- GC content between 40 and 60%
- Local self-complimentarity score (indicating the tendency of the primer to anneal to itself or form secondary structure) of 6 (PRIMER_MAX_SELF_ANY).
- Maximum self-complimentarity score of the 3 primer (a measure of its tendency to form a primer-dimer with itself) of 1 (PRIMER_MAX_SELF_ANY).

I then tested the primer specificity by aligning the primers obtained by Primer3 to the reference genome, using BLAT [59]. Only pairs with a single alignment were selected. PCR reactions and sequencing was performed by the Max Planck-Genome-Centre Cologne, Germany (http://mpgc.mpipz.mpg.de/home/) using myBudget Polymerase (BioBudget) and standard PCR conditions with 55^{o}C annealing temperature and 2.30 minutes extension time. All PCR products were sequenced from both ends and sequence data were automatically assembled by using DNAStar Seqman version 10.0.1 (LASERGENE), and manually edited.

The reads obtained from both ends were assembled into contigs. I then analysed the contigs as follows:

1. To test the presence of genomic rearrangements, I aligned the contigs to the reference genome and found that 26 spanned the corresponding breakpoint.

2. I genotyped the contigs at informative sites within them, in particular at sites where the predicted haplotypes flanking the breakpoints were different and found that in 25 of them the capillary sequence variants were identical to the ones predicted by the high-coverage short-read sequence data.

3. I tested the primer products for heterozygosity, by manually checking in the sequence trace files for overlaying signals.

Overlapping forward and reverse strand capillary reads were assembled into contigs $500-1000$bp long. In 27 out of 29 cases, the PCR products could be sequenced, and in 26 cases mapped to the correct locus in the reference, thereby indicating that structural rearrangement had not occurred at these breakpoints. Of these, all contained at least one sequence variant on each side of the breakpoint, diagnostic of the two haplotypes, and in 25 of them the capillary data are identical with the sequence data, thereby verifying the junctions and the flanking genotypes (e.g. in Appendix F, Table F.2 row 1 confirms a transition from founder Oy-0 to Wil-2). I found no evidence of heterozygosity, with the exception of some ambiguous nucleotides in 3 capillary reads. In two of them the ambiguities were within the 5 first bases of a read, therefore they were most likely sequencing errors, while in the other the ambiguities did not affect variant sites. Figure 5.1b shows the twelve breakpoints in a cluster in the right arm of chromosome 3 in MAGIC.287 and Figure 5.1c shows eleven in a cluster in the left arm of chromosome 3 in MAGIC.446.

Overall, PCR and capillary sequencing confirmed 12 out of 14 breakpoints from the first cluster and 11 out of 12 breakpoints from the second. Taken together with the 169 cluster breakpoints confirmed by de-novo assembly (see Section 5.6.2), this analysis suggests the majority of predicted cluster breakpoints are both contiguous and have the expected flanking haplotypes.

5.6.4 Heterozygosity

Residual heterozygosity in the MAGIC lines

To test for residual heterozygosity in the MAGIC lines, I made a read pileup for each high-coverage genome at known SNP sites using Samtools [74], masking out sites not having exactly two alleles,

Line	**Heterozygosity**
MAGIC.105	1.10%
MAGIC.149	13.17%
MAGIC.175	1.50%
MAGIC.287	1.57%
MAGIC.329	13.86%
MAGIC.338	1.39%
MAGIC.426	11.44%
MAGIC.433	15.34%
MAGIC.446	0.75%

Table 5.3: Heterozygosity levels, defined as the fraction of known SNPs in each genome that were called heterozygous, for the 9 high-coverage genomes.

or within transposons or known repeats. A site was considered heterozygous if it was covered by n reads, of which r were for the minor allele, and the probability that the number of minor alleles was r or fewer exceeded 0.1, under the assumption that r followed a binomial distribution $B(n, 0.5)$. This definition avoids calling false heterozygotes where the read coverage is low. Heterozygosity within a region was defined as the fraction of known variant sites that were called as heterozygotes.

Heterozygosity was low throughout the genome (1.27% heterozygous sites on average) in five of the nine lines sequenced at high coverage, and therefore could be excluded as a cause of the clusters in these lines. 13.70% of loci in the remaining four lines were heterozygous (Table 5.3). Figures 5.10, 5.11 and 5.12 illustrate heterozygosity levels and cluster locations across 3 of the 9 high-coverage genomes.

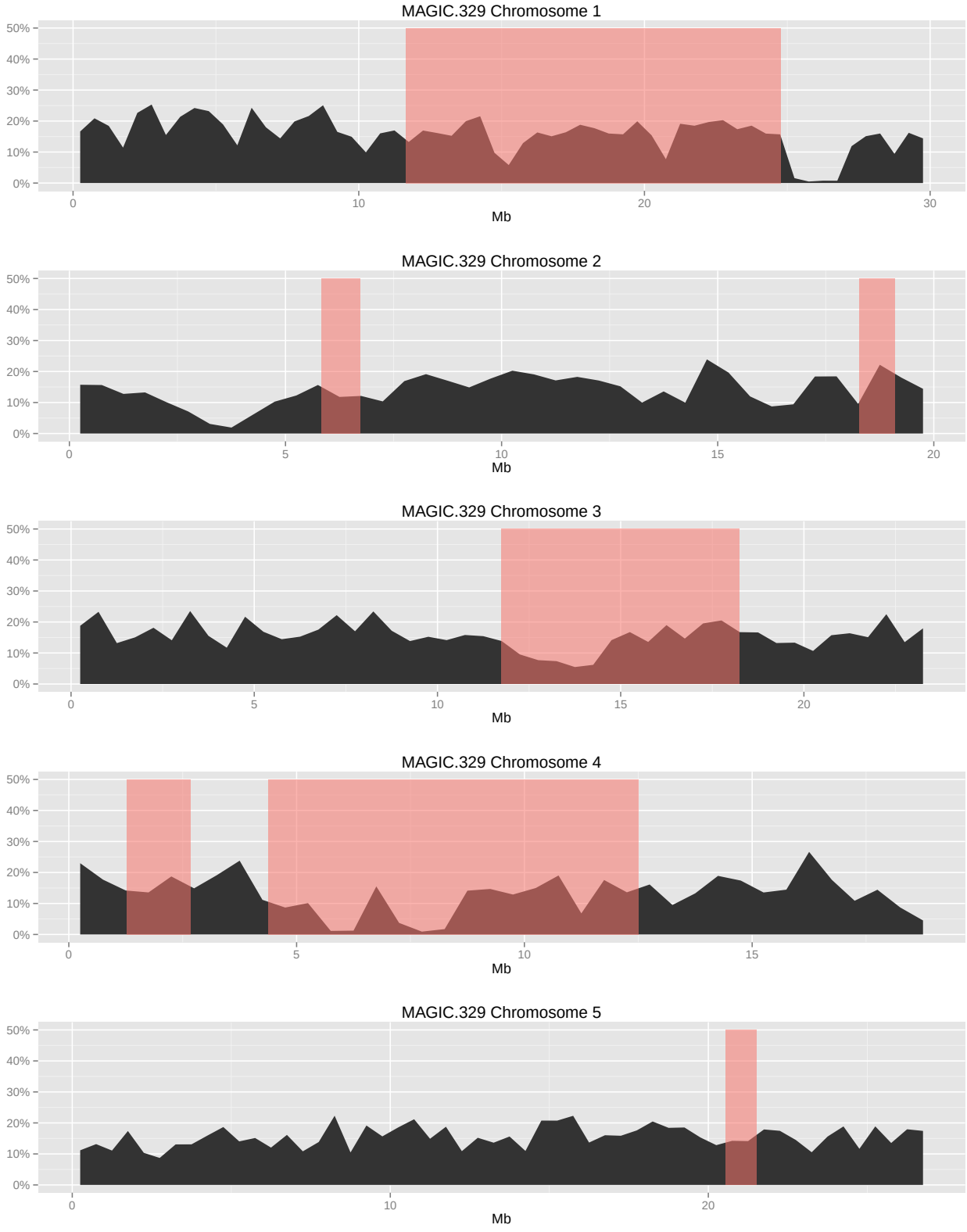


Figure 5.10: Distribution of heterozygosity and cluster location across MAGIC.329 sequenced at 12x coverage. The x-axis is the genomic position in Mb, and the y-axis the percentage of heterozygous sites computed in bins of 100kb along each chromosome. The pink rectangles mark the regions in which clusters are observed.

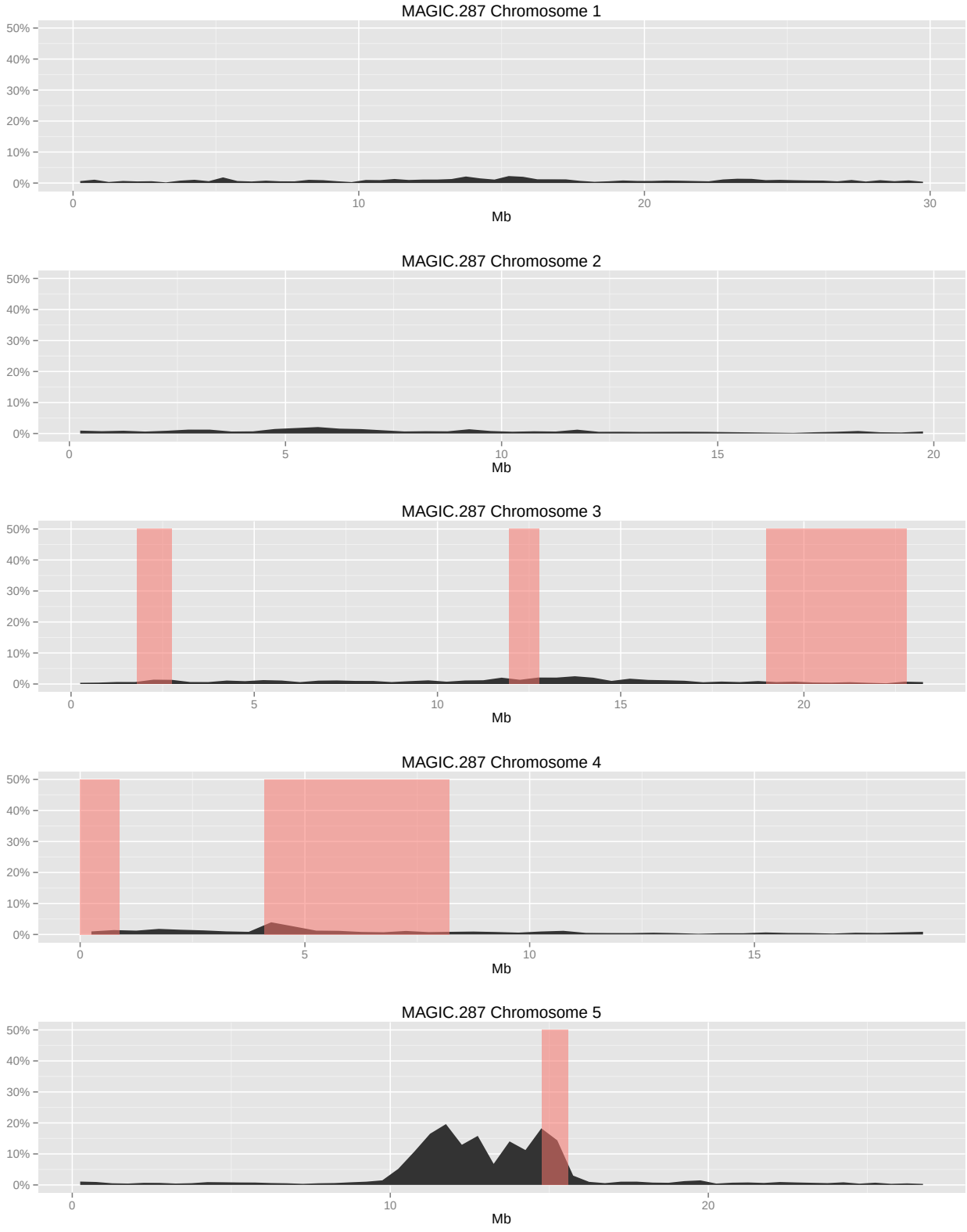


Figure 5.11: Distribution of heterozygosity and cluster location across MAGIC.287 sequenced at 12x coverage. The figure can be interpreted as Figure 5.10.

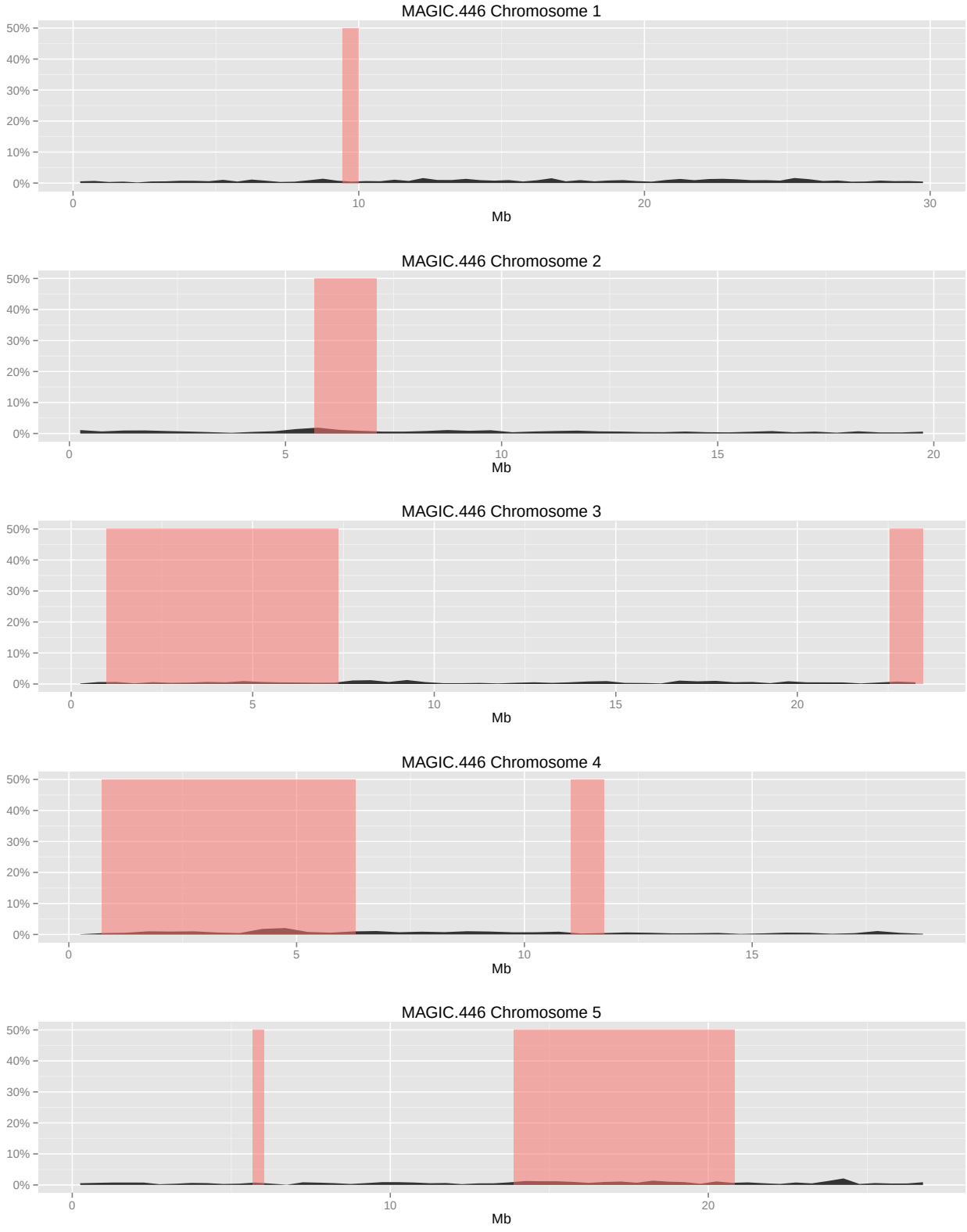


Figure 5.12: Distribution of heterozygosity and cluster location across MAGIC.446 sequenced at 12x coverage. The figure can be interpreted as Figure 5.10.

To test if there are significant differences in heterozygosity levels near cluster breakpoints in comparison to non-cluster breakpoints and random regions, I computed the fraction of heterozygous SNPs detected in 10kb windows around cluster, non-cluster breakpoints and random regions. The distributions across the three classes of regions can be seen in Figure 5.13a. For each genome, we performed two FETs with null hypotheses:

Hypothesis 6a: The fraction of regions with over 10% heterozygous sites in a 10kb window around cluster breakpoints is equal or lower than that of non-cluster breakpoint regions.

Hypothesis 6b: The fraction of regions with over 10% heterozygous sites in a 10kb window around cluster breakpoints is equal or lower than that of random regions.

The p-values of the FETs for each line are shown in Tables 5.4 and 5.5. In eight out of nine lines heterozygosity around cluster breakpoints was not significantly different from non cluster breakpoints or random loci. Overall, heterozygosity at only 7.6% of cluster breakpoints differed significantly from random control or non-cluster breakpoint regions. Furthermore, capillary sequence data collected in two lines around 29 breakpoints (described in Section 5.6.3), showed no heterozygosity in variant sites. However, 4 out of 9 lines have unexpectedly high levels of heterozygosit. The clusters in these lines mostly involve two oscillating founder states (similar to the clusters in Figure 5.5b). Indeed, these 4 lines were among the 8 highly heterozygous lines reported in Section 3.6.2, which were reconstructed as largely heterozygous by the diploid reconstruction algorithm. Over heterozygous regions the haploid algorithm encounters alleles corresponding to two different haplotypes, and reconstructs a mosaic with frequent oscillations between the two haplotype states. Reconstructing the low-coverage mosaics using the diploid algorithm led to the disappearance of 284 (52.9%) clusters, which were reconstructed as heterozygous states. We conclude that there is evidence of residual heterozygosity in some of MAGIC lines and that it is the cause of half of the clusters.

	FET p-value in cluster vs non-cluster breakpoints					
MAGIC	**N** (H 1a)	**C** (H 2a)	**PP** (H 3a)	**S** (H 4a)	**SC** (H 5a)	**H** (H 6a)
MAGIC.105 244 | 38	0.148(42)	0.634(14)	0.32(8)	0.703(23)	0.242(10)	0.32(8)
MAGIC.149 91 | 63	0.165(8)	0.644(7)	0.221(9)	0.39(10)	0.668(8)	0.36(67)
MAGIC.175 49 | 45	1(0)	0.15(3)	0.192(6)	0.195(9)	0.907(2)	**0.057**(**9**)
MAGIC.287 73 | 47	0.939(1)	1(0)	0.448(5)	0.91(8)	0.75(5)	0.44(5)
MAGIC.329 121 | 62	0.465(8)	0.55(28)	**0.04**(**8**)	**0.052**(**18**)	**0.011**(**16**)	0.84(72)
MAGIC.338 104 | 28	0.658(11)	0.412(8)	0.93(3)	0.47(11)	0.199(7)	0.93(3)
MAGIC.426 103 | 53	0.234(9)	0.885(1)	0.372(7)	0.109(12)	0.561(5)	0.19(77)
MAGIC.433 168 | 52	0.152(17)	0.265(5)	0.747(16)	**0.041**(**30**)	0.518(12)	0.77(108)
MAGIC.446 179 | 32	0.814(12)	1(0)	0.544(9)	0.905(16)	0.445(5)	1(0)

Table 5.4: P-values of one-sided Fisher exact tests (FET) comparing cluster breakpoints with non-cluster breakpoints for six quality metrics. P-values < 0.1 are in bold. In parentheses, the number of abnormal cluster breakpoints for each test. Below each MAGIC id the total number of cluster breakpoints and non-cluster breakpoints for each line is shown. N - number of novel SNPs (Hypothesis 1a), C - median coverage (Hypothesis 2a), PP - % of properly paired reads (Hypothesis 3a), S - % of strand errors (Hypothesis 4a), SC - split contigs (Hypothesis 5a), H - heterozygosity (Hypothesis 6a).

	FET p-value in cluster vs random					
MAGIC	**N** (H 1b)	**C** (H 2b)	**PP** (H 3b)	**S** (H 4b)	**SC** (H 5b)	**H** (H 6b)
MAGIC.105 244 | 100	**0.009**(**42**)	0.121(14)	0.599(8)	**0.004**(**23**)	0.548(10)	0.59(8)
MAGIC.149 91 | 100	0.248(8)	0.632(7)	**0.027**(**9**)	**0.074**(**10**)	**0.059**(**8**)	0.11(67)
MAGIC.175 49 | 100	1(0)	0.809(3)	**0.002**(**6**)	**0.005**(**9**)	0.133(2)	**0.12**(**9**)
MAGIC.287 73 | 100	0.888(1)	1(0)	0.433(5)	0.21(8)	**0.068**(**5**)	0.43(5)
MAGIC.329 121 | 62	**0.041**(**8**)	0.989(28)	0.648(8)	**0.074**(**18**)	0.256(16)	0.18(72)
MAGIC.338 104 | 100	0.133(11)	0.993(8)	0.917(3)	0.133(11)	0.584(7)	0.86(3)
MAGIC.426 103 | 100	0.157(9)	0.758(1)	0.105(7)	**0.052**(**12**)	0.412(5)	**0.0006**(**77**)
MAGIC.433 168 | 100	0.126(17)	0.978(5)	0.242(16)	**0.004**(**30**)	0.254(12)	0.405(102)
MAGIC.446 179 | 100	0.471(12)	1(0)	0.192(9)	0.192(16)	0.871(5)	1(0)

Table 5.5: P-values of one-sided Fishers exact tests comparing cluster breakpoint regions with random regions for all 6 quality control checks (see Table 5.4 for description). The total number of cluster breakpoints and the number of random regions considered for each line are shown below the MAGIC ids.

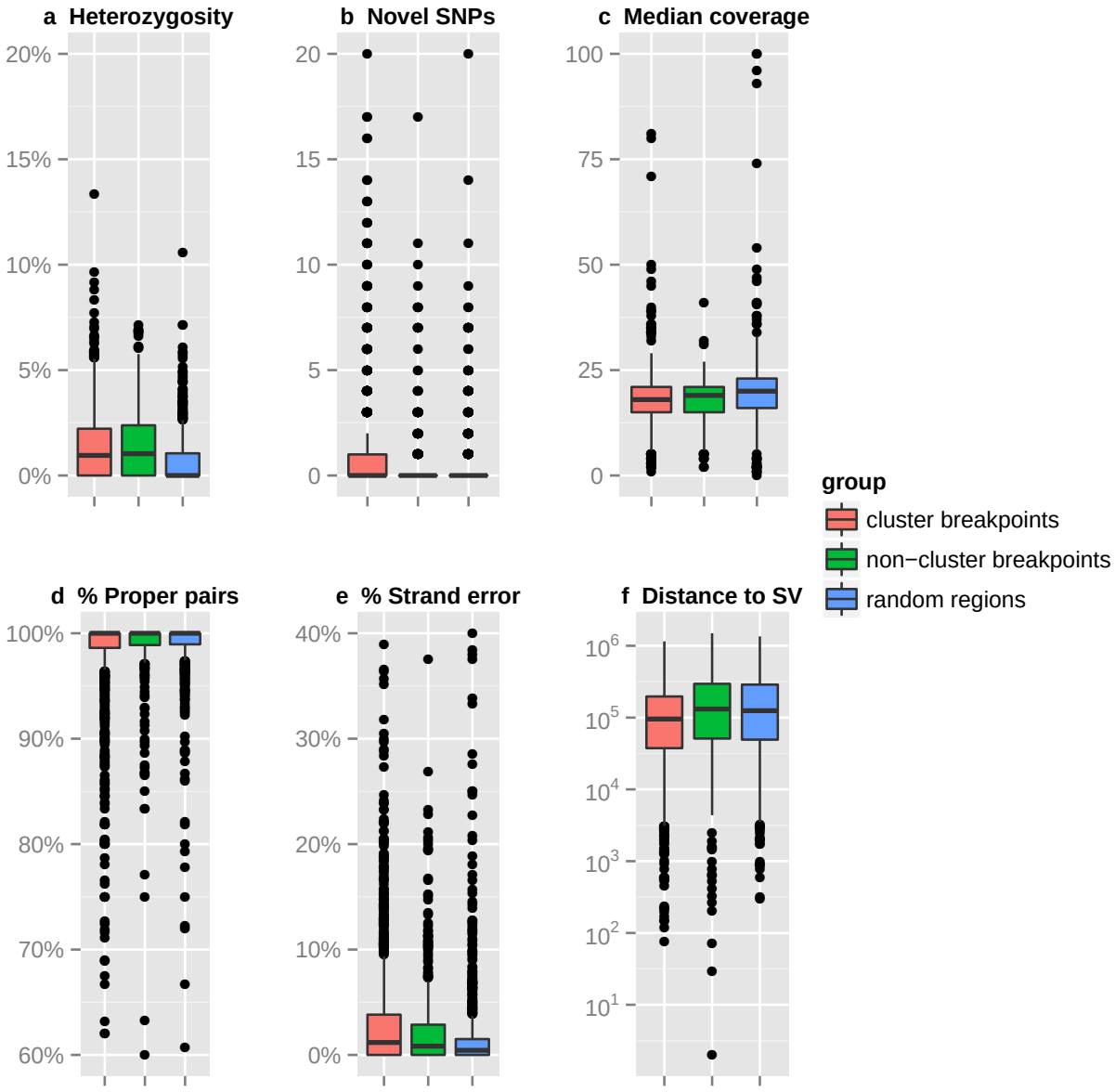


Figure 5.13: Characteristics of 1132 cluster breakpoints (salmon), 420 classical non-cluster breakpoints (green), and 900 randomly chosen control regions (blue) with respect to quality control metrics, in nine MAGIC genomes sequenced at high coverage (15x). Shown are box and whisker plots for (a) heterozygosity in a 10 kb window around each breakpoint, (b) number of novel SNPs in a 10kb window, (c) median read coverage in a 2kb window, (d) percentage of properly paired reads (correct insert size and both reads present) either flanking or within 1kb of breakpoint, (e) percentage of read pairs mapping in error to same strand instead of opposite strands, (f) distance of breakpoint to nearest candidate structural variant, based on *de novo* contigs mapping to split regions in the reference.

Heterozygosity in the 19 founders

Undetected heterozygosity in the 19 founders could be another source of the clusters, as that would introduce extra unknown haplotypes. Reconstructing an unknown haplotype in terms of the 19 would have the appearance of a dense cluster, with breakpoints corresponding ancient recombinants. The introgression analysis in 5.6.1 would have covered this possibility only if there were untyped variants in the original 19 founders. However, it could also be possible that the set of SNPs in the 19 founders was complete, yet heterozygous alleles were underestimated.

To test this, I re-called SNPs at the same sites using GATK [89]. Heterozygosity levels in the 19 founders estimated by GATK were very different than those estimated by IMR/DENOM, as explained in Section 3.2. I reconstructed the genome mosaics again using the new set of alleles. In total, 166 (30.9%) of clusters were not in the revised mosaics: many of the clusters are replaced by a Po-0 haplotype (Figure 5.14), indicating that some clusters were signatures to the of the second Po-0 haplotype. This partially explains the shared haplotype segments across multiple lines, although because of sequencing at low coverage the clusters did not overall look identical. This also explains similar clusters shared between multiple (over 3) lines, described in Section 5.5.

5.6.5 Revised clusters

Excluding clusters over regions with residual heterozygosity (identified by the diploid reconstruction algorithm), in total there are 86 clusters, affecting 60 lines. The mean size of these clusters is

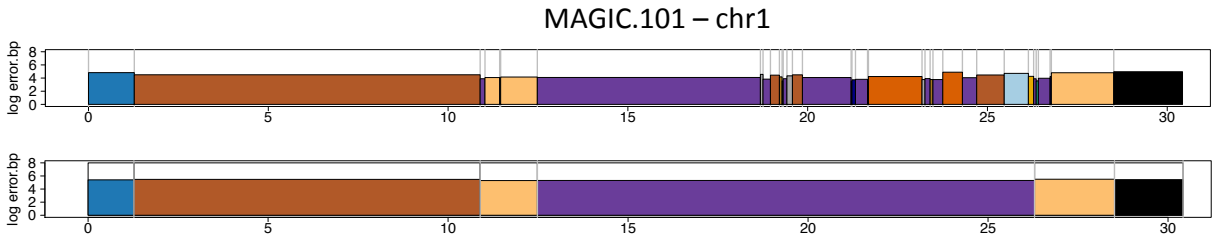


Figure 5.14: Mosaics of chromosome 1 from MAGIC.101 with the original (IMR/DENOM) set of sequence variants (top panel) and with the revised (GATK) set. The purple haplotype which has replaced the clusters in the bottom genome is Po-0. Other dense breakpoints (such as the one at $\sim$ 11Mb) are also absent from the revised mosaic.

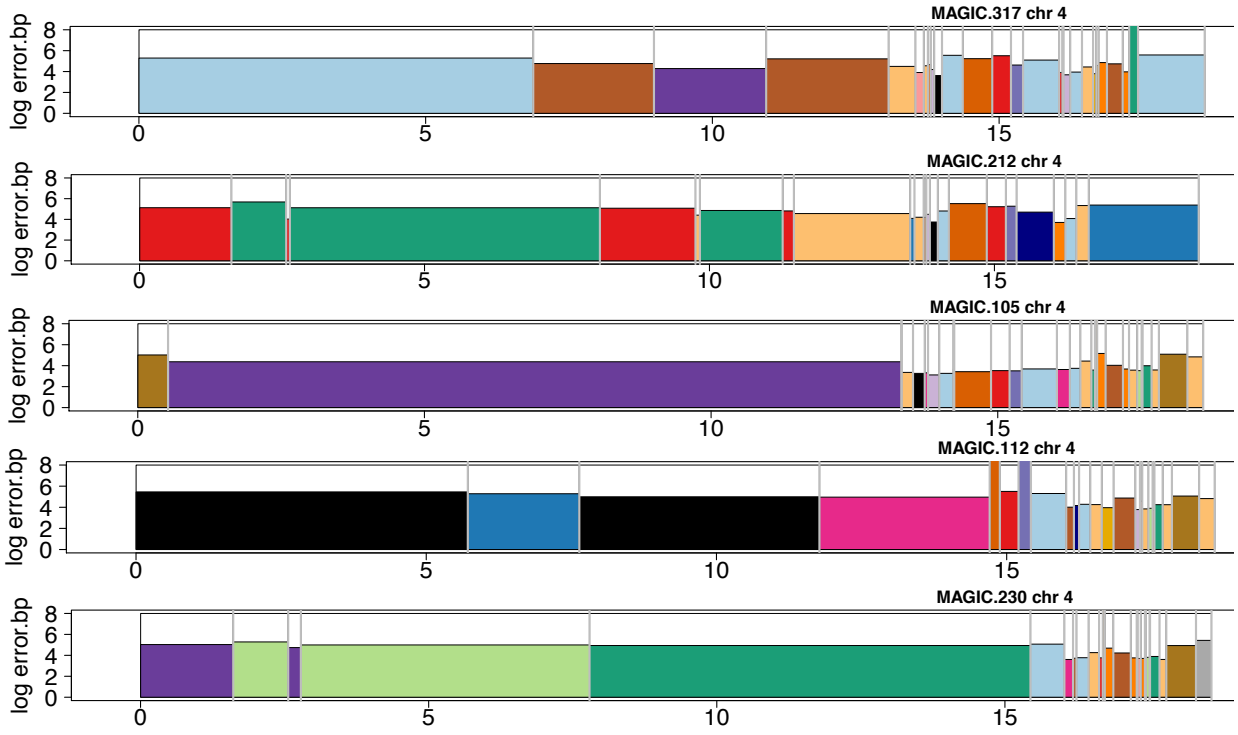


Figure 5.15: Clusters present in the GATK mosaics with shared haplotype signatures. Part of the haplotypes comprising the clusters are shared between lines. 12 lines in total have a similar cluster in the same region, only 5 are shown here.

1.03Mb, each cluster comprising 7 breakpoints on average. However, 49 (56.9%) of these remaining clusters have haplotype signatures that are similar to more than 5 lines. These clusters are concentrated around three genomic regions, namely Chr1 : $4.2 - 10.1$Mb, Chr4 : $13.8 - 18.9$Mb and Chr5 : $17.6 - 23.2$Mb. An example of such clusters is illustrated in Figure 5.15. Such identical signatures are unlikely to be caused by recombination and we have also ruled out introgression or presence of additional haplotypes in the 19 founders. Extensive rearrangements could be a potential cause, although the size of these events is too large to explain solely by rearrangement. Excluding these shared clusters, the mosaics are left with 37 clusters which are unique, found in 36 lines. Their average size is 0.8Mb and they may be recombination-related in the sense that they have passed all validation tests.

5.7 Discussion

This chapter presented the analysis and validation of the MAGIC genomic mosaics. Our main goal was to use the mosaics in order to analyse recombination in the population, however obtaining accurate mosaics would be essential for genomic analysis using low-coverage populations. In a very large part of the population the mosaics greatly differed to what was originally expected, as 280 (58.8%) genomes appeared having a much higher number of recombination breakpoints than expected given the breeding generations, which also tended to form clusters in specific positions. Although unexpected, clusters could be consistent with evidence from other studies on recombination in *A. thaliana*, so we did not immediately dismiss them as artefacts. Firstly, the existence of some clusters in recombinant inbred lines is to be expected, simply because visible recombinants can only accumulate in the few remaining regions of heterozygosity [13]. Secondly, non-interfering crossovers in a single meiosis have been reported in male meiosis in *A. thaliana* [8] and reduction in crossover interference would be expected in order to generate a cluster at a single event. Furthermore, gene conversion has been suggested as a major cause of recombination in *A. thaliana*, with multiple gene conversions concentrated around the centromeres [138].

However, after detailed data analysis, in which we considered a large catalogue of sources of artefactuality, we discovered that the majority of clusters can be explained by residual heterozygosity in some of the MAGIC lines and by undetected heterozygosity in the 19 founders. A smaller fraction was caused by structural variants. There is a small number of clusters left in the final mosaics which may be recombination-related and would be interesting to investigate further. Although these clusters have survived extensive attempts to eliminate them they may still be artifactual, but they could potentially be caused by recombination or by some other interesting phenomenon. An approach to investigate this in the future would be to sequence these clusters with very high coverage and use *de novo* assembly to form large contigs that accurately capture the region. An alternative and, perhaps, superior approach to this would be to use novel next-generation sequencing technologies that produce very long reads, such as Oxford Nanopore MINION [54] or PAC-Bio sequencing [19]. The main questions to ask then would be whether the genotype of these regions confirms the mosaic predictions and whether there are extensive structural variants (possibly *de*

novo) over the regions.

Finally, despite the fact that the reconstruction algorithm was able to accurately reconstruct simulated genetic mosaics with virtually no error, even in the presence of 1% purposefully-inserted allele errors (Chapter 3), real data were considerably more complex, In this particular case, the quality of the mosaics was affected by a combination of complex genetic features (structural variants and rearrangements) and technical errors (assumptions about the expected levels of heterozygosity, lines that had selfed for fewer generations than expected). Eventually, after careful data analysis we were able to clear the data of errors and produce reliable mosaics. The next chapter gives a detailed account on the final mosaics and describes the recombination hotspot analysis of the MAGIC lines.

Chapter 6

Recombination hotspot analysis of the MAGIC lines

This chapter presents the MAGIC mosaics after revising the SNP calls in the founder accessions, as described in Chapters 3 and 5. I compute recombination rates in each of the five chromosomes and discuss identification of hotspots. I also compare the hotspots estimated in MAGIC with the *A. thaliana* genetic map, estimated by two independent studies using linkage disequilibrium in accessions [50, 23].

6.1 Recombination hotspots

As we saw in Section 3.6, each MAGIC line had on average 29 recombination breakpoints and there were $14,260$ recombination breakpoints. Assuming that each mosaic breakpoint represents a recombination event, I computed the average recombination rate ρ of each chromosome, defined as number of breakpoints per kilobase (Table 6.1).

To identify recombination hotspots, I divided the genome into 1kb bins and counted the total number of breakpoints within each one. I then amagalmated bins containing breakpoints using a recursion based on the Smith-Waterman (SW) algorithm, similar to the one used to define cluster breakpoints in Chapter 5, Algorithm 3. This approach maximises the number of breakpoints that are in close proximity and also determines hotspot size dynamically. The algorithm scored each

Chr	**Size (bp)**	**Total recombinants**	ρ **(breakpoints/kb)**
1	$30,427,671$	$3,508$	0.1153
2	$19,698,289$	$2,009$	0.1019
3	$23,459,830$	$2,330$	0.0993
4	$18,585,056$	$2,880$	0.1549
5	$26,975,502$	$3,533$	0.1309
Genome	$119,146,348$	$14,260$	0.1196

Table 6.1: Numbers of recombination breakpoints and estimated recombination rates ρ per chromosome.

bin by the number of breakpoints it contained. Bins with no breakpoints were given a negative score. Each iterative application of the recursion returned the maximum scoring island of bins, i.e. a potential hotspot. After each iteration the maximum-scoring island was given a negative score, until all positive-scoring islands were detected. The algorithm and the parameters used are summarised in Algorithm 4.

From this algorithm I obtained regions with a positive number of recombination breakpoints. To determine which were statistically significant given their size and the number of breakpoints they contained I used the binomial distribution Bin to compute the probability of observing an equal or higher recombination rate within a region of the same size given the recombination rate of the corresponding chromosome. Thus, for a segment i if l_i the length of the segment, a_i the total number of recombination breakpoints in it and ρ the expected chromosomal recombination rate (recombinants/kb) the p-value of the segment was given by $P_{\mathsf{Bin}}(n \geq a_i)$ for the binomial distribution with parameters $\mathsf{Bin}(l, \rho)$, where l the length of the segment in kb. I adjusted the p-values using the Benjamini-Hochberg adjustment [10], thus by ranking all the p-values $P_{(1)}, P_{(2)}, \ldots P_{(n)}$ and computed the q-values: $Q_{(i)} = P_{(i)} \frac{n}{i}$. I selected as recombination hotspots all regions that had q-values at FDR≤ 0.01. In total, I identified 448 hotspots, of average width 4.40kb and with average recombination rate 2.16 recombinants/kb. In total $3,476$ breakpoints (24.38%) were within these hotspots. Figure 6.1 displays the genome-wide distribution of recombinants in MAGIC and the recombination hotspots and Figure 6.2 shows the distribution of recombination hotspot lengths.

Data: N - total number of bins in the genome

a_i - number of recombination breakpoints in i-th bin

Initialisation:

$$X(i) = \begin{cases} a_i & \text{if } a_i > 0 \\ -1 & \text{otherwise} \end{cases}$$

$S(0) = 0$

$B(0) = 0$

Recursion:

$$S(i) = \begin{cases} S(i-1) + X(i), & \text{if } S(i) + X(i) > 0 \\ 0, & \text{otherwise} \end{cases}$$

$$B(i) = \begin{cases} B(i-1), \text{if } S(i) > 0 \\ i, \text{otherwise} \end{cases}$$

Algorithm 4: The Smith-Waterman type recursion used to dynamically determine hotspot size.

