Detection of Inversion Polymorphisms in the Human Genome

Elísabet Linda Þórðardóttir

Final paper in MA–degree in anthropology

School of Social Sciences

HÁSKÓLI ÍSLANDS

Detection of Inversion Polymorphisms in the Human Genome

Elísabet Linda Þórðardóttir

Final paper in MA–degree in anthropology

Supervisor: Agnar Helgason

Faculty of Social and Human Sciences

School of Social Sciences

October 2016

Reykjavík, Iceland 2016

Abstract

While genome variation such as single nucleotide polymorphisms (SNPs), insertions and deletions (indels) is well characterized in humans, little is known about the number of inversion polymorphisms. In heterozygotes for different orientations of inversions, it is thought that the normal pairing of homologous chromosomes is impeded during meiotic cell division and thereby leading to a local suppression of recombination. This, in turn, leads to an accumulation of orientation specific mutations over time. The objective of this study was to devise and implement a novel method to detect inversion polymorphisms using dense microarray genotype data available at deCODE Genetics from 39,616 parents, for whom 2.4 million recombination events have been inferred based on meiotic transmission to 79,708 genotyped offspring. The aim was to discover inversions through the application of a novel test, wherein local recombination rates are compared between homozygotes and heterozygotes of potential tagging SNPs, such that a statistically significant suppression of recombination in the latter group is taken as evidence for the presence of an inversion. Our results demonstrate that this test has considerable specificity, detecting common and known inversions. They also indicate that no other inversions of comparable size and frequency are found in the Icelandic gene pool. However, a handful of candidates for smaller and/or less frequent inversions were detected.

Preface

This thesis outlines the results of a research-based Master's project in biological anthropology conducted at the University of Iceland and deCODE Genetics in the years 2014-2016. The study was supervised by Dr. Agnar Helgason. The project and the thesis accounted for 90 ECTS of 120 ECTS units required to obtain a MA degree in Anthropology from the University of Iceland.

Acknowledgements

First, I would like to thank my supervisor, Agnar Helgason. His support and guidance during my studies have gone above and beyond his duty. I also like to thank my team at deCODE Genetics, Anuradha, Ellen, Margrét, Sunna and Valdís, for their invaluable support, both professionally and personally. Finally, I want to thank my family, Darri, Snorri, Auður Ýr, Marinó and Elísabet Ylfa. Your encouragement and understanding have made this possible.

Table of Contents

List of Figures

List of Tables

1 Introduction

Inversions are one of several types of structural variations that have been discovered in genomes. With the advance of DNA sequencing technologies, it has become clear that structural variation in the human genome is far more extensive than previously thought (Alves et al., 2012). A few large inversions are known to be polymorphic in humans. However, due to their balanced nature, no reliable high-throughput genotyping assay provides the means to detect inversions in the genomes of humans or other organisms. Therefore, while genome variation attributable to SNPs and insertions and deletions (indels) is well characterized, the prevalence of inversions in the human genome is not known.

The objective of this project is to devise and implement a novel *in silico* method to detect inversion polymorphisms using dense genotype data and information about the location of recombination events in a large set of individuals. Our approach makes use of the expectation that polymorphic inversions will be tagged by sequence variants and that heterozygotes for different orientations of an inversion (and thereby of the tagging SNPs) experience diminished recombination across the region spanned by it. We implemented a statistical test based on this principle, the Suppression of Heterozygote Recombination (SHR) test, using a vast resource of Illumina microarray genotype data from 39,616 distinct parents available at deCODE Genetics, for whom 2.4 million recombination events have been inferred based on meiotic transmissions to 79,708 genotyped children.

To our knowledge, data on recombination events have not before been used to infer the presence of inversion polymorphisms. Several studies have focused on linkage disequilibrium (LD) as an indication of suppressed recombination that might be due to inversions. The advantage of our approach is the availability of information about recombination events for each genotyped individual, which makes it possible to directly detect the impact of suppressed recombination caused by inversions in heterozygotes. With this method we are able to avoid

false positive signals due to selective sweeps or unusually low region specific recombination rates that are detected by methods based on LD in population data.

The thesis is organized into five chapters, starting with this introduction and followed by a chapter on the state of knowledge on the subject of inversions and their formation, their impact and role in speciation. This chapter furthermore gives a short overview of known polymorphic inversions and of methods that have been developed for detecting inversions. Chapter 3 describes the methods applied in this study, while chapter 4 presents the results and compares our findings to previous knowledge of polymorphic inversions. Finally, chapter 6 summarizes the main conclusions of the study and their implications.

2 Background

Inversions are one of several types of structural variations (SVs) that are found in genomes. SVs are defined as changes in the genome involving contiguous segments of variable sizes. These segments are in some cases microscopically visible (>3 Mb), but are mostly submicroscopic (~50 bp to 3 Mb) (Feuk et al., 2006; Sudmant et al., 2015). Due to these microscopically visible variations, the existence of SVs has been known for decades. However, as these large variants are rare, SVs in general were also thought to be relatively rare (Escaramís, Docampo and Rabionet, 2015). With the advance of DNA sequencing technologies, it has become clear that structural variation in the human genome is far more extensive than previously thought. In addition to single nucleotide polymorphisms (SNPs), which were the main subject of interest at the onset of whole genome sequencing (WGS), duplications, deletions, insertions, copy number variations (CNVs), translocations, and inversions have now become a more prominent concern in human genome studies (Alves et al., 2012).

Structural variations are categorized as either balanced or unbalanced. Unbalanced variations affect the size of the genome, that is, the carrier has either increased or decreased length of sequence when compared to a non-carrier. These include CNVs, duplications and deletions. Because of this effect on sequence length these variations are relatively easy to detect with current technology, such as array-based methods (Feuk et al., 2006). Balanced variations however, do not alter DNA sequence length, but rather just the position or orientation of DNA fragments in the genome. Few reliable methods exist at present to detect such variations, which include inversions and some translocations (Feuk et al., 2006).

2.1 Inversions

Inversions are stretches of sequence that are present in different orientations on homologous chromosomes. They can arise as a result of erroneous repair of a double strand breakage of DNA. This can happen in the course of chromosomal recombination in meiosis, where

programmed double strand breaks (DSB) play an essential role in promoting meiotic recombination. DSBs may also occur spontaneously at any time, for example because of radiation, replication across a nick on one DNA strand, or enzyme malfunction (Lieber et al., 2003). In humans, DSB are predominantly repaired by non-homologous end-joining (NHEJ), a mechanism where the broken ends are simply rejoined, causing a loss of one or more nucleotides. During S and G_2 phase of cell division, however, sister-chromatids, and, in the case of meiosis, homologs, are available as templates for error-free repair of DSBs (Lieber et al., 2003). The use of homologs for DSB repair, whether programmed or due to damage, is termed homologous recombination (HR) and takes place only between DNA sequences that have high similarity, although they need not be perfectly matched (Barzel and Kupiec, 2008).

While HR is considered an error-free repair mechanism as opposed to NHEJ, the prevalence of highly homologous repeats across the genome implies that the homology search may locate a paralogous sequence that, although highly similar, is not located at the same position in the genome. This is termed non-allelic homologous recombination (NAHR), and can cause various rearrangements, such as duplications, deletions and inversions, depending on the location and the orientation of the paralog. NAHR between repeats that are located on different chromatids and have the same orientation can cause either a duplication or a deletion of the intervening sequence, while NAHR between repeats on the same chromatid cause a deletion. If the repeats are inverted, however, it may result in an inversion of the intervening sequence (Figure 1) (Sharp et al., 2006; Parks et al., 2015). The majority of known large inversions in the human genome are thought to be due to NAHR (Feuk, 2010).

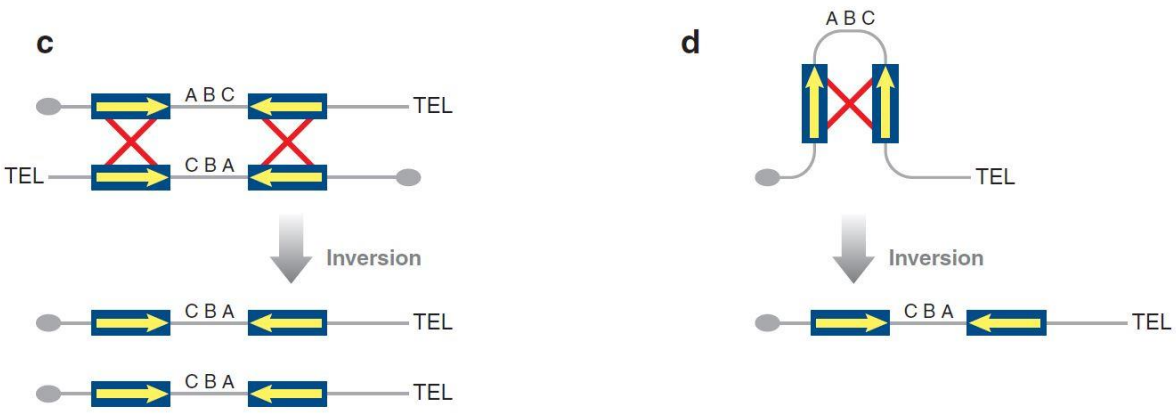

Figure 1. Non-allelic homologous recombination between inverted repeats result in an inversion of the intervening sequence. Inter-chromatid (left) and intra-chromatid (right). From Sharp et al., 2006.

2.2 Effects of inversions

2.2.1 Recombination suppression

It has been shown that recombination is locally suppressed in heterozygotes for orientations of inversion polymorphisms. This is due to the fact that during pairing in meiosis, the different orientations of the inversion induce a loop in one of the chromosomes. Crossing over events where the breakpoint falls within an inversion, can result in aneuploidy and thus non-viable zygotes (Andolfatto et al., 2001). Crossing over within paracentric inversions (i.e. inversions that do not encompass the centromere region), generates a dicentric bridge between homologous chromosomes, which breaks at a random position, and an acentric fragment, that will be lost (Figure 2, left). Such a meiosis will give rise to two normal, non-recombinant chromosomes, and two chromosomes containing deletions, whose sizes rely on the location of the crossover and the breakpoint of the bridge. If the inversion is pericentric, that is, the centromere lies within the inverted region, crossing over within the inversion does not cause a loss of fragments, but will result in two non-recombinant chromosomes, and two chromosomes, both containing a duplication and a deletion of sequence (Figure 2, right) (Griffiths et al., 2014). The effective suppression of recombination in heterozygotes for different orientations of inversions leads to the diversification of the orientations with time. As orientation specific mutations accumulate, they may further contribute to the obstruction of meiotic pairing in heterozygotes across the region spanning the inversion.

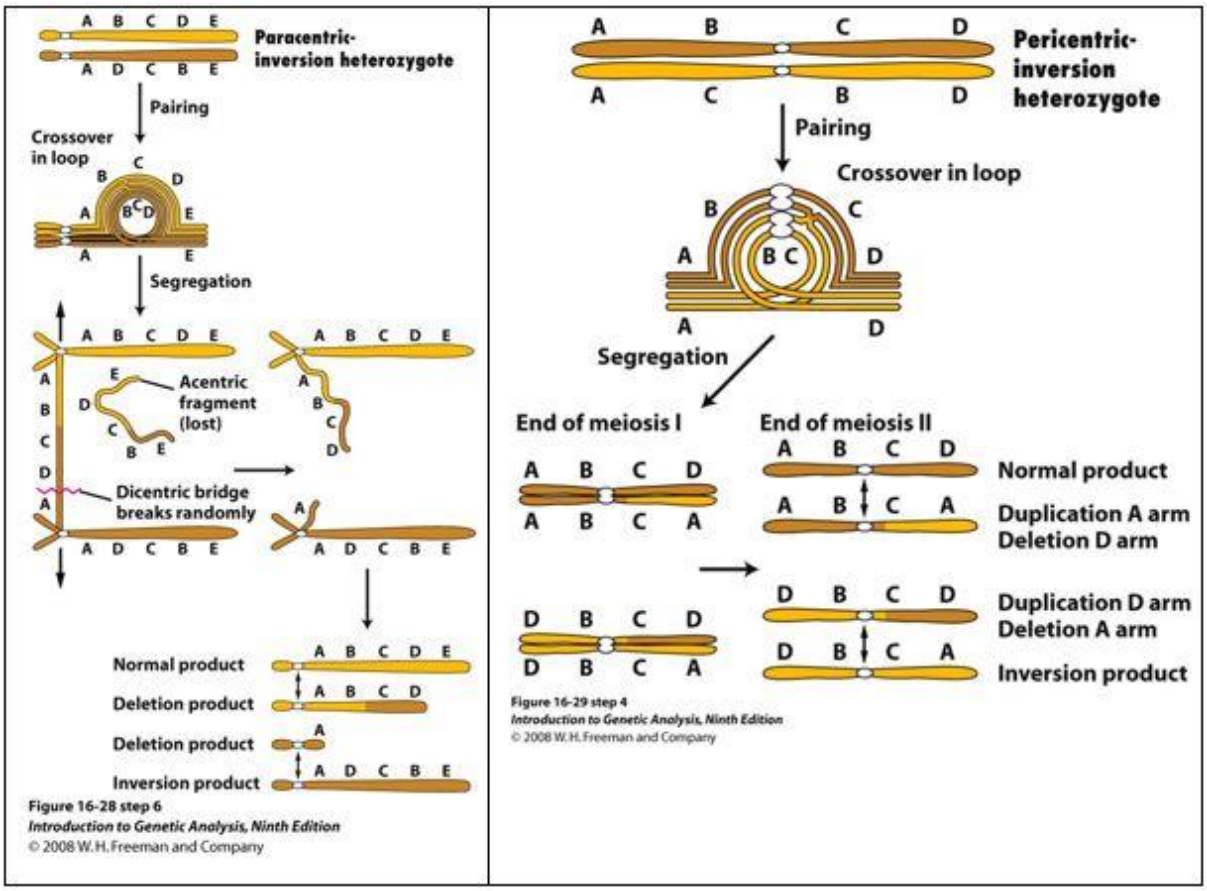

Figure 2. Result of crossovers within pericentric (left) and paracentric (right) inversions. From Griffiths et al., 2014.

2.2.2 Predisposition to other rearrangements

In addition to recombination suppression, since inversions are frequently flanked by segmental duplications, heterozygosity for inversion orientations may give rise to incorrect crossing over (NAHR) during meiosis, resulting in gain or loss of DNA in the zygote, and, as a consequence, a genetic disorder in the offspring of a carrier. For example, a study on patients with Williams-Beuren syndrome, which in most cases is caused by a deletion on chromosome 7, reported that an inversion was found in a parent of 33% of the patients (Osborne et al., 2001). Another study found that four out of six mothers of patients with deletions causing Angelman syndrome carried an inversion at that region, which was found in only 4.5% of the general population (Gimelli et al., 2003). Other examples of predisposition to rearrangements include a recurrent deletion associated with Sotos syndrome, and a recurrent translocation

involving the X and Y chromosomes, associated with sex reversal. These rearrangements have only been found on chromosomes that have an inversion at that particular region (Sharp, 2008).

2.2.3 Disruption of gene expression and phenotype association

Apart from the impact on pairing and segregation of chromosomes in meiosis and the gradual accumulation of orientation specific mutations, the inversions are themselves thought to be mostly phenotypically silent. That is, they are not thought to affect the phenotype of the carrier, unless the inversion breaks fall within genes and hence disrupt gene function or alter expression. Inversions that affect gene expression have been found in isolated cases or restricted to one family (Puig et al., 2015). An example is a case report from 2013 (Jones et al., 2013) that described a pericentric inversion disrupting the AP3B1 gene on chromosome 5, causing Hermansky-Pudlak syndrome. The patient was homozygous for the inversion, which was found in a heterozygous state in both parents, who were related. There have, however, been reports of recurrent disease-causing inversion events. Perhaps the best studied example is a recurrent 0.6 Mb inversion, of which one breakpoint is within intron 22 of the factor VIII gene on the X chromosome, causing a truncated transcript of the gene. These events explain about 40-45% of severe cases of Haemophilia A, which has a prevalence of around 1 in 5,000 male births, around half of which are classified as severe (Peyvandi, Garagiola and Young, 2016). Another example, also an X-linked disorder, is Hunter syndrome. It has been shown that around 13% of individuals with this syndrome carry an inversion in intron 7 of the IDS gene, caused by a recombination event between a segment of the gene and its pseudogene, resulting in a truncation of the gene (Lagerstedt et al., 1997).

Although intergenic inversions are unlikely to directly affect the phenotype of the carrier, there are reports of associations between such inversions and phenotypes. One study shows, for example, that a 0.45 Mb large inversion on chromosome 16, which is found at a frequency ranging from 10% in Africa to 49% in Northern Europe, is associatied with reduced risk of joint asthma and obesity (Gonzáles et al., 2014). There have also been reports of an association between the orientation of a common 4.5 Mb inversion on chromomsome 8, with frequency up to 70% in Africa, and risk of lupus in populations of European ancestry (Namjou et al., 2014;

Salm et al., 2012). The cause for such associations is likely to be found in the combination of alleles on the background of different inversion orientations, rather than being due to the different orientations per se.

2.3 Rise in frequency and role in speciation

Inversions that occur in the germline may be passed on to the carrier‘s offspring. The fate of inversions, like other types of mutations, depends on genetic drift, gene flow and natural selection. When not affected by selection, new inversions may linger in populations for a few generations, but will most likely disappear from the population with time due to random drift. Some may rise in frequency by chance, particularly in smaller populations. However, there are many examples of inversions reaching fixation in the human lineage, as well as in other species (Salm et al., 2012). It has been suggested that the suppression of recombination between inversion orientations may play a role in promoting a rise in frequency. The inversion may capture a combination of alleles that are well adapted to a particular environment. Because of recombination suppression, the haplotype will not be broken up and may thus become subject to positive selection over haplotypes that are continually recombining and therefore affected by gene flow that may not be as well adapted to that particular environment (Kirkpatrick and Barton, 2006). It has also been pointed out that this may happen the other way around, positively selected alleles may arise on one orientation of an inversion and because of recombination suppression they remain linked only to that particular orientation (Navarro and Barton, 2003). Although it is difficult to determine whether such fixation events are due to genetic drift alone or if selection has played a part as well, the fact that there are reports of association between inversion orientation and phenotypes suggests that they can be subject to natural selection, despite their supposed neutrality. Moreover, a study on 29 thousand Icelanders reported a statistically significant difference in number of offspring of women carrying different orientations of a common inversion on chromosome 17, suggesting that it is subject to selection (Stefansson et al., 2005).

Comparison of human and chimpanzee genomes have revealed many fixed inverted sequences that differ between the species, including nine cytogenetically visible pericentric inversions, indicating that polymorphic inversions may be common in both species (Feuk et

al., 2005). Since heterokaryotypes for chromosomal rearrangements (i.e. individuals that carry a rearrangement on one chromosome) are often infertile, it has been proposed that they may contribute to speciation. As well as other rearrangements, the role of inversions has been discussed (e.g. Rieseberg, 2001; Navarro and Barton, 2003; Zhang, Wang and Podlaha, 2004; Kirkpatrick, 2010), and the possibility that the reduced fitness of heterozygotes for an inversion, due to the creation of unbalanced gametes in meiosis, might contribute to speciation. This model has been criticised on the grounds that the fixation of new alleles in strong underdominance, i.e. where there is strong selection against heterozygotes, is highly unlikely, and on the other hand, if underdominance were weak, the barriers of gene flow would not be strong enough to promote speciation (Rieseberg, 2001; Navarro and Barton, 2003). An alternative model has been proposed where the suppression of recombination between different orientations is the main cause for speciation. The divergence of inversion orientations allows genes to accumulate differences, which in time can lead to incompatibilities between the different alleles and eventually reproductive isolation (Rieseberg, 2001; Navarro and Barton, 2003). A study comparing human and chimpanzee sequence data found that protein evolution was significantly faster in segments that had undergone rearrangements, suggesting that the separation process of the two lineages were facilitated by chromosomal changes (Navarro and Barton, 2003). A later study with substantially more data failed to replicate these results (Zhang, Wang and Podlaha, 2004), and the role of inversion in speciation remains controversial (Kirkpatrick, 2010; Alves et al., 2012).

2.4 Common inversion polymorphisms

Although more than 1,000 inversions have been reported in the Database of Genomic Variants (Alves et al., 2012), little is known about their frequency and only a handful have been characterized in detail. InvFEST, a database in which information on all inversions reported in the human genome are collected and merged into a non-redundant set, lists 1,092 inversions that have been predicted by one or more high-throughput methods. Of these, 85 are reported as validated, such that at least one of the inversion breakpoints has been experimentally validated (Martínez-Fundichely et al., 2014). These validations are performed using various methods, for example polymerase chain reaction (PCR), fluorescence *in situ* hybridization

(FISH), or karyotype analysis. The size of these inversions ranges from 660 base pairs (bp) to 22.6 Mb. For 54 of the validated inversions, a frequency estimate is reported, whereof 47 are reported to have a global minor allele frequency (MAF) of 1% or higher. These estimates are based on highly various sample sizes from various populations. (Martínez-Fundichely et al., 2014).

Three of the best studied inversion polymorphisms are on chromosomes 8, 15 and 17.

2.4.1 8p23.1

The inversion on chromosome 8 (8p23.1) spans a 4.5 Mb region and is one of the largest polymorphic inversions that has been found in the human genome (Salm et al., 2012). In a study of recurring rearrangements on chromosome 8, Giglio et al. (2001) discovered this inversion when they found that the mothers of all eight subjects of a rearrangement consisting of a deletion and a duplication, that causes among other things severe mental retardation, were heterozygous for the inversion. Despite its size, the inversion is submicroscopic, but was detected with FISH. Further study revealed that the inversion is flanked by large, complex blocks of low-copy repeats (LCR), 1.3 Mb and 400 kb long. The proximal block (i.e. closer to the first position on chromosome 8), which is longer, turned out to be a mixture of forward and reverted segments when compared to the shorter distal block (Sugawara et al., 2003). The inverted orientation, when compared to the reference genome, is thought to be the ancestral state in humans and is found at a frequency of around 70% in sub-Saharan Africa, declining with geographical distance from Africa to around 1.3% in the Americas (Salm et al., 2012).

To date, no marker has been identified that is perfectly correlated with inversion orientation, although Bosch et al. (2009) reported haplotypes that serve as surrogate markers for the inversion in some populations. This is unexpected in the light of its frequency and distribution, which suggests that it is old enough to have accumulated orientation specific mutations. The cause of this lack of tagging SNPs may be due to the size of the inversion. As mentioned earlier, a crossover event within an inverted region in a heterozygous individual typically results in aneuploidy. However in the case of two, or an even number of events within an inversion, the crossing over is unlikely to affect ploidy. Crossover events affect the probability of another crossover in adjacent regions, through a mechanism called interference

(Griffiths et al., 2014). This means that the probability of two events occurring within an inversion is low, although it increases with the size of the inversion. In the light of the size of the inversion on chromosome 8, the possibility of two recombination events occurring within the inverted segment cannot be ruled out. Antonacci et al. (2009) have also suggested that the lack of SNP tags may be due to frequent gene conversion events. Multiple inversion events can not be excluded either, since a single universal breakpoint has not been identified, although Salm et al. (2012) concluded that the correlation they observed between inversion status and genetic substructure suggested that an inversion event was not highly recurring.

Analysis of the inversion in samples of other primate species found no derived allele in three gorillas (*Gorilla*), three orangutans (*Pongo pygmaeus*) and one macaque (*Macaca mulatta*). However, eight chimpanzees (*Pan troglodytes*) were homozygous for the derived allele and one bonobo (*Pan paniscus*) was heterozygous (Antonacci et al., 2009). Estimates of the time to most recent common ancestor of the two orientations in humans range from 315-420 thousand years ago, which suggests independent events in the Homo and Pan lineages (Salm et al., 2012).

2.4.2 17q21.31

Another well-characterised inversion is on chromosome 17 (17q21.31). It was discovered when generating a chromosome-specific assembly from BAC clones from one individual who turned out to be heterozygous for the inversion (Stefansson et al., 2005). Previous studies (Skipper et al., 2004; Oliveira et al., 2004) had reported the existence of two highly divergent haplotypes (H1 and H2) in the MAPT gene, which is located within the inverted region, and the chromosomes from the BAC clones represented the two different haplotypes. When assembling and mapping to the reference genome, which contains H1, Stefansson et al. (2005) discovered a structural difference between the two haplotypes. While the H1 chromosome mapped to the reference genome, when assembled, a 970 kb segment on the H2 chromosome was found to map in the opposite direction. Furthermore, a partial duplication of the NSF gene, which is located 100 kb upstream from the full-length gene on H1, was seperated from the gene by 900 kb on H2 and mapped on the reverse strand (Stefansson et al., 2005).

Further study of the two haplotypes showed that of 95 SNPs genotyped within the breakpoints of the inversion, 36 of them were fixed for different alleles on different

orientations of the inversion. This suggests that no recombination has occurred between the two haplotypes (Stefansson et al., 2005), which is consistent with the size of the inversion, although Steinberg et al. (2012) reported a 30 kb sequence within the inversion region that was strikingly similar between the two orientations in a group of 728 unrelated individuals from all major continental groupings of HapMap, and proposed that this was due to a double recombination event.

It has been estimated that the divergence between the two orientation occurred two to three million years ago (Stefansson et al., 2005; Zody et al., 2008; Steinberg et al., 2012). Orthologous polymorphic inversions have been found in both *Pan* species, chimpanzees and bonobos, but are thought to have occurred independently in the *Pan* and *Homo* lineages. More distantly related primates have been found to be homozygous for the H2 orientation, with the exception of a single heterozygous Bornean orangutan (Zody et al., 2008). In order to determine the evolutional history of the inversion in humans, Zody et al. (2008) selected SNPs that were fixed in one orientation and polymorphic in the other, and compared them to the chimpanzee alleles. They found that for 90% of SNPs that were polymorphic in H1, the allele that was fixed in the H2 orientation matched the chimpanzee, but only 60% of SNPs polymorphic in H2 had a fixed H1 allele that matched the chimpanzee. This points to the H2 orientation being the ancestral state in humans (Zody et al., 2008), which seems inconsistent with the observation that the H2 haplotype is rare in Africa, with a frequency lower than 1% in sub-Saharan populations as opposed to 20% frequency in Europeans (Zody et al., 2008), and that the sequence diversity is much lower than that of H1. In order to explain how both alleles were maintained in the gene pool for so long, Stefansson et al. (2005) suggested that some kind of balancing selection may have been acting, followed by strong positive selection. The H2 allele may thus have been maintained in the population at a low frequency for a long time before it rose to higher frequency relatively recently. According to Stefansson et al. (2005), a more likely explanation was that the H2 orientation was eliminated from the human lineage entirely, then introduced again into the ancestral human gene pool, before or soon after the migration from Africa, through interbreeding with another hominin species. It has also been proposed that the increased frequency of the H2 allele in Europeans may be a result of founder effect rather than selection (Stefansson et al., 2005; Zody et al., 2008).

2.4.3 15q13.3

The identification of a recurrent deletion on chromosome 15, and previous reports of microdeletions arising preferentially from chromosomes carrying an inversion, led Sharp et al. (2008) to investigate the possibility that the 15q13.3 region harboured an inversion. These deletions cause mild to moderate retardation and mild dysmorphic features. They found that two patients carrying the deletion had a parent who was heterozygous for an inversion at the site. Testing of eight HapMap individuals from various populations revealed that seven of the sixteen alleles were inverted, which suggested that the inversion was quite common.

However, a recent study by Antonacci et al. (2014) estimated the frequency of the inversion to be 6%, which is lower than previous estimates of 20% (Antonacci et al., 2009) and 44% (Sharp et al., 2008). They found that two gorillas had an inverted orientation when compared to the human reference genome, but three chimpanzees and three orangutans had the same orientation as the reference genome, suggesting that there were seperate inversion events in gorillas and humans. Sequencing of the non-human primates revealed a much simpler structure of the region than in humans, lacking most of large duplications present in humans.

2.4.4 16p11.2

The inversion on chromosome 16 was found by Martin et al. (2004), when comparing two haplotypes of the region. It is around 450 kb and it is flanked by two large blocks of segmental duplications (Martin et al., 2004). A recent study reported an association between the inversion and the risk for joint asthma and obesity (González et al., 2014). According to the same study, the frequency of the inverted allele, that is the allele not present in the reference genome, ranges from 10% in East Africa to 49% in Northern Europe. Although it is rarer than the reference allele, it is believed to be ancestral, as it is found in all non-human primates, as well as the Neanderthal and Denisova genomes (González et al., 2014).

2.5 Methods for detecting inversions

Before the advent of whole genome sequencing, large balanced rearrangements such as inversions were primarily identified microscopically with the G-banding technique, provided

that the different orientations yielded distinguishable differences in the chromosome's banding pattern (Feuk, 2010). However, the lower size limit for such methods is 2-3 Mb (Jobling et al., 2014). Despite the fast development of sequencing techniques, detection of submicroscopic inversions has proven tricky, owing to the short length of sequence reads generated by current methods. Sanger sequencing can produce contiguous sequence reads up to 900 base pairs in length (Morozova and Marra, 2008), which makes it possible to detect small inversions that fall within this size range. Next generation sequencing (NGS) yields even smaller sequence reads (<160bp), although paired-end sequence reads, where both ends of the DNA fragments are sequenced, extend the inferential reach of such data by a few hundred nucleotides (Jobling et al., 2014).

Thus, Tuzun et al. (2005) developed a method for identifying breakpoints from NGS data by comparing the paired ends of each sequenced fragment to the human reference genome. If the ends map to the reference genome in opposite directions, the fragment might encompass a breakpoint, which can then be verified with other methods, for example FISH or PCR. Paired-end mapping (PEM) was a major breakthrough in the detection of inversions and has been applied in various studies with good results (e.g. Korbel et al., 2007; Kidd et al., 2008; Wang et al., 2008; Ahn et al., 2009; McKernan et al., 2009; Pang et al., 2010; Arlt et al., 2011). However, inversions are prone to occur between inverted low copy repeats, therefore these signals are usually effectively invisible since the sequences flanking the two breakpoints are often identical. The read lengths provided by NGS technologies provide limited resolution to directly detect and locate inversion breakpoints. Other types of genotyping methods that do not yield consecutive reads of DNA sequence are even less informative when it comes to detecting inversions.

