Molecular Mechanisms of Recombination Hotspots in Humans

Nudrat Noor

Wellcome Trust Centre for Human Genetics

University of Oxford

A thesis submitted for the degree of

Doctor of Philosophy

Hilary 2013

Acknowledgements

I wish to express my deepest gratitude to my supervisor Dr Simon Myers. His constant guidance, encouragement to explore on my own while providing unfailing support when experiments failed, his constructive scientific discussions that ingrained the importance of critical thinking have been invaluable. Simon's passion for science will always be a source of inspiration for me; I am extremely fortunate to have had him as a mentor.

I would like to thank Dr Julian Knight and Professor Gil McVean for co-supervising my research work. Their supervision, constructive discussions and feedback have been extremely valuable during the period of my DPhil. I am grateful to Dr. Radu Aricescu at the Structural Biology Unit, for his collaboration and guidance in synthesising PRDM9. I would also like to express my gratitude to Professor Peter Donnelly for his support and for funding part of my research work.

I am grateful to all my colleagues at the Wellcome Trust Centre for Human Genetics for their input and advice at various stages of my DPhil. I am specially thankful to Anjali Hinch, Afidalina Tumian and Nick Altemose for the very many helpful discussions during these years. I would also like to thank Kate Wicks and Seiko Makino who helped me learn experimental techniques like EMSA and cloning; Benjamin Bishop, Amber Clayton, Weixian Lu and Yuguang Zhao for the support required for PRDM9 synthesis.

This work would not have been possible without the love and support of my family. I am extremely grateful to my parents, Mehtab

Ahmed Qureshi and Naushaba Mehtab; I am specially thankful my father who has been my inspiration for higher education - it is to him that I dedicate this thesis. To my brother and sister-in-law, Naveed and Habiba, thank you for your care and encouragement; I am so fortunate to have you as family. I am grateful to my husband, Ahmed Rattani, for his patience and unfaltering support. He has been my closest friend and confidant; I cannot thank him enough for always being there.

Finally, I am thankful to the Wellcome Trust and Clarendon Fund for funding my DPhil studies and the Genomic Medicine and Statistics program at the Wellcome Trust Centre for its structured training.

Abstract

Meiotic recombination involves the exchange of DNA between two homologous chromosomes, forming cross-overs and gene conversion events. The cross-over process is important for the proper segregation of chromosomes during meiosis, and drives genetic diversity. Human hotspots are enriched for a 13-bp motif, CCNCCNTNNCCNC; a close match to this motif occurs in about 40% of our cross-over hotspots. A DNA binding protein called PRDM9, having histone trimethyltransferase (H3K4me3) activity, binds the motif and is becoming established as a major determinant of recombination hotspots (narrow regions with high cross-over activity). This research aimed to understand the mechanisms involved in promoting PRDM9 binding to its target sites, and subsequently, initiating cross-over hotspot activity.

We first explored the relationship between PRDM9 binding and DNA sequence, to directly confirm whether PRDM9 binds to the 13-bp hotspot motif using in-vitro gel-shift assays, and found that it does bind sequence specifically to the canonical 13-mer motif. PRDM9 is able to bind the motif in a highly selective manner, with certain single base pair changes abolishing binding. However, we observe that it is also able to tolerate degeneracy in its binding sites, as demonstrated by strong in-vitro binding to degenerate versions of the 13-bp motif. Hence, these results confirmed that PRDM9 is able to directly bind to the 13-bp hotspot motifs, and given that it can also tolerate degeneracy, this raised the question of why PRDM9 is able to bind only a subset of all such potential binding sites in the genome.

To address this, a ChIP-seq analysis was performed to identify genome wide binding sites for PRDM9. This information also helped us to characterise binding sites and investigate if factors such as the local chromatin environment play a role in specifying PRDM9 binding targets and hotspot formation. We were able to identify over 170,000 PRDM9 binding sites in the genome. Surprisingly, these binding sites were also enriched in promoter regions, however, bound sites in these regulatory regions showed low recombination activity. We found that PRDM9 is able to confer the H3K4me3 mark on all bound sites, even those without a pre-existing H3K4me2 mark. We also investigated the role of other chromatin related marks on PRDM9 binding and found that binding occurs in chromatin accessible, but nucleosome rich regions, whereas heterochromatin regions tend to inhibit binding. Further, for hotspot formation, it was seen that less chromatin accessible, nucleosome dense regions away from transcribed sites, are preferred. Hotspots tend to avoid regions marked by transcription activating histone modifications, however, these regions do not appear to inhibit PRDM9 binding itself. These results show how PRDM9 binding in the genome is dependent on both primary DNA sequence and the surrounding epigenetic factors. Together these factors promote binding and, with additional downstream factors, positioning of hotspot locations in the human genome.

Contents

Chapter 1

Introduction

1.1 Meiotic Cell Division: A brief overview

It was the insightful experiments of Robert Brown followed by Schwann and Schleiden in the early 1800s which showed that plant and animal tissues are made of numerous cells, each containing a circular structure called a nucleus [1]. Rudolph Virchow and Robert Remak in the mid 1800s demonstrated the ability of cells to reproduce by cell division, an observation that marked a cell as the functional unit of life [2, 3]. The work of Walther Flemming in the late 1800s further explained the process of cell division. Using an aniline dye to stain the cells of a tadpole's tail, known to divide rapidly, he identified structures called chromosomes which distributed equally between two dividing cells- a cell division process which he called as Mitosis [4].

At around about the same time, Gregor Mendel showed the inheritance patterns of traits explaining variability and evolution in pea plants [5]. By the late 19th century, the process of fertilisation by union of sex cells (sperm and eggs) had been described, but how these cells contributed to forming new life remained unknown. Edouard Van Beneden had observed that all cells of a species of round worm contained four chromosomes (diploid), but its sex cells contained half that number (haploid) [6]. What led these germ cells to contain only half the number of chromosomes? August Weismann, in 1890, explained this by reporting

that two cell divisions are required to transform diploid cells into haploid cells [7]. This process of reduction division was later named as Meiosis. Further, Thomas Morgan in the early 20th century first observed the process of crossing over in Drosophila Melanogaster, and presented the chromosome theory of heredity, which explained that genes are transmitted on chromosomes [8].

1.1.1 Meiosis

Meiosis is an essential process for sexual reproduction in eukaryotes which creates new combinations of genetic material, leading to variations in each generation [9]. This is a highly regulated process during which chromosomes undergo a round of replication, followed by the first nuclear division (Meiosis I or reduction division) where homologous chromosomes segregate and move to opposite poles, and finally a second nuclear division (Meiosis II) where sister chromatids segregate yielding four haploid gametes [10]. The timing of meiosis in mammals differs in males and females. In males, the spermatogonia (male germ cells) enter meiosis after puberty. Whereas in females, the oogonia (females germ cells) initiate meiosis within the fetal ovary, forming oocytes, which later develop into eggs; meiosis does not complete until fertilisation. [11, 12].

Initially the germ cell is diploid (containing pairs of homologous chromosomes) and before entering meiosis the DNA replicates, leading to four copies of DNA (i.e. two almost identical sister chromatids for each chromosome). This is the S (Synthesis) phase of Meiosis, after which the cell enters the phase of Meiosis I. Meiosis I begins with Prophase I, divided into the Leptotene, Zygotene, Pachytene and Diplotene phases. During Prophase, the chromosomes first appear within the nuclear envelope with tightly bound sister chromatids [13, 11]. Next, joining of homologous chromosomes occurs, after which sister chromatids separate but non-sister chromatids remain attached as they undergo recombination or exchange of genetic material, and finally non-sister chromatids or homologs begin to separate [13]. The recombination process will be described in more detail in the next section (1.2). During Metaphase I, homologous chromosomes align at the metaphase plate (mid way of opposite poles of the cell), spindle fibres attach to kinetochores

such that the sister chromatids face the same poles (an important feature as it allows the chromatids of each homologous chromosome pair to separate). During Anaphase I the cell lengthens and homologous chromosomes separate, with the help of microtubules and spindle fibre. Telophase I leads to division of the parent cell into two daughter cells, which then prepare for the second meiotic division. This time, the cells enter Prophase II without the replication step, and continue on to produce the final products i.e., four haploid cells, each cell containing a single copy of the chromosomes [13, 14, 12].

1.2 Initiation of meiotic recombination in mammals

Proper meiotic division depends upon complex interactions between homologous (parental) chromosomes during the first meiotic Prophase. These interactions take place through recombination, a process during which an exchange of genetic material takes place [15]. This exchange is an important feature as it enables connections (or chiasmata) between homologous chromosomes, leading to proper alignment on the spindle and accurate segregation of homologous chromosomes [16]. Hence, recombination has two important mechanistic roles: 1) In the early stages of meiosis, it helps to search for homologs leading to synapsis (explained in section 1.2.2), and 2) Later in meiosis, when this exchange of genetic material binds non-sister chromatids, together providing the resistance required for the spindle to pull chromatids to their respective poles. In the absence of this tension, i.e. without at least one cross-over event within each chromosome, proper segregation of chromosomes does not occur leading to detrimental consequences for gamete formation [15, 17].

Recombination initiates with double strand break (DSB) formation, a process considered to be a universal feature for meiotic recombination which occurs at the Leptotene stage of Prophase I, after which the DSB is marked for DNA repair. In the zygotene stage there is a search for homologous chromosomes along with a process called synapsis, which aids in DSB repair. In meiotic recombination,

DNA is repaired using an intact homolog as template, contrary to somatic recombination, in which the sister chromatid in used for repair. Finally, the exchange of genetic material is initiated in the zygotene stage and continues on into the diplotene stage. This process terminates in desynapsis along the length of the homologs, other than the points where reciprocal exchange had occurred [12, 15].

1.2.1 SPO11 forms double strand breaks

As mentioned previously, the formation of programmed DSBs marks the initiation of recombination. In meiotic cells these DSBs are essential for proper synapsis or pairing of homologous chromosomes. An evolutionarily conserved protein called SPO11 is required to make these DSBs. SPO11 cleaves double stranded DNA through an attack mediated by its tyrosine side chain on the phosphodiester backbone of DNA, forming a phosphodiester bond between SPO11 and the 5' end, and creating a free 3' single stranded DNA overhang [18, 19] (Figure 1.1 a). There are about 200-250 DSbs per nucleus on average, based on RAD51 and DMC1 foci, with about 300 DSBs reported for females and about half the number (150) reported in males (as shown by immunoflorescent experiments) [15].

The specificity of SPO11 in directing DSB formation can be highlighted by the fact that Keeney et al. performed site directed mutagenesis of the *Spo11* gene and were able to see a decrease in cleavage specificity, and also reported these mutations to have caused changes in DSB patterns at a recombination hotspot [19, 15]. Also, in Yeast, *spo11* null mutations lead to loss of DSB formation, along with defects in synapsis [15]. Further, experiments by Camerini Otero et al. showed that both male and female *Spo11* knockout mice (*Spo11-/-*) are infertile. In mice spermatocytes arrest and enter apoptosis before the pachytene stage of meiotic Prophase I [19]. One study in humans has shown that a SNP in the *SPO11* gene affects DSB formation in male reproduction [20], however more research in the area is needed to confirm these associations.

SPO11 interacts with other proteins to form DSBs, using its non-conserved N and C- terminals [15]. RAD50 and MRE11, two important proteins of the MRN

complex, a complex which serves to stabilise DSBs by forming a scaffold like structure, have been suggested to work with SPO11 for double strand break formation [21]. It is also reported that these proteins are involved in the initial steps of DSB repair, as MRE11 makes the cleavage for end resection, resulting in 3' ssDNA which is then utilised in the repair process. Failure of SPO11 to interact with these proteins, and mutations in these proteins, have been shown to reduce DSB formation and in some cases cause infertility [22, 15].

After the formation of DSBs, SPO11 attached to the 5' end of DNA is removed by an endonucleolytic cleavage, cleaving several bases downstream of 5' end [23] to produce 3' end overhangs. This asymmetric cleavage is suggested to differentiate the two ends of a DSB for strand invasion and DNA repair, the next steps of meiotic recombination [24, 25].

1.2.2 Synapsis and DSB repair

DSB repair in meiosis is accompanied by the progression of synapsis (pairing of homologous chromosomes). The formation and repair of DSBs plays an important role in achieving chromosome pairing, which is evident from reports of abnormalities in synapsis in mice lacking SPO11, or having mutant DNA repair genes (*Dmc1* and *Rad51C*) [26, 27, 28]. Synapsis is a process where two homologous chromosomes are connected together along their entire length and form the lateral elements of a zipper like proteinaceous structure, called the Synaptonemal Complex (SC) The SC, discovered over 40 years ago, is composed of two lateral elements which are cohesin based homolog axes, and a central element with transverse filaments. Cohesin, a multi-subunit complex comprising of SMC1, SMC3 and (in meiosis) REC8 proteins [29], is shown to be important for the assembly of SC.

DSB repair begins with the single stranded 3' ends generated by DSBs, being bound by RAD51 and DMC1 proteins. These proteins together form nucleoprotein filaments that aim to identify a complementary sequence on the homologous chromosome. This helps to initiate a process called strand invasion (Figure 1.1

b) where one single stranded end interacts with a homologous chromatid, and as a result the unpaired strand forms a D-loop structure (Figure 1.1 c). Following strand invasion and synthesis the DSB can be repaired in two possible ways: by the DSBR (Double strand break repair) pathway [30] or the SDSA (synthesis dependant strand annealing) pathway. In case of the DSBR pathway, the second DSB resected end is captured which in turn forms two Holliday Junctions (Figure 1.1d), this structure can then be resolved as a cross over (figure 1.1d top; with one horizontal and one vertical resolution), or as a non cross-over event (figure 1.1d bottom; with two horizontal resolutions). Repair by the SDSA takes place by strand displacement and annealing of extended end to the DNA on the other break end; products of the SDSA pathway are always non cross-overs (Figure 1.1 e) [30, 31, 32, 33].

1.2.3 Gene Conversion or Non-Crossovers

Gene conversion or Non-Crossover events are the most common outcome of homologous recombination. During gene conversion, genetic information is transferred from the intact homologous sequence to the homolog containing the DSB, as gene conversion efficiency depends upon homology of interacting regions [34]. In other words, this is a nonreciprocal transfer of sequence information from one homolog to another. This process can occur between both homologous chromosomes and sister chromatids [35, 36]. When the transfer occurs between identical sister chromatids, there is no impact on genetic variation the offspring, so such events are difficult to detect. Cross-overs, on the contrary, involve a reciprocal exchange of both DNA strands between two homologous chromosomes. Non-reciprocal gene conversion still occurs near the DSB site, and thus accompanies the cross-over. The cytological manifestation of cross-overs is called Chiasma. Non-Crossover events are the more dominant form with a frequency of about 90% as opposed to 10% for crossovers in humans [37, 38].

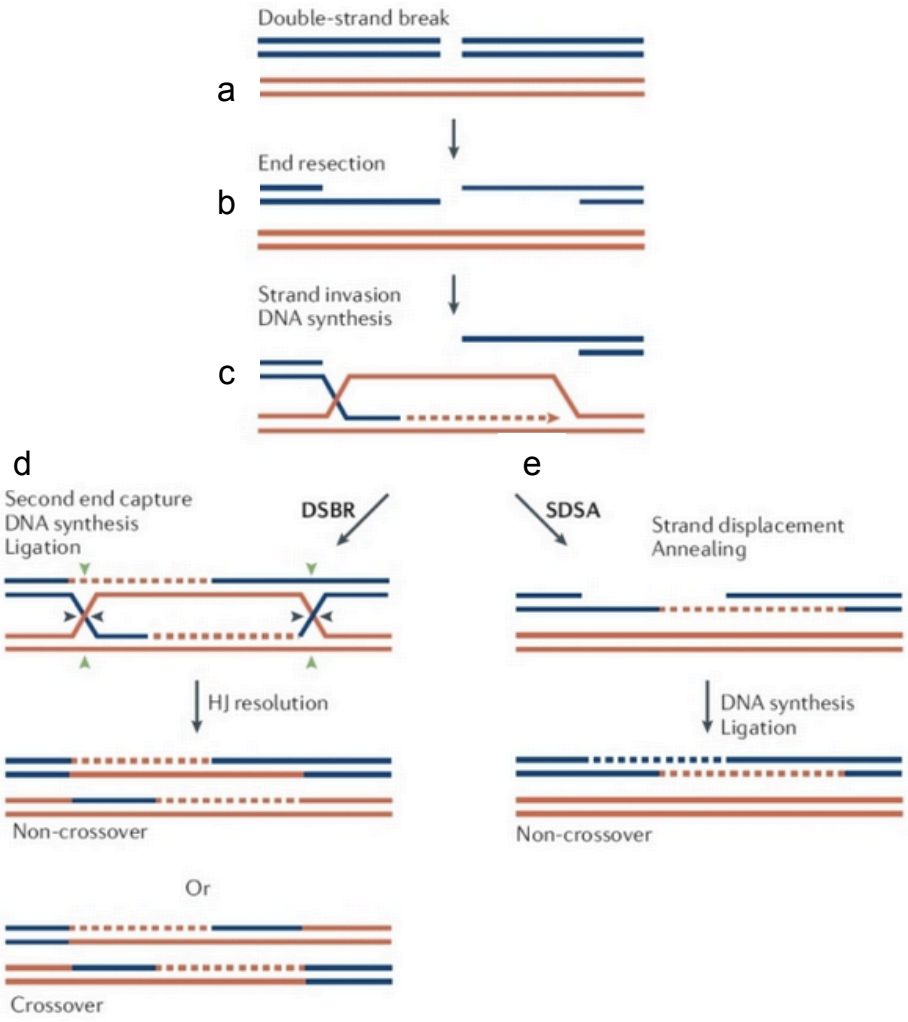


Figure 1.1: ***Meiotic Recombination.*** *A diagrammatic representation of the recombination pathway. [Figure adapted from Sung and Klein [39]]*

1.2.4 Crossovers

Crossovers, as mentioned previously, aid the correct orientation of homologous chromosomes, and their subsequent migration to opposite poles of the spindle fibre. Crossover is important not only due to its critical mechanical role in meiosis, but also because of its role in evolution, since it produces genetic variation by breaking haplotypes and reshuffling alleles [37, 40, 41]. Lack of crossovers results in incorrect segregation of homologous chromosomes, leading to aneuploidy [15]. Studies have shown a significantly reduced rate of recombination and synapsis in infertile men, and reduction in recombination has been shown to be a strong risk factor in the formation of aneuploid gametes [42, 43, 44]. These imbalanced gametes, if fertilised, may either lead to early embryonic death or pregnancy loss, or produce an offspring with chromosome imbalance and developmental issues e.g. Trisomy 21 [45, 42]. In addition, if the meiotic recombination genetic machinery is faulty, this can lead to infertility by activating meiotic checkpoints and triggering apoptosis [22].

The number of crossover events in spermatocytes or oocytes is variable, and can be measured using microscopy by counting the number of MLH1 foci (MLH1 is a DNA repair protein which repairs mismatches, and has been shown to be involved in recombination) [46]. In humans, there are about 200-400 DSBs that occur in each meiosis, and of those only 25% resolve as cross-overs. Studying MLH1 foci in humans has shown about 50 cross-overs occur per cell. In females the level of crossing over has been shown to be about 1.4 times greater than males [47, 48, 20].

1.2.5 Recombination rates in the genome

The variation in recombination activity in expressed as cM/Mb, where cM (centiMorgan) is the unit of genetic distance and Mb (one million base pairs) is the unit of physical distance. This recombination intensity can be used to identify hotspots and coldspots (i.e. regions of low recombination activity) e.g. by looking at the deviation that comes from comparing recombination intensity in a certain region to the intensity of the whole genome, or by comparing it to ad-

jacent regions i.e. if the recombination intensity of a region is higher than its surroundings, it would be marked as a cross-over hotspot [41, 49]. The average cross-over frequency is reported as 1.1cM/Mb in humans and 0.5cM/Mb in mice [40, 50, 51]. Sex averaged crossover rates are seen to be higher towards telomeres and lower near the centromeres, with these rates being strongly associated with features including gene density and GC content [41]. It is interesting to note that diversity within humans, and between humans and other species, tends to increase with broad scale recombination rates [49]. In females, chromosomes have longer genetic maps, lower crossover rates at telomeres and higher rates near centromeres compared to males. Crossovers are not randomly scattered at fine scales, however, but rather tend to be very strongly clustered into 1-2kb regions of the genomes, called as recombination hotspots [41, 49, 52].

Hotspots in humans occur at a higher rate near genes, but preferentially outside transcribed regions [37, 53, 54]. Both hotspot density and recombination intensity is greater in telomeric regions, compared to centromeric regions [55]. Although recombination is a process involving extensive DNA repair, it has been shown by fine mapping that exchange points of crossovers are precise to the base pair [37, 56].

1.3 Recombination Hotspots

In mammals, the first recombination hotspot was discovered in the MHC region in mice, on chromosome 17 [57]. A few years later, while analyzing familial inheritance patterns, the first hotspots in humans were discovered, in the *beta*-globin and insulin regions. The number of hotspots identified in humans gradually grew further over the following 25 years [37].

1.3.1 Methods to detect Recombination Hotspots

The three main approaches that have been used to identify human meiotic recombination hotspots are 1) The use of pedigree analysis 2) Statistical analysis using data for genetic variations and 3) Sperm typing [58, 41, 37], and we briefly

discuss these here.

Pedigree analysis in humans is used to identify currently active hotspots, and involves mapping cross-over events among offsprings of families in the pedigree at high resolution. The drawback of this method is the cost associated with the collection and genotyping of thousands of samples [59, 60, 50]. This approach has historically been used only for broad-scale rates, and has only recently been applied to find hotspots.