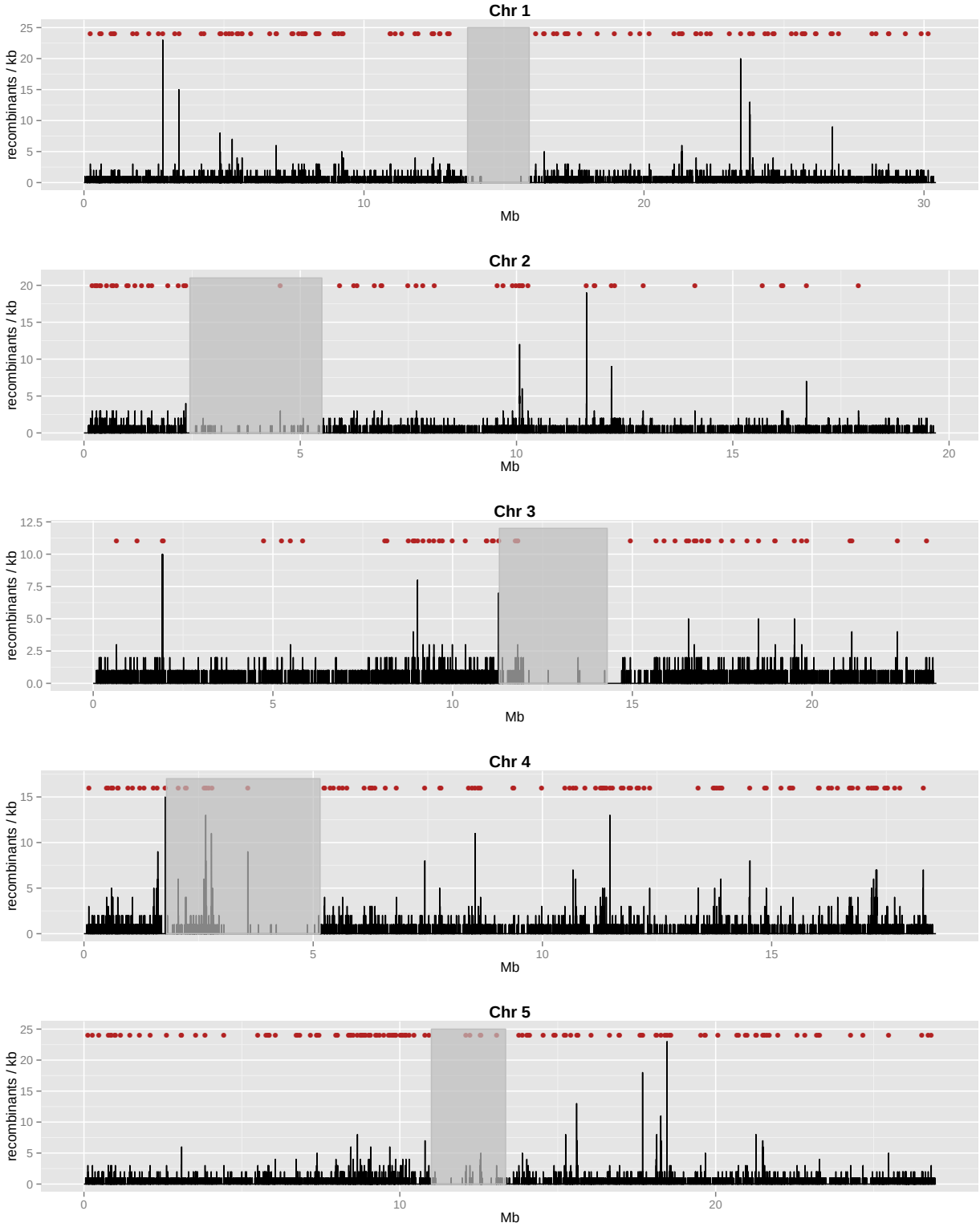


Figure 6.1: Distribution of genome-wide recombination rates and hotspot positions. The x-axis shows genomic position in Mb and the y-axis recombination rates (ρ). Bars are 1kb genomic region and their heights show the total number of recombinants they contain. Red dots show hotspot positions. Dark grey areas mark centromeres.

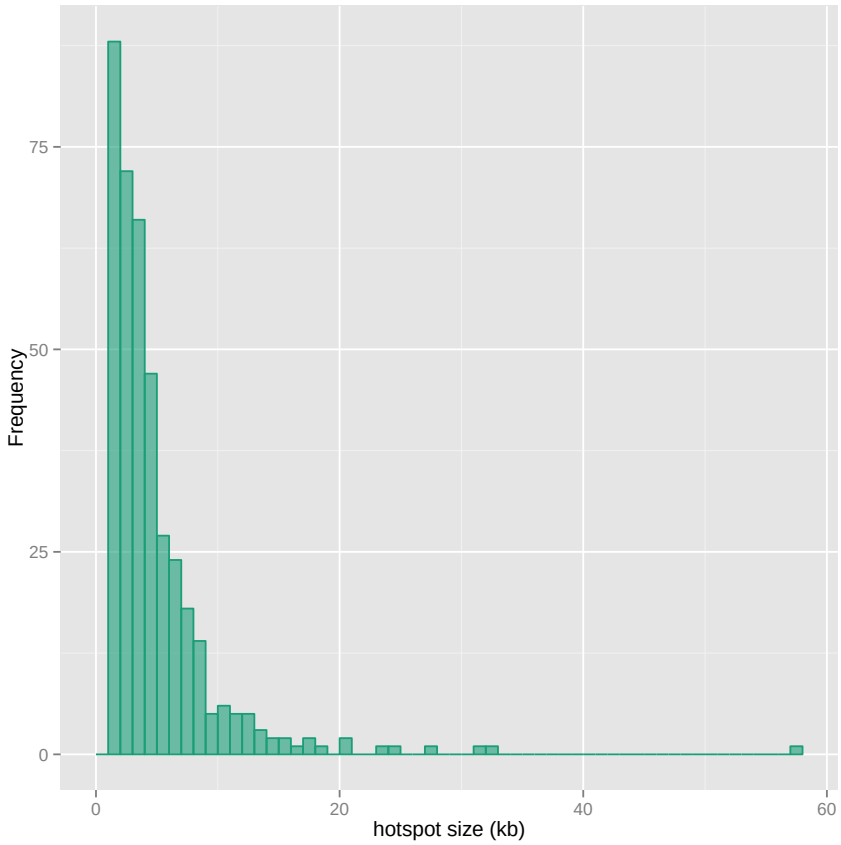


Figure 6.2: Distribution of hotspot sizes in kb.

6.2 Comparison of hotspots with the *A. thaliana* genetic map

Recombination hotspots were compared to two independent coalescent-based studies of recombination for natural accessions of *A. thaliana* from ancestral recombination data, which both estimated hotspot locations [50, 23]. The genetic map from [50] was based on 1,307 worldwide accessions, including samples with highly differentiated *A. thaliana* genomic regions, that were genotyped at 250k sites. Recombination rates in that study were estimated by computing linkage disequilibrium (LD) within regions of 2.5k SNPs and 2,809 hotspots were reported with $\rho > 3$ recombinants/kb. In [23], the genetic map was inferred from 80 Eurasian accessions genotyped at over 2M SNPs, again using genome-wide linkage and 8,448 hotspots were identified.

To test the overlap of hotspots identified in MAGIC with the published hotspots, I computed

the recombination rates estimated by each of the two studies on the midpoints of the MAGIC hotspots and whether these midpoints were within reported hotspots. I assessed overlap using the following three tests:

1. Binomial tests with MAGIC hotspots as targets and random regions of the same size as controls. The null hypothesis was that the number of MAGIC hotspots that were also in the list of hotspots of each of the two independent studies is the same as the number that would have occurred by chance. I estimated the expected overlap with random locations using 100 random samples without replacement, each of the same length as the MAGIC hotspots. I found that 277 (61.8%) MAGIC hotspots overlap with those from [50] (Binomial Test (BT) p-value $< 10^{-15}$) and 427 (87.5%) overlap with those from 29 (BT p-value $< 10^{-17}$). A Venn diagram showing overlap across the three studies is in Figure 6.3

2. Students t-tests of the null hypothesis that the mean recombination rates in the two studies are the same in MAGIC hotspots and random regions. I found that recombination rates estimated by [50, 23] at the 448 MAGIC hotspots are significantly elevated compared to the rest of the genome (T-test p-values $< 10^{-5}$, 10^{-10}, respectively). Therefore, recombination rate estimated on the MAGIC hotspots by the independent studies is elevated compared to random.

3. Correlation of the hotspot recombination rates estimated in MAGIC and the recombination rates of the same hotspots estimated by the two independent studies. I report no correlation; the correlation coefficient in the comparison with [50] was -0.07 (one-sided p-value < 0.93) and with [23] was -0.06 (one-sided p-value < 0.89).

Therefore, the MAGIC hotspots are reasonably consistent with the *A. thaliana* genetic map, as they significantly overlap with the two independent studies but there is variation in hotspot recombination rates.

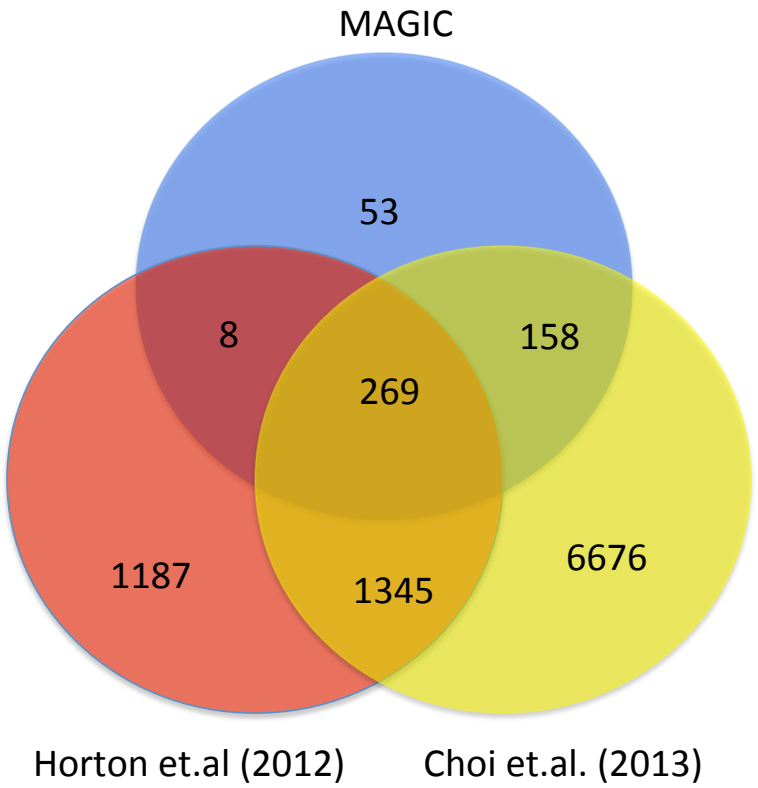


Figure 6.3: Venn diagram showing overlap of MAGIC hotspots with hotspots estimated by 2 independent LD-based studies [50, 23]

6.3 Conclusion

The MAGIC genome mosaics can be used for recombination hotspot inference. Recombination events were predicted from mosaic breakpoints, mapped with high precision to within 2.5kb, and were then used to compute chromosomal recombination rates and identify recombination hotspots. In total I detected 448 hotspots and compared them to the *A. thaliana* genetic map, described by two independent LD-based studies [50, 23].

Overall, the MAGIC hotspots are consistent with the known recombination landscape of *A. thaliana*. There was high overlap of hotspots estimated in MAGIC with the genetic map as 89.1% of MAGIC hotspots overlap with at least one study. Therefore, the genome mosaics can be used as a reliable alternative to the genetic map.

I observed variation in the recombination rates of hotspots between studies, as recombination

rates estimated on the MAGIC hotspots by the independent studies did not correlate with the rate estimated from the mosaics. In contrast, recombination rates on overlapping hotspots were highly correlated between the two LD-based studies ($r = 0.29$, $P < 10^{-10}$). This lack of correlation of recombination rates in MAGIC might be related to the different study design. Decay in linkage disequilibrium observed in naturally-occurring populations sometimes corresponds to ancient recombinants shared by a large fraction of genomes, so some hotspots may be inactive at present or have reduced activity. In contrast, recombinants inferred in MAGIC are recent and so MAGIC hotspots probably indicate presently active loci.

No functional sequencing motif has yet been proposed for recombination in *A. thaliana*, the most highly associated motif reported being a poly-A [50]. Discovering a functional motif would help understand the mechanism of recombination in *A. thaliana*, and since MAGIC hotspots are active, they comprise a suitable dataset. Using the MAGIC hotspots to search for enriched sequence motifs would therefore be a useful next step.

Chapter 7

Genetic mapping of loci associated with endosymbionts using unmapped host sequencing reads

Bacterial and viral endosymbionts live inside and on the surface of the cells of other organisms. Many are essential to the survival of the host organism, such as rhizobia, which are nitrogen-fixing bacteria involved in the formation of root systems [29], while others are pathogenic. Response to endosymbionts differs between ecotypes [125] and probably has a genetic component. I noticed that unmapped reads from the MAGIC sequence data sometimes align to the genomes of endosymbionts and might correspond to organisms hosted in the genomes of the sequenced plants. This chapter explains this observation.

Using unmapped reads I quantified presence of specific endosymbionts in the genomes of the MAGIC lines using number of reads mapped to the 16S rRNA of each organism. I compared these read mapping levels in 9 MAGIC lines that were sequenced at both high and low coverage (from Chapter 5) and observed high correlation between high and low-coverage counts. Consequently, low-coverage data can be used to study variation in endosymbiont levels. I scanned the genome for endosymbiont levels for 24 organisms and found marginally significant genome-wide associations. Interestingly, different endosymbionts, most of them related to response to salt-stress, shared the

same QTL, which was close to a gene regulating response to saline toxicity. Repeating the mapping with a larger population size or use of a multivariate model that would combine several endosymbiont levels would be required for conclusive results.

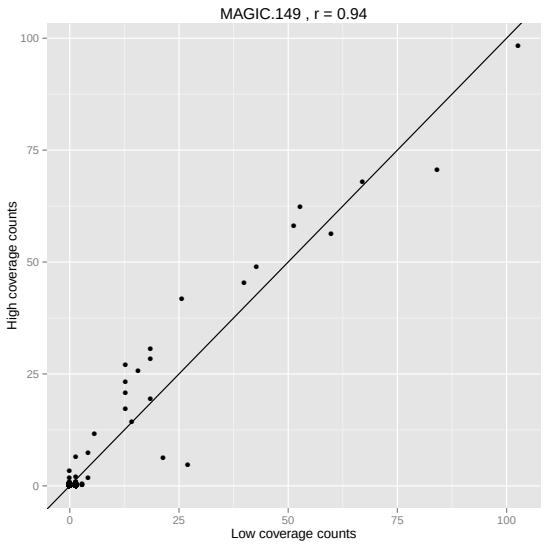
7.1 Detection of endosymbionts using unmapped reads

DNA for sequencing the MAGIC lines was extracted from leaf tissue. Part of the DNA in the leaves does not belong to the sequenced plant, but rather to endosymbiont organisms that are present on or within the cells of the leaf. Endosymbiont DNA is sometimes accidentally sequenced and is reflected by those reads that do not map to the reference genome. 16S rRNA sequences are commonly used for distinguishing between prokaryotic species because they are found universally yet are usually unique to each species. A 16S rRNA is an RNA gene that is a component of the 30S small subunit of prokaryotic ribosomes.

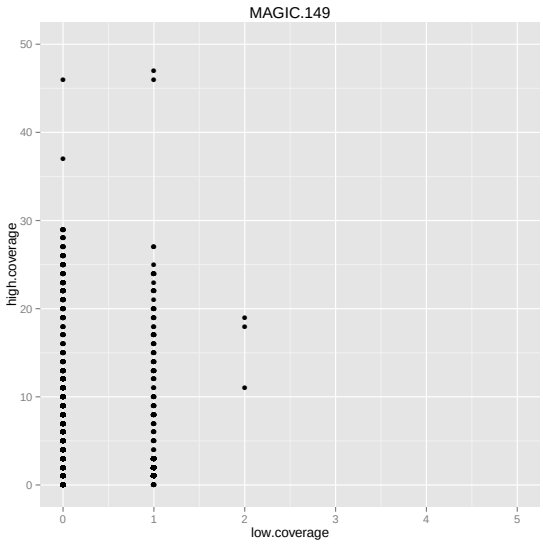
To detect endosymbionts present in the MAGIC genomes at the time of sequencing, I collected all unmapped sequencing reads, i.e. reads that did not align to the reference TAIR10, and mapped them against the genomes of a 16S rRNA database [135] using BLAT [60]. For each MAGIC line, I counted the number of reads corresponding to each endosymbiont in the database, which defined a score for that organism. I repeated the mapping for the 9 lines that were resequenced at high (12x) coverage (see Chapter 5, Section 5.6). In total $115,326$ endosymbionts mapped to at least one MAGIC line.

I checked whether low-coverage sequencing is suitable for analysis of endosymbiont levels, by comparing counts of endosymbionts in the 9 lines for which both high and low-coverage data was available. For a given line i, I computed the read counts L_{ij} and H_{ij}, respectively, denoting the score for endosymbiont j from low-coverage and high-coverage data, respectively. Correlation between L_{ij} and H_{ij} was high (mean $r = 0.84$) in all 9 lines (Figures 7.1a, 7.1c). This was reflected mainly in endosymbionts with higher read counts, since for $L_{ij} \leq 2$ there was little correlation (Figures 7.1b, 7.1d).

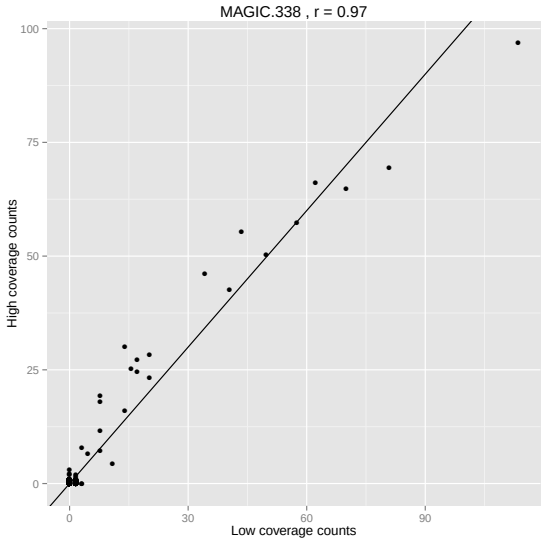
Based on this analysis, I selected 24 endosymbionts which were supported by at least 5 reads in at least 20 individuals and so their endosymbiont levels inferred from low-coverage reads would



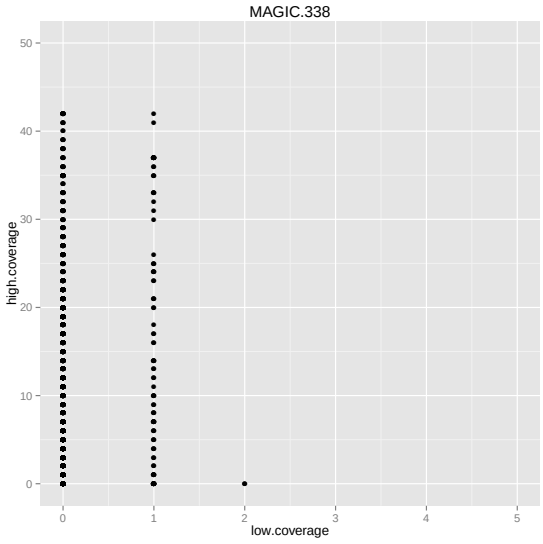
(a) MAGIC.149, scaled low and high coverage counts



(b) MAGIC.149, unscaled counts, $L_{ij} < 5$



(c) MAGIC.338, scaled low and high coverage counts



(d) MAGIC.338, unscaled counts, $L_{ij} < 5$

Figure 7.1: Comparison of endosymbiont levels extracted from unmapped read counts for 9 lines sequenced at both low and high-coverage. Each point represents an endosymbiont j, with x-axis indicating low-coverage reads supporting it (L_{ij}) and y-axis high-coverage reads H_{ij}. Figures 7.1a and 7.1c show counts of endosymbionts in low-and high coverage data in lines MAGIC.149 and MAGIC.338, scaled so that both range from 0 to 100 (in reality low-coverage counts can range up to $\sim$110 and high-coverage up to $\sim$5500). Figures 7.1b and 7.1d show unscaled versions of the same plots, only for endosymbionts with low read count (up to 5) in the low-coverage data.

be representative of high-coverage read counts.

7.2 Genome scan results for endosymbiont levels

Endosymbiont levels were variable across lines and so it is possible that variance can be explained by host genetic variation. For example, different host resistance alleles may allow different levels of pathogenic endosymbionts in leaf tissue. To investigate, I analysed the 24 endosymbionts with sufficiently variable read counts. Because read coverage between MAGIC lines vary (mean coverage is 0.3x but it ranges from $0.15 - 1.1$x), endosymbiont levels may be confounded by coverage. To control for this I regressed out coverage K_i from the read counts L_{ij} and extracted the residuals by fitting the linear model:

$$L_{ij} = aK_i + R_{ij} \tag{7.1}$$

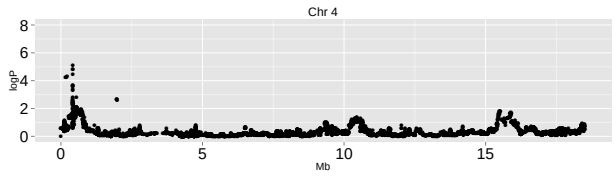
I mapped the residuals of the read counts R_{ij} using the haplotype genome scan described in Section 4.5 and performed 1000 phenotype permutations for each trait to determine genome-wide significance at each locus. The genome-wide significance threshold was set at $-\log P > 4$. 9 endosymbionts had a marginally significant QTL, at the same locus (chr 4: 0.425Mb), shown in Figure 7.2. I also combined the $\log P$ values of each locus using Fisher's method, i.e. using the fact that the sum of the natural logarithms of the p-values follows a chi-squared distribution:

$$2\sum_{1}^{k} lnP \sim X^2_{2k} \tag{7.2}$$

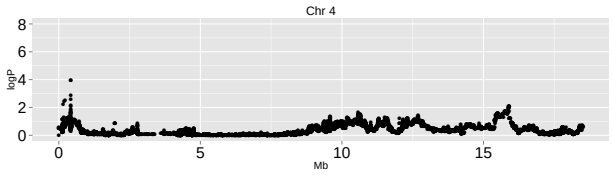
which increased the QTL signficance. However, Fisher's method assumes independent tests, which may not be the case for all 9 endosymbionts, as there are sequence similarities between 4 of the 9 endosymbionts so reads may have been counted twice.

CNGC13, a gene of the cyclic nucleotide-gated channels (CNGC) family which regulate salt uptake in the roots [56, 73] is 10kb upstream of the QTL. *Arabidopsis* is sensitive to high salinity, which has been shown to affect the root system architecture [56]. Rhizobacteria and fungal

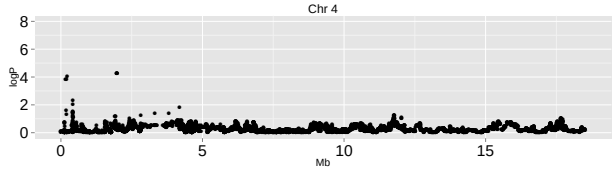
endosymbionts have been shown to reduce saline toxicity and alleviate salt stress [113, 31]. Most of the 9 endosymbionts that are associated with the locus are rhizobacteria that regulate nitrogen channels and that live in hyper-saline environments, and so are possibly associated with the host organism's response to salt stress. *Sinorhizobium meliloti (Ensifer meliloti)*, one of the associated endosymbionts, is a root bacterium, known to be associated with salinity stress tolerance in *A. thaliana* [46, 106]. *Acinobacter calcoaceticus* is also a plant rhizobacterium controlling nitrogen channels and controlling salt stress levels [57], while *Palucibacter propionicigenes* is not viable in some saline environments but also allows calcium salt production [126]. Two microorganisms known to live in hyper-saline soil are also among the 9 associated endosymbionts, namely the archaeon *Methanoalophilus mahii* [117] and the bacterium *Thioalkalimicrobium cyclicum*, which is highly similar to 4 of the 9 endosymbionts [116]. Finally, the last associated bacterium is *Phaeobacterium gallaeciensis* which is important in cell lignification and can also act as a pathogen, but is not known to be associated with response to salinity [105, 112].



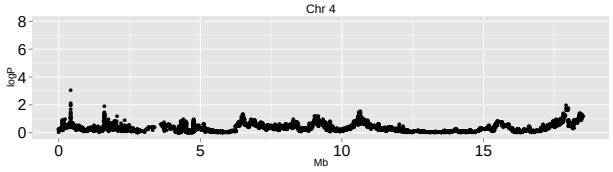
(a) *Sinorhizobium meliloti (ensifer meliloti)* - S001328289



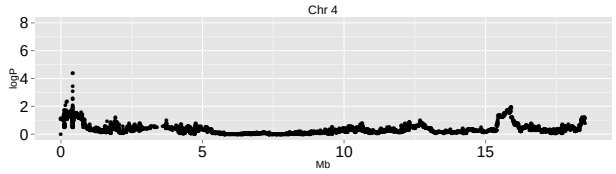
(b) *Acinobacter calcoaceticus* - S001036299



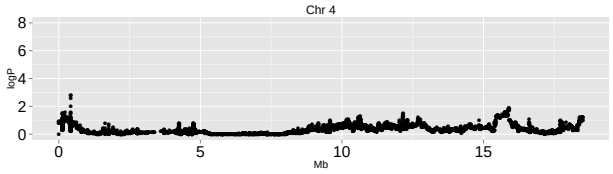
(c) *Palucibacter propionicigenes* - S001215453



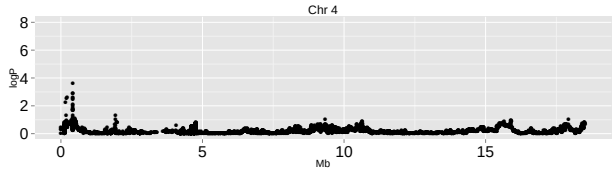
(d) *Methanoalophilus mahii* - S001337072



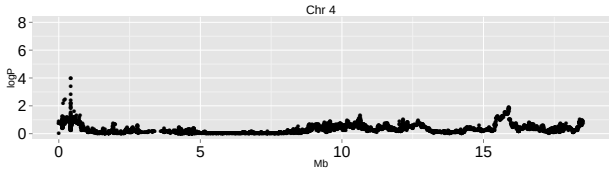
(e) *Thioalkalimicrobium cyclicum* - S001337659



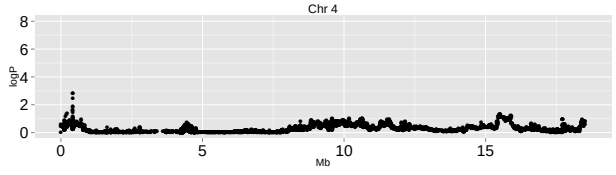
(f) *Thioalkalimicrobium cyclicum* - S001337931



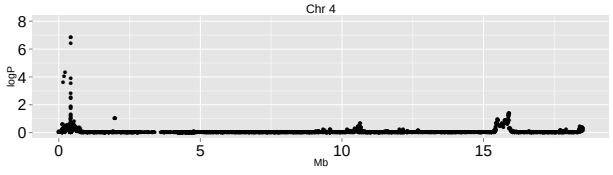
(g) *Thioalkalimicrobium cyclicum* - S001337899



(h) *Thioalkalimicrobium cyclicum* - S001338015



(i) *Phaeobacterium gallaeciensis* - S001337962



(j) Combined P-values using Fisher's method

Figure 7.2: 9 endosymbionts hosted in MAGIC genomes, controlled by the same QTL. Each figure is a manhattan plot of a single GWAS with phenotype the endosymbiont level of each organism.

7.3 Discussion

Unmapped reads from short-read sequencing data occasionally originate from endosymbionts that use the sequenced sample as host organism. Endosymbiont levels corresponding to low-coverage read count can capture the variation that would have been explained by high-coverage data, at least for abundant endosymbionts with large read counts. However, most endosymbionts in the low coverage data had very few (typically zero or one) reads per sample. I chose not to map these endosymbionts, because their scores might only reflect differences in read coverage and because they did not correlate very well with high-coverage counts. However, after controlling for differences in coverage, differences reflected in a binary endosymbiont score could be real. Therefore, it might be useful to complete the mapping for the full set of $115,326$ endosymbionts detected in MAGIC using the existing data.

From the mapping of 24 highly variant endosymbionts, I observed a single QTL shared across 9 endosymbionts. The genome-wide significance achieved by any of these traits was not impressive, however its recurrence indicates that it might be genuine. Endosymbionts of the same family, or with similar functions are likely regulated by the same gene(s), so they are likely highly related, rather than independent variables. Therefore, one could approach the problem by employing a multivariate approach that would combine several dependent traits to a single variable [63, 64]. Multivariate approaches have been shown to increase power in similar studies, and therefore would help validate the QTL.

Overall, I have shown proof-of-concept for this type of approach. A further analysis would be performed in a larger population of individuals sequenced at higher coverage. For example, the 3000 rice genomes recently sequenced by the Gates foundation [108] could be useful for this purpose.

Chapter 8

Conclusions

This thesis explored computational and statistical methods for the analysis of genetic variation segregating in a population, using *Arabidopsis thaliana* as a model organism, focusing on two aspects of genetic variation: structural variation and recombination. Throughout the thesis, I used low-coverage sequence data from the MAGIC *A. thaliana* population, so my analysis highlights the capabilities of low-coverage data as well as of the MAGIC population design. I presented algorithms for genetic imputation in MAGIC, which can reconstruct each line as a genetic mosaic at high precision with low-coverage data. Using the mosaics and the low-coverage sequence, I developed three novel applications of next-generation sequencing data. First, I described a method which combines anomalous read alignments and genetic mapping for the mapping of structural variants (SVs). The method can distinguish between short and long-range SVs and was used to identify, for the first time, a large number of translocations and inversions in a population. Second, I analysed the recombination history of MAGIC from the imputed genome mosaics and showed that it recapitulates the known *Arabidopsis* genetic map. Third, I displayed that some unmapped sequencing reads correspond to endosymbionts and gave evidence that variance in endosymbiont level is explained by genetics. In this chapter I review the results of the thesis and discuss its contribution to the field. I also suggest directions of future work.

8.1 Imputation of genome mosaics in MAGIC

Genetic imputation is necessary for QTL mapping with low-coverage data to complete missing genotypes. Several unsupervised algorithms have been proposed [122, 71], however the problem of haplotype inference in MAGIC is simpler as the population history is known. Haplotypes in MAGIC are resolved using dynamic programming or, equivalently, Hidden Markov Models (HMMs), either to recover the best annotated sequence or to compute posterior probabilities of every haplotype at each locus. I showed by simulation that both methods achieve high imputation accuracy in MAGIC, even with low coverage data. Furthermore, the imputation is reliable in fully or partially heterozygous genomes, so the algorithms can be employed for imputation in outbred populations, such as F2 and heterogeneous stocks.

8.2 Mapping SVs as quantitative traits

Detection of SVs from next-generation sequencing short reads is challenging, as it is difficult to distinguish true signal from noise. This is particularly the case over long-range structural variants involving sequence from remote loci, such as translocations. I presented a novel framework for the identification of SVs segregating in a population, that utilises genetic mapping. The method first defines quantitative traits from misaligned reads at a given locus, then searches for positions in the genome whose haplotype correlates with the extent of anomalous reads. If anomalous reads are random noise then their association with haplotypes are also random. However, if they are caused by segregating SVs, then specific haplotypes are associated with them at the locus and so they are controlled by QTLs. By systematically classifying anomalous reads throughout the genome and performing genetic mapping at each locus we can call segregating SVs at the loci that had SV QTLs. The position of the QTL reveals whether the SV affects the local sequence only (cis-QTL) or if remote loci are related (trans-QTL) and so can identify translocations.

The method detected $6,502$ SVs in Arabidopsis, of which 25% are translocations. The mapping resolution in MAGIC is too low ($\sim$200kb [67]) to pinpoint exact SV breakpoints, so I used high-coverage sequence from the 19 founders [36] for this purpose. An alternative for populations in

which the founders are unknown would be to collect high-coverage reads in a small fraction of genomes, covering the entire haplotype space. The SVs were validated in silico and experimentally, wherever possible, with high (82%) success rate.

I examined the relationship of SVs with gene expression and physiological phenotypes and discovered an important impact to both. Regarding gene expression, I found that SVs substantially disturb gene expression. Besides disrupted genes on the boundaries of SVs, genes that are transposed or inverted in their entirety are very often silenced. Furthermore, SVs contribute significantly to physiological traits. For example, a germination time QTL in *Arabidopsis* can be explained by a single SV. Anomalous reads at certain loci also explained large fractions of phenotypes even when they are not SV-related (in the sense of not mapping to an SV QTL), probably indicating functional regions missing from the reference genome.

Looking forward, this general framework can be applied to any sequenced population, synthetic or natural. For instance, it could be used to infer SVs in the human genome and detect associations with disease and other phenotypes. Furthermore, predicted SV QTLs can be used to improve the quality of sequencing read alignments. Because SVs dysregulate gene expression and, in many cases, explain very large fractions of trait heritability, designing a model that combines SNPs, SVs and anomalous read information may be worth pursuing to better understand the impact of SVs to phenotypes; such an analysis could reveal sources of missing heritability.

8.3 Recombination

Simulations of genome mosaics showed that the presence of standard sequencing errors does not affect the accuracy of the algorithms to reconstruct genome mosaics. Nevertheless, despite the soundness of the algorithm, we have seen that several factors, genomic and technical, can have serious impact on mosaic quality. I showed that unusual clusters of mosaic breakpoints, which were consistent with some of the literature on recombination in *A. thaliana* [23, 8, 138], and which appeared to affect a very large fraction of the population were mostly artifactual. Notably, there were multiple different error sources in the data that gave rise to clusters and should be taken into consideration in similar experiments, including residual heterozygosity in some of the MAGIC lines,

undetected heterozygosity of the founders and structural variation causing read misalignment. A fraction of the clusters are still present in the data, in spite of exhaustive efforts to explain them as artefacts. We are currently in the process of growing MAGIC lines containing clusters in order to re-sequence them with long reads using the Oxford Nanopore MinION platform [54].

After clearing artefacts, the mosaics revealed a partial recombination history of the MAGIC lines and were used to identify recombination hotspots. Comparing these hotspots with those previously detected by LD-based independent studies [50, 23], I showed that the MAGIC data are consistent with the genetic map, as there is significant overlap in hotspot positions. Therefore, recombination analysis using genomic mosaics is a reliable resource for recombination analysis and can serve as an alternative of LD analysis. Although the MAGIC recombination hotspots co-localised with the genetic map hotspots, there was variation in the recombination rates of hotspots between studies. Since LD does not distinguish recombination events with respect to timing, some hotspots in the genetic map may correspond to a single ancient event or have reduced activity at present. In contrast, recombination hotspots in MAGIC are inferred from recent recombinants thus indicate active hotspots. Therefore, the MAGIC hotspots can be used for further studies on the mechanism of recombination in plants, for example by searching for sequence motifs associated with recombination in breakpoint regions.

8.4 Mapping of endosymbionts in MAGIC genomes

I investigated whether variability in the number of endosymbionts hosted in the MAGIC genomes has a genetic component. Reads that do not align to the reference genome (unmapped reads) sometimes correspond to endosymbionts living on or inside the sequenced tissue. I designed a pilot study of genetic mapping of endosymbiont levels; in particular, I selected 24 endosymbionts whose presence was variable within MAGIC and also detectable with low-coverage data and constructed quantitative traits from read counts of every endosymbiont at each line. After mapping the traits, I observed that a subset of endosymbionts shared the same, marginally significant, QTL. Some of the associated endosymbionts are rhizobia controlling nitrogen-channels and regulating salt tolerance while others live in hyper-saline environments. Supporting this, a gene controlling salt-uptake at

the roots is near the QTL, lending support to the hypothesis that groups of endosymbionts may be controlled by the same gene. I conclude that the study of endosymbionts from unmapped reads is feasible. Designing multivariate methods that would combine multiple traits to a single one might help increase statistical power with low-coverage data.

8.5 Conclusion

Complex traits are often associated with a large number of polymorphisms of small effects that require large population samples to detect them; current GWAS are often underpowered due to small sample sizes. Low-coverage sequencing has emerged as an approach to deal with this problem as obtaining fewer sequencing reads per sample allows to increase population samples at the same cost. This thesis has shown that low-coverage sequence and imputation is a suitable design for a variety of genomics studies, including structural variation discovery, recombination analysis and mapping of QTLs. The benefits of using low-coverage sequencing in regard to study cost, efficiency and, critically, statistical power are great. Although prediction accuracy is sometimes compromised with low-coverage sequence, the differences are negligible for many problems (for example in imputation accuracy), while others can be dealt with by obtaining high-coverage sequence for a few samples at a fraction of the cost. Therefore, large-scale low-coverage sequencing studies is a viable study design which can advance our understanding of complex traits and has the potential of becoming the main GWAS paradigm in the future.