Some researchers have tried to overcome this problem by employing indirect methods for identifying inversions. For example, Bansal et al. (2007) suggested a statistical method to detect large inversions by patterns of linkage disequilibrium (LD) from SNP data. In the case of an inversion there should be unusually strong LD between two markers that have become physically close due to the inversion, but are distant in the ancestral sequence. Conversely, weaker LD would also be detected between markers that are normally close but have become distant in the inverted sequence. Their method was designed to search for such signals. However, it was only able to detect inversions with high frequency and, as the authors

observed, unlikely to detect variants that are inverted, compared to the reference sequence, but have lower frequency than the normal variant. This method relies on the assumption that inversions constitute regions of strong LD, but as Alves et al. (2012) point out, signs of extended LD may also be due to low recombination rates or selective sweeps, rather than just due to inversions, rendering LD-based methods for detecting inversions unreliable.

The emergence of the so-called third generation sequencing technology shows promise of advancement in the discoveries of inversion polymorphisms. This new technology, unlike NGS, generates sequence data from single molecules, without the amplification and library preparations. It entails a huge increase in read-lengths, with the latest systems generating an average read length over 10 kb, albeit at the expense of lower throughput, higher error rate and higher cost per base (Rhoads and Au, 2015). Application of single-molecule real-time (SMRT) sequencing technology has for instance shed light on the complex architecture of the 15q13.3 region (Antonacci et al., 2014), and Sudmant et al. (2015) were also able to validate 208 of the 786 inversions found in the 1000 Genomes data, using targeted single molecule sequencing. Despite the improved read length, reads within long, highly identical repeats cannot be unambiguously assembled (Chaisson et al., 2014).

3 Methods

The aim of this study is to identify polymorphic inversions in the human genome by making use of the fact that inversions suppress recombination in heterozygotes, and searching for signs of such suppression. We used dense microarray SNP genotype data from 39,616 distinct parents available at deCODE Genetics, for whom over 2.4 million recombination events have been inferred based on a total of 79,708 meioses. Informed consent had been obtained from all genotyped individuals by deCODE Genetics. The microarray data derives from a combination of Illumina microarray chip types that probe a range of 300,000 to 5 million SNPs per individual. Table 1 shows the microarray chip types and the number of individuals genotyped on each type.

Table 1. Overview of genotyping arrays used and the number of individuals genotyped on each chip type. In some cases, individuals were genotyped on more than one chip type, therefore the total sum is higher than the total number of parents.

Microarray chip types	*Number of individuals*
HumanHap 300k	20,185
HumanHap 1M	628
HumanHap 610k	248
Infinium Omni1	8,148
Infinium Omni2.5	2,147
Infinium Omni5	69
Infinium OmniExpress	10,902
Total	*42,327*

3.1.1 Phasing

The genotypes of the individuals in the dataset were phased using the so-called long-range phasing approach that also yields information about the parental origin of each allele (Kong et al., 2008). Family data, where both parents of a proband are genotyped, provide a way to accurately phase the proband's genotypes. However, if all members of the trio are heterozygous at a locus, accurate phasing is not possible based on the trio's data alone. The long-range phasing method is based on the same principle as phasing with family data, but instead of using only the parents, it considers all individuals who share long haplotypes, identical by descent (IBD), with the proband. When shared haplotypes have been identified, the haplotype carriers can serve as the proband's surrogate mothers or fathers, depending on which haplotype they carry, when phasing the particular IBD region. In the case of a heterozygous SNP, the surrogate mothers, or fathers, are scanned until an individual, who shares the haplotype but is homozygous for the particular SNP, is found, enabling the genotype to be determined. The extensive genotype data available at deCODE Genetics results in a large number of surrogate parents for each proband, making it possible to phase the entire genome with high accuracy (Kong et al., 2008).

3.1.2 Recombination calling

When all individuals had been phased with this method, recombination events that took place during the production of the germ cells could be detected, by comparing the phased chromosomes of each parent-offspring pair (Kong et al., 2014). Recombination events were localized to a region between the two closest markers that were heterozygous in the parent, as heterozygosity is necessary for distinguishing the haplotype origin of alleles, and thus for determining a shift of origin from one chromosome of the parent to the other (Kong et al., 2010). This process yielded information about the positions of more than 2.4 million recombination events from 79,708 meioses, of which 33,870 are paternal and 45,838 maternal. The average male autosomal recombination rate per meiosis was 18.93, and the average female autosomal recombination rate was 38.52, while the female overall recombination rate was 40.06. The number of meioses (offspring) per parent in the data set ranged from 1 to 14, and the average number of meioses per parent was 2.01.

3.2 Microarray genotype data

As the microarray data of the parents used in this study were obtained using various different Illumina microarray chips, we merged the data from all arrays in order to obtain the complete number of genotyped individuals for each SNP. The complete number of autosomal SNPs typed with the different arrays was 1,282,653. Table 2 shows the number of SNPs per chromosome.

Table 2. Number of SNPs on each chromosome.

Chromosome	*Total N of SNPs*
1	107,352
2	105,682
3	86,657
4	78,816
5	79,437
6	92,319
7	71,087
8	68,604
9	58,922
10	67,621
11	65,191
12	64,418
13	47,649
14	41,857
15	38,890
16	39,612
17	36,119
18	36,821
19	26,827
20	31,622
21	18,040
22	19,110
Total	*1,282,653*

3.3 Testing for recombination suppression

A key premise underlying our approach to detect inversions is that individuals who are heterozygotes for an inversion polymorphism are less likely to experience a recombination event within the inverted region relative to individuals who are homozygous. After an

inversion arises, orientation specific mutations will accumulate over time. If a SNP tags an inversion perfectly, that is, different alleles are fixed on different orientations of an inversion, we expect a lower mean number of recombination events in heterozygotes around that SNP, as it follows that the individual is also heterozygous for the inversion. Given this premise, we applied a simple statistical test to assess the relationship between heterozygosity and recombination rates. For each locus, we grouped parents with valid genotypes into homozygotes and heterozygotes, and performed a t-test to compare the mean number of recombination events per individual across these two groups that occurred within 500 kb of the locus. We call this the Suppression of Heterozygote Recombination (SHR) test. Assuming that there is a sufficient number of both heterozygous and homozygous parents, and that an inversion suppresses recombination in heterozygotes, it follows that SNPs that tag such an inversion should yield significant differences under the proposed test (Figure 3). As we were looking to the two inversions on chromosomes 8, and 17, and considering their sizes, we concluded that a 500 kb radius around each SNP would be small enough to detect inversions of both sizes, yet large enough to have sufficient number of recombination events to detect differences between the two groups.

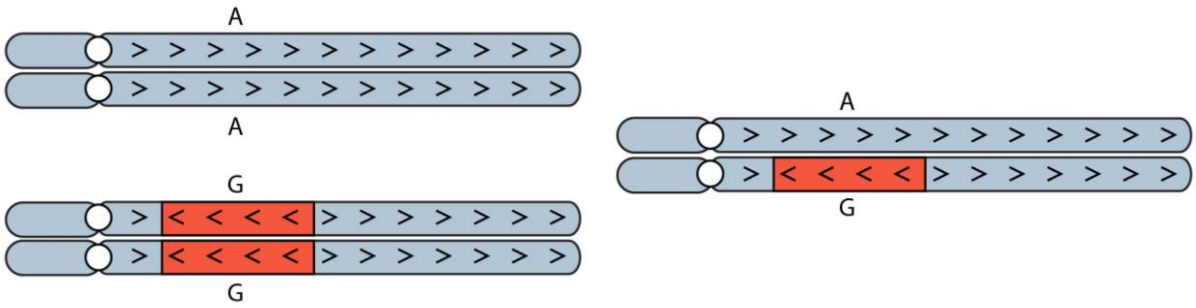

Figure 3. Grouping of parents according to genotypes. At each locus the genotyped individuals were split into groups of homozygotes (left) and heterozygotes (right). The groups were then tested for difference in recombination rates within 500 kb radius from the SNP. This is an example of a SNP tagging an inversion. At this locus we expect to see a significantly lower mean number of recombination events in heterozygotes than in homozygotes, as a recombination event within the inverted region can result in aneuploidy.

We processed each locus in turn, dividing parents into two groups of homozygotes and heterozygotes, tallying the number of recombination events per transmission within a 500kb radius for each group.

3.3.1 Identifying candidate regions for inversions

The resulting table of results from the per-locus SHR tests was then examined in the quest for clusters of loci with a significantly lower mean number of recombination events in heterozygotes, possibly due to the presence of an inversion. We scanned the genome for multiple clusters of SNPs with SHR p-values under a certain threshold, set a maximum distance between markers in order to categorize them as a group, and a minimum count of markers within a group, for the group to count as a candidate region of an inversion. The position of the first and the last SNP within the group was used to demarcate the candidate region.

All other things being equal, our approach of using suppressed recombination in heterozygotes to identify inversion polymorphisms leads to an expectation of stronger signals from common inversions than rarer ones. This is because rare inversions will have accumulated fewer orientation specific mutations over time, as they are more likely to be recent than common inversions. In addition, the SHR test will yield weaker significance when comparing means when one of the two groups has very few individuals, as in the case of rare inversions.

Also, we expect that large inversions will be easier to detect than smaller ones. This is because firstly, we expect that a recombination event between two orientations of a small inversion is less likely to affect the viability of a zygote, and therefore it is less likely to show recombination suppression. Also, the probability of a recombination event to occur within the inversion breakpoints in inversion orientation homozygotes increases with its size, and thus increases the possibility of detecting a recombination difference between homozygotes and heterozygotes. Thirdly, our method considers a 500 kb radius around each SNP, which means that inversions smaller than 1 Mb will show weaker signs of suppression in the SHR test. Thus, as the effect on recombination is only within the inversion breakpoints, the smaller the inversion, the greater the probability of a recombination event within 1 Mb that falls outside its breakpoints. The same goes for regions close to the breakpoints of an inversion of any size, we expect that the signal will be diluted as the proportion of the 1 Mb under consideration not within the inversion grows larger. This means that our method may not be optimal for identifying inversions that are, for example, smaller than 100 kb, given that there will be no

suppression of recombination for 90% of the 1 Mb surrounding the SNP. Identification of smaller inversions is also strongly dependent on the density of SNPs under study, which varies throughout the genome.

Apart from the regions closest to the breakpoints of an inversion, our expectation is that the signals of suppressed recombination should be distributed evenly within the breakpoints. However, in the case of very large inversions, as the probability of two recombination events in orientation heterozygotes within the breakpoints increases, there may be weaker suppression around the middle of the inversion, as a double recombination event would break up the genetic isolation of the two orientations without affecting ploidy. Thus, the power to detect suppression of recombination in inversion orientation heterozygotes with our method will increase with growing size of an inversion until it starts to fade concurrent with increased probability of two recombination events to occur within its breakpoints, although we expect the signals to be detectable near the breakpoints.

4 Results

The SHR test was applied to all autosomal SNPs in our microarray data from the 39,616 parents with information about recombination events. A total of 894,012 autosomal SNPs had reports of recombination events within 500 kb radius and were therefore suitable for comparison of recombination rates in SHR tests. After eliminating all SNPs that yielded fewer than 10 heterozygotes or homozygotes, the total number of SNPs left for analysis was 852,542. The mean recombination rate was lower in the group of heterozygotes than in the group of homozygotes around 482,075 (58.3%) of the 852,542 SNPs. The numbers of SNPs per chromosome before and after filtering are shown in Table 3.

Table 3. The distribution of SNPs with reports of recombination events within 500 kb before and after filtering.

Chromosome	*No. of SNPs*	*No. of SNPs after filtering*	*No. of SNPs w. a lower mean no of rec. events in heterozygotes*
1	79,555	74,987	42,208 (56.3%)
2	78,537	74,774	41,886 (56.0%)
3	63,870	60,880	34,556 (56.8%)
4	57,317	54,519	31,136 (57.1%)
5	57,487	55,433	31,422 (56.7%)
6	61,989	59,528	34,737 (58.4%)
7	51,419	49,018	27,189 (55.5%)
8	47,782	45,599	25,872 (56.7%)
9	39,232	37,640	21,088 (56.0%)
10	46,813	44,532	25,077 (56.3%)
11	45,233	43,065	24,658 (57.3%)
12	44,977	42,831	24,247 (56.6%)
13	32,141	30,854	17,530 (56.8%)
14	28,489	27,145	14,811 (54.6%)
15	26,917	25,426	14,207 (55.9%)
16	26,097	24,899	13,857 (55.7%)
17	24,839	23,341	13,390 (57.4%)
18	24,580	23,435	13,166 (56.2%)
19	16,949	16,126	9,102 (56.4%)
20	19,890	19,252	10,899 (56.6%)
21	9,404	9,100	5,054 (55.5%)
22	10,495	10,158	5,983 (58.9%)
Total	*894,012*	*852,542*	*482,075 (56.5%)*

4.1 Distribution of p-values

Of the 852,542 SNPs, 47,947, or 5.6% had a p-value lower than 0.05 from the SHR test. Bonferroni correction for multiple comparisons yielded a threshold of statistical significance (α) of 5.86×10^{-8}. Tests for 600 SNPs passed this threshold, or 0.07% of all SNPs under consideration. Of these 600 SNPs, 574 (95.7%) yielded a lower mean number of recombination events in heterozygotes. Table 4 presents an overview of the distribution of p-values of SNPs under analysis.

Table 4. Overview of the distribution of significance of SNPs under analysis. The second column shows the number (and percentage) of SNPs with a lower mean number of recombination events in heterozygotes than in homozygotes.

	All SNPs	*SNPs w. a lower mean no. of rec events in heterozygotes*
SNPs under analysis	852,542	482,075 (56.5%)
p-value < 0.05	47,947 (5.6%)	28,436 (59.3%)
p-value < 5.86 x 10^{-8} (Bonferroni correction)	600 (0.07%)	574 (95.7%)

The genomic locations by chromosome band of SNPs with a significantly lower mean number of recombination events in heterozygotes after Bonferroni correction are shown in Table 5.

Table 5. Locations by chromosome band of SNPs with a lower mean number of recombination events in heterozygotes significant after Bonferroni correction.

Chromosome band	*No. of SNPs*
1p33	1
1q21.1	1
2q21.3	4
5p11	1
7q11.21	2
8p23.1	487
8q11.1	6
10p11.1	1
10q22.3	6
12q24.13	2
14q23.3	1
15q13.3	13
16p11.2	4
16p11.1	3
17q21.31	41
18q11.1	1
	574

It is clear, from the numbers of SNPs with significant p-values after Bonferroni correction, that the strongest signal of heterozygote recombination suppression in our data comes from the location of the known inversion on chromosome 8 (8p23.1). Of the 574 SNPs that have a lower mean number of recombination events in heterozygotes and are significant after Bonferroni correction, 487 (84.8%) are located within its breakpoints. The known inversion on chromosome 17 has 41 (7.1%) SNPs within its breakpoints and the known inversions on chromosomes 15 and 16 have 13 (2.3%) and 4 (0.7%) SNPs respectively. Thus, the four known inversions collectively account for 545 (94.9%) of the SNPs that survive Bonferroni correction.

Assuming that no suppression of recombination in heterozygotes is present in the data, we expect p-values from the SHR tests to be uniformly distributed. In order to assess whether there is evidence for polymorphic inversions other than those on chromosomes 8, 15, 16, and 17, we made a Q-Q plot of the p-values with and without the known inverted regions (Figures 4 and 5).

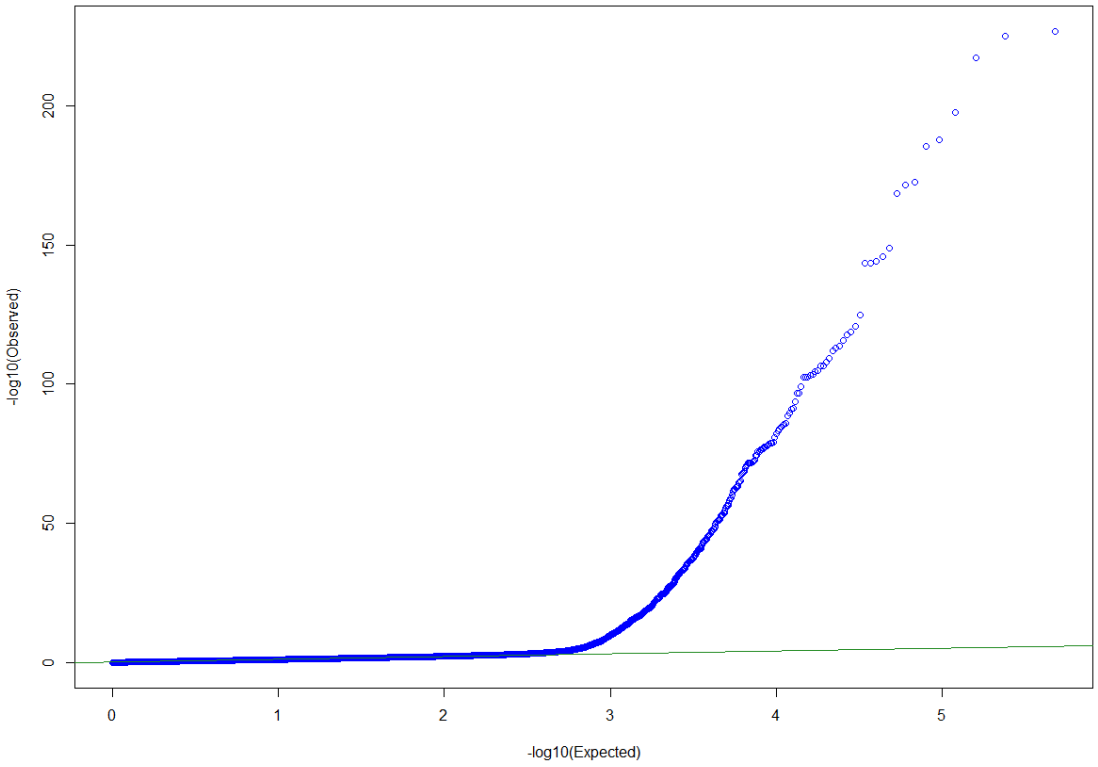

Figure 4. Q-Q plot of p-values from SHR tests. All 482,075 SNPs that yield a lower mean number of recombination events in heterozygotes.

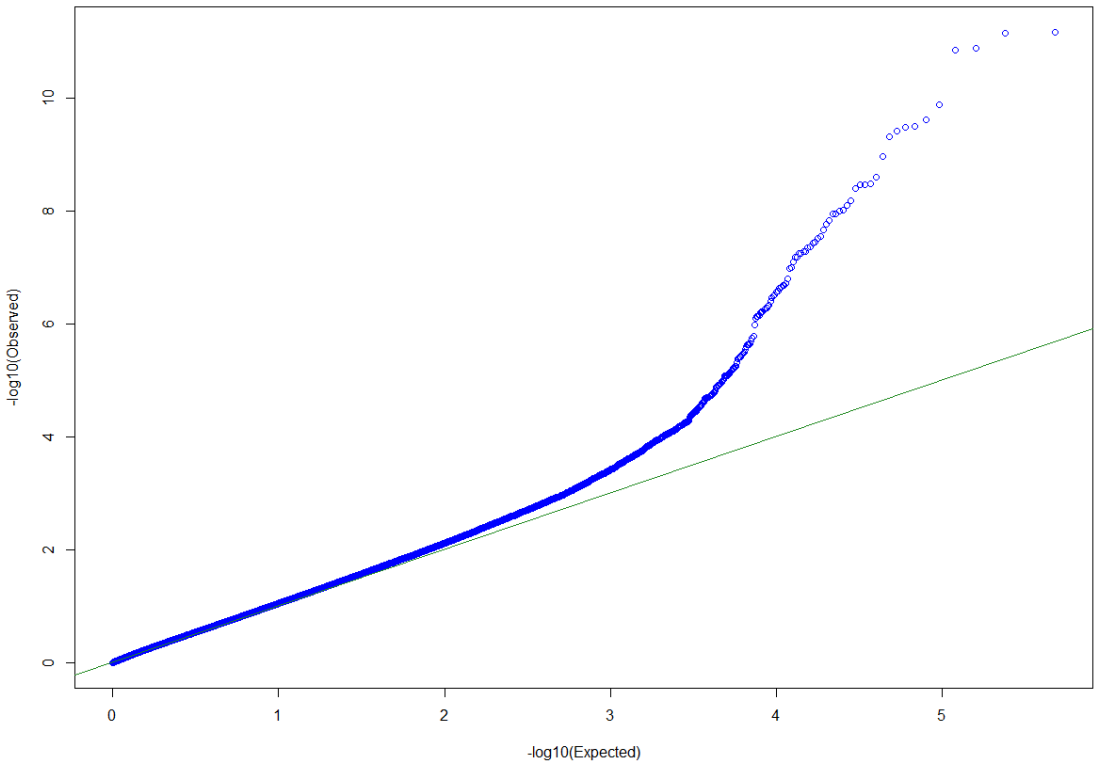

Figure 5. Q-Q plot of p-values from SHR tests. SNPs that yield a lower mean number of recombination events in heterozygotes, excluding 1,899 SNPs located within the breakpoints of the four known inversions.

A comparison of Figures 4 and 5 reveals that the four known inversions explain the majority of the skewness of the Q-Q plot. However, after excluding the SNPs located within the breakpoints of the known inversions, the distribution of p-values is substantially skewed relative to the null hypothesis of no additional inversions. An overview of the distribution of p-values, including and excluding SNPs within the four known inversions, is shown in Table 6.

Table 6. The 482,075 SNPs that yield a lower mean number of recombination events in heterozygotes. Overview of the number of SNPs with a p-value lower than 0.05, 0.001 and 5.86 x 10-8 (Bonferroni correction).

P-value	*N of SNPs*	*N of SNPs outside the four known inversions*	*N of SNPs within the four known inversions*
All	482,075	480,176	1,899
< 0.05	28,436 (5.90%)	27,343 (5.69%)	1,093 (57.56%)
<0.001	1,688 (0.35%)	884 (0.18%)	804 (42.34%)
< 5.86 x 10^{-8} (Bonferroni)	574 (0.12%)	35 (0.007%)	539 (28.38%)

A Manhattan plot of all SNPs with a lower mean number of recombination events in heterozygotes (Figure 6) shows very clearly that the inversion on 8p23.1 is unique, not only with regards to its size and the number of SNPs tagging it, but also the significance of the difference in recombination events between heterozygotes and homozygotes. Way behind the 8p23.1 inversion, although with multiple SNPs of strong significance within their breakpoints, the two known inversions on chromosomes 15 and 17 tower over a few other visible rises in the plot (Figure 7).

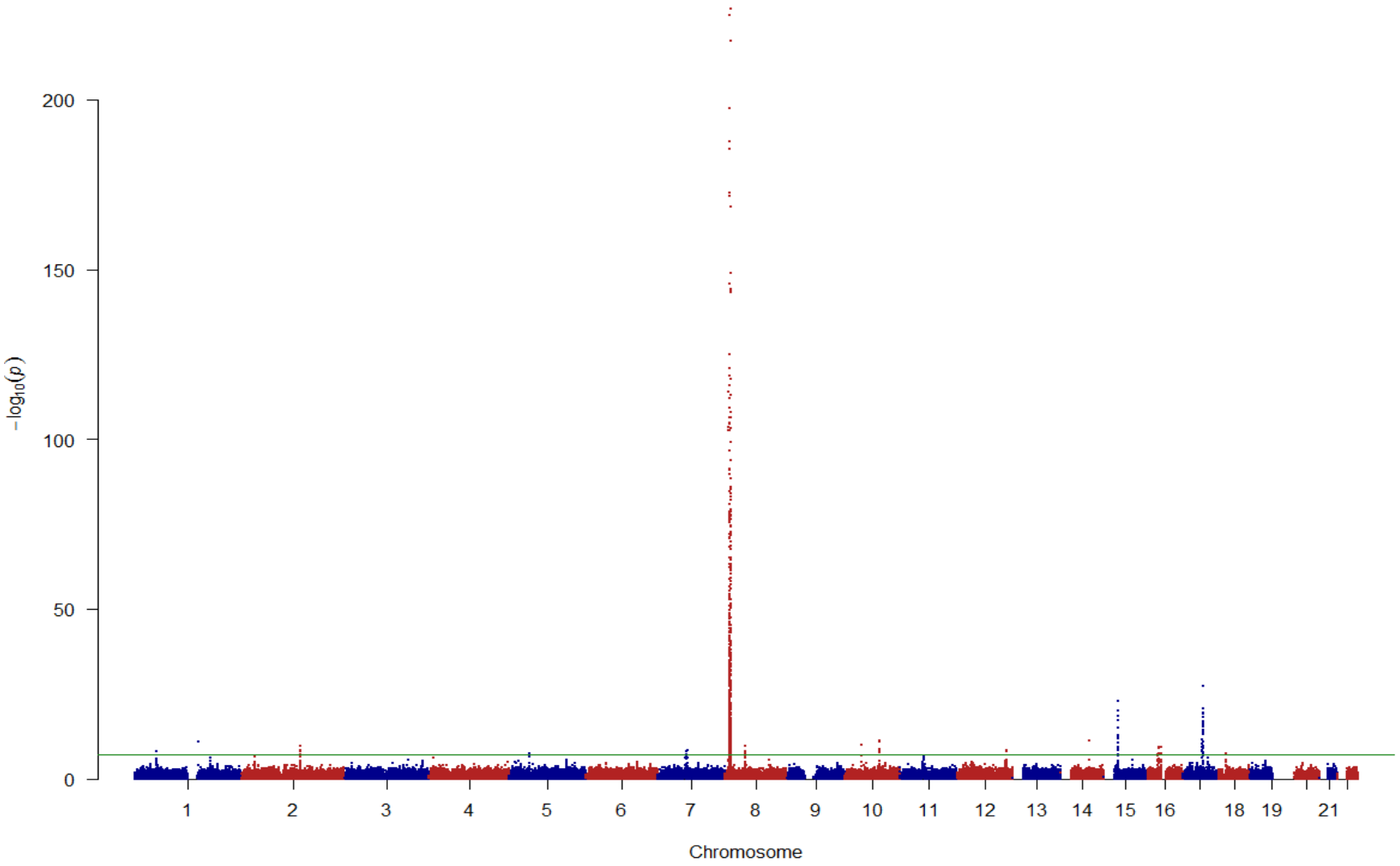

Figure 6. A Manhattan plot of –log10 transformed p-values from the SHR test for 482,075 SNPs with lower recombination in heterozygotes. Green horizontal line marks the Bonferroni threshold of statistical significance.

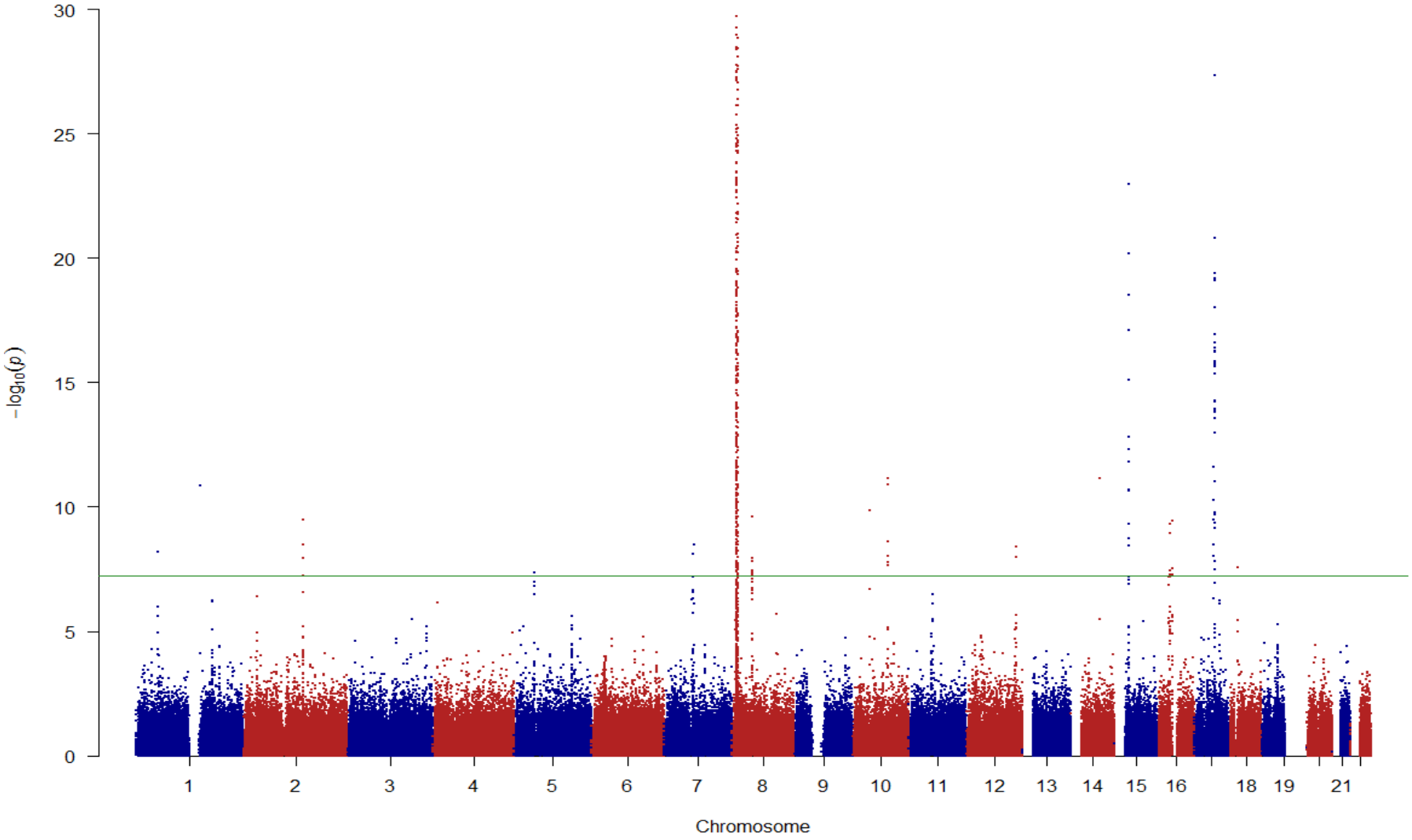

Figure 7. A Manhattan plot of –log10 transformed p-values from the SHR test for 482,075 SNPs with lower recombination in heterozygotes – with Y axis adjusted to a maximum value of 30. Green horizontal line marks the Bonferroni threshold of statistical significance.