The statistical approach uses the fact that adjacent SNPs form haplotype blocks (or clusters of SNPs) of about 10-100kb and within each of these blocks, most SNPs are in linkage disequilibrium (LD) i.e. allelic types are associated with those nearby SNPs, therefore, positions in the genome when association breaks down often correspond to cross-over recombination hotspots, because recombination is the only genetic force that can break down association between markers, provided (as in humans) repeat mutation events occur only rarely [37]. With the help of coalescence based statistical models, LD data can be used to infer locations where boundaries of haplotypes mark historical hotspots. This approach has been very valuable in identifying recombination events distributed in the human genome, and has also been applied in other species [60]. The LD based hotspots identified by this method have been validated by the sperm-typing approach, and known hotspots in humans identified by sperm-typing, conversely, have also been discovered by the LD based methods. Therefore, this approach is a powerful tool to determine broad scale and high resolution recombination rates [61, 60].

Sperm typing uses pools of sperm drawn from one or more men, and uses allele-specific PCR to selectively amplify recombinant sperm, which can then be fully sequenced [62]. Sperm typing has almost always been used to investigate individual hotspots, and enables to define hotspots at a width of 1-2kb, while identifying hotspot centres within tens of bases. It has helped to identify SNPs that influence the initiation role at specific hotspots, and also allowed us to explore the variability in hotspot intensity among individual men [63, 64]. Sperm typing has allowed exquisite precision of inference of cross-over, and non-cross-over patterns

at about 40 human hotspots [62, 58]. This approach, however, is very challenging (as these are expensive and technically demanding experiments) to extend to the whole genome. More recently, whole genome sequencing of individual sperm has been attempted [65], with future possibilities of learning genome-wide maps for individual humans. We will study 10 hotspots characterised by sperm-typing in chapter 2.

1.4 What determines the location of mammalian recombination hotspots?

The factors that determine the location of recombination hotspots are only beginning to become clear. From ongoing research in this area, there is evidence of both cis and trans-acting factors are known to play roles in determining hotspot activity [37]. A number of very important discoveries have been made over the last few years that have greatly contributed to this field of research. I describe below some of these notable experiments and their findings that aimed to investigate these questions in detail.

In 2005, Myers et al. [66] constructed the first high-resolution cross-over map in humans using genetic variation data, and identified around 35,000 recombination hotspots [55]. Recombination hotspots were identified using a method called LDhot, while recombination rates were estimated using the program LDhat. LDhat was applied to an existing data on genetic diversity which included 1.6 million SNPs from three human populations (set of 24 European Americans, 23 African Americans and 24 Han Chinese). This map revealed that recombination hotspots at about 50kb intervals throughout the human genome tend to occur outside transcribed regions [66, 67].

1.4.1 Sequence association with Hotspots

Myers et al. also investigated whether factors including genomic repeats or specific sequence motifs, are associated with human recombination hotspots. Among repeats, they showed that long terminal repeats of two retrovirus-like retrotrans-

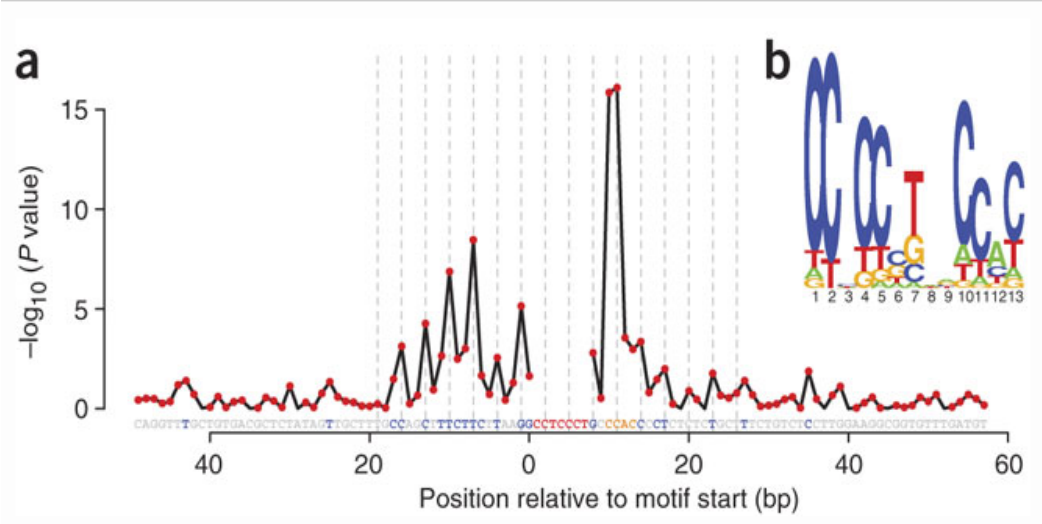


Figure 1.2: ***Predicted PRDM9 binding site and flanking sequence.*** *a) Role of flanking sequence and motif degeneracy on LD based hotspot activity. The motif is shown in red/orange letters. Vertical dotted lines spaced at 3-bp intervals show periodicity of strongly signaled bases, while bases showing evidence of association with recombination rate suggest a 36bp binding motif, highly degenerate at many sites. b) Degeneracy within the core 13-bp motif estimated by comparing counts of each motif mismatching exactly 1 bp of the 13-bp core CCTCCCTNNCCAC in hotspots and matched coldspots [reproduced from Myers et al. [68]]*

posons, THE1A and THE1B, are strongly associated with hotspots [68]. On further investigation, a 7-nucleotide motif CCTCCCT was found to be enriched 5-fold in those repeats within hotspots, relative to other THE1A/B repeats that did not mark hotspots. Computational analysis revealed that if this motif occurs within the THE1A/B repeat background, it results in an identified hotspot 60% of the time [55]. Further, results from sperm typing studies showed that disrupting this 7-mer could directly reduce rates at hotspot DNA2 on chromosome 6. This implied the role of a distinct sequence feature in recombination activation. In addition, the identical 7-bp motif was found to mark hotspot locations even outside these repeat regions [55].

This analysis was further refined by using greater hotspot resolution available with the HapMap data [68]. Taking the previously identified 7-mer (CCTCCCT) in non-repeat DNA, Myers et al. identified 50 bases flanking the region that contributes to marking the presence of hotspots (figure 1.2a). This analysis revealed an additional 4bp sequence, CCAC, just two bases downstream of the 7-mer. The new extended 13-bp "core" motif "CCTCCCTNNCCAC" (figure 1.2b) was strongly correlated with hotspot activity, when it occurs in both repeat and non-repeat DNA [68].

The core 13-mer motif showed degeneracy at five positions: 3, 6, 8, 9 and 12. Mismatches at positions 3, 6 and 12 of the core motif still showed some level of hotspot activity. The consensus sequence, after accounting for degeneracy at those positions, is referred to in this thesis as the 13-mer "degenerate" motif: "CCNCCNTNNCCNC" (see figure 1.2). This degenerate 13-mer motif marks the location of at least 40% of all human hotspots identified by LD, with notable variation in motif penetrance across different genetic backgrounds e.g. In THE1A elements, this motif makes a hotspot 73% of the time [68]. This motif is similarly present near the centre of most of the hotspots identified by sperm-typing e.g. DNA2, DNA3, MS32, DMB2 [69, 70, 71], and is now to play a functional role specifying human hotspot location (see below).

This 13-bp motif, however, on its own is not sufficient to confer hotspot activity, as this sequence can also be found in non-hotspot regions. As stated above the motif is also not found in about 60% of the hotspots so is apparently not necessary for hotspot activity. Together these facts raise raise the question: Are there other cis or trans-acting factors in the genome that play a role, together with the motif, in marking the location of human recombination hotspots?

Suggested cis-acting factors, other than the 13-mer motif sequence, include GC content, repeat elements, methylation marks, gene density and sequence variation [69, 70]. For example, sequence differences in mouse strains at two hotspots, Psmb9 and Hlx1, are seen to affect the recombination rates [72]. Initial evidence

of *trans-acting* features being associated with recombination hotspots came from Shiroishi et al. in 1991, who showed that sequences flanking the Psmb9 hotspot in mice affect both sex-specific and absolute recombination activity [73]. Similar evidence in humans was reported by Neumann et al. in 2006, who showed that although a 10kb sequence around the MSTM1b hotspots was exactly the same in sampled men, hotspot activity at these hotspots differed significantly. This pointed to a potential role for external factors that may be responsible for regulating hotspots [74, 37].

Smagulova et al. looked at other genomic features that might influence hotspot locations, and reported some interesting findings. They noted that sequences 5' of the centre of hotspots tend to be rich in purines, but this bias changes in the centre of the hotspots where the 3' sequences tend to be rich in pyrimidines [75]. This purine-pyrimidine skew, observed in both mice and humans, has been noted previously in other functional elements such as transcription start sites or origin of replication [76, 34].

1.4.2 PRDM9 marks mammlian hotspot location

In 2010, PRDM9, a trimethyl transferase, was reported to be associated with hotspot activity in both humans and mice, by three independent groups [53, 77, 78]. The role of PRDM9 in hotspot specification has since then been established by many groups [79, 80, 81, 82].

Myers et al. on further exploring the function of the 13-mer motif [68], observed a pattern of threefold periodicity within and outside the 13-mer motif. As it was unlikely for this to be driven by coding sequence, given hotspots are located away from transcribed regions, this suggested a role of a trans-regulatory DNA motif binding protein [68].

They next aimed to determine potential candidate proteins that may bind to the 13-mer hotspot motif, thereby directly or indirectly influencing crossover activity. Based on the 3-bp periodicity observation and the longer sequence (over 30bp; fig-

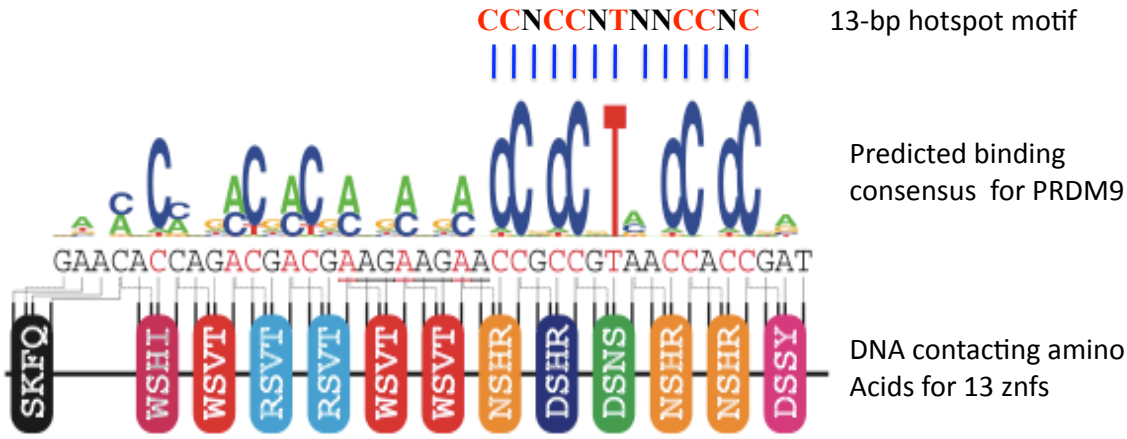


Figure 1.3: ***PRDM9 binds the 13-mer motif.*** *a) Predicted binding sequence of PRDM9, aligned to 13-mer hotspot motif. Red letters in the 13-bp motif indicate non-degenerate bases, black letters indicate degenerate bases. Below the 13-mer motif, is the PRDM9 predicted binding sequence; the logo plot is aligned to positions of the motif with vertical blue lines (the height of letters in the logo plot indicate estimated probability of hotspot activity). Lastly, below the PRDM9 binding consensus are 13 ovals indicating the 13 zinc fingers of PRDM9, with lines showing their predicted base contacts within the 13-mer motif; the letters in the ovals are the DNA contacting amino acid residues of each zinc finger used to make the binding target prediction [Figure from Myers et al. [53]].*

ure 1.2) containing the motif, they guessed the binding protein to be a zinc finger protein with at least 12 zinc fingers. Possible zinc finger proteins were explored, with the condition that candidates should have a region with predicted sequence specificity matching the degenerate 13-mer motif. On examining the predicted consensus binding targets for all of the 691 human C2H2 zinc finger proteins, this motif was present within the predicted binding sequence of five proteins. Further, comparing the predicted motif degeneracy of each candidate protein with degeneracy patterns in the 13-mer motif, PRDM9 was the sole candidate whose prediction exactly matched observed degeneracy at positions 3,6,8,9, and 12 of the 13-mer motif [68] (see figure 1.3).

In parallel, two additional lines of evidence supporting the role of PRDM9 in recombination came from experimental approaches reported by Baudat et al. and Parvanov et al. [77, 78]. Baudat et al. examined the Dsbc locus on chr 17, which has been shown to regulate the distribution of crossovers in several regions on different chromosomes, partly by regulating recombination at the initial steps of DNA double-strand break formation [16, 83]. On further fine-mapping the Dsbc1 locus, Baudat et al. observed that this region contains only the Prdm9 gene [77].

As earlier findings had shown that a high level of H3K4me3 is associated with recombination at the psmb9 and Hlx1 hotspots, and also that different Dsbc1 alleles exhibit varying levels of H3K4me3 at these hotspots [84], Baudat et al. were able to hypothesise that the Prdm9 gene, known to encode for histone methyltransferase [85], was the most likely potential candidate causing the Dsbc1 effect, and that Prdm9 zinc fingers could bind specific targeted areas in the genome. Firstly, they noted that polymorphisms in the DNA binding zinc finger region of Prdm9 are associated with differences in hotspot usage in two mouse strains. Secondly, they confirmed through an in-vitro experiment, that the human PRDM9 reference allele is able to recognise the 13-mer motif enriched in hotspots. This suggested the role of PRDM9 might be conserved between species, which led them to sequence PRDM9 in humans, revealing variations in the zinc finger array which they found to associate with differences in hotspot usage in humans. These ob-

servations established that PRDM9 binding to specific sequences is associated with recombination hotspots in humans and mice [77].

1.4.3 Properties of PRDM9

The above mentioned evidence confirms a role for PRDM9 as a trans-regulatory factor involved in specifying hotspot location [53, 77, 78, 79, 80]. The features of this protein add insights into this role. PRDM9 is a Histone 3 Lysine 4 trimethyl transferase (H3K4me3) and is expressed solely during prophase in meiosis [86]. It contains at least 6 separate domains: KRAB domain, which is likely to serve for protein-protein binding, a PR/SET domain responsible for the trimethylation activity, a zinc knuckle, an upstream zinc finger, an SSX domain and an array of Cys2-His2 zinc fingers for DNA binding [87, 86]. Based on the canonical model of how C2H2 zinc fingers bind DNA, within the zinc finger array each zinc finger is predicted to bind trinucleotides of DNA sequence targets (figure 1.4). In humans, the reference PRDM9 allele contains 13 zinc fingers [87, 86]. The PRDM9 zinc finger array is encoded by a minisatellite sequence, where 28 amino acids code for each zinc finger [88]. The C2H2 zinc fingers each have a β-hairpin, unusually, followed by a single α-helix. The amino acids of the N-terminal of the helix are the ones that come in contact with the major groove of the double stranded DNA. Relative to the first amino acid of the α-helix (taken as position 1), the three amino acids at positions -1, 3 and 6 are the ones that contact DNA and bind to three DNA bases sequentially [86], while base 2 is involved in contacting a base overlapping the target of the adjacent zinc finger. Zinc finger repeats are almost identical in sequence, except at positions which encode for amino acids at coordinates -1, 3, 6 of zinc finger alpha helices [87, 86, 89].

The zinc finger array of PRDM9 shows extremely rapid evolution in mammals [86], and differs between humans and chimpanzees [54]. This explains the fact that despite 99% sequence identity between the two species, the 13-bp motif is not associated with hotspot activity in chimpanzee, and more generally, humans and chimps do not share hotspot locations [54][53]. Myers et al. looked at 22 human hotspot regions that contained the 13-mer motif in both humans and

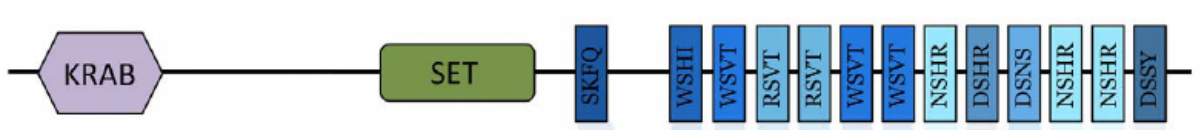


Figure 1.4: ***A subset of PRDM9 Domains.*** *N-Terminal KRAB and SET domains and C-terminal zinc fingers. Figure adapted from Sgurel et al. [90]*

chimps, and observed that no hotspot was conserved between the species [53]. The differences between humans and chimps in PRDM9 are concentrated in the DNA-contacting residues within the zinc finger regions. In another study Auton et al. [54] sequenced 48 western chimpanzee PRDM9 haplotypes and showed PRDM9 to have extensive variation within western chimpanzee. Variation was found in both the number of zinc fingers (the most common PRDM9 alleles contain 6,16 and 18 zinc fingers) and the type of DNA contacting residues within zinc fingers [54, 91].

Loss of PRDM9 produces total sterility, with an inability to repair meiotic DSBs and pachytene arrest in both sexes [92]. PRDM9 has been suggested to play a role in male sterility in both mouse and humans [87, 93, 94]. Given its important function, it is interesting that this protein does not appear to exist in the genomes of dogs (where it is apparently a pseudogene), birds and fruit flies [87, 86, 95, 96]. The PRDM9 zinc finger array is also variable in human populations [79]. Ancient hotspot activity appears to be strongly associated with the common (or similar) versions of PRDM9 alleles [37, 90], which are very similar to the reference sequence allele whose properties we explore in this thesis. Intriguingly PRDM9 is the only known speciation gene in vertebrates [87, 86].

1.4.4 PRDM9 variation influences hotspot activity

The property of PRDM9 variants to affect hotspot activity was shown by Baudat et al. and Berg et al.[79, 77]. Baudat et al. showed that in humans, hotspot usage (fraction of crossovers that occur in recombination hotspots based on LD data, which are expected from above to mainly reflect particular PRDM9 alleles)

varies with the variation of *PRDM9* alleles. On sequencing and genotyping the zinc finger arrays from individuals of European ancestory in a Hutterite population, a number of *PRDM9* alleles differing in both the number and sequence of zinc finger repeats were identified. The major allele found was the "A" allele, which is similar to the reference "B" allele (the difference being at the one amino acid in the 6th zinc finger). Other alleles found were called C, D, E, K and I. Three of these alleles, allele "A" , "B" (reference allele) and "I" occurred at 94%, 4% and 2% frequencies, respectively. Allele "I" had amino acid changes in the zinc finger array which do not predict the 13-mer motif to be this allele's binding site. On testing interactions of PRDM9 A and I allelic variants with their respective predicted binding motifs in-vitro, Baudat et al. were able to confirm that the A allele had high binding affinity to the 13-mer DNA motif; whereas binding of the I allele was specific to its own predicted motif and this allele had a lower affinity for the 13-mer motif, as expected [77].