Appendices

Appendix A

Genome-wide distributions of read anomaly traits

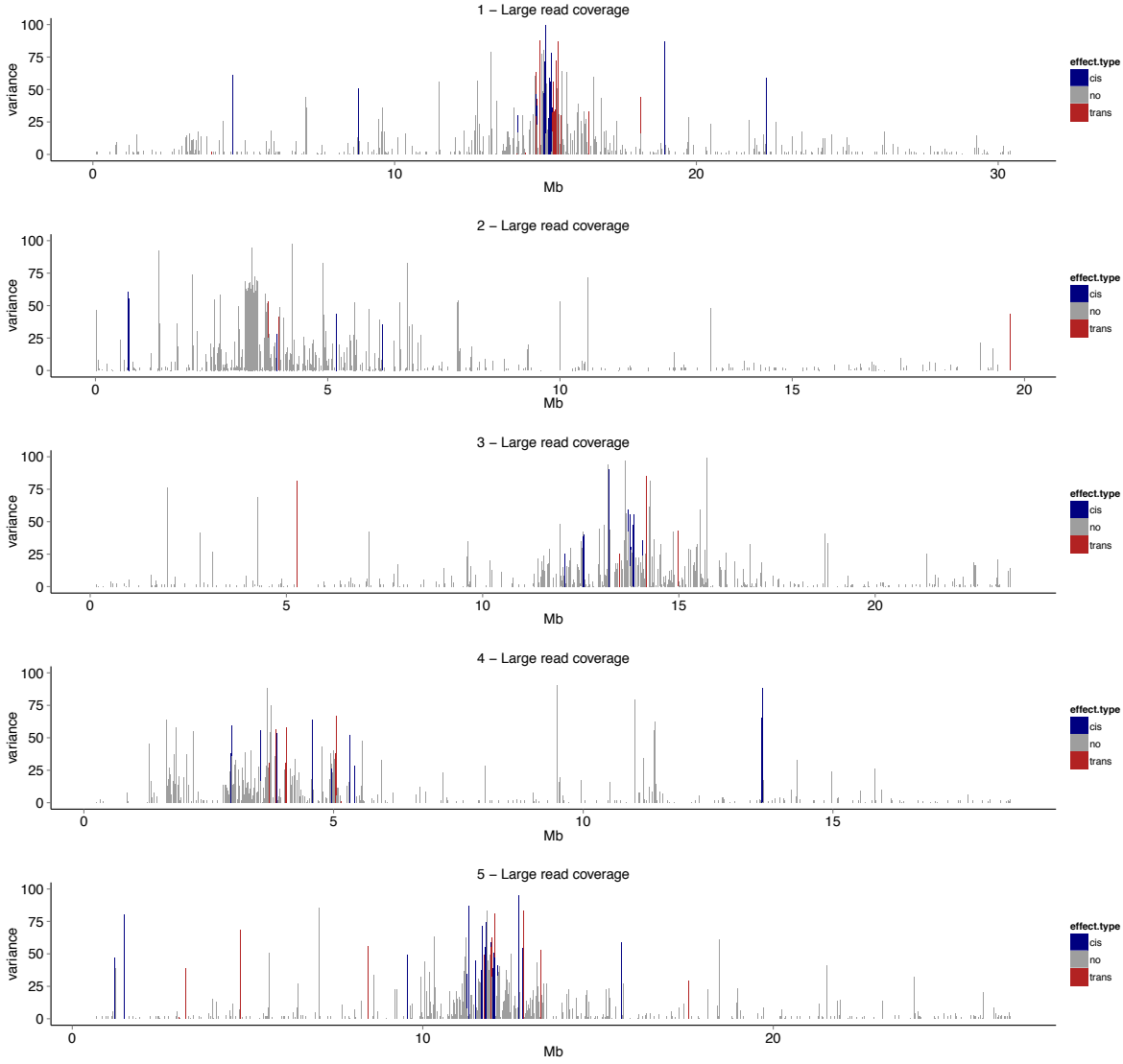


Figure A.1: High read coverage

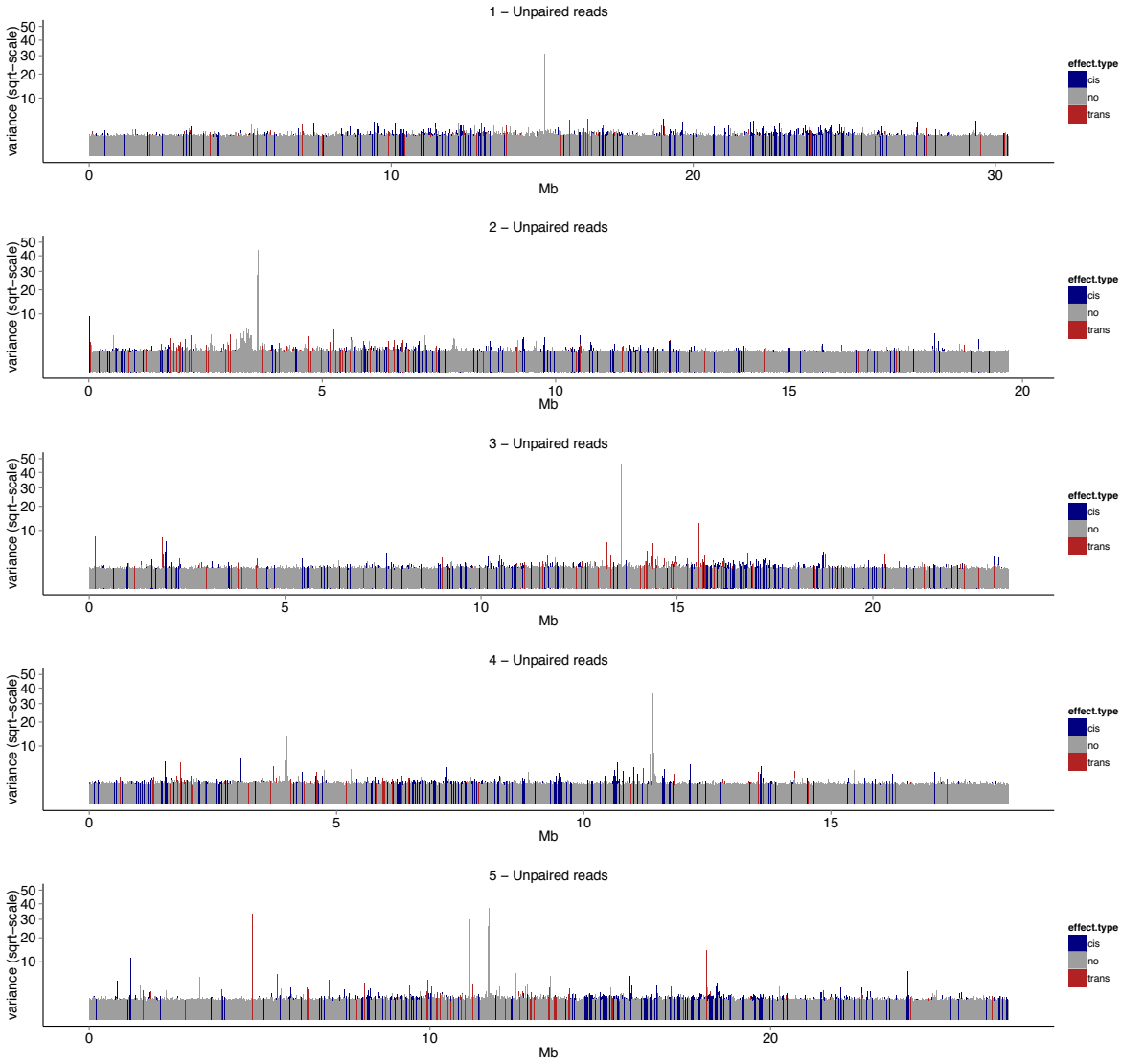


Figure A.2: Unpaired reads

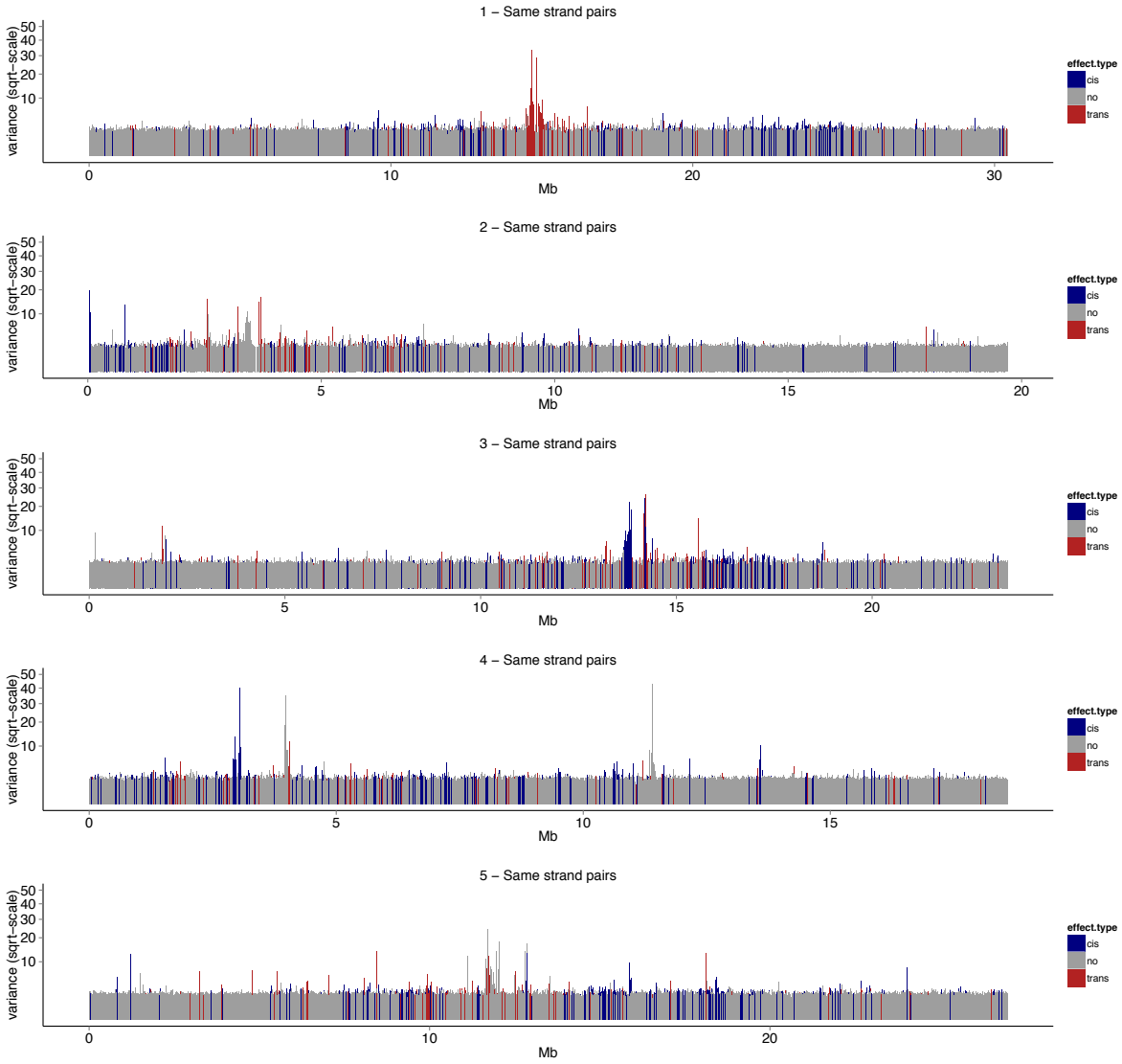


Figure A.3: Read pairs on the same strand

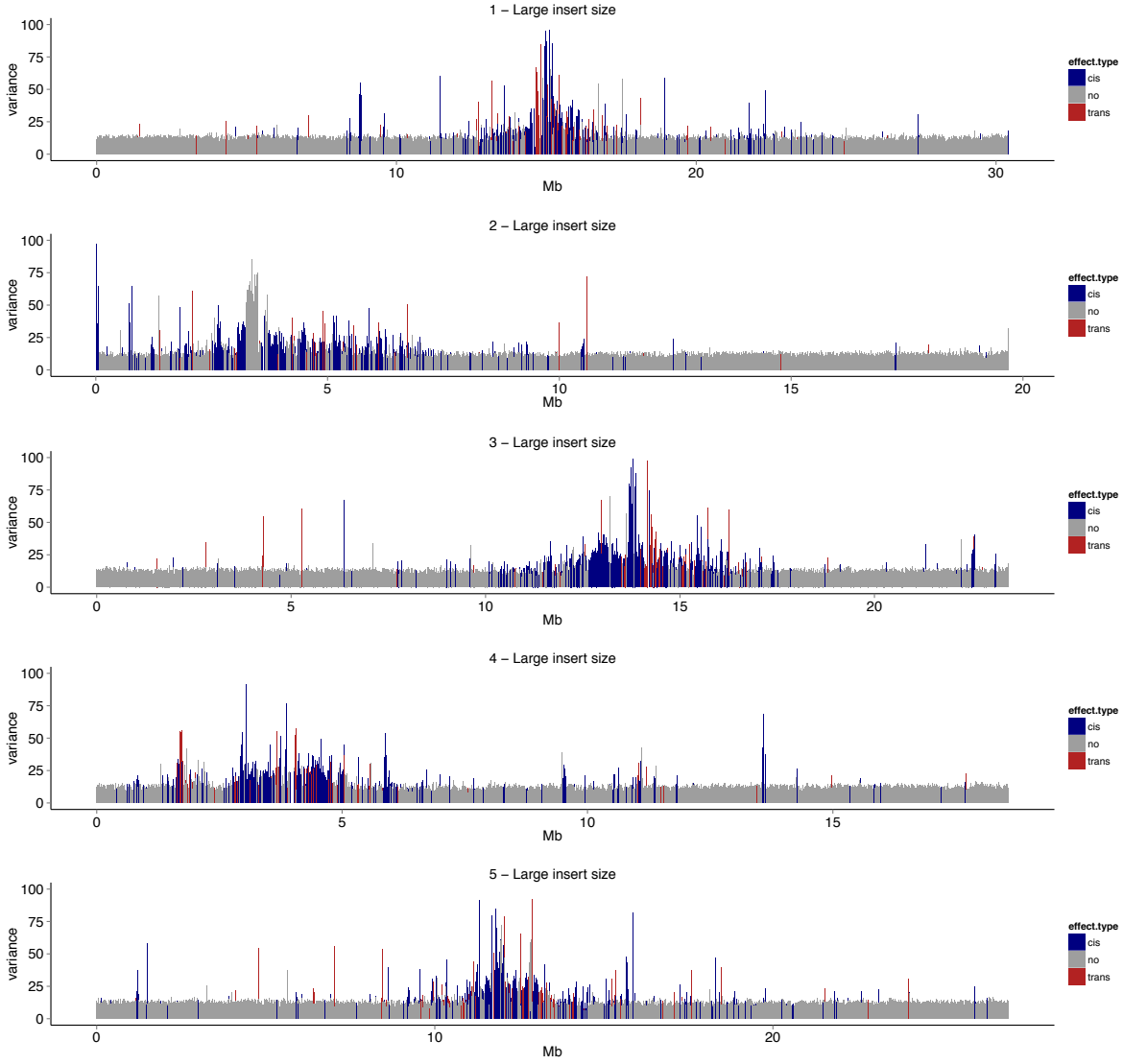


Figure A.4: Large insert size

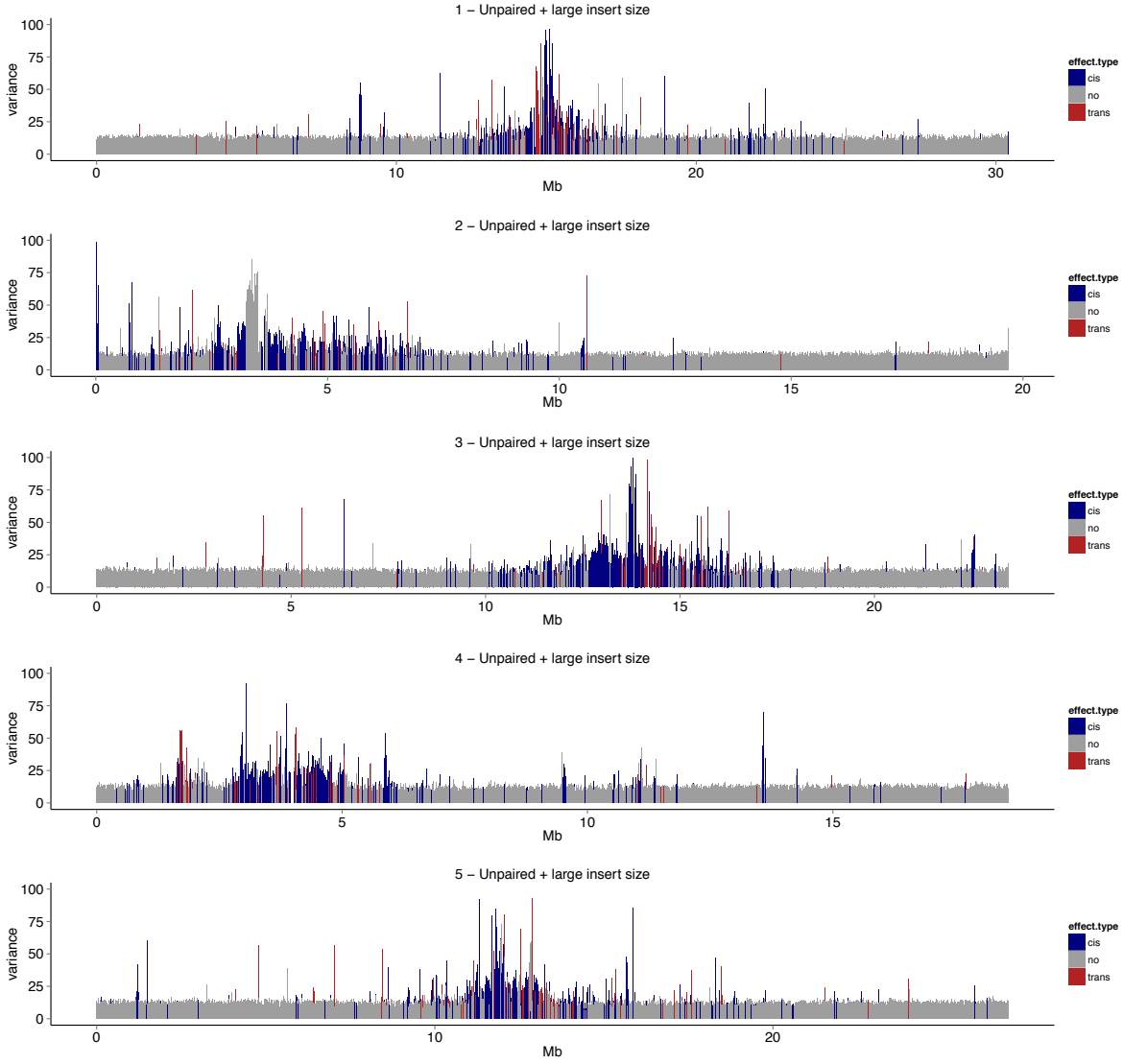


Figure A.5: Unpaired reads or with large insert size

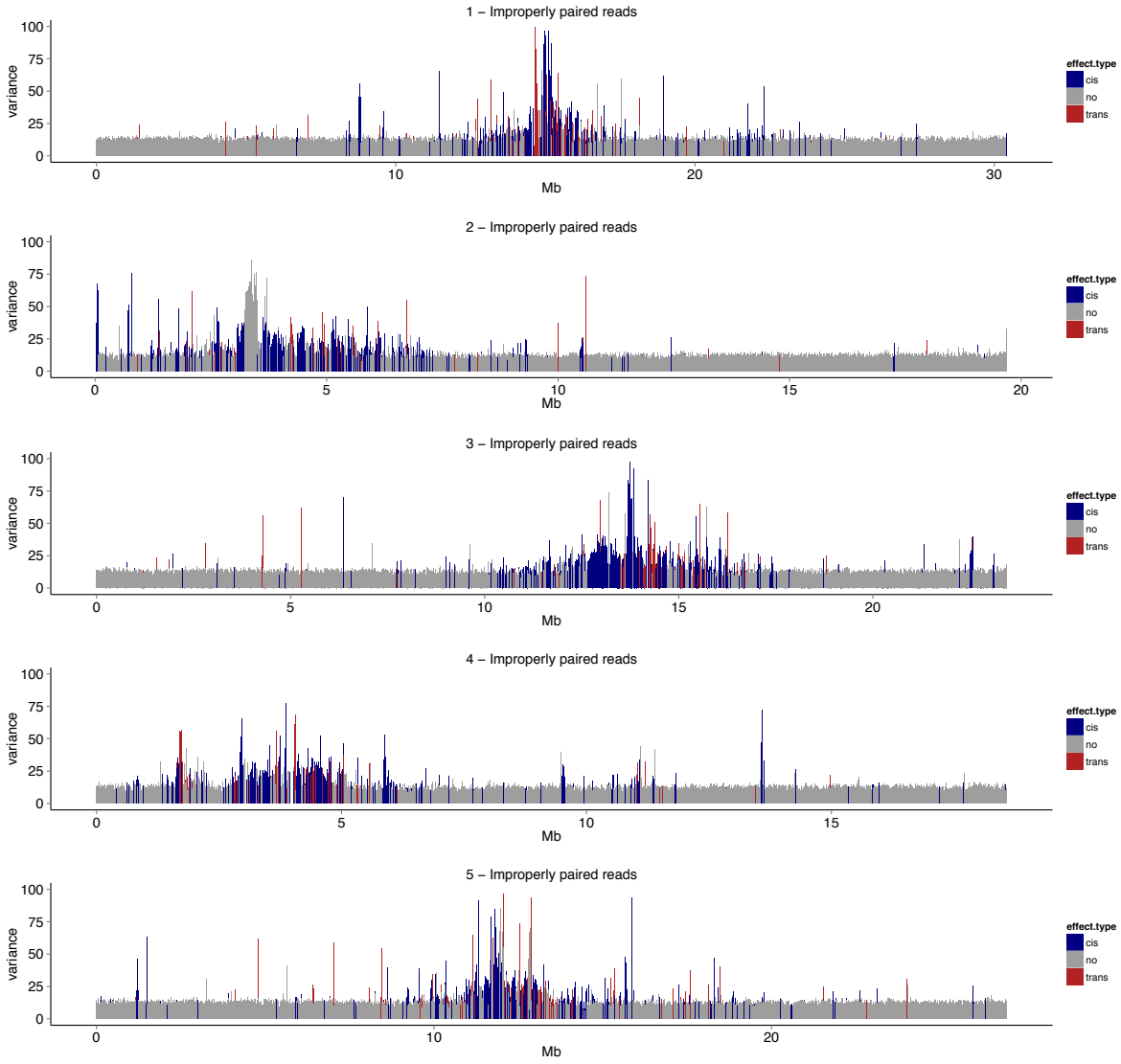
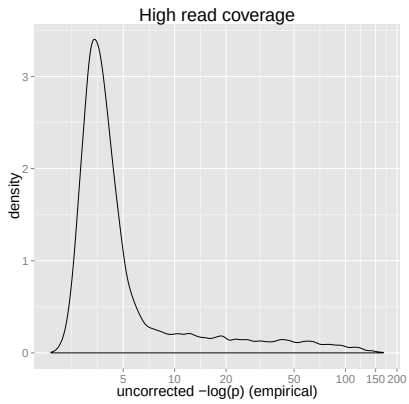


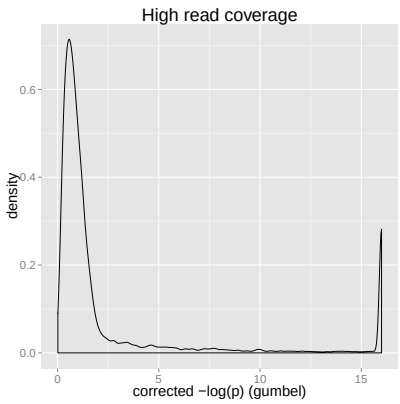
Figure A.6: Improperly paired reads

Appendix B

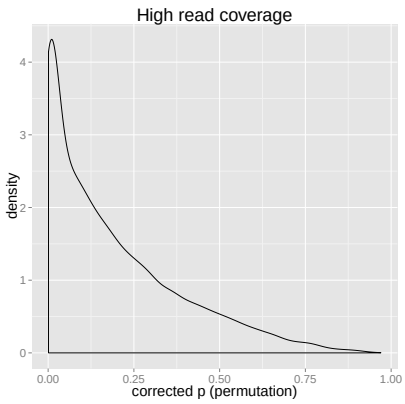
Density of uncorrected empirical p-values (λ_A), gumbel p-values (γ_A) and permutation p-values (π_A) for all traits in the six types of anomalous reads



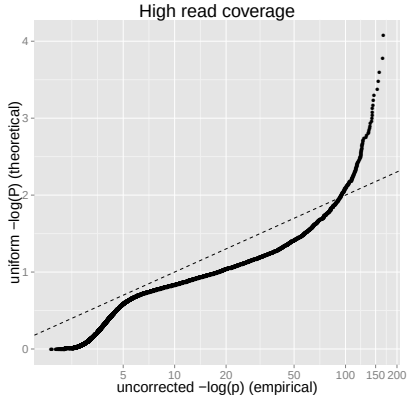
(a) Density of λ_A



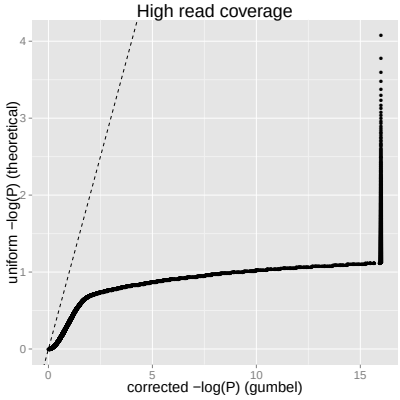
(b) Density of γ_A



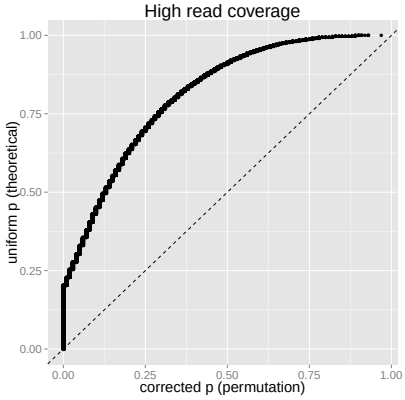
(c) Density of π_A



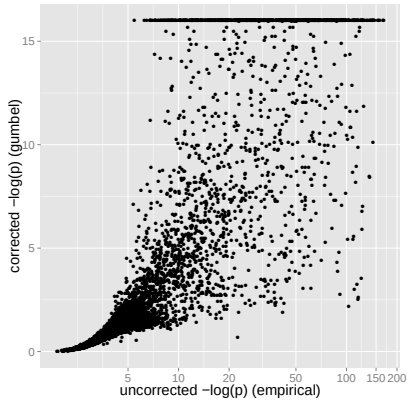
(d) QQplot: x: λ_A, y: Uniform



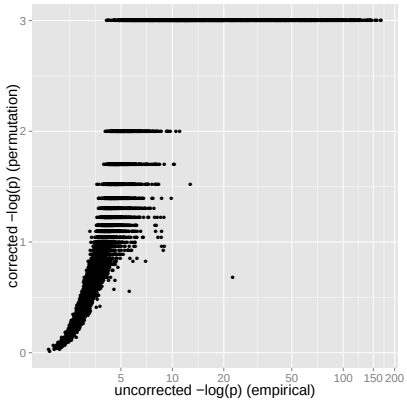
(e) QQplot: x: γ_A y: Uniform



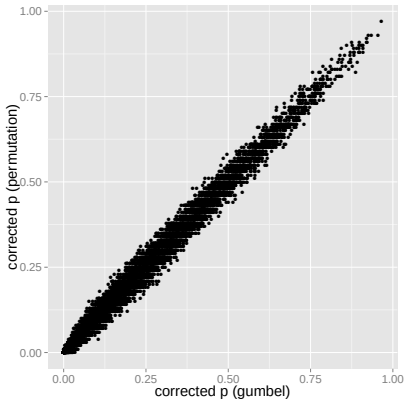
(f) QQplot: x: π_A y: Uniform



(g) Scatterplot: x: λ_A (log) y: γ_A (log)



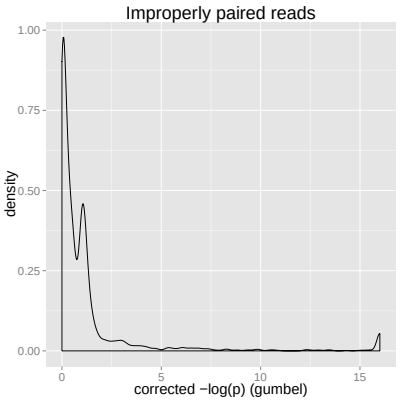
(h) Scatterplot: x: λ_A (log) y: γ_A (log)



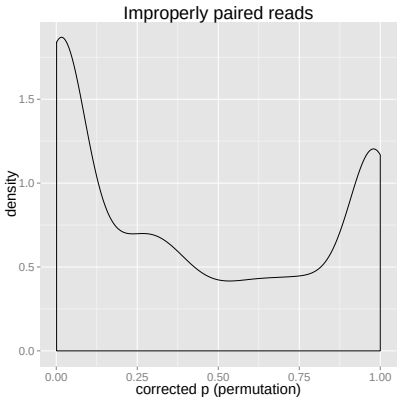
(i) Scatterplot: x: γ_A y: π_A

Figure B.1: High read coverage

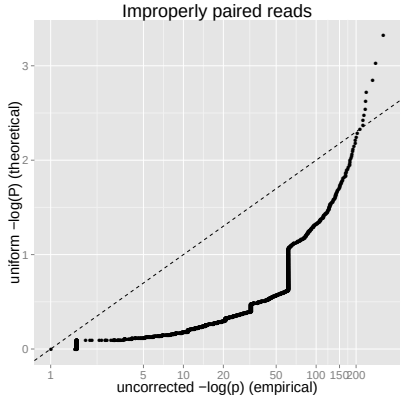
(a) Density of λ_A



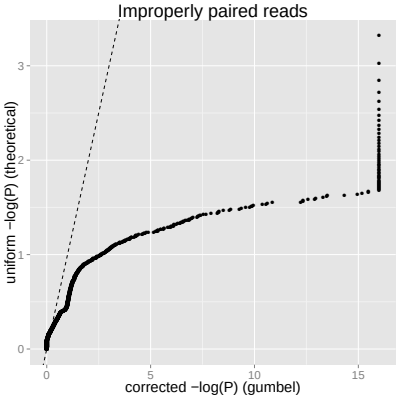
(b) Density of γ_A



(c) Density of π_A

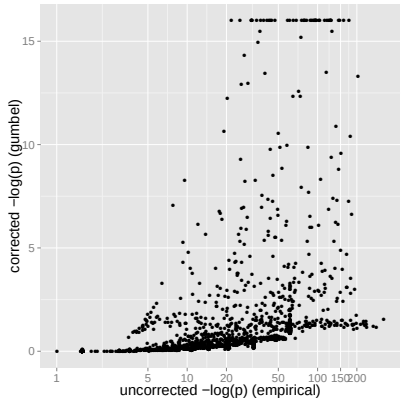


(d) QQplot: x: λ_A, y: Uniform

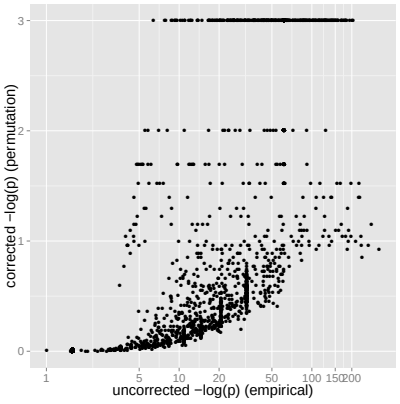


(e) QQplot: x: γ_A y: Uniform

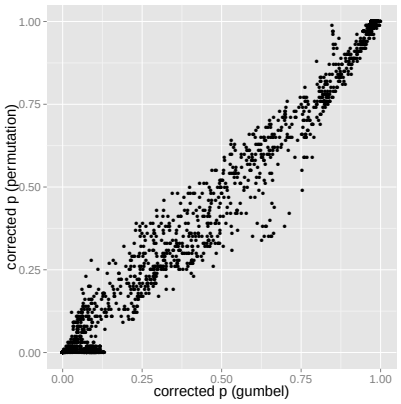
(f) QQplot: x: π_A y: Uniform



(g) Scatterplot: x: λ_A (log) y: γ_A (log)

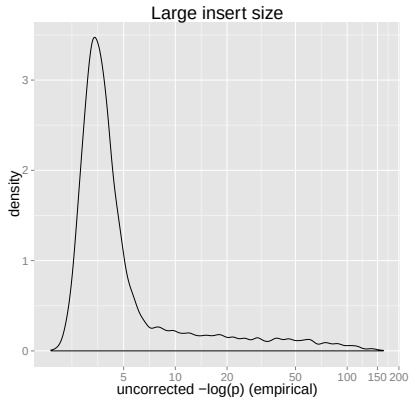


(h) Scatterplot: x: λ_A (log) y: γ_A (log)

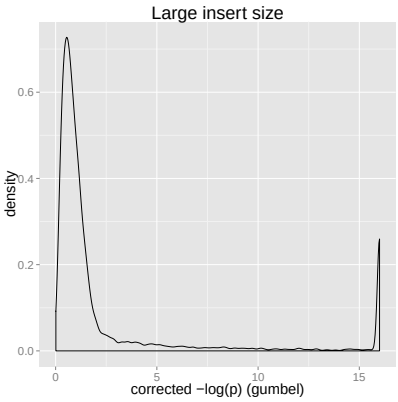


(i) Scatterplot: x: γ_A y: π_A

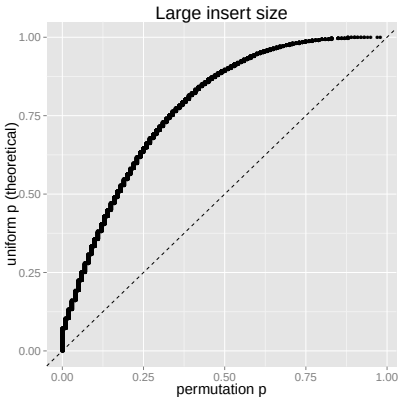
Figure B.2: Improperly paired reads



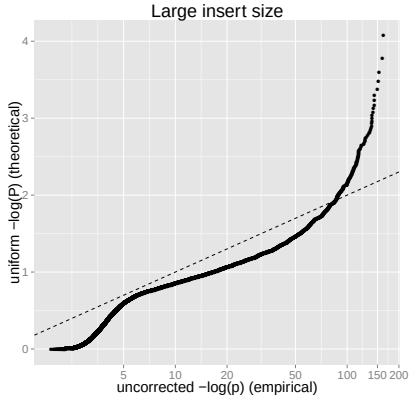
(a) Density of λ_A



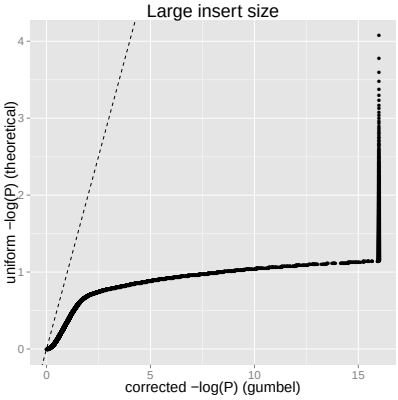
(b) Density of γ_A



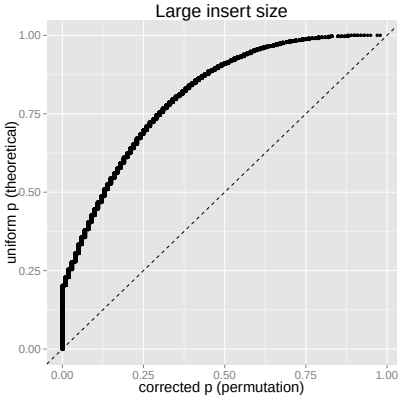
(c) Density of π_A



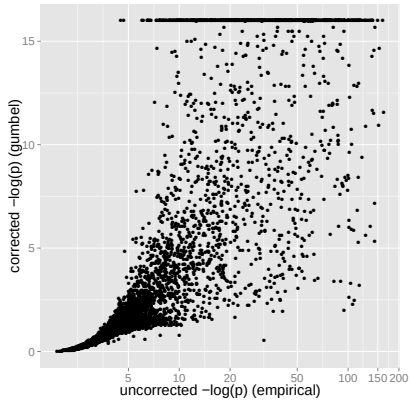
(d) QQplot: x: λ_A, y: Uniform



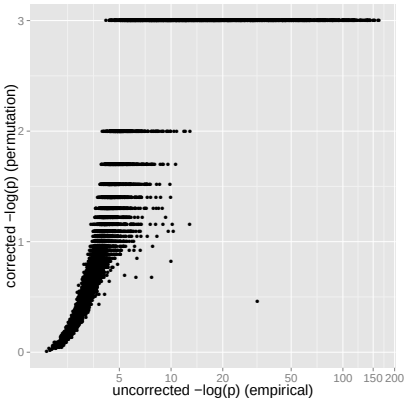
(e) QQplot: x: γ_A y: Uniform



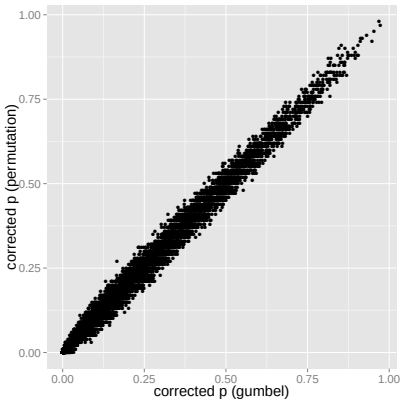
(f) QQplot: x: π_A y: Uniform



(g) Scatterplot: x: λ_A (log) y: γ_A (log)

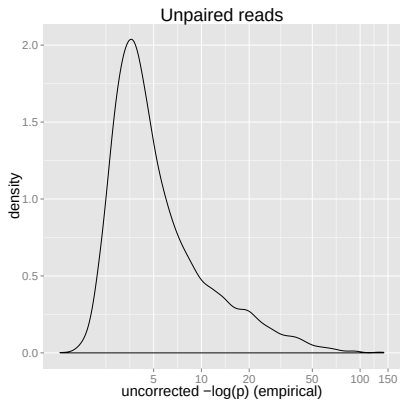


(h) Scatterplot: x: λ_A (log) y: γ_A (log)

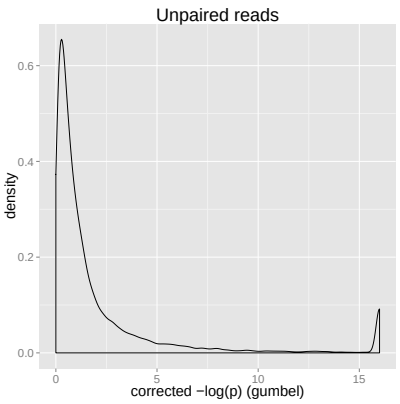


(i) Scatterplot: x: γ_A y: π_A

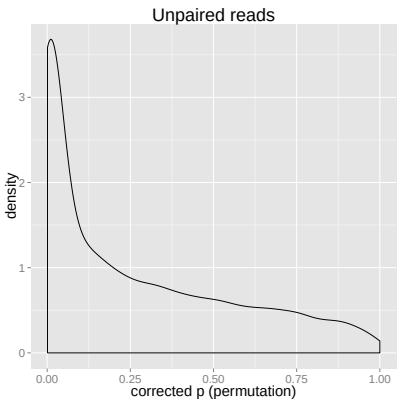
Figure B.3: Reads with large insert size



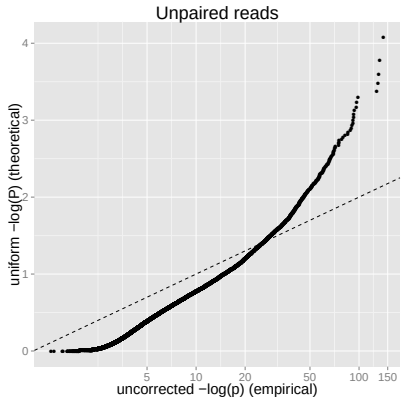
(a) Density of λ_A



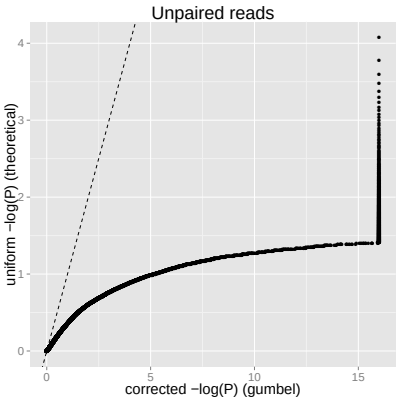
(b) Density of γ_A



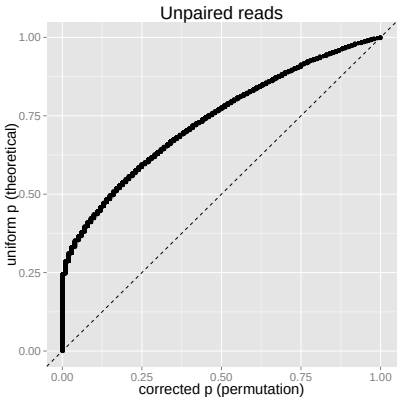
(c) Density of π_A



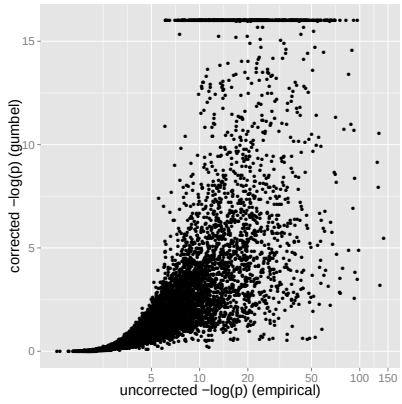
(d) QQplot: x: λ_A, y: Uniform



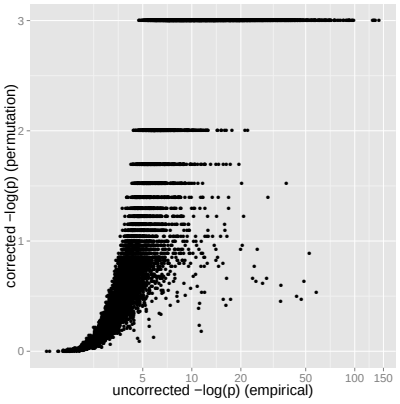
(e) QQplot: x: γ_A y: Uniform



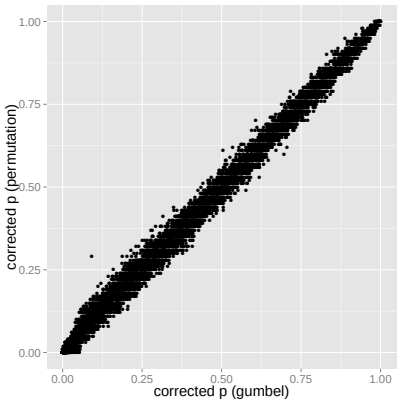
(f) QQplot: x: π_A y: Uniform



(g) Scatterplot: x: λ_A (log) y: γ_A (log)