4.2 Identifying regions of heterozygote suppression

4.2.1 Microarray genotype data

Given the strong pattern of local correlation of alleles, i.e. linkage disequilibrium (LD), in the genome, the number of independent tests is somewhat smaller than the total number of individual SNPs. As a result, Bonferroni correction based on the total number of SNPs tested is overly conservative. Taking this into account, we used an alternative approach to identify candidate inversions in the microarray data, scanning the genome for local regions where groups of at least five SNPs within 250 kb distance from one another, had p-values <0.001. This resulted in a list of 1,144 SNPs in 34 regions (see Table 7). The size of the regions (hereafter referred to as SHR regions), demarcated by the position of the first and last SNP within the group, ranged from 32 kb to 3.8 Mb and the number of SNPs per region under the threshold ranged from 5 to 719. The number of genotyped individuals behind each SNP ranged from 602 to 39,610 individuals (with a mean of 25,395 and a standard deviation (SD) of 10,254) (see Figure 8).

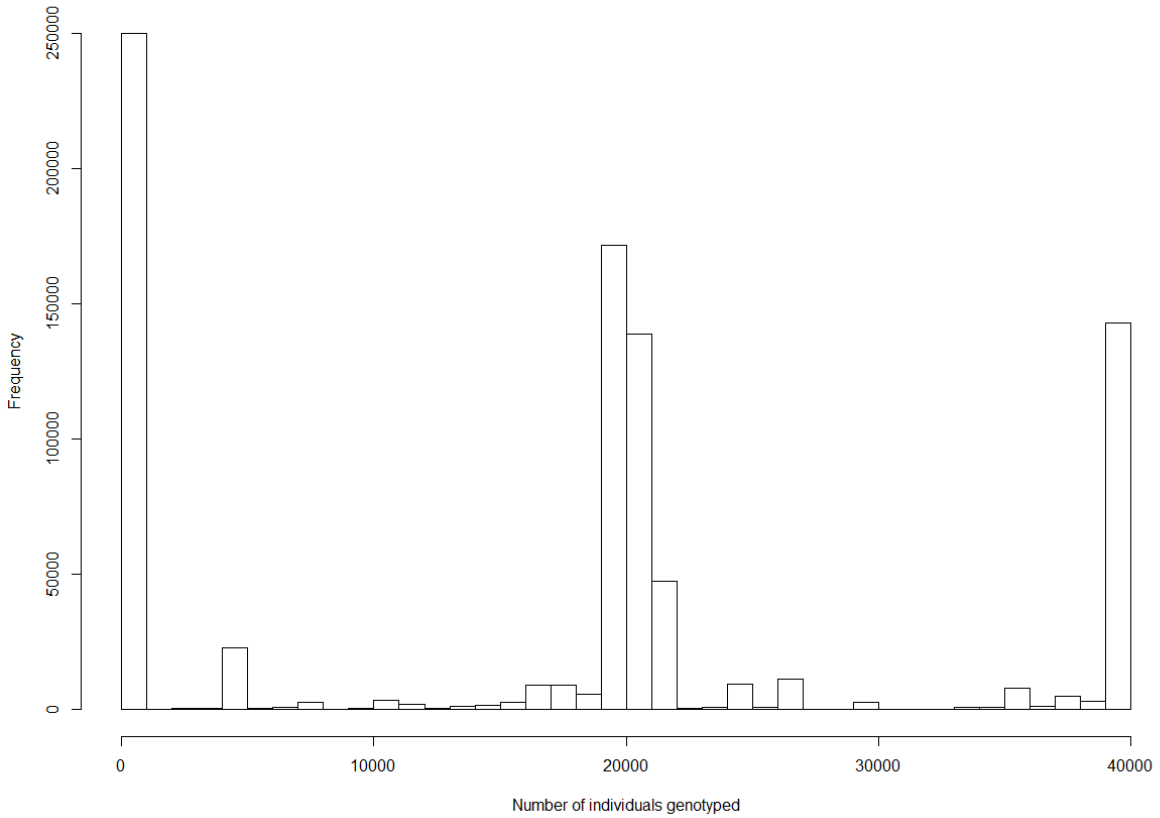

Figure 8. The distribution of individuals genotyped for each SNP.

4.2.2 Validation with WGS data

Of the 39,616 parents in our data, WGS data was available for 6,175. In order to shed further light on the 34 candidate regions of inversions, we ran the SHR test on all single nucleotide polymorphisms in the WGS data with a MAF higher than 1%. After filtering out SNPs with less than 10 heterozygotes or homozygotes, there were 10,664,075 SNPs left with recombination events within 500 kb radius. To ensure the quality of the data, we filtered out SNPs with >2% missing genotypes, leaving 10,084,112 SNPs. A total of 5,647,371 (56.0%) had a lower mean number of recombination events in heterozygotes and 201,140 of those, or 3.6%, had a p-value lower than 0.05. As expected, the signal of recombination suppression was weaker in the WGS data than in the microarray data due to substantially smaller sample size. While the Bonferroni correction is overly conservative, we recognize that an α level of 0.05 will yield a large number of false positives when using the SHR test to seek SNPs associated with inversions. However, we expect an enrichment for SNPs with real association among SNPs

with a p-value under 0.05 and a high proportion of SNPs that pass that threshold may be an indication of such enrichment. The number of SNPs in the microarray data and the WGS data with $p<0.05$ from the SHR test, in each of the 34 SHR regions identified using the microarray data, is shown in Table 7.

Table 7. Summary of results from microarray and WGS data for the 34 SHR regions. All analyses were performed using genomic coordinates from NCBI Build 38. Note that all the regions are demarcated by the first and last SNPs identified with our approach in the microarray data, not the reported breakpoints of the inversions.

Band	*Position of first SNP*	*Position of last SNP*	*Size*	*SNPs w. lower rec rates in het*	*SNPs w p<0.05*	*SNPs w p<0.001*	*Bonferroni*	*WGS loci w. lower rec rates in het*	*WGS loci w p<0.05*	*Previous reports*
1p33	49,362,991	49,662,526	299,535	96.1% (49/51)	24.5% (12/49)	7	1	81.1% (492/607)	1.2% (6/492)	
1q25.1	174,159,002	174,374,567	215,565	100.0% (71/71)	28.2% (20/71)	14	0	80.5% (458/569)	0.0% (0/458)	
1q31.1	189,289,778	189,409,595	119,817	89.7% (26/29)	57.7% (15/26)	7	0	71.1% (345/485)	11.3% (39/345)	
2p22.3	31,875,990	32,232,468	356,478	72.7% (88/121)	29.5% (26/88)	9	0	85.0% (1086/1277)	27.7% (301/1086)	
2q21.3	135,004,774	135,741,596	736,822	84.4% (135/160)	21.5% (29/135)	18	4	89.1% (1350/1515)	8.7% (117/1350)	
3q26.32	178,638,481	178,899,463	260,982	67.1% (55/82)	18.2% (10/55)	5	0	72.3% (834/1153)	10.9% (91/834)	
5p11	45,783,404	46,378,855	595,451	97.4% (74/76)	48.6% (36/74)	14	1	82.7% (2739/3310)	1.2% (32/2739)	
5q14.1	80,647,183	80,787,812	140,629	69.8% (37/53)	29.7% (11/37)	7	0	70.4% (373/530)	7.5% (28/373)	
5q23.3-q31.1	130,775,461	131,515,245	739,784	97.3% (183/188)	34.4% (63/183)	21	0	76.2% (1813/2379)	8.1% (146/1813)	
5q31.2	136,158,385	136,190,161	31,776	100.0% (12/12)	75.0% (9/12)	9	0	76.5% (75/98)	64.0% (48/75)	
6p22.1	29,008,294	29,644,108	635,814	77.6% (420/541)	29.5% (124/420)	20	0	58.2% (1834/3152)	2.0% (36/1834)	
6p21.33	30,870,911	30,953,113	82,202	78.2% (122/156)	9.0% (11/122)	6	0	67.5% (280/415)	0.0% (0/280)	
6q24.3	145,797,089	145,886,521	89,432	77.8% (14/18)	57.1% (8/14)	8	0	88.3% (203/230)	38.4% (78/203)	
7q11.21	65,305,569	66,200,069	894,500	82.3% (116/141)	29.3% (34/116)	17	2	81.0% (3097/3824)	7.8% (243/3097)	Predicted
7q21.2	92,064,523	92,130,283	65,760	69.2% (18/26)	50.0% (9/18)	9	0	74.9% (125/167)	12.8% (16/125)	
8p23.1	8,236,884	12,002,342	3,765,458	77.6% (1447/1864)	64.1% (928/1,447)	719	487	67.9% (15149/22321)	40.8% (6186/15149)	Validated
8q11.1	46,031,220	46,426,613	395,393	91.4% (53/58)	64.2% (34/53)	24	6	86.0% (1637/1903)	7.2% (118/1637)	
9q33.1	116,538,497	116,691,176	152,679	80.5% (33/41)	36.4% (12/33)	6	0	80.5% (529/657)	0.9% (5/529)	Predicted
10q22.3	79,637,253	80,180,945	543,692	45.9% (90/196)	23.3% (21/90)	15	6	74.3% (2250/3027)	13.0% (292/2250)	Predicted
11p11.12	49,866,070	50,662,884	796,814	96.0% (120/125)	29.2% (35/120)	15	0	82.4% (3909/4745)	5.9% (232/3909)	Predicted
11q11	54,590,278	54,804,507	214,229	97.4% (37/38)	43.2% (16/37)	7	0	89.8% (1502/1673)	20.6% (309/1502)	
12p12.1	23,327,394	23,432,709	105,315	86.4% (19/22)	47.4% (9/19)	5	0	74.9% (334/446)	15.3% (51/334)	
12p11.1	34,605,331	34,673,639	68,308	100.0% (11/11)	54.5% (6/11)	6	0	91.7% (397/433)	0.0% (0/397)	
12q11-q12	37,533,312	37,851,950	318,638	95.5% (21/22)	57.1% (12/21)	6	0	90.7% (1975/2177)	26.2% (517/1975)	Predicted
12q24.12-q24.13	111,693,894	112,087,269	393,375	91.6% (87/95)	40.2% (35/87)	19	2	88.5% (787/889)	13.6% (107/787)	
13q21.1	53,703,693	53,884,196	180,503	88.9% (56/63)	42.9% (24/56)	7	0	72.0% (415/576)	0.2% (1/415)	
14q21.1	40,968,475	41,240,508	272,033	71.6% (58/81)	19.0% (11/58)	5	0	54.9% (767/1397)	17.1% (131/767)	
15q13.3	30,906,769	32,089,406	1,182,637	52.4% (209/399)	22.0% (46/209)	27	13	50.3% (2428/4823)	15.7% (382/2428)	Validated
16p11.2	28,473,820	28,976,948	503,128	84.8% (95/112)	42.1% (40/95)	23	4	68.9% (779/1130)	28.6% (223/779)	Validated
16p11.1	35,452,449	35,867,952	415,503	92.3% (24/26)	62.5% (15/24)	11	3	78.4% (820/1046)	8.0% (66/820)	Predicted
17q21.31	45,415,735	46,290,846	875,111	69.1% (141/204)	66.7% (94/141)	52	41	82.6% (4262/5162)	79.9% (3405/4262)	Validated
17q22	58,437,097	58,597,544	160,447	79.6% (39/49)	23.1% (9/39)	6	0	57.0% (259/454)	10.8% (28/259)	
18q11.1	20,990,274	21,177,462	187,188	100.0% (34/34)	14.7% (5/34)	5	1	61.4% (213/347)	1.9% (4/213)	
19q13.12	36,950,209	37,395,428	445,219	99.1% (105/106)	41.9% (44/105)	15	0	79.1% (1181/1493)	0.3% (4/1181)	

Overall, 5.9% of SNPs with a lower mean number of recombination events in heterozygotes yield $p<0.05$ from the SHR tests in the microarray data, while 3.6% meet that criteria in the SHR tests based on the WGS data. As can be observed in Table 7, within most of the SHR regions, a higher percentage of SNPs have $p<0.05$ in both datasets. Not surprisingly, the four known inversions have considerably higher percentage of SNPs with $p<0.05$ than expected by chance and within the 17q21.31 region, an astounding 79.9% of the SNPs in the WGS data have $p<0.05$. Other interesting regions are for example 5q31.2, the smallest of the SHR regions, spanning just over 30 kb, where 75% of SNPs in the microarray data and 64% of SNPs in the WGS data have $p<0.05$, and 6q24.3, a region of 90 kb, where 57.1% and 38.4% of SNPs pass that threshold in the microarray and WGS data respectively. No reports on inversions were found within these two regions.

4.2.3 Correlation of SNPs with inversion orientation

As the sample sizes behind each SNP in the microarray data vary widely, the p-values from the SHR test are not sufficiently comparable to use for identifying the SNPs with the strongest association with candidate inversions in the 34 SHR regions. In contrast, while there is less power for the SHR tests in the WGS data, due to smaller sample sizes, the p-values are comparable across all SNPs due to the same underlying sample size in all cases. We used the following approach to identify the best tagging SNPs in each of the 34 SHR regions for the putative inversion polymorphisms. In the microarray data, we applied a p-value threshold of 0.001 and then sought the SNP that provided the greatest difference in average number of recombination events per individual between heterozygotes and homozygotes (hereafter referred to as δ_{SHR}). The same approach was used for the WGS data, but with a p-value threshold of 0.05. For three of the 34 SHR regions, no SNP had a p-value that passed the threshold in the WGS data. In these instances, we used the SNP with the lowest p-value. In order to avoid excessive sampling error for SNPs with small sample sizes in the microarray data, we only considered SNPs with at least 10,000 genotypes. The results are shown in Table 8.

Table 8. **SNPs with the greatest δ_{SHR} within each region, after applying a threshold of p-value<0.001 in the microarray data (left) and 0.05 in the WGS data (right). Within three regions, no SNP had p<0.05 in the WGS data, in which case the lowest p-value was used (*).**

	Microarray							*WGS data*						*WGS most*
Band	*SNP*	*Position*	*N het*	*N hom*	δ_{SHR}	*p-value*	*MAF*	*Position*		*N het*	*N hom*	δ_{SHR}	*p-value*	*common MAF*
1p33	rs4926814	49,491,464	9,652	10,151	0.0024	1.06E-06	0.43	49,511,623	rs1167293	2634	3563	0.0029	3.59E-03	0.22
1q25.1	rs7555067	174,297,504	9,053	10,742	0.0024	6.13E-07	0.36	174,333,802	rs78683861	1435	4765	0.0015	9.73E-02	*0.14
1q31.1	rs12747712	189,332,781	12,876	13,889	0.0042	4.26E-05	0.42	189,314,577	rs201754002	2640	3559	0.0056	3.60E-03	0.42
2p22.3	rs212708	32,226,377	11,828	12,955	0.0036	1.19E-05	0.40	32,205,825	rs212678	2712	3482	0.0065	2.44E-04	0.12
2q21.3	rs1561277	135,334,491	5,379	15,596	0.0086	1.14E-08	0.15	135,072,022	rs12469098	892	5304	0.0105	4.41E-03	0.09
3q26.32	rs13064262	178,853,361	7,936	13,031	0.0084	6.23E-06	0.26	178,739,597	rs74385196	409	5786	0.0140	3.00E-02	0.23
5p11	rs10941704	45,783,404	7,666	13,324	0.0016	9.84E-04	0.24	46,323,031	rs11950489	3072	3107	0.0007	5.71E-03	0.41
5q14.1	rs33010	80,787,812	4,682	16,344	0.0080	9.71E-04	0.13	80,675,991	rs836812	2419	3774	0.0085	1.30E-02	0.22
5q23.3-q31.1	rs4705889	130,800,404	8,416	12,552	0.0027	9.41E-05	0.28	131,037,397	rs836812	2078	4116	0.0027	9.78E-03	0.22
5q31.2	rs12719482	136,178,528	5,650	14,161	0.0070	3.95E-04	0.17	136,190,161	rs7731417	2472	3720	0.0091	1.06E-02	0.29
6p22.1	rs3117329	29,259,866	9,803	9,991	0.0030	1.81E-04	0.46	29,636,347	rs1233378	2355	3824	0.0049	3.71E-03	0.39
6p21.33	rs1264333	30,876,537	18,771	20,813	0.0038	1.00E-04	0.40	30,912,379	rs1233378	353	5837	0.0092	8.55E-02	*0.03
6q24.3	rs9390358	145,824,542	9,852	10,193	0.0027	6.83E-05	0.46	145,879,888	rs2253886	3088	3076	0.0049	3.40E-03	0.47
7q11.21	rs4718225	65,305,569	10025	11014	0.00231	2.38E-04	0.40	65,430,185	rs66918658	3015	3139	0.0039	3.18E-04	0.36
7q21.2	rs6465347	92,064,523	9,613	11,412	0.0045	4.56E-04	0.35	92,130,283	rs4644173	2817	3380	0.0045	3.67E-02	0.36
8p23.1	rs9657521	11,972,993	7441	12376	0.08057	7.23E-84	0.25	11,970,691	rs4841659	3051	3140	0.1084	1.87E-44	0.48
8q11.1	rs4873062	46,256,532	8,698	12,069	0.0009	4.79E-04	0.30	46,134,510	rs10866884	2769	3424	0.0013	8.77E-03	0.30
9q33.1	rs803892	116,645,267	9,589	10,099	0.0088	9.09E-05	0.44	116,597,466	rs2093324	2969	3227	0.0130	3.69E-03	0.39
10q22.3	rs2395594	79,637,253	1,440	8,907	0.0120	1.29E-04	0.07	79,799,784	-	199	5962	0.0265	3.75E-02	0.11
11p11.12	rs2007068	50,136,389	9,920	9,892	0.0011	4.92E-04	0.50	50,428,834	rs691329	2719	3470	0.0009	3.08E-02	0.47
11q11	rs1603756	54,804,507	10,398	10,599	0.0005	3.47E-06	0.50	54,614,031	rs4447158	3188	2967	0.0006	1.90E-04	0.50
12p12.1	rs1867520	23,419,968	6,602	13,441	0.0087	2.71E-04	0.21	23,371,131	rs10743468	2088	4105	0.0119	1.64E-02	0.22
12p11.1	rs9705474	34,673,639	9,418	10,340	0.0009	1.91E-05	0.39	34,636,091	rs71459549	2268	3925	0.0008	5.87E-02	*0.25
12q11-q12	rs8189549	37,851,950	5,088	5,538	0.0012	1.28E-04	0.40	37,533,312	rs12230545	3019	3173	0.0009	4.03E-05	0.41
12q24.12-q24.13	rs4767364	112,083,644	9,605	11,365	0.0038	2.33E-06	0.36	111,751,197	rs118018677	740	5452	0.0049	2.52E-02	0.23
13q21.1	rs9568954	53,884,196	4,490	14,786	0.0061	4.37E-04	0.13	53,877,515	rs2051121	2921	3267	0.0059	4.42E-02	0.40
14q21.1	rs1954451	41,102,533	5,787	15,231	0.0051	1.23E-04	0.17	41,070,236	rs8016864	1718	4470	0.0090	1.94E-03	0.17
15q13.3	rs1075232	31,449,013	2,469	18,573	0.0396	6.54E-21	0.06	31,681,426	rs72709326	504	5685	0.0415	4.60E-06	0.07
16p11.2	rs4788069	28,605,344	8329	7629	0.00898	1.01E-06	0.42	28,588,700	rs117985404	200	5996	0.0192	4.66E-02	0.44
16p11.1	rs2163977	35,681,763	10,212	10,636	0.0005	1.21E-05	0.46	35,622,117	rs1973278	3143	3050	0.0005	1.71E-04	0.11
17q21.31	rs17660132	46,088,437	4731	12397	0.01313	4.38E-16	0.17	45,629,062	rs142822273	219	5972	0.0173	4.74E-02	0.18
17q22	rs11650710	58,506,186	5,360	14,459	0.0076	8.04E-07	0.16	58,507,147	rs2302190	1626	4565	0.0073	2.22E-02	0.16
18q11.1	rs11660183	21,177,462	10,259	10,622	0.0016	1.55E-04	0.43	21,177,462	rs11660183	3042	3149	0.0018	1.70E-02	0.45
19q13.12	rs1644634	36,951,078	9,663	10,116	0.0022	7.53E-05	0.45	37,256,322	rs12971925	1774	4409	0.0029	3.78E-02	0.18

In almost all cases, δ_{SHR} is greater in the WGS data than in the microarray data, which indicates that although there is less power per SNP in the WGS data, it provides greater resolution than the microarray data, due to the much greater number of SNPs tested.

The SNPs that show the strongest signal of association with suppression of recombination in heterozygotes can be assumed to be the best taggers of the putative underlying inversions. Their allele frequencies may therefore also be considered to provide a strong indication of the orientation frequencies of the putative inversions. However, an estimation of frequency solely based on the frequency of the SNP that yields the greatest δ_{SHR}, may be too conservative, as there may be strong suppression of recombination in heterozygotes for a SNP which one allele is only found on one inversion orientation, although the allele is only found within a subset of inversion carriers.

To circumvent this problem, we examined the MAF of all SNPs within each SHR region in the WGS data with $p<0.05$. Although we expect some false positives, we should have an enrichment of SNPs with strong correlation with the putative inversions. The stronger the correlation of SNPs with an inversion, the more correspondence we should see between their frequencies. Therefore, in order to assess the frequencies of the putative inversions, we classified all SNPs within each region by their frequency and found the most common MAF within each region. The results are displayed in Table 8.

Table 9 shows a comparison of the most frequent MAF within the regions of the four known and validated inversions with the reported frequency of the inversions, revealing a concordance between the MAF of the four inversions in Europeans and the most frequent MAF within the regions in the Icelandic WGS data.

Table 9. MAF of the four validated inversions in Europeans (left) and the most frequent MAF of SNPs within the region in the WGS data (see text for details).

	MAF of inversion (previous studies)	*Most frequent MAF in WGS data*
8p23.1	0.43 (Salm et al., 2012)	0.48
15q13.3	0.06 (Antonacci et al., 2014)	0.07
16p11.2	0.49 (González et al., 2014)	0.44
17q21.31	0.20 (Stefansson et al., 2005)	0.18

4.2.4 Regions of common inversions

Comparing the list of candidate inversions to the InvFEST database and scanning the literature revealed reports of ten inversions overlapping the SHR regions, four of which were the common inversions on chromosomes 8, 15, 16, and 17. No published reports of inversions were found for the rest of the 34 SHR regions.

4.2.4.1 8p23.1

According to InvFEST, the proximal breakpoint of the 8p23.1 inversion is located between 7,064,966 and 8,239,446, and the distal breakpoint between 11,922,365 and 12,716,088 (Martínez-Fundichely et al., 2014). In the microarray data, there was a clear increase in statistical significance of lower mean number of heterozygote recombination events around SNPs within the breakpoints of the inversion (see Figure 9). A total of 1,900 SNPs were positioned within the breakpoints, of which 1,466 (77,2%) yielded a lower mean number of recombination events in heterozygotes than homozygotes.

The region identified through the SHR test at 8p23.1, demarcated by the first and last SNP under the given threshold, is 3.8 Mb. The coordinates of the first and last SNPs are 8,236,884 and 12,002,342, both within the regions of the reported breakpoints, although close to the given inner boundaries. The SNP that yielded the most significant difference in the mean number of recombination events between heterozygotes and homozygotes in the microarray data was rs2898290, with the p-value 1.82×10^{-227} (N=39,547). The mean number of recombination events within a 500kb radius of the SNP in homozygotes for the SNP was 0.0811 while it was 0.0116 in heterozygotes, yielding the δ_{SHR} 0.0695. The SNP with the greatest δ_{SHR}, after filtering out those with fewer than 10,000 genotypes, and may thus have the strongest

correlation with the inversion, was rs9657521, at position 11,972,993. The δ_{SHR} around the SNP was 0.08057 (N=19,817, p=7.23^{-84}) (see Table 8).

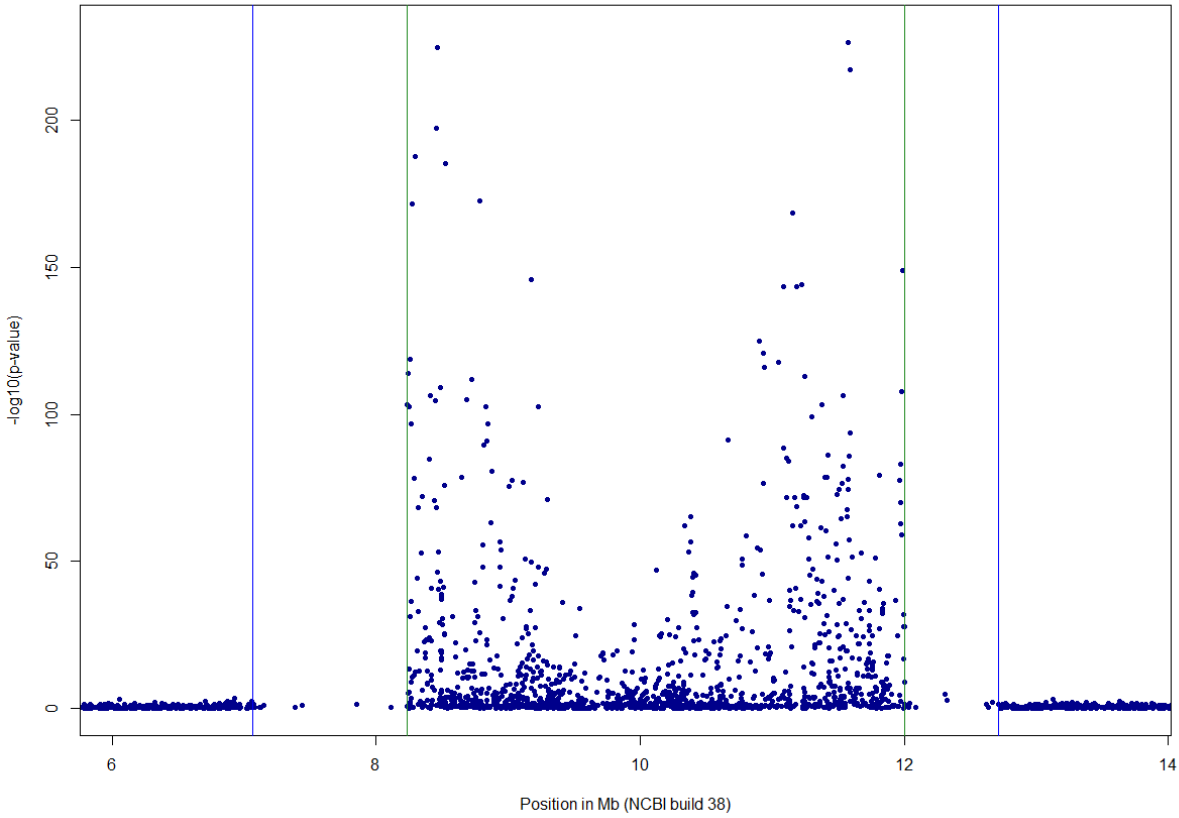

Figure 9. Results from the SHR test in the region containing the 8p23.1 inversion. Blue dots represent the p-values of SNPs in the microarray data that yield a lower mean number of recombination events in heterozygotes. Blue vertical lines mark the inversion breakpoints according to the InvFEST database and green lines mark the positions of the first and the last SNP of the SHR region identified through our test.

Within the reported breakpoints in the WGS data, 65.4% (17,934 of 27,420) of the SNPs yielded a lower mean number of recombination events in heterozygotes, of which 6,591, or 36.8% had a p-value lower than 0.05. The most significant SNP, rs10097870 at position 11,587,007, had a p-value of 6.81 x 10^{-52} and a δ_{SHR} of 0.0861 (N=6,193). Overall, we observe the same pattern of suppressed recombination in heterozygotes across the 8p23 inversion in the WGS data, although with less statistical significance due to smaller sample size (see Figure 10). The greatest δ_{SHR} within the inversion breakpoints in the WGS data was 0.1084, for rs4841659, at position 11,970,691 (N=6,191, p=1.87 x 10^{-44}).

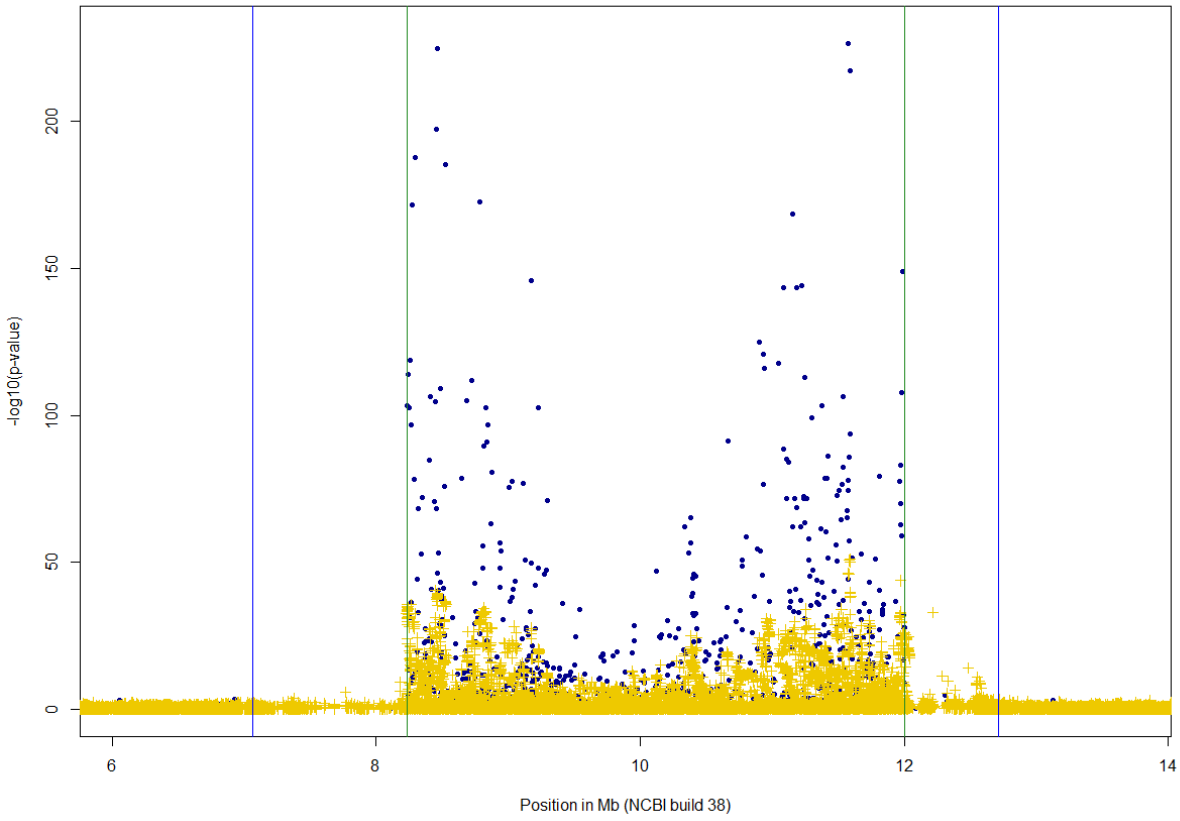

Figure 10. Results from the SHR test in the region containing the 8p23.1 inversion. Blue dots represent the p-values of SNPs in the microarray data that yield a lower mean number of recombination events in heterozygotes and yellow crosses SNPs with MAF>1% and a lower mean number of recombination events in heterozygotes in the WGS data. Blue vertical lines mark the inversion breakpoints according to the InvFEST database and green lines mark the positions of the first and the last SNP of the SHR region.