The role of PRDM9 further investigated by Berg et al. who aimed to explore the impact of *PRDM9* allelic variation on crossover activity at hotspots with and without 13-mer motif [79]. They hypothesised that recombination activity would be high in sperm samples carrying PRDM9 alleles that bind the 13-mer motif as opposed to males lacking these alleles. They typed *PRDM9* zinc finger alleles in European and African individuals, and found a diversity of alleles which ranged from having 8 to 18 zinc fingers (see figure 1.5). They selected ten active hotspots, five containing a central 13-mer motif match, and five without a clear motif match. For each of these hotspots, semen donors were identified carrying either A/A, A/N or N/N *PRDM9* alleles, N alleles being defined as non-A alleles. On comparing crossover frequencies in men of these three genotypes, they observed that all N/N men tested (with one exception) showed crossover activity of less than 5% the activity in AA individuals. Interestingly, this pattern was observed at all hotspots i.e., even those that do not contain clear motifs. These results support the previous findings which suggest PRDM9 as a major regulator of hotspot activity. These findings were not however able to establish a relationship between the activating allelic versions of *PRDM9* and presence of the 13-mer sequence motif, except for an over-representation of this motif in the hotspots

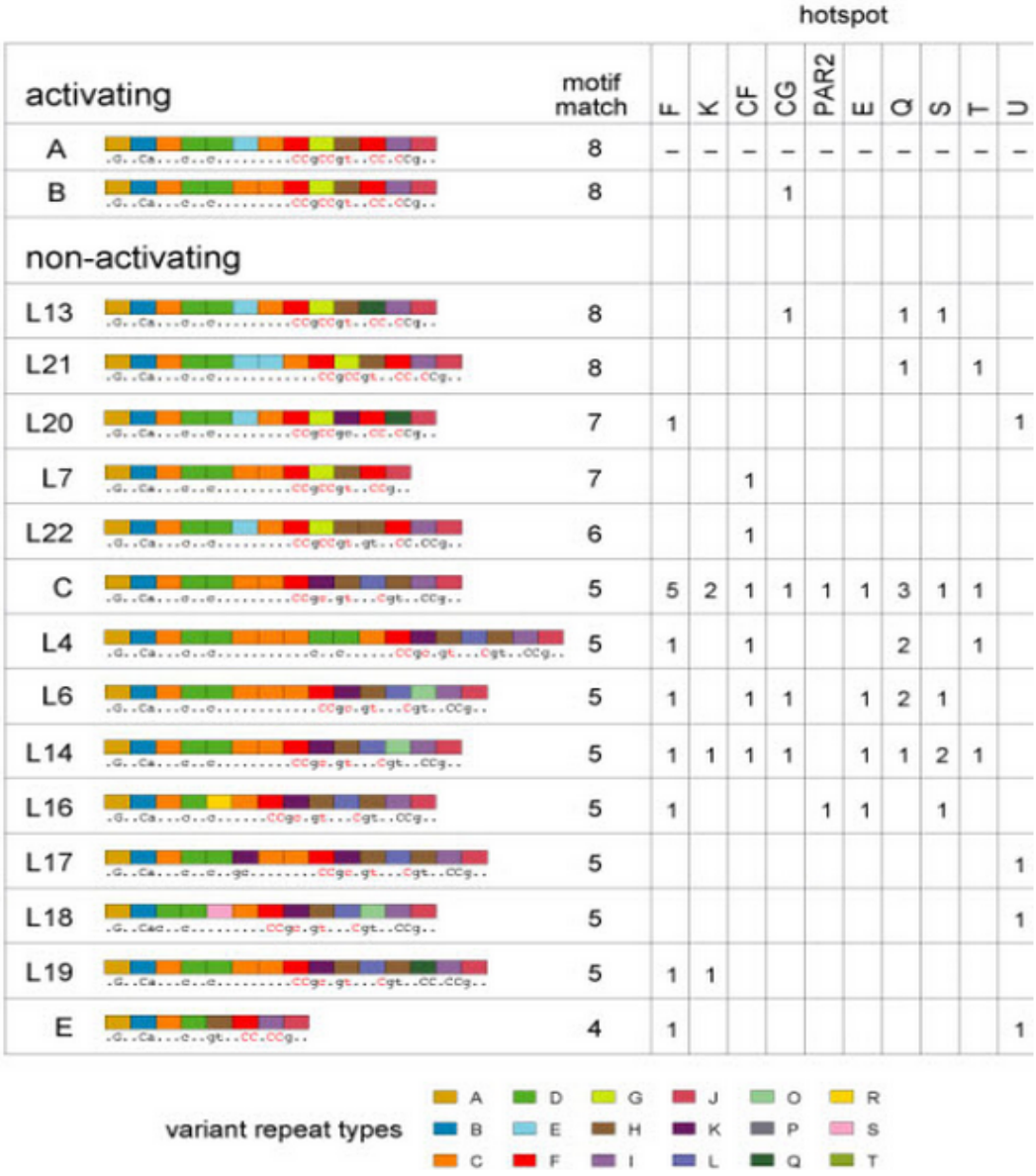


Figure 1.5: ***PRDM9 Alleles:*** *Variation in PRDM9 alleles activating or not activating 10 tested hotspots. Figure adapted from berg et al. [79]. Only the A and B alleles activate these hotspots; entries in the table show number of men tested.*

tested [79]. Hence, they determined that PRDM9 on its own has a strong effect on recombination activity in sperm, irrespective of the presence of the 13-mer hotspot motif, also finding that even small changes in the zinc finger array are capable of creating hotspot activating, or non-activating, variants. Similarly, in mouse Smagulova et al. have shown a mouse motif, and Brick et al. have further demonstrated that the different PRDM9 alleles in mice change 99% of the DSB sites [20, 75].

1.4.5 PRDM9 variation produces hotspots of high activity in African populations

Berg et al. tested whether PRDM9 variants are able to activate a different set of hotspots more common in Africans [80]. They observed that certain allelic variants (Ct) were more commonly found in this population. Individuals with these PRDM9 alleles were reported to be able to use a different set of hotspots. Notably, although a number of PRDM9 variants are predicted to share similar target motifs, these hotspots were shown to be activated only by distinct PRDM9 variants, further implying complex interplay between the zinc finger binding regions and hotspots [80].

In another study, Hinch et al. mapped genome-wide differences in recombination landscape between African American individuals and Europeans [82], and found about 2500 hotspots active only in African Americans but not Europeans. On mapping variants genome-wide they found that only PRDM9 type reflected which landscape an individual used, with Ct type alleles as a group largely predicting the landscape in African Americans. This implies a dominant role for PRDM9 in hotspot landscape differences among humans. Strikingly, the authors found that individuals carrying only Ct alleles did not use any of the European hotspots, implying PRDM9 controls all, or almost all, human autosomal hotspots. This also agrees well with the findings that PRDM9 controls almost all hotspots in mouse [20, 75].

1.5 Epigenetic marks associated with hotspots

The discussion in the earlier sections highlights some important research findings, implying that PRDM9 has a central role in defining the location of recombination hotspots. But a key question that remains is what genomic features set the stage for PRDM9 to bind targeted regions of the genome, causing them subsequently to become hotspots? We know that 13-mer motifs are enriched in hotspots. However, there are over 300,000 such words in the human genome [90], and only a fraction of them produce recombination hotspots (i.e. around 40% of the hotspots contain these motifs [68]). Motifs occurring in THE1A/B repeat elements in humans almost always make hotspots, implying a role for additional sequence features, whereas mouse hotspots also appear to be associated with repeat elements and GC rich regions, but these cases only contribute to a fraction of the total hotspots [68, 75].

This implies that the sequence motif itself, is not necessary or sufficient to make hotspots. What other factors could be involved? Although not much is known at present, there have been reports pointing in the direction of additional genomic factors, such as epigenetic marks, that can serve as predictors for recombination activity. This evidence comes from studies done in yeast, mouse and humans [75, 97, 12]. Research from these different model organisms points to an independent or collaborative role of primary sequence, chromatin accessibility, higher order chromosome structure, nucleosome occupancy, histone modification marks and transcription factors in generating a conducive environment for DSBs and meiotic recombination to occur [97, 12]. In the next few sections, I will describe in more detail what each of these factors are, and how they may contribute to marking hotspot locations.

1.5.1 Chromatin accessibility

The structure of chromatin only started to be understood in the 1970s with the finding that chromatin is built with nucleosomes [98]. Later on, in the 1980s it was found that chromatin structure plays a role in regulating genes and that

active genes tend to have a more open or accessible chromatin structure. In general, it was observed that the regulatory regions of genomes tend to be associated with open chromatin and nucleosome free regions, and these regions were later referred to as DNase I hypersensitive regions (DHSs) [99].

Chromatin accessibility is suggested to be an important factor in determining the location of recombination hotspots. There have not been many studies aiming to look at this association directly. However, the ones that have been reported agree with the idea that accessible chromatin serves as a background that promotes meiotic recombination. In yeast, studies have provided evidence that recombination sites have open chromatin structure [100, 101, 102, 103]. For example, Ohta et al. reported that in budding yeast, DSB variability and recombination frequencies were found to correspond with modifications in chromatin structure. It was seen that hypersensitivity to DNase increased at hotspots earlier on in meiotic prophase i.e. before the formation of DSBs. One interpretation for this was given by Bergerat et al. [104], suggesting that a chromatin modification is necessary early in meiosis which serves as a substrate for DSB inducing enzyme SPO11. Another explanation was given by Kleckner et al., who suggested that change in chromatin accessibility could be the result of a recombination complex that assembles at the given site before a DSB is formed [18].

Wu and Lichten presented the evidence for an association between DNase hypersensitive sites and DSB sites. However, they also noted that accessible chromatin is necessary but not sufficient to generate DSBs, as regions with lower DSB frequency at the two promoter regions tested retained their chromatin accessible patterns, and deletions that increased the frequency of DSBs at these sites did not alter DNase accessibility [105].

In mammals, evidence for a similar association has been reported by a few studies. Shenkar et al. tested if an increase in recombination activity results from increased DNA accessibility. They looked at a well established meiotic recombination hotspot in mice and found a DNase hypersensitive site in this region, near which they located binding sites for three Transcription Factors (TFs). This led

them to conclude that binding of TFs may contribute to increase in recombination activity by modifying chromatin structure making it more accessible for the recombination machinery [106]. However in humans and mice (unlike in yeast) few, if any, hotspots are found at DNase accessible promoter regions. Interestingly, Smagulova et al. [75] found that in Prdm9-/- mice DSBs re-localised to H3k4me3 peaks, often at promoters, which suggests that PRDM9 may actually play a role in keeping DSBs away from promoters and DNase hypersensitive sites if marked by H3K4me3. A resolution to these differences has not yet been found.

1.5.2 Nucleosomes

Chromatin by nature is not accessible and normally occurs as repressed by nucleosomes [107, 108]. Nucleosomes contain about 147bp of DNA wrapped around a histone ocatmer. The histone octamer contains two copies of the core histone proteins H3A, H2B, H3 and H4 [107]. The nucleosomes are separated by about 20 base pairs of DNA, which is referred to as the linker region [109, 110].

The genome contains binding sites for TFs, whose recruitment or assembly leads to a change in chromatin conformation. Since the mid 90s there has been a lot of research in the area of chromatin modifications and modifying enzymes. There is now considerable information on the specific kinds of modifications found at promoters, genic regions, intron and exon boundaries etc. Recent studies [111, 112] exhibit evidence of a positive association between nucleosomes and recombination hotspots. Castro et al. [113] generated a genome-wide nucleosome profile for fission yeast, and noted that there was a significant amount (over 80%) of colocalization between the origin of replication and DSBs in intergenic regions [113]. They also reported that the nucleosome depleted regions were the only common feature that was shared by all meiotic recombination hotspots in the fission yeast [111, 114, 113].

Getun et al. looked at nucleosome profiles at four recombination hotspots in mice and their findings suggested that nucleosome occupancy appears to direct recom-

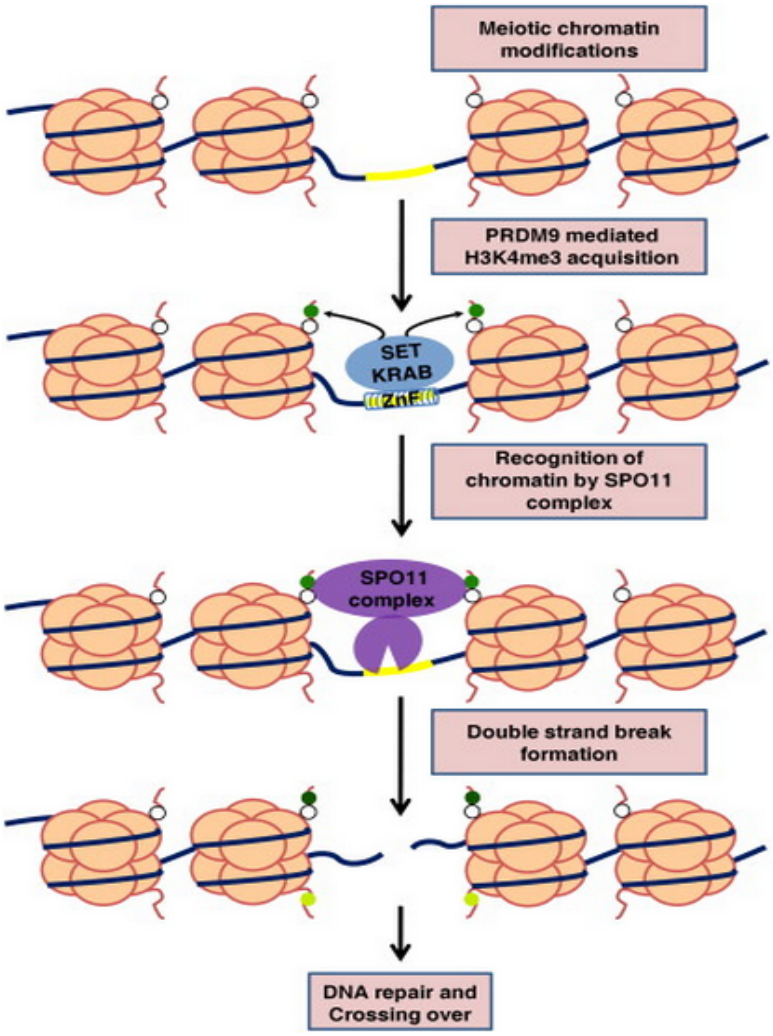


Figure 1.6: ***Model for PRDM9 mediating meiotic recombination.*** *PRDM9 may potentially be recruited to recombination hotspot sites by existing favourable histone modifications (white circles) or other chromatin related factors. It makes an H3K4me3 mark and modifies chromatin. This chromatin modification triggers SPO11 activity to produce double strand breaks. Figure adapted from Satya et al. [115]*

bination activity. They observed that the centres of these hotspots were mainly DNase accessible, but interestingly, at sites within the hotspot centre where nucleosomes occurred, these sites showed little or no cross-over activity [112]. On the contrary, Smagulova et al. reported that higher nucleosome occupancy was associated with recombination hotspots in mice, and hypothesized that hotspots may be involved in nucleosome positioning [75]. There is still, therefore, much about the mechanics of this association (and that of histone modifications) that remains to be understood (see figure 1.6 for one possible model).

Another interesting recent finding by Smagulova et al. [20] was that the H3K4me3 marks at mouse hotspots are specific to hotspots i.e. recombination sites were seen to be associated with testis specific H3K4me3, and were distinctly different compared to the H3K4me3 marks at transcription sites.

1.5.3 Histone modifications

Histones are able to undergo a number of modifications, made by histone modifying enzymes, which correlate with cellular processes like transcription activation and repression. To date over 100 such modifications have been reported. For example, histone marks associated with transcription activation include H3K4me3, H3K36me3, H3K9ac, H3K14ac, H4K12ac whereas transcription repressing marks include H3K27me3, H3K9me2, H3K9me3 etc. [116, 117, 118].

In non-mammalian species a number of mutations in enzymes that have a role in post translational histone modifications have been reported to reduce DSB activity, which is suggestive of an association between these histone modifications and DSB formation or recombination activity. In yeast, Yamada et al. and others noted that the H3K4me3 mark occurs at meiotic DSB sites [119, 120]. DSBs are reduced in the absence of H3K4 methyltransferase enzyme, Set1, in yeast [121]. Borde et al. [84] showed that in yeast, H3K4me3 is enriched at DSB sites, independently of gene transcript levels. They also showed that in the absence of Set1, DSB formation dropped at 84% of the hotspots.

De Massy et al. [122] examined the relationship between post-translational histone modifications and the psmb9 recombination hotspot in mice, aiming to differentiate between marks that occur before recombination activity and ones that occur after. They showed that there were a number of histone modification marks that were enriched at the hotspot. H3K4me3, H3K4me2 and H3K9ac were enriched at the recombination initiating site and were also present in *Spo11-/-* mice, suggesting that these precede the initiation of recombination at the psmb9 hotspot. They also noted that hyperacetylation of H3K4 was enriched at both recombining chromatids, implying their increase is a result of DSB formation. These features were also common to another mouse hotspot Hlx1 [85].

1.5.4 Transcription Factors

Evidence for non-histone (other than PRDM9 and other known parts of the recombination machinery) proteins interacting with recombination hotspots to mediate their activity has been supported by a number of groups. In E.coli, an enzyme called RecBCD was reported to interact with a particular motif called Chi, to increase recombination [123, 124]. In yeast, White et al. performed an experiment to show that the binding of transcription factors is able to stimulate recombination activity at the HIS4 hotspot. They noted that the HIS4 hotspot contains binding sites for transcription factors Rap1p, Gcn4p, Bas1p and Bas2p and mutations that abolish the binding sites for Rap1p or Bas2p can eliminate hotspot activity at this site. This implies that transcription factor binding and recombination appear to be associated. However, the authors noted that removal of the TATA sequence which is required for HIS4 expression does *not* alter recombination activity [125, 126, 127].

Kirkpatrick et al. further followed up the findings of White et al.[125] and reported that HIS4 hotspot activity is dependent on Rap1p transcription activators, but transcription at HIS4 is not associated with recombination activity. However, they noted that this recombination inducing activity of Rap1p does require the

transcription activation domain. They showed that a hybrid protein containing the Gal4p-DNA binding domain with Rap1p transcription activation domain is capable of stimulating recombination by binding to the Gal4 binding site upstream of the HIS4 region. The authors termed hotspots in yeast cells that are transcription factor dependant as "alpha hotspots" and those not requiring transcription factors as "beta" hotspots [126].

In yeast, Kon et al. studying the M26 hotspot, established that the Mst1/Mst2 heterodimer is important for the activity of M26 hotspot. Other similar studies have reported that this heterodimer is also responsible for regulating the timing along with activation of many additional hotspots. This finding again provides evidence that specific transcription factors may be required to bind cis-acting sites to activate recombination hotspots [128]. Yamada et al. also showed the role of histone acetyl transferases, coordinating with ATP-dependant chromatin remodelling factors, in selective activation of transcription to mediate recombination activity at the M26 hotspot [129].

Together these findings indicate that a complex network of transcription factors may be involved in activation of recombination hotspots. Possible mechanisms that bring about this collaboration may be that 1) transcription factors are able to interact with some other recombination initiating complexes to stimulate recombination (e.g. the activation domain of a TF may be able to help recruit recombination machinery) or 2) binding of transcription factors may result in modification of chromatin rendering the chromatin accessible, which may in turn activate transcription and/or recombination [130, 131]. The next chapters will describe in more detail about how we looked into the association of the epigenetic marks mentioned above, and others, with recombination hotspots in humans. We will also find evidence for interactions between TF binding and recombination activity of PRDM9.

1.6 Aims and objectives

In the light of recent findings, that have contributed to highlighting the role of PRDM9 in hotspot regulation, broadly, this research work aims to address two main questions: 1) Is PRDM9 able to directly bind exact matches to the 13bp motifs, as well as the less exact, or degenerate versions of the motif in the genome? and 2) What features determine PRDM9 binding at its target sites, and subsequent hotspot formation? To investigate this we will map PRDM9 binding sites in the genome, and look at various epigenetic marks surrounding the binding sites with an aim to understand the type of local chromatin environment that is likely to be most favourable for PRDM9 binding and cross-over hotspot activity. To answer these questions, we will use both in-vitro and in-vivo approaches.

Chapter 2

PRDM9 sequence targets in-Vitro

2.1 Introduction

We know that PRDM9 seems to contest all hotspots [79, 77, 53], however, no identified motif targets are seen in 60% of the hotspots. This is highlighted by Berg et al.'s work mentioned in the previous chapter, which tested PRDM9 activity at 10 hotspots, 5 of which contained the motif and 5 did not. We aimed to use an experimental approach, electrophoretic mobility shift assays (EMSAs) [132], described below, to ask if PRDM9 binding of specific targets could be confirmed in these hotspots. We also aimed to understand PRDM9 binding to good or poor motif matches. Hence, the primary aim of work described in this chapter was, broadly, to address the following questions: 1) Is PRDM9 able to bind specifically to the 13-mer motif? and 2) Is it able to tolerate degeneracy by exhibiting binding to more degenerate versions of the motif? We used an in-vitro approach (EMSAs) to attempt to answer these questions.

The EMSA assay is a sensitive technique, which detects if the protein of interest can interact with a chosen DNA sequence on a polyacrylamide gel. It makes use of the fact that free running DNA will have a higher electrophoretic mobility than that of a larger protein-DNA complex. Radioisotope labelled (^{32}P) DNA probes

are used, the mobility of which can be detected by autoradiography [132]. The speed at which these complexes move through the gel is determined by features including their size and charge. The control lane is run with the radiolabelled "probe only" i.e. it is run without the protein present and will therefore contain only the unbound DNA. In the lanes with both probe and protein, if the protein is capable of binding to the DNA fragment, this lane will show a band that represents the complex of DNA bound to protein. Additional lanes are run with competitor probes to determine the most favorable binding sequence for the protein in question. A competitor probe is not radiolabelled, and could be a related (e.g. an identical "self" probe) or an unrelated sequence to the radiolabelled probe which contains the binding target for the protein. Competition with related sequence should result in a decrease in band intensity, as less protein is then available to bind to the radiolabelled probe. In contrast, an unrelated competitor, which is not bound by the protein of interest is not expected to have any affect on band intensity, as the protein should only bind to the radiolabelled probe being tested. These lanes can then confirm binding that is specific to the sequence being tested. Further, a super-shift lane is able to recognise identity of the protein binding the tested target sequence, using an antibody against it.

To apply this approach to study PRDM9 binding requires recombinant PRDM9, or nuclear extracts with sufficient amount of PRDM9. Hence, we designed gel-shift experiments with both commercially available testis nuclear extract, and recombinant protein. The task of synthesizing recombinant PRDM9 was initially assigned to a company (GenWay Bio); in the case they were unable to successfully deliver full-length purified protein, a parallel objective was to synthesize full-length PRDM9, in collaboration with the Structural Biology Unit (STRUBI) at the Wellcome Trust Centre for Human Genetics.

In the first section of this chapter I will discuss the initial experiments performed using testis nuclear extract and company synthesized PRDM9, and the caveats involved. These experiments were somewhat unsuccessful in expressing PRDM9, though allowed us to optimize experimental conditions. In the later sections, I will discuss our findings from the gel-shifts using full length PRDM9, that was

synthesized in collaboration with STRUBI, and which did allow us to study binding properties of the protein.

2.2 Initial Gel-shift experiments

2.2.1 EMSAs with testis nuclear extract

In line with our primary objectives discussed above, we aimed to test PRDM9 binding to various motif-containing and non-motif containing DNA sequences. We initiated a first set of gel-shift assays using commercially available human testis nuclear extract. As the extract is composed of a mesh of cells at all stages of the cell cycle, we may hope that PRDM9 is present in sufficient amounts to be detectable by sensitive gel shift assays. We used two probes for the initial experiments, referred to as: THE1 and Cold-1 (Table 2.1). The THE1 probe contains the 13-bp core motif present in the THE1 repeat background sequence, and is expected to bind PRDM9 based on LD predictions (THE1 repeat elements have been reported to show elevated recombination rates in the presence of this canonical motif [68], as discussed in more detail in chapter 1). Cold-1 probe is a sequence from a random recombinationally inert region showing no evidence of hotspot activity and is not expected to bind PRDM9. Further details of the full list of probes tested for this and subsequent experiments are given in Table 2.1.