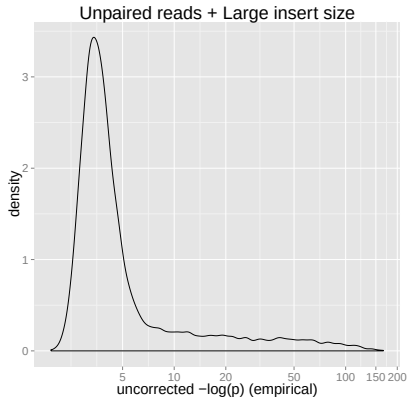


(h) Scatterplot: x: λ_A (log) y: γ_A (log)

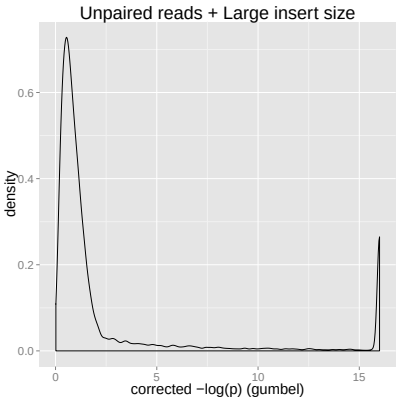


(i) Scatterplot: x: γ_A y: π_A

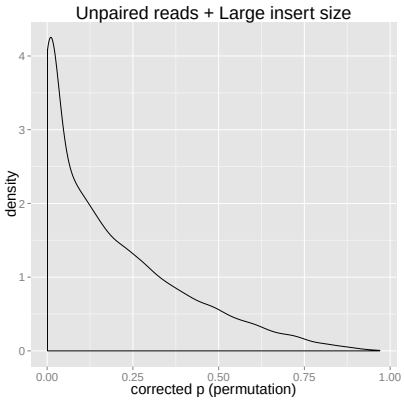
Figure B.4: Unpaired reads



(a) Density of λ_A

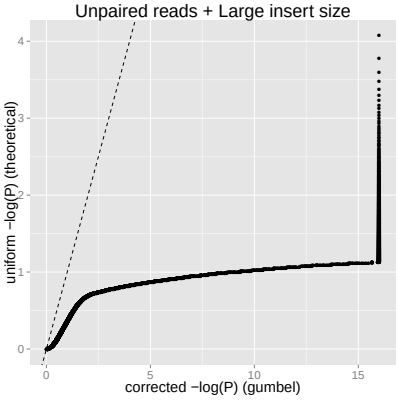


(b) Density of γ_A

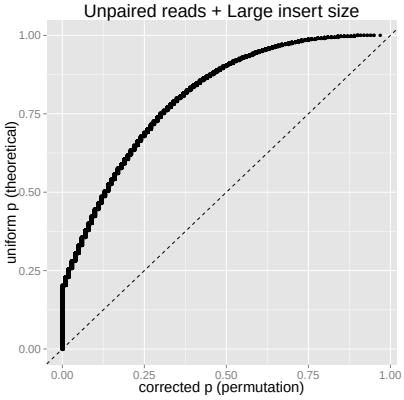


(c) Density of π_A

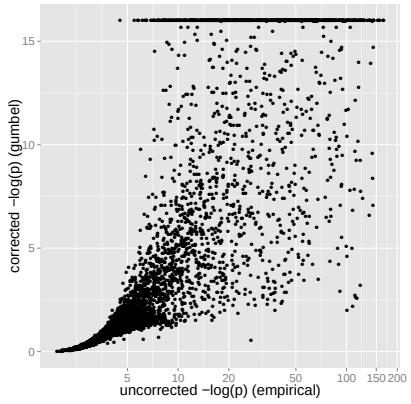
(d) QQplot: x: λ_A, y: Uniform



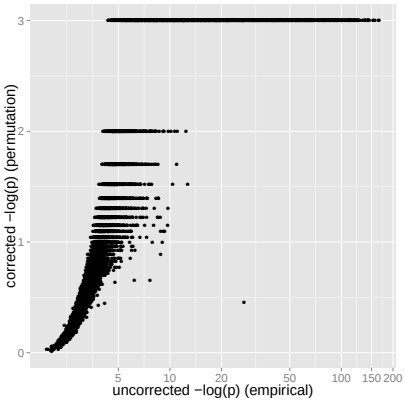
(e) QQplot: x: γ_A y: Uniform



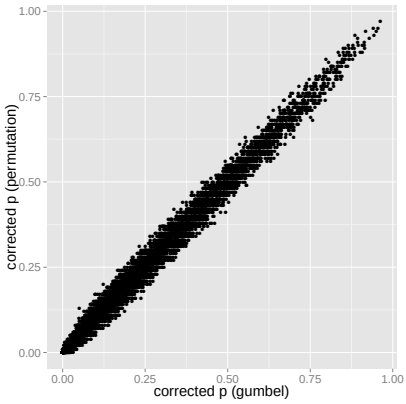
(f) QQplot: x: π_A y: Uniform



(g) Scatterplot: x: λ_A (log) y: γ_A (log)

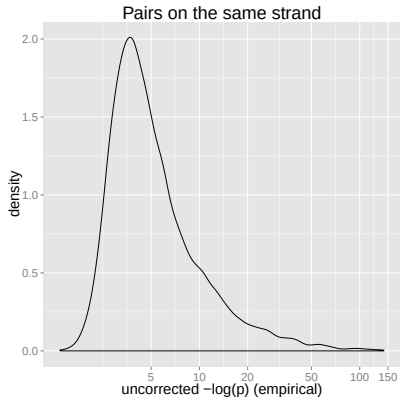


(h) Scatterplot: x: λ_A (log) y: γ_A (log)

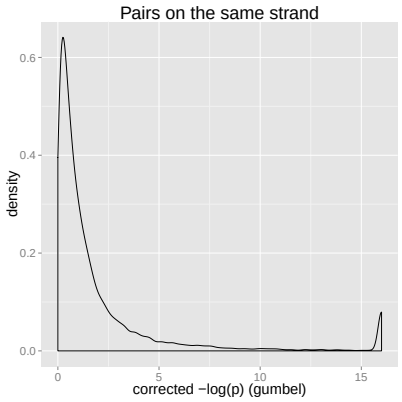


(i) Scatterplot: x: γ_A y: π_A

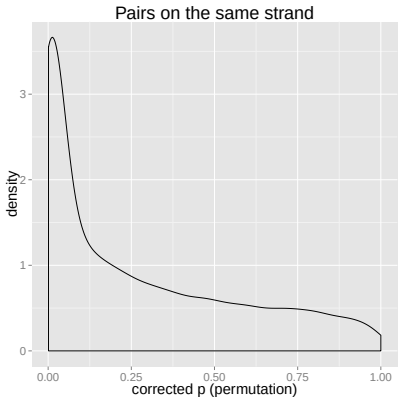
Figure B.5: Unpaired reads + reads with large insert size



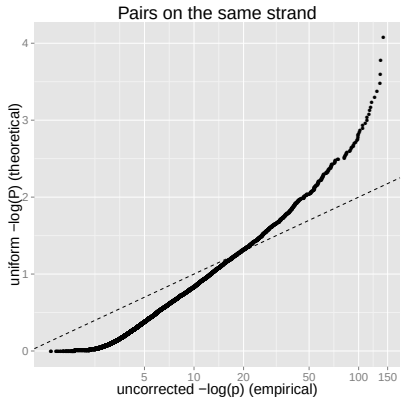
(a) Density of λ_A



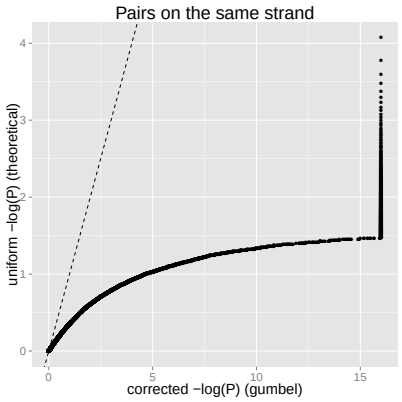
(b) Density of γ_A



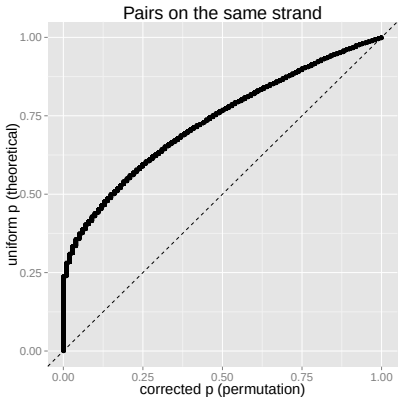
(c) Density of π_A



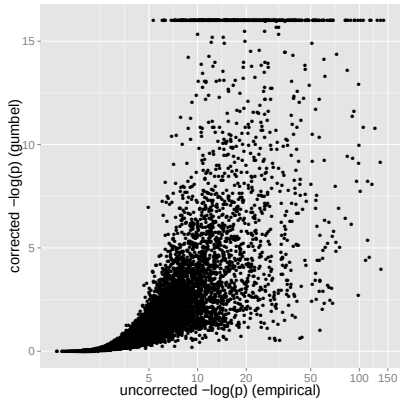
(d) QQplot: x: λ_A, y: Uniform



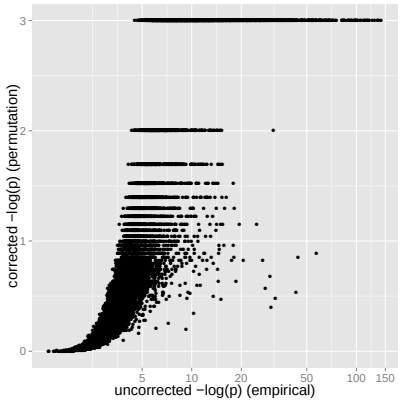
(e) QQplot: x: γ_A y: Uniform



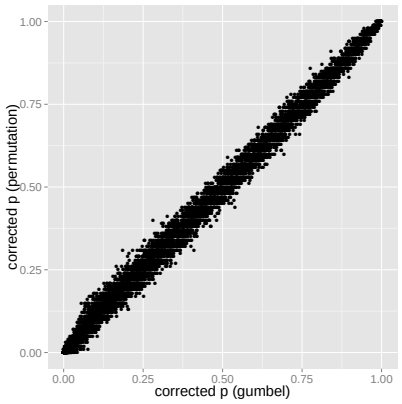
(f) QQplot: x: π_A y: Uniform



(g) Scatterplot: x: λ_A (log) y: γ_A (log)



(h) Scatterplot: x: λ_A (log) y: γ_A (log)

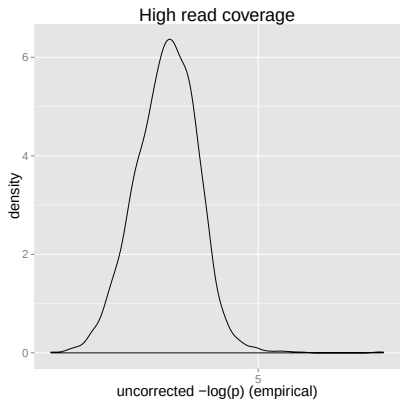


(i) Scatterplot: x: γ_A y: π_A

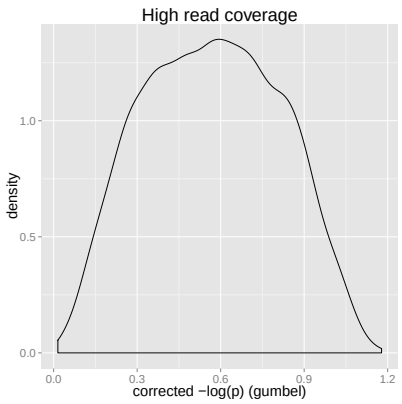
Figure B.6: Read pairs on the same strand

Appendix C

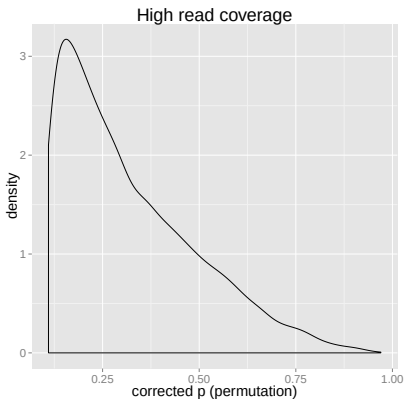
Density of uncorrected empirical p-values (λ_A), gumbel p-values (γ_A) and permutation p-values (π_A) for traits mapped to different chromosomes and with $P_p > 0.1$



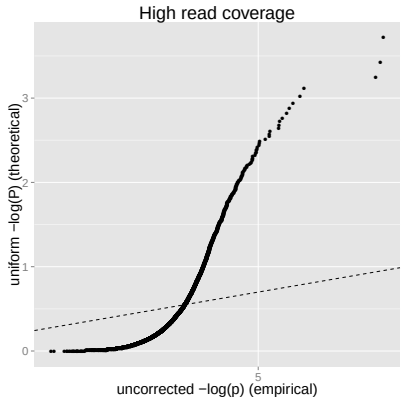
(a) Density of λ_A



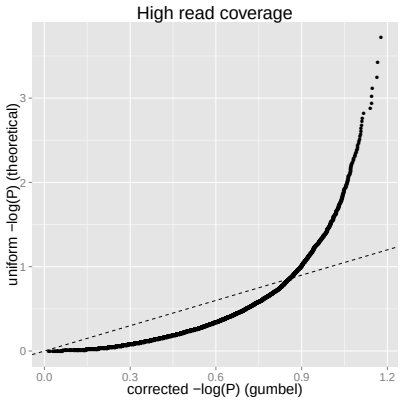
(b) Density of γ_A



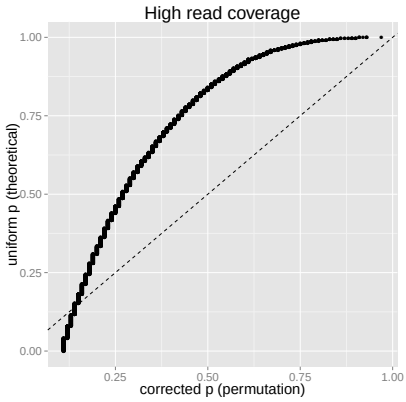
(c) Density of π_A



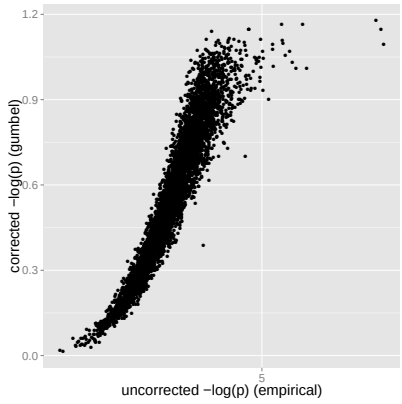
(d) QQplot: x: λ_A, y: Uniform



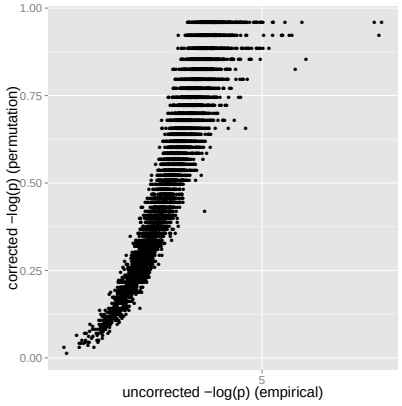
(e) QQplot: x: γ_A y: Uniform



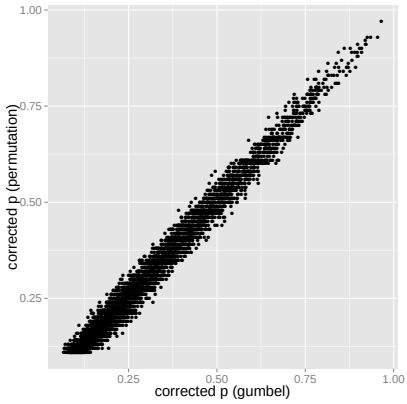
(f) QQplot: x: π_A y: Uniform



(g) Scatterplot: x: λ_A (log) y: γ_A (log)

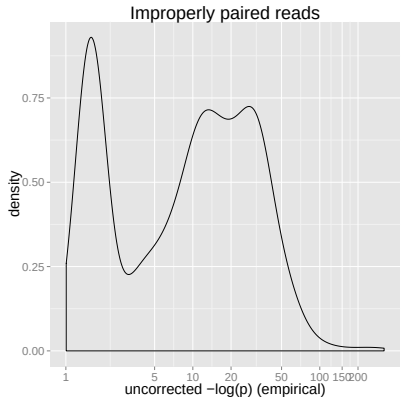


(h) Scatterplot: x: λ_A (log) y: γ_A (log)

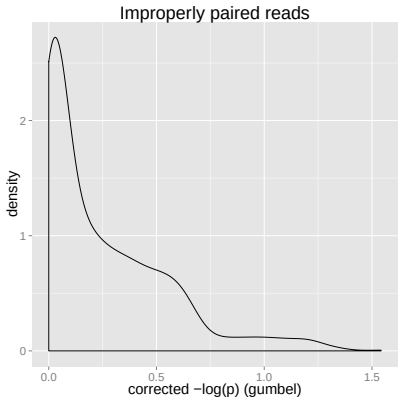


(i) Scatterplot: x: γ_A y: π_A

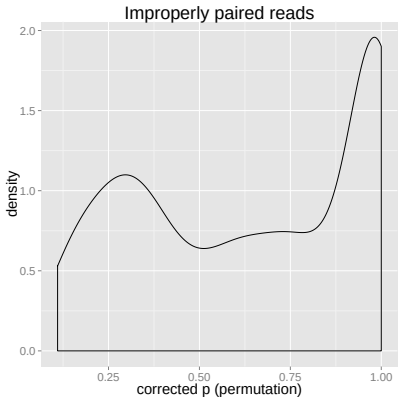
Figure C.1: High read coverage



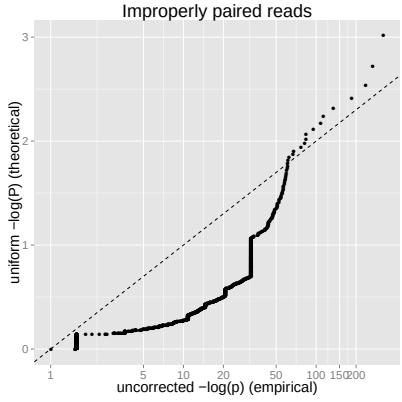
(a) Density of λ_A



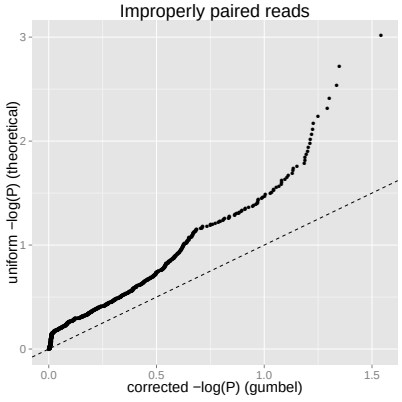
(b) Density of γ_A



(c) Density of π_A

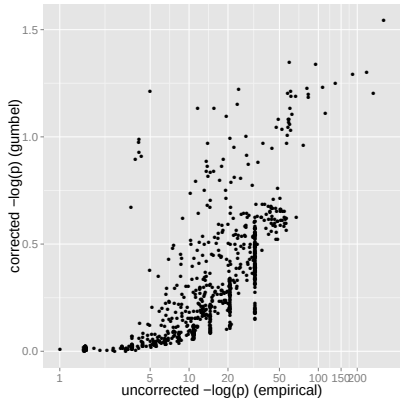


(d) QQplot: x: λ_A, y: Uniform

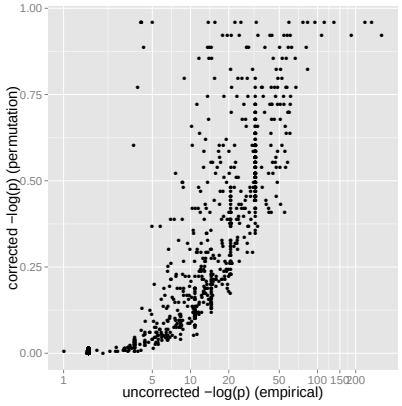


(e) QQplot: x: γ_A y: Uniform

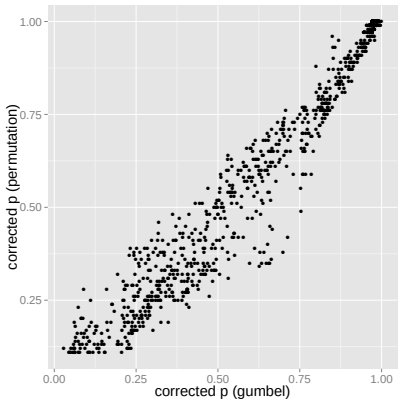
(f) QQplot: x: π_A y: Uniform



(g) Scatterplot: x: λ_A (log) y: γ_A (log)

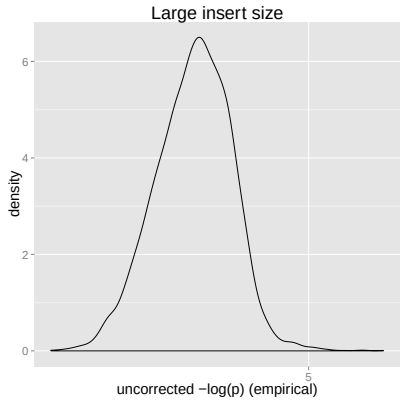


(h) Scatterplot: x: λ_A (log) y: γ_A (log)

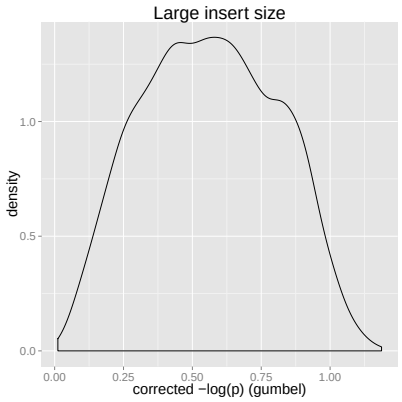


(i) Scatterplot: x: γ_A y: π_A

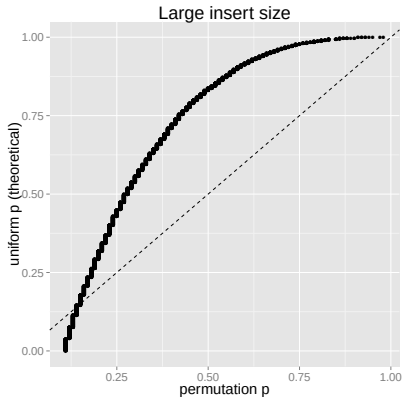
Figure C.2: Improperly paired reads



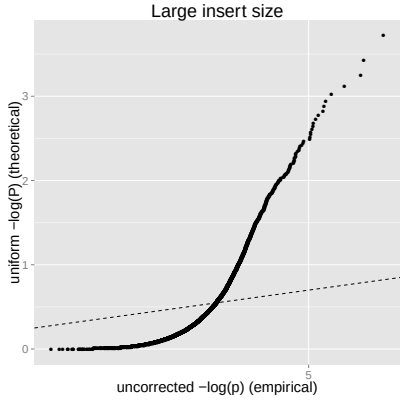
(a) Density of λ_A



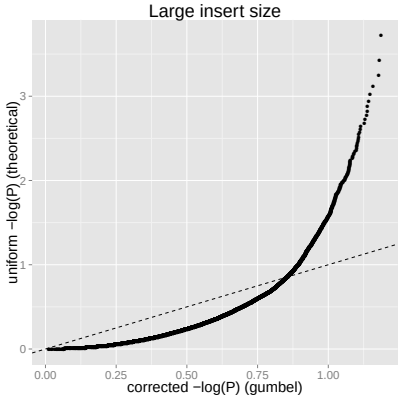
(b) Density of γ_A



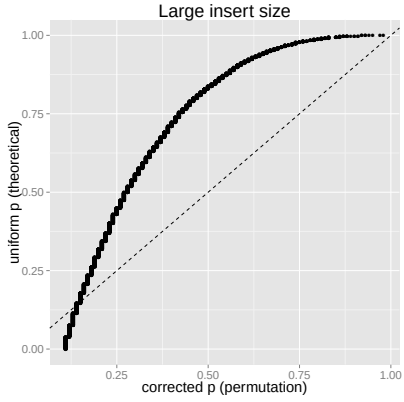
(c) Density of π_A



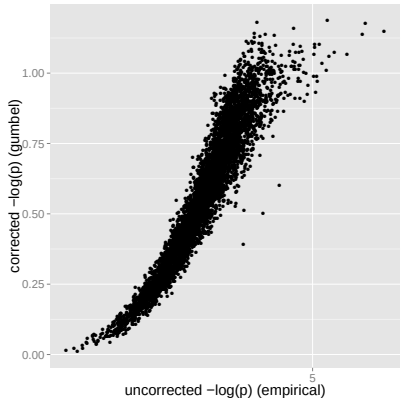
(d) QQplot: x: λ_A, y: Uniform



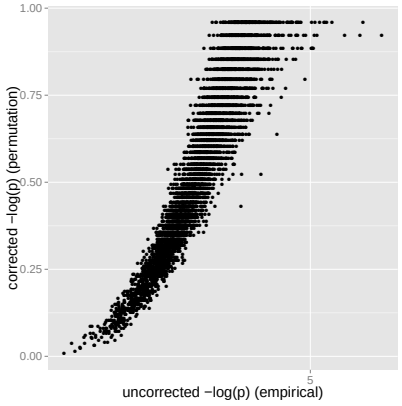
(e) QQplot: x: γ_A y: Uniform



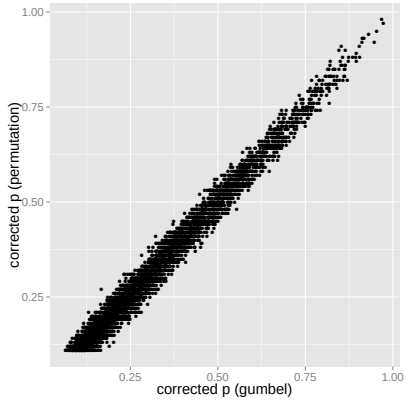
(f) QQplot: x: π_A y: Uniform



(g) Scatterplot: x: λ_A (log) y: γ_A (log)

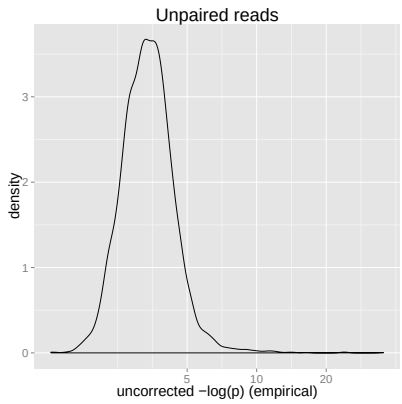


(h) Scatterplot: x: λ_A (log) y: γ_A (log)

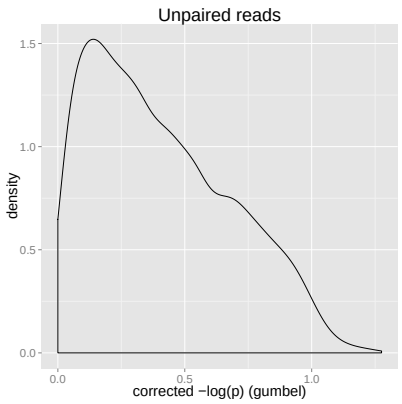


(i) Scatterplot: x: γ_A y: π_A

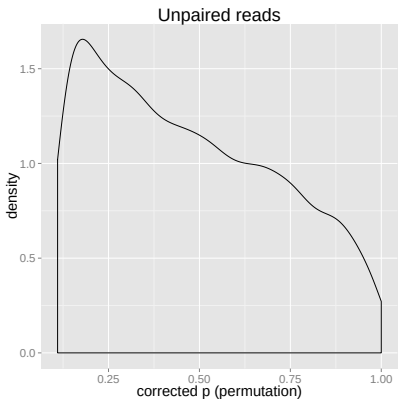
Figure C.3: Reads with large insert size



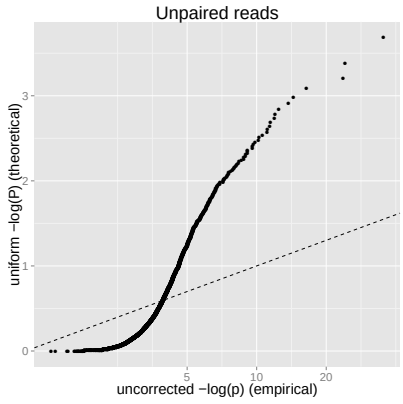
(a) Density of λ_A



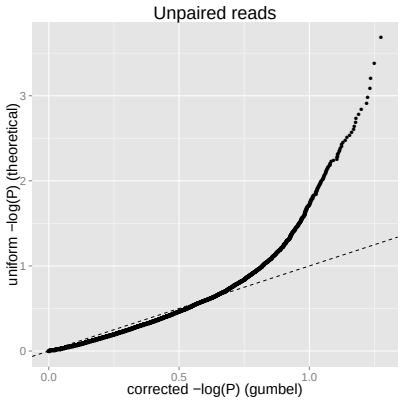
(b) Density of γ_A



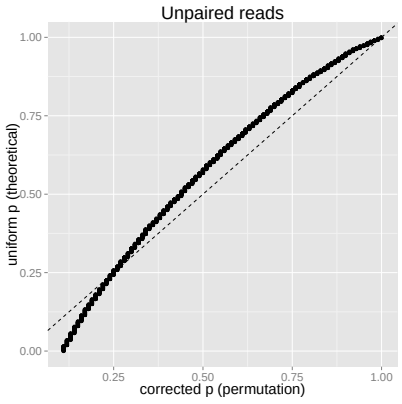
(c) Density of π_A



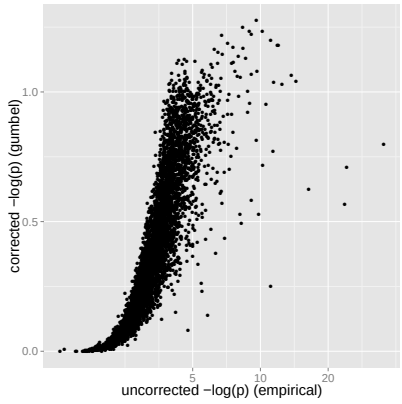
(d) QQplot: x: λ_A, y: Uniform



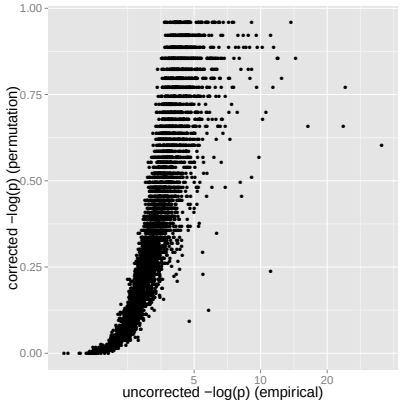
(e) QQplot: x: γ_A y: Uniform



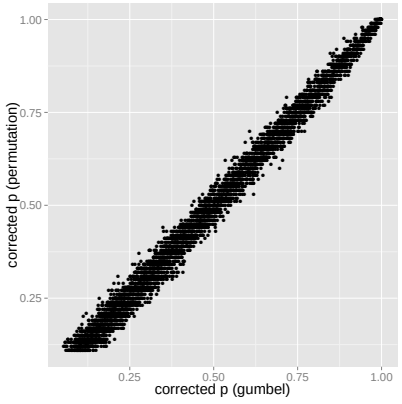
(f) QQplot: x: π_A y: Uniform



(g) Scatterplot: x: λ_A (log) y: γ_A (log)

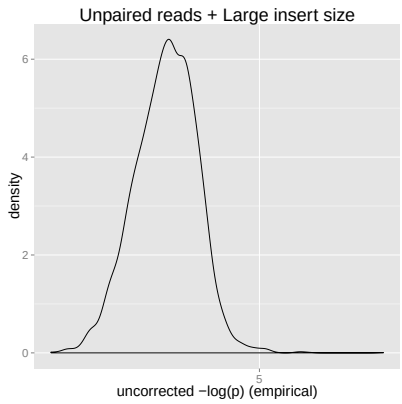


(h) Scatterplot: x: λ_A (log) y: γ_A (log)

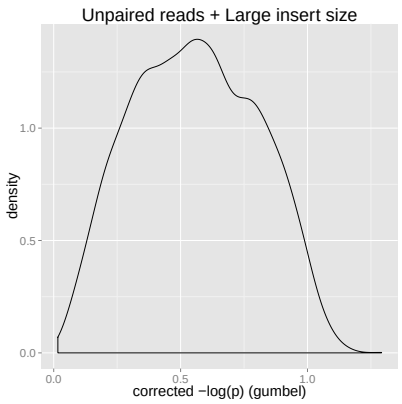


(i) Scatterplot: x: γ_A y: π_A

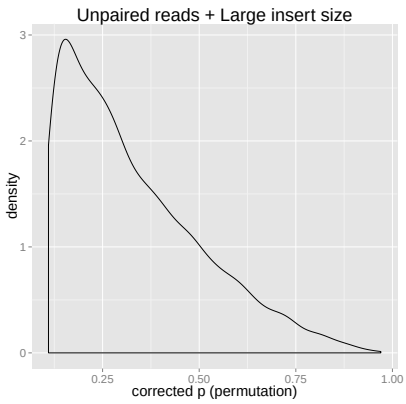
Figure C.4: Unpaired reads



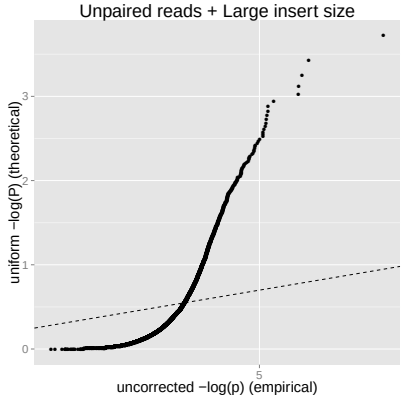
(a) Density of λ_A



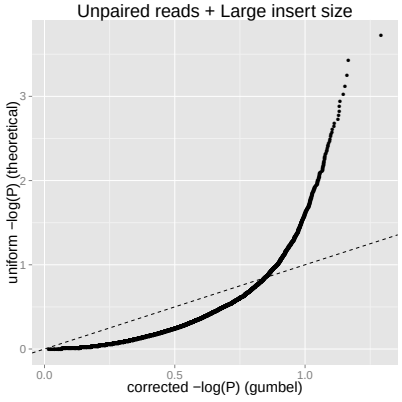
(b) Density of γ_A



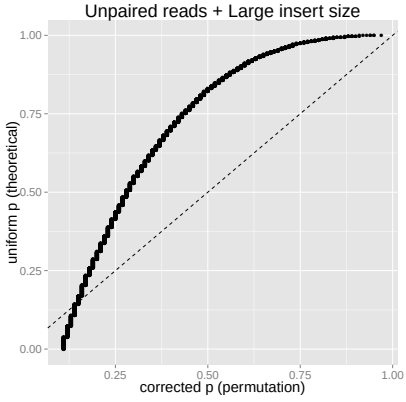
(c) Density of π_A



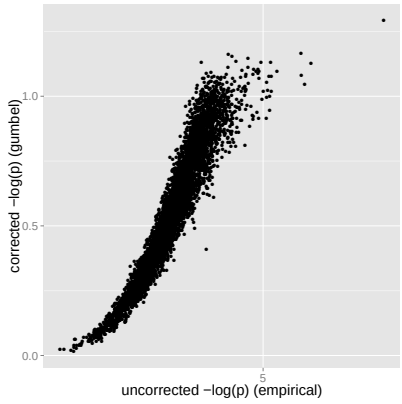
(d) QQplot: x: λ_A, y: Uniform



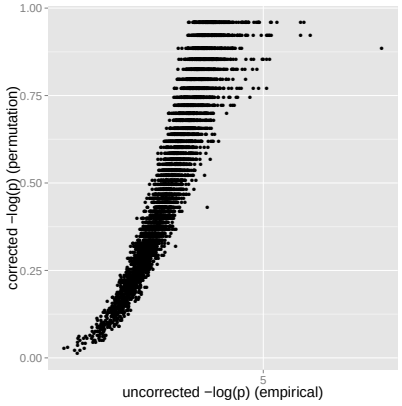
(e) QQplot: x: γ_A y: Uniform



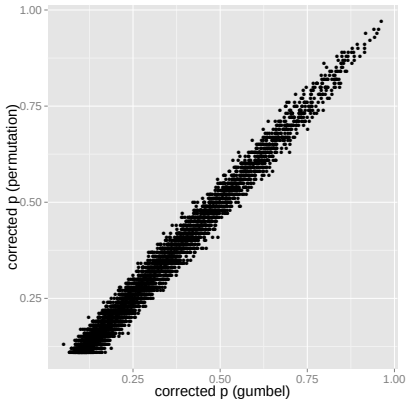
(f) QQplot: x: π_A y: Uniform



(g) Scatterplot: x: λ_A (log) y: γ_A (log)

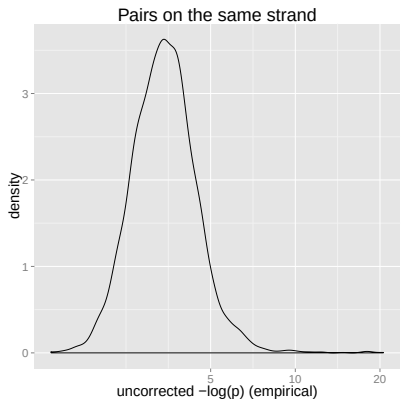


(h) Scatterplot: x: λ_A (log) y: γ_A (log)

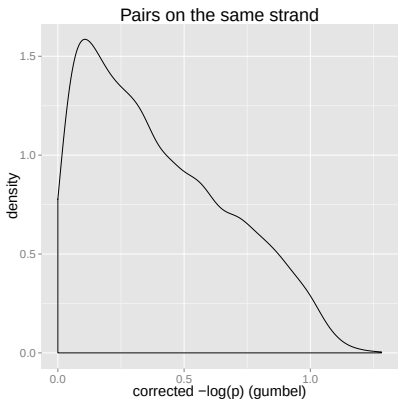


(i) Scatterplot: x: γ_A y: π_A

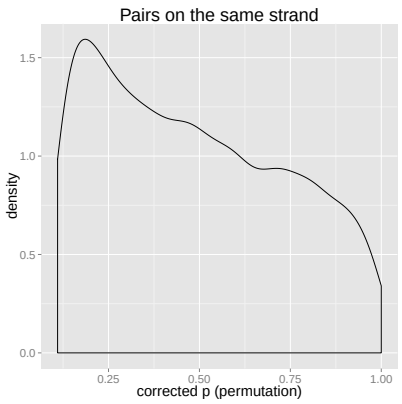
Figure C.5: Unpaired reads + reads with large insert size