The reported breakpoints of the 8p23 inversion are located within regions of segmental duplications. As can be observed in Figure 9, there is a decrease in the density of SNPs in the microarray data in these regions due to these duplications, as genotype calling within such duplicated regions is often challenging and microarray chips typically target SNPs at unambiguous positions. This accounts for the smaller estimate of inversion size through the SHR test, when compared to the reported length of the inversion.

4.2.4.2 17q21.31

The size of the 17q21.31 inversion was initially reported as 970 kb (Stefansson et al., 2005), but its size according to the InvFEST database is 835 kb. Its breakpoints lie within the regions

45,495,836-45,627,799 and 46,087,894-46,707,123, respectively (Martínez-Fundichely et al., 2014). In the microarray data, there were 197 SNPs located within the reported breakpoints, whereof 137 of these yielded a lower mean number of recombination events in heterozygotes. Of the 137 SNPs, 89 (65.0%) had a p-value lower than 0.05. The most significant SNP within the breakpoints was rs12185268 at position 45,846,317, with $p=4.52 \times 10^{-28}$ and δ_{SHR} 0.0119 (N=39,433) (see Figure 11). The greatest δ_{SHR} was 0.0131 for rs17660132 at position 46,088,437 (N=17,128, $p=1.15 \times 10^{-5}$). Both SNPs are located close to the middle of the reported inversion, where the window tested for recombination encompasses the inversion almost entirely. The recombination suppression thus affects most of the region within the window. Accordingly, a rise in significance can be observed in Figure 11 around the middle of the inverted region, both in the microarray data and in the WGS data. The same rise can also be detected within the 8p23.1 inversion with growing distance from the breakpoints, although the significance drops around the middle of the inversion. This drop in significance may be due to double recombination events, which has been suggested to explain the problems finding SNPs that tag the inversion (see section 2.5.1 for details).

The positions of the first and last SNPs of the 17q21.31 SHR region are 45,415,735 and 46,290,846, respectively. The last SNP is within the range given by InvFEST, but the first SNPs are located outside the given proximal breakpoint of the inversion. In our implementation of the SHR test, SNPs that are located outside an inversion, but within 500 kb from its breakpoints, may yield signs of recombination suppression (for more details, see section 3.3.1). We do not observe such results for the 8p23.1 inversion, most likely because of the greater size of its segmental duplications, which harbour almost no SNPs.

In the WGS data, the most significant SNP was rs112454267 at position 45,927,963, with $p=5.85 \times 10^{-7}$ and δ_{SHR} 0.0146 (N=6,077), while the greatest δ_{SHR} was 0.0173 for rs142822273 at 45,629,062 (N=6,191). Remarkably, in the WGS data, 80.0%, of the 4,114 SNPs with a lower mean number of recombination events in heterozygotes within the reported breakpoints of the 17q21.31 inversion yielded p-values below 0.05 from the SHR test. This is a considerably higher percentage than within the 8p23.1 inversion, despite the greater statistical significance of its top SNPs from the SHR test. As can be observed in Figure 11, a large number of SNPs within the 17q21.31 region from 45.6 Mb to 46.25 Mb give similar results from the test. This distinguishes the inversion at 17q21.31 from the other inversions that show a more varied

distribution of p-values, and is due to the remarkably low sequence diversity of the H2 orientation and the extensive divergence of H2 from the H1 orientation, which is estimated to span more than two million years (Stefansson et al., 2005; Zody et al., 2008; Steinberg et al., 2012). As a result, we get similar outcome from our test for multiple SNPs as they have similar sample sizes and MAFs. The reason why we don't see such a clear pattern in the microarray data is because of varying number of individuals genotyped for each SNP, due to the different Illumina chip types used for genotyping.

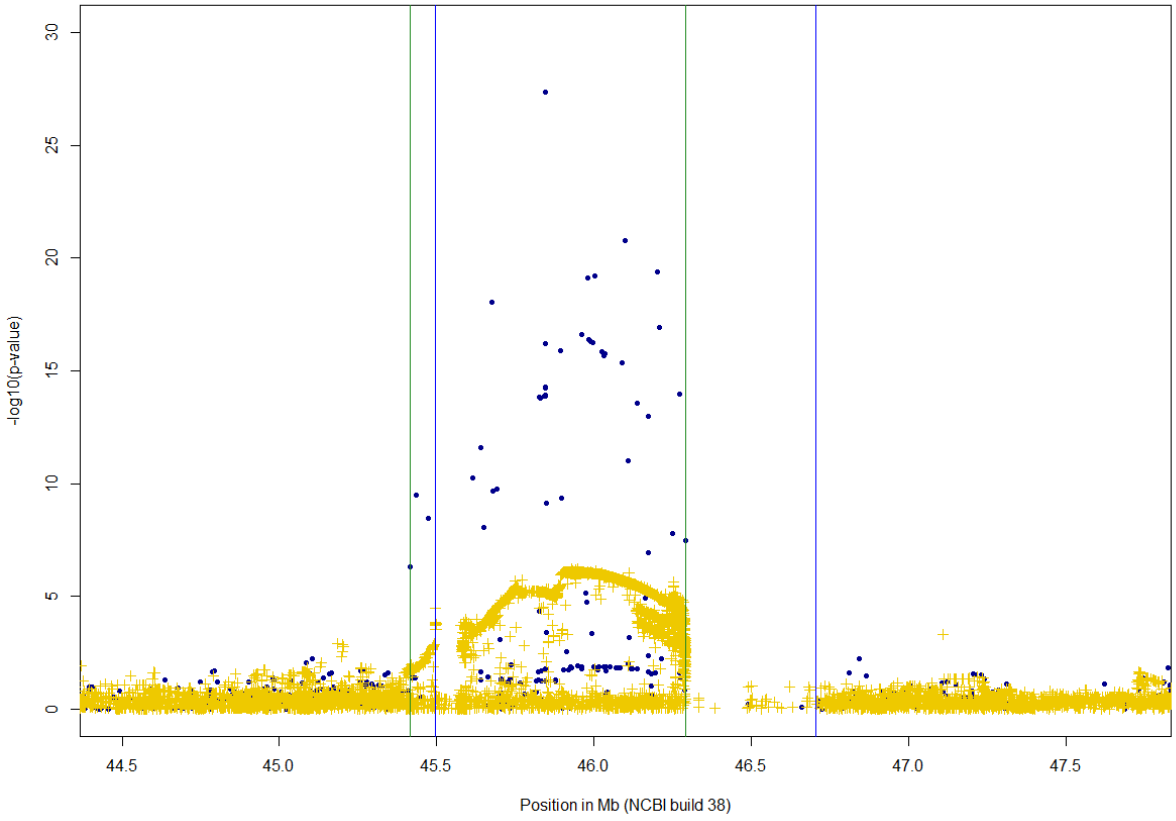

Figure 11. Results from the SHR test in the region containing the 17q21.31 inversion. See Figure legend 10 for details.

4.2.4.3 15q13.3 and 16p11.1

According to previous reports, the 15q13.3 inversion spans around 1.8 Mb (Antonacci et al., 2014). The InvFEST database reports breakpoints between 30,077,909-30,618,102 and 32,153,207-32,607,507 (Martínez-Fundichely et al., 2014), encompassing 492 SNPs in our

microarray data. Of the 253 SNPs that yielded a lower mean number of recombination events in heterozygotes, 50 or 19.8% had a $p<0.05$ from the SHR test. The most significant p-value was 1.07×10^{-23} for rs12442141 (N=39,600, $\delta_{SHR}=0.0323$), and the maximum δ_{SHR} was 0.0396 for rs1075232 (N=21,042, $p=6.54 \times 10^{-21}$).

The first and last SNPs of the 15q13.3 SHR region are at positions 30,906,769 and 32,089,406. Neither SNP is located within the range of the respective reported breakpoints, although both SNPs are within the reported inverted region. As with the other inversions, there is a sparsity of SNPs around the breakpoints of the inversion, due to segmental duplications (see Figure 12).

In the WGS data, 6,946 SNPs were within the 15p13.3 inversion breakpoints, of which 3,724 (53.6%) had a lower mean number of recombination events in heterozygotes than in homozygotes, and of these 465 SNPs, or 12.5% yielded p-values <0.05 from the SHR test. The most significant SNP was rs34959140 at position 31,433,020, with a p-value of 1.37×10^{-9} and the δ_{SHR} 0.0394 (N=6,189). The SNP with the most δ_{SHR}, of 0.0415, was rs72709326 at position 31,681,426 (N=6,189, $p=4.6 \times 10^{-6}$). Both SNPs are close to the centre of the inversion, as was the case within the 17q21.31 inversion.

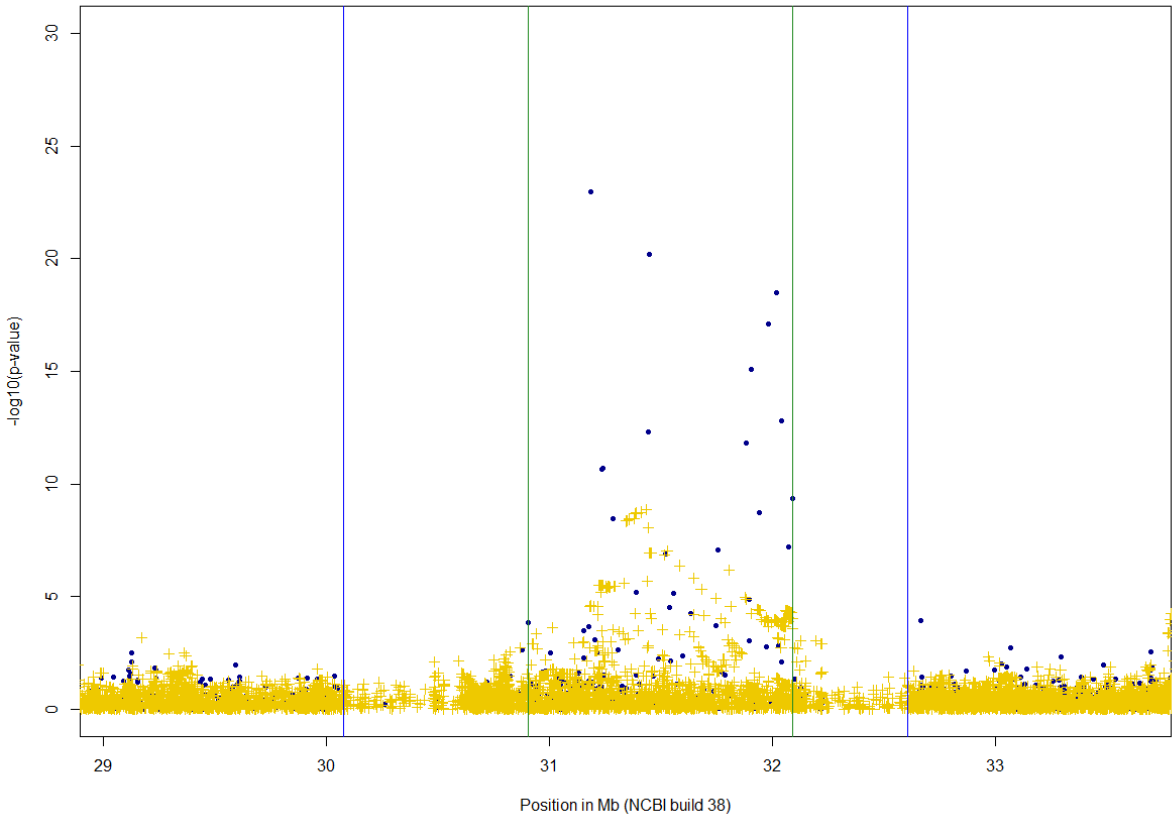

Figure 12. Results from the SHR test in the region containing the 15q13.3 inversion. See Figure legend 10 for details.

The inversion at 16p11.2 is the smallest of the four known ones, spanning around 450 kb (González et al., 2014). The breakpoints are located between 28,337,952-28,471,892 and 28,643,181-28,777,130 (Martínez-Fundichely et al., 2014). A total of 54 SNPs in the microarray data are within the inversion breakpoints, of which 43 have a lower mean number of recombination events in heterozygotes, and 20 (46.5%) yielded p<0.05 in the SHR tests. The most significant p-value was 4.87 x 10^{-10} for rs8049439 (N=39,564, δ_{SHR} 0.0074), and the greatest δ_{SHR} was 0.0090 for rs4788069 (N=15,958, p=1.01 x 10^{-4}).

In the WGS data, 754 SNPs were within the reported breakpoints of the inversion, of which 535 (71.0%) had a lower mean number of recombination events in heterozygotes than homozygotes. 124 of these (23.2%) had a p-value under 0.05. The most significant SNP, rs3020804 at position 28,607,315, had p= 1.02 x 10^{-3} (N=6,177, δ_{SHR}=0.0112), and the greatest δ_{SHR} was 0.0192 for rs117985404 at 28,588,700 (N=6,196, p=4.66 x 10^{-2}). The suppression of

recombination in heterozygotes within the inversion at 16p11.2 is therefore not as evident in the WGS data as in the microarray data although we see a greater δ_{SHR} in the WGS data (see Figure 13).

The SNPs that demarcate the 16p11.2 SHR region are located at 28,473,820, which is within the inversion breakpoints, and 28,976,948, well outside the breakpoints according to InvFEST (Martínez-Fundichely et al., 2014). Figure 13 shows that multiple SNPs located outside the reported breakpoints show recombination suppression in heterozygotes, among them four SNPs that yield greater statistical significance than the strongest SNP within the reported breakpoints. This may be due to their proximity to the inversion. Due to its small size, a relatively large number of SNPs associated with suppression of recombination may be located outside the inversion. It is noteworthy, however, that we don't see this effect of proximity to the inversion on SNPs located close to the proximal breakpoint of the inversion, although there are multiple SNPs within 500 kb from the breakpoint.

Of the four known inversions, 16p11.2 yields the smallest δ_{SHR} values, and the least statistical significance despite being the most common one. Thus, the reported frequency of the inversion in Northern Europe is 49% (Gonzáles et al., 2014), and the most frequent MAF of SNPs with p-values under 0.05 in the WGS data is 44%. One reason why results for the 16p11.2 inversion are not as clear as the other three known inversions is the radius of 500kb around each SNPs, which is likely to be less effective for smaller inversions, as the recombination suppression will only affect part of the sequence within the window.

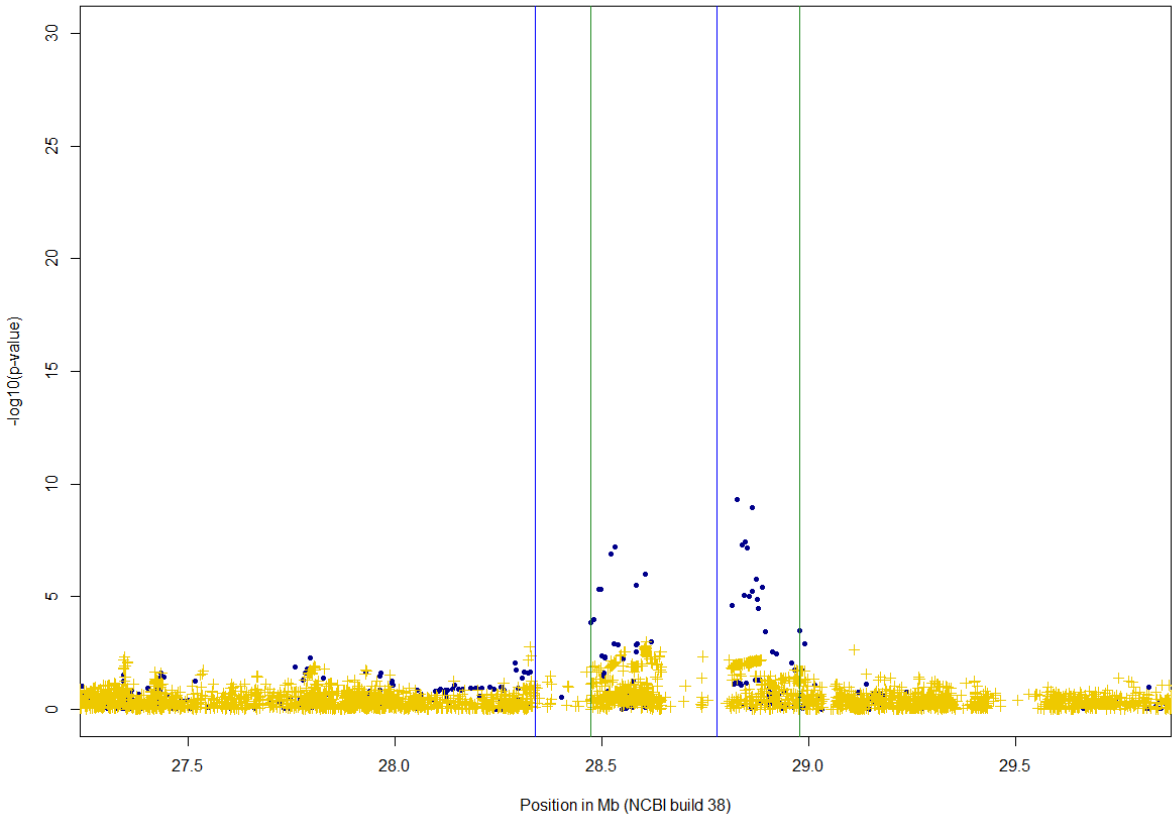

Figure 13. Results from the SHR test in the region containing the 16p11.2 inversion. See Figure legend 10 for details.

4.2.4.4 Pattern of results for the known common inversions

From this initial examination of these four known inversions and comparison with their reported sizes and breakpoints, it is evident that we are able to detect clear signals of all four known inversions using the SHR test in the microarray data and three of the four in the WGS data. We see considerably stronger results from the test within the 8p23.1 inversion than the other three, which is likely to be due to the combined impact of its size and its high MAF. Due to the low frequency of the 15q13.3 inversion, which according to the latest estimates is only around 6% (Antonacci et al., 2014), its signal according to the SHR test is relatively weak, even though it is the second largest of the known inversions. Conversely, the smallest of the four known inversions at 16p11.2, has a relatively high reported MAF of 0.49, but evidently requires large sample sizes for detection through the SHR test – as witnessed by the weak signal detected in the WGS data. We do however observe greater δ_{SHR} for the most significant

SNPs within the 15q13.3 inversion than within the two smaller ones. Thus, the δ_{SHR} reveals correlation of the SNP to the inversion, but it is also affected by the size of the inversion. The ability to determine the significance of the results is then affected by the MAF of the SNP, along with the sample size in the microarray data.

All plots show a sparsity of SNPs close to the breakpoints of the inversions due to segmental duplications surrounding the breakpoints. Inversion breakpoints are frequently positioned within such areas (see section 2.2), therefore, the paucity of SNPs within such regions, and the variable density of SNPs throughout the genome entails that the assessment of breakpoint positions based on our results from the microarray data cannot be accurate.

4.2.5 Regions with previous reports of inversions not experimentally validated

In addition to the four known inversions, there are six candidates identified through our approach which overlap with previous reports based on sequencing or paired-end mapping (PEM) (Kidd et al., 2008, The 1000 Genomes Project Consortium, 2010), that have not been experimentally validated with FISH or similar methods.

In 2008, Kidd et al. generated an extensive map of structural variation in the human genome by comparing the genomes of eight individuals to the human reference using PEM. This study identified 224 inversions, of which 7 overlap with one of our candidate regions for inversions. Of these, two were validated with FISH, namely the known inversions on chromosomes 15 and 17. The other 5 inversions are in chromosome bands 7q11.21, 9q33.1, 10q22.3, 11p11.12 and 16p11.1.

The results of the 1000 Genomes pilot project, which were published in 2010, based on sequence data from 179 individuals and two mother-father-child trios and identified a large number of structural variants in the human genome, including inversions (The 1000 Genomes Project Consortium). Three of these inversions coincide with our candidate regions, located on chromosome bands 11p11.12, 12q11-q12 and 16p11.1, two of which also overlap with candidates found by Kidd et al. (2008).

The region that shows the most significant difference in recombination rates, apart from the four known and validated inversions, is at chromosome band 10q22.3. The positions of the SNPs that mark the region are 79,637,253 and 80,180,945. Of the 90 SNPs that had a lower

mean number of recombination events in heterozygotes, 21 (23.3%) had a SHR p-value lower than 0.05, and 6 SNPs lower than the Bonferroni corrected significance level. An inversion in this region was found by Kidd et al. (2008) in three samples. Its breakpoints were reported 79,500,916 and 80,221,876 (see Figure 14).

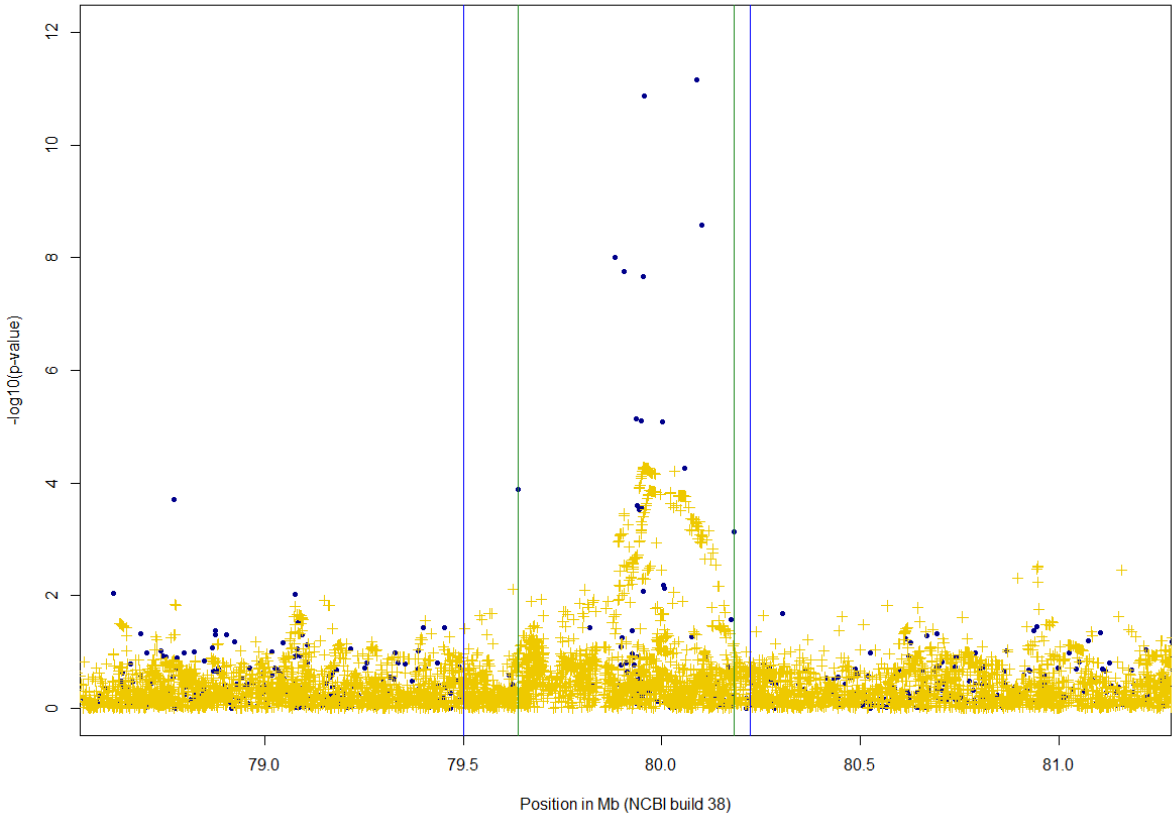

Figure 14. Results from the SHR test in the region containing the 10q22.3 inversion. See Figure legend 10 for details.

It must be taken into consideration that in their study, Kidd et al. (2008) referred to NCBI build 35, while the identification of recombination events used in our study, was done in NBCI build 36 (Kong et al., 2014). While there were few modifications within this region between builds 35 and 36, considerable changes were made between builds 36 and 37, so that the sequence length between the breakpoints of the inversion reported by Kidd et al. (2008) dropped from 1 Mb in builds 35 and 36 to just over 500 kb in builds 37 and 38 (Kent et al., 2002). It is therefore possible that errors in the previous assemblies affected both the results

of the PEM and the recombination calling. However, the δ_{SHR} observed within the 10q22.3 region in the WGS data is 0.0265 (N=6,161, p=3.75 x 10^{-2}), greater than both within the 16p11.2 and the 17q21.31 inversions (see Table 8), and is surpassed only by 8p23.1 and 15.13.3, the largest of the known inversions. This suggests that the inversion is larger than 500 kb, which raises the question if the more recent assemblies are incorrect, or if the size of this region may be polymorphic.

The SNPs that demarcate the SHR region at 7q11.21 are positioned at 65,305,569 and 66,200,069. The region overlaps by 399 kb with a 556 kb inversion, also found by Kidd et al. (see Figure 15), with breakpoints 65,149,138 and 65,704,935. The regions around the inversion breakpoints are rich with segmental duplications. Although our SHR test results indicate that the effect on recombination may be detected for SNPs outside the breakpoints of an inversion, the SHR region stretches quite far beyond the reported distal breakpoint. The SNP that demarcates the distal end of the region, which is one of two SNPs within the region that survive a Bonferroni correction, is positioned 495 kb from the breakpoint, suggesting that the inversion may be larger than previously reported.

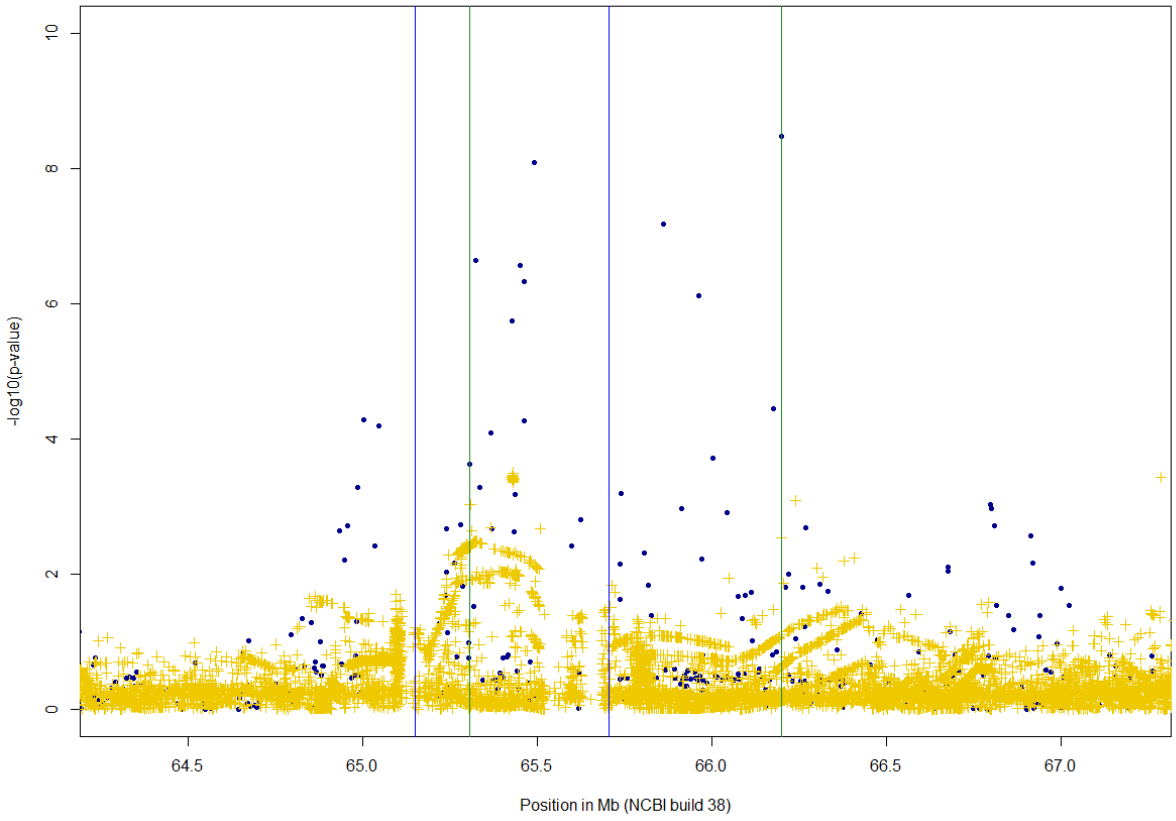

Figure 15. Results from the SHR test within the 7q11.21 region. See Figure legend 10 for details.

Another inversion reported by Kidd et al. (2008) is on chromosome band 9q33.1. It is almost 6 Mb in length, with reported breakpoints of 114,900,117 and 120,805,395. As it was only found in one sample, we do not know if this inversion is polymorphic. The SHR region that falls within the reported breakpoints is considerably smaller, from position 116,538,497 to 116,691,176, which suggests that our signal may not be due to an inversion of the same size as that found by Kidd et al. (see Figure 16).