Supplementary figure 1 shows results from one of these experiments, which tests binding of PRDM9 (potentially present) in the nuclear extract to THE1 labelled probe. A DNA-protein complex was observed, which implied sequence specific binding, as indicated by fainter bands when competed with 10 and 100 fold excess of unlabeled THE1 oligos. On competing with an unrelated and unlabeled "cold" competitor probe, we saw no difference on binding affinity, suggesting binding is probe specific. However, we were not able to see a super-shift in the PRDM9 antibody lanes which aim to test if the complex is in fact bound by PRDM9.

The main concern with these initial experiments using testes nuclear extract was the fact that PRDM9 is only expressed in the prophase stage of cells undergoing

Core 13-mer motif: ‘CCTCCCTNNCCAC
Degenerate 13-mer motif: ‘CCNCCNTNNCCNC

Probes	Sequence	Description	LD based predictions
THE1	TGCCTTCCATCATGATTATGAGGCCTCCCTAGCCACATGTA	Motif in THE1 repeat background	Binds; localizes LD based hotspots
THE1_snp2	TGCCTTCCATCATGATTATGAGGCATCCCTAGCCACATGTA	Targeted disruption in THE1 motif at position 2	Does not bind
THE1_snp5	TGCCTTCCATCATGATTATGAGGCCTCTCTAGCCACATGTA	SNP in THE1 motif at non-degenerate position 5	Does not bind
L2	AATGTCACCTCCTCAGTGAGGCCCTCCCTGACCACCCAGTT	Motif in L2 repeat background	Binds; localizes LD based hotspots
SNP2_L2	AATGTCACCTCCTCAGTGAGGCCTTCCCTGACCACCCAGTT	SNP in L2 motif at non-degenerate position 2	Does not bind; No hotspot in LD
SNP12_L2	AATGTCACCTCCTCAGTGAGGCCCTCCCTGACCGCCCAGTT	SNP in L2 motif at degenerate position 12	Binds
THE1_perm	CACCATCGTGTGCCAAAATGTTTTCTCTACCCGTCTCAAGG	Permuted version of THE1 motif [Control: cold sequence]	Does not bind
In-Silico	GAACACCAGACGACGAAGAAGAACCGCCGTAACCACCGAT	In-silico sequence match, on comparing predicted motif degeneracy for PRDM9 with degenerate patterns in 13-mer motif	Binds
LD_consensus	TGCCAGCTTTCTTCTTAAGGCCTCCCTAACCACCCCTCT	LD consensus sequence indicating the best binding sequence for PRDM9	Binds
Cold-1	CTCAAATGATCTGGCTGGCAGTGATCTCATGTGACCTGTCA	Random cold DNA sequence	Does not bind

Table 2.1: Probes designed for gel shift assays. Sequences above the table are the core and degenerate versions of the 13-mer motif, with N indicating degenerate positions. The table provides oligo sequences with red letters indicating the location of the 13-mer motif within each oligo sequence. The blue underlined letters indicate mismatches to the core motif.

meiosis. To ascertain if PRDM9 was present in the extract, would require one to perform Western Blots to confirm, which in turn require a "good" antibody against PRDM9; these confirmation experiments were not carried out at this stage (given negative results in any case; Later we *do* perform such experiments). The antibody requirement would also hold to carry out super-shifts in EMSAs to ensure that the protein-DNA complex detected by these assays did in fact include PRDM9. We initially used a commercially available PRDM9 antibody. However at that stage, it was unknown how well the antibody could work under these assay conditions. Another concern was that the protein-DNA complex observed appeared to be moving very quickly through the gel, which implies the complex is formed by a protein having a low molecular weight. We would expect this complex to be higher up on the gel i.e. with lower mobility, owing to the fact that PRDM9 is a large protein of predicted molecular weight around 100 KDa (see later results). Therefore, at this point it seemed unlikely that we were observing true PRDM9 binding to our probe, and we felt it was in any case difficult to validate results, given the issues listed above. The next set of experiments then used the company synthesized PRDM9, which we hoped might provide greater concentration and purity of protein.

2.2.2 EMSAs with company synthesized PRDM9

The next gel-shifts were performed using PRDM9 synthesized by GenWay Bio. [Brief description of protocol: HEK293 cells had been transfected with vector DNA encoding for Myc (C-terminal) and Flag (N-terminal) tagged PRDM9. PRDM9 was provided in three batches. Batch 1 contained PRDM9 isolated from cytosolic fraction, with additional protein contaminants (Western Blot showed there was a contaminant band which was not recognized by MYC or FLAG antibody). Batch 2 and 3 were described as containing PRDM9 isolated from cytosolic and nuclear fraction respectively, with no additional protein contaminants visible on SDS-PAGE].

Supplementary figure 2 shows results from one of the gel shift experiments using

PRDM9 from Batch 1 (containing full length PRDM9 and also contaminant proteins). We investigated PRDM9 binding with two probes: L2 and L2-snp2. Both probes contain motifs, but incase of the latter a SNP at non-degenerate base 2 of the motif, is expected to disrupt binding. A protein-DNA complex was seen in both lanes with these probes. There is in-vivo evidence that a SNP at the second non-degenerate base of the motif is expected to reduce binding activity. Myers et al. [68] showed that the L2 repeat carrying either core or degenerate version of 13-bp motif showed a narrow peak in average recombination rate centered at the motif. However, in the case where this motif in L2 repeat region was disrupted at position 2 from "C" to "T", there was no such peak observed indicating no recombination activity as a result of this change. This evidence, however, appears inconsistent with the band we observed in the L2-snp2 lane. Further, we looked at PRDM9 binding with two cold probes i.e. THE1-perm and Cold-1, which do not confer any hotspot activity. Again, both cold probes show the same pattern of binding as seen in the case of the motif containing probe L2. Finally, an assessment of all three batches of PRDM9 sent by the company showed that only batch 1 shows binding with the motif containing THE1 probe, whereas batches 2 and 3 (containing truncated PRDM9) show no evidence of binding.

These equivocal results led to concerns about the quantity and quality of PRDM9 present in these batches. On further examination and discussion with company researchers, we discovered that batch 1 was the only one containing full length PRDM9, but was also the only batch with a contaminant protein. Batches 2 and 3 contained a truncated version of PRDM9 (about 50-60kDa in size). The truncated protein was recognized by Flag antibody in western blots. Since the Flag tag was attached to the N-terminal end of PRDM9, this meant the truncated PRDM9 present in batches 2 and 3 did not contain the DNA binding zinc finger domain. The company researchers were not able to see any positive results on western blot using Myc-antibody to detect the C-terminal myc tag of PRDM9, which was quite concerning. Given these problems, it became very unclear whether the DNA binding protein from batch 1 was actually full-length PRDM9, or the contaminant protein. In light of the existing issues, to be able to make any meaningful inferences from in-vitro experiments, it was decided to

initiate synthesizing full-length PRDM9 protein, in collaboration with STRUBI at the Wellcome Trust Centre.

2.3 Gel-shifts with full-length PRDM9

All subsequent gel-shift experiments were performed using the full length recombinant PRDM9. The full length protein was synthesized in four different versions, with respect to tag type and the terminal to which it was attached. The four constructs prepared were 1) PRDM9 with N-terminal His-tag, 2) PRDM9 with C-terminal His-tag, 3) PRDM9 with N-terminal GFP-tag and 4) PRDM9 with C-terminal GFP-tag. The reason for preparing these different constructs was to ascertain that at least one would have stable conformation without compromising functional changes owing to the presence of tags, and also to facilitate the extraction/ purification process which may be more efficient for proteins expressed with tags at the N-terminal rather than the C-terminal, or vice-versa. In order to express PRDM9, transfection was carried out in insect cells as opposed to mammalian cells, as in the latter these cells appeared to have a low rate of transfection and expression, when tested, and so were not able to produce enough PRDM9 for purification.

2.4 Methods

2.4.1 DNA cloning

For expression in insect cells baculo virus expression system (pBacPak9) was used. pBacPac9 can efficiently produce a large amount of recombinant protein in insect cells. Also, the post-translational processing of mammalian proteins expressed in insect cells is similar to processing in mammalian cells, which makes the biological activities of mammalian recombinant proteins similar to those expressed in mammalian cells. Constructs were cloned into pBacPac9 vector using the cloning plan (supplementary figures 3 and 4).

2.4.2 Large Scale Transfection

Constructs were transfected into insect cells, and transfected cells were visualised under the microscope. Figure 2.1 shows expressed Venus tagged PRDM9 localizes inside the nucleus, which confirms that the protein functions as expected, as it is able to be transported into the nucleus after post-translational modifications.

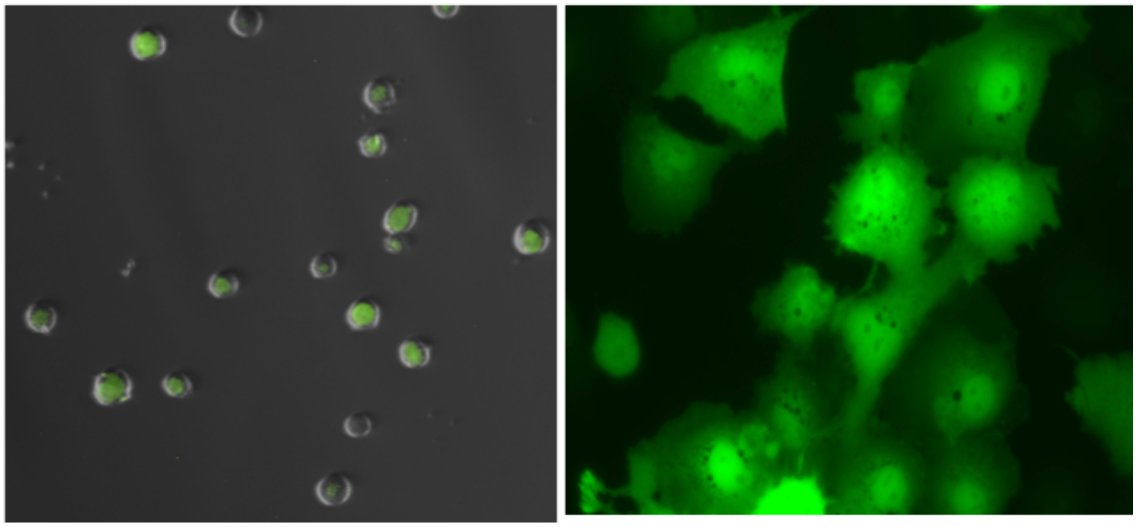


Figure 2.1: *Image showing GFP-tagged PRDM9 expressed in the nucleus of insect and mammalian COS7 cells.*

2.4.3 Extraction/Purifiction of PRDM9

Cells were lysed with high salt lysis buffer and chromatin was sheared by sonicating at 40% amplitude for 30 seconds with a total of 10 cycles on ice. The supernatant, after spinning cells at 4000 rpm for 10 minutes and containing cytoplasmic material was discarded. The nuclear pellet was further treated with extraction buffer, to extract nuclear proteins and was spun at 18000 rpm for 45 minutes. The supernatant from this spin containing nuclear extract was then used for further purification. Partial purification of His-Tagged PRDM9 was done using a column of cobalt-coated talon beads, which have a high affinity for the His tag. Briefly, the beads in column were first equilibrated with water. The

supernatant was passed through the column (collected as flow through), the column was then washed with approximately 400ml of extraction buffer to remove non-specifically bound proteins (collected as wash). Finally, beads attached with proteins were eluted out with Imidazole buffers of varying concentrations (20mM, 50mM, 250mM and 500mM), with higher concentrations aiming to elute more tightly bound proteins. A western blot was run to determine which of these elutions (if any) contained the His-tagged PRDM9. PRDM9 was eluted at 20mM concentration of Imidazole (see figure 2.2). Eluted protein was aliquoted and stored at -80^0, and later used to perform gel-shift assays.

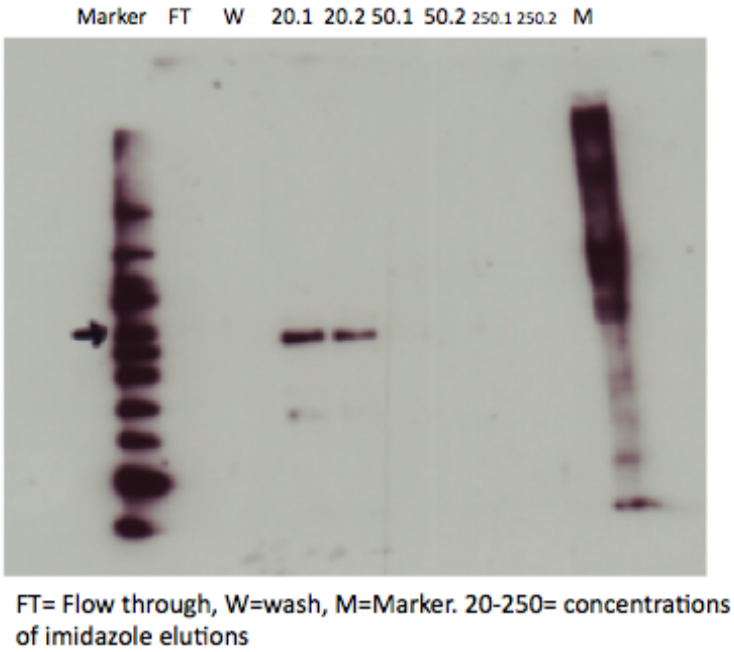


Figure 2.2: *Western blot results for PRDM9: A 100KDa band (indicated by arrow) corresponding to molecular mass of full-length PRDM9, is observed at 20mM imidazole elutions.*

2.4.4 Labelling DNA probes for Gel-shift experiments

Oligos (designed with agct overhangs to facilitate labeling reaction) were first annealed with NEB buffer 2 set in boiling water and left overnight to anneal. The annealed oligos were incubated with dCTP, buffer, klenow enzyme and 32P

dCTP for 45 mins. The enzyme was deactivated at 65^0 for 10 mins. Reaction mix was passed through a resin column and spun for 2 mins to remove any radioactive nucleotides below 5bp in size. Probe radiation activity was then measured with a scintillation counter as counts per minute. Finally, water was added to get a final 6,667 cpm/ul for each probe. We used 3ul of this labeled probe for each binding reaction in EMSA (i.e 20,000 cpm/ reaction or lane).

2.4.5 Binding reaction

Binding reactions were carried out by incubating recombinant protein in binding buffer, dIdC, BSA and the radiolabelled probe for 15 minutes at room temperature. Competition experiments were carried out as following: First, a 15-20 minute incubation of PRDM9 with unlabelled competitor (e.g. Anti-His Antibody for supershifts or related / un-related competitor at different dilutions) in the binding buffer in cold room (4^0). The 2nd incubation was a 10-15 minute incubation of PRDM9+unlabelled competitor mix with radiolabelled probe at room temperature.

2.4.6 Preparing EMSA Gels

The reaction mix was then loaded on to a 5% polyacrylamide gel, which was set at 4^0 and run at 200V for 3 hours. The gel was then dried on 3mm blotting paper covered with saran film at 80^0 for 1 hour. The dried gel was then exposed to film in a cassette overnight ($\sim$ 16-18 hours).

2.5 Results

2.5.1 PRDM9 binds the canonical 13-mer Hotspot Motif

The first set of gel shift assays were performed to test the concept that PRDM9 binds in-vitro to the 13-mer motif. As discussed previously, an LD based consensus sequence, which we used to produce an "LD" probe, is predicted to be a strong binding sequence for PRDM9 zinc fingers; this probe contains the stringent

core (CCTCCCTNNCCNC) motif. This "core" motif was obtained by investigating repeat (THE1A/B and L2 elements) and non-repeat DNA, for flanking bases within 50 bp that are influential in determining hotspot occurrence. This revealed the sequence "CCTCCCTNNCCAC" to be the strongest determinant of hotspot occurrence. On further testing of motif occurrence outside repeat elements, and mismatching a single base of the 13-bp core motif, an additional degeneracy at positions 3, 6 and 12 within the motif was observed. Mismatches at these positions still confer some level of hot-spot activity. The "degenerate" consensus sequence was CCNCCNTNNCCNC. As the first exploratory experiment, the LD probe was tested along with an in-silico sequence match to the bioinformatically predicted PRDM9, containing the degenerate version (CCNCCNTNNCCAC) of the motif [68]. Figure 2.3 shows that PRDM9 binds strongly with the LD probe; the unlabeled self competitor at 5 and 20 fold excess competes with the labelled LD probe, while the 20-fold excess of an unrelated control sequence does not compete, implying binding to the LD probe is sequence specific. The protein-DNA complex is recognized by the anti-His antibody, causing a supershift, which confirms that it is in fact His-tagged PRDM9 that is bound to the LD probe. A similar binding pattern with PRDM9 is seen for the in-silico probe. The above mentioned motif containing probes are compared to a control, the THE1-permuted probe, which shows no binding to PRDM9 and no supershift with anti-His antibody, as expected.

Next, two more core motif containing probes were tested, whose sequence comes from recombinationally active THE1B and L2 repeat backgrounds of DNA as determined by LD predictions. Motifs in the THE1 background within the human genome have been shown to result in hotspots 60% of the time, and are predicted to have a two-fold greater level of recombination activity compared to the L2 background. As both probes contain the stringent 13-mer motif and show hotspot activity in humans, they would be expected to bind well with PRDM9, with certain disruptions, specially in non-degerate positions of the motif sequence, causing a disruption in binding. Figure 2.4 shows that THE1 probe binds to PRDM9. With increasing concentrations of unlabeled THE1 probe, we see a decrease in band intensity, implying competition is effective. The unlabeled cold

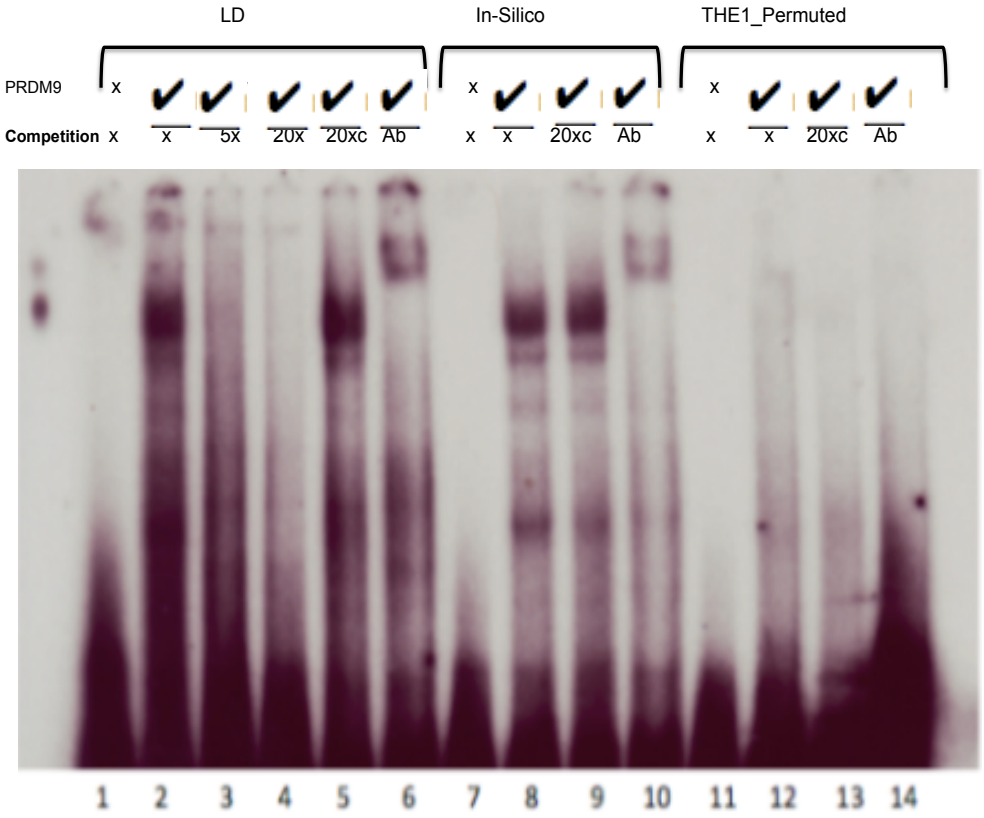


Figure 2.3: *PRDM9 binding to LD and in-silico based consensus sequence containing the 13-mer motif; Lane 1: LD probe only, Lane 2: LD probe and PRDM9, Lane 3: LD probe with 5 fold excess of self unlabeled competitor, Lane 4: LD probe with 20 fold excess of self unlabeled competitor, Lane 5: LD probe with 20 fold excess of unlabeled THE1-permuted competitor, Lane 6: LD probe with PRDM9 super-shifted with anti-His antibody, Lane 7 : In-silico probe only, Lane 8: In-silico probe and PRDM9, Lane 9: In-silico probe with PRDM9 and 20 fold unlabeled THE1-permuted competitor, Lane 10: In-silico probe with PRDM9 super-shifted with anti-His antibody. Lane 11: THE1-permuted probe only, Lane 12: THE1-permuted probe with PRDM9, Lane 13: THE1-permuted probe with PRDM9 and 20 fold unlabeled Cold-1 competitor, Lane 14: THE1-permuted probe with PRDM9 super-shifted with anti-His antibody.*

competitor, THE1-permuted probe is unable to compete with THE1, and finally supershift with anti-His antibody confirms the band as a PRDM9 and THE1 probe complex.