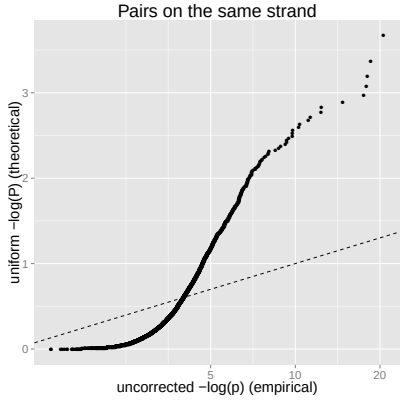
(a) Density of λ_A



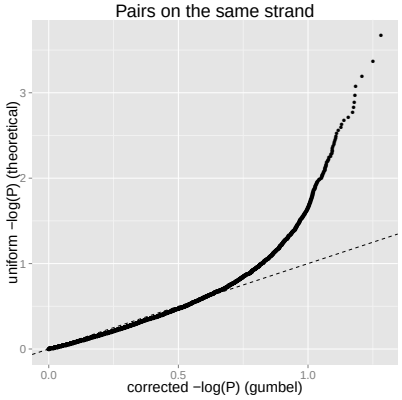
(b) Density of γ_A



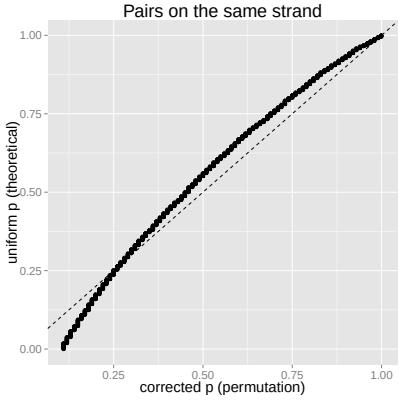
(c) Density of π_A



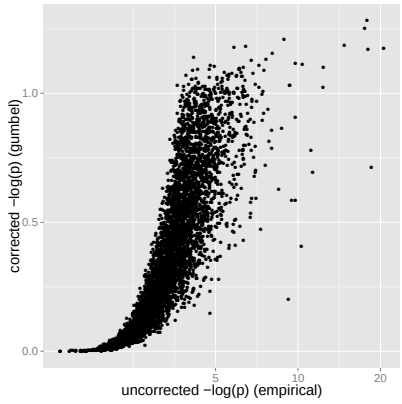
(d) QQplot: x: λ_A, y: Uniform



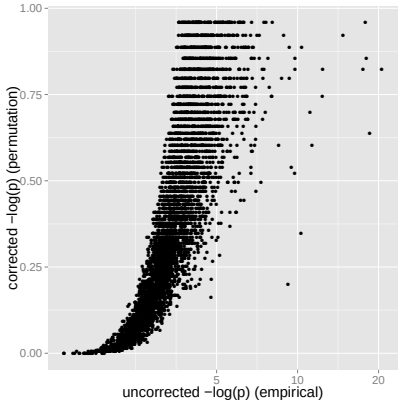
(e) QQplot: x: γ_A y: Uniform



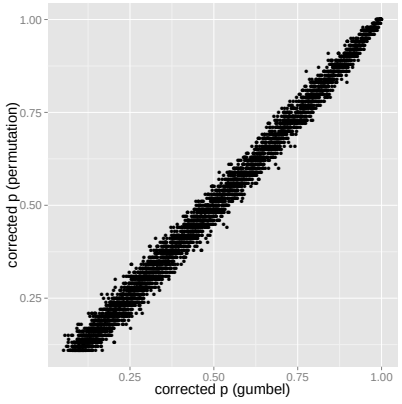
(f) QQplot: x: π_A y: Uniform



(g) Scatterplot: x: λ_A (log) y: γ_A (log)



(h) Scatterplot: x: λ_A (log) y: γ_A (log)



(i) Scatterplot: x: γ_A y: π_A

Figure C.6: Read pairs on the same strand

Appendix D

Circos plots of Structural Variation relative to col-0 (TAIR10) in the 18 founders