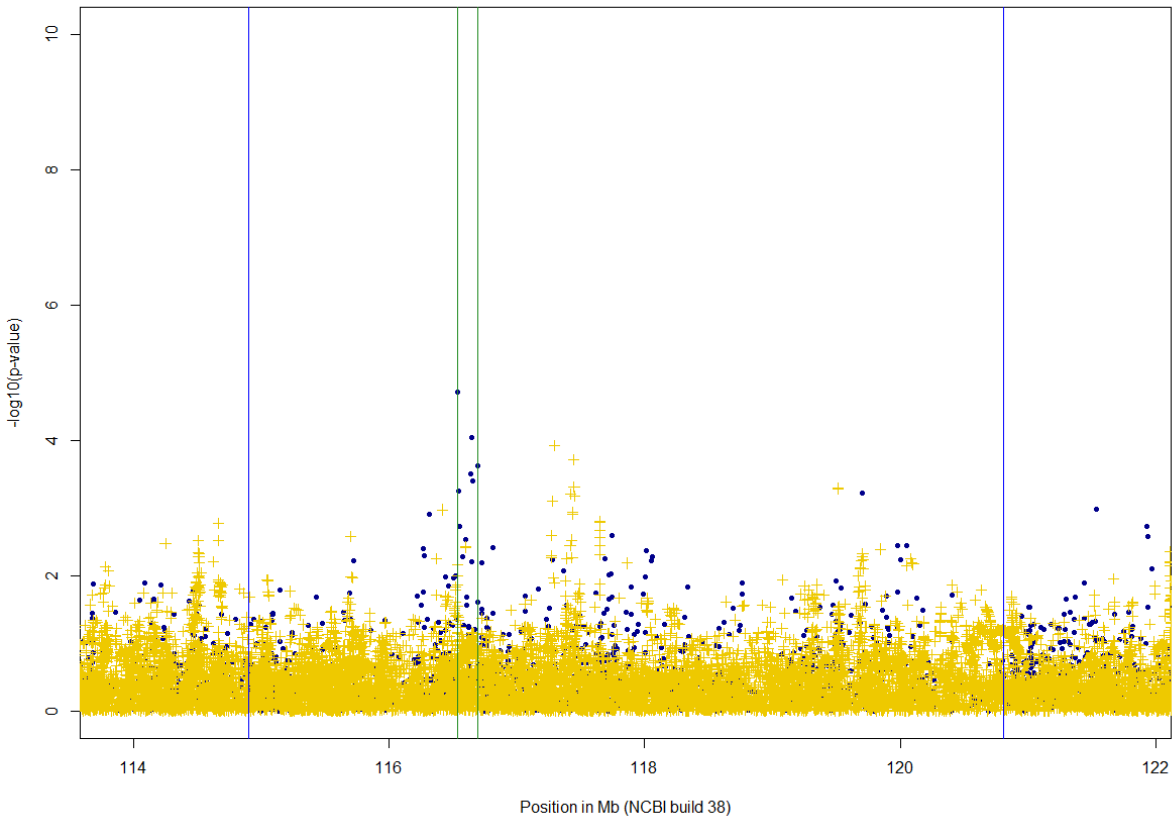

Figure 16. Results from the SHR test within the 9q33.1 region. X-axis adjusted to the size of the reported inversion. See Figure legend 10 for details.

4.2.6 Centromeric regions

Of the 34 SHR regions, 8 are positioned close to centromeres. We found previous reports of inversions within three of these regions. As the centromeres are large regions of repetitive DNA, these regions are difficult to map, and mapping errors may affect the results of the SHR testing. The observation that the SNPs within the centromeric SHR regions show little difference in recombination between the two groups, both in the microarray data and in the WGS data, raises suspicion that something other than inversion polymorphisms may be causing this effect. However, in the case of the SHR region on chromosome band 11p11.12, there are two separate reports of an inversion found, although in the same sample, at this location.

The breakpoints of the 11p11.12 inversion reported by Kidd et al. (2008) are 50,131,500 and 50,421,805, and the breakpoints reported by the 1000 Genomes Project Consortium are close to the previously reported breakpoints, at 50,165,565 and 50,424,957 (MacDonald et al., 2014). Our region comprises a slightly larger area, from position 49,866,070 to 50,662,884 (see Figure 17). Within the SHR region there were 125 SNPs, of which 120 (96%) had a lower mean number of recombination events in heterozygotes. In the WGS data, 82.4% of the SNPs also had a lower mean of recombination events in heterozygotes, although only 5.9% of these had $p<0.05$. The maximum δ_{SHR} value found within the region was 0.0011 in the microarray data. Figure 17 reveals that there is a rise in significance in both datasets with growing proximity to the centromere, although this starts considerably closer to the centromere in the WGS data, and is mostly outside the breakpoints of the reported inversion. The significance we observe within this region is not particularly strong and the δ_{SHR} seems small for an inversion of the size reported, although we may be detecting signs of a smaller inversion than previously reported. Figure 17 shows that there are multiple SNPs in the WGS that are correlated, similar to what we saw for 17q21.31, which is a pattern that can be observed in most of the centromeric regions.

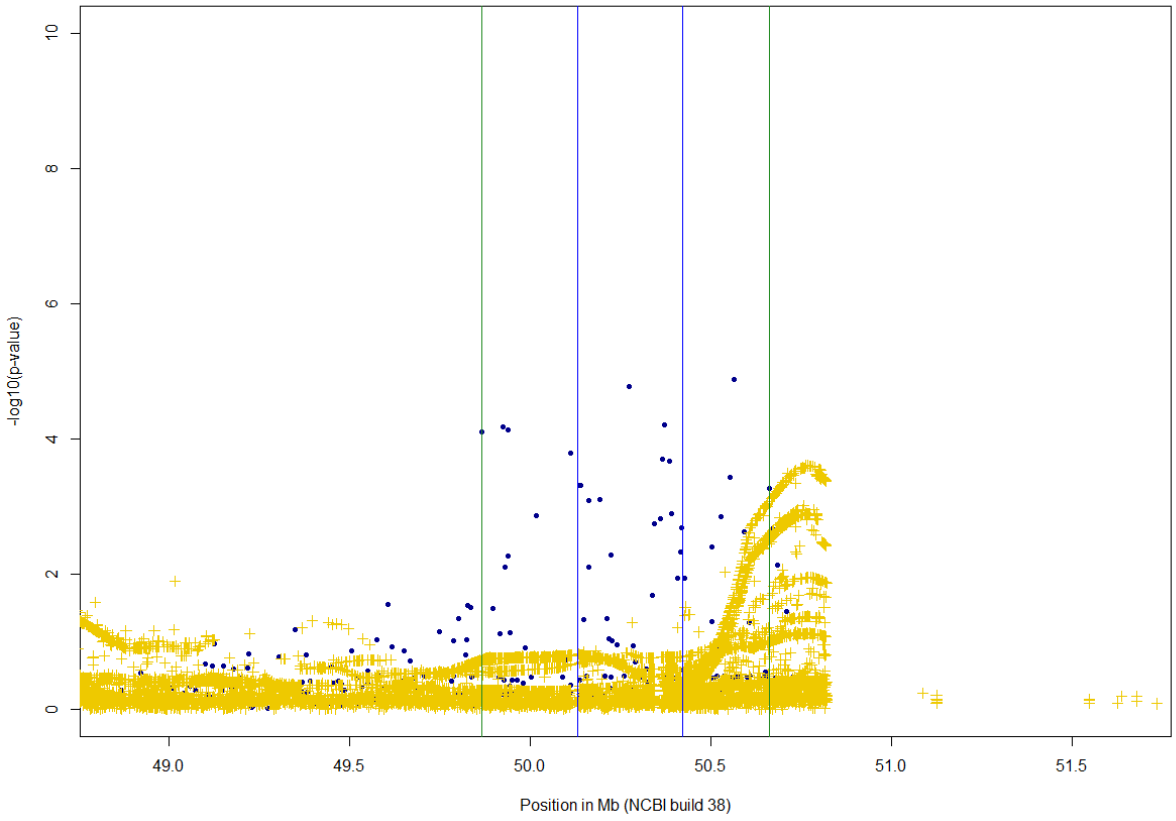

Figure 17. Results from the SHR test within the 11p11.12 region. See Figure legend 10 for details.

Another inversion detected in the 1000 Genomes pilot project that coincides with one of the SHR regions is located close to the centromere on chromosome 12, at 12q11-q12. According to the Database of Genomic Variants (n.d.), the inversion was found in one sample and was not verified with other methods. The size of the inversion is 34,420 bp, with breakpoints at positions 37,704,131 and 37,738,550 (see Figure 18).

The SNPs that demarcate our SHR region at 12q11-q12 are at positions 37,533,312 and 37,851,950, spanning almost 319 kb. In the microarray data, 21 of 22 SNPs have a lower mean of recombination events in heterozygotes and 12 have $p<0.05$. In the WGS, 1,975, or 90.7% of 2,177 SNPs have fewer recombination events in heterozygotes, and 26.2% have $p<0.05$. The greatest δ_{SHR} detected was in the microarray data, 0.0012, which is not a particularly strong effect, although if the difference is due to an inversion of only 34 kb, the effect would be small because of the small size of the inversion compared to the window size under consideration.

As in the case of the 11p11.2 region, we see a rise in significance with proximity to the centromere, and pattern of “layers” of SNPs can also be observed for 12q11-q12 (Figure 18).

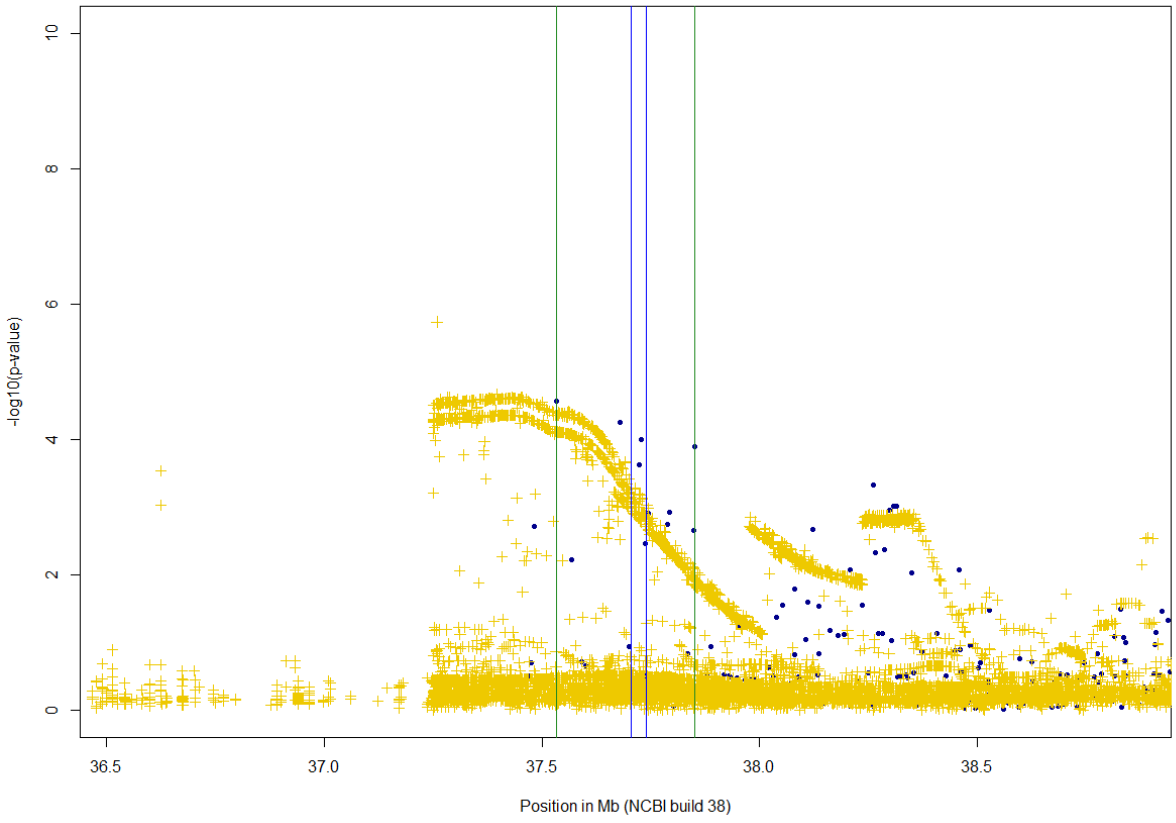

Figure 18. Results from the SHR test within the 12q11-q12 region. See Figure legend 10 for details.

The region on chromosome band 16p11.1 is also close to the centromere, spanning from position 35,452,449 to 35,867,952. There are 26 SNPs located within the region and 24 have a lower mean of recombination events in heterozygotes. Of these, 15 yielded $p<0.05$ in the SHR test, with the most significant being rs11646602 ($p=3.83 \times 10^{-10}$). Kidd et al. (2008) reported an inversion found in two samples with breakpoints 35,124,161-35,518,326, and the 1000 Genomes pilot project also revealed an inversion in one of these two samples (Database of Genomic Variants, n.d.). The breakpoints were, according to the Database of Genomic Variants, at positions 35,157,199 and 35,522,883.

As can be observed in Figure 19, our region only partially coincides with the inversion reported by Kidd et al. and 1000 Genomes Project Consortium. The sparsity of SNPs at the

location, however, affects our ability to estimate the size and shape of the region under recombination suppression. The rise in significance we detect may be the result of the inversion previously reported, as the last SNP within our region is within 500 kb from the distal breakpoint. However, the rise in significance we observe in the WGS data seems to peak close to the distal end of our SHR region, around the position of one of three microarray Bonferroni survivors within the region. Thus the pattern in the WGS data suggests that this signal is not due to the inversion reported by Kidd et al.

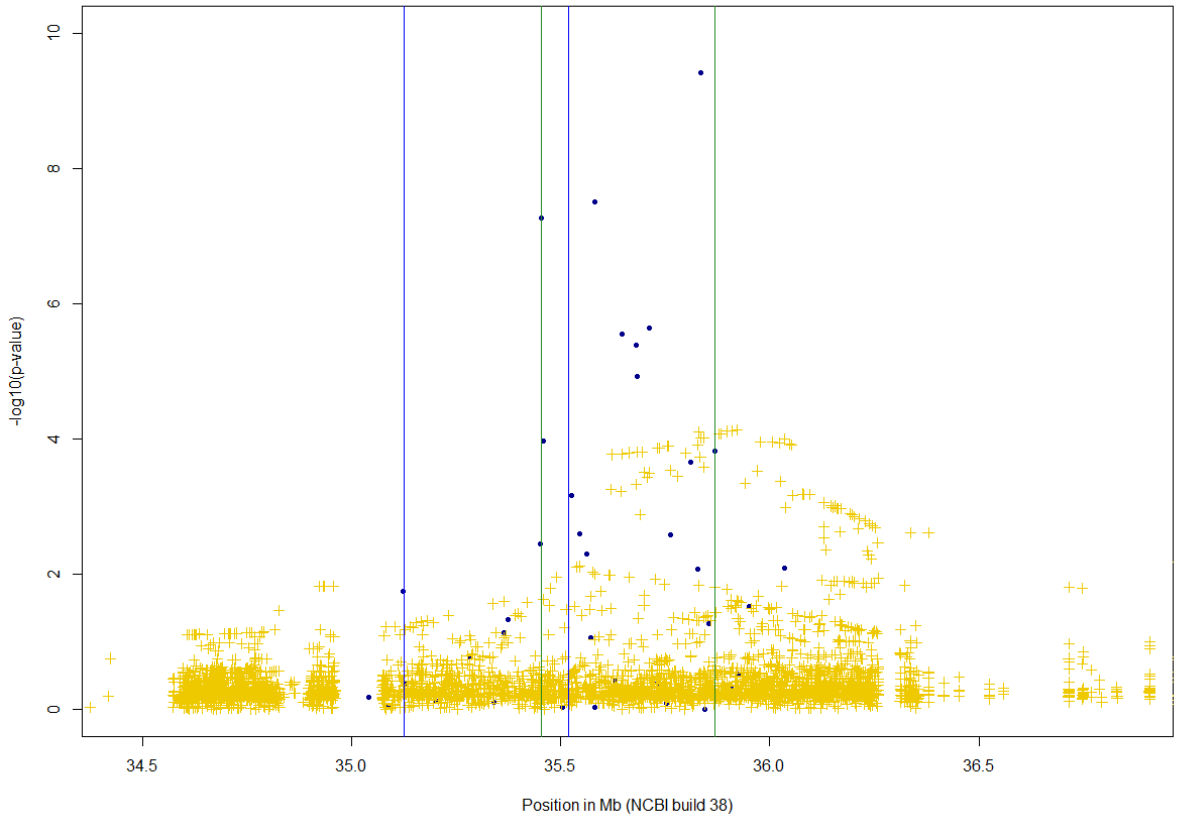

Figure 19. Results from the SHR test within the 16p11.1 region. See Figure legend 10 for details.

Other centromeric regions identified through our test were at chromosome bands 5p11, 8q11.1, 11q11, 12p11.1 and 18q11.1 (see Figures 20 through 24). We were not able to find previous reports of inversions within these regions. Three of them harbour SNPs with p-values lower than the Bonferroni corrected significance level, one is positioned within the 18q11.1 region, one within 5p11 and six SNPs within the 8q11.1 region. Although the difference in

recombination rate is significant, the δ_{SHR} values are relatively small, as is the case for all the centromeric SHR regions (see Table 8).

Within most of these centromeric regions we observe a rise in significance towards the centromeres and groups of SNPs with similar p-values. Such groups of SNPs can be observed within some of the other SHR regions, although nowhere as prominent as within the 17q21.31 inversion. Within the centromeric regions we see the recurrent pattern of SNPs with similar results from the SHR test, indicating long regions of strong LD. It should be noted that recombination rates are generally low in the centromeric regions, and LD is therefore strong (Kong et al., 2010). While it is possible that these properties of centromeres contribute to the results of the SHR tests in these regions, it is not clear why strong LD would suppress recombination in heterozygotes relative to homozygotes.

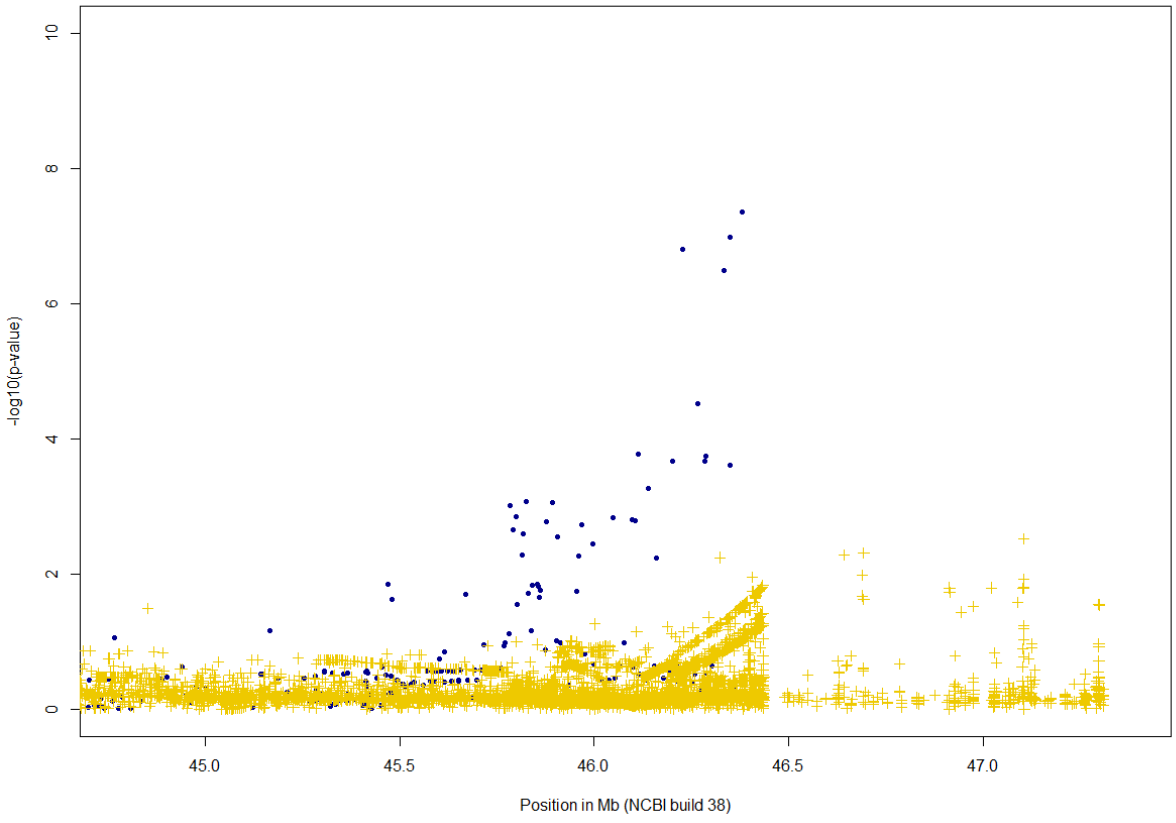

Figure 20. Results from the SHR test within the 5p11 region. See Figure legend 10 for details.

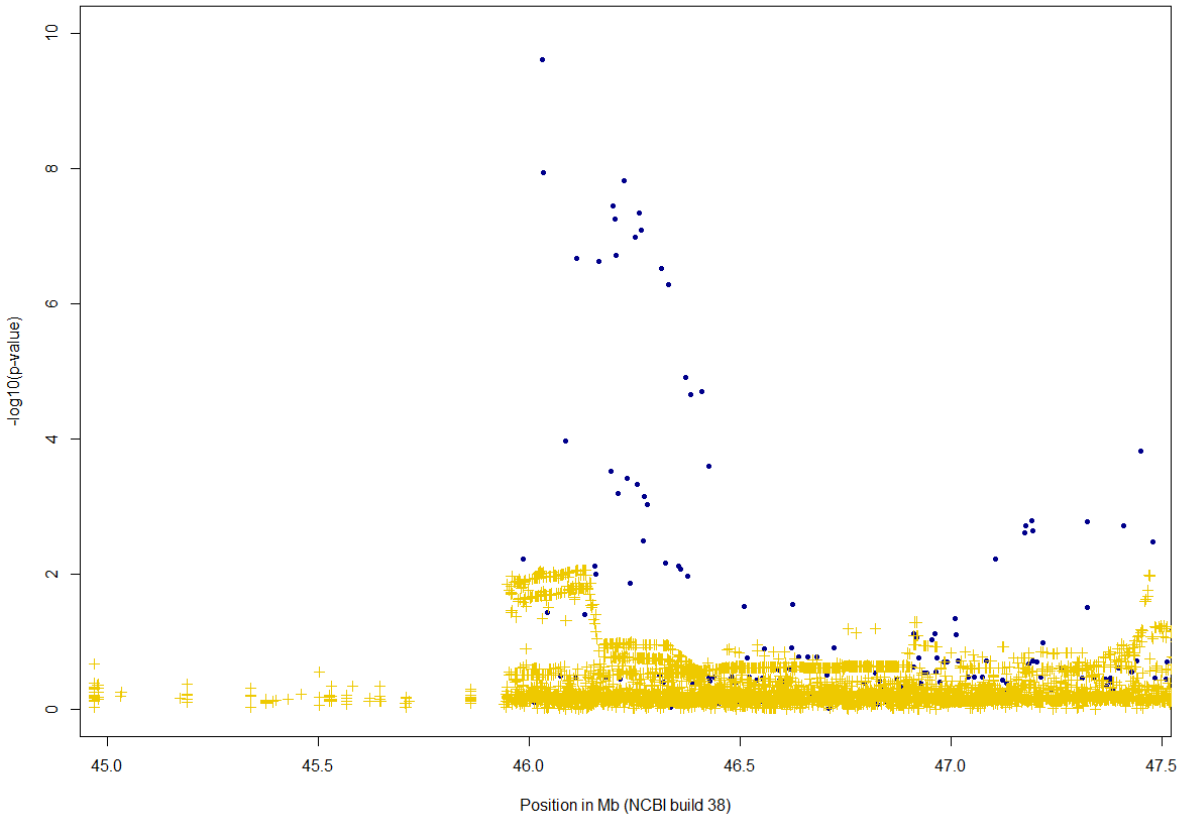

Figure 21. Results from the SHR test within the 8q11.1 region. See Figure legend 10 for details.

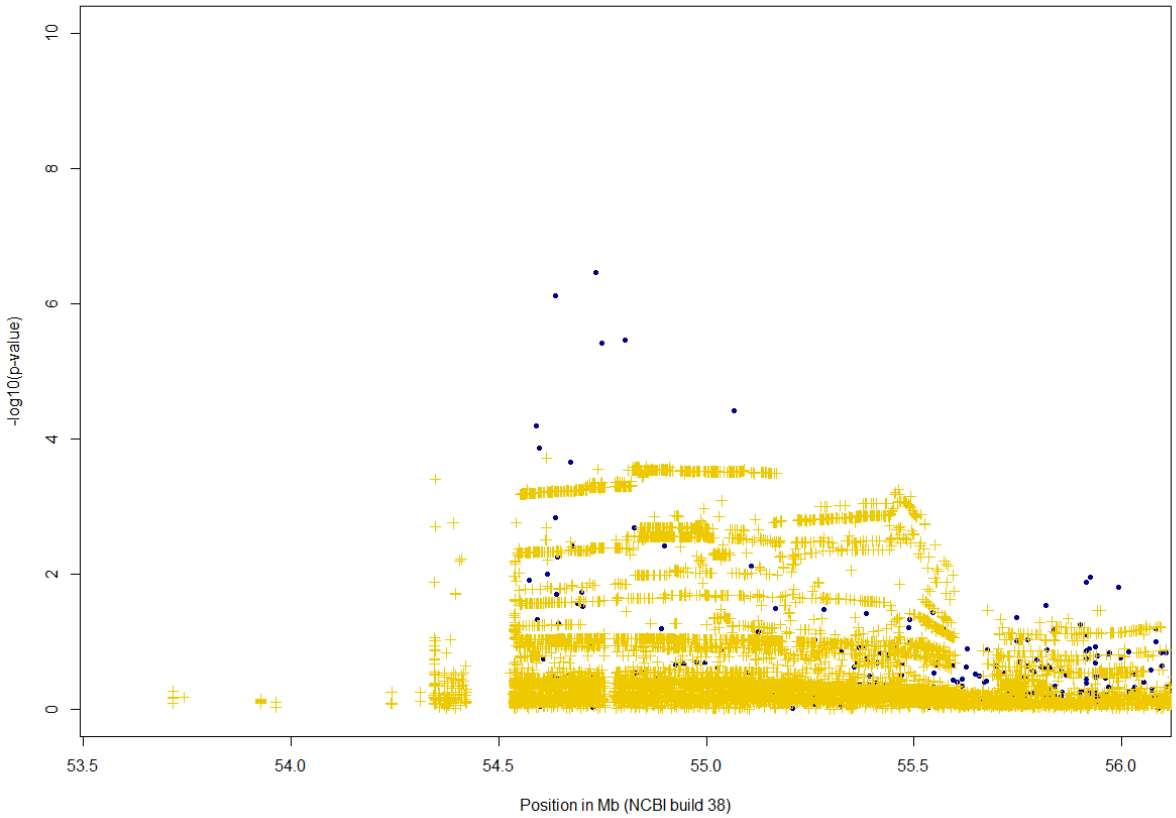

Figure 22. Results from the SHR test within the 11q11 region. See Figure legend 10 for details.

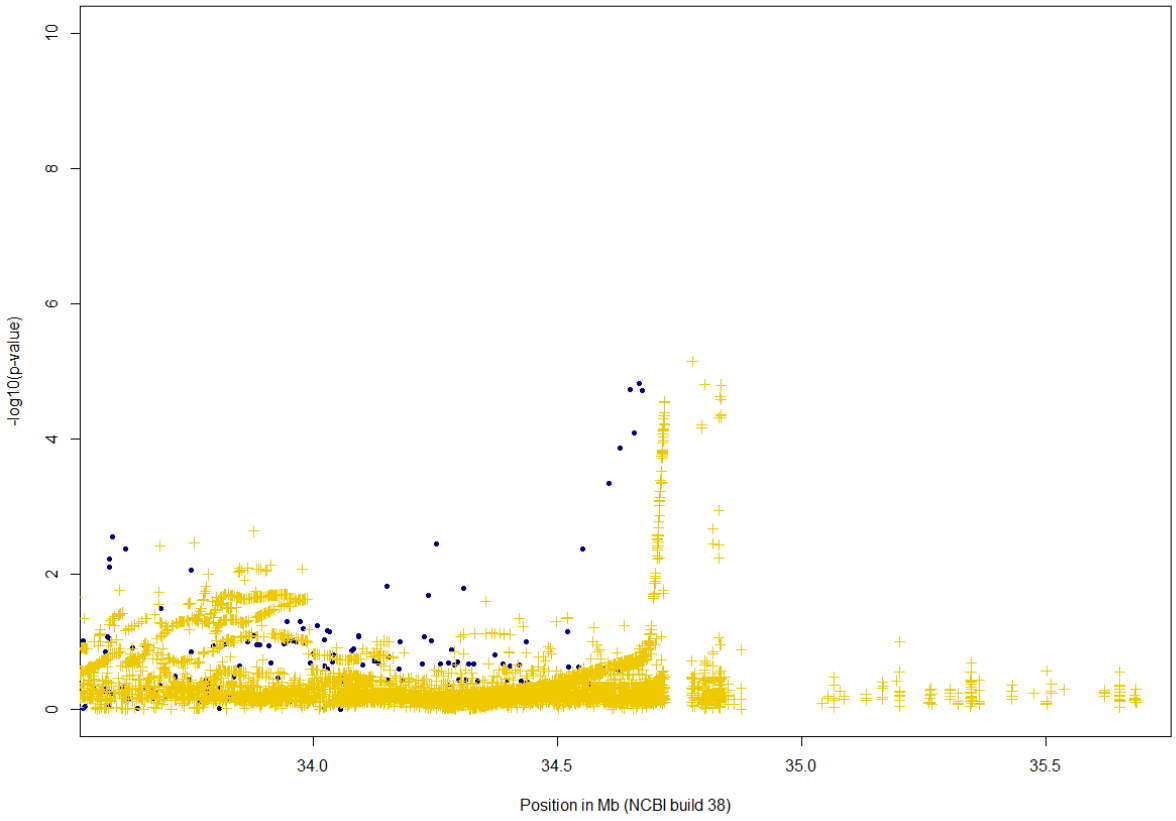

Figure 23. Results from the SHR test within the 12p11.1 region. See Figure legend 10 for details.