2.5.2 Single base disruptions in the motif can abolish PRDM9 binding

Next, we performed targeted single base disruptions on THE1 motif to determine if they have an impact on PRDM9 binding. Probes were designed so that in one case the motif was changed at base 2 from C to A (probe: THE1-snp2), and in another case there was a C to T change at position 5 (probe: THE1-snp5), i.e the motif in both cases was disrupted at a non-degenerate position. It was observed that PRDM9 binding was almost abolished with THE1-snp2, but retained in the case of THE1-snp5, suggesting some (previously unknown) degeneracy at position 5 within the motif.

On referring to computational evidence [53] which provides the estimated probability for each nucleotide at a given position in the motif, a possible explanation for PRDM9 by retaining binding to the THE1-snp 5 probe may be that while a T in place of C at position 5 is predicted to reduce activity by 5-fold on average, this is also the most degenerate of all 8 non-degenerate bases.

Testing the L2 probe shows that it binds with PRDM9 as expected, approximately as strongly as the THE1 probe. Changes in the L2 probe were made at base 2 (probe L2-snp2) and base 12 (probe L2-snp12) i.e at a non-degenerate and degenerate position in the motif, respectively. PRDM9 binding was strongly disrupted in both cases. Any changes at non-degenerate positions are not expected to strongly influence PRDM9 binding, however going back to computational evidence, it is seen that a G at base 12 is the *least* degenerate of all degenerate changes. Additionally, there could be a role of nucleotide preference at certain positions, which may be specific to a given background. There is convincing evidence from these gel-shifts that a mutation at base 2 of the 13-mer motif (from C to A or T), at least in the L2 and THE1 backgrounds, is likely to almost

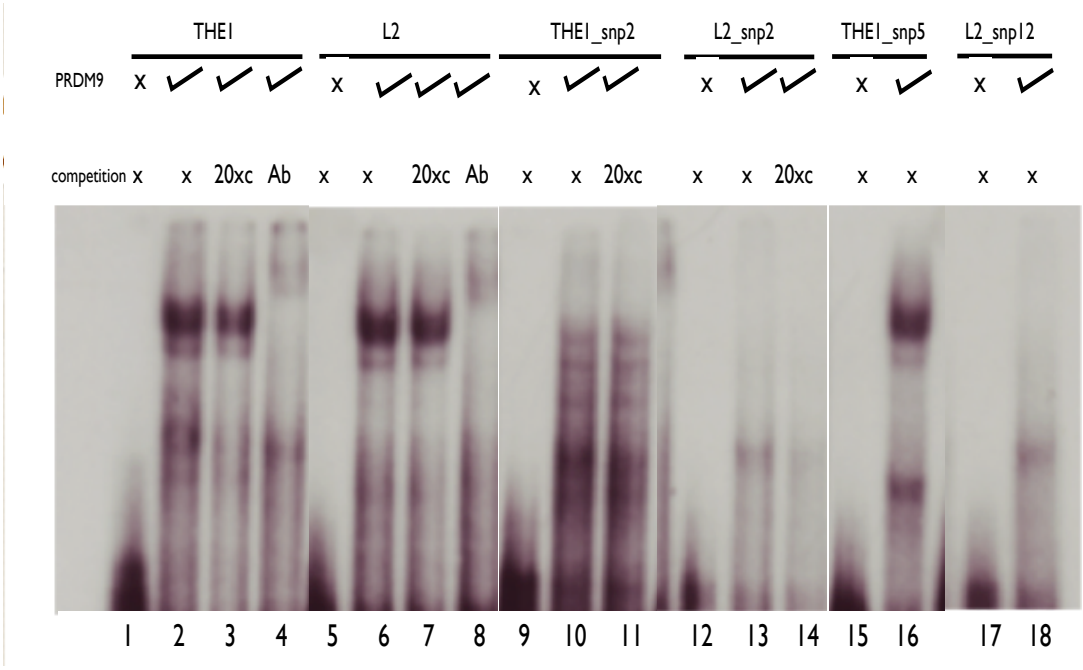


Figure 2.4: *PRDM9 binds to a probe containing the 13-mer motif within the THE1 and L2 repeat backgrounds; Lane 1: THE1 probe only, Lane 2: THE1 probe and PRDM9, Lane 3: THE1 probe with PRDM9 and 20 fold excess of THE1-permuted unlabeled competitor, Lane 4: THE1 probe with PRDM9 super-shifted with anti-His antibody, Lane 5: L2 probe only, Lane 6: L2 probe and PRDM9, Lane 7: L2 probe with PRDM9 and 20 fold excess of THE1-permuted unlabeled competitor, Lane 8: L2 probe with PRDM9 super-shifted with anti-His antibody, Lane 9: THE1-snp2 probe only, Lane 10: THE1-snp2 probe with PRDM9, Lane 11: THE1-snp2 probe with PRDM9 and 20 fold excess of THE1-permuted unlabeled competitor, Lane 12: L2-snp2 probe only, Lane 13: L2-snp2 probe with PRDM9, Lane 14: L2-snp2 probe with PRDM9 and 20 fold excess of THE1-permuted unlabeled competitor, Lane 15: THE1-snp5 probe only, Lane 16: THE1-snp5 probe with PRDM9, Lane 17: L2-snp12 probe only, Lane 18: L2-snp12 probe with PRDM9.*

completely abolish PRDM9 binding.

2.5.3 PRDM9 binds all recombination hotspots

With this first set of experiments, we are therefore able to conclude that we can determine PRDM9 binding efficiently through this in-vitro assay, and that this binding is highly specific to target sequences of interest. Our next goal was to determine if PRDM9 can bind only to the obvious 13-mer matching sites in hotspots or to sequences within other recombination hotspots. As mentioned previously, Berg et al. have shown that small allelic PRDM9 variations are sufficient to trigger or turn off hotspot activity, and that these changes affect recombination activity at both motif and non-motif containing hotspots [79]. To test the hypothesis that PRDM9 can bind in the absence of clear motif occurrences, probes were designed to test PRDM9 binding to all 10 of the hotspots examined using sperm typing by Berg et al. (Table 2.2).

Of these ten hotspots F, K, CF, CG and PAR2 are motif containing, whereas hotspots E, S, T, Q and U lack an obvious motif at the centre of the hotspot. To design these probes, the following criteria was used: Berg et al. [79] had reported matches to the 13-bp motif in their 10 selected hotspots (eg. 0-3 mismatches with core motif; or 0-2 mismatches to the "degenerate" motif), located within from the centre of hotspots. To construct our probes for each of the 10 hotspots, we selected the closest motif match, by finding the 13-bp word that was located within 50bp of the centre of these hotspots and was the closest match to the 13 bp motif, i.e. in the case of motif containing hotspots, the designed probes contained 0 or 1 mismatches with the core or degenerate motif; whereas for hotspots without a clear motif, the probes had 2 mismatches to the degenerate motif in all 5 cases. We then added 21 bases upstream and about 5 bases downstream of the motif sequence to provide sufficient bases for PRDM9 to bind without any issues of possible edge effects, given the 13-mer occurs off centre in the PRDM9 binding target, as determined either in silico or based on LD patterns. Details of number of mismatches for each of these probes is given in Table 2.2. PRDM9 showed evidence of binding to probes taken from all five out of the five motif containing

Core 13-mer motif: 'CCTCCCTNNCCAC'
Degenerate 13-mer motif: 'CCNCCNTNNCCNC

F	CAGAAAGTTACTTCCTTCTAAGCACCACCCTGACCCCTTCAT	Probes for Motif-containing hotspots
K	GACAATGCACTTCCTCTGTGATGCCCTCCCTGATCACTAGGCA	
CF	CTCAATGCAAGTTAGTTTGCGTTCCACCCTCACCACAGCTGG	
CG	TTCGCTTTCTGCCATGATTGTGAGGCCTCCGTAGCCATGTGG	
PAR2	TAAGCTGAACACACTGCCAACTCACTCCCTGCCCCCACTGCA	
E	AACCAGAAGCTGGGAGAAATAACCTCCTCTCTCCGCCTTCTAT	Probes for Hotspots lacking clear motif
Q	ACGTCACAATAGTGGATACAGAGTGCCACTGTGGCCTTTTCA GTACGTTT	
S	GTGATTTTCACCAAGATTACTAAGACCCTCCATTCACTAGGAAC	
T	TTGGTTTTACTGCCTCCATTCAGATCCTCCTTTTCTCTCTTGC	
U	CATGGGGGATTGATTGGTTCTAGGACCCCCTTGGATACCAAAA	

Table 2.2: Probes designed to test Berg et al.'s hotspots. The first five oligos are sequences from hotspots containing central motifs, the next five oligos are sequences from hotspots reported to lack a clear motif. Sequences above the table are of the core and degenerate forms of the 13-mer motif, with the letters in grey indicating degenerate positions. The table provides oligo sequences, with red letters indicating location of approximate matches to the 13-mer motif within each oligo sequence. The blue letters indicate bases not matching the consensus within the motif.

hotspots tested.

Interestingly, PRDM9 was also clearly seen to bind with all the five probes with only a weak 13-mer motif. Figure 2.5 shows binding results for 5 of the 10 probes tested (F, K, S, T, E). The red X marked next to probes S, T and E indicates probes that are the more degenerate versions of the canonical motif (allowing for at most 2 mismatches within the degenerate motif). The two rows below "PRDM9" and "Competition" are marked with checks and crosses to indicate the presence or absence of each, respectively. For example, lane 1 contains probe F only, which serves as a blank/control lane. Lane 2 contains probe and

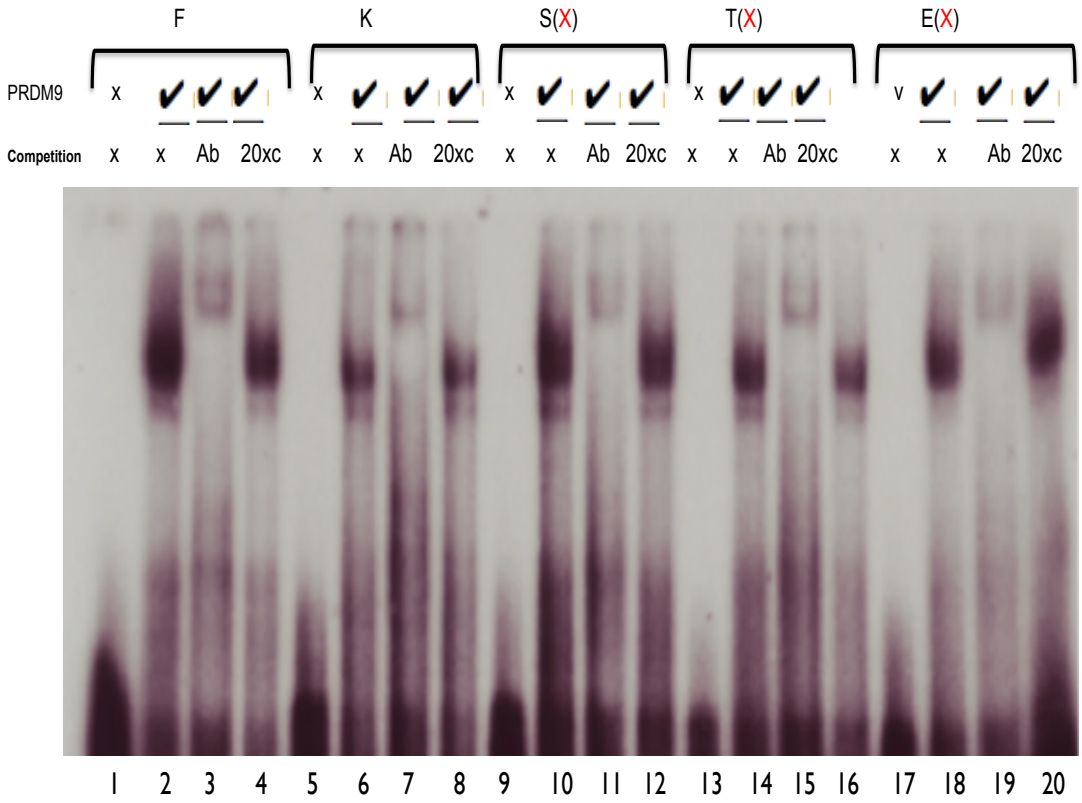


Figure 2.5: *PRDM9 binding to sequence within Berg et al.'s hotspots. The X in red next to hotspot names indicates the hotspots that are reported not to contain a match to the motif by Berg et al.: Lane1: Probe F only, Lane 2: Probe F with PRDM9, Lane 3: Supershift with anti-His antibody, Lane 4: Probe F with 20 fold excess of unlabeled THE1-permuted probe, Lane 5: Probe K only, Lane 6: Probe K with PRDM9, Lane 7: Supershift with anti-His antibody, Lane 8: Probe K with 20 fold excess of unlabeled THE1-permuted probe, Lane 9: Probe S only, Lane 10: Probe S with PRDM9, Lane 11: Supershift with anti-His antibody, Lane 12: Probe S with 20 fold excess of unlabeled THE1-permuted probe, Lane 13: Probe T only, Lane 14: Probe T with PRDM9, Lane 15: Supershift with anti-His antibody, Lane 16: Probe with 20 fold excess of unlabeled THE1-permuted probe, Lane 17: Probe E only, Lane 18: Probe E with PRDM9, Lane 19: Supershift with anti-His antibody, Lane 20: Probe E with 20 fold excess of unlabeled THE1-permuted probe.*

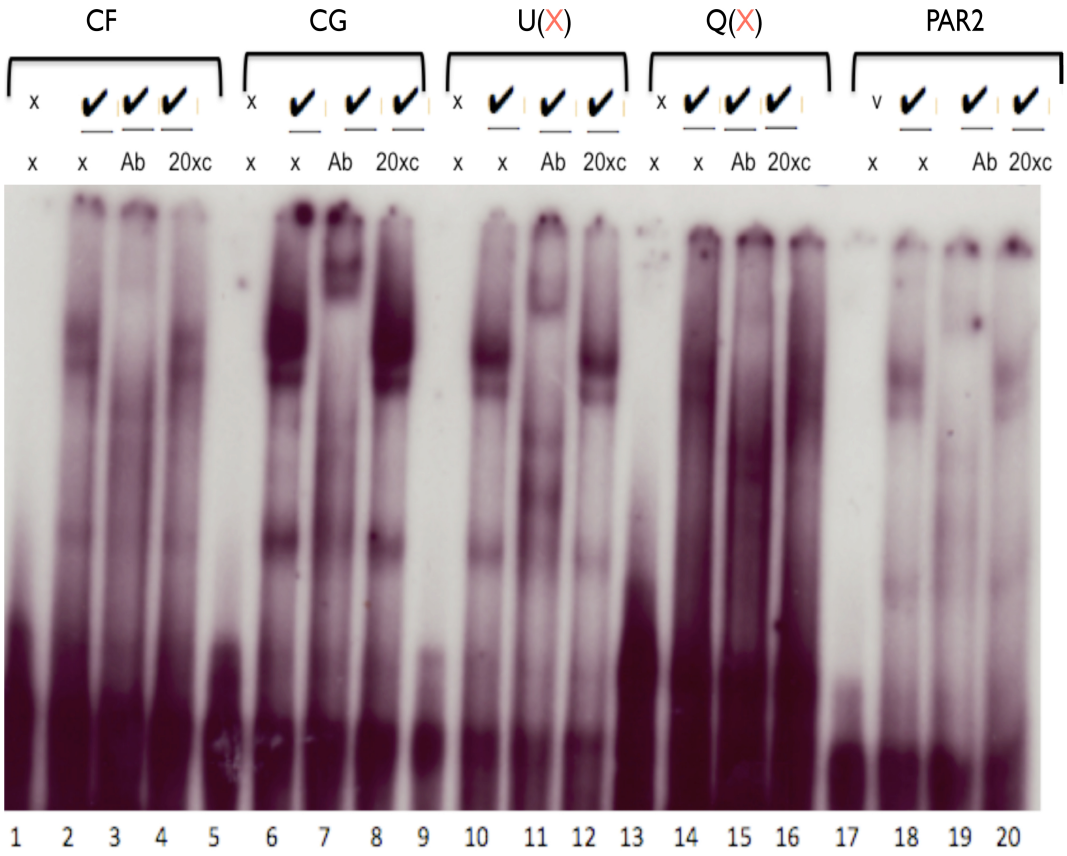


Figure 2.6: *PRDM9 binding to sequences within Berg et al.'s hotspots (continued): Lane1: Probe CF only, Lane 2: Probe CF with PRDM9, Lane 3: Supershift with anti-His antibody, Lane 4: Probe CF with 20 fold excess of unlabeled THE1-permuted probe, Lane 5: Probe CG only, Lane 6: Probe CG with PRDM9, Lane 7: Supershift with anti-His antibody, Lane 8: Probe CG with 20 fold excess of unlabeled THE1-permuted probe, Lane 9: Probe U only, Lane 10: Probe U with PRDM9, Lane 11: Supershift with anti-His antibody, Lane 12: Probe U with 20 fold excess of unlabeled THE1-permuted probe, Lane 13: Probe Q only, Lane 14: Probe Q with PRDM9, Lane 15: Supershift with anti-His antibody, Lane 16: Probe Q with 20 fold excess of unlabeled THE1-permuted probe, Lane 17: Probe PAR2 only, Lane 18: Probe PAR2 with PRDM9, Lane 19: Supershift with anti-His antibody, Lane 20: Probe PAR2 with 20 fold excess of unlabeled THE1-permuted probe.*

PRDM9; the strong band indicates PRDM9 is able to bind specifically to the sequence in the F probe. Lane 3 contains the anti-his antibody as competitor, the antibody is against the his-tag of PDM9, which in turn is bound to probe F. Together, the antibody-protein-probe complex causes a "super-shift". Finally, lane 4 contains a cold and unrelated probe as a competitor. This sequence is not recognised by PRDM9, which leaves the intensity of the protein-probe complex unaffected and is similar to that in lane 2. We were also able to see similar results for the remaining 4 probes, i.e. even ones without a clear motif match (marked with X in red) (figure 2.5, figure 2.6). Probe Q however in figure 2.6, appears to have a very weak (and noisy) binding signal; this Probe had a caveat that it contained two motif like sequences very close together on opposite strands, hence this probe contains a shorter upstream sequence than in the other 9 cases. This lack of sequence context may in turn, be influencing PRDM9 binding in-vitro.

These results provide overwhelming evidence of interaction between PRDM9 and the 13-mer motif. As shown in figure 1.6, PRDM9 is suggested to be involved in initiating recombination, by marking chromatin with the H3K4me3 mark. We thus need to further understand details of where and how PRDM9 binds and the consequences of such binding. It may bind at locations with additional sequence or chromatin features that mark the position of hotspots.

2.6 Summary

Our goal was to determine if we can assess sequence specific binding of PRDM9 to the 13-bp motif, and also examine degeneracy changes that disrupt this binding. Using gel shift assays, we were able to determine that PRDM9 shows sequence specific binding to the LD based consensus sequence containing a strong match for zinc finger binding sites, with a similar binding pattern for other core motif containing sequences in the THE1 and L2 backgrounds. A SNP or mismatch at base 2 of the motif showed complete abolition of binding (as seen in the cases of THE1-snp2 and L2-snp2 probes), a prediction which is supported by LD-based evidence from recombination hotspots in the case of L2-snp2, as mentioned pre-

viously. There were, however, also cases where a mismatch at a non-degenerate position (THE1-snp5) or a mismatch at degerate position (L2-snp12) would retain or abolish binding respectively, contrary to what we might previously have expected. In such instances, there may be a more pronounced role of sequence features within or flanking hotspot motifs, which remain unclear at this point (In chapter 4, we further explore PRDM9 binding preferences via ChIP-seq). On examining motif-like sequences at the hotspots investigated by Berg et al., we were able to observe clear PRDM9 binding to motif sequences near the centre of all hotspots tested. Even the 5 hotspots without a clear motif showed binding by PRDM9; the sequences tested at these hotspots were 6/8 matches to the degenerate 13-bp motif, placed at the centre of the hotspot in question. Together the results from these binding assays suggest that although PRDM9 is able to bind specifically to the 13-mer canonical motif, it is also able to tolerate a certain level of degeneracy within the motif sequence. It is likely that there is nucleotide preference for each position in the motif, as well as an interaction between bases, to promote PRDM9 binding. Further, along with nucleotide preference, the background context surrounding the motif is also likely to play a strong role in PRDM9 binding and subsequent hotspot formation e.g. the local chromatin environment and various epigenetic marks may influencing PRDM9 binding at recombination hotspots. These features, in light of the 13-mer motif, will be explored in the next chapter.

Chapter 3

Exploring the Chromatin Landscape around Recombination Hotspots

3.1 Introduction

We know broadly, that a protein resembling a transcription factor, PRDM9, binds the 13-bp motif, but we do not know why only some motifs are selectively bound by PRDM9 to form crossover hotspots. In order to address this, we further explored motifs present within hotspots to get a better understanding of how these motifs differ from other non-hotspot motifs in the genome. Chromatin is likely to be an important factor in determining this difference between hotspot and non-hotspot motifs [133, 111, 83], and may act in two ways: either by allowing or inhibiting PRDM9 binding, or by allowing or inhibiting subsequent crossover hotspot formation.

Hence, following the in-vitro experiments that determined binding specificity of PRDM9 with the canonical motif, we next aimed to computationally explore the association of various chromatin related features (like DNase hypersensitivity, histone modification marks, transcription factor binding sites etc.) to provide clues to factors involved in triggering crossover hotspot activity. In this chapter, we

use public data to begin to learn about these chromatin features correlating with whether a motif forms a hotspot. In the next chapter, we use PRDM9 binding information to separate binding from hotspot formation.