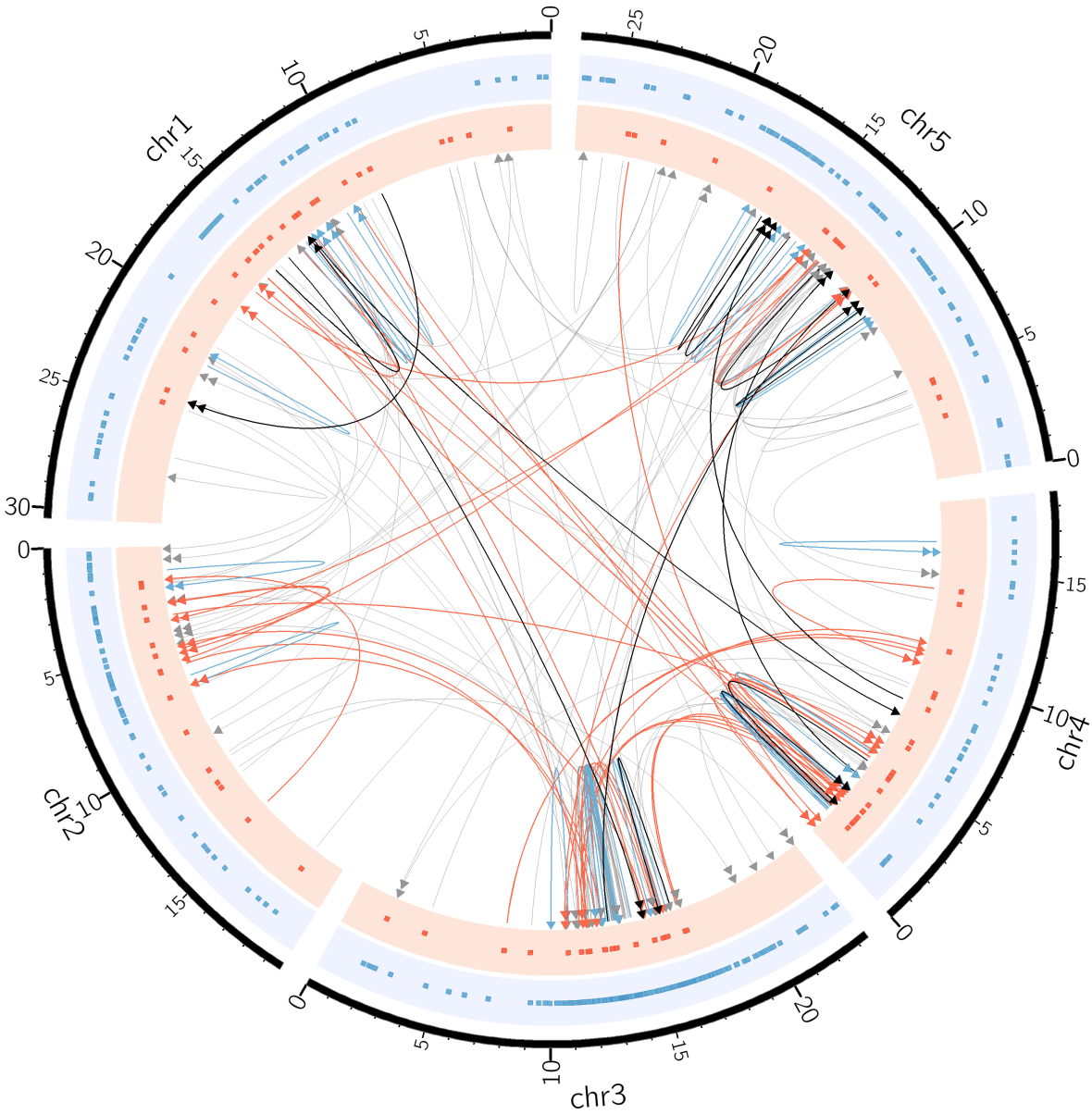


Figure D.1: Circos: bur-0 SVs

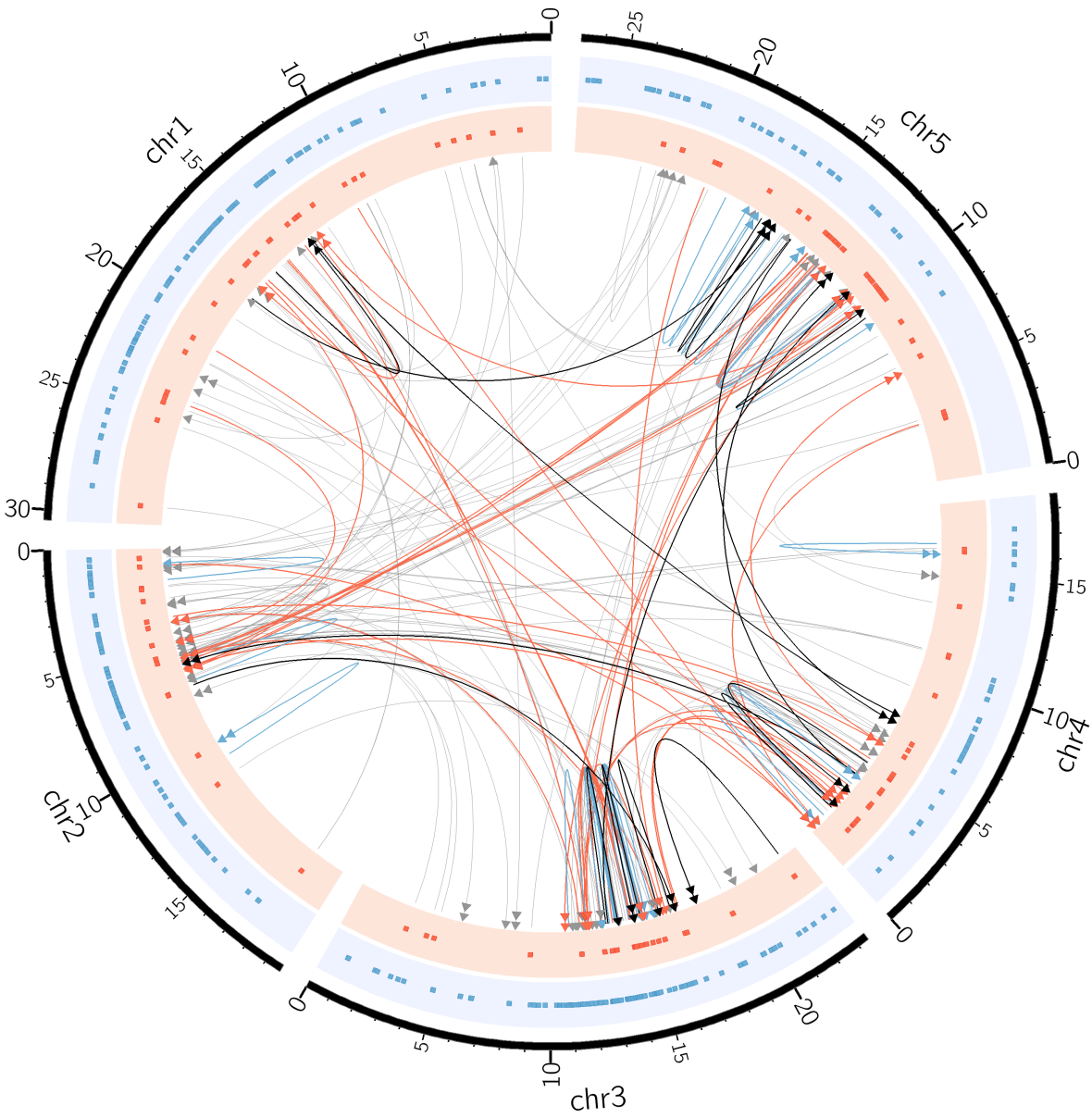


Figure D.2: Circos: can-0 SVs

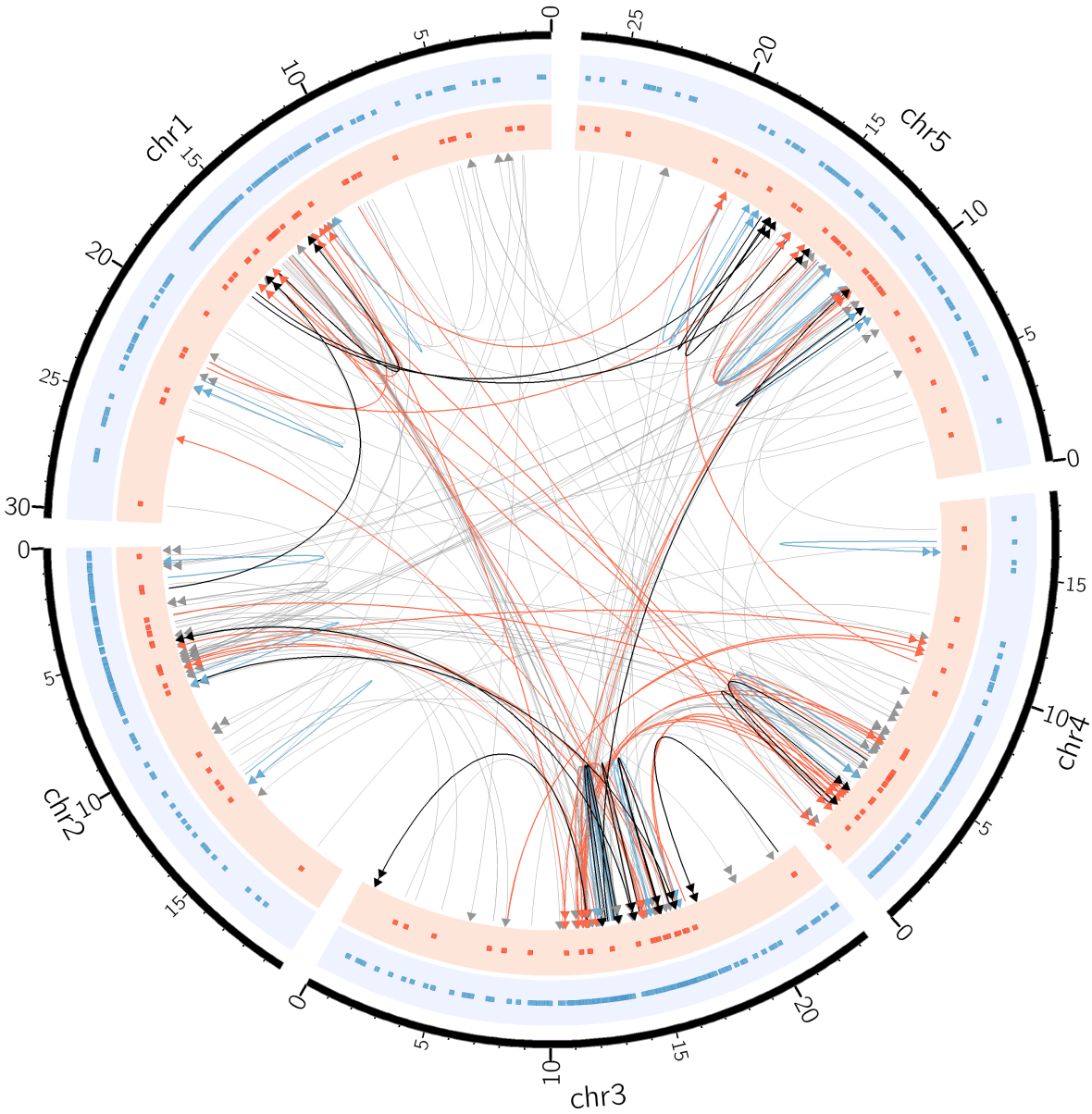


Figure D.3: Circos: ct-1 SVs

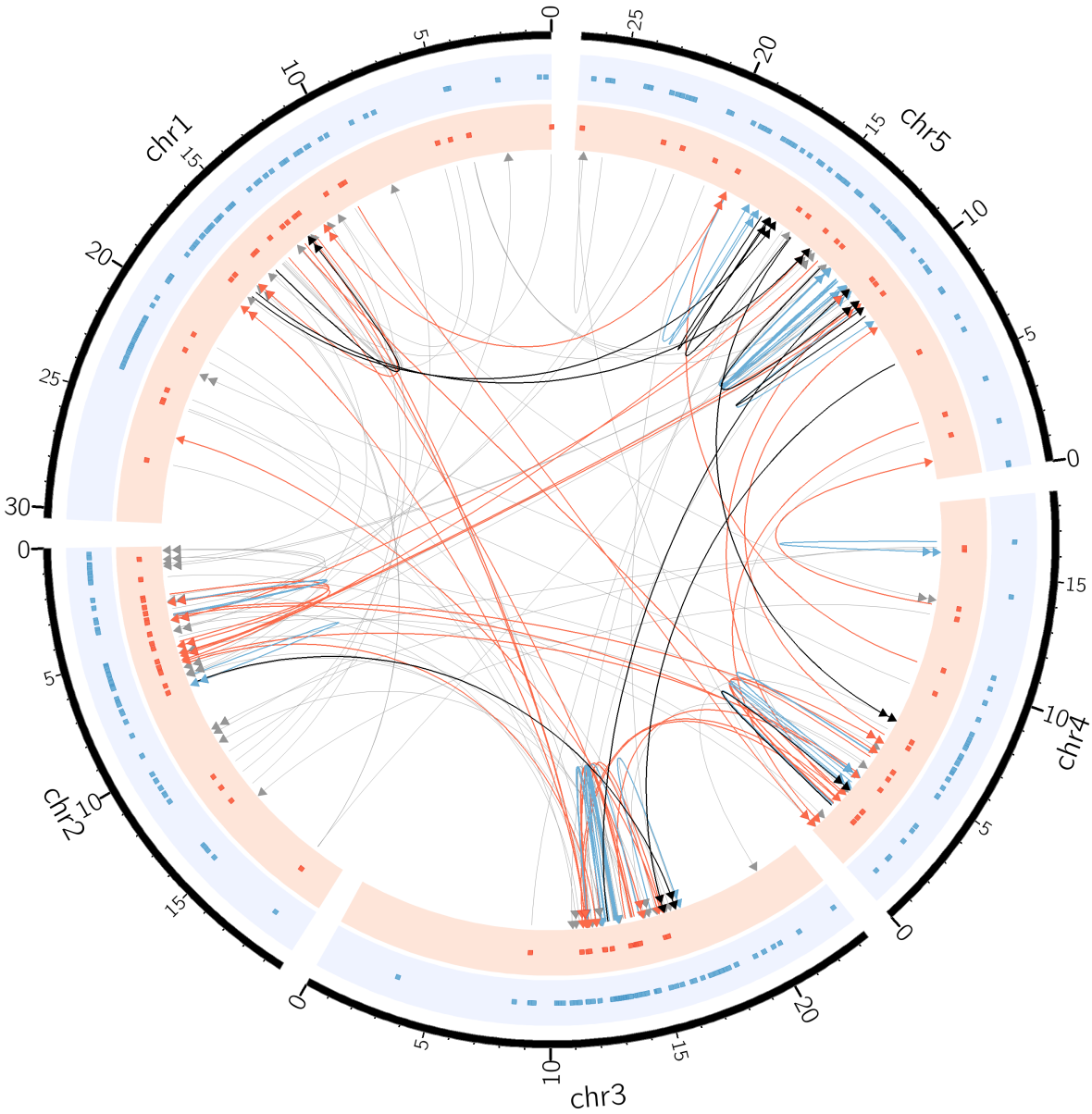


Figure D.4: Circos: edi-0 SVs

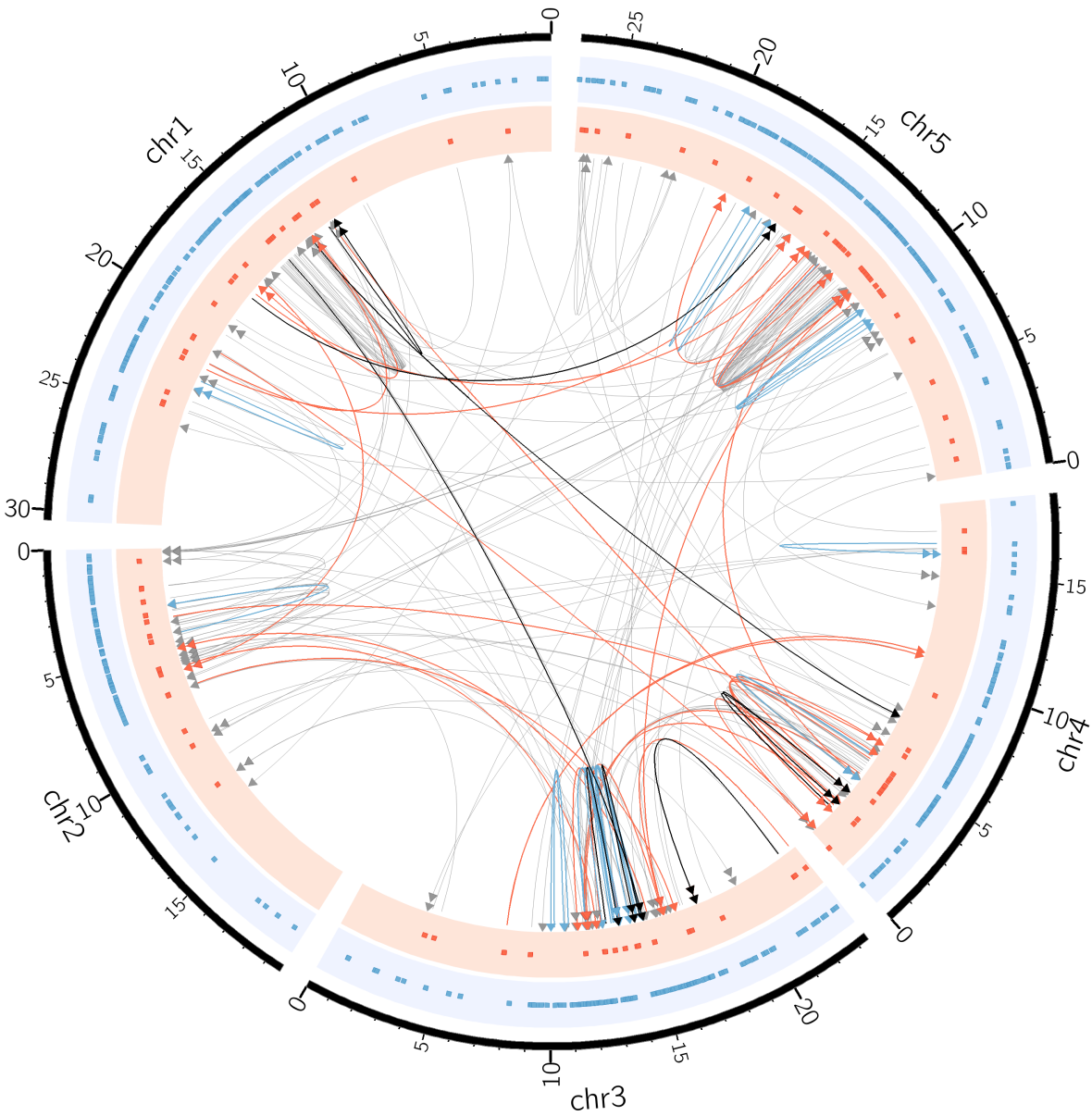


Figure D.5: Circos: hi-0 SVs

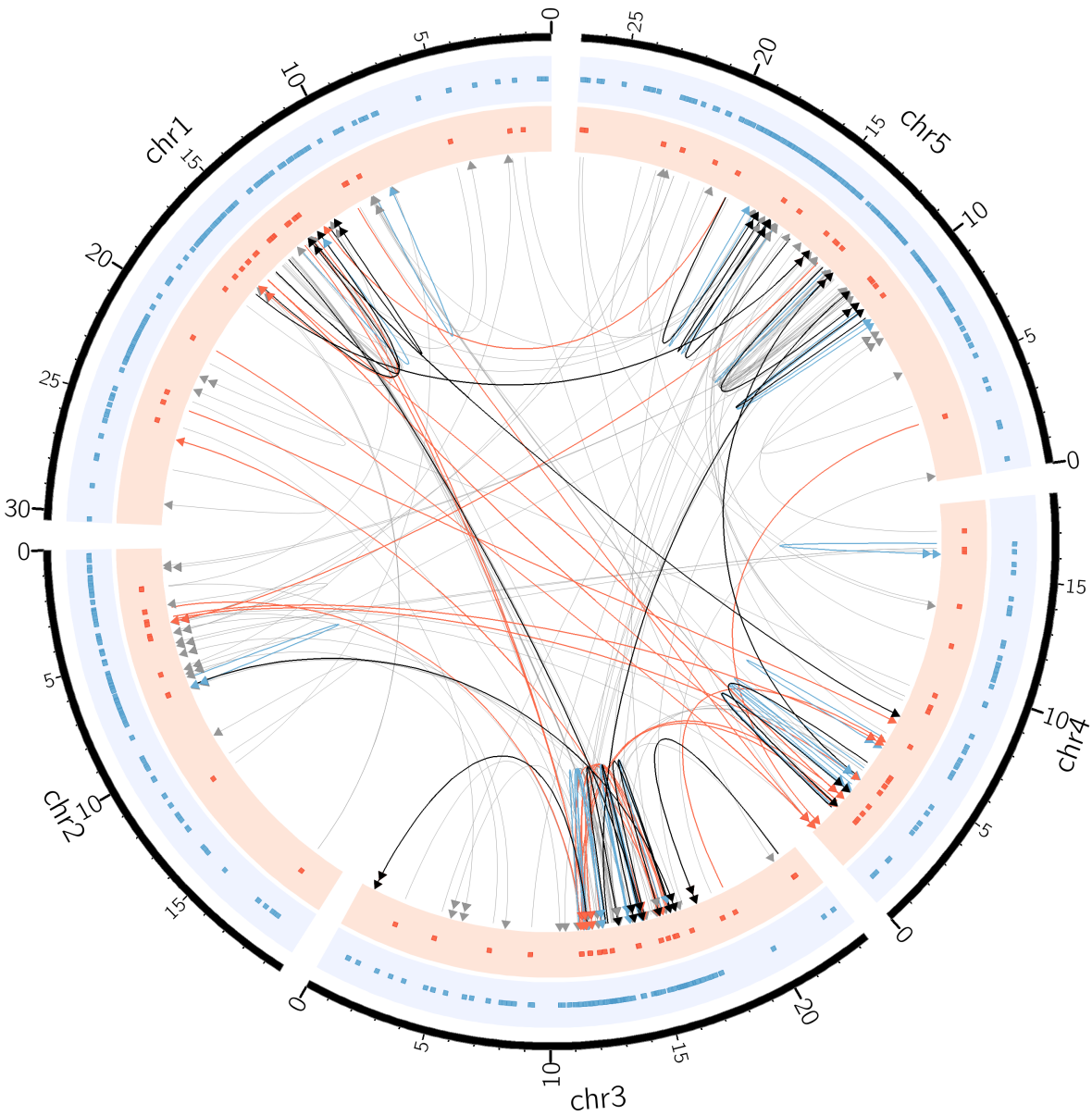


Figure D.6: Circos: kn-0 SVs

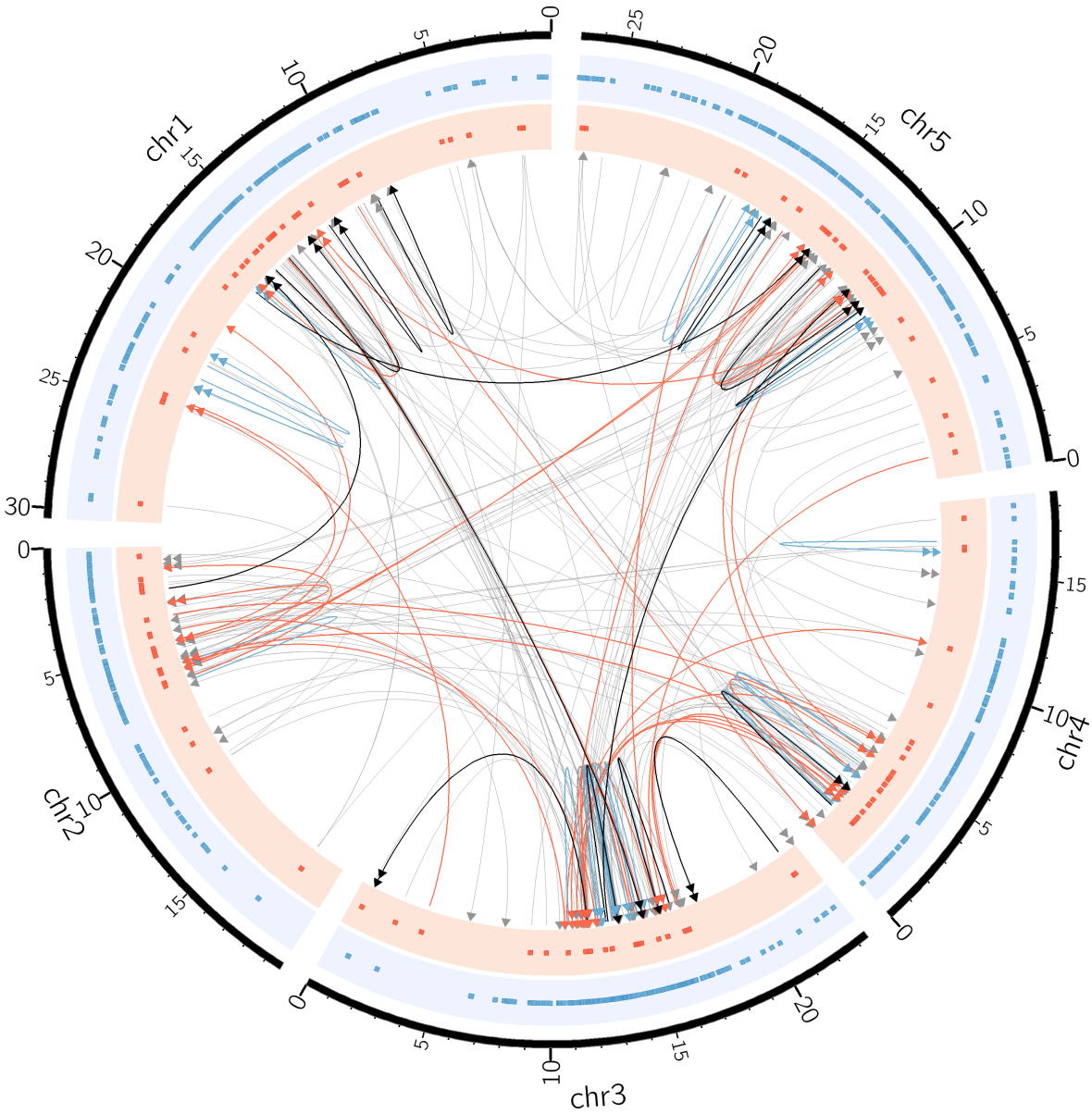


Figure D.7: Circos: ler-0 SVs

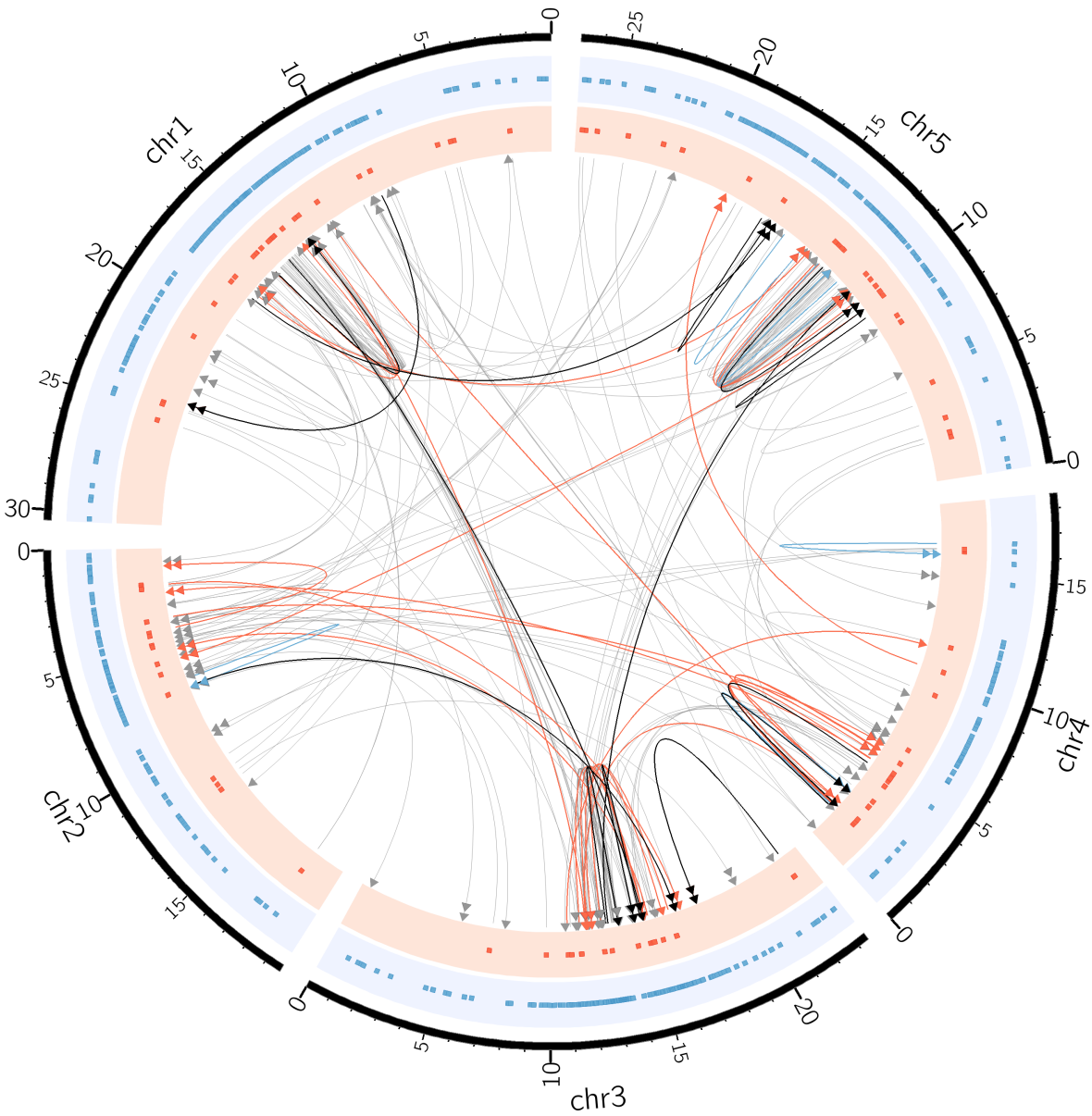


Figure D.8: Circos: mt-0 SVs

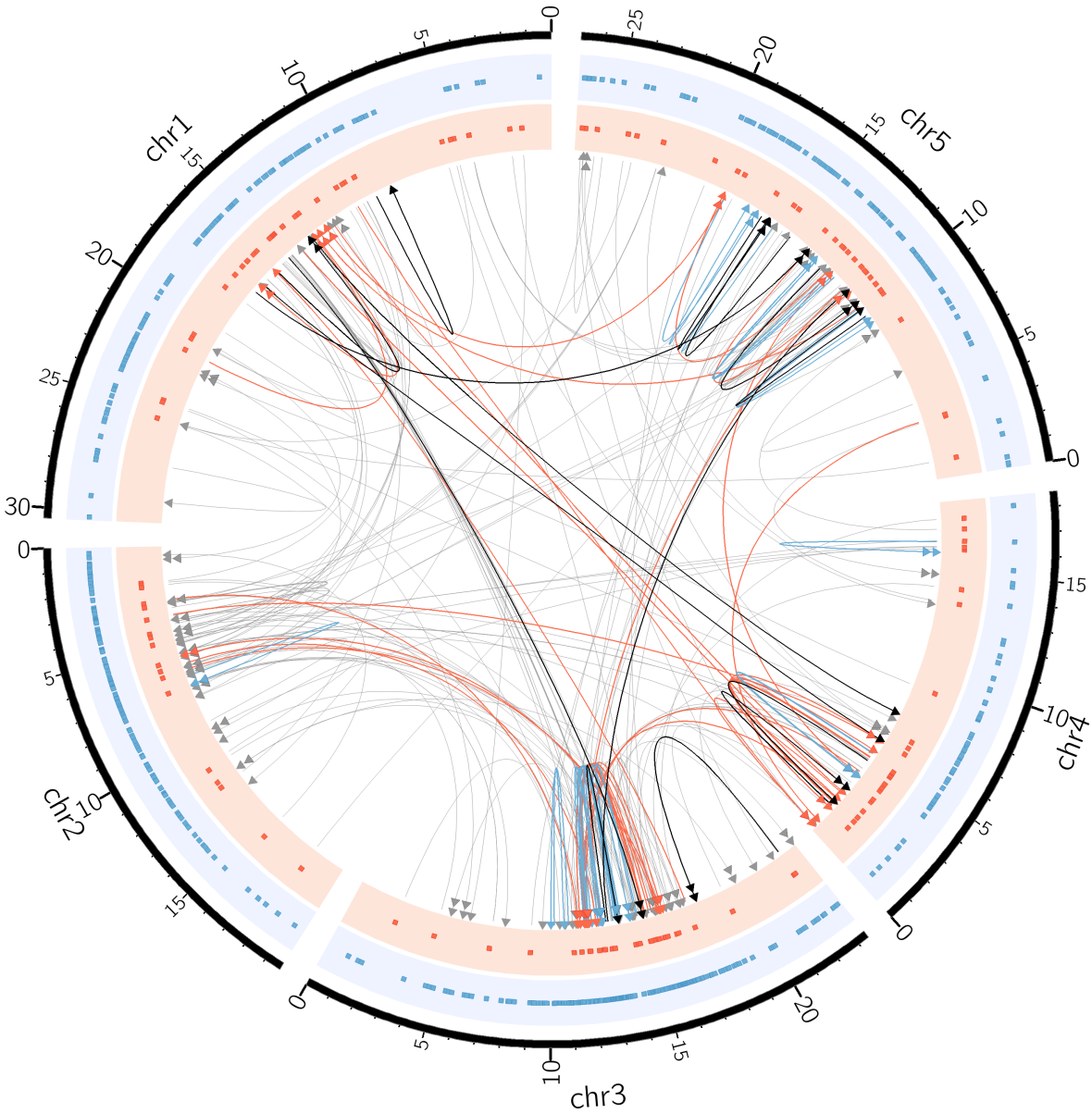


Figure D.9: Circos: no-0 SVs

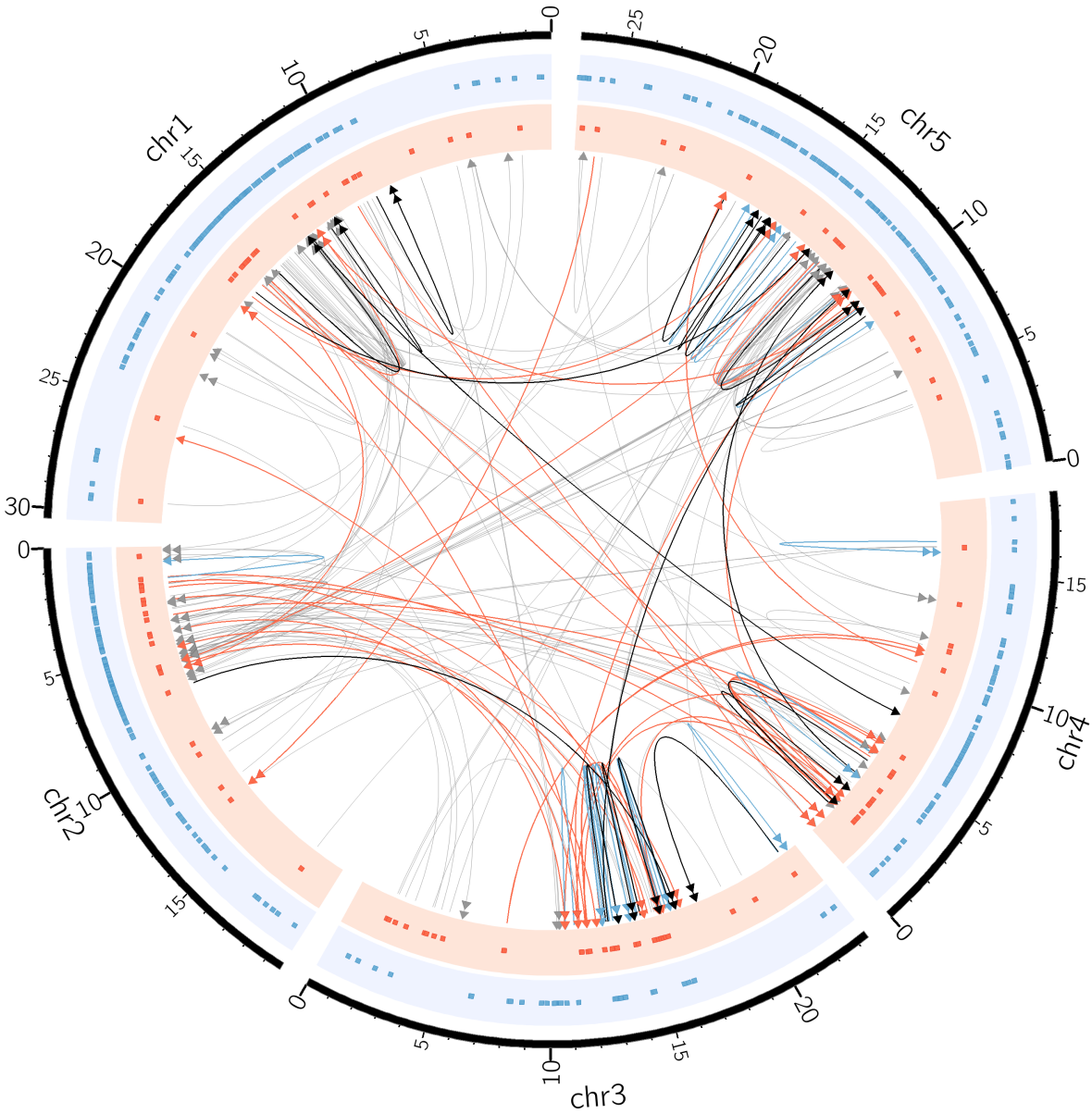


Figure D.10: Circos: oy-0 SVs

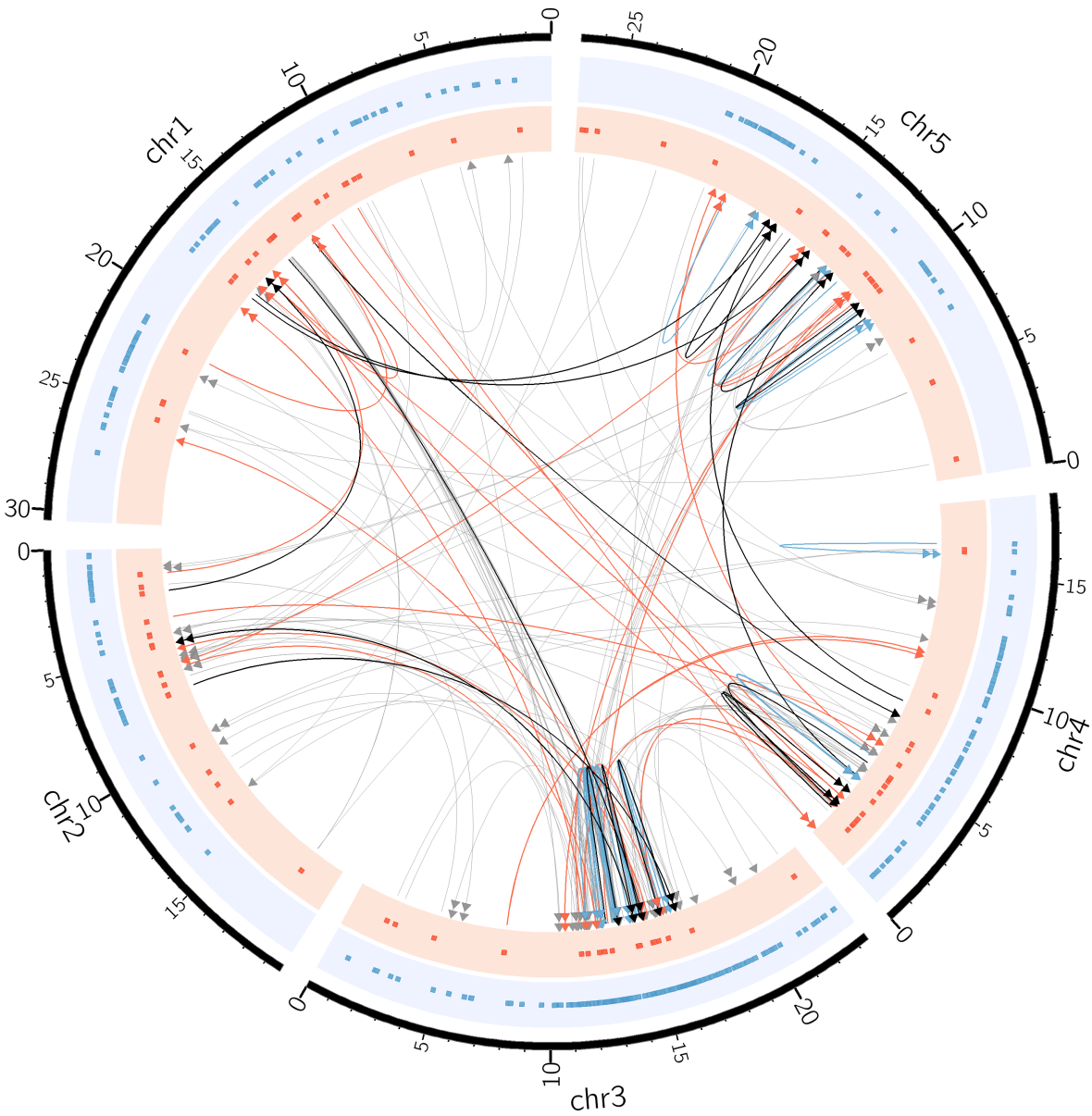


Figure D.11: Circos: po-0 SVs

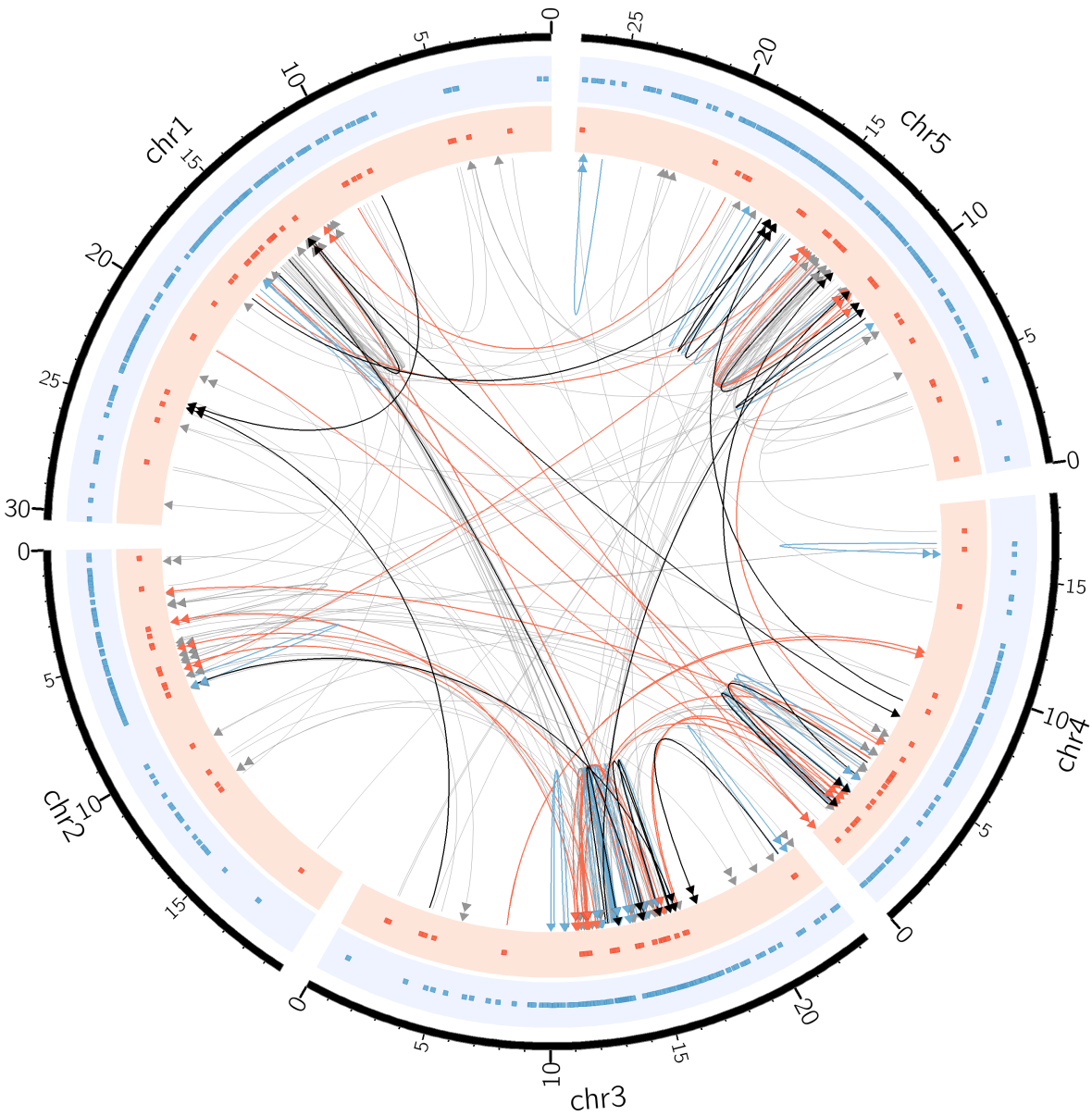


Figure D.12: Circos: rsch-4 SVs

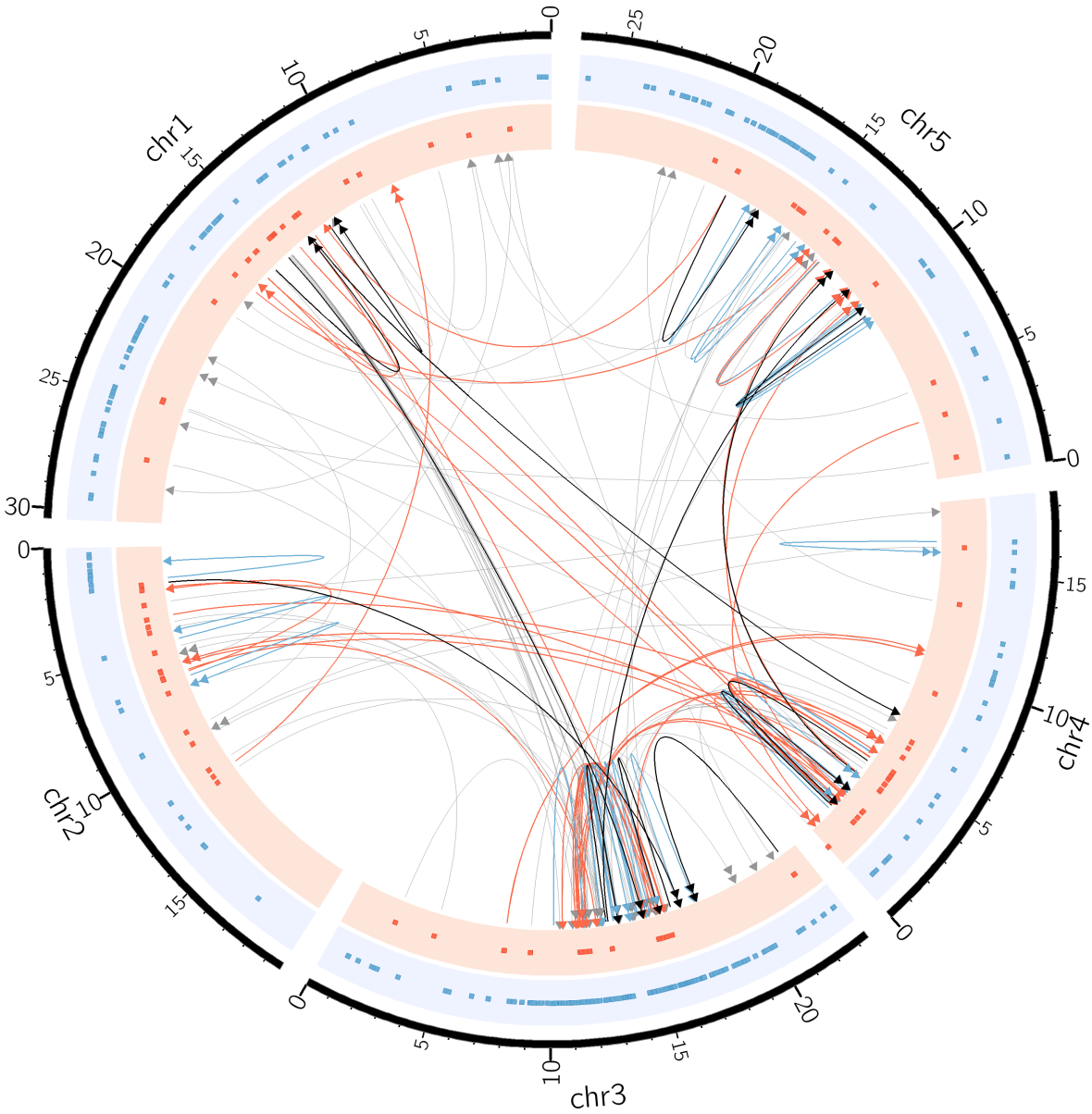


Figure D.13: Circos: sf-2 SVs

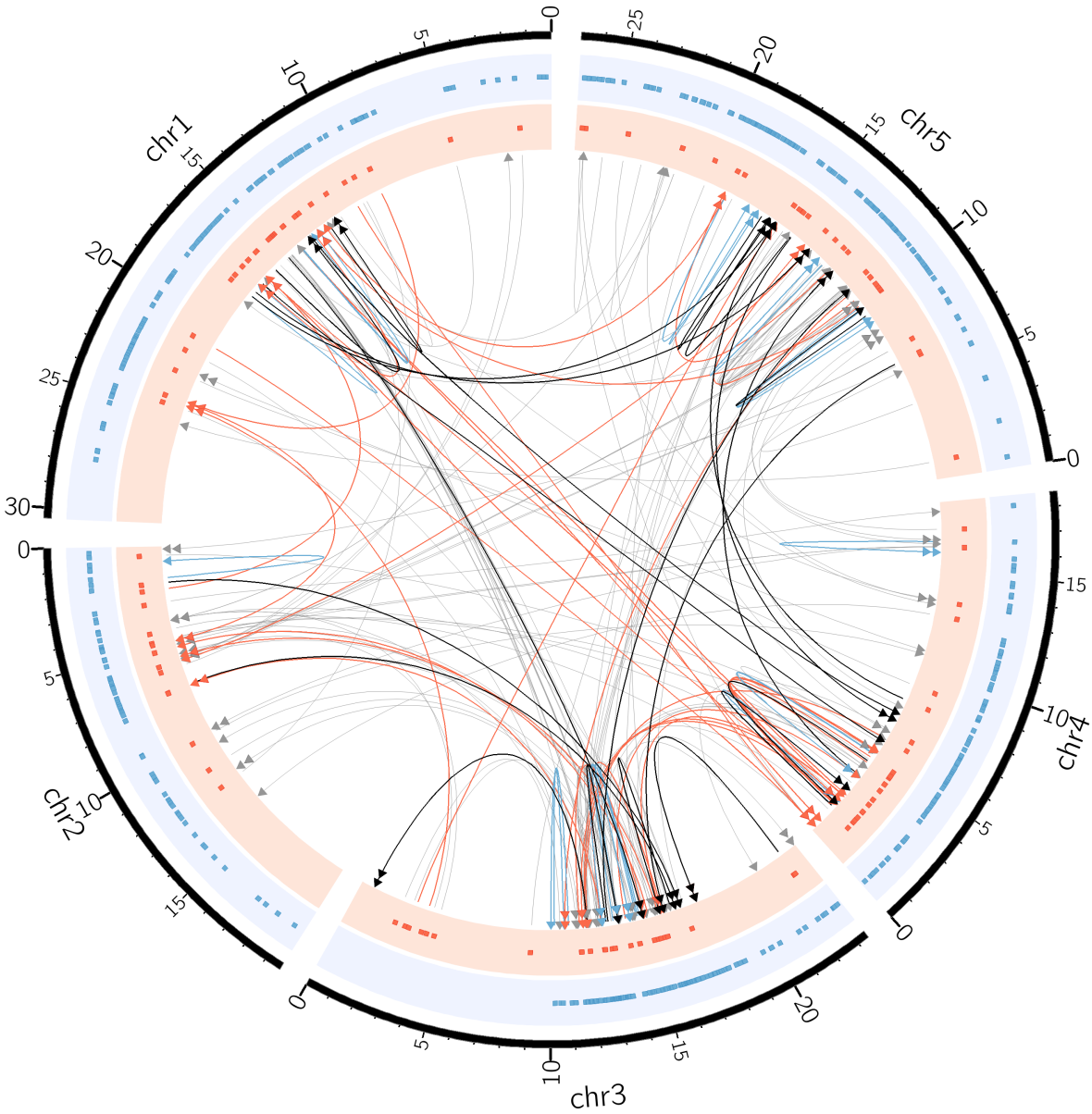


Figure D.14: Circos: tsu-0 SVs

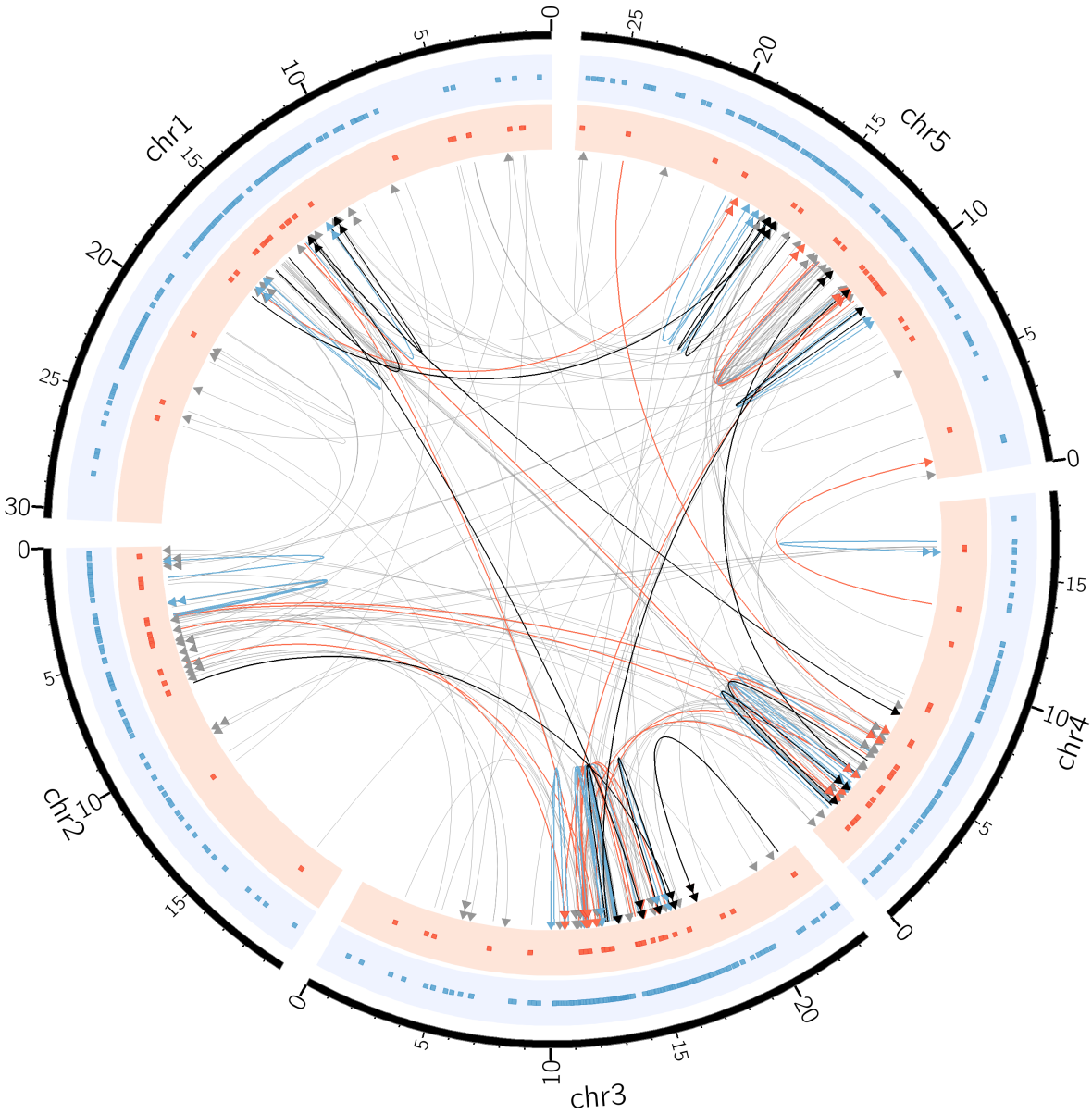


Figure D.15: Circos: wil-2 SVs

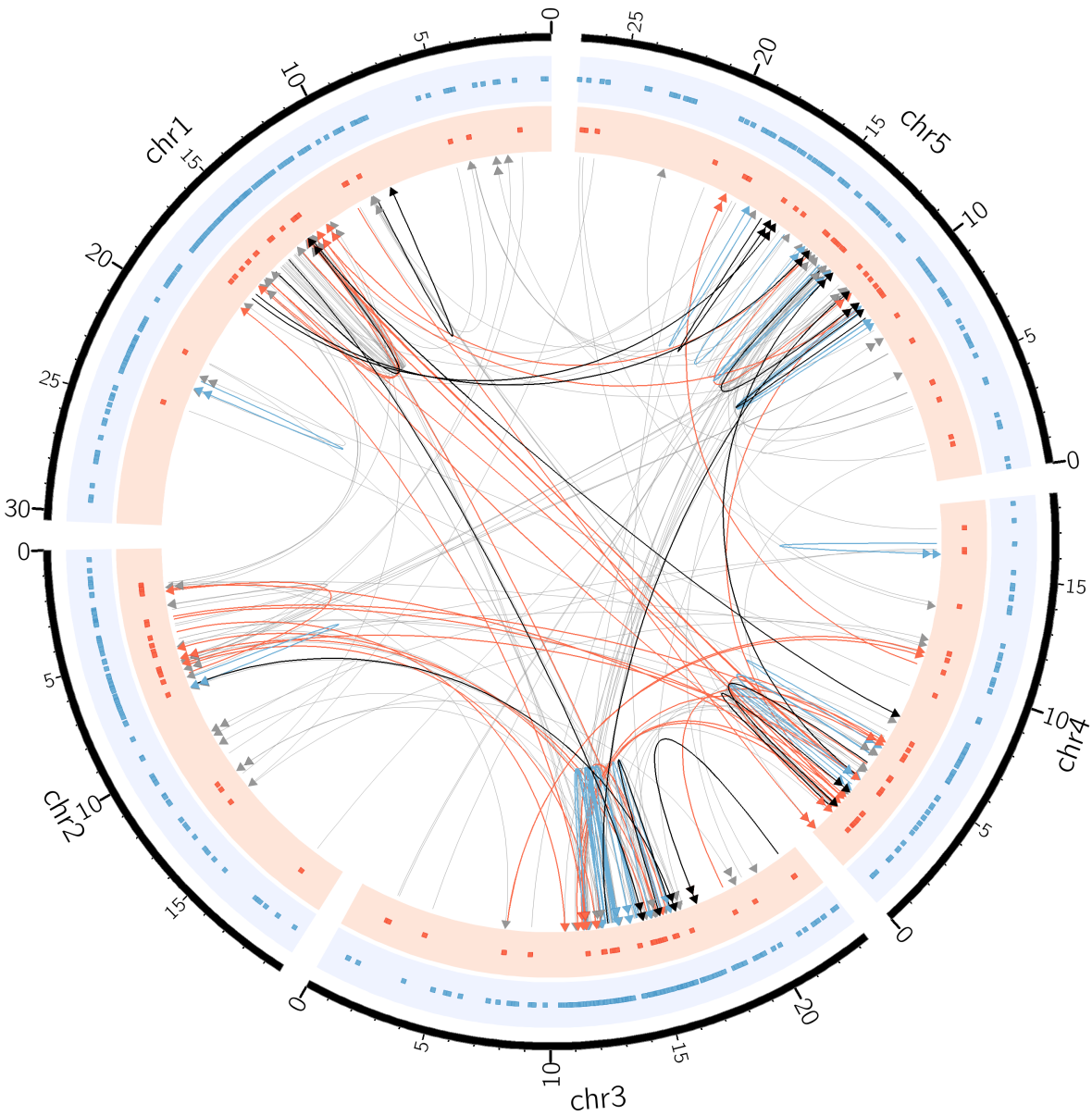


Figure D.16: Circos: wu-0 SVs

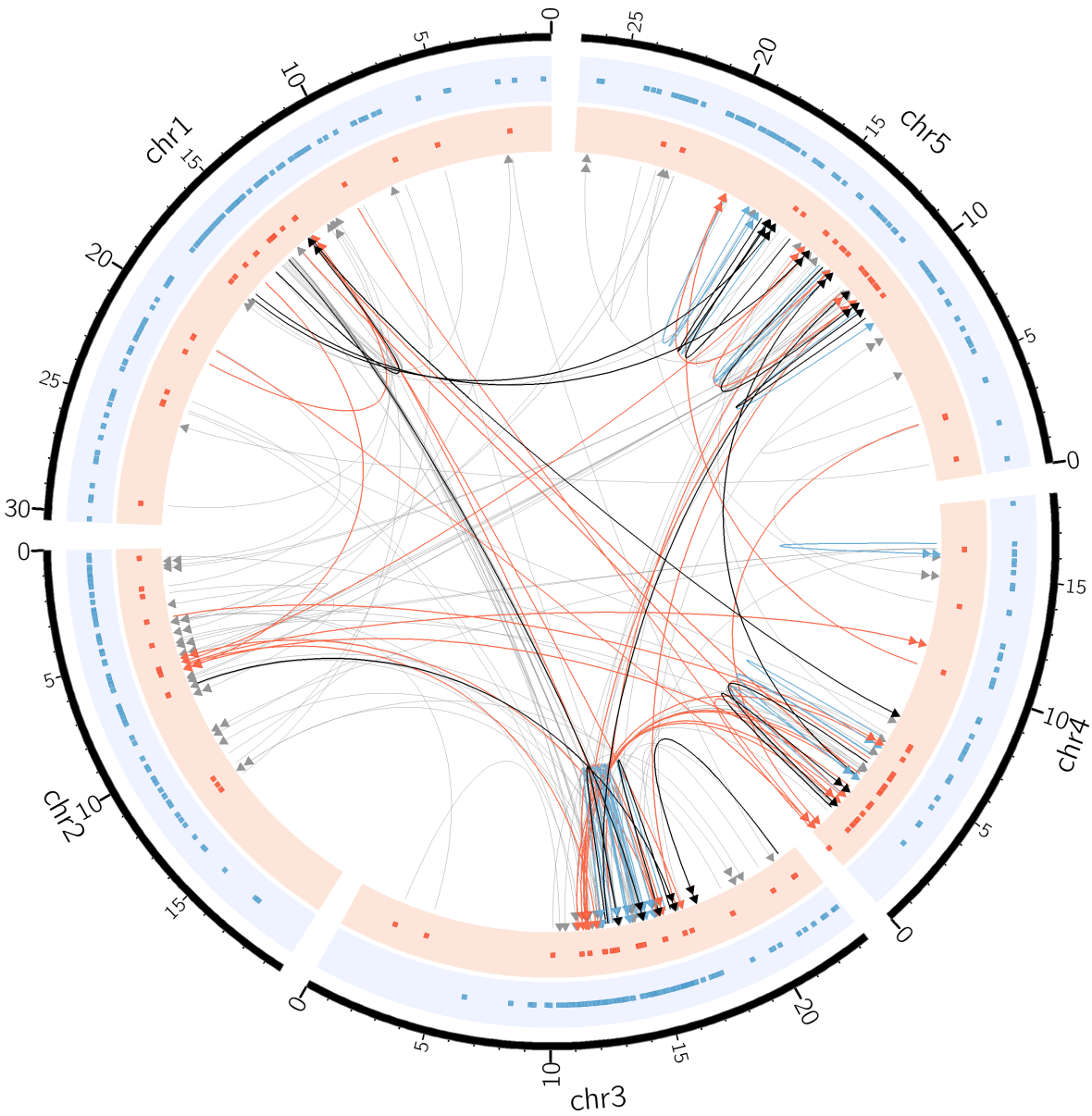


Figure D.17: Circos: ws-0 SVs

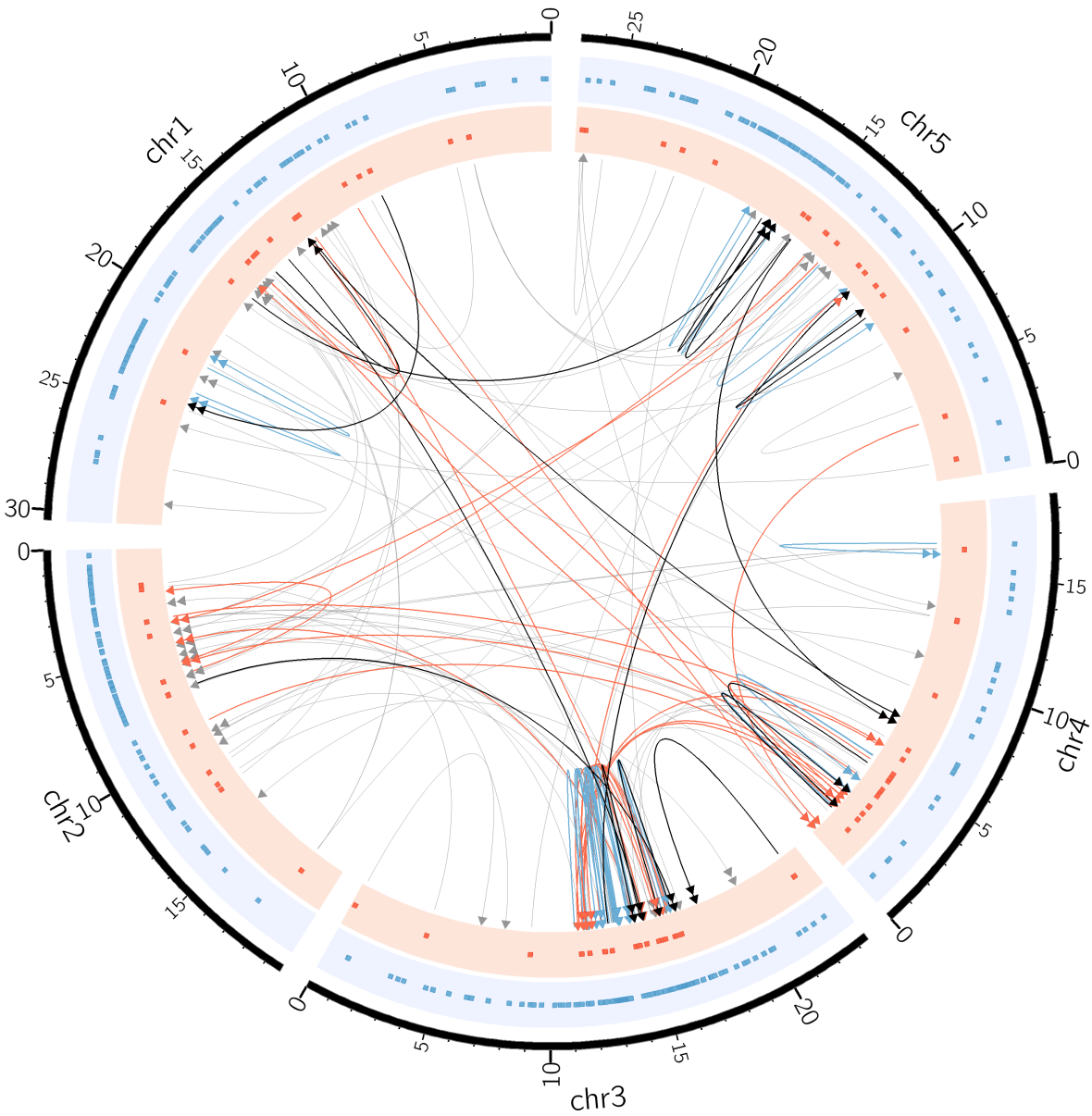


Figure D.18: Circos: zu-0 SVs

Appendix E

PCR results for SV breakpoint validation

N	type	Forward	Reverse	bur-0	edi-0	ct-1	col-0	can-0	hi-0	kn-0	ler-0	mt-0	no-0	oy-0	po-0	rsch-4	sf-2	tsu-0	wil-2	ws-0	wu-0	zu-0
1	1	3_16193837F	3_16051029R	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1	1	3_16056244F	3_16097096R	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1	1	3_16091476F	3_16045728R	1	0	1	1	1	1	1	1	0	0	1	0	1	1	1	1	1	1	1
1	1	3_16092006F	3_16221181R	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0
1	2	3_16193837F	3_16194636R	1	1	0	1	1	0	0	1	1	1	0	1	1	1	1	1	0	1	1
2	1	5_13625861F	1_17366573R	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	1	0	0
2	1	5_13625861F_INV	1_17366573R_INV	0	1	1	0	0	0	1	1	0	1	0	1	0	0	1	0	1	1	0
2	2	5_13625861F_INV	5_13626302R	0	1	1	1	1	0	0	0	0	1	1	1	1	1	0	1	1	0	1
2	2	1_17366003F	1_17366573R_INV	1	1	1	1	1	0	0	1	1	1	0	1	1	1	1	1	1	1	1
3	1	3_13172116F	3_14474741R	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
3	1	3_14474741R	3_13172116F	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
4	1	4_2586165F	2_2184567R	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
5	1	3_11973019F	3_904402R	0	0	1	0	0	0	1	1	0	0	0	0	0	0	1	0	0	0	0
5	1	3_904402R_INV	3_11973019F_INV	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
5	2	3_903246F	3_904402R_INV	1	1	1	1	1	0	0	0	1	0	0	1	1	0	0	1	1	1	0
6	1	5_12686938F	5_10272973R	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0
6	1	5_12686938F_INV	5_10272973R_INV	1	1	0	0	0	0	0	1	1	1	1	1	1	0	0	0	1	1	0
6	1	5_12650905F	5_10076009R	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
6	2	4_1611938F	4_1612571R	1	1	0	1	1	0	0	1	1	1	1	1	1	1	1	1	0	0	1
7	1	3_16709398F	2_7171492R	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
7	1	3_16709399F	2_7171492R_INV	0	1	1	1	1	0	1	0	1	0	1	1	1	0	1	1	1	1	1
8	1	1_9071525F	1_24214408R	1	0	0	0	0	0	0	0	1	0	0	0	1	0	0	0	0	0	1
8	1	1_9071525F	1_24215441R	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
9	1	1_12226078F	1_11592551R	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
9	1	1_11592551R_INV	1_12226078F_INV	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
9	2	1_12226078F_INV	1_12227492R	1	1	1	1	1	0	0	0	1	1	0	1	1	0	0	0	1	1	1
9	2	1_11591181F	1_11592551R	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
10	1	5_14838471F	5_15937125R_INV	1	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0
10	1	5_14838452F	5_15937125R	1	1	1	1	1	0	1	0	0	1	0	1	1	0	1	1	1	0	1
10	2	5_14838452F	5_14839317R	1	1	1	1	1	0	0	1	1	1	0	1	0	0	1	1	1	1	1
10	2	5_15936544F	5_15937125R	1	1	1	1	1	0	1	0	1	1	0	1	1	1	1	1	1	1	1
11	1	3_14320859F	3_14748377R	0	0	1	1	1	1	1	0	1	0	1	1	0	0	0	0	0	0	1
12	1	4_1612571F	4_2782483R	1	1	1	1	1	0	1	1	1	1	1	0	1	1	1	1	1	1	1
12	1	4_1612532F	4_2781917R	1	1	1	1	1	0	1	1	1	1	0	0	1	1	1	1	1	1	1
12	1	4_1612532F	4_2782548R	1	1	1	1	1	0	1	0	0	1	1	1	1	1	1	0	1	0	1
12	1	4_1612532F	4_2781335R	0	0	1	0	1	0	0	0	0	1	1	0	1	1	0	0	0	0	1
12	1	4_1612037F	4_2781917R	1	1	1	0	1	1	0	1	0	1	0	1	1	0	1	1	1	1	1
12	1	4_2781917R	4_1612037F	1	1	1	1	1	0	0	1	1	1	0	1	1	0	1	1	1	1	1

12	1	4_2781917R	4_1612532F	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1
12	1	4_2781335R	4_1612532F	0	1	1	0	1	0	0	0	0	0	0	0	1	0	0	0	1	0	1
12	1	4_2580475F	4_1526396R	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
12	2	4_2781917R	4_2783051R	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
12	2	4_2780506F	4_2781917R_INV	1	1	1	1	1	0	0	0	1	1	0	1	1	0	0	0	1	0	1
12	2	4_2781917R	4_2782548R_INV	1	1	1	1	1	0	0	1	1	1	0	1	1	1	1	1	1	1	1
13	1	5_10209174F	5_14200929R	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
14	1	5_6298188F	3_16049768R	0	1	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0
15	1	4_3772295F	3_12631061R	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
16	1	4_1853989F	4_4449454R	0	0	0	0	0	0	1	0	0	0	0	0	0	1	0	0	0	1	0
16	1	4_1784891F	4_4497930R	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
16	1	4_4499865F	4_1821194R	1	1	1	1	1	0	0	1	0	1	1	1	1	0	1	1	1	1	1
16	1	4_4497930R_INV	4_1784891F_INV	1	1	1	1	1	0	0	0	1	1	0	1	0	0	0	0	1	0	1
16	2	4_1784891F_INV	4_1786363R	1	1	1	1	1	0	1	1	1	1	0	1	1	0	1	1	1	0	1
17	1	1_16757998F	4_6194184R	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1	0	0	0	0
18	1	3_12837849F	3_13224910R	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
19	1	4_8167657F	5_15952489R	0	0	0	0	0	0	0	0	0	0	0	1	0	0	1	0	0	0	0
19	1	4_8166395F	5_15952747R	1	0	0	1	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0
20	1	3_16193837F	3_16051336R	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
21	1	3_17354752F	3_22556429R	0	0	1	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	1
21	1	3_22562633F	3_17358916R	0	0	0	0	1	1	1	1	1	1	1	0	1	1	1	1	1	1	0
21	1	3_17359034F	3_22563595R	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0
21	2	3_17359034F	3_17359469R	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
21	2	3_22562635F	3_22563028R	0	0	1	1	0	0	1	0	1	1	1	1	0	1	0	0	0	0	0
21	2	3_22562635F	3_22563045R	0	0	0	1	0	0	1	0	0	1	1	0	0	1	0	0	0	0	0
21	2	3_22227690F	3_22228732R	1	1	1	1	1	0	1	1	1	1	1	1	1	0	1	1	1	1	1
22	1	2_9080636F	4_2605928R	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
23	1	1_14918256F	3_14591539R	1	0	0	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1
23	1	3_14589209F	1_14920456R	1	0	0	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	0
24	1	5_17121950F	5_18835849R	0	0	0	0	0	0	1	0	0	0	1	0	0	1	0	0	0	0	0
24	2	5_18834369F	5_18835849R_INV	1	1	1	1	1	0	0	0	1	0	0	1	1	0	1	1	1	0	1
24	1	5_18835849R_INV	5_17121950F_INV	1	1	1	0	1	0	1	0	0	1	0	1	1	1	1	1	1	1	1
25	1	5_9231155F	5_9228216R	0	1	0	0	0	0	0	1	1	0	1	0	0	0	1	0	1	0	0
26	1	3 _13038437F	5_10918979R	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1
27	1	1 _17695673F	5_15871143R	0	1	0	1	1	0	0	0	1	0	0	0	1	0	1	1	1	0	1
27	1	5_15871143R	1_17695673F	0	1	1	1	1	1	0	0	1	0	0	1	1	0	0	1	1	1	1
28	1	3_16769437F	2_1840188R	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0
29	1	3_13183503F	3_12868316R	0	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	0
30	1	2_6123558F	4_3651495R	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0

31	1	1_13350338F	4_7609008R	0	0	0	0	0	0	1	0	0	0	0	1	0	0	1	0	0	0	0
31	1	4_7611256F	1_13351653R	1	0	0	0	1	1	1	0	0	1	1	1	1	1	1	1	1	1	1
32	1	1_15920942F	1_13658231R	1	1	1	1	1	0	1	1	1	1	1	0	1	1	1	1	1	1	1
33	1	1_16115474F	2_2089357R	0	0	1	1	0	0	0	1	0	0	0	1	0	0	0	0	0	0	0
34	1	5_12223580F	4_3573677R	1	0	0	0	1	0	1	0	0	0	1	1	1	1	1	1	1	1	0
35	1	4_6763117F	5_14776120R	0	1	0	1	1	0	0	0	0	0	0	0	0	0	1	0	0	0	1
36	1	3_15900790F	3_14841574R	0	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0	0
37	1	3_14255170F	2_4888798R	0	0	1	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0
38	1	1_16965574F	1_12645883R	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
39	2	5_16401082F	5_16620111R	1	1	1	1	1	0	1	1	1	1	1	0	1	0	1	1	1	1	1
40	1	3_3643868F	1_23974142R	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0
40	1	1_23974142R_INV	3_3643868F_INV	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
40	2	1_23972663F	1_23974142R_INV	1	1	1	1	1	0	1	0	1	1	0	1	1	0	1	1	1	1	1
41	1	4_2804713F	4_1772122R	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0	0	0
42	1	1_9473516F	1_8419370R	0	0	0	0	0	0	0	1	0	1	0	0	0	0	0	0	0	0	0
42	1	1_9473516F_INV	1_8419370R_INV	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	1	0
43	1	3_14474741R	3_13172091R	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
44	1	2_4790156F	3_15005840R	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
45	1	2_5605866F	2_2894614R	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Table E.1: Results of PCR validating SV breakpoints. Each row corresponds to a primer pair used to confirm a single SV breakpoint. Columns: **N**: SV QTL id tested (source and sink for each SV QTL tested can be seen in Table 4.3), **type**: type of experiments - type 1 experiments are expected to amplify only in the founders carrying an SV, type 2 experiments should amplify in the reference (col-0) and to genomes not carrying an SV (the col-0 genome used here was contaminated so not every type 2 experiment has worked - however, in most cases it has worked in at least three other genomes, which act as reference surrogates, **Forward, Reverse**): primer ids used for the experiment, their name reveals genomic position, thus 2_5605866F means that the primer sequence starts at chromosome 2 at position $5,605,866$ at the reference and that it is a forward primer, hence having forward orientation in the reference. The remaining columns correspond to each of the 19 founders DNA: value 1 means that the PCR amplified and produced a product, while 0 shows no amplification. The primer sequences used are in Table E.2.