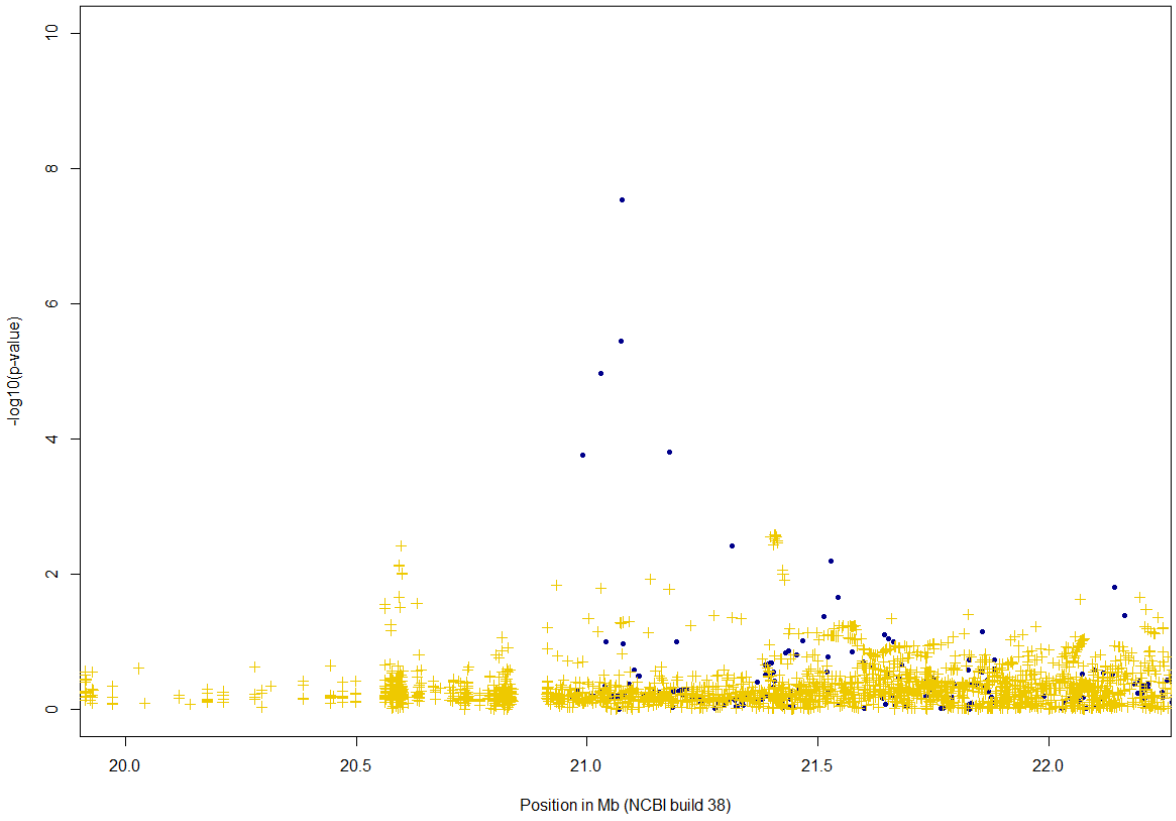

Figure 24. Results from the SHR test within the 18q11.1 region. See Figure legend 10 for details.

4.2.7 Regions with no previous reports of inversions

We were unable to find previous reports of inversions for 24 of the 34 SHR regions. Five of these regions are the aforementioned centromeric regions. Two of the regions are within the major histocompatibility complex (MHC), on chromosome bands 6p22.1 and 6p21.33 (see Figures 25 and 26). The significance is not strong within these regions and there is not much support in the WGS data, where only 2% of 1,834 SNPs within 6p22.1 and none within 6p21.33 have $p<0.05$. There is a high degree of variation within the MHC, as well as strong LD (Jobling et al., 2014), which may produce these signals in some way. Also, because it is a region of great scientific interest, there is an unusually high density of microarray SNPs in the MHC region. This may explain why these regions are picked up by our approach, as it targets regions by the

number of SNPs under a certain level, not proportion, rendering regions of high SNP density more likely to be identified by chance.

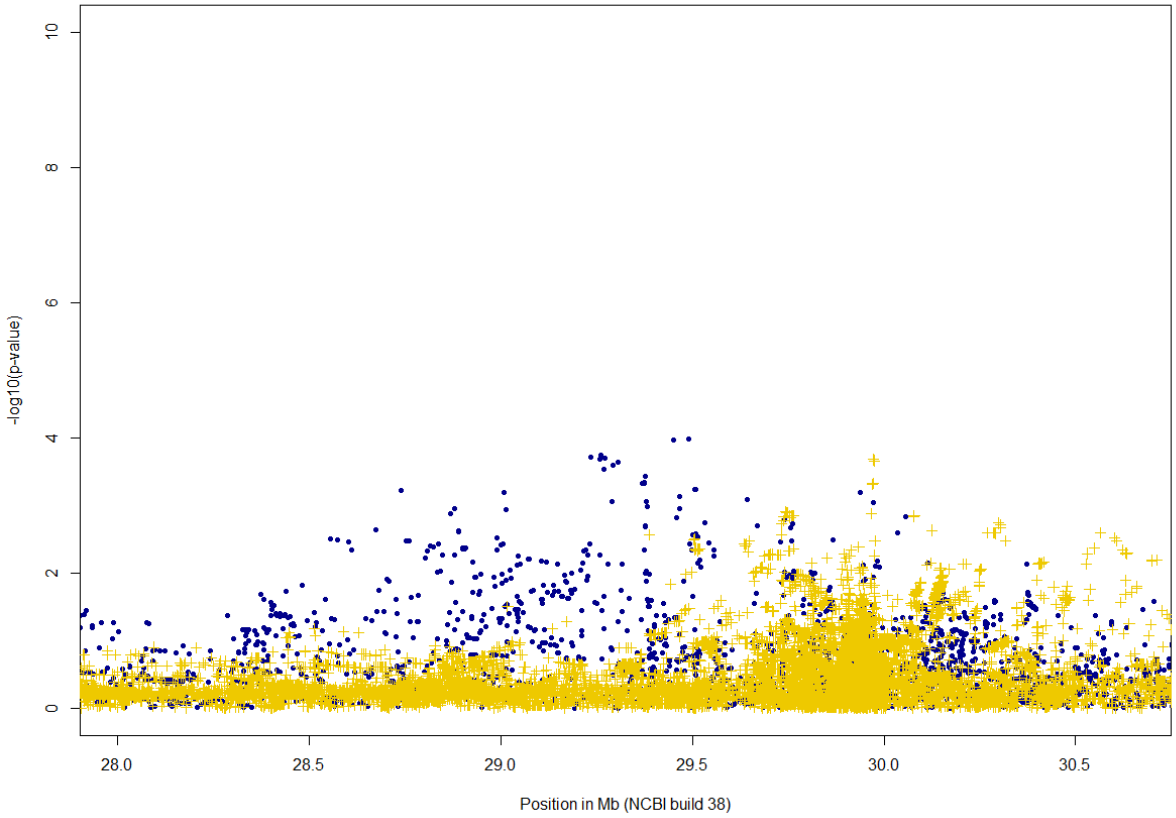

Figure 25. Results from the SHR test within the 6p22.1 region. See Figure legend 10 for details.

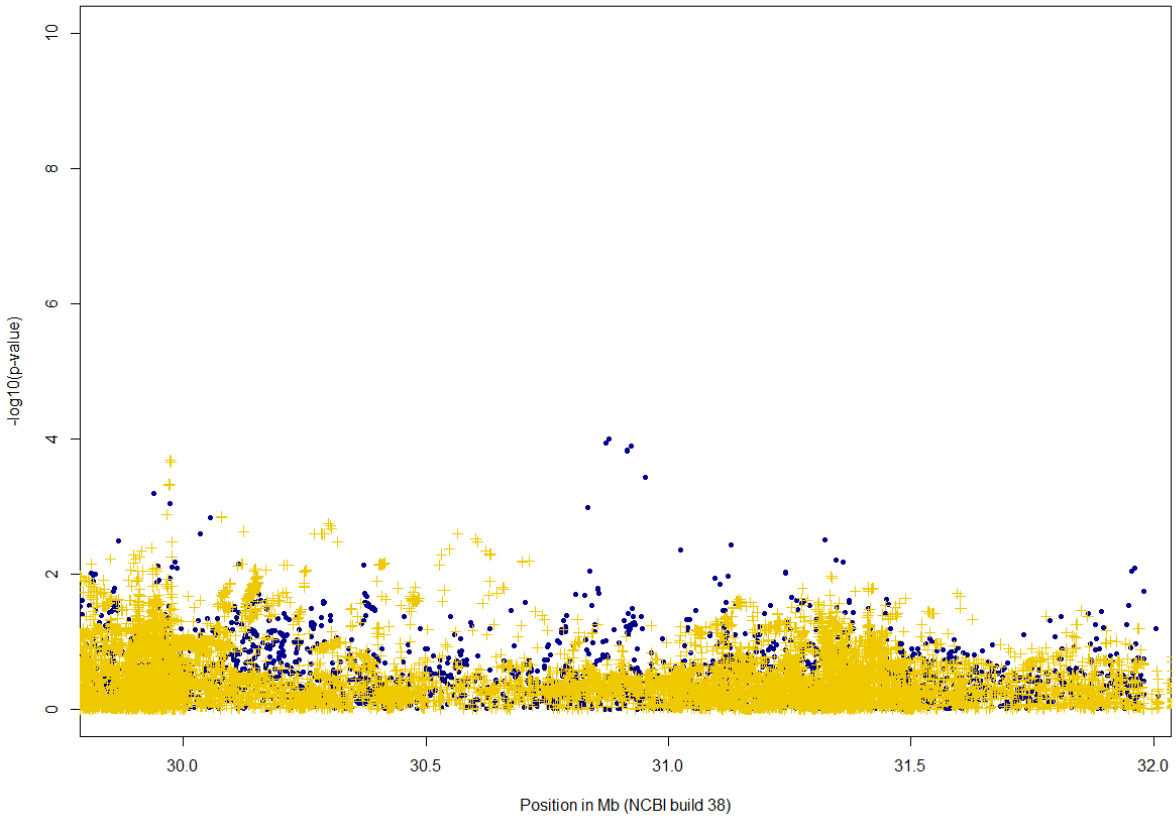

Figure 26. Results from the SHR test within the 6p21.33 region. See Figure legend 10 for details.

For many of these regions, we observe only a very weak signal of lower recombination in heterozygotes in the WGS data. However, there are a few exceptions. For example, within the region on chromosome band 2p22.3, 27.7% of 1,086 SNPs have $p<0.05$ in the SHR test (see Figure 27). Although the signal of suppressed recombination in heterozygotes is not particularly strong within the region on chromosome band 6q24.3 in the microarray data (lowest $p=6.83 \times 10^{-5}$), the majority of SNPs within the region (77.8% in the microarray data and 88.3% in the WGS data) have a lower mean of recombination events in heterozygotes, and a high percentage, 57.1% in the microarray data and 38.4% in the WGS data, yielded $p<0.05$ in the SHR tests (see Figure 28). The δ_{SHR} is small, however, but the region is one of the smallest ones identified, just under 90 kb.

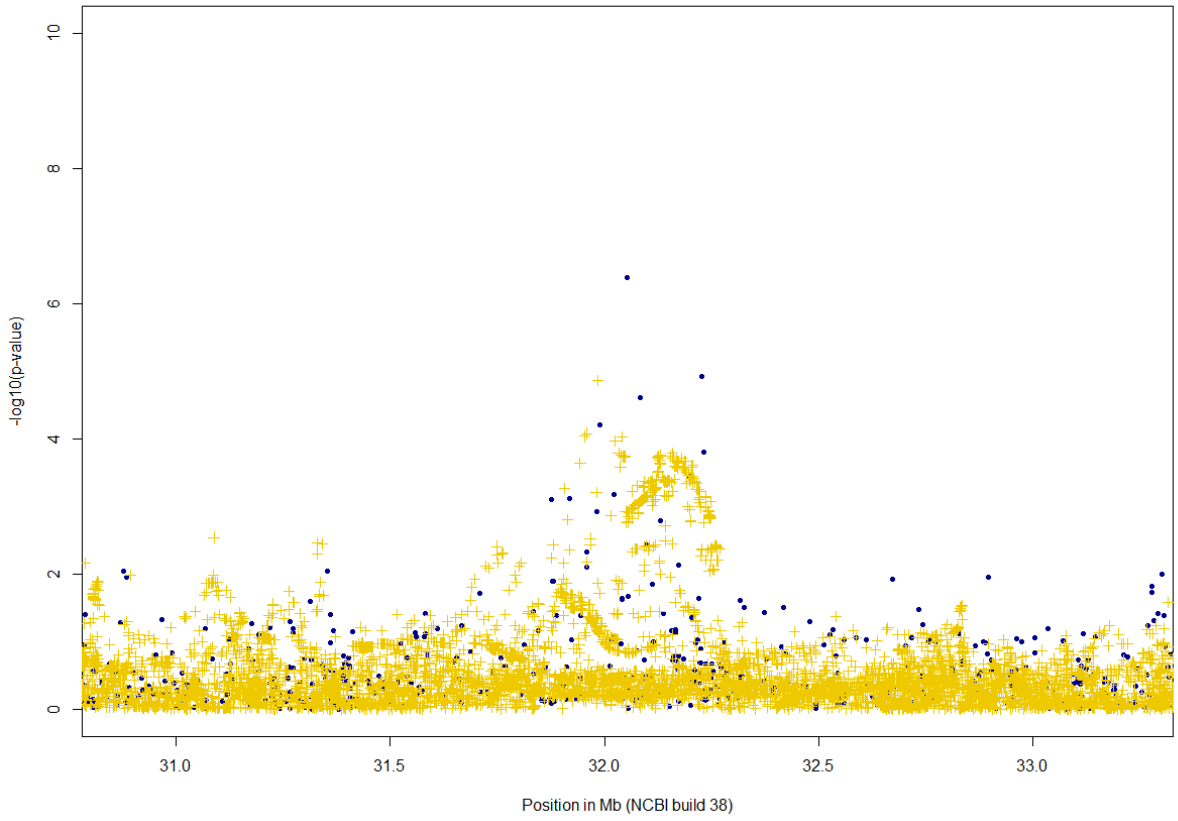

Figure 27. Results from the SHR test within the 2p22.3 region. See Figure legend 10 for details.

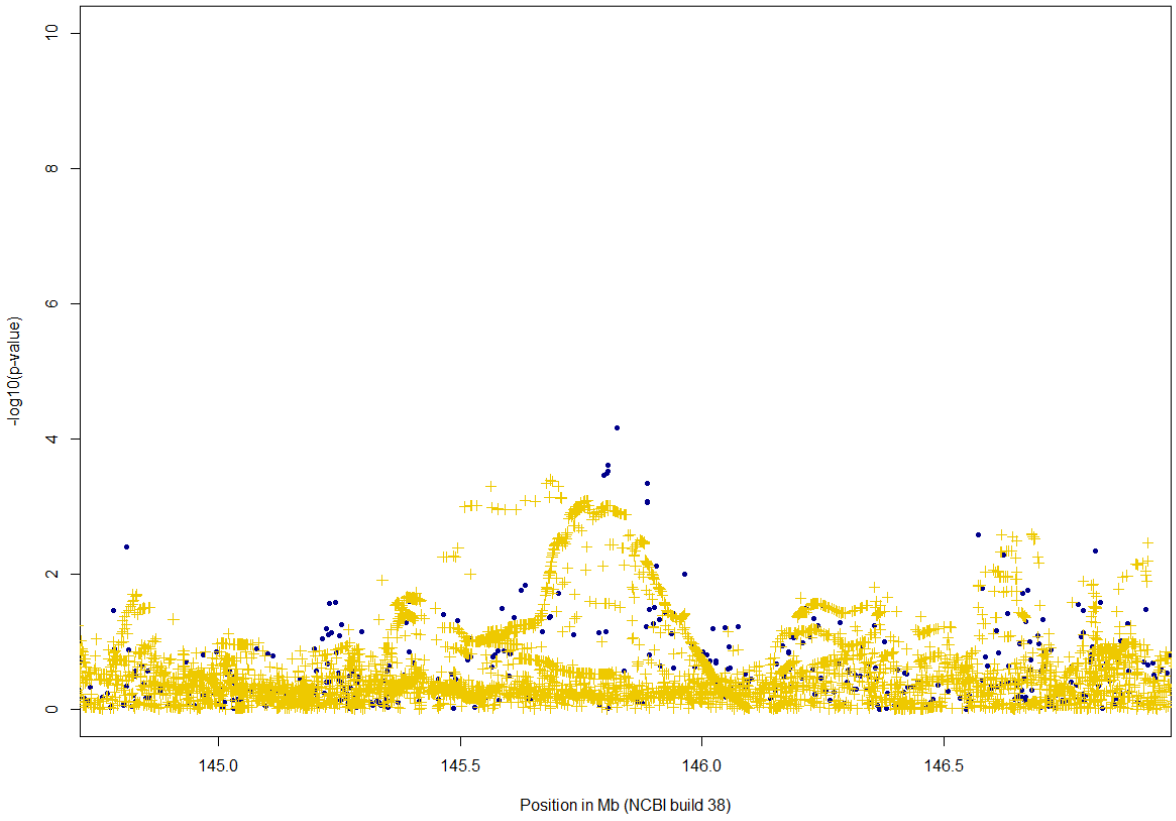

Figure 28. Results from the SHR test within the 6q24.3 region. See Figure legend 10 for details.

Among the lowest p-values in the microarray data, within the SHR regions that do not have validated inversion, is found at 2q21.3, where four SNPs survive Bonferroni correction. The most significant attained p-values were 3.26×10^{-10} (N=39,573, δ_{SHR}=0.0069) and 1.9×10^{-5} (N=6,194, δ_{SHR}=0.0091) in the microarray data and in the WGS data, respectively (see Figure 29), and 85% of the 1,277 SNPs within the region in the WGS yielded a lower mean number of recombination events in heterozygotes. The maximum δ_{SHR} was 0.0105 in the WGS data (N=6,196, $p=4.41 \times 10^{-3}$). The significance of the results in the microarray data, along with a relatively large δ_{SHR} make this an interesting candidate for an inversion. Other regions where we see SNPs under the Bonferroni significance level are 1p33, and 12q11-q12. One SNP passes the Bonferroni threshold within the 1p33 region. Although only 1.2% of the 492 SNPs in the WGS data yields p<0.05, Figure 30 shows a small rise in significance within the region. The maximum δ_{SHR} is quite small, only 0.0029 in the WGS data (N=6,197, $p=3.59 \times 10^{-3}$), which suggests that we may be observing a signal of a small inversion, explaining the weak signal in

the WGS data. The region at 12q24.12-q24.13 has two Bonferroni survivors, with the lowest p-value of 4.12 x 10^{-9} (N=39,574, δ_{SHR}=0.0033) (Figure 31). In the WGS data, 13.6% of SNPs have a p-value under 0.05, and the lowest p-value we see is 4.89 x 10^{-4} (N=6,191, δ_{SHR}=0.0039) The maximum δ_{SHR} is small however, only 0.0049 (N=6,192, p=2.52 x 10^{-2}), but the size of the SHR region is just under 400 kb. Given the significance in the microarray data, and the support we see in the WGS data, we consider this a good candidate for an inversion, although in the light of the δ_{SHR}, likely a small one.

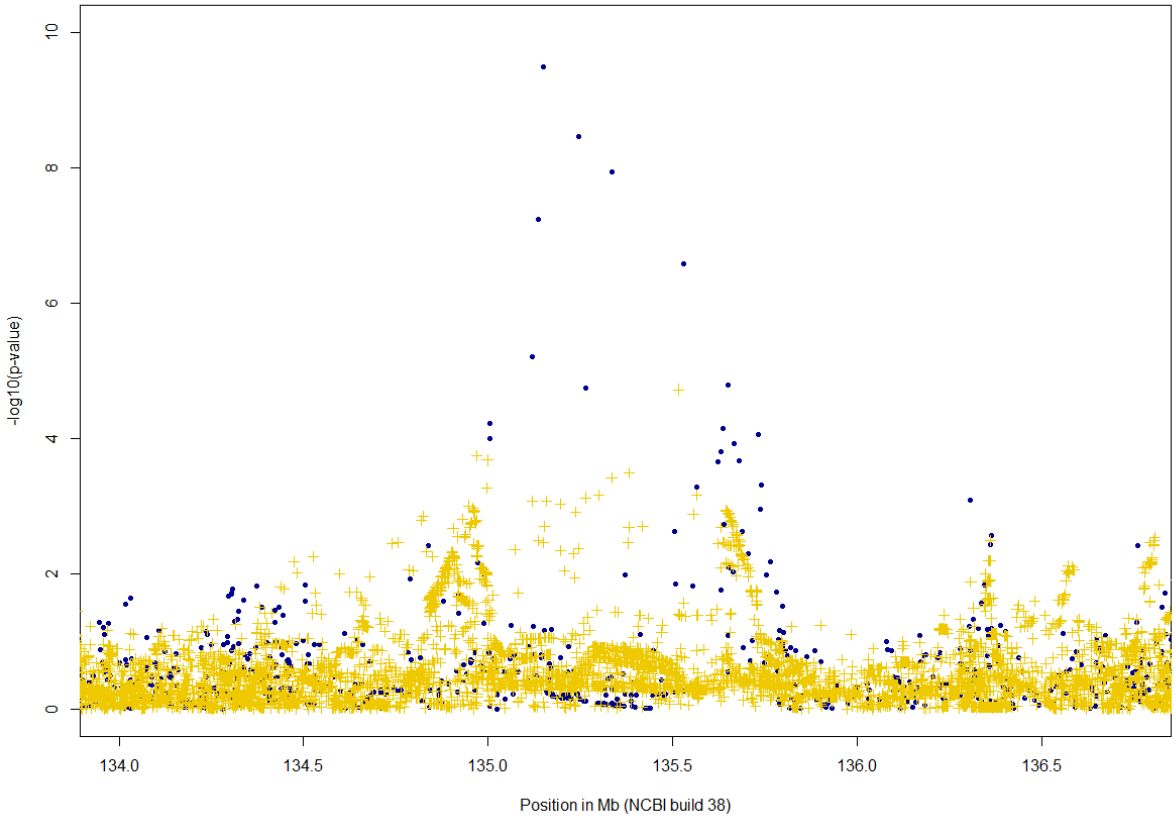

Figure 29. Results from the SHR test within the 2q21.3 region. See Figure legend 10 for details.

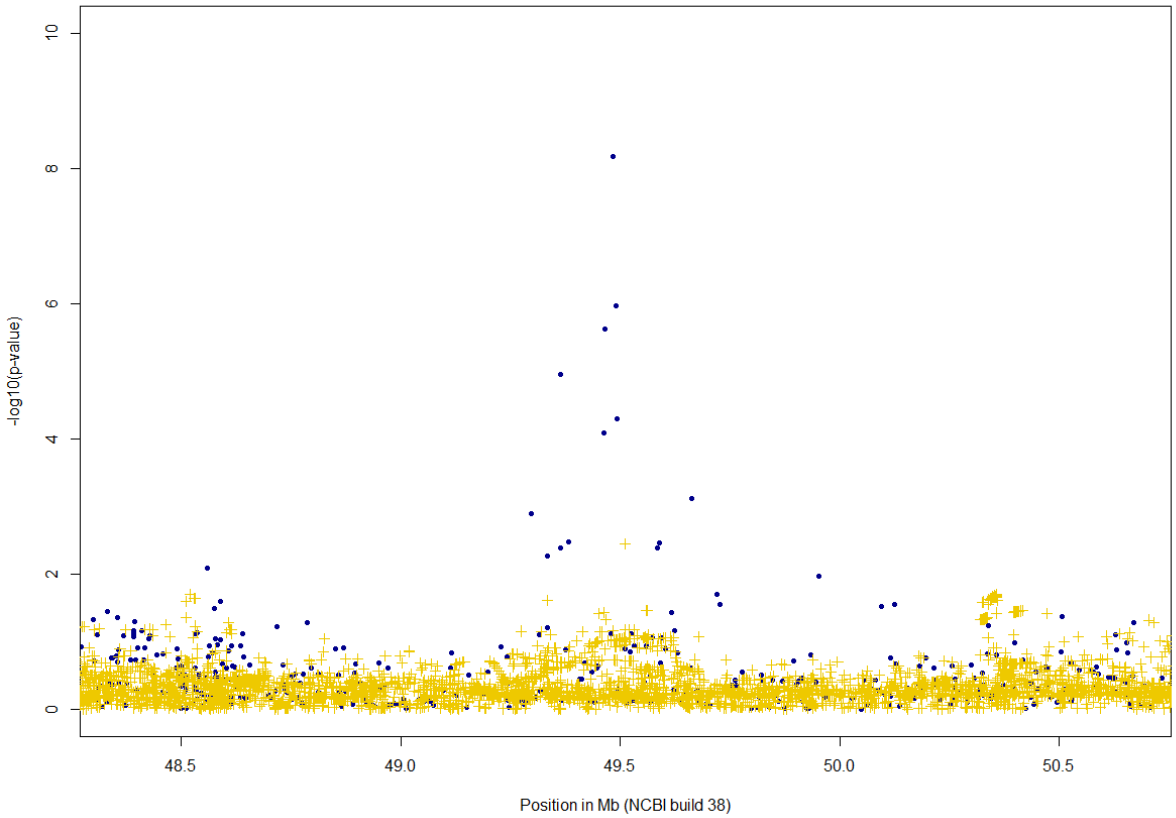

Figure 30. Results from the SHR test within the 1p33 region. See Figure legend 10 for details.

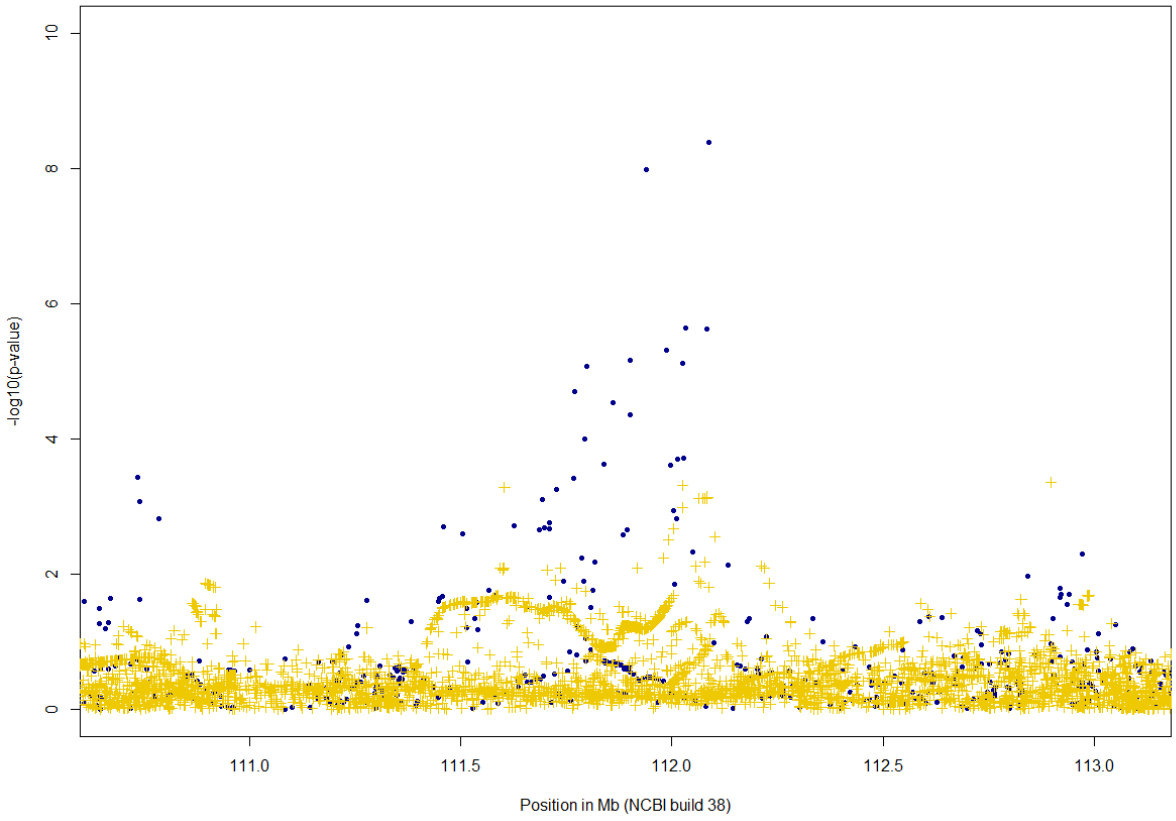

Figure 31. Results from the SHR test within the 12q24.12-q24.13 region. See Figure legend 10 for details.

Our approach to identify regions of recombination suppression revealed 34 candidate regions, of which there are likely some false positives. For example, the little support we see in the WGS data for the two region within the MHC seems a bit suspicious. The centromeric regions are also a bit of a puzzle, as we observe strong significance within them, but a very small δ_{SHR}. It is difficult to say whether we are observing real signals of very small inversions, or if there is some other reason for these signs of heterozygote recombination suppression. However, quite a few of the SHR regions show promise. For example, all regions that contain SNPs with p-values from the SHR test under the Bonferroni significance level, particularly those that harbour more than one, can be considered strong candidates. Other regions of less significance where we see a rise in significance in both datasets, for example at 2p22.3 and 6q24.3, may also harbour real inversions of smaller sizes. This approach has therefore yielded quite a few candidates that deserve further study.

Plots of other regions identified through our approach can be viewed in Figures 32 through 43.

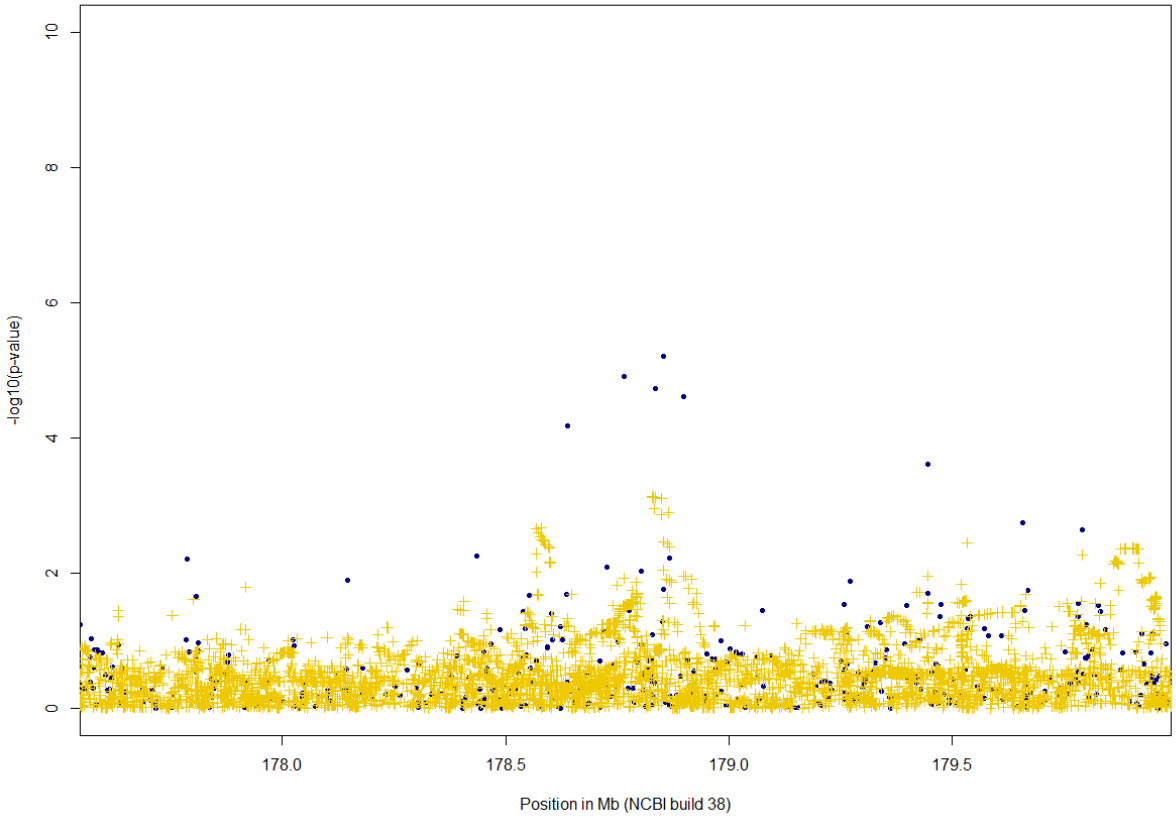

Figure 32. Results from the SHR test within the 3q26.32 region. See Figure legend 10 for details.