Here, we will identify marks enriched around 13-mer motifs occurring within recombination hotspots, relative to motifs in recombinationally inert non-hotspot regions (or coldspots) which we use as controls. Conversely, we also look for features enriched in these coldspots, which would point towards factors that prevent crossover formation. This investigation may enable us to make inferences about possible biology underlying observed association and how they may play a role in marking hotspot locations. We will revisit this topic in chapter 4, using PRDM9 binding sites identified using ChIP-seq.

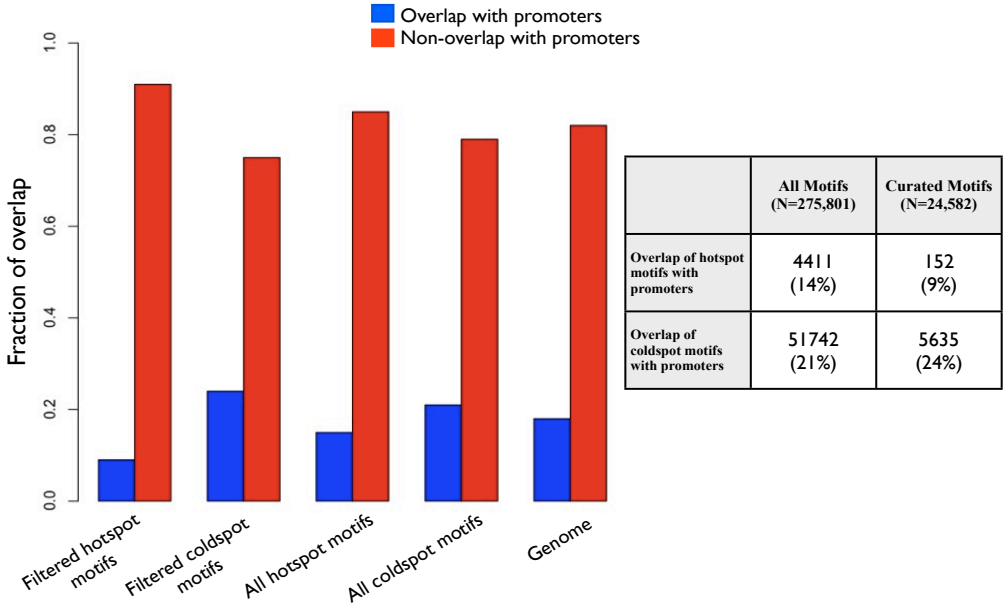
We compared the presence of any enriched epigenetic marks (available from public databases including the Encyclopedia of DNA Elements (ENCODE) [134], which aim to identify functional elements in the human genome sequence) within hotspots. The ENCODE project was initiated with the goal of cataloguing all functional elements of the genome [134]. Using powerful sequencing technologies and precise analyses, ENCODE has proved to be a useful public resource for exploring chromatin related factors through a wide range of experiments provided by participating labs. The experiments include data on various cell types (lymphoblastoid, kidney embryonic etc.), and the types of data produced broadly include data on chromatin accessibility (DNase1 hypersensitivity experiments), nucleosome positions (MNase data), histone modifications and transcription factor binding sites (ChIP-seq) [134, 135]. For the analysis performed in this chapter, we use data on chromatin accessibility, nucleosome positioning and transcription factor binding generated by labs part of the ENCODE project. In addition, data on nucleosome positioning and histone modifications was also taken from MNase-seq and ChIP-seq experiments published by Schones and Barski et al. on CD4+ cells [136, 137].

3.2 Defining hotspot and coldspot motifs

For our analysis in this chapter, we identified 13-bp motifs throughout the genome (Build 36, hg18) by searching for exact matches to the degenerate motif (CCNCCNTNNCCNC) [68]. A total of 270,000 motifs were found of which 29,410 motifs overlapped LD based crossover hotspots (n=34,137), whereas 246,393 motifs did not overlap hotspots. A motif is annotated to overlap a hotspot, when the full length of the 13-bp word is contained within hotspot boundaries. Conversely motifs that are located outside hotspot boundaries are annotated as coldspots.

In addition to this *full set* of hotspot and coldspot motifs, we also created a *curated* (or *filtered*) set of motifs. This carefully curated set of motifs was generated to ensure that in *hotspot motif* cases, each motif is the cause of the hotspots it is present in, and that *coldspot motifs* are actually cold recombinationally inert control regions [data provided by Simon Myers]. The newly curated set contained 1700 hotspot motif cases and about 23,000 controls. Both case and control motifs were in non-repeat regions and were the only degenerate motifs occurring in a given hotspot. All hotspot motif cases had estimated recombination crossover intensity of at least 5cM/Mb over the 2kb surrounding the motif position and all hotspots cases were equal to or lower than 5kb in width i.e. narrow hotspot containing motifs in their central region (with at least 750bp from the ends for 3kb hotspots, and 1kb from the end for 4kb or 5kb hotspots). Coldspot motif cases had estimated recombination crossover intensity at most 0.2cM/Mb over the 2kb background (i.e. about 6-fold reduced, relative to the genome-wide average recombination rate).

Prior to investigating the association between various chromatin features and 13-mer motifs in hotspots, we used both "full motif set" and the "curated" motif set to look at the overlap of hotspot and coldspot motif cases with promoter sites surrounding transcription start sites (TSS), a basic marker of gene regulation. This also helps us to gain an initial sense about the relationship between genes and recombination. The locations for transcription start sites were taken from the UCSC genome browser, and promoters were defined as regions within 2kb of



	All Motifs (N=275,801)	Curated Motifs (N=24,582)
Overlap of hotspot motifs with promoters	4411 (14%)	152 (9%)
Overlap of coldspot motifs with promoters	51742 (21%)	5635 (24%)

Figure 3.1: ***Fraction of hotspot and coldspot motifs overlapping promoters.*** *Overlap of full set of hotspot and coldspot motifs (before curation) with promoters, compared to overlap of curated motifs with promoters.*

a transcription start site. Figure 3.1 shows the fraction of motifs overlapping or not overlapping with promoters (motifs are annotated as overlapping a promoter if they fall within 2kb from the start of a transcription start site), taking the "full set" of hotspot and coldspot motifs, as well as motifs from the "filtered" data. The fraction of overlapping and non overlapping motifs is similar in the case of the full set of hotspot and coldspot motifs, but in the curated hotspots there is a far smaller proportion of motifs overlapping promoters, whereas a higher fraction of coldspot cases overlaps promoters.

These results strongly validate our filtering as helping to remove those motifs that overlap hotspots, specially broader hotspots, purely by chance. It illustrates the benefit of curation and shows a strong enrichment of coldspot motifs in promoters. These results indicate that although motifs tend to be enriched in regions

near transcription start sites, they seem not to cause hotspots there. This raises a question, addressed later in chapter 4, of whether such motifs are actually bound by PRDM9. We next proceeded to exploring the association between the canonical motifs and various chromatin marks, using the curated set of hotspot and coldspot motifs (which will be used for all subsequent analysis in this chapter, unless stated otherwise).

3.3 Chromatin accessibility surrounding 13-mer motifs

We first asked if chromatin accessibility plays a role in differentiating motifs that are able to form hotspots from those that do not. Chromatin accessible sites are sensitive to cleavage by the nuclease enzymes like DNase1, and are therefore also called DNAse hypersensitive sites. A hypersensitive site is a region of chromatin accessibility (implying an active cis-regulatory sequence) where the nucleosome structure may not be organised. Hence, using data on DNase hypersensitive sites would help us to uncover the relationship between crossover hotspot motifs and open chromatin regions [138, 139].

To investigate this, we used chromatin accessibility data produced by Boyle et al. as part of the ENCODE project which was generated using Lymphoblastoid cell lines. These cells were grown in accordance with ENCODE cell culture protocols and were digested by DNase I. DNase cut fragments were isolated and fragment ends were sequenced generating 27bp reads using the Solexa platform. Uniquely mappable reads of high quality were mapped to the genome (hg18), with DNase I hypersensitive site signal being reflected by raw tag intensity [140]. As mentioned previously, ENCODE provides data on multiple cell lines. We examined the distribution of chromatin accessible sites surrounding our case and control motifs in all available cell lines, and noted that the pattern of this distribution remained similar across cell lines. This may be expected, given previous reports demonstrating similarities in chromatin structure between various mitotic cell lines as well as between mitotic and meiotic landscape [84, 141, 113, 142]. Our choice

to use the lymphoblastoid cell lines was based on the fact that the DNase-seq data was available on six similar Lymphoblastoid cell lines (gm12878, gm12891, gm12892, gm19238, gm192878, gm19239) generated from the same lab, allowing us to sum across these cell lines, thereby increasing power and resolution to detect signals of association.

DNase hypersensitivity signals also tend to give us an idea as to the probable positioning of nucleosomes. If we expect to see a hint of nucleosome signal, we would see regions of low chromatin accessibility spread around 160-200 bp implying the presence of a nucleosome [143, 144, 145]. To understand if the DNase-seq data has the power to establish this relationship, and also for the purpose of validating this data, we looked at a transcription factor, CTCF, for which the pattern of chromatin and nucleosomes is established [140]. We examined the DNase-seq signals around the CTCF binding motif. It is established that CTCF, an insulator protein, is able to position an array of nucleosomes around it [137]. Using DNase data summed across all 6 lymphoblastoid cell lines, and taking CTCF motifs (genome-wide positions taken from extracting the closest match to the 20bp CTCF binding motif, from build 36, hg18), strand specific plots were generated. Figure 3.2 shows that at long range distance there is an array of small peaks and dips around the motif centre. The dips appear to be $\sim$ 200bp in width indicating chromatin inaccessible regions and implying nucleosomes positioned upstream and downstream of the CTCF motif centre. These results help to validate the DNase hypersensitivity data and also show that it has the resolution to determine nucleosome positioning.

3.3.1 Chromatin accessibility is enriched around recombination hotspot motifs

After establishing the principle that our approach works, we next examined the association between DNase hypersensitivity and our canonical motif. To get the exact cutting positions by DNase I enzyme, read positions were used to assign cut position and strand orientation. We looked at the distribution of DNase cuts around the centre of 13-mer motifs after correcting for strand information. Plots

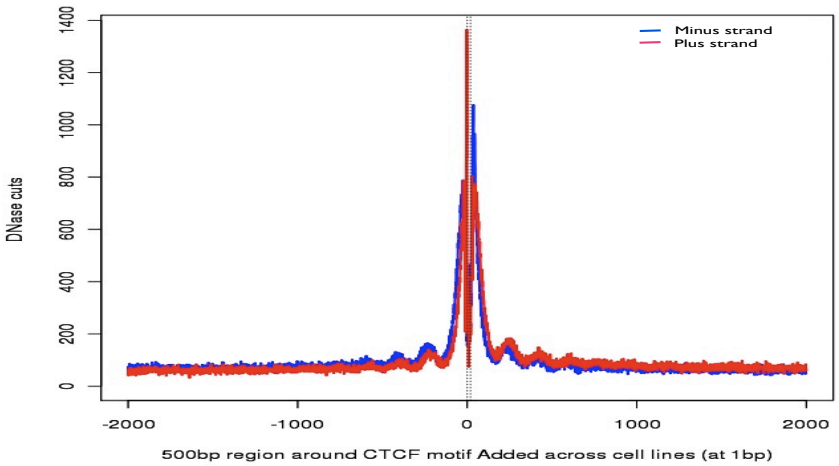


Figure 3.2: ***Chromatin inaccessibility signals hint nucleosome positioning around CTCF motifs.*** *Dip in DNase hypersensitivity around CTCF motifs imply chromatin inaccessible regions, which may be owing to nucleosomes positioned in those regions.*

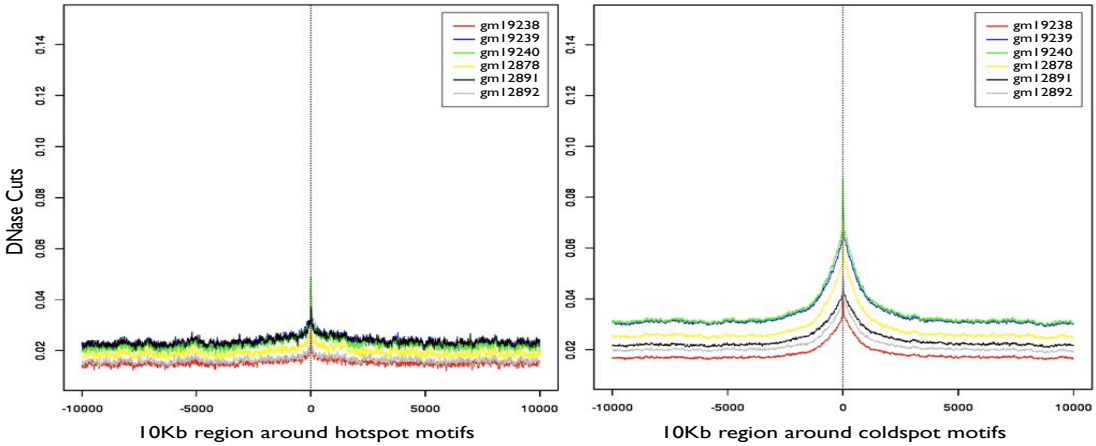


Figure 3.3: ***Chromatin accessibility compared across cell types.*** *Distribution of DNase cuts in 6 lymphoblastoid cell lines around the curated set of hotspot and coldspot motifs (left and right plots).*

were produced by accounting for strand information at both the motifs and in DNase cuts, so e.g. we can look at the plus strand cuts around plus strand motifs. In this section, to look at chromatin accessibility around motifs, we use the full set of hotspot and coldspot motifs, along with the curated set of motifs, to enable comparison and determine if the curated motifs are able to exhibit clearer signals. Looking at each of the six lymphoblastoid cell lines, we noted a strong enrichment of DNase cuts around the curated set of coldspot motif cases, compared to the much lower peak of DNase accessibility around hotspot motif cases (Figure 3.3).

To avoid artifacts due to genomic repeats, we next analysed the canonical 13-bp motifs in the non-repeat genomic background. Figure 3.4 shows the distribution of DNase hypersensitive sites around 13-mer motifs in hotspots and coldspots in non-repeat regions. Plots account for strand, where the red line indicates plus strand i.e. an aggregate of plus strand cuts on plus strand motifs (+m+c) and minus strand cuts on minus strand motifs (-m-c), and the blue line indicates minus strand cuts i.e. the sum of plus strand cuts on minus strand motifs (-m+c) and minus strand cuts on plus strand motifs (+m-c). On the x-axis, the centre 0, marks the middle base (T) of the 13-mer motif (CCNCCNTNNCCNC). Broad scale distribution of DNase hypersensitive sites (on a 10kb and 500bp window) appears to be similar around both hotspot and coldspot motif cases, with a longer range effect extending about 1000bp upstream and downstream from the centre, and a strong spike localised within a few tens of bases of the motif centre. The plus and minus strand distribution appears to agree closely at longer distances from the motif, but there appears to be some local strand specific difference about 200bp around the motif centre. It might be that this difference in strands is attributed to features of the DNase-seq assay, like mapability, sequencing biases on the Solexa platform used, given the motif is GC rich, or the DNase1 enzyme cutting preferentially at certain bases in the genome. In order to account for bias introduced by mapability or nucleotide cutting preference, we normalised the observed DNase cuts around 13-mer motifs by the estimated or predicted cuts around the motifs.

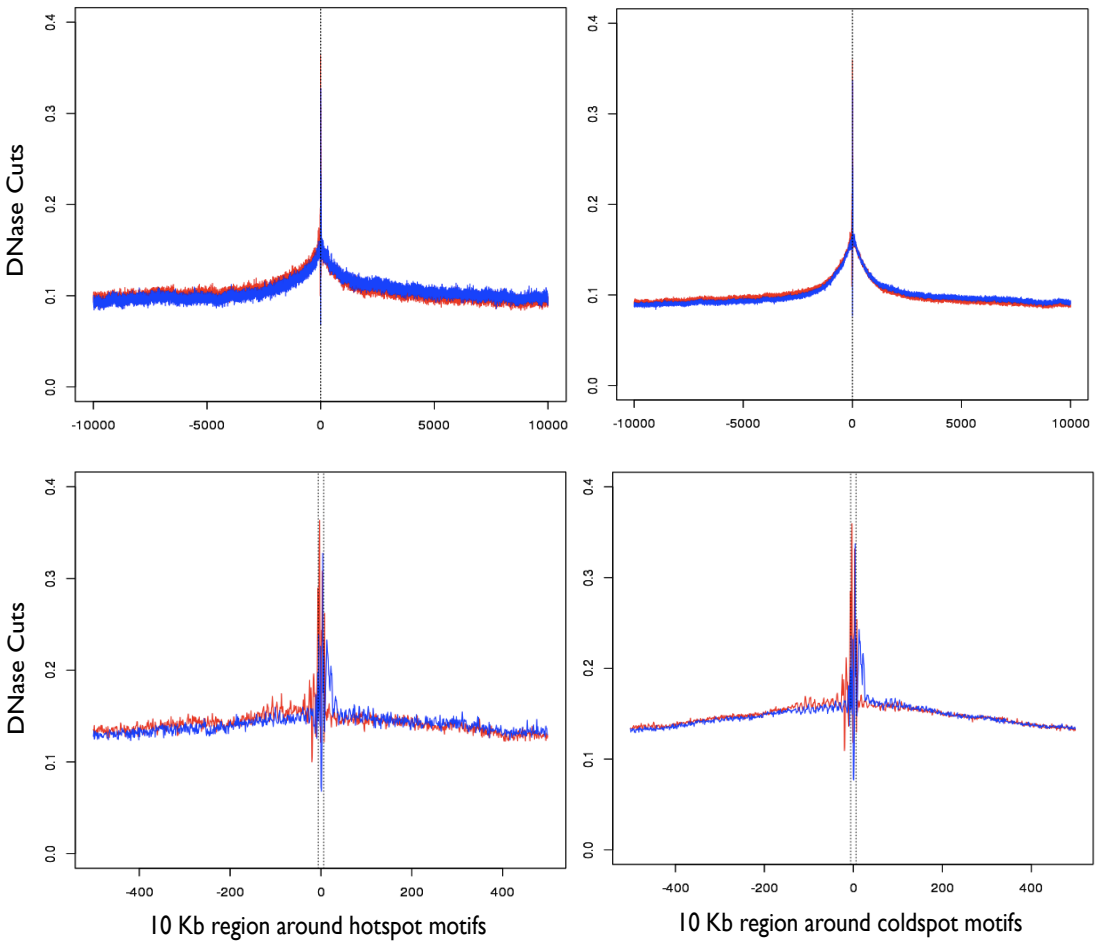


Figure 3.4: ***Chromatin accessibility (summed across 6 cell lines) around recombination hotspot and coldspot motifs.*** *Red line represents + strand (sum of + strand DNase cuts on + strand motifs and - strand cuts on - strand motifs). Blue line represents - strand (sum of +motif -cuts and - motifs +cuts). Top row: Distribution of DNase hypersensitive sites around 13-bp motifs in hotspots and coldspots (10kb window). Long range affects extend beyond 2.5 kb observed in both hotspot and coldspot motif cases, Bottom row: Zoomed in plots (500bp window) show strand specific peaks within and around the centre of the motif.*

3.3.2 Normalizing DNase-seq Data by enzyme cutting preference and mapability

We aimed to normalize the DNase data by a normalising factor i.e. predicted cuts based on mappable bases in the genome and DNase cut preference. To calculate predicted cuts, we computed how often dinucleotides are cut by the DNase 1 enzyme. First, the ***Total number of mappable DNase cuts*** at each dinucleotide (e.g. cuts at AA, GG etc.) were counted, for both plus and minus strands, through out the whole genome. Similarly, the ***Total number of cuts at the Non-Repeat mappable bases*** were counted, giving e.g. for each strand N_{AA+} or N_{GG+} etc. (i.e. 16 dinucleotide cut counts on the + strand and 16 dinucleotide cut counts on the - strand). For counting dinucleotides corresponding to + strand cuts: the base at the read start position and one base upstream of that position were taken, and the nucleotides at those two positions were noted as the dinucleotide corresponding to a given cut; and for - strand cuts, the read start position and a base downstream was taken. (For example: a DNase cut on plus strand at position 237 corresponding to AA dinucleotide, means that there is a base A at position 237 and a base A at position 236. Similarly an AA dinucleotide cut on minus strand at position 81311 means that there is an A base at position 81311 and at 81312). Hence, in doing so we know exactly where the enzyme made the cut, as the DNase would cut in between the two nucleotide positions. We expect cut preferences to mirror each other on complimentary strands, so e.g. the frequency of + strand TT cuts should be that of - strand AA cuts.