3_3643868F	TGGATTCTGGGGATTCTTTG
3_13038437F	AAATGTTGACGGTTACTTTTT
5_12686938F	GATGAATCGTTGTCGGGATC
4_1612571F	AAACCGTCCTTCTTTCAGCC
2_9080636F	TTTGGTAAGGATAAAAATAAAAATACCA
4_3772295F	TGAAAACTTTGGATTCTTGGC
1_9473516F	CCTTATATGAATTACTAAATTGA
1_16115474F	TCATGCCAAGAAGAAGGCTT
1_15920942F	TTGGGTGGGATTAATCTTGC
5_13625861F	TCGAAATAGCGTGTGTGGAA
4_2586165F	CATCGTCCCTGTGGTCTTTT
5_14838471F	AGCAGTACGGTTCAGCCAAT
1_12226078F	AGCACCCTCAATGGTTAAGAAC
5_10209174F	ACACCAACATCTCCAAAGGC
3_16056244F	GAAATTCAGAATATTTTCATTGAA
3_17354752F	TGGCCTTGACTTCAACATCA
4_8167657F	GCCTCGCACTAAGGTTGAAG
3_13183503F	TGTGGTAGTTTTGGACTTTTGG
1_17695673F	GAGTGGCCCTGTTGGTCTAG
1_13350338F	CGGAAATGTCCCTGAACTACT
3_16091476F	ACAGGATACAAACAACAACAA
4_2580475F	TGGAAATACTGCTCTGCACTTG
4_1853989F	ACCATCCCACGTCCTCAGCT
5_17121950F	CTAGATCGCGAGTGGGCTAC
1_16965574F	TACCTCATCATTCATGCCCA
4_7611256F	TGAATCTATGACATTTGCCTAATC
4_6763117F	CGCTTGCATCTCCATCAGAT
5_12223580F	GATTTTGAAGAATTCGCCGA
1_14918256F	TTTTGGCAGCAATTTTCCCT
3_11973019F	CAATTCATAGTCATTTTGTGCA
3_16193837F	AATCTCCAACTTTTCAGGGGA
3_22562633F	TTCAGAAAACAAGTAGAAAGAGTCCA
2_6123558F	ATTTCGAGGGCCGATAATTAAC
3_14589209F	ATATTGAAACCACCGCCTGA
3_12837849F	GTGCCATCAAAATCTTGCAA
5_12650905F	GAGATGGTTTACTCAAATGAAAAA
4_1784891F	GGAGAAGAATCAACGAAGCG
3_16195965F	TATAGACCGTTGACGAGTCT
5_6298188F	TTAAGTTATTCTCACGCATCGC
1_9473516F_INV	TCAATTTAGTAATTCATATAAGG
5_12686938F_INV	GATCCCGACAACGATTCATC
4_1612532F	TATAGGCCACTCTTGCCCAC
5_14838452F	CTGCCACAGAACGAGTCTGA
3_15900790F	TCAACATGACGTTCACCGTT
3_17359034F	TCCAGACATTTTTCCATGCA
2_4790156F	TCGACATGAGACTATGTTTGGG
3_16769437F	GCGGCCAAGATGATTATTGT
5_16401082F	AGCAGAGATGGAGCAAAGCT
2_5605866F	GCGGTTTAGGGGTGAATTTC
3_14320859F	GCCGTCTTTAAAAGCCCAAT
3_16709398F	AAACCATAATGGGACACCTGA
3_14255170F	TGCCCCTTCAATGTTAGGAG
4_2804713F	GTGTTTCGGGTAACGATGCT
3_13172116F	GTTGAGGTAGGGCGATTTCA

4_1612037F	GGTACACTTCGGCAACGATT
3_16092006F	CCCATTAGGAGAAATCGAGTCA
4_4499865F	CAAGAAAATTGCATGCGTGT
1_9071525F	TGCGATCATATGGGATGTTG
5_9231155F	TTATCAATAAAAATGAAACTCTCCG
3_16709399F	CAGGTGTCCCATTATGGTTTC
4_8166395F	AGAAACGTGGCAAACTAGCA
1_16757998F	TTGTACAATTGTTCTTACTCATGCA
3_22562635F	CAGAAAACAAGTAGAAAGAGTCCAG
3_22227690F	AGCCTTGTCAGGGGGTAGTT
4_2780506F	GGCTCGGTTTAGCTTCAGTG
4_1611938F	CGCAAAATCGACATGAGAAA
1_17366003F	ATTCCACCAATATGCAGNGTG
5_13625861F_INV	TTCCACACACGCTATTTCGA
3_903246F	GCATTAGGGCACGATTCACT
3_11973019F_INV	TGCACAAAATGACTATGAATTG
1_12226078F_INV	GTTCTTAACCATTGAGGATG
1_11591181F	TCTCACTTGCCAGTTGATGC
5_15936544F	TTGCCGAATTACGTAGTCCC
4_1784891F_INV	CGCTTCGTTGATTCTTCTCC
5_18834369F	CCAACAAGCTTGGTAGGTTG
5_17121950F_INV	GTAGCCCACTCGCGATCTAG
1_23972663F	AAAAGCTGAGAAGTCACTTGGG
4_1784891F_INV	CGCTTCGTTGATTCTTCTCC
3_3643868F_INV	CAAAGAATCCCCAGAATCCA
3_13172091R	ATGGTTCGGAGGAGGTTAGG
5_15871143R	TGTGGCATTGTCTTGCTCTC
3_14591539R	CCACGTGGATTACAACAATTT
5_15937125R_INV	TTGATAAATTGGATGTTCGCAT
4_2605928R	GGAAATTTGGAAAACTAACCTT
3_16051029R	TCAATACCGATCGTCGACAA
5_15952747R	CCCTCAGATTTGAGAAAAGCA
1_11592551R	TAGGCCCAAATTTGCAGTTC
1_13658231R	GTTCGATAACATTTTGTTTACTT
3_16097096R	ACAACAACAAACACCAACGGTA
1_24215441R	CTTGTTGAGGTGTTGTGCGT
2_4888798R	GCACAACTAACCGAGGTTGC
3_12631061R	TTCCTAATTACCTAATGATCATCA
5_15937125R	ATGCGAACATCCAATTTATCAA
3_13224910R	GGAAAAATCCATGTTACAGCAA
3_904402R	TCAACATGGAGATAACGAGCC
4_2782548R	GTCAGTCCGTTCCCAAAAAG
3_14841574R	CCTTTAGCGCTGGGAGAGAT
3_14474741R	TGCCCTTGACCAAGACTAGG
3_16045728R	GCCATGAAAGTCGTGTAAAGC
4_6194184R	TAATGTTTACCCACGCATCG
4_3573677R	CAGCTTTCATCGCACCTACC
5_9228216R	TCAGGGCATGACCCTTACTC
4_4449454R	GTAACAAGAGGGATGTCAATTCATT
3_16194636R	CGCAGCACTTCTCATTCTTG
2_1840188R	TCTCTCTCTTGCGACACGAA
4_1526396R	CTCTCCCTATCTTCTTCAGCC
5_16620111R	GAGCTTCAAGTCCGACCATC
3_14748377R	ACTATCTTGCTGGGTCGTCG
3_16051336R	TAAAATGATGCTTGGGGAGC

4_1772122R	TGTTCACTGCCAAAAACCAA
4_2781917R	AACCAGGTGAGGACAACGTC
5_10272973R	TTGCACATAAAAATGGGGGT
3_22563595R	TTATTCCACGGTACATCGCA
1_8419370R_INV	ATGCGGATGCGGTTACATAC
3_17358916R	GAAAGGTTTATAATGCGGTTCG
3_15005840R	TGTTAATGAAGATGCCCACTCA
4_4497930R	CCGTCATTGTAAAACGGGAA
3_12868316R	AAAAAGTGAACCATTGCCTAA
5_14776120R	AGAGGTTTAAGCGGTTGGGT
1_13351653R	GCTTCCAGTACGTTATTTGGG
1_8419370R	GTATGTAACCGCATCCGCAT
4_3651495R	GGATACGTCATCGGGGTTCC
2_7171492R_INV	TTAACACCGCCAAATGTTGA
1_12645883R	AAGGATCTAGAAGATCAACAAGTT
5_10076009R	AGAAATCGTAAGAAAGTGGATG
1_14920456R	CAACTTACACACGCCTCTAACA
2_2894614R	AAAACATAAAGCTTCCTCTCTCT
5_15952489R	CCCAGAAACGACCAACATCT
4_7609008R	TTTTGTGTTTCTTCCCAAAAAGA
1_17366573R	ACCGAGATCGATTCCAGGAA
1_23974142R	TTGTGATGGGAACGTTTTGA
4_2782483R	AGCGGCGTAAATCAAGATGT
5_10272973R_INV	ACCCCCATTTTTATGTGCAA
4_2781335R	TTGACGACAACGAAGACGAC
5_14200929R	CACCATACTTTTCATCTTATTAAAAA
2_2089357R	TAGCAGTATGGCCCGTGCTA
3_16221181R	TCACTCTCCCCTTCTATCTCTTTG
5_10918979R	CCAAAATGTTGCGATAAGAAA
1_24214408R	ACATGGATCCCATGACCATT
5_18835849R	GCCTCTTCAAGCTGGTGTCT
2_7171492R	TCAACATTTGGCGGTGTTAA
2_2184567R	CCTCAGTAAGCCAGCGTTTT
3_22556429R	TTCTCGTGGCTGATTGTTTCT
4_1821194R	ACCGGTGAGGTACTAGCGAA
3_16049768R	CTCCATGGATGTTTCCTAGAA
3_16196032R	TGTGTGATTCTTCAACTCTCTGG
3_22563028R	TAGCCGACGCTACTTCGATT
3_17359469R	TCAACAACCCACATTCGAAA
3_22563045R	CCGACGCTACTTCGATTCTC
3_22228732R	GGTTTCGCAGTCTTTCTCGT
4_2781917R_INV	GACGTTGTCCTCACCTGGTT
4_2782548R_INV	CTTTTTGGGAACGGACTGAC
4_2783051R	GCCTAAATGTGCAACATTCTCA
4_1612571R	GGCTGAAAGAAGGACGGTTT
1_17366573R_INV	TTCCTGGAATCGATCTCGGT
5_13626302R	GGTTTCATCGCCACTGTTTT
3_904402R_INV	GNCTCGTTATCTCCATGTTGA
1_12227492R	ACCATCTCCTGAAACGCACT
1_11592551R_INV	GAACTGCAAATTTGGGCCTA
5_14839317R	CTCAGATCCCAGTGACTGCA
4_1786363R	GCCAACCACACAACCTNTAGC
4_4497930R_INV	TTCCCGTTTTACAATGACGG
5_18835849R_INV	AGACACCAGCTTGAAGAGGC

1_23974142R_INV	TCAAAACGTTCCCATCACAA

Table E.2: Oligo sequences for SV validation by PCR

Appendix F

Capillary sequences for breakpoint validation

Id	Line	Chr	Forward oligo (F)	Reverse oligo (R)	PS	SP (F)	SP (R)
1	MAGIC.287	3	CGAGGCATATTTTACAAGCA	TTCGTGAAACTTTTAGTGGGG	795	18976543	18977317
2	MAGIC.287	3	TCGCACGGATGAATAATGAC	TTCGATCGTAGAAACCTCCC	818	19171418	19172216
3	MAGIC.287	3	AGGACTCGATGTGGAAAGCA	CAATGCCACATTCACCATCT	942	19404571	19405493
4	MAGIC.287	3	GGTATCGAAAACGAATGGGA	CTCCATTCCTCAAAAACACTGA	996	19466727	19467701
5	MAGIC.287	3	ATCCTCGAGTGTTTTGGGTG	AACTACTCGGTGGCACATCA	911	21434906	21435797
6	MAGIC.287	3	GAACACACGATCGAAAATGG	TAGGTCGCAAGCGCAAGTAG	824	21556316	21557120
7	MAGIC.287	3	AGGTAATGCGCACAAAAAGG	AATCACCGAGATCACCGAAC	749	21715914	21716643
8	MAGIC.287	3	GAATCCCCATGTAAGGGTCA	CATCACGGTCAGTCATCCAC	943	22107605	22108528
9	MAGIC.287	3	GTGTATTGCCGCCAGTTTTT	ACGAGAAATGAGCTGTGCTT	709	22140327	22141016
10	MAGIC.287	3	AACCTTCCGAATAAAGCCAA	CCTGAATTATCGGAAGCGAG	594	22187409	22187983
11	MAGIC.287	3	CCGTCTTCGGCCTCTATCAA	AGATGCAATGACTTTTGGAC	984	22442279	22443243
12	MAGIC.287	3	GTCCCCAAGTGCATTCTGTC	TTTGTTACGTGCTCCACCAG	835	22465859	22466674
13	MAGIC.287	3	CAGGTGAAGCTTATTGGGGA	TCCGGGCTAGACTTTTTCTG	774	22493961	22494715
14	MAGIC.287	3	ACACGCACTCACTGTCAGAA	CTTCTGACACCCCGATGAAC	844	22816058	22816882
15	MAGIC.446	3	GGAGCTTTTGGACGAGACAG	GCAAAAGAGTCGGACCGTAG	850	1727988	1728818
16	MAGIC.446	3	TGGGATAAATCATGGAAGCC	CATGAGCCAAGGTTGATGTG	725	2627432	2628137
17	MAGIC.446	3	CTGGGTCGGTTTACTTGTGG	TGGATATTTTGCACGGCTTC	717	4633722	4634419
18	MAGIC.446	3	GTTTTGAACGTAGCTGGGGA	CAGAGCGATCAGAAAGGAATG	965	5372625	5373569
19	MAGIC.446	3	CGATGTTATCGTGGGAGATG	TTTTGTCTCACCGCTCAATC	771	6118672	6119423
20	MAGIC.446	3	CATTATGCACCAGGAGGAAG	AACCTAGCCAGGGATGCTTT	822	6121670	6122472
21	MAGIC.446	3	AAAGTGTCGGGGCTTCTTCT	TAAATGCGCGAGAGACAAAA	604	6160037	6160621
22	MAGIC.446	3	AACAACCAGCACAATCAACA	TTGCCGTCTCCTTTATCGTC	831	6265318	6266129
23	MAGIC.446	3	CTAGTGTCGCCAAACCATCA	TGACACACTCTTCACCGACC	852	6465846	6466678
24	MAGIC.446	3	TATCACGGAAGGTACGGCTC	GCGTGTGCTATCTTAGCTTCC	942	6490157	6491078
25	MAGIC.446	3	TTTGGCTGTCTTCTTGTCCC	AGAGCACAATCGGAACCAAC	951	6652074	6653005
26	MAGIC.446	3	CCAAATTGATTCGACATTCC	TTTCTTCGAAATCACACCCC	726	6925595	6926301
27	MAGIC.446	4	AAGGACAGAACCGACTTCACA	ATTGTTACCTGTCCACACGAC	568	1244873	1245420
28	MAGIC.446	4	TAGTCCCTTGATTAAAAGCCC	GCATCATTTCCGCACTCTTT	875	714625	715480
29	MAGIC.446	5	GAAACACCTCAAGGGAAGCA	ATCAACCATTCCCCATGTTC	883	19304361	19305224

Table F.1: Primer sequences used for validation. Id indicates the sequence id, Line the MAGIC line id, Chr the chromosome, Forward oligo (F) and Reverse oligo (R) the two primer oligos, PS is the product size and SP the starting position of each oligo in the chromosome.

Id	Line	Chr	start.bp	end.bp	l.hap	r.hap	spans.bp	l.sites	r.sites	sup.sites	confirm
1	MAGIC.287	3	18977057	18977127	oy-0	wil-2	TRUE	7	1	8	TRUE
2	MAGIC.287	3	19171742	19172025	ws-0	po-0	TRUE	1	1	2	TRUE
3	MAGIC.287	3	19404954	19405256	po-0	can-0	TRUE	8	6	14	TRUE
4	MAGIC.287	3	19466843	19467573	hi-0	po-0	TRUE	2	1	3	TRUE
5	MAGIC.287	3	21435482	21435731	po-0	hi-0	TRUE	6	3	7	FALSE
6	MAGIC.287	3	21556521	21556784	hi-0	ct-1	TRUE	1	1	2	TRUE
7	MAGIC.287	3	21716088	21716592	ct-1	ler-0	TRUE	1	1	2	TRUE
9	MAGIC.287	3	22140861	22140966	bur-0	oy-0	TRUE	9	2	11	TRUE
10	MAGIC.287	3	22187581	22187889	oy-0	mt-0	TRUE	2	1	3	TRUE
11	MAGIC.287	3	22442457	22443184	wil-2	bur-0	TRUE	1	1	2	TRUE
12	MAGIC.287	3	22466007	22466444	bur-0	col-0	TRUE	1	1	2	TRUE
13	MAGIC.287	3	22494608	22494627	col-0	ct-1	TRUE	1	3	4	TRUE
14	MAGIC.287	3	22816278	22816509	edi-0	bur-0	TRUE	1	2	3	TRUE
15	MAGIC.446	3	1728604	1728606	po-0	wu-0	TRUE	2	1	3	TRUE
16	MAGIC.446	3	2627870	2627920	ws-0	po-0	TRUE	1	1	2	TRUE
17	MAGIC.446	3	4633961	4634153	bur-0	ct-1	TRUE	1	3	4	TRUE
18	MAGIC.446	3	5373046	5373192	can-0	edi-0	TRUE	2	1	3	TRUE
19	MAGIC.446	3	6119115	6119116	po-0	ler-0	TRUE	6	1	7	TRUE
20	MAGIC.446	3	6121798	6122288	ler-0	oy-0	TRUE	1	1	2	TRUE
21	MAGIC.446	3	6160418	6160559	ct-1	po-0	TRUE	2	1	3	TRUE
23	MAGIC.446	3	6466292	6466492	wu-0	bur-0	FALSE	0	0	0	FALSE
24	MAGIC.446	3	6490573	6490952	bur-0	hi-0	TRUE	4	2	6	TRUE
25	MAGIC.446	3	6652377	6652560	tsu-0	col-0	TRUE	1	4	5	TRUE
26	MAGIC.446	3	6925725	6926217	wu-0	can-0	TRUE	1	1	2	TRUE
27	MAGIC.446	4	1245093	1245370	col-0	ct-1	TRUE	2	1	3	TRUE
28	MAGIC.446	4	714731	714762	po-0	kn-0	TRUE	1	1	2	TRUE
29	MAGIC.446	5	19304522	19304889	tsu-0	rsch-4	TRUE	2	3	5	TRUE

Table F.2: Results of sequencing of breakpoint regions after PCR amplifications. Only primers that produced a products that formed a contig are included. The columns are: Id the index of the contig tested, Line the MAGIC line name, Chr the chromosome of the region, start.bp, end.bp the coordinates of the breakpoint being verified, l.hap and r.hap the accession names of the two haplotypes flanking the breakpoint, spans.bp is a boolean variable indicating whether the sequence spans the breakpoint, l.sites, r.sites the number of informative sites (different in the two haplotypes) on each side of the breakpoint, sup.sites the total number of genotypes supporting the haplotype predictions, confirm Boolean variable showing whether the region is consistent to the predictions.

Bibliography

[1] Avigail Agam, Binnaz Yalcin, Amarjit Bhomra, Matthew Cubin, Caleb Webber, Christopher Holmes, Jonathan Flint, and Richard Mott. Elusive copy number variation in the mouse genome. *PloS one*, 5(9):e12839, January 2010.

[2] Imranul Alam, Daniel L. Koller, Qiwei Sun, Ryan K. Roeder, Toni Cañete, Gloria Blázquez, Regina López-Aumatell, Esther Martínez-Membrives, Elia Vicens-Costa, Carme Mont, Sira Díaz, Adolf Tobeña, Alberto Fernández-Teruel, Adam Whitley, Pernilla Strid, Margarita Diez, Martina Johannesson, Jonathan Flint, Michael J. Econs, Charles H. Turner, and Tatiana Foroud. Heterogeneous stock rat: A unique animal model for mapping genes influencing bone fragility. *Bone*, 48:1169–1177, 2011.

[3] D M Altshuler, R M Durbin, G R Abecasis, D R Bentley, A Chakravarti, A G Clark, P Donnelly, E E Eichler, P Flicek, S B Gabriel, R A Gibbs, E D Green, M E Hurles, B M Knoppers, J O Korbel, E S Lander, C Lee, H Lehrach, E R Mardis, G T Marth, G A Mcvean, D A Nickerson, J P Schmidt, S T Sherry, J Wang, R K Wilson, H Dinh, C Kovar, S Lee, L Lewis, D Muzny, J Reid, M Wang, X D Fang, X S Guo, M Jian, H Jiang, X Jin, G Q Li, J X Li, Y R Li, Z Li, X Liu, Y Lu, X D Ma, Z Su, S S Tai, M F Tang, B Wang, G B Wang, H L Wu, R H Wu, Y Yin, W W Zhang, J Zhao, M R Zhao, X L Zheng, Y Zhou, N Gupta, L Clarke, R Leinonen, R E Smith, X Zheng-Bradley, R Grocock, S Humphray, T James, Z Kingsbury, R Sudbrak, M W Albrecht, V S Amstislavskiy, T A Borodina, M Lienhard, F Mertes, M Sultan, B Timmermann, M L Yaspo, L Fulton, R Fulton, G M Weinstock, S Balasubramaniam, J Burton, P Danecek, T M Keane, A Kolb-Kokocinski, S McCarthy, J Stalker, M Quail, C J

Davies, J Gollub, T Webster, B Wong, Y P Zhan, A Auton, F Yu, M Bainbridge, D Challis, U S Evani, J Lu, U Nagaswamy, A Sabo, Y Wang, J Yu, L J M Coin, L Fang, Q B Li, Z Y Li, H X Lin, B H Liu, R B Luo, N Qin, H J Shao, B Q Wang, Y L Xie, C Ye, C Yu, F Zhang, H C Zheng, H M Zhu, E P Garrison, D Kural, W P Lee, W F Leong, A N Ward, J T Wu, M Y Zhang, L Griffin, C H Hsieh, R E Mills, X H Shi, M von Grotthuss, C S Zhang, M J Daly, M A DePristo, E Banks, G Bhatia, M O Carneiro, G del Angel, G Genovese, R E Handsaker, C Hartl, S A McCarroll, J C Nemesh, R E Poplin, S F Schaffner, K Shakir, S C Yoon, J Lihm, V Makarov, H J Jin, W Kim, K C Kim, T Rausch, K Beal, F Cunningham, J Herrero, W M McLaren, G R S Ritchie, S Gottipati, A Keinan, J L Rodriguez-Flores, P C Sabeti, S R Grossman, S Tabrizi, R Tariyal, D N Cooper, E V Ball, P D Stenson, B Barnes, M Bauer, R K Cheetham, T Cox, M Eberle, S Kahn, L Murray, J Peden, R Shaw, K Ye, M A Batzer, M K Konkel, J A Walker, D G MacArthur, M Lek, R Herwig, M D Shriver, C D Bustamante, J K Byrnes, F M De la Vega, S Gravel, E E Kenny, J M Kidd, P Lacroute, B K Maples, A Moreno-Estrada, F Zakharia, E Halperin, Y Baran, D W Craig, A Christoforides, N Homer, T Izatt, A A Kurdoglu, S A Sinari, K Squire, C L Xiao, J Sebat, V Bafna, E G Burchard, R D Hernandez, C R Gignoux, D Haussler, S J Katzman, W J Kent, B Howie, A Ruiz-Linares, and E T Dermitzakis. An integrated map of genetic variation from 1,092 human genomes. *Nature*, 491:56–65, 2012.

[4] Simon Anders and Wolfgang Huber. Differential expression analysis for sequence count data. *Genome Biology*, 11:R106, 2010.

[5] The Arabidopsis and Genome Initiative. Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. *Nature*, 408(6814):796–815, 2000.

[6] S Atwell, Y S Huang, B J Vilhjalmsson, G Willems, M Horton, Y Li, D Meng, A Platt, A M Tarone, T T Hu, R Jiang, N W Muliyati, X Zhang, M A Amer, I Baxter, B Brachi, J Chory, C Dean, M Debieu, J de Meaux, J R Ecker, N Faure, J M Kniskern, J D Jones, T Michael, A Nemri, F Roux, D E Salt, C Tang, M Todesco, M B Traw, D Weigel, P Marjoram, J O

Borevitz, J Bergelson, and M Nordborg. Genome-wide association study of 107 phenotypes in Arabidopsis thaliana inbred lines. *Nature*, 465(7298):627–631, 2010.

[7] S Barth, A E Melchinger, B Devezi-Savula, and T Lübberstedt. Influence of genetic background and heterozygosity on meiotic recombination in Arabidopsis thaliana. *Genome / National Research Council Canada = Génome / Conseil national de recherches Canada*, 44(6):971–8, December 2001.

[8] Sayantani Basu-Roy, Franck Gauthier, Laurène Giraut, Christine Mézard, Matthieu Falque, and Olivier C Martin. Hot Regions of Noninterfering Crossovers Coexist with a Nonuniformly Interfering Pathway in Arabidopsis thaliana. *Genetics*, 195(3):769–79, November 2013.

[9] F Baudat, J Buard, C Grey, A Fledel-Alon, C Ober, M Przeworski, G Coop, and B De Massy. PRDM9 is a major determinant of meiotic recombination hotspots in humans and mice. *Science*, 327(5967):836–40, 2010.

[10] Y Benjamini and Y Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society Series B Methodological*, 57:289–300, 1995.

[11] K W Broman and J L Weber. Characterization of human crossover interference. *American journal of human genetics*, 66:1911–1926, 2000.

[12] Karl W Broman, Lucy B Rowe, Gary A Churchill, and Ken Paigen. Crossover interference in the mouse. *Genetics*, 160:1123–1131, 2002.

[13] KW Broman. The genomes of recombinant inbred lines. *Genetics*, 169:1133–1146, 2005.

[14] Jonathan Butler, Iain MacCallum, Michael Kleber, Ilya A Shlyakhter, Matthew K Belmonte, Eric S Lander, Chad Nusbaum, and David B Jaffe. ALLPATHS: de novo assembly of whole-genome shotgun microreads. *Genome research*, 18(5):810–20, May 2008.

[15] Claudia P Cabrera, Pau Navarro, Jennifer E Huffman, Alan F Wright, Caroline Hayward, Harry Campbell, James F Wilson, Igor Rudan, Nicholas D Hastie, Veronique Vitart, and

Chris S Haley. Uncovering networks from genome-wide association studies via circular genomic permutation. *G3 (Bethesda, Md.)*, 2(9):1067–75, September 2012.

[16] Patrick Cahan, Yedda Li, Masayo Izumi, and Timothy A Graubert. The impact of copy number variation on local gene expression in mouse hematopoietic stem and progenitor cells. *Nature genetics*, 41:430–437, 2009.

[17] Jun Cao, Korbinian Schneeberger, Stephan Ossowski, Torsten Günther, Sebastian Bender, Joffrey Fitz, Daniel Koenig, Christa Lanz, Oliver Stegle, Christoph Lippert, Xi Wang, Felix Ott, Jonas Müller, Carlos Alonso-Blanco, Karsten Borgwardt, Karl J Schmid, and Detlef Weigel. Whole-genome sequencing of multiple Arabidopsis thaliana populations. *Nature genetics*, 43:956–63, 2011.

[18] Mark J Chaisson and Glenn Tesler. Mapping single molecule sequencing reads using Basic Local Alignment with Successive Refinement (BLASR): Theory and Application. *BMC Bioinformatics*, 13:238, 2012.

[19] Mark J. P. Chaisson, John Huddleston, Megan Y. Dennis, Peter H. Sudmant, Maika Malig, Fereydoun Hormozdiari, Francesca Antonacci, Urvashi Surti, Richard Sandstrom, Matthew Boitano, Jane M. Landolin, John A. Stamatoyannopoulos, Michael W. Hunkapiller, Jonas Korlach, and Evan E. Eichler. Resolving the complexity of the human genome using single-molecule sequencing. *Nature*, 514, November 2014.

[20] Jarrod A Chapman, Martin Mascher, Ayd N Buluç, Kerrie Barry, Evangelos Georganas, Adam Session, Veronika Strnadova, Jerry Jenkins, Sunish Sehgal, Leonid Oliker, Jeremy Schmutz, Katherine A Yelick, Uwe Scholz, Robbie Waugh, Jesse A Poland, Gary J Muehlbauer, Nils Stein, and Daniel S Rokhsar. A whole-genome shotgun approach for assembling and anchoring the hexaploid bread wheat genome. *Genome biology*, 16(1):26, January 2015.

[21] Ken Chen, John W Wallis, Michael D McLellan, David E Larson, Joelle M Kalicki, Craig S Pohl, Sean D McGrath, Michael C Wendl, Qunyuan Zhang, Devin P Locke, Xiaoqi Shi,

Robert S Fulton, Timothy J Ley, Richard K Wilson, Li Ding, and Elaine R Mardis. BreakDancer: an algorithm for high-resolution mapping of genomic structural variation. *Nature methods*, 6(9):677–81, September 2009.

[22] Derek Y Chiang, Gad Getz, David B Jaffe, Michael J T O'Kelly, Xiaojun Zhao, Scott L Carter, Carsten Russ, Chad Nusbaum, Matthew Meyerson, and Eric S Lander. High-resolution mapping of copy-number alterations with massively parallel sequencing. *Nature methods*, 6:99–103, 2009.

[23] Kyuha Choi, Xiaohui Zhao, Krystyna A. Kelly, Oliver Venn, James D. Higgins, Nataliya E. Yelina, Thomas J. Hardcastle, Piotr A. Ziolkowski, Gregory P. Copenhaver, F Chris H Franklin, Gil McVean, and Ian R. Henderson. Arabidopsis meiotic crossover hot spots overlap with H2A.Z nucleosomes at gene promoters. *Nature genetics*, 45:1327–36, 2013.

[24] Kyuha Choi, Xiaohui Zhao, Krystyna A Kelly, Oliver Venn, James D Higgins, Nataliya E Yelina, Thomas J Hardcastle, Piotr A Ziolkowski, Gregory P Copenhaver, F Chris H Franklin, Gil McVean, and Ian R Henderson. Arabidopsis meiotic crossover hot spots overlap with H2A.Z nucleosomes at gene promoters. *Nature genetics*, 45(11):1327–36, November 2013.

[25] R Coco and V B Penchaszadeh. Cytogenetic findings in 200 children with mental retardation and multiple congenital anomalies of unknown cause. *American journal of medical genetics*, 12(2):155–73, June 1982.

[26] Donald F Conrad, Dalila Pinto, Richard Redon, Lars Feuk, Omer Gokcumen, Yujun Zhang, Jan Aerts, T Daniel Andrews, Chris Barnes, Peter Campbell, Tomas Fitzgerald, Min Hu, Chun Hwa Ihm, Kati Kristiansson, Daniel G Macarthur, Jeffrey R Macdonald, Ifejinelo Onyiah, Andy Wing Chun Pang, Sam Robson, Kathy Stirrups, Armand Valsesia, Klaudia Walter, John Wei, Chris Tyler-Smith, Nigel P Carter, Charles Lee, Stephen W Scherer, and Matthew E Hurles. Origins and functional impact of copy number variation in the human genome. *Nature*, 464(7289):704–12, April 2010.

[27] Rat Genome Sequencing Consortium, Mapping, Amelie Baud, Roel Hermsen, Victor Guryev, Pernilla Stridh, Delyth Graham, Martin W McBride, Tatiana Foroud, Sophie Calderari, Margarita Diez, Johan Ockinger, Amennai D Beyeen, Alan Gillett, Nada Abdelmagid, Andre Ortlieb Guerreiro-Cacais, Maja Jagodic, Jonatan Tuncel, Ulrika Norin, Elisabeth Beattie, Ngan Huynh, William H Miller, Daniel L Koller, Imranul Alam, Samreen Falak, Mary Osborne-Pellegrin, Esther Martinez-Membrives, Toni Canete, Gloria Blazquez, Elia Vicens-Costa, Carme Mont-Cardona, Sira Diaz-Moran, Adolf Tobena, Oliver Hummel, Diana Zelenika, Kathrin Saar, Giannino Patone, Anja Bauerfeind, Marie-Therese Bihoreau, Matthias Heinig, Young-Ae Lee, Carola Rintisch, Herbert Schulz, David A Wheeler, Kim C Worley, Donna M Muzny, Richard A Gibbs, Mark Lathrop, Nico Lansu, Pim Toonen, Frans Paul Ruzius, Ewart de Bruijn, Heidi Hauser, David J Adams, Thomas Keane, Santosh S Atanur, Tim J Aitman, Paul Flicek, Tomas Malinauskas, E Yvonne Jones, Diana Ekman, Regina Lopez-Aumatell, Anna F Dominiczak, Martina Johannesson, Rikard Holmdahl, Tomas Olsson, Dominique Gauguier, Norbert Hubner, Alberto Fernandez-Teruel, Edwin Cuppen, Richard Mott, and Jonathan Flint. Combined sequence-based and genetic mapping analysis of complex traits in outbred rats. *Nature genetics*, 45:767–775, 2013.

[28] A Darvasi and M Soller. Advanced intercross lines, an experimental population for fine genetic mapping. *Genetics*, 141:1199–1207, 1995.

[29] Frans J. de Bruijn, editor. *Molecular Microbial Ecology of the Rhizosphere*. John Wiley & Sons, Inc., Hoboken, NJ, USA, May 2013.

[30] Seth Debolt. Copy number variation shapes genome diversity in arabidopsis over immediate family generational scales. *Genome Biology and Evolution*, 2:441–453, 2010.

[31] Christian Dimkpa, Tanja Weinand, and Folkard Asch. Plant-rhizobacteria interactions alleviate abiotic stress conditions. *Plant, cell & environment*, 32(12):1682–94, December 2009.

[32] P Ferragina and G Manzini. Opportunistic data structures with applications. In *Proceeding FOCS '00 Proceedings of the 41st Annual Symposium on Foundations of Computer Science-*

Proceeding FOCS '00 Proceedings of the 41st Annual Symposium on Foundations of Computer Science, pages 390–398, 2000.

[33] J M Fletcher, K Evans, D Baillie, P Byrd, D Hanratty, S Leach, C Julier, J R Gosden, W Muir, and D J Porteous. Schizophrenia-associated chromosome 11q21 translocation: identification of flanking markers and development of chromosome 11q fragment hybrids as cloning and mapping resources. *American journal of human genetics*, 52(3):478–90, March 1993.

[34] Jonathan Flint, Yiping Chen, Shenxun Shi, and Kenneth S Kendler. Epilogue: Lessons from the CONVERGE study of major depressive disorder in China. *Journal of affective disorders*, 140(1):1–5, September 2012.

[35] Michael Freeling. Bias in plant gene content following different sorts of duplication: tandem, whole-genome, segmental, or by transposition. *Annual Review of Plant Biology*, 60(January):433–453, 2009.

[36] X C Gan, O Stegle, J Behr, J G Steffen, P Drewe, K L Hildebrand, R Lyngsoe, S J Schultheiss, E J Osborne, V T Sreedharan, A Kahles, R Bohnert, G Jean, P Derwent, P Kersey, E J Belfield, N P Harberd, E Kemen, C Toomajian, Paula X Kover, R M Clark, G Ratsch, and R Mott. Multiple reference genomes and transcriptomes for Arabidopsis thaliana. *Nature*, 477(7365):1–5, 2011.

[37] Giulio Genovese, Robert E Handsaker, Heng Li, Nicolas Altemose, Amelia M Lindgren, Kimberly Chambert, Bogdan Pasaniuc, Alkes L Price, David Reich, Cynthia C Morton, Martin R Pollak, James G Wilson, and Steven a McCarroll. Using population admixture to help complete maps of the human genome. *Nature genetics*, 45:406–14, 414e1–2, 2013.

[38] Sante Gnerre, Iain Maccallum, Dariusz Przybylski, Filipe J Ribeiro, Joshua N Burton, Bruce J Walker, Ted Sharpe, Giles Hall, Terrance P Shea, Sean Sykes, Aaron M Berlin, Daniel Aird, Maura Costello, Riza Daza, Louise Williams, Robert Nicol, Andreas Gnirke, Chad Nusbaum, Eric S Lander, and David B Jaffe. High-quality draft assemblies of mammalian genomes from

massively parallel sequence data. *Proceedings of the National Academy of Sciences of the United States of America*, 108(4):1513–1518, 2011.