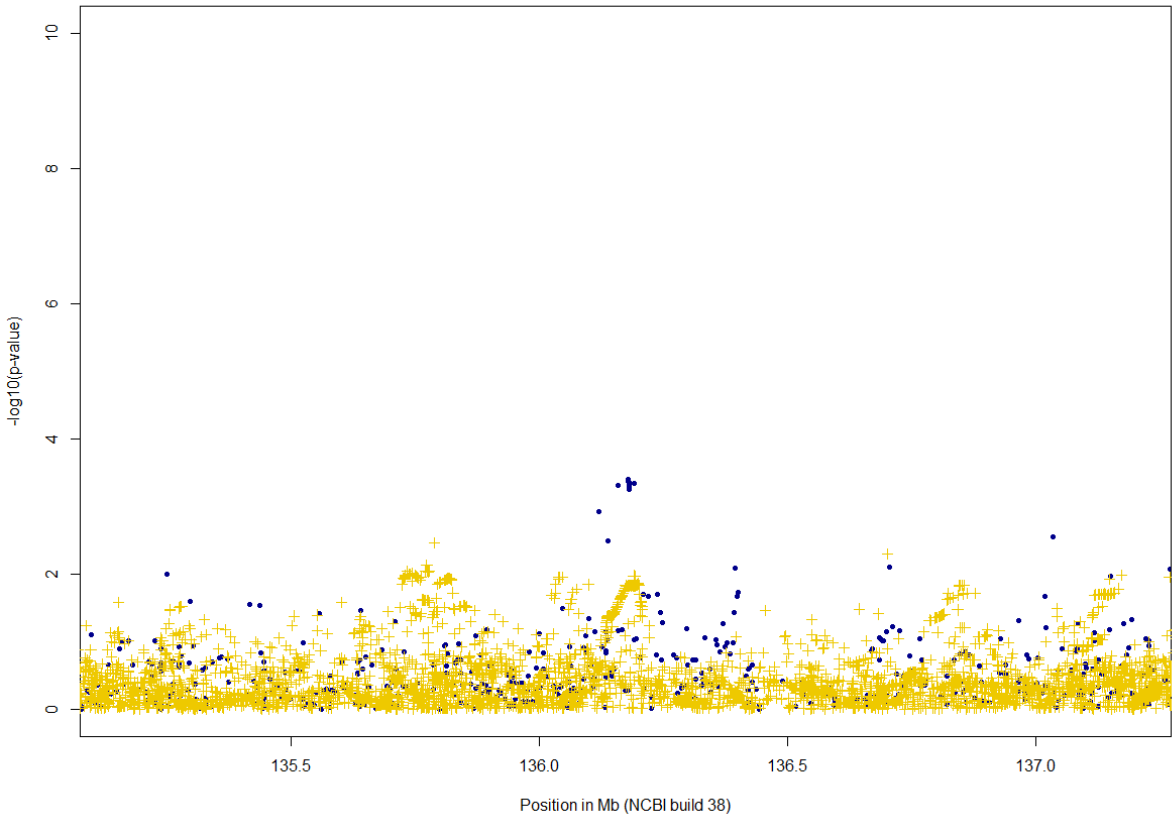

Figure 33. Results from the SHR test within the 5q31.2 region. See Figure legend 10 for details.

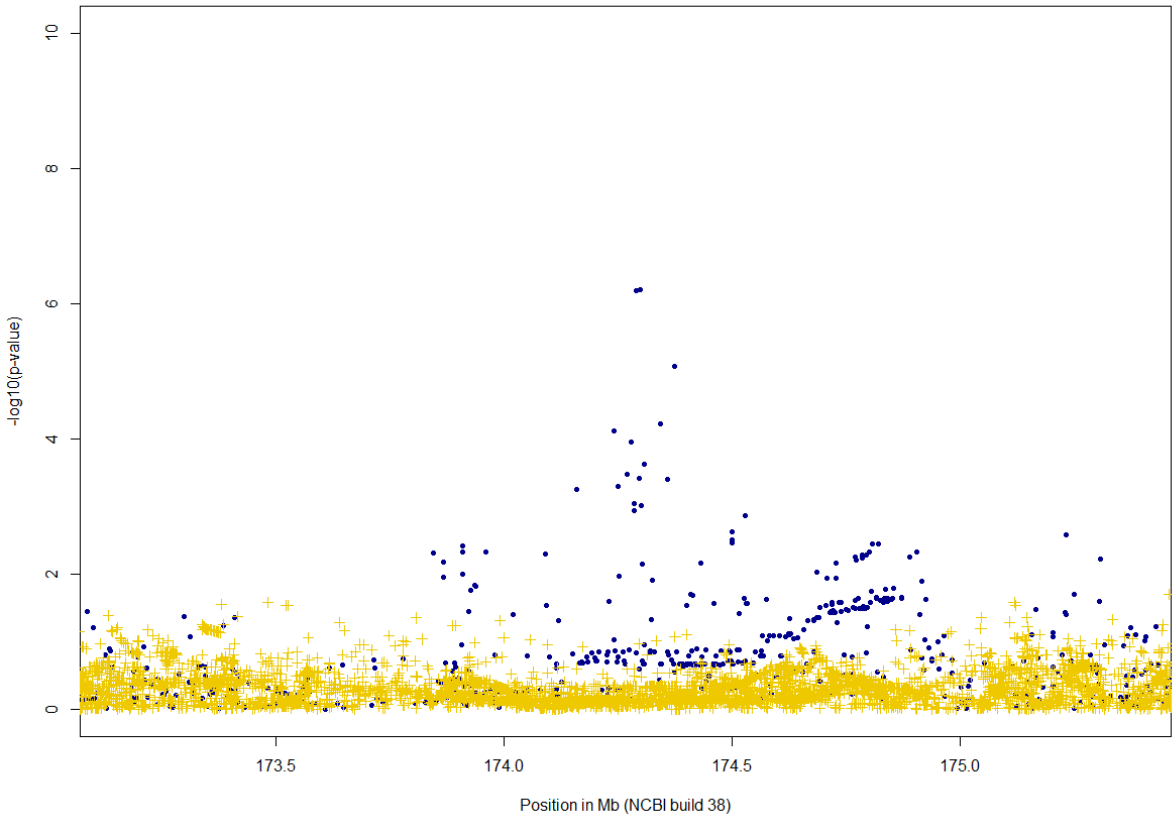

Figure 34. Results from the SHR test within the 1q25.1 region. See Figure legend 10 for details.

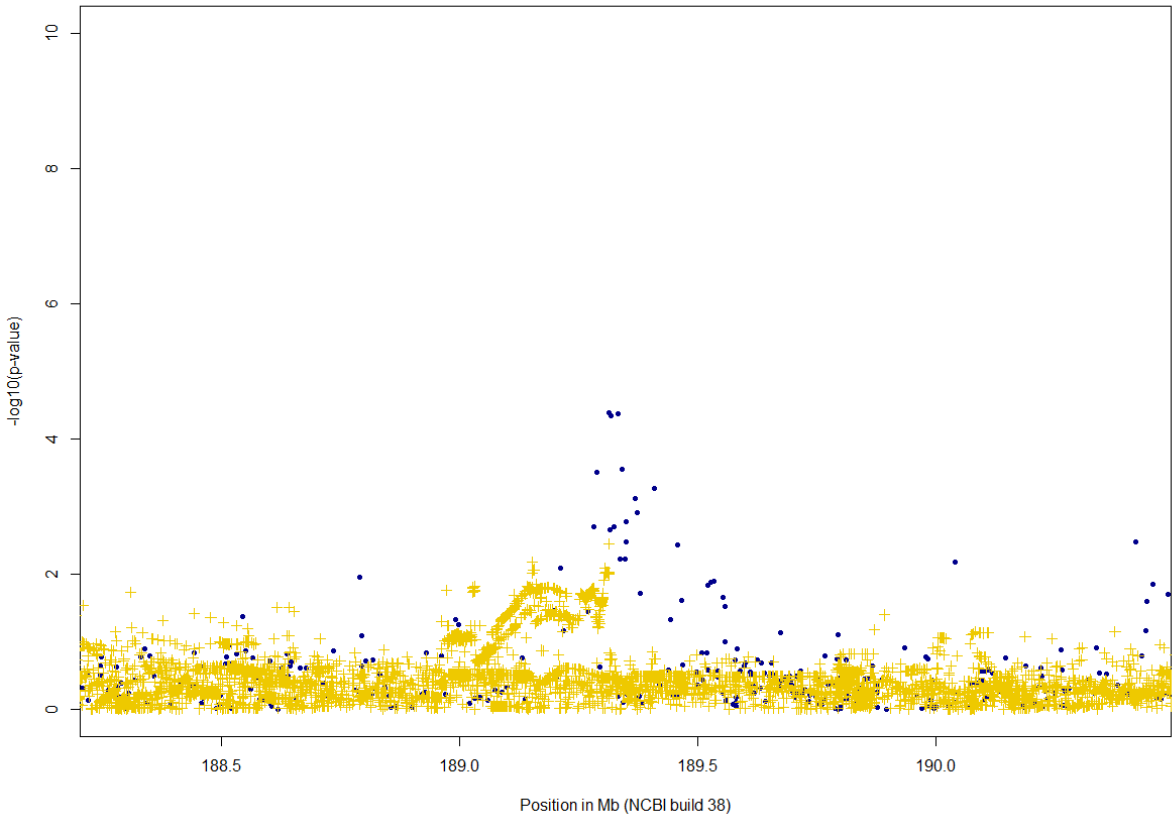

Figure 35. Results from the SHR test within the 1q31.1 region. See Figure legend 10 for details.

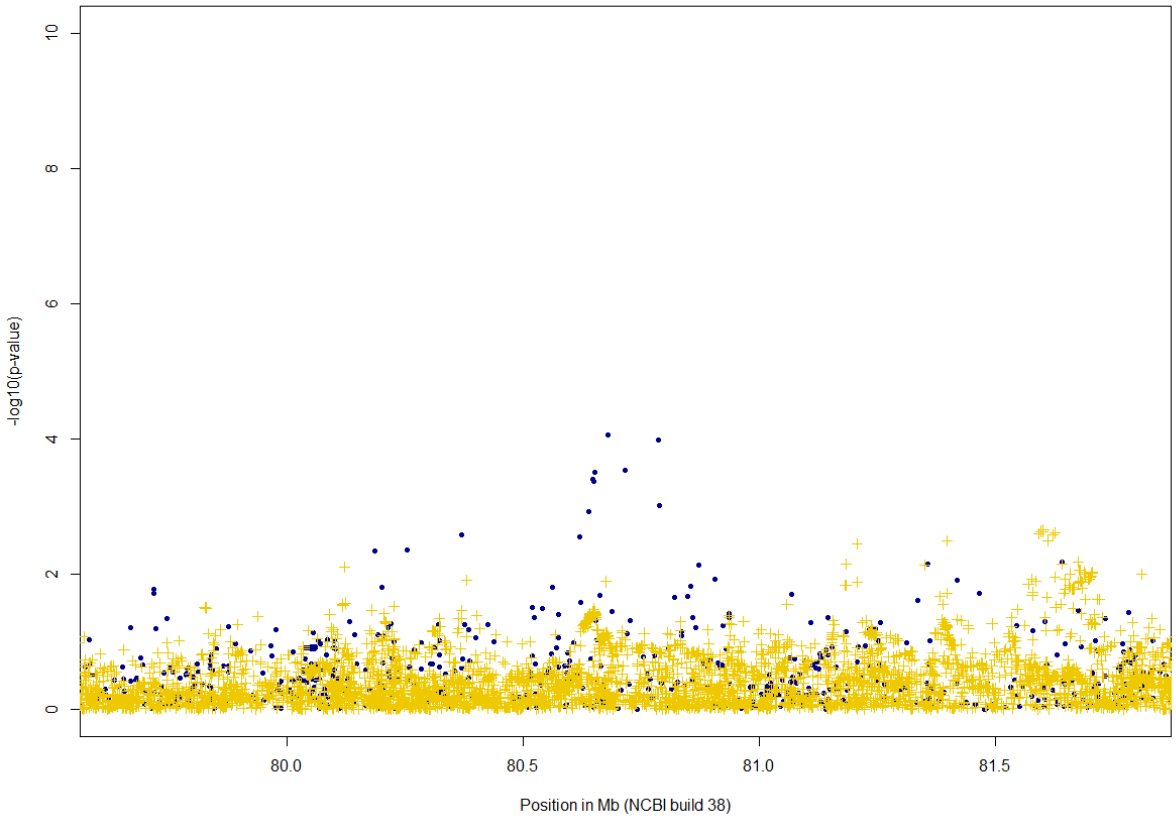

Figure 36. Results from the SHR test within the 5q14.1 region. See Figure legend 10 for details.

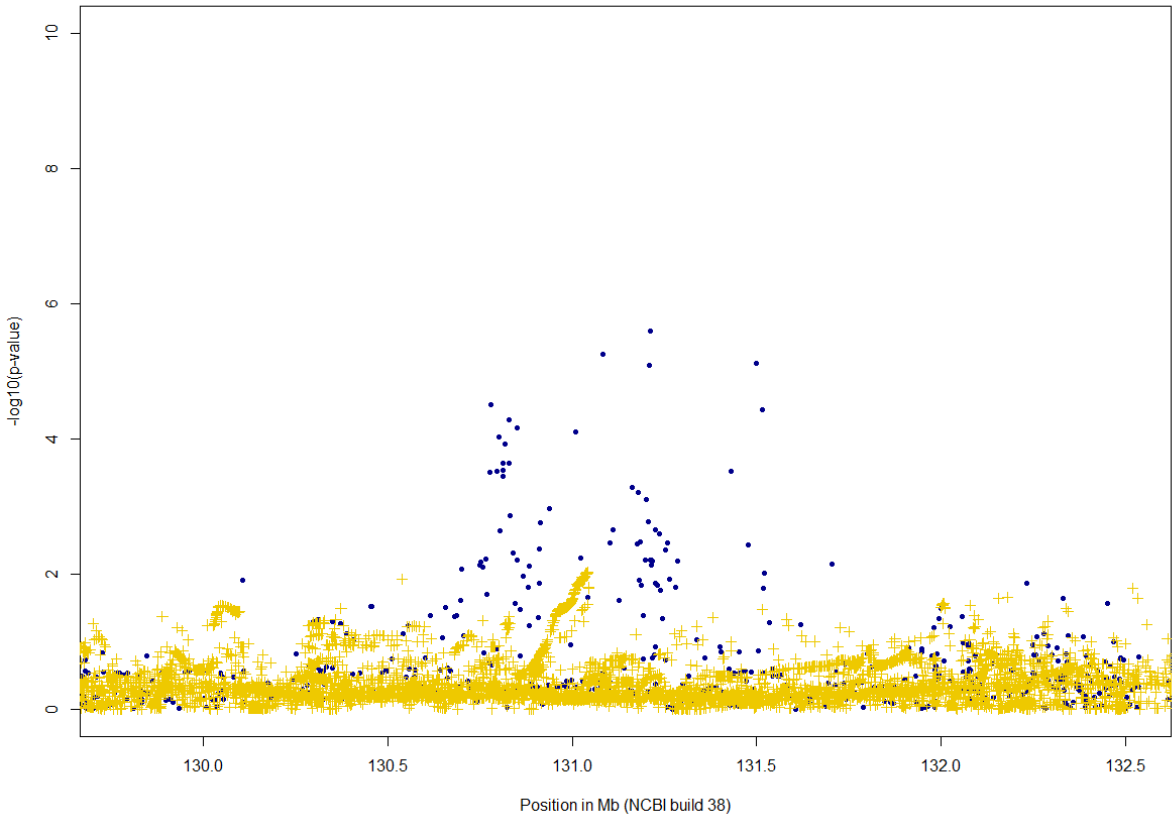

Figure 37. Results from the SHR test within the 5q23.3-q31.1 region. See Figure legend 10 for details.

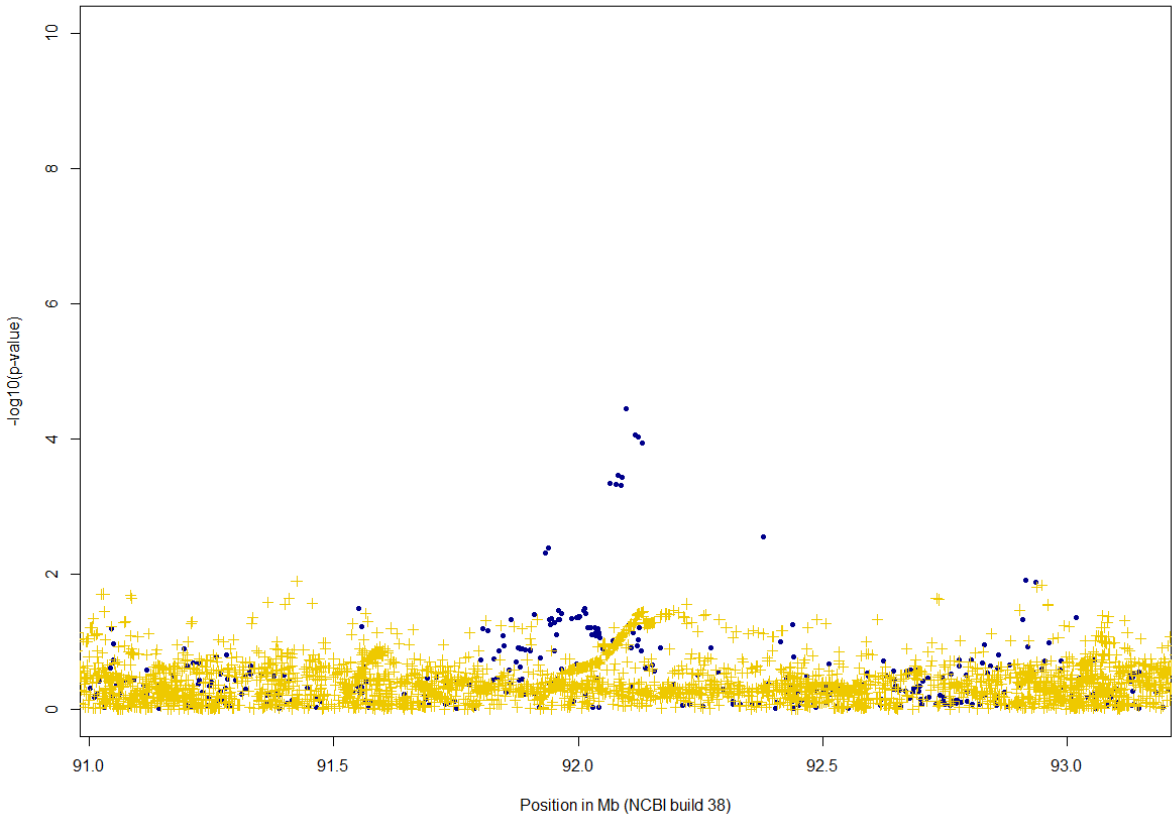

Figure 38. Results from the SHR test within the 7q21.2 region. See Figure legend 10 for details.

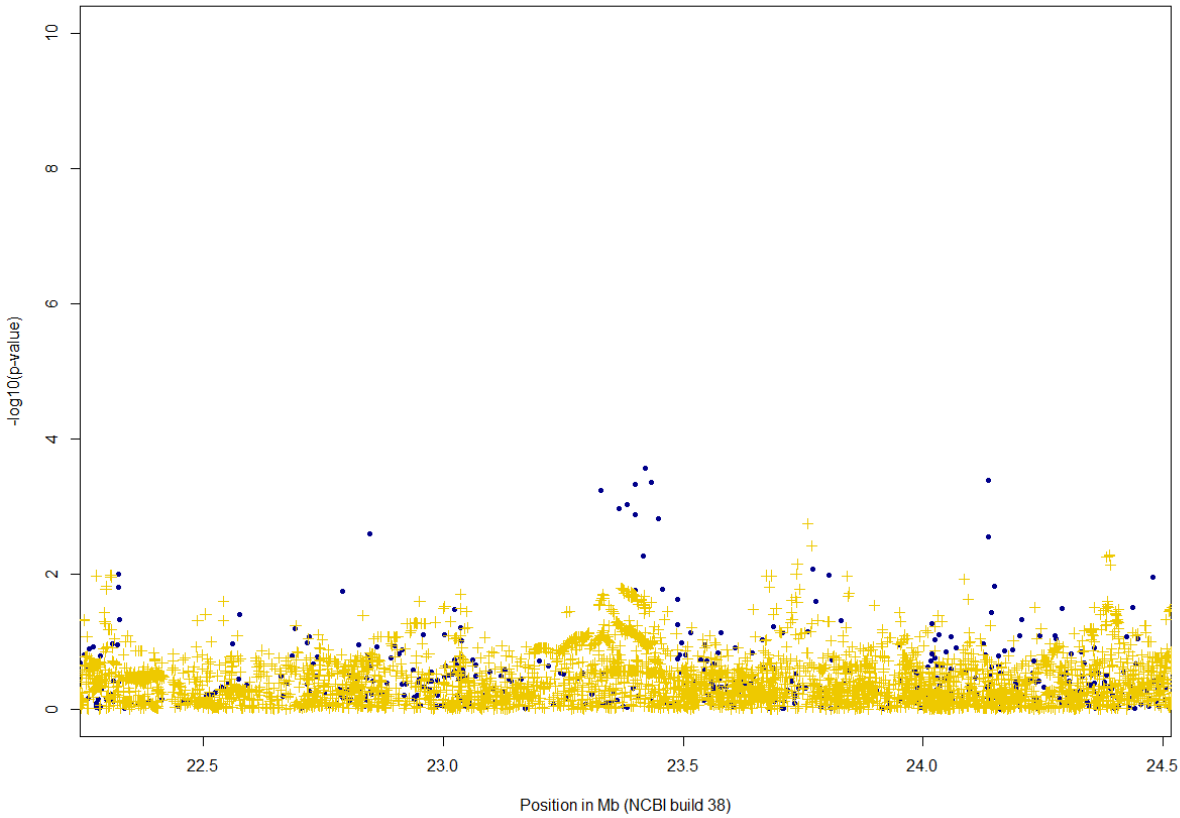

Figure 39. Results from the SHR test within the 12p12.1 region. See Figure legend 10 for details.

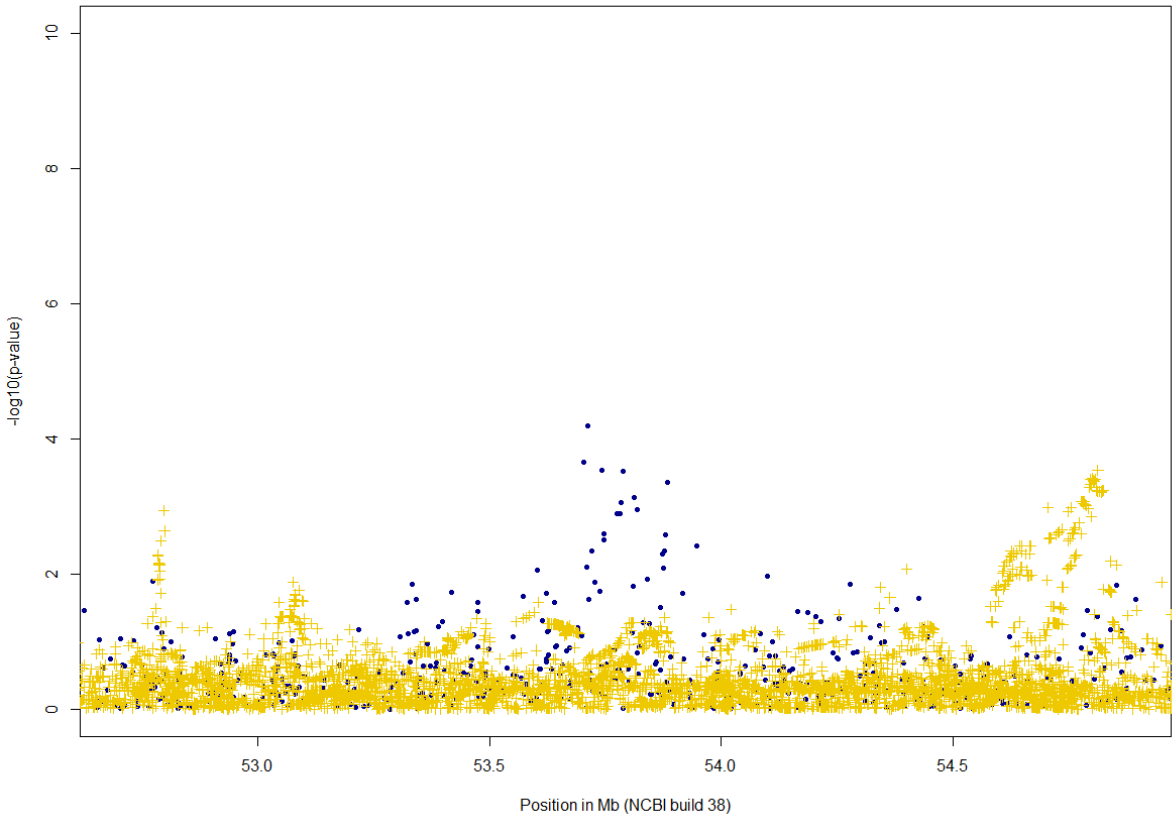

Figure 40. Results from the SHR test within the 13q21.1 region. See Figure legend 10 for details.

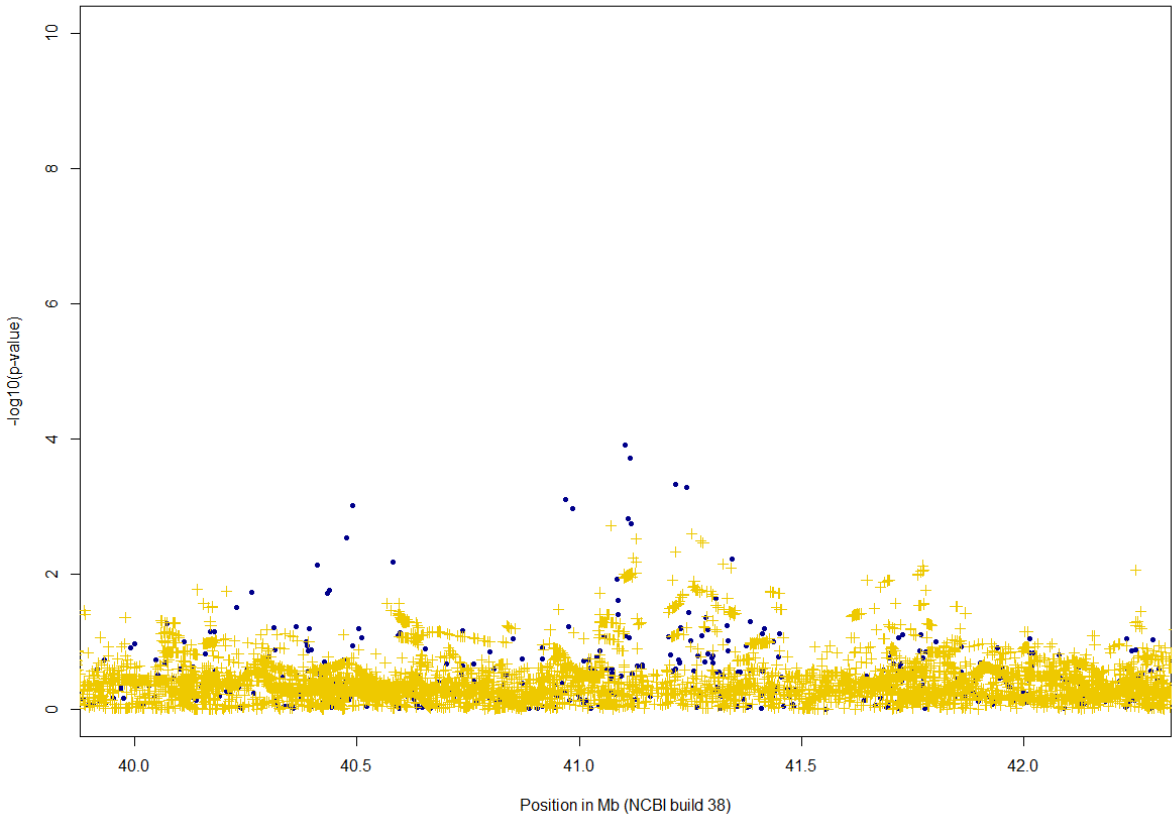

Figure 41. Results from the SHR test within the 14q21.1 region. See Figure legend 10 for details.