Next, we calculated the ***total number of mappable dinucleotides*** across the whole genome, giving us for each strand T_{AA+} or T_{GG+} etc. With these numbers we computed ***the average number of cuts per mappable base*** for each dinucleotide. By repeating across all dinucleotides, we got 16 estimates of cuttability on plus strand and similarly 16 estimates for the minus strand. (As a check, on comparing all complementary pairs e.g. P_{AA+} and P_{TT-}, the numbers were nearly identical, as expected). The ***average cuttability*** value was then computed as:

$$P_{AA} = [N_{AA+} + N_{TT-}]/[T_{AA+} + T_{TT-}].$$

Dinucleotides	No. Mappable Dinucleotides In genome	No. Mappable Dinucleotide cuts + strand	No. Mappable Dinucleotide cuts - strand	% of dinucleotide cuts + strand	% of dinucleotide cuts - strand
AA	132983091	686650	752347	0.6	0.6
AT	103872329	609538	610746	0.6	0.6
AC	65771734	1037918	471106	1.6	0.8
AG	93270754	1657138	967244	1.8	1.1
CA	94206041	821104	1712005	0.9	1.9
CG	12336173	447539	478452	3.7	3.9
CC	67312338	1335366	1264125	2	1.9
CT	93258867	967354	1655418	1.1	1.8
GA	79308566	505292	1441338	0.7	1.9
GG	67326613	1267465	1332657	1.9	2
GC	54723826	859940	856818	1.6	1.6
GT	65898197	473963	1040202	0.8	1.6
TA	89745630	538506	535674	0.7	0.6
TG	94331752	1720316	822532	1.9	0.9
TC	79245205	1444422	505237	1.9	0.7
TT	133067688	755802	689967	0.6	0.6

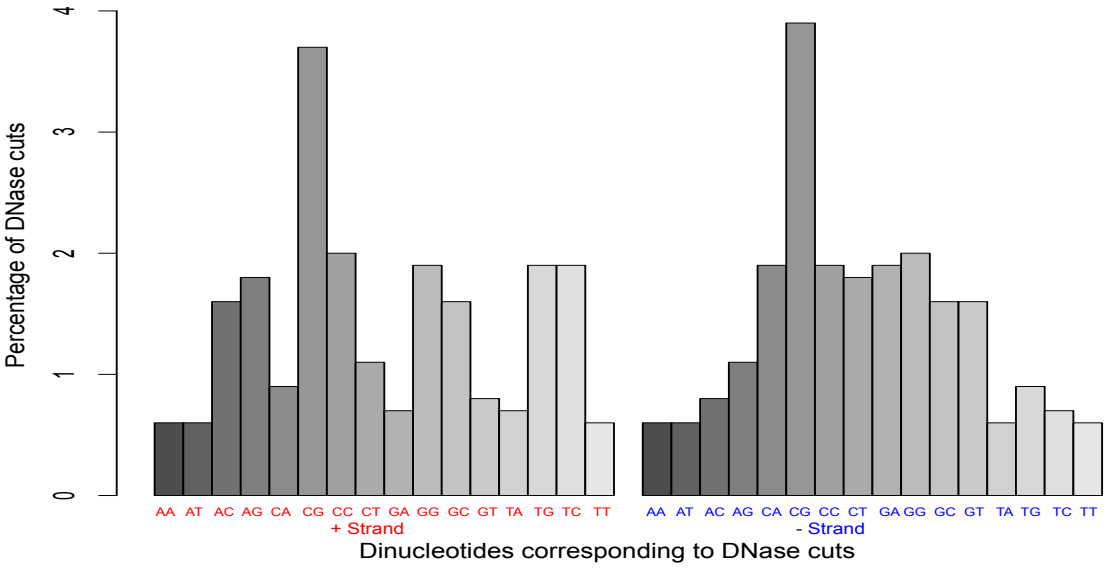


Figure 3.5: ***Percentage of Mappable DNase cuts in the genome.*** *Top: Counts for mappable dinucleotides corresponding to DNase cuts on + and - strand. Bottom: Distribution of DNase cuts corresponding to each dinucleotide combination, illustrating a clear preference of DNase 1 enzyme to cut at CG dinucleotides.*

This estimate of cuttability does the strand averaging separately on top and bottom of the equation, i.e. how often AA is cut on the plus strand, or equivalently its compliment TT is cut on the minus strand (figure 3.5). It was noted that dinucleotides had a big effect on cut frequencies, with e.g. CC bases 4 times more likely to be cut than TT bases, and CG bases up to 6-fold more likely to be cut than AT pairs.

Creating strand-specific plots: Using the above numbers we then adjusted the DNase plots by our normalising factor (expected cuts, based on mappable bases and cuttability estimates). We, therefore, normalised our observed values by how many values we would expect to see across the entire genome. To make the chromatin accessibility plots for each position relative to the 13-mer motifs in the non-repeat region of the genome (for which we also have strand information for each motif occurrence), say, 500bp upstream and downstream, we counted cuts on each strand. Hence, for each position i relative to the motif, we had N_i cuts for a given strand, and this position had a mappable dinucleotide A_i.

We then formed two sums given a set of 13-bp motifs. The first sum was the ***total number of cuts seen at a position on a given strand relative to the motif*** ($\sum_{i=1}^{M} N_i$; where M is the total number of motifs and for a given position the ith motif has N cuts relative to the motif). For each motif, we then defined cuttability C_i, where C_i is equal to the average number of cuts per mappable base seen given a certain dinucleotide A_i, i.e. $C_i = P_{A_i}$ (C_i=0 if this position is not mappable).

The second sum we needed to compute was the ***Total number of cuts that we would expect at a position relative to the motif***. We worked out expected cuts for a given position relative to motifs as: $\sum_{i=1}^{M} C_i$ (*Number of AA mappable bases * P(AA bases cut)+Number of AC mappable bases * P(AC bases cut)...Number of TT mappable bases * P(TT bases cut)*).

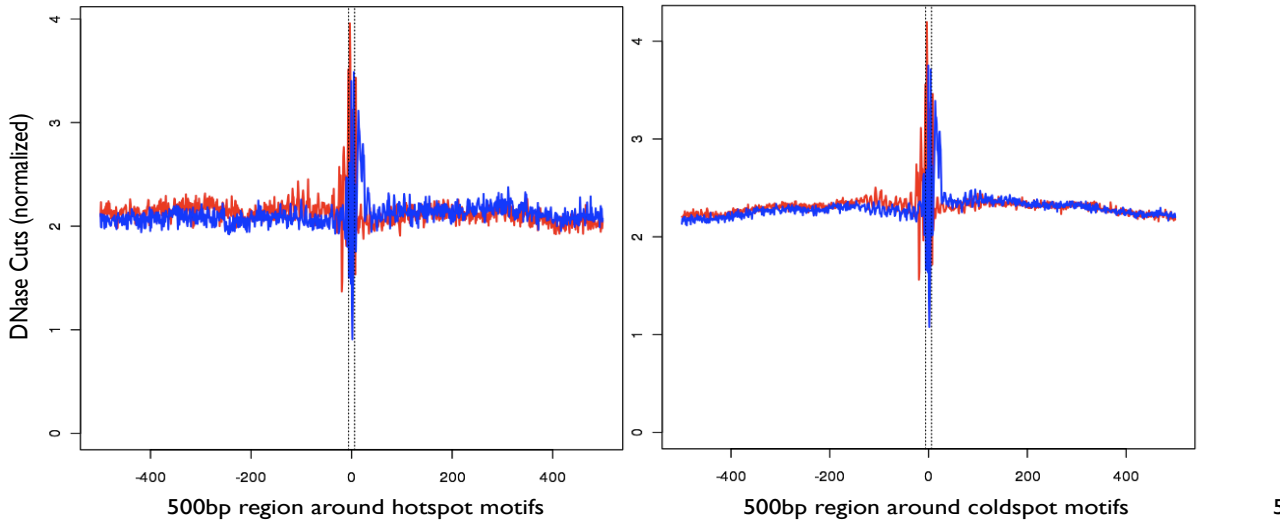


Figure 3.6: ***DNase Hypersensitivity around the 13-mer motif in hotspots and coldspots normalised by enzyme cutting preference and mapability. Red line represents + strand and blue line shows - strand cuts. The decay appears to diminish at a 500bp range after normalising; long-range effects still remain.***

Finally, we normalised the first sum by the second sum ($\sum_{i=1}^{M} N_i / \sum_{i=1}^{M} C_i$), thereby giving us the relative chromatin accessibility at a base. We could now take this as a more robust estimate of accessibility as this averages over both cuttability (or dinucleotide preference) and mapability. The above process was repeated for all motif positions and the final normalised plus strand and minus strand cutting plots were generated (Note that this process does not account for sequencing biases, discussed below).

After getting the observed and predicted cuts, these were summed up across all six lymphoblastoid cell lines to reduce noise, before computing the ratio of hotspots

to coldspots. Figure 3.6 shows DNAse plots normalised by predicted cuts at a 500bp range. On a 500bp scale, it appears that the normalisation slightly adjusts the decay, however, strand differences and long range decay remain in case of both 500bp and 10kb windows (data not shown), respectively. The strand difference is puzzling, and extends only a short distance 25bp around the motif. It might mean that there are more reads containing the 13-mer, or part of it, than expected. This might reflect sequencing preferences, e.g. if 13-mer occurrences are more readily sequenced than other genomic regions, possibly due to their high GC content. As observed, this effect might be reduced by dinucleotide normalisation, but not eliminated, since it effects all of the reads. Thus we are hesitant in ascribing a biological meaning to this difference.

Figure 3.7 shows DNase hypersensitivity in hotspots *relative* to coldspots, normalised by mapability and cuttability and combined over strands. We would expect that at big distances from the motif, the ratio would converge to 1 i.e. the average expectation for the genome. This ratio is at long distances very close to 1, i.e. as expected, and there is a very localised signal (about 2-3 kb, roughly the size of hotspots) for *less* DNase cutting in the hotspot motif cases. This appears to demonstrate a hotspot-specific epigenetic effect in humans, the first such yet seen.

In Fig 3.7b, at fine scales, there is a hint of a very narrow region of open chromatin around the centre of the motif. This chromatin accessible region at the motif seems plausible as being the site where PRDM9 is able to bind. Further, there also appears to be a hint of a nucleosome signal, positioned immediately downstream of the motif ($\sim$ 200 bp), and possibly also upstream of the motif. This observation appears to be true for both the full set of motifs and the curated set, as seen in figure 3.8. However, using the curated set, we note that even at longer ranges, the normalised signal for chromatin accessibility around hotspots is lower relative to coldspots. Similar to the full set of hotspot and coldspot motif cases, this filtered set of motifs also shows a narrow region of chromatin accessibility at the motif, and evidence of a nucleosome dense region upstream and downstream of the motif centre.This suggests that hotspots are situated in much

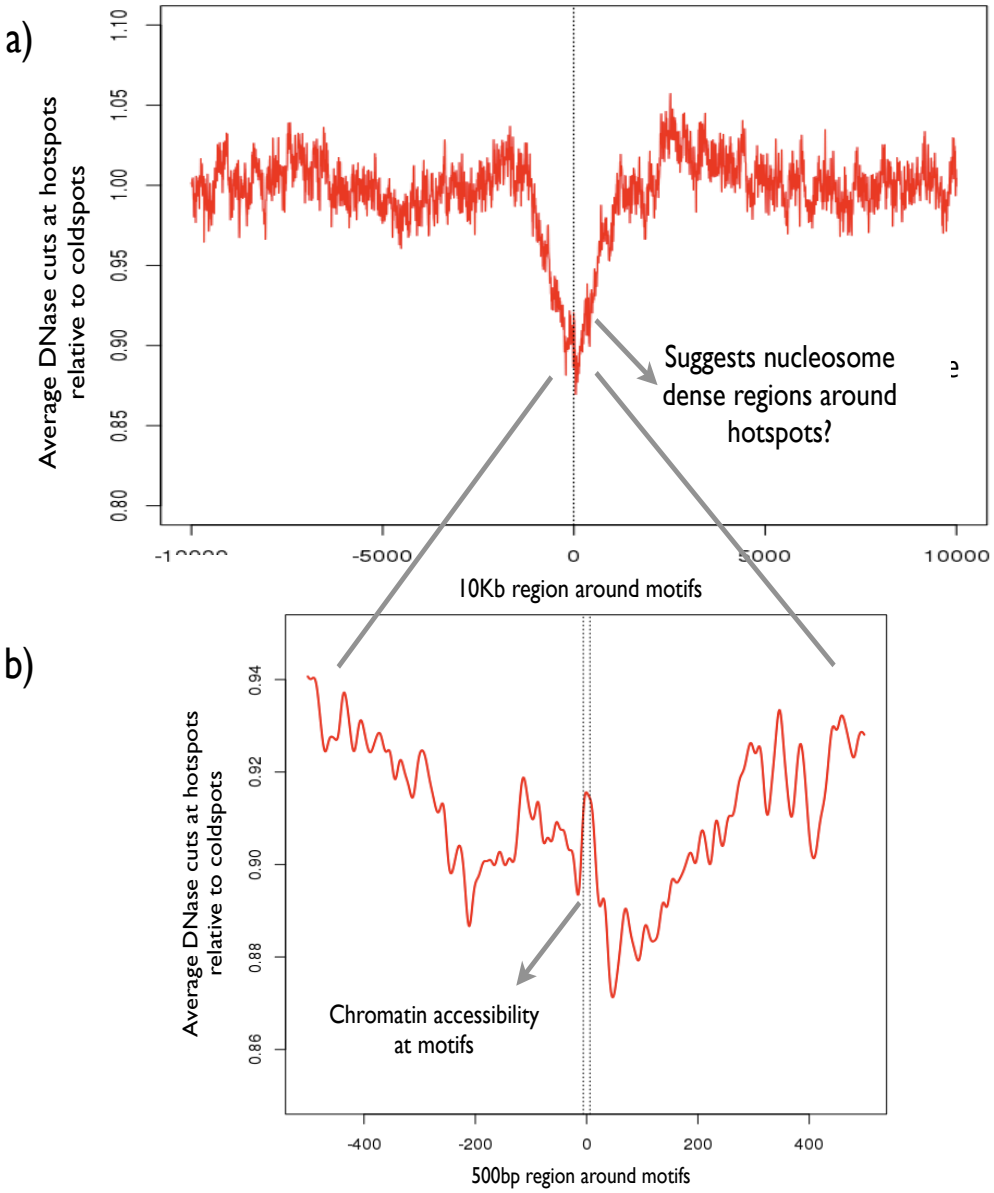


Figure 3.7: ***Chromatin accessibility is depleted in hotspots relative to coldspots*** *Distribution of chromatin accessible sites in: Top) a 10kb region around the 13-mer motifs in hotspots relative to coldspots. Bottom) Zoomed in 500bp region around the 13-mer motifs smoothed at 10bp, in hotspots relative to coldspots. This reveals a narrow region of chromatin accessibility at the motif, and evidence of a nucleosome dense region upstream and downstream of the motif centre.*

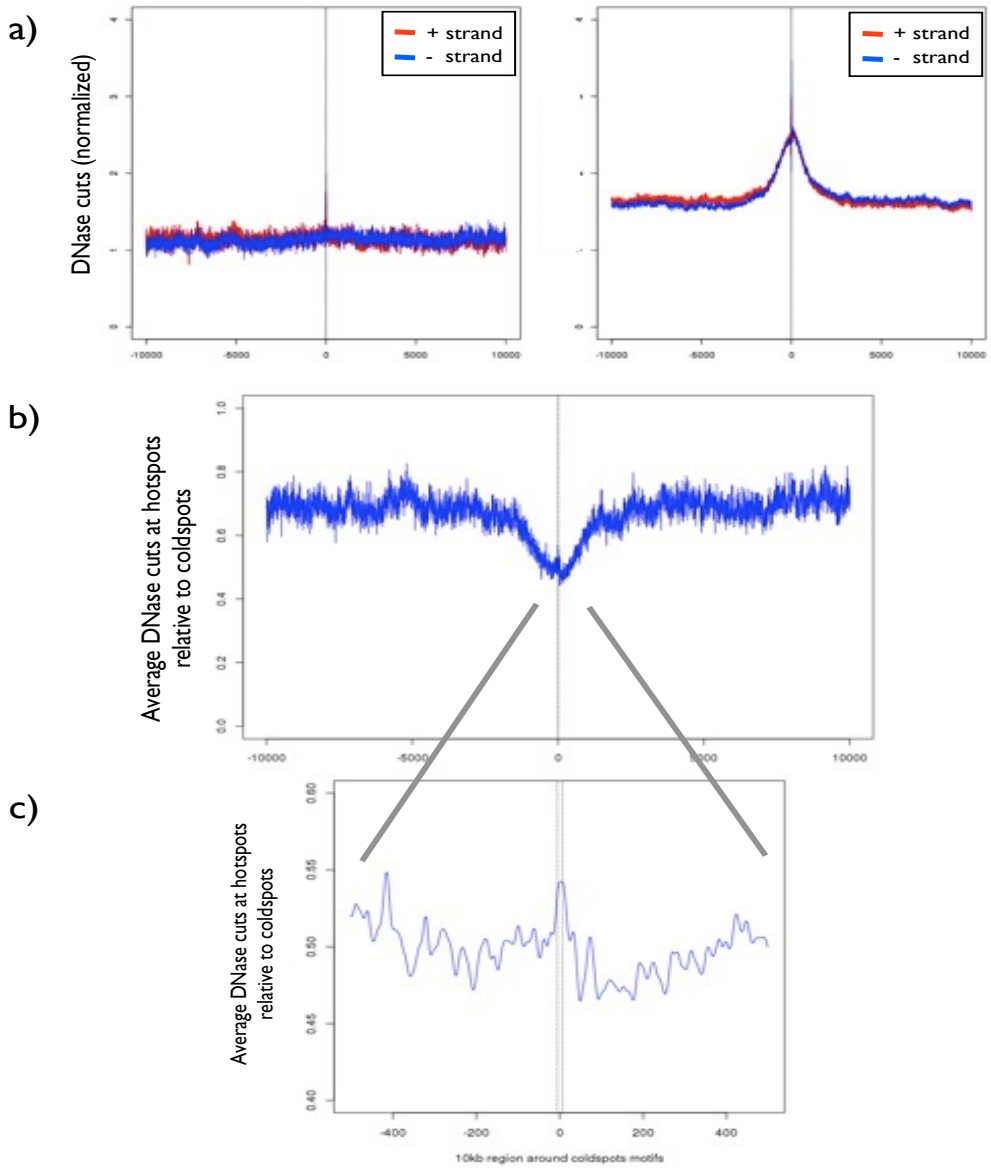


Figure 3.8: ***Chromatin accessibility around the curated set of hotspots relative to coldspots.*** *a) Top left and top right plots show DNase hypersensitivity normalised by mapability and cutability around the curated hotspot and coldspot motifs, respectively, b) Distribution of chromatin accessible sites in a 10kb region around the 13-mer motifs in the curated hotspots relative to coldspots, c) Distribution of chromatin accessible sites in a 500bp region around the 13-mer motifs smoothed at 10bp, in the curated hotspots relative to coldspots.*

more chromatin inaccessible, or nucleosome rich regions relative to coldspots.

3.4 Nucleosomes

We next aimed to further explore the relationship between 13-mer motifs and the positioning of nucleosomes around them. Following the DNase sensitivity results suggesting inaccessible chromatin around hotspot motif sites, we aimed to ask if hotspot motif sites are in fact rich in nucleosomes. We know that nucleosomes play an important role in the regulation of transcription, specifically transcription factor binding and histone modification marks [146]. In mouse work, it has been shown that recombination hotspot motifs are preferentially occupied by nucleosomes [75, 85, 84], and we also know from previous work that the post-translational histone modification, H3K4me3, is implicated in regulating recombination activity [75]. As it is established that PRDM9 confers the histone mark, H3K4me3, it seems likely that it must need nucleosomes to be positioned nearby to add the mark which potentially triggers recombination activity by recruiting the recombination machinery. Hence, we investigated if the DNA around hotspot motifs displays evidence of nucleosome occupancy.

In order to understand the relationship between nucleosomes and recombination hotspots, we first used data generated by Schones et al. [136]. They produced a genome-wide map of nucleosome positions using CD4+ cells, sequencing the ends of nucleosomes with the help of Solexa high-throughput sequencing technique. CD4+ cells were digested by MNase and mononucleosome sized fragments ($\sim$ 200bp of DNA) were isolated and the ends of the DNA sequenced and mapped to the genome (hg18). A scoring function was used to create a nucleosome profile across the genome [136]. As there was no data on nucleosome positioning available from ENCODE at the time, our decision to use this specific dataset was based on the fact that it was the only available map of nucleosome positions in a human cell line mapped to the hg18 reference genome.

We first validated these data by looking at nucleosome positions around CTCF

motifs, following the same principle as applied earlier for DNase hypersensitivity plots. CTCF, as mentioned previously, is already known to exhibit nucleosome positioning capability. It has been reported that an array of nucleosomes are positioned both upstream and downstream of the CTCF motifs in humans [147]. Figure 3.9 clearly illustrates the presence of 10 nucleosomes positioned upstream of the CTCF motif and 10 downstream of the motif, with stronger signals being seen for the 6 peaks present within 1kb of the centre of the CTCF motif. Each peak, about 170bp in width, represents a positioned nucleosome surrounding the CTCF motif, with the centre of the peak being the midpoint of the nucleosome, and + and - strand cuts flanking the nucleosome.

We next used these data to generate plots for nucleosome positioning around the 13-mer motifs in hotspots and coldspots. Strand separated and normalised plots were produced for hotspot and coldspot motifs in non-repeat genomic regions. We were not, however, able to see any clear evidence of a nucleosome positioning signal (Supplementary figure 5).The CTCF motif plot for nucleosome positioning showed that our chosen MNase data did have the power to detect real associations. However, we were not able to establish this association with our canonical motifs. A possible explanation for this could be that nucleosomes are positioned only around 13-mer motifs which are already bound by PRDM9, or it might simply be that these data lack the power to detect associations with our motifs.