[39] Anthony Gobert, Graeme Park, Anna Amtmann, Dale Sanders, and Frans J M Maathuis. Arabidopsis thaliana cyclic nucleotide gated channel 3 forms a non-selective ion transporter involved in germination and cation transport. *Journal of experimental botany*, 57(4):791–800, January 2006.

[40] Gideon Grafi, Assa Florentin, Vanessa Ransbotyn, and Yakov Morgenstern. The stem cell state in plant development and in response to stress. *Frontiers in plant science*, 2:53, January 2011.

[41] Kevin M Haigis and William F Dove. A Robertsonian translocation suppresses a somatic recombination pathway to loss of heterozygosity. *Nature genetics*, 33(1):33–9, January 2003.

[42] Robert E Handsaker, Vanessa Van Doren, Jennifer R Berman, Giulio Genovese, Seva Kashin, Linda M Boettger, and Steven A McCarroll. Large multiallelic copy number variations in humans. *Nature genetics*, 47(3):296–303, January 2015.

[43] Ross C Hardison. Comparative Genomics. *PLoS Biology*, 1(2):E58, 2003.

[44] P. J. Hastings, Grzegorz Ira, and James R. Lupski. A microhomology-mediated break-induced replication model for the origin of human copy number variation. *PLoS Genetics*, 5, 2009.

[45] Charlotte N Henrichsen, Nicolas Vinckenbosch, Sebastian Zöllner, Evelyne Chaignat, Sylvain Pradervand, Frédéric Schütz, Manuel Ruedi, Henrik Kaessmann, and Alexandre Reymond. Segmental copy number variation shapes tissue transcriptomes. *Nature genetics*, 41:424–429, 2009.

[46] Casandra Hernández-Reyes, Sebastian T Schenk, Christina Neumann, Karl-Heinz Kogel, and Adam Schikora. N-acyl-homoserine lactones-producing bacteria protect plants against plant and human pathogens. *Microbial biotechnology*, 7(6):580–8, November 2014.

[47] Kenneth J Hillers. Crossover interference. *Current biology : CB*, 14:R1036–R1037, 2004.

[48] Anjali G Hinch, Arti Tandon, Nick Patterson, Yunli Song, Nadin Rohland, Cameron D Palmer, Gary K Chen, Kai Wang, Sarah G Buxbaum, Ermeg L Akylbekova, Melinda C Aldrich, Christine B Ambrosone, Christopher Amos, Elisa V Bandera, Sonja I Berndt, Leslie Bernstein, William J Blot, Cathryn H Bock, Eric Boerwinkle, Qiuyin Cai, Neil Caporaso, Graham Casey, L Adrienne Cupples, Sandra L Deming, W Ryan Diver, Jasmin Divers, Myriam Fornage, Elizabeth M Gillanders, Joseph Glessner, Curtis C Harris, Jennifer J Hu, Sue A Ingles, William Isaacs, Esther M John, W H Linda Kao, Brendan Keating, Rick A Kittles, Laurence N Kolonel, Emma Larkin, Loic Le Marchand, Lorna H McNeill, Robert C Millikan, Adam Murphy, Solomon Musani, Christine Neslund-Dudas, Sarah Nyante, George J Papanicolaou, Michael F Press, Bruce M Psaty, Alex P Reiner, Stephen S Rich, Jorge L Rodriguez-Gil, Jerome I Rotter, Benjamin A Rybicki, Ann G Schwartz, Lisa B Signorello, Margaret Spitz, Sara S Strom, Michael J Thun, Margaret A Tucker, Zhaoming Wang, John K Wiencke, John S Witte, Margaret Wrensch, Xifeng Wu, Yuko Yamamura, Krista A Zanetti, Wei Zheng, Regina G Ziegler, Xiaofeng Zhu, Susan Redline, Joel N Hirschhorn, Brian E Henderson, Herman A Taylor, Alkes L Price, Hakon Hakonarson, Stephen J Chanock, Christopher A Haiman, James G Wilson, David Reich, and Simon R Myers. The landscape of recombination in African Americans. *Nature*, 476(7359):170–175, 2011.

[49] Fereydoun Hormozdiari, Can Alkan, Evan E. Eichler, and S. Cenk Sahinalp. Combinatorial algorithms for structural variation detection in high-throughput sequenced genomes. *Genome Research*, 19:1270–1278, 2009.

[50] M W Horton, A M Hancock, Y S Huang, C Toomajian, S Atwell, A Auton, N W Muliyati, A Platt, F G Sperone, B J Vilhjalmsson, M Nordborg, J O Borevitz, and J Bergelson. Genome-wide patterns of genetic variation in worldwide Arabidopsis thaliana accessions from the RegMap panel. *Nature Genetics*, 44(2):212–216, 2012.

[51] E A Housworth and F W Stahl. Crossover interference in humans. *American journal of human genetics*, 73:188–197, 2003.

[52] F A Iraqi, M Mahajne, Y Salaymah, H Sandovski, H Tayem, K Vered, L Balmer, M Hall, G Manship, G Morahan, K Pettit, J Scholten, K Tweedie, A Wallace, L Weerasekera, J Cleak, C Durrant, L Goodstadt, R Mott, B Yalcin, D L Aylor, R S Baric, T A Bell, K M Bendt, J Brennan, J D Brooks, R J Buus, J J Crowley, J D Calaway, M E Calaway, A Cholka, D B Darr, J P Didion, A Dorman, E T Everett, M T Ferris, W F Mathes, C P Fu, T J Gooch, S G Goodson, L E Gralinski, S D Hansen, M T Heise, J Hoel, K J Hua, M C Kapita, S Lee, A B Lenarcic, E Y Liu, H D Liu, L McMillan, T R Magnuson, K F Manly, D R Miller, D A O'Brien, F Odet, I K Pakatci, W Q Pan, F P M de Villena, C M Perou, D Pomp, C R Quackenbush, N N Robinson, N E Sharpless, G D Shaw, J S Spence, P F Sullivan, W Sun, L M Tarantino, W Valdar, J Wang, W Wang, C E Welsh, A Whitmore, T Wiltshire, F A Wright, Y Y Xie, Z N Yun, V Zhabotynsky, Z J Zhang, F Zou, C Powell, J Steigerwalt, D W Threadgill, E J Chesler, G A Churchill, D M Gatti, R Korstanje, K L Svenson, F S Collins, N Crawford, K Hunter, S N P Kelada, B C E Peck, K Reilly, U Tavarez, D Bottomly, R Hitzeman, S K McWeeney, J Frelinger, H Krovi, J Phillippi, R A Spritz, L Aicher, M Katze, E Rosenzweig, A Shusterman, A Nashef, E I Weiss, Y Houri-Haddad, M Soller, R W Williams, K Schughart, H N Yang, J E French, A K Benson, J Kim, R Legge, S J Low, F R Ma, I Martinez, J Walter, K W Broman, B Hallgrimsson, O Klein, G Weinstock, W C Warren, Y V Yang, D Schwartz, and Collaborative Cross Consortium. The Genome Architecture of the Collaborative Cross Mouse Genetic Reference Population. *Genetics*, 190:389–U159, 2012.

[53] P A Jacobs, A G Baikie, W M Court Brown, and J A Strong. The somatic chromosomes in mongolism. *Lancet*, 1:710, 1959.

[54] Miten Jain, Ian T Fiddes, Karen H Miga, Hugh E Olsen, Benedict Paten, and Mark Akeson. Improved data analysis for the MinION nanopore sequencer. *Nature Methods*, advance on, February 2015.

[55] Alec J Jeffreys, Liisa Kauppi, and Rita Neumann. Intensely punctate meiotic recombination in the class II region of the major histocompatibility complex. *Nature Genetics*, 29(2):217–222, 2001.

[56] Janelle K H Jung and Susan McCouch. Getting to the roots of it: Genetic and hormonal control of root architecture. *Frontiers in plant science*, 4:186, January 2013.

[57] Sang-Mo Kang, Gil-Jae Joo, Muhammad Hamayun, Chae-In Na, Dong-Hyun Shin, Hak Youn Kim, Jin-Kyu Hong, and In-Jung Lee. Gibberellin production and phosphate solubilization by newly isolated strain of Acinetobacter calcoaceticus and its effect on plant growth. *Biotechnology letters*, 31(2):277–81, February 2009.

[58] S Karlin and S F Altschul. Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes. *Proceedings of the National Academy of Sciences of the United States of America*, 87(6):2264–2268, 1990.

[59] W J Kent. BLAT–the BLAST-like alignment tool. *Genome Research*, 12(4):656–664, 2002.

[60] W James Kent. BLAT–the BLAST-like alignment tool. *Genome Research*, 12(4):656–664, 2002.

[61] Jeffrey M Kidd, Gregory M Cooper, William F Donahue, Hillary S Hayden, Nick Sampas, Tina Graves, Nancy Hansen, Brian Teague, Can Alkan, Francesca Antonacci, Eric Haugen, Troy Zerr, N Alice Yamada, Peter Tsang, Tera L Newman, Eray Tüzün, Ze Cheng, Heather M Ebling, Nadeem Tusneem, Robert David, Will Gillett, Karen A Phelps, Molly Weaver, David Saranga, Adrianne Brand, Wei Tao, Erik Gustafson, Kevin McKernan, Lin Chen, Maika Malig, Joshua D Smith, Joshua M Korn, Steven A McCarroll, David A Altshuler, Daniel A Peiffer, Michael Dorschner, John Stamatoyannopoulos, David Schwartz, Deborah A Nickerson, James C Mullikin, Richard K Wilson, Laurakay Bruhn, Maynard V Olson, Rajinder Kaul, Douglas R Smith, and Evan E Eichler. Mapping and sequencing of structural variation from eight human genomes. *Nature*, 453:56–64, 2008.

[62] S Kim, V Plagnol, TT Hu, C Toomajian, RM Clark, S Ossowski, Ecker, D Weigel, and M Nordborg. Recombination and linkage disequilibrium in Arabidopsis thaliana. *Nature Genetics*, 39(9):1151–5, 2007.

[63] Seyoung Kim and Eric P. Xing. Statistical estimation of correlated genome associations to a quantitative trait network. *PLoS Genetics*, 5(8), 2009.

[64] Ewen F Kirkness, Brian J Haas, Weilin Sun, Henk R Braig, M Alejandra Perotti, John M Clark, Si Hyeock Lee, Hugh M Robertson, Ryan C Kennedy, Eran Elhaik, Daniel Gerlach, Evgenia V Kriventseva, Christine G Elsik, Dan Graur, Catherine A Hill, Jan A Veenstra, Brian Walenz, José Manuel C Tubío, José M C Ribeiro, Julio Rozas, J Spencer Johnston, Justin T Reese, Aleksandar Popadic, Marta Tojo, Didier Raoult, David L Reed, Yoshinori Tomoyasu, Emily Kraus, Emily Krause, Omprakash Mittapalli, Venu M Margam, Hong-Mei Li, Jason M Meyer, Reed M Johnson, Jeanne Romero-Severson, Janice Pagel Vanzee, David Alvarez-Ponce, Filipe G Vieira, Montserrat Aguadé, Sara Guirao-Rico, Juan M Anzola, Kyong S Yoon, Joseph P Strycharz, Maria F Unger, Scott Christley, Neil F Lobo, Manfredo J Seufferheld, Naikuan Wang, Gregory A Dasch, Claudio J Struchiner, Greg Madey, Linda I Hannick, Shelby Bidwell, Vinita Joardar, Elisabet Caler, Renfu Shao, Stephen C Barker, Stephen Cameron, Robert V Bruggner, Allison Regier, Justin Johnson, Lakshmi Viswanathan, Terry R Utterback, Granger G Sutton, Daniel Lawson, Robert M Waterhouse, J Craig Venter, Robert L Strausberg, May R Berenbaum, Frank H Collins, Evgeny M Zdobnov, and Barry R Pittendrigh. Genome sequences of the human body louse and its primary endosymbiont provide insights into the permanent parasitic lifestyle. *Proceedings of the National Academy of Sciences of the United States of America*, 107(27):12168–73, July 2010.

[65] Jan O Korbel, Alexej Abyzov, Xinmeng Jasmine Mu, Nicholas Carriero, Philip Cayting, Zhengdong Zhang, Michael Snyder, and Mark B Gerstein. PEMer: a computational framework with simulation-based error models for inferring genomic structural variants from massive paired-end sequencing data. *Genome biology*, 10:R23, 2009.

[66] Jan O Korbel, Alexander Eckehart Urban, Jason P Affourtit, Brian Godwin, Fabian Grubert, Jan Fredrik Simons, Philip M Kim, Dean Palejev, Nicholas J Carriero, Lei Du, Bruce E Taillon, Zhoutao Chen, Andrea Tanzer, A C Eugenia Saunders, Jianxiang Chi, Fengtang Yang, Nigel P Carter, Matthew E Hurles, Sherman M Weissman, Timothy T Harkins, Mark B Ger-

stein, Michael Egholm, and Michael Snyder. Paired-end mapping reveals extensive structural variation in the human genome. *Science (New York, N.Y.)*, 318(5849):420–6, October 2007.

[67] Paula X Kover, William Valdar, Joseph Trakalo, Nora Scarcelli, Ian M Ehrenreich, Michael D Purugganan, Caroline Durrant, and Richard Mott. A Multiparent Advanced Generation Inter-Cross to Fine-Map Quantitative Traits in Arabidopsis thaliana. *PLoS Genetics*, 5(7):15, 2009.

[68] Alvina G Lai, Matthew Denton-Giles, Bernd Mueller-Roeber, Jos H M Schippers, and Paul P Dijkwel. Positional information resolves structural variations and uncovers an evolutionarily divergent genetic locus in accessions of Arabidopsis thaliana. *Genome biology and evolution*, 3:627–40, January 2011.

[69] Sandy Y Lam, Sarah R Horn, Sarah J Radford, Elizabeth A Housworth, Franklin W Stahl, and Gregory P Copenhaver. Crossover interference on nucleolus organizing region-bearing chromosomes in Arabidopsis. *Genetics*, 170(2):807–12, June 2005.

[70] Ben Langmead and Steven L Salzberg. Fast gapped-read alignment with Bowtie 2, 2012.

[71] Daniel John Lawson, Garrett Hellenthal, Simon Myers, and Daniel Falush. Inference of population structure using dense haplotype data. *PLoS genetics*, 8(1):e1002453, January 2012.

[72] Wan-Ping Lee, Michael P Stromberg, Alistair Ward, Chip Stewart, Erik P Garrison, and Gabor T Marth. MOSAIK: a hash-based algorithm for accurate next-generation sequencing short-read mapping. *PloS one*, 9(3):e90581, January 2014.

[73] Qiang Leng, Richard W Mercier, Bao-Guang Hua, Hillel Fromm, and Gerald A Berkowitz. Electrophysiological analysis of cloned cyclic nucleotide-gated ion channels. *Plant physiology*, 128(2):400–10, February 2002.

[74] H Li, B Handsaker, A Wysoker, and T Fennell. The sequence alignment/map format and SAMtools. *Bioinformatics*, 25(16):2078–2079, 2009.

[75] Heng Li and Richard Durbin. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics*, 26:589–595, 2010.

[76] Heng Li, Jue Ruan, and Richard Durbin. Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome research*, 18:1851–1858, 2008.

[77] J Y Li, F Gaillard, A Moreau, J L Harousseau, C Laboisse, N Milpied, R Bataille, and H Avet-Loiseau. Detection of translocation t(11;14)(q13;q32) in mantle cell lymphoma by fluorescence in situ hybridization. *The American journal of pathology*, 154(5):1449–52, May 1999.

[78] Ruiqiang Li, Chang Yu, Yingrui Li, Tak-Wah Lam, Siu-Ming Yiu, Karsten Kristiansen, and Jun Wang. SOAP2: an improved ultrafast tool for short read alignment. *Bioinformatics*, 25(15):1966–7, 2009.

[79] M Lichten and A S Goldman. Meiotic recombination hotspots. *Annual Review of Genetics*, 29:423–444, 1995.

[80] Xiangtao Liu, Shizhong Han, Zuoheng Wang, Joel Gelernter, and Bao Zhu Yang. Variant Callers for Next-Generation Sequencing Data: A Comparison Study. *PLoS ONE*, 8, 2013.

[81] Quan Long, Fernando A Rabanal, Dazhe Meng, Christian D Huber, Ashley Farlow, Alexander Platzer, Qingrun Zhang, Bjarni J Vilhjálmsson, Arthur Korte, Viktoria Nizhynska, Viktor Voronin, Pamela Korte, Laura Sedman, Terezie Mandáková, Martin A Lysak, Umit Seren, Ines Hellmann, and Magnus Nordborg. Massive genomic variation and strong selection in Arabidopsis thaliana lines from Sweden. *Nature genetics*, (November 2012), 2013.

[82] Jan M Lucht, Brigitte Mauch-Mani, Henry-York Steiner, Jean-Pierre Metraux, John Ryals, and Barbara Hohn. Pathogen stress increases somatic recombination frequency in Arabidopsis. *Nature genetics*, 30(3):311–4, March 2002.

[83] Gerton Lunter and Martin Goodson. Stampy: a statistical algorithm for sensitive and fast mapping of Illumina sequence reads. *Genome Research*, 21(6):936–939, 2011.

[84] Ruibang Luo, Binghang Liu, Yinlong Xie, Zhenyu Li, Weihua Huang, Jianying Yuan, Guangzhu He, Yanxiang Chen, Qi Pan, Yunjie Liu, Jingbo Tang, Gengxiong Wu, Hao Zhang, Yujian Shi, Yong Liu, Chang Yu, Bo Wang, Yao Lu, Changlei Han, David W Cheung, Siu-Ming Yiu, Shaoliang Peng, Zhu Xiaoqian, Guangming Liu, Xiangke Liao, Yingrui Li, Huanming Yang, Jian Wang, Tak-Wah Lam, and Jun Wang. SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. *GigaScience*, 1(1):18, January 2012.

[85] Jeffrey R. MacDonald, Robert Ziman, Ryan K C Yuen, Lars Feuk, and Stephen W. Scherer. The Database of Genomic Variants: A curated collection of structural variation in the human genome. *Nucleic Acids Research*, 42, 2014.

[86] Christopher A Maher, Chandan Kumar-Sinha, Xuhong Cao, Shanker Kalyana-Sundaram, Bo Han, Xiaojun Jing, Lee Sam, Terrence Barrette, Nallasivam Palanisamy, and Arul M Chinnaiyan. Transcriptome sequencing to detect gene fusions in cancer. *Nature*, 458:97–101, 2009.

[87] Jonathan Marchini, Bryan Howie, Simon Myers, Gil McVean, and Peter Donnelly. A new multipoint method for genome-wide association studies by imputation of genotypes. *Nature genetics*, 39(7):906–13, July 2007.

[88] Elaine R Mardis. Next-generation DNA sequencing methods. *Annual Review of Genomics and Human Genetics*, 9(1):387–402, 2008.

[89] Aaron McKenna, Matthew Hanna, Eric Banks, Andrey Sivachenko, Kristian Cibulskis, Andrew Kernytsky, Kiran Garimella, David Altshuler, Stacey Gabriel, Mark Daly, and Mark A DePristo. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome research*, 20:1297–1303, 2010.

[90] Gil McVean and Simon Myers. PRDM9 marks the spot. *Nature genetics*, 42(10):821–2, October 2010.

[91] Jenelle D F Meyer, Danielle C G Silva, Chunling Yang, Kerry F Pedley, Chunquan Zhang, Martijn van de Mortel, John H Hill, Randy C Shoemaker, Ricardo V Abdelnoor, Steven A

Whitham, and Michelle A Graham. Identification and analyses of candidate genes for rpp4-mediated resistance to Asian soybean rust in soybean. *Plant physiology*, 150(1):295–307, May 2009.

[92] Ryan E. Mills, Christopher T. Luttig, Christine E. Larkins, Adam Beauchamp, Circe Tsui, W. Stephen Pittard, and Scott E. Devine. An initial map of insertion and deletion (INDEL) variation in the human genome. *Genome Research*, 16:1182–1190, 2006.

[93] Ryan E Mills, Klaudia Walter, Chip Stewart, Robert E Handsaker, Ken Chen, Can Alkan, Alexej Abyzov, Seungtai Chris Yoon, Kai Ye, R Keira Cheetham, Asif Chinwalla, Donald F Conrad, Yutao Fu, Fabian Grubert, Iman Hajirasouliha, Fereydoun Hormozdiari, Lilia M Iakoucheva, Zamin Iqbal, Shuli Kang, Jeffrey M Kidd, Miriam K Konkel, Joshua Korn, Ekta Khurana, Deniz Kural, Hugo Y K Lam, Jing Leng, Ruiqiang Li, Yingrui Li, Chang-Yun Lin, Ruibang Luo, Xinmeng Jasmine Mu, James Nemesh, Heather E Peckham, Tobias Rausch, Aylwyn Scally, Xinghua Shi, Michael P Stromberg, Adrian M Stütz, Alexander Eckehart Urban, Jerilyn A Walker, Jiantao Wu, Yujun Zhang, Zhengdong D Zhang, Mark A Batzer, Li Ding, Gabor T Marth, Gil McVean, Jonathan Sebat, Michael Snyder, Jun Wang, Kenny Ye, Evan E Eichler, Mark B Gerstein, Matthew E Hurles, Charles Lee, Steven A McCarroll, and Jan O Korbel. Mapping copy number variation by population-scale genome sequencing. *Nature*, 470(7332):59–65, February 2011.

[94] R Mott, C J Talbot, M G Turri, A C Collins, and J Flint. A method for fine mapping quantitative trait loci in outbred animal stocks. *Proceedings of the National Academy of Sciences of the United States of America*, 97(23):12649–54., 2000.

[95] Simon Myers, Rory Bowden, Afidalina Tumian, Ronald E Bontrop, Colin Freeman, Tammie S MacFie, Gil McVean, and Peter Donnelly. Drive against hotspot motifs in primates implicates the PRDM9 gene in meiotic recombination. *Science (New York, N.Y.)*, 327(5967):876–9, February 2010.

[96] Simon Myers, Colin Freeman, Adam Auton, Peter Donnelly, and Gil McVean. A common sequence motif associated with recombination hot spots and genome instability in humans. *Nature genetics*, 40(9):1124–9, September 2008.

[97] Robert K. Neely, Jochem Deen, and Johan Hofkens. Optical mapping of DNA: Single-molecule-based methods for mapping genomes. *Biopolymers*, 95:298–311, 2011.

[98] Rasmus Nielsen, Joshua S Paul, Anders Albrechtsen, and Yun S Song. Genotype and SNP calling from next-generation sequencing data. *Nature reviews. Genetics*, 12:443–451, 2011.

[99] Bogdan Pasaniuc, Nadin Rohland, Paul J McLaren, Kiran Garimella, Noah Zaitlen, Heng Li, Namrata Gupta, Benjamin M Neale, Mark J Daly, Pamela Sklar, Patrick F Sullivan, Sarah Bergen, Jennifer L Moran, Christina M Hultman, Paul Lichtenstein, Patrik Magnusson, Shaun M Purcell, David W Haas, Liming Liang, Shamil Sunyaev, Nick Patterson, Paul I W de Bakker, David Reich, and Alkes L Price. Extremely low-coverage sequencing and imputation increases power for genome-wide association studies. *Nature Genetics*, 44:631–635, 2012.

[100] Petko M. Petkov, Karl W. Broman, Jin P. Szatkiewicz, and Kenneth Paigen. Crossover interference underlies sex differences in recombination rates. *Trends in Genetics*, 23:539–542, 2007.

[101] Dalila Pinto, Elsa Delaby, Daniele Merico, Mafalda Barbosa, Alison Merikangas, Lambertus Klei, Bhooma Thiruvahindrapuram, Xiao Xu, Robert Ziman, Zhuozhi Wang, Jacob A S Vorstman, Ann Thompson, Regina Regan, Marion Pilorge, Giovanna Pellecchia, Alistair T. Pagnamenta, Bárbara Oliveira, Christian R. Marshall, Tiago R. Magalhaes, Jennifer K. Lowe, Jennifer L. Howe, Anthony J. Griswold, John Gilbert, Eftichia Duketis, Beth A. Dombroski, Maretha V. De Jonge, Michael Cuccaro, Emily L. Crawford, Catarina T. Correia, Judith Conroy, Inês C. Conceição, Andreas G. Chiocchetti, Jillian P. Casey, Guiqing Cai, Christelle Cabrol, Nadia Bolshakova, Elena Bacchelli, Richard Anney, Steven Gallinger, Michelle Cotterchio, Graham Casey, Lonnie Zwaigenbaum, Kerstin Wittemeyer, Kirsty Wing, Simon Wallace, Herman Van Engeland, Ana Tryfon, Susanne Thomson, Latha Soorya, Bernadette

Rogé, Wendy Roberts, Fritz Poustka, Susana Mouga, Nancy Minshew, L. Alison McInnes, Susan G. McGrew, Catherine Lord, Marion Leboyer, Ann S. Le Couteur, Alexander Kolevzon, Patricia Jiménez González, Suma Jacob, Richard Holt, Stephen Guter, Jonathan Green, Andrew Green, Christopher Gillberg, Bridget A. Fernandez, Frederico Duque, Richard Delorme, Geraldine Dawson, Pauline Chaste, Cátia Café, Sean Brennan, Thomas Bourgeron, Patrick F. Bolton, Sven Bölte, Raphael Bernier, Gillian Baird, Anthony J. Bailey, Evdokia Anagnostou, Joana Almeida, Ellen M. Wijsman, Veronica J. Vieland, Astrid M. Vicente, Gerard D. Schellenberg, Margaret Pericak-Vance, Andrew D. Paterson, Jeremy R. Parr, Guiomar Oliveira, John I. Nurnberger, Anthony P. Monaco, Elena Maestrini, Sabine M. Klauck, Hakon Hakonarson, Jonathan L. Haines, Daniel H. Geschwind, Christine M. Freitag, Susan E. Folstein, Sean Ennis, Hilary Coon, Agatino Battaglia, Peter Szatmari, James S. Sutcliffe, Joachim Hallmayer, Michael Gill, Edwin H. Cook, Joseph D. Buxbaum, Bernie Devlin, Louise Gallagher, Catalina Betancur, and Stephen W. Scherer. Convergence of genes and cellular pathways dysregulated in autism spectrum disorders. *American Journal of Human Genetics*, 94:677–694, 2014.

[102] Hannes Ponstingl and Zemin Ning. SMALT - A new mapper for DNA Sequencing Reads. 2013.

[103] Shaun M Purcell, Jennifer L Moran, Menachem Fromer, Douglas Ruderfer, Nadia Solovieff, Panos Roussos, Colm O'Dushlaine, Kimberly Chambert, Sarah E Bergen, Anna Kähler, Laramie Duncan, Eli Stahl, Giulio Genovese, Esperanza Fernández, Mark O Collins, Noboru H Komiyama, Jyoti S Choudhary, Patrik K E Magnusson, Eric Banks, Khalid Shakir, Kiran Garimella, Tim Fennell, Mark DePristo, Seth G N Grant, Stephen J Haggarty, Stacey Gabriel, Edward M Scolnick, Eric S Lander, Christina M Hultman, Patrick F Sullivan, Steven a McCarroll, and Pamela Sklar. A polygenic burden of rare disruptive mutations in schizophrenia. *Nature*, 506:185–90, 2014.

[104] L. R. Rabiner and B. H. Juang. An introduction to hidden Markov models. *IEEE ASSP Magazine*, 3:4–16, 1986.

[105] Jeroen Raes, Antje Rohde, Jørgen Holst Christensen, Yves Van de Peer, and Wout Boerjan. Genome-wide characterization of the lignification toolbox in Arabidopsis. *Plant physiology*, 133(3):1051–71, November 2003.

[106] Francisco J Redondo, Teodoro Coba de la Peña, M Mercedes Lucas, and José J Pueyo. Alfalfa nodules elicited by a flavodoxin-overexpressing Ensifer meliloti strain display nitrogen-fixing activity with enhanced tolerance to salinity stress. *Planta*, 236(6):1687–700, December 2012.

[107] Estelle Remy, Tânia R Cabrito, Rita A Batista, Miguel C Teixeira, Isabel Sá-Correia, and Paula Duque. The major facilitator superfamily transporter ZIFL2 modulates cesium and potassium homeostasis in Arabidopsis. *Plant & cell physiology*, 56(1):148–62, January 2015.

[108] The 3000 rice genomes project. The 3,000 rice genomes project. *GigaScience*, 3(1):7, January 2014.

[109] Lars Rönnegå rd and William Valdar. Recent developments in statistical methods for detecting genetic loci affecting phenotypic variability. *BMC Genetics*, 13(1):63, 2012.

[110] S Rozen and H Skaletsky. Primer3 on the WWW for general users and for biologist programmers. *Methods in molecular biology (Clifton, N.J.)*, 132:365–386, 2000.

[111] Stephen W Scherer, Charles Lee, Ewan Birney, David M Altshuler, Evan E Eichler, Nigel P Carter, Matthew E Hurles, and Lars Feuk. Challenges and standards in integrating surveys of structural variation. *Nature genetics*, 39:S7–S15, 2007.

[112] Mohammad R Seyedsayamdost, Rebecca J Case, Roberto Kolter, and Jon Clardy. The Jekyll-and-Hyde chemistry of Phaeobacter gallaeciensis. *Nature chemistry*, 3(4):331–5, April 2011.

[113] L Signora, I De Smet, C H Foyer, and H Zhang. ABA plays a central role in mediating the regulatory effects of nitrate on root branching in Arabidopsis. *The Plant journal : for cell and molecular biology*, 28(6):655–62, December 2001.

[114] Jared T Simpson, Rebecca E McIntyre, David J Adams, and Richard Durbin. Copy number variant detection in inbred strains from short read sequence data. *Bioinformatics (Oxford, England)*, 26(4):565–7, February 2010.

[115] Suzanne Sindi, Elena Helman, Ali Bashir, and Benjamin J Raphael. A geometric approach for classification and comparison of structural variants. *Bioinformatics (Oxford, England)*, 25:i222–i230, 2009.

[116] Dimitry Yu Sorokin, Vladimir M Gorlenko, Tat'yana P Tourova, Alexandre I Tsapin, Kenneth H Nealson, and Gijs J Kuenen. Thioalkalimicrobium cyclicum sp. nov. and Thioalkalivibrio jannaschii sp. nov., novel species of haloalkaliphilic, obligately chemolithoautotrophic sulfur-oxidizing bacteria from hypersaline alkaline Mono Lake (California). *International journal of systematic and evolutionary microbiology*, 52(Pt 3):913–20, May 2002.

[117] Nikos C Kyrpides Stefan Spring, Carmen Scheuner, Alla Lapidus, Susan Lucas, Tijana Glavina del Rio, Hope Tice, Alex Copeland, Jan-Fang Cheng, Feng Chen, Matt Nolan, Elizabeth Saunders, Sam Pitluck, Konstantinos Liolios, Natalia Ivanova, Konstantinos Mavromatis, Athanasios and Hans-Peter Klenk. Frontiers — The genome sequence of Methanohalophilus mahii SLP(T) reveals differences in the energy metabolism among members of the Methanosarcinaceae inhabiting freshwater and saline environments. *Archaea*, 2010, 2010.

[118] Philip J Stephens, David J McBride, Meng-Lay Lin, Ignacio Varela, Erin D Pleasance, Jared T Simpson, Lucy A Stebbings, Catherine Leroy, Sarah Edkins, Laura J Mudie, Chris D Greenman, Mingming Jia, Calli Latimer, Jon W Teague, King Wai Lau, John Burton, Michael A Quail, Harold Swerdlow, Carol Churcher, Rachael Natrajan, Anieta M Sieuwerts, John W M Martens, Daniel P Silver, Anita Langerø d, Hege E G Russnes, John A Foekens, Jorge S Reis-Filho, Laura van 't Veer, Andrea L Richardson, Anne-Lise Bø rresen Dale, Peter J Campbell, P Andrew Futreal, and Michael R Stratton. Complex landscapes of somatic rearrangement in human breast cancer genomes. *Nature*, 462(7276):1005–10, December 2009.

[119] David Stoddart, Andrew J Heron, Ellina Mikhailova, Giovanni Maglia, and Hagan Bayley. Single-nucleotide discrimination in immobilized DNA oligonucleotides with a biological nanopore. *Proceedings of the National Academy of Sciences of the United States of America*, 106(19):7702–7, May 2009.

[120] Karen L Svenson, Daniel M Gatti, William Valdar, Catherine E Welsh, Riyan Cheng, Elissa J Chesler, Abraham A Palmer, Leonard McMillan, and Gary A Churchill. High-resolution genetic mapping using the Mouse Diversity outbred population. *Genetics*, 190(2):437–47, February 2012.

[121] Dingzhong Tang, Jules Ade, Catherine A Frye, and Roger W Innes. Regulation of plant defense responses in Arabidopsis by EDR2, a PH and START domain-containing protein. *The Plant journal : for cell and molecular biology*, 44(2):245–57, October 2005.

[122] The International HapMap Consortium. A haplotype map of the human genome. *Nature*, 437(7063):1299–320, October 2005.

[123] Helga Thorvaldsdóttir, James T. Robinson, and Jill P. Mesirov. Integrative Genomics Viewer (IGV): High-performance genomics data visualization and exploration. *Briefings in Bioinformatics*, 14(2):178–192, 2013.

[124] Feng Tian, Peter J Bradbury, Patrick J Brown, Hsiaoyi Hung, Qi Sun, Sherry Flint-Garcia, Torbert R Rocheford, Michael D McMullen, James B Holland, and Edward S Buckler. Genome-wide association study of leaf architecture in the maize nested association mapping population. *Nature genetics*, 43(2):159–162, 2011.

[125] Pablo Tornero and Jeffery L. Dangl. A high-throughput method for quantifying growth of phytopathogenic bacteria in Arabidopsis thaliana. *The Plant Journal*, 28(4):475–481, January 2002.

[126] Atsuko Ueki, Hiroshi Akasaka, Daisuke Suzuki, and Katsuji Ueki. Paludibacter propionicigenes, a novel strictly anaerobic, Gram-negative, propionate-producing bacterium isolated

from plant residue in irrigated rice-field soil in Japan. *International journal of systematic and evolutionary microbiology*, 56(Pt 1):39–44, January 2006.

[127] William Valdar, Jonathan Flint, and Richard Mott. Simulating the Collaborative Cross: Power of Quantitative Trait Loci Detection and Mapping Resolution in Large Sets of Recombinant Inbred Strains of Mice. *Genetics*, 172(3):1783–1797, 2006.

[128] William S J Valdar, Jonathan Flint, and Richard Mott. QTL fine-mapping with recombinant-inbred heterogeneous stocks and in vitro heterogeneous stocks. *Mammalian genome : official journal of the International Mammalian Genome Society*, 14:830–838, 2003.

[129] Erik A van der Biezen, Cecilie T Freddie, Katherine Kahn, Jane E Parker, and Jonathan D G Jones. Arabidopsis RPP4 is a member of the RPP5 multigene family of TIR-NB-LRR genes and confers downy mildew resistance through multiple signalling components. *The Plant journal : for cell and molecular biology*, 29(4):439–51, February 2002.

[130] Matthew W. Vaughn, Miloš Tanurdžić, Zachary Lippman, Hongmei Jiang, Robert Carrasquillo, Pablo D. Rabinowicz, Neilay Dedhia, W. Richard McCombie, Nicolas Agier, Agnès Bulski, Vincent Colot, R. W. Doerge, and Robert A. Martienssen. Epigenetic natural variation in Arabidopsis thaliana. *PLoS Biology*, 5:1617–1629, 2007.

[131] T J Vision, D G Brown, and S D Tanksley. The origins of genomic duplications in Arabidopsis. *Science*, 290(5499):2114–2117, 2000.

[132] Erik Wijnker, Geo Velikkakam James, Jia Ding, Frank Becker, Jonas R Klasen, Vimal Rawat, Beth A Rowan, Daniël F de Jong, C Bastiaan de Snoo, Luis Zapata, Bruno Huettel, Hans de Jong, Stephan Ossowski, Detlef Weigel, Maarten Koornneef, Joost Jb Keurentjes, and Korbinian Schneeberger. The genomic landscape of meiotic crossovers and gene conversions in Arabidopsis thaliana. *eLife*, 2:e01426, January 2013.

[133] Kim Wong, Thomas M Keane, James Stalker, and David J Adams. Enhanced structural variant and breakpoint detection using SVMerge by integration of multiple detection methods and local assembly. *Genome biology*, 11(12):R128, January 2010.

[134] Lai Ping Wong, Rick Twee Hee Ong, Wan Ting Poh, Xuanyao Liu, Peng Chen, Ruoying Li, Kevin Koi Yau Lam, Nisha Esakimuthu Pillai, Kar Seng Sim, Haiyan Xu, Ngak Leng Sim, Shu Mei Teo, Jia Nee Foo, Linda Wei Lin Tan, Yenly Lim, Seok Hwee Koo, Linda Seo Hwee Gan, Ching Yu Cheng, Sharon Wee, Eric Peng Huat Yap, Pauline Crystal Ng, Wei Yen Lim, Richie Soong, Markus Rene Wenk, Tin Aung, Tien Yin Wong, Chiea Chuen Khor, Peter Little, Kee Seng Chia, and Yik Ying Teo. Deep whole-genome sequencing of 100 southeast Asian malays. *American Journal of Human Genetics*, 92:52–66, 2013.

[135] P C Y Woo, S K P Lau, J L L Teng, H Tse, and K-Y Yuen. Then and now: use of 16S rDNA gene sequencing for bacterial identification and discovery of novel bacteria in clinical microbiology laboratories. *Clinical microbiology and infection : the official publication of the European Society of Clinical Microbiology and Infectious Diseases*, 14(10):908–34, October 2008.

[136] Thomas Wu and Burrows Wheeler. Suffix arrays and the Burrows-Wheeler Transform Motivation with. *Bioinformatics*, 2009.

[137] Binnaz Yalcin, Kim Wong, Avigail Agam, Martin Goodson, Thomas M Keane, Xiangchao Gan, Christoffer Nellå ker, Leo Goodstadt, Jérôme Nicod, Amarjit Bhomra, Polinka Hernandez-Pliego, Helen Whitley, James Cleak, Rebekah Dutton, Deborah Janowitz, Richard Mott, David J Adams, and Jonathan Flint. Sequence-based characterization of structural variation in the mouse genome. *Nature*, 477(7364):326–9, September 2011.

[138] Sihai Yang, Yang Yuan, Long Wang, Jing Li, Wen Wang, Haoxuan Liu, Jian-Qun Chen, Laurence D Hurst, and Dacheng Tian. Great majority of recombination events in Arabidopsis are gene conversion events. *Proceedings of the National Academy of Sciences of the United States of America*, 109(51):20992–7, December 2012.

[139] Kai Ye, Marcel H Schulz, Quan Long, Rolf Apweiler, and Zemin Ning. Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. *Bioinformatics (Oxford, England)*, 25(21):2865–71, November 2009.

[140] N E Yelina, K Choi, L Chelysheva, M Macaulay, B de Snoo, E Wijnker, N Miller, J Drouaud, M Grelon, G P Copenhaver, C Mezard, K A Kelly, and I R Henderson. Epigenetic Remodeling of Meiotic Crossover Frequency in Arabidopsis thaliana DNA Methyltransferase Mutants. *Plos Genetics*, 8(8):e1002844, 2012.

[141] Daniel R Zerbino and Ewan Birney. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome research*, 18(5):821–9, May 2008.

[142] Thomas Zichner, David A. Garfield, Tobias Rausch, Adrian M. Stütz, Enrico Cannavo, Martina Braun, Eileen E M Furlong, and Jan O. Korbel. Impact of genomic structural variation in Drosophila melanogaster based on population-scale sequencing. *Genome Research*, 23:568–579, 2013.