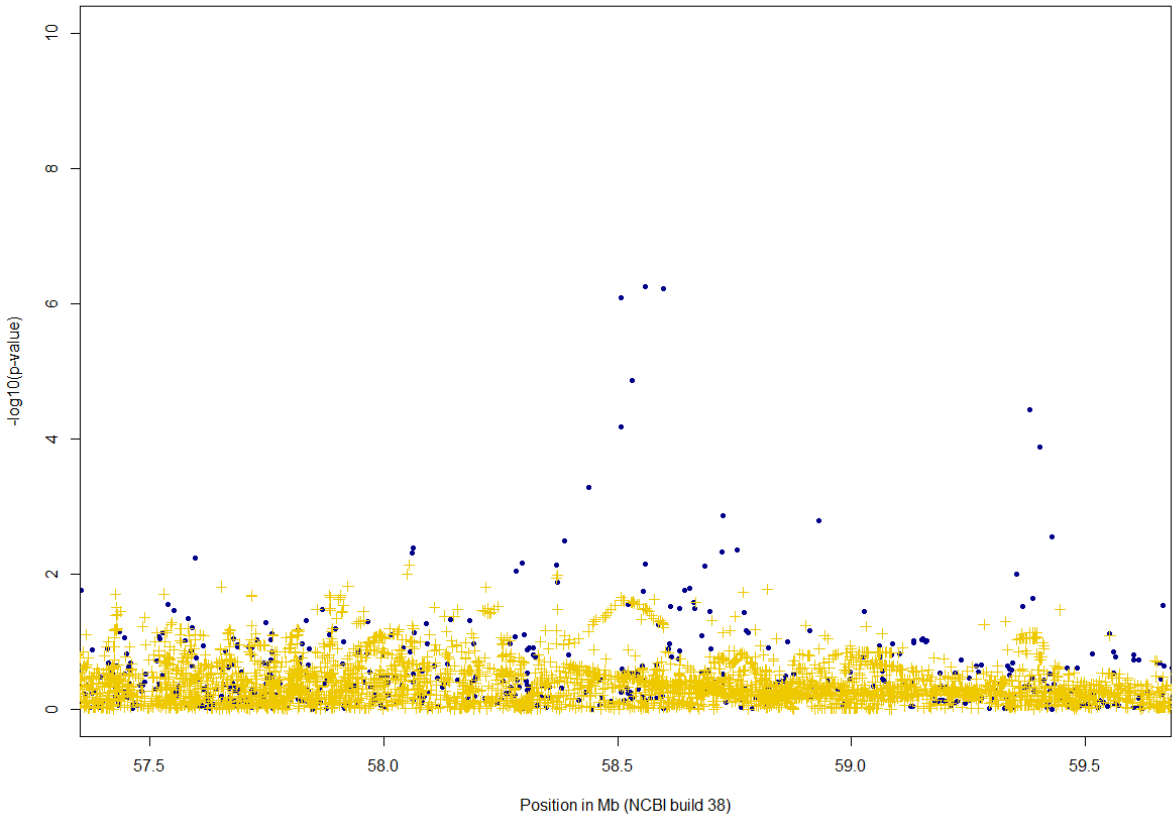

Figure 42. Results from the SHR test within the 17q22 region. See Figure legend 10 for details.

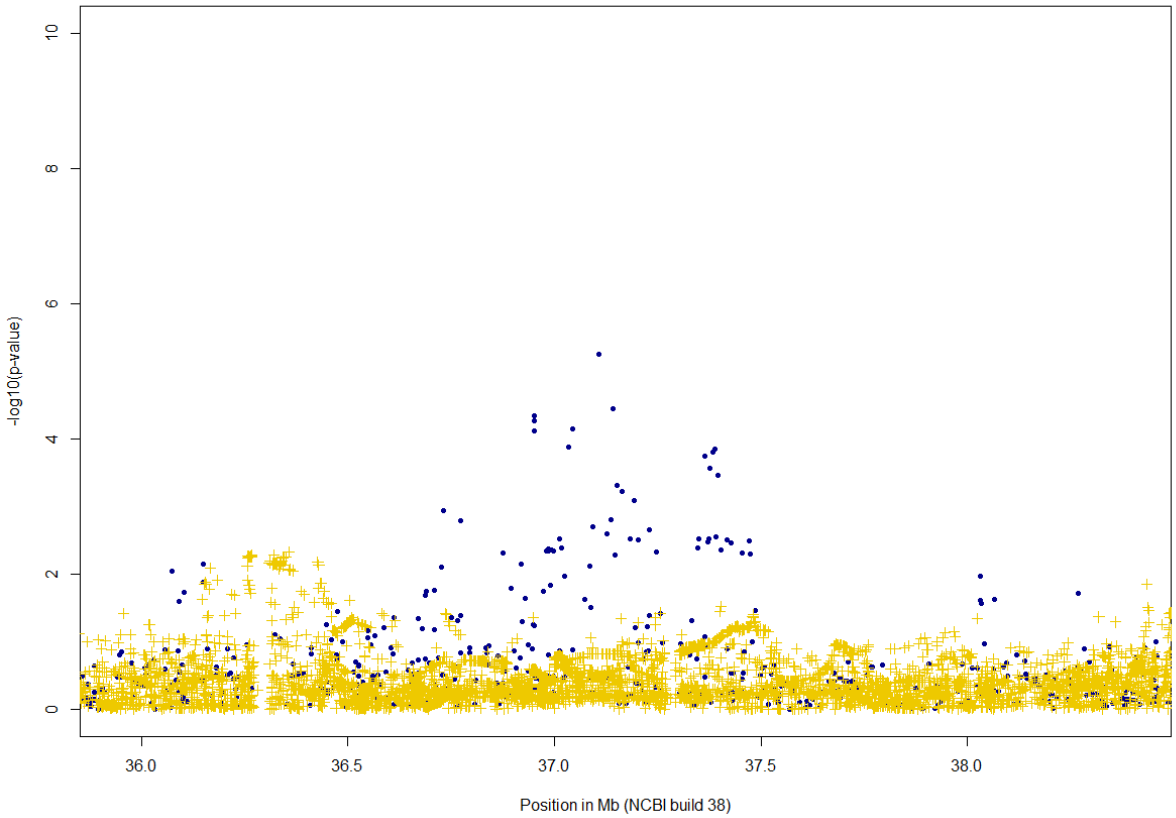

Figure 43. Results from the SHR test within the 19q13.12 region. See Figure legend 10 for details.

5 Discussion

Our aim in this study was to devise and implement a method to detect inversion polymorphisms in the Icelandic gene pool. The SHR test used for this purpose relies on the observation that inversions suppress recombination in heterozygotes for different orientations. The advantage of this approach, when compared to previously published methods (primarily based on detecting regions of strong LD), is it will not yield many false positives due to regions of low recombination rates and selective sweeps. This is because the SHR test compares two groups of individuals, heterozygotes and homozygotes for particular loci, that are expected to be equally affected by such factors.

It is clear, however, that our SHR test will not detect all the inversions present in the Icelandic gene pool. In particular, it is very hard to use this approach to identify inversions that are very recent, small, rare or recurrent. Our method relies on SNPs whose alleles exhibit a strong correlation with inversion orientation. All other things being equal, the closer in time the underlying SNP mutation was to the inversion event itself, the stronger the correlation between the SNPs alleles and the inversion orientations. We postulated that such SNPs would provide the strongest signal of suppressed recombination in heterozygotes. If an inversion is new, there will not be many mutations that are orientation specific. And more importantly, for rare inversion polymorphisms, the number of heterozygotes may be so small, such that there is insufficient statistical power to identify the inversion using the SHR test. In the case of smaller inversions, there will also be relatively few orientation specific mutations, and a smaller number of recombination events than might be needed for sufficient power in the SHR test. The varying density of SNPs, particularly in the microarray data, may also cause smaller inversions within regions of SNP sparsity to go undetected. If an inversion is recurrent, that is, more than one inversion event has occurred in the same region in the population, it will be harder to find SNPs with a strong correlation to one of the three or more orientations and therefore such inversions will be harder to detect using the SHR test.

The results from the SHR test for the regions harbouring the four known polymorphic inversions demonstrate that it has considerable specificity. Thus, the strongest signals and the vast majority of SNPs associated with suppressed recombination are found in these four regions. Using these known inversions as examples, we also see the effect of inversion frequency and size on the results. In particular, we see the signal for the relatively rare but large 1.8 Mb inversion at 15q13.3, and the small but common 16p11.2 inversion. When designing the SHR test we looked to the known inversions at 8p23.1 and 17q21.31 as a reference. Considering their sizes, we concluded that a 500 kb radius would be suitable for comparison of recombination events, estimating that it was small enough to show suppression for inversions of similar sizes, yet wide enough to harbour enough recombination events to yield significant difference between the groups. Our results show that this window size is effective in detecting the four known inversions and possibly others. Moreover, they indicate that the SHR test has considerable sensitivity, i.e. that inversions of similar size and frequency as those at 8p23.1 and 17q21.31 are not to be found in the Icelandic population. Considering the results for the 16p11.2 inversion, however, and taking into account its frequency, we suspected that the 500 kb radius might be sub-optimal for smaller inversions. Thus, the signal from the SHR test for this inversion is relatively weak, particularly in the WGS data. Moreover, a large proportion of significant SNPs in the microarray data are located outside the reported breakpoints. One possibility is that a smaller radius around putative tagging SNPs, for tallying recombination events, could be more effective for detecting smaller inversions. The SHR signal of such inversions could be heavily diluted with larger window sizes. Due to time constraints in the preparation of this thesis, we were not able to run the SHR test with different window sizes for the entire autosomal genome. However, in order to verify our suspicion for the 16p11.2 region, we ran the SHR test using a 250kb (rather than 500kb) radius in the WGS data to see if the SHR signal would become stronger. The results can be viewed in Figure 44.

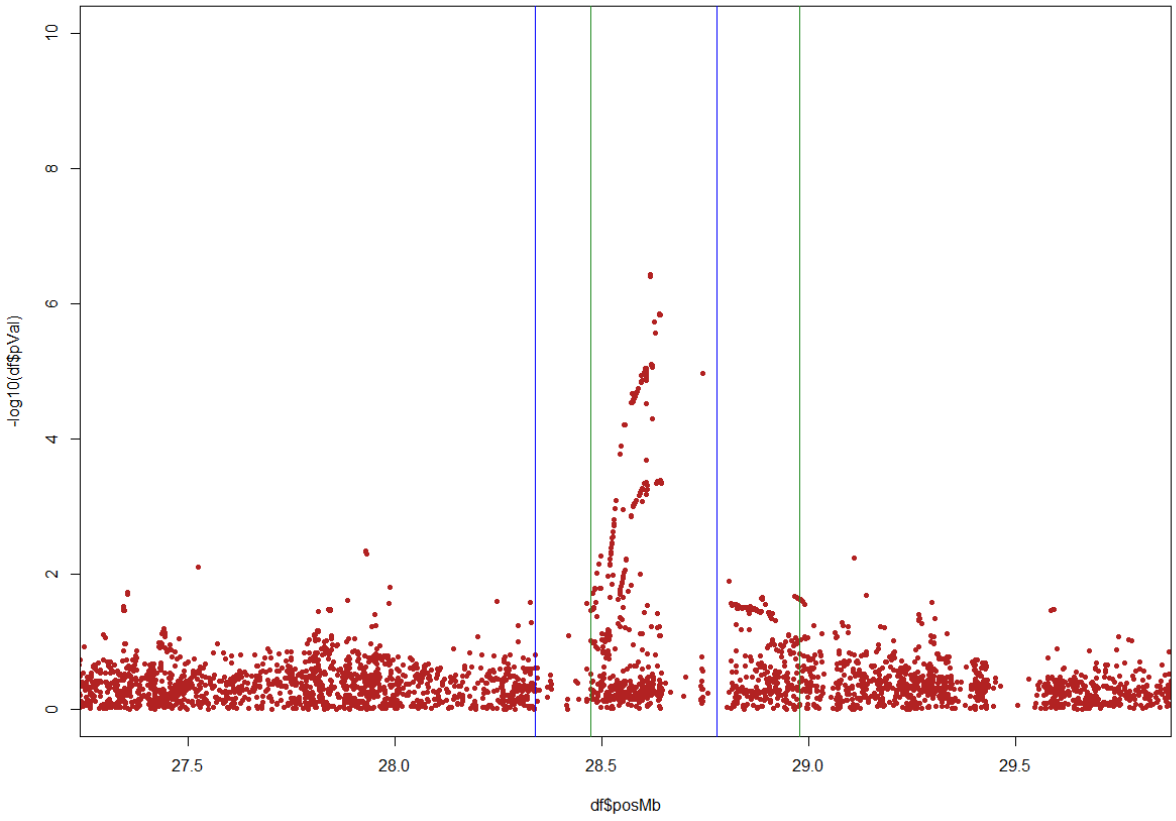

Figure 44. Results from the SHR test within the 16p11.2 region, using 250 kb radius. Red dots represent the p-values of SNPs in the WGS data that yield a lower mean number of recombination events in heterozygotes. Blue vertical lines mark the inversion breakpoints according to the InvFEST database and green lines mark the positions of the first and the last SNP of the SHR region identified through our test.

A comparison of Figures 13 and 44 shows clearly that a smaller radius provides a stronger SHR signal for the rather small 16p11.2 inversion. The lowest p-value within the region based on a radius of 250kb is 3.65×10^{-7} as opposed to 1.02×10^{-3} for the 500kb radius. While we detect greater statistical significance, and a greater relative difference in recombination rates between heterozygotes and homozygotes, we do observe smaller absolute values of δ_{SHR}. The greatest δ_{SHR} observed drops from 0.0192, when using a 500 kb radius, to 0.0109, as the number of recombination events drops. The regions identified through our methods are all smaller than 1 Mb, with the exception of those harboring the 8p23.1 and 15q13.3 inversions, suggesting that the different radius might strengthen the signals within these regions, and perhaps reveal other regions we were not able to detect using the 500 kb radius.

The results from the SHR test within the regions of the four known inversions did not only establish the efficacy of our test, identify the best tagging SNPs for the inversions, and give us information about their frequencies in the population. They also revealed some distinctive features of the 8p23.1 inversion and the 17q21.31 inversion. It has been suggested that there has been some gene flow between the two orientations of the 8p23.1 inversion. The considerable drop in significance around the middle of the inversion, where we might expect it to peak, supports the hypotheses that one or more double recombination events may have occurred between the two orientations in this region – due to its large size.

The unusual pattern of SNPs we observe within the 17q21.31 inversion, where the majority of SNP within the region have the same MAF, emphasizes the extensive divergence of the two orientations, and the homogeneity of the H2 haplotype. This divergence seems unlikely to have come about unless the haplotype was reintroduced to the human lineage as suggested by Stefansson et al. (2005).

5.1 Identifying candidate inversions

After running the SHR test on all SNPs we needed to apply some rules in order to identify candidate regions of inversions. As the multiple tests were not independent, we found that the Bonferroni correction for multiple testing was too strict, although the results from the Bonferroni correction give us a clue about the prevalence of large, common inversions in the population. It revealed the singularity of the 8p23.1 inversion, and according to the results, we should not expect to find more inversions of similar age, size and frequency as the known inversions on chromosomes 8, 15 and 17. There were, however, a few other regions of multiple SNPs with a p-value lower than the Bonferroni correction significance level, and may therefore be considered strong candidates for previously unknown inversion polymorphisms. Of the 34 regions, 10 regions harboured more than one SNP with a p-value lower than the Bonferroni significance level.

It is however difficult to assess the validity of the regions of smaller significance. The experiment on the 16p11.2 has shown that it is well worth a try to run the test again using smaller radiuses. The SHR test has proven to work on inversions of different sizes and different frequencies although time did not allow us to fine-tune the test to examine its effect on the

detection of smaller inversions genomewide. We are confident that running the SHR test on both the microarray and WGS datasets using smaller radiuses will reveal a more detailed picture of other regions.

5.2 Further research

The conditions we set for identifying candidate regions with polymorphic inversions are somewhat arbitrary and it is difficult to assess which thresholds to use. We have already discussed the impact of age, size and frequency on our ability to detect inversions through the SHR test. Other issues that affect our attempt to identify these regions, which effect varies throughout the genome, is the different density of SNPs, the different sample sizes, and the differing recombination rates throughout the genome. The problem of different recombination rates is hard to circumvent, but the WGS data remedies the SNP density problem, and the problem of variable sample sizes.

Where we do find support in the WGS data for recombination suppression, it would be interesting to apply other experimental methods to validate our findings. If we are indeed picking up signals of real inversions through tagging SNPs, then it follows that these SNPs can be used to enrich samples of individuals who are either homozygotes or heterozygotes for the different orientations. It would be interesting to examine these regions further using other methods, for example to perform FISH on chromosomes of the different genotypes to see if it reveals tangible evidence for the presence of different orientations. As explained earlier, we expect the significance of the SHR test to fade out within the 500 kb radius around the breakpoints of an inversion. The observation that inversion breakpoints are frequently positioned within areas of inverted repeats, and the variable density of SNPs throughout the genome means that the assessment of breakpoint positions based on our results in the microarray data cannot be fully accurate. Despite the weaker significance in the WGS data, it can give a more detailed picture of regions found in the microarray data. It is therefore possible to make a more precise estimation of breakpoints with the help of the WGS data. A smaller radius will also provide better resolution for predicting breakpoints. It would therefore be interesting to run the test on both datasets with a 250kb radius, apply the same conditions as before and see the affect these changes have on our list of regions. The results from the

WGS data can then be used for a more precise estimation of breakpoints and tagging SNPs when designing tests to verify the findings with other methods, and may also reveal smaller inversions that are not detectable using a 500 kb radius.

References

The 1000 Genomes Project Consortium. (2010). A map of human genome variation from population scale sequencing. *Nature, 467*(7319), 1061–1073.

Ahn, S. M., Kim, T. H., Lee, S., Kim, D., Ghang, H., Kim, D. S. … Kim, S. J. (2009). The first Korean genome sequence and analysis: full genome sequencing for a socio-ethnic group. *Genome Research, 19*(9), 1622-1629.

Altshuler, D., Durbin, R. M. Abercasis, G. R., Bentley, D. R., Chakravarti, A., Clark, A. G., … Wilson, R. K. (2010). A map of human genome variation from population-scale sequencing. *Nature, 467*(7319), 1061–1073.

Alves, J. M., Lopes, A. M., Chikhi, L., & Amorim, A. (2012). On the structural plasticity of the human genome: chromosomal inversions revisited. *Current Genomics, 13*(8), 623-632.

Andolfatto, P., Depaulis, F., & Navarro, A. (2001). Inversion polymorphisms and nucleotide variability in Drosophila. *Genetical Research, 77*(1), 1-8.

Antonacci, F., Dennis, M. Y., Huddleston, J., Sudmant, P. H., Steinberg, K. M., Rosenfeld, J. A., & Eichler, E. E. (2014). Palindromic GOLGA8 core duplicons promote chromosome 15q13.3 microdeletion and evolutionary instability. *Nature Genetics, 46*(12), 1293-1302.

Antonacci, F., Kidd, J. M., Marques-Bonet, T., Ventura, M., Siswara, P., Jiang, Z., & Eichler, E. (2009). *Human Molecular Genetics, 18*(14), 2555-2566.

Arlt, M. F., Ozdemir, A. C., Birkeland, S. R., Lyons Jr., R. H., Glover, T. W., & Wilson, T. E. (2011). Comparison of constitutional and replication stress-induced genome structural variation by SNP array and mate-pair sequencing. *Genetics, 187*(3), 675-683.

Bansal, V., Bashir, A., & Bafna, V. (2007). Evidence for large inversion polymorphisms in the human genome from HapMap data. *Genome Research, 17*(2), 219-230.

Barzel, A., & Kupiec, M. (2008). Finding a match: how do homologous sequences get together for recombination? *Nature Reviews Genetics, 9*(1), 27-37.

Chaisson, M. J. P., Huddleston, J., Dennis, M. Y., Sudmant, P. H., Malig, M., Hormozdiari, F., … Eichler, E. E. (2014). Resolving the complexity of the human genome using single-molecule sequencing. *Nature, 517*(7536), 608-611.

Escaramís, G., Docampo, E., & Rabionet, R. (2015). A decade of structural variants: description, history and methods to detect structural variation. *Briefings in Functional Genomics, 14*(5), 305-314.

Feuk, L. (2010). Inversion variants in the human genome: role in disease and genome architecture. Genome Medicine 2(11), 1-8.

Feuk, L., Carson, A. R., & Scherer, S. W. (2006). Structural variation in the human genome. *Nature Reviews Genetics, 7*(2), 85-97.

Feuk, L., MacDonald, J. R., Tang, T., Carson, A. R., Li, M., Rao, G., … Scherer, S. W. (2005). Discovery of human inversion polymorphisms by comparative analysis of human and chimpanzee DNA sequence assemblies. *PLOS Genetics, 1*(4), e56.

Giglio, S., Broman, K. W., Matsumoto, N., Calvari, V., Gimelli, G., Neumann, T., … Zuffardi, O. (2001). Olfactory receptor–gene clusters, genomic-inversion polymorphisms, and common chromosome rearrangements. *American Journal of Human Genetics, 68*(4), 874-883.

Gimelli, G., Pujana, M. A., Patricelli, M. G., Russo, S., Giardino, D., Larizza, L., … Zuffardi, O. (2003). Genomic inversions of human chromosome 15q11–q13 in mothers of Angelman syndrome patients with class II (BP2/3) deletions. *Human Molecular Genetics, 12*(8), 849-858.

Gonzáles, J. R., Cáceres, A., Esko, T., Cuscó, I., Puig, M., Esnaola, M., … Pérez-Jurado, L. A. (2014). A common 16p11.2 inversion underlies the joint susceptibility to asthma and obesity. *American Journal of Human Genetics, 94*(3), 361-372.

Griffiths, A. J. F., Wessler, S. R., Carroll, S. B., Doebley, J. (2014). *Introduction to genetic analysis* (10th ed.). New York, NY: W. H. Freeman and Company.

Jobling, M., Hollox, E., Hurles, M., Kivisild, T., and Tyler-Smith, C. (2014). *Human evolutionary genetics* (4th ed.). New York, NY: Garland Science.

Jones, M. L., Murden, S. L., Brooks, C., Maloney, V., Manning, R. A., Gilmour, K. C., … Mumford, A. D. (2013). Disruption of AP3B1 by a chromosome 5 inversion: a new disease mechanism in Hermansky-Pudlak syndrome type 2. *BMC Medical Genetics, 14*(42).

Kent, W. J., Sugnet, C. W., Furey, T. S., Roskin, K. M., Pringle, T. H., Zahler, A. M., Haussler, D. (2002). The human genome browser at UCSC. *Genome Research, 12*(6), 996-1006.

Kidd, J. M., Cooper, G. M., Donahue, W. F., Hayden, H. S., Sampas, N., Graves, T., … Eichler, E. E. (2008) Mapping and sequencing of structural variation from eight human genomes. *Nature, 453*(7191), 56-64.

Kirkpatrick, M. (2010). How and why chromosome inversions evolve. *PLOS Biology 8*(9), e1000501.

Kirkpatrick, M., Barton, N. (2006). Chromosome inversions, local adaptation and speciation. *Genetics, 173*(1), 419-434.

Kong, A., Masson, G., Frigge, M. L., Gylfason, A., Zusmanovich, P., Thorleifsson, G., … Stefansson, K. (2008). Detection of sharing by descent, long-range phasing and haplotype imputation. *Nature Genetics, 40*(9), 1068-1075.

Kong, A., Steinthorsdottir, V., Masson, G., Thorleifsson, G., Sulem, P., Besenbacher, S., … Stefansson, K. (2009). Parental origin of sequence variants associated with complex diseases. *Nature, 462*(7275), 868-874.

Kong, A., Thorleifsson, G., Frigge, M. L., Masson, G., Gudbjartsson, D. F., Villemoes, R., … Stefansson, K. (2014). Common and low-frequency variants associated with genome-wide recombination rate. *Nature Genetics, 46*(1), 11-16.

Kong, A., Thorleifsson, G., Gudbjartsson, D. F., Masson, G., Sigurdsson, A., Jonasdottir, A., … Stefansson, K. (2010). Fine-scale recombination rate differences between sexes, populations and individuals. *Nature, 467*(7319), 1099-1103.

Korbel, J. O., Urban, A. E., Affourtit, J. P., Godwin, B., Grubert, F., Simons, J. F., … Snyder, M. (2007). Paired-End Mapping Reveals Extensive Structural Variation in the Human Genome. *Science, 318*(5849), 420-426.

Lagerstedt, K., Karsten, S. L., Carlberg, B. M. Kleijer, W. J., Tönnesen, T., Pettersson, U., & Bondeson, M. L. (1997). Double-strand breaks may initiate the inversion mutation causing the Hunter syndrome. *Human Molecular Genetics, 6*(4), 627-633.

Lieber, M. R., Ma, Y., Pannicke, U., & Schwarz, K. (2003). Mechanism and regulation of human non-homologous DNA end-joining. *Nature Reviews Molecular Cell Biology, 4*(9), 712-720.

MacDonald, J. R., Ziman, R., Yuen, R. K., Feuk, L., & Scherer, S. W. (2013). The database of genomic variants: a curated collection of structural variation in the human genome. *Nucleic Acids Research, 42*(D1), D986-D992.

Martin, J., Han, C., Gordon, L. A., Terry, A., Prabhakar, S., She, X., … Pennacchio, L. A. (2004). The sequence and analysis of duplication-rich human chromosome 16. *Nature, 432*(7020), 988-994.

Martínez-Fundichely, A., Casillas, S., Egea, R., Ràmia, M., Barbadilla, A., Pantano, L., … Cáceres, M. (2014). InvFEST, a database integrating information of polymorphic inversions in the human genome. *Nucleic Acids Research, 42*(D1), D1027-D1032.

McKernan, K. J., Peckham, H. E., Costa, G. L., McLaughlin, S. F., Fu, Y., Tsung, E. F. … Blanchard, A. P. (2009). Sequence and structural variation in a human genome uncovered by short-read, massively parallel ligation sequencing using two-base encoding. *Genome Reserarch, 19*(9), 1527-1541.

Morozova, O., & Marra, M. A. (2008). Applications of next-generation sequencing technologies in functional genomics. *Genomics, 92*(5), 255-264.

Namjou, B., Ni, Y., Harley, I. T. W., Chepelev, I., Cobb, B., Kottyan, L. C., … Harley, J. B. (2014). The effect of inversion at 8p23 on BLK association with lupus in caucasian population. *PLOS ONE, 9*(12), e115614.

Navarro, A., & Barton, N. H. (2003). Chromosomal Speciation and Molecular Divergence—Accelerated Evolution in Rearranged Chromosomes. *Science, 300*(5617), 321-324.

Oliveira, S. A., Scott, W. K., Zhang, F., Stajich, J. M., Fujiwara, K., Hauser, M., … Martin, E. R. (2004). Linkage disequilibrium and haplotype tagging polymorphisms in the Tau H1 haplotype. *Neurogenetics, 5*(3), 147-155.

Osborne, L. R., Li, M., Pober, B., Chitayat, D., Bodurtha, J., Mandel, A., … Scherer, S. W. (2001). A 1.5 million-base pair inversion polymorphism in families with Williams-Beuren syndrome. *Nature Genetics, 29*(3), 321-325.

Pang, A. W., MacDonald, J. R., Pinto, D., Wei, J., Rafiq, M. A., Conrad, D. F., … Scherer, S. W. (2010). Towards a comprehensive structural variation map of an individual human genome. *Genome Biology, 11*(5), R52.

Parks, M. M., Lawrence, C. E., & Raphael, B. J. (2015). Detecting non-allelic homologous recombination from high-throughput sequencing data. Genome Biology, 16, 72.

Peyvandi, F., Garagiola, I., & Young, G. (2016). The past and future of haemophilia: diagnosis, treatments, and its complications. *Lancet, 388*(10040), 187-197.

Puig, M., Casillas, S., Villatoro, S., & Cáceres, M. (2015). Human inversions and their functional consequences. *Briefings in Functional Genomics, 14*(5), 369-379.

Rhoads, A., & Au, K. F. (2015). PacBio sequencing and its applications. *Genomics Proteomics Bioinformatics, 13*(5), 278-289.

Rieseberg, L. H. (2001). Chromosomal rearrangements and speciation. Trends *in Ecology & Evolution, 16*(7), 351-358.

Salm, M. P. A., Horswell, S. D., Hutchison, C. E., Speedy, H. E., Yang, X., Liang, L., … Shoulders, C. C. (2012). The origin, global distribution, and functional impact of the human 8p23 inversion polymorphism. *Genome Research, 22*(6), 1144-1153.

Sharp, A. J. (2008). Emerging themes and new challenges in defining the role of structural variation in human disease. *Human Mutation, 30*(2), 135-144.

Sharp, A. J., Cheng, Z., & Eichler, E. E. (2006). Structural variation of the human genome. *Annual Review of Genomics and Human Genetics, 7*, 407-442.

Sharp, A. J., Mefford, H. C., Li, K., Baker, C., Skinner, C., Stevenson, R. E., … Eichler, E. E. (2008). A recurrent 15q13.3 microdeletion syndrome associated with mental retardation and seizures. *Nature Genetics, 40*(3), 322-328.

Skipper, L., Wilkes, K., Toft, M., Baker, M., Lincoln, S., Hulihan, M., … Farrer, M. (2004). Linkage disequilibrium and association of MAPT H1 in Parkinson disease. *American Journal of Human Genetics, 75*(4), 669-677.

Stefansson, H., Helgason, A., Thorleifsson, G., Steinthorsdottir, V., Masson, G., Barnard, J., … Stefansson, K. (2005). A common inversion under selection in Europeans. *Nature Genetics, 37*(2), 129-137.

Steinberg, K. M., Antonacci, F., Sudmant, P. H., Kidd, J. M., Campbell, C. D., Vives, L., … Eichler, E. E. (2012). Structural diversity and African origin of the 17q21.31 inversion polymorphism. *Nature Genetics, 44*(8), 872-880.

Sudmant, P. H., Rausch, T., Gardner, E. J., Handsaker, R. E., Abyzov, A., Huddleston, J., … Korbel, J. O. (2015). An integrated map of structural variation in 2,504 human genomes. *Nature, 526*(7571), 75-81.

Sugawara, H., Harada, N., Ida, T., Ishida, T., Ledbetter, D. H., Yoshiura, K., … Matsumoto, N. (2003). Complex low-copy repeats associated with a common polymorphic inversion at human chromosome 8p23. *Genomics, 82*(1), 238-244.

Tuzun, E., Sharp, A. J., Bailey, J. A., Kaul, R., Morrison, V. A., Pertz, L. M., … Eichler, E. E. (2005) Fine-scale structural variation of the human genome. *Nature Genetics, 37*(7), 727-732.

Wang, J., Wang, W., Li, R., Li, Y., Tian, G., Goodman, L, … Wang, J. (2008). The diploid genome sequence of an Asian individual. *Nature, 456*(7218), 60-65.

Zhang, J., Wang, X., & Podlaha, O. (2004). Testing the chromosomal speciation hypothesis for humans and chimpanzees. *Genome Research, 14*(5), 845-851.

Zody, M. C., Jiang, Z., Fung, H. C., Antonacci, F., Hillier, L. W., Cardone, M. F., … Eichler, E. E. (2008). Evolutionary toggling of the MAPT 17q21 inversion region. *Nature Genetics, 40*(9), 1076-1083.