To check this, we repeated this analysis using a more recently generated MNase-seq data (mapped to Build 37, hg19) provided by ENCODE. Figure 3.10 shows that nucleosomes are in fact positioned similarly around both hotspot and coldspot motifs. This implies that about 11 positioned nucleosomes exist around all 14-mer motifs, irrespective of hotspot status, with the motif itself typically occurring at a high nucleosome occupancy position. However, regions surrounding the hotspot motif cases appear to be more enriched in nucleosomes compared to coldspot motifs. The hotspot to coldspot ratio plot shows that hotspot motifs tend to exist in nucleosome rich regions, however, a relatively depleted signal exactly at the motif sites implies accessible DNA relative to cold motifs, which also agrees with

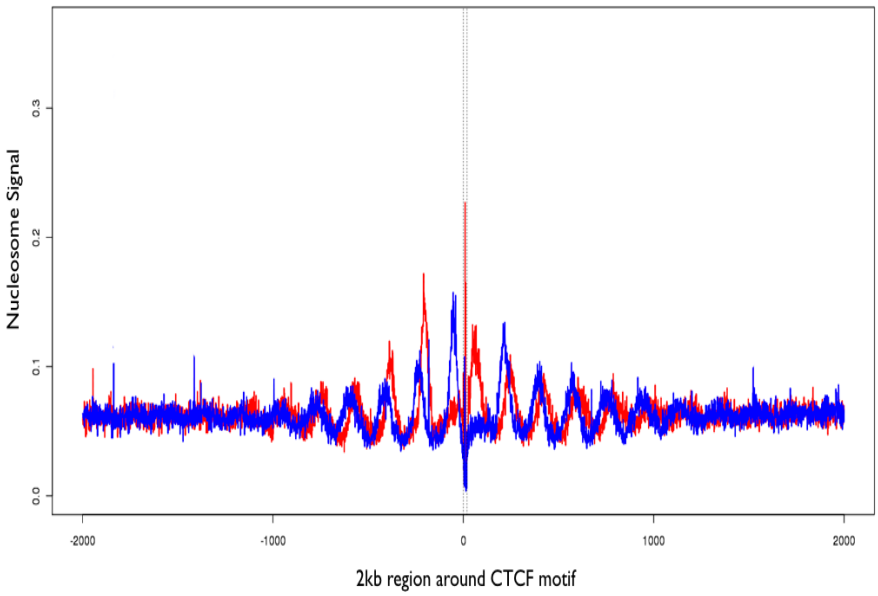


Figure 3.9: ***Nucleosome positioning around CTCF binding motifs.*** *MNase cuts in a 2kb region surrounding CTCF motifs separated by + (red) and -(blue) strands.*

our DNase findings in the previous section. This accessible region may be required for PRDM9 binding. These results also suggest that nucleosomes must be positioned at, or near, the canonical motif site to facilitate PRDM9 binding, and for it to be able to confer the H3K4me3 mark to trigger crossover recombination activity. Our next follow-up question, given that hotspot motifs tend to occur in nucleosome rich regions, was to understand if any histone modifications are enriched around or provide a conducive environment for motifs to form hotspots. Looking at these marks would also help establish if transcription has an effect on crossover activity [148].

3.5 Histone Modification Marks

Histone modification marks are important features of the genome as they have been implicated in dictating gene expression in the genome [149, 150]. We explored the link between recombination and transcription in the light of the 13-bp motif, and determine whether histone modification marks are able to suggest such a link. The data we used for analysing the association between recombination

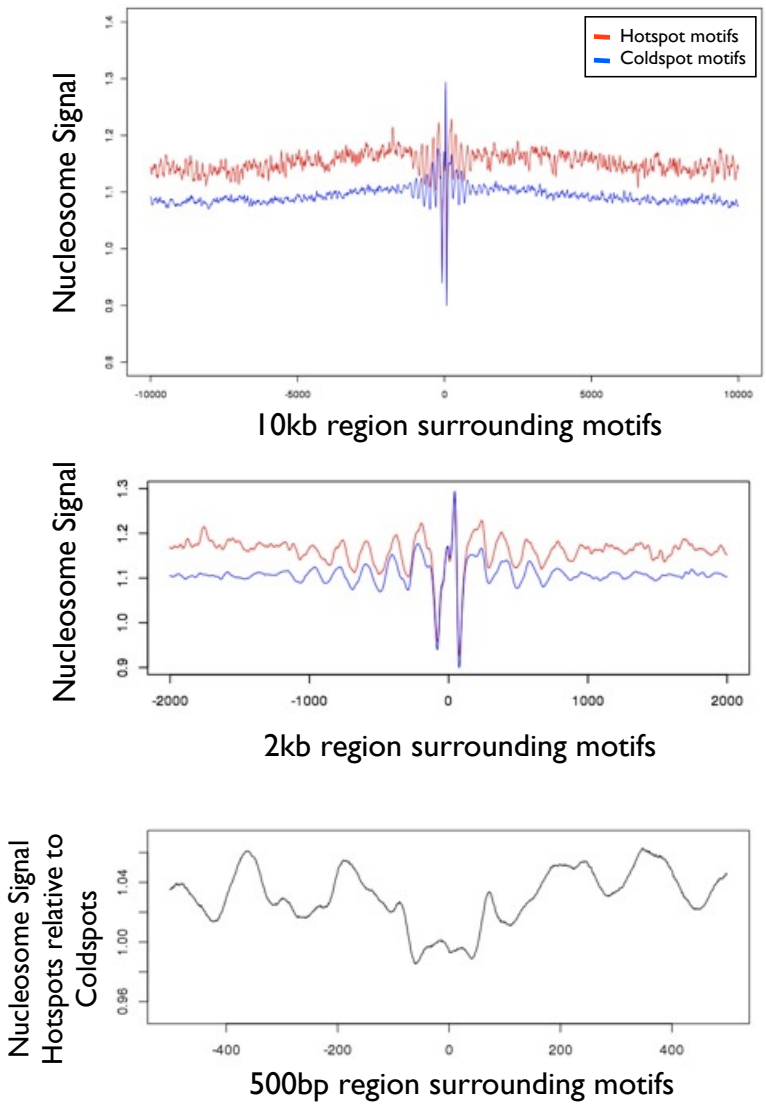


Figure 3.10: ***Nucleosome positioning around 13-mer motifs with updated MNase-seq data.*** *MNase cuts in: top row) a 10kb region around non-repeat, non-promoter, 13-mer motif occurances in hotspots (red line) and coldspots (blue line), middle row) a 500bp region around 13-mer motifs in hotspots (red line) and coldspots (blue line), bottom row) Ratio plot of hotspots relative to coldspots.*

motifs and histone modification marks was generated by Barski et al [137]. They generated high resolution maps for the genome-wide distribution of a number of histone modifications using the Solexa sequencing technology. This data was produced using the CD4+ cells, that were purified from human blood. To map enzyme targets, these cells were cross linked with formaldehyde, and fragmented into 200-300bp fragments by sonication. These fragments were then digested by MNase to get mononucleosomes used to map histone modifications. Samples were analysed using the Solexa 1G Genome Analyzer, with which each run was able to generate over 20 million sequence tags of (typically) 36 bp each, with the number of tags of a nucleosome being proportional to the modification level of that nucleosome [137]. We chose to use the above data over data on histone marks provided by ENCODE, as this was the most comprehensive set of data available on histone marks, generated by experiments performed on one cell line, from the same lab. ENCODE, at the time, did not cover such an extended list of marks on any one cell line.

A total of 25 different histone modification marks were studied. We know that histone marks indicate (on average) specific types of genomic regions. Hence we classified all the marks studied into three groups, being: 1) marks of promoters (e.g. H3K4me1/2/3 etc. [151, 152]) , 2) marks of gene bodies or transcribed regions(e.g. H3K79me1/2 etc. [153]), and 3) marks of heterochromatin (e.g. H3K9me2/3, etc. [154, 155, 156, 155, 157, 118]) [See table 3.1]. We then examined these marks around hotspot and coldspot motifs, to understand the relationship between recombination and these functions, given potential PRDM9 binding sites. Plots for selected marks, representative of each of these groups, are discussed below.

3.5.1 Transcription activating marks are enriched in coldspot motifs and depleted in hotspot motifs

The distribution of histone marks around the hotspot and coldspot motifs was analysed. We first looked at the association of histone marks belonging to each of the three groups described above around the 13-mer motifs i.e. plots were made

to look at the enrichment signals of histone modifications marking promoters, gene bodies or heterochromatin, in a 10kb region surrounding the centre of the 13-bp motifs.

We looked at a total of 16 histone modifications that mark promoter regions in the genome. Figure 3.11 shows the average value of these marks around hotspot and coldspot motifs. On examining the modifications in this group, we noted that all the marks that are involved in activating genes, were enriched in coldspot motifs but not in hotspot motifs. Further, removing promoters, did not remove the pronounced differences between hotspot and coldspot motifs. For example, the localised H3K4me3 peak, is still present in the coldspots outside promoter regions as well, suggesting this mark can generally suppress crossovers. Figure 3.12 shows the ratio plot of hotspots over coldspots for H3K4me3 cases. One interesting mark was H4K16ac, which is suggested to be involved in activating transcription and was more enriched around coldspot motifs, but displayed a very local spike not only at coldspot cases but also hotspot cases.

Next, we looked at histone modifications reported to mark gene bodies (Figure 3.13). On examining a total of 6 such marks, all of which are reported to be involved in transcription activation or present in transcribed regions, we again observed an enrichment of these marks surrounding coldspot motif cases. For example, the H4K20me1 modification marks actively transcribed genes, and is about 2.5-fold enriched in the coldspot cases, as opposed to hotspot cases. Finally, we looked at the 3 histone modifications found in heterochromatin regions (Figure 3.14). Of these modifications, 1 marks constitutive heterochromatin (H3K27me2) and 2 mark facultative heterochromatin (H3K9me2 and H3K9me3), and each of these are present in transcriptionally silent regions. All these marks were more enriched in hotspot cases but depleted in coldspot motifs.

These results suggest that high levels of transcription may be a strong factor contributing to repression of recombination, either through the histone marks themselves or indirectly through the process of transcription. Notably, as transcription is measured outside meiosis, the causality would appear to be in the

		Activating	Repressing
Promoter		H3K4me1, H4K8ac, H3K4ac, H3K9me1 H3K4me2, H4K91ac, H3K36me1, H3K4me3, H3K27ac, H9K18ac, H3K12ac, H4K5ac, H4K16ac, H3K36ac, H3K27me1	H3K27me3
Gene Body		H4K20me1, H3K79me1, H3K79me3, H4K20me3 H3K79me2 H3K36me3	
Hetero-chromatin	**Facultative**		H3K27me2 H3K27me3
	Constitutive		H3K9me2 H3K9me3

Table 3.1: Histone modifications analysed, grouped by location and suggested function of each mark.

said direction. These results were very interesting as they suggest that hotspot activity is going to be reduced throughout both actively transcribed genes in a large fraction of the genome, and other regions with similar histone modifications. The results are consistent with, and extend on a previous report by McVicker and Green which show that crossover rate shows a strong negative correlation with gene expression in meiotic tissues, suggesting that crossover is inhibited by transcription [148].

3.6 Transcription Factors

We explored the role of a number of transcription factors like Rad21, Nfkb, Pol2, CFos, Pol3, Yy1, Cmyc, Tr4, Jund and Zz3. All these TFs were much more enriched in the coldspot motif cases, as opposed to hotspots (which were either not enriched or showed a very modest increase in TF signal). This may just be

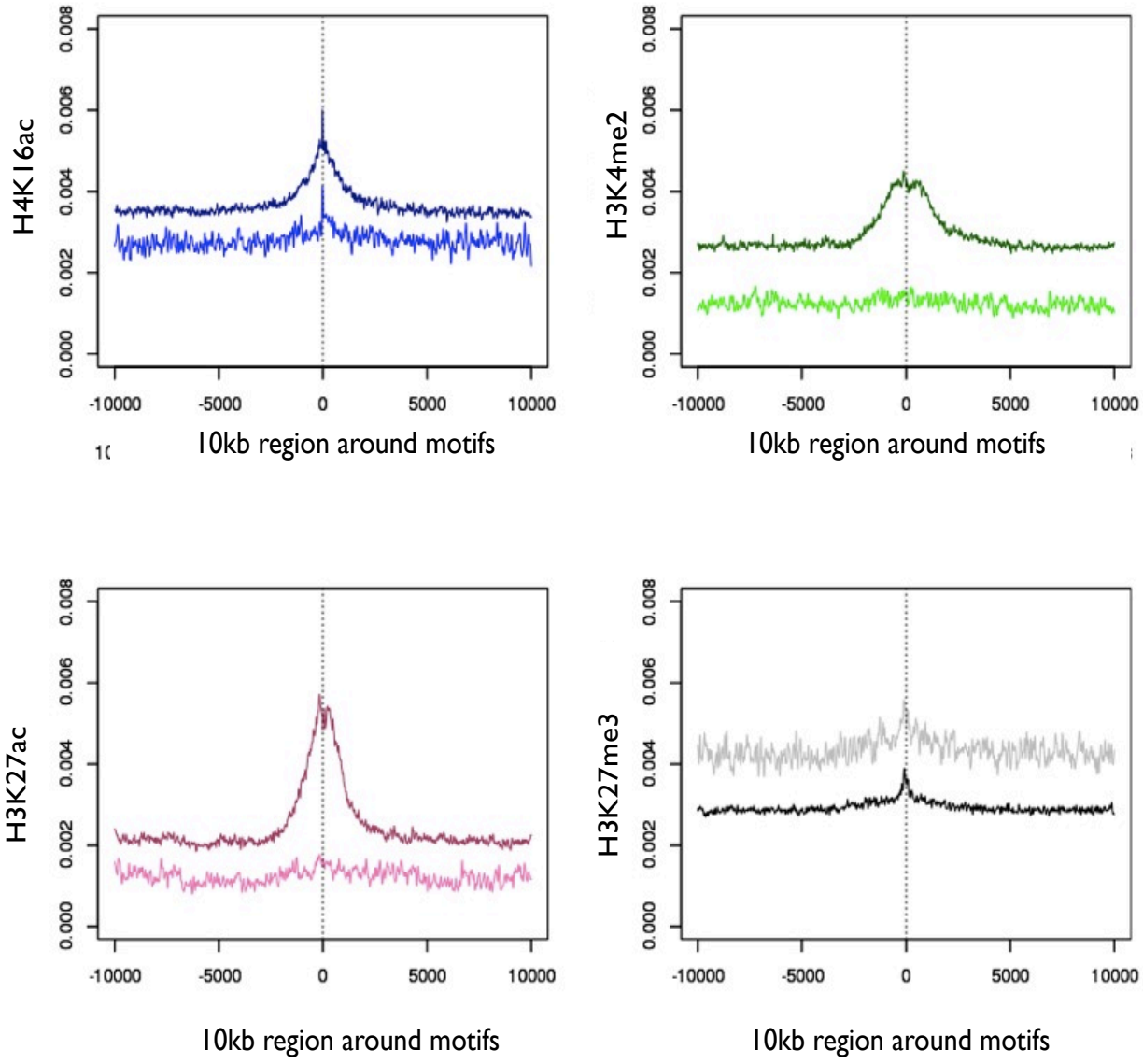


Figure 3.11: ***Histone modifications marking promoter regions surrounding 13-mer motifs.*** *Darker shaded lines (dark blue, dark green, maroon and black) show signals of histone marks around coldspot motifs, and lighter colours (light blue, light green, light pink and grey) show signals of histone marks around hotspot motifs. Histone modifications that activate genes are enriched in coldspot cases, whereas, repressing marks (H3K27me3) appear to be enriched around hotspot cases.*

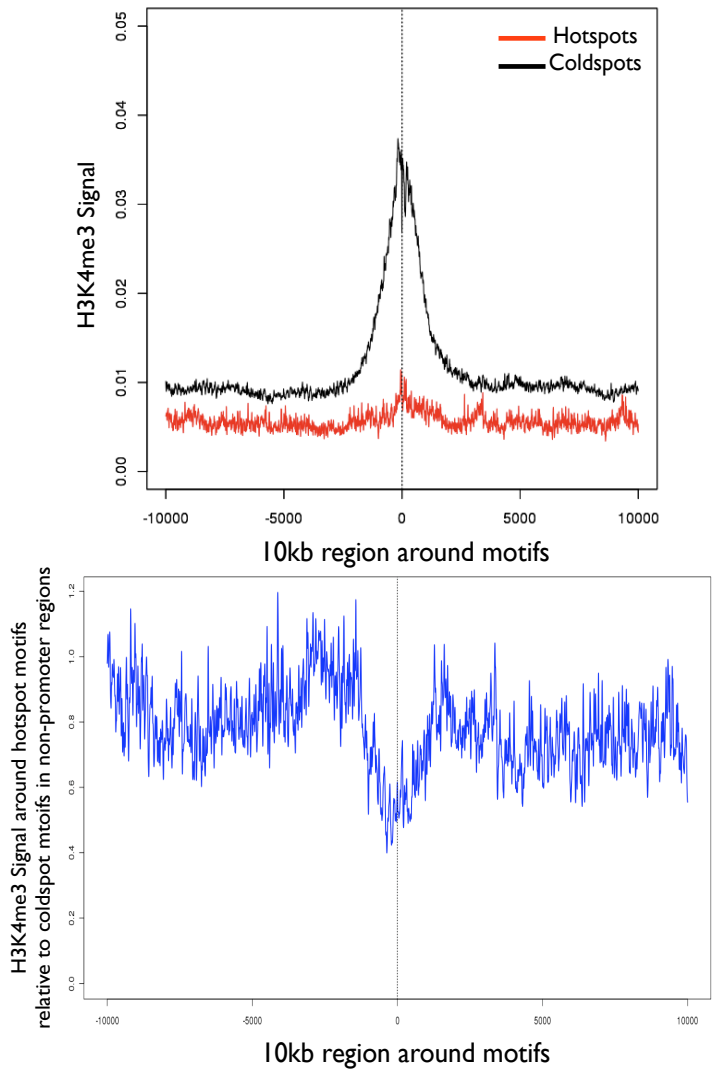


Figure 3.12: ***H3K4me3 mark around 13-mer motifs.*** *Top: H3K4me3 mark surrounding motifs in hotspots (red) and coldspots (black) where there is a clear coldspot specific peak for this mark. Bottom: H3K4me3 mark distributed around hotspots relative to coldspot motifs, after excluding promoter regions.*

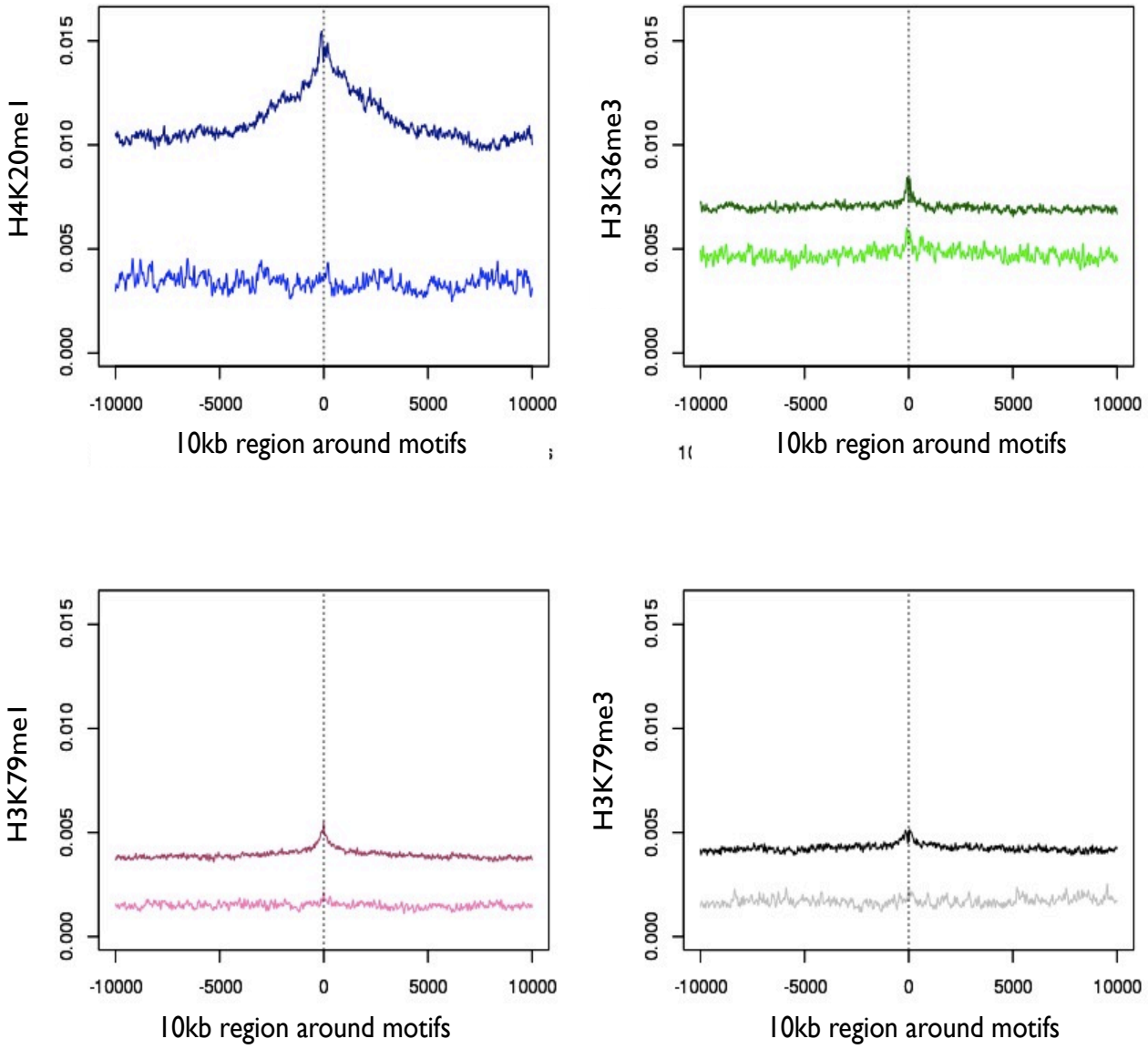


Figure 3.13: ***Histone modifications marking the body of genes depleted around hotspot motifs.*** *Darker shaded lines (dark blue, dark green, maroon and black) show signals of histone marks around coldspot motifs, and lighter colours (light blue, light green, light pink and grey) show signals of histone marks around hotspot motifs.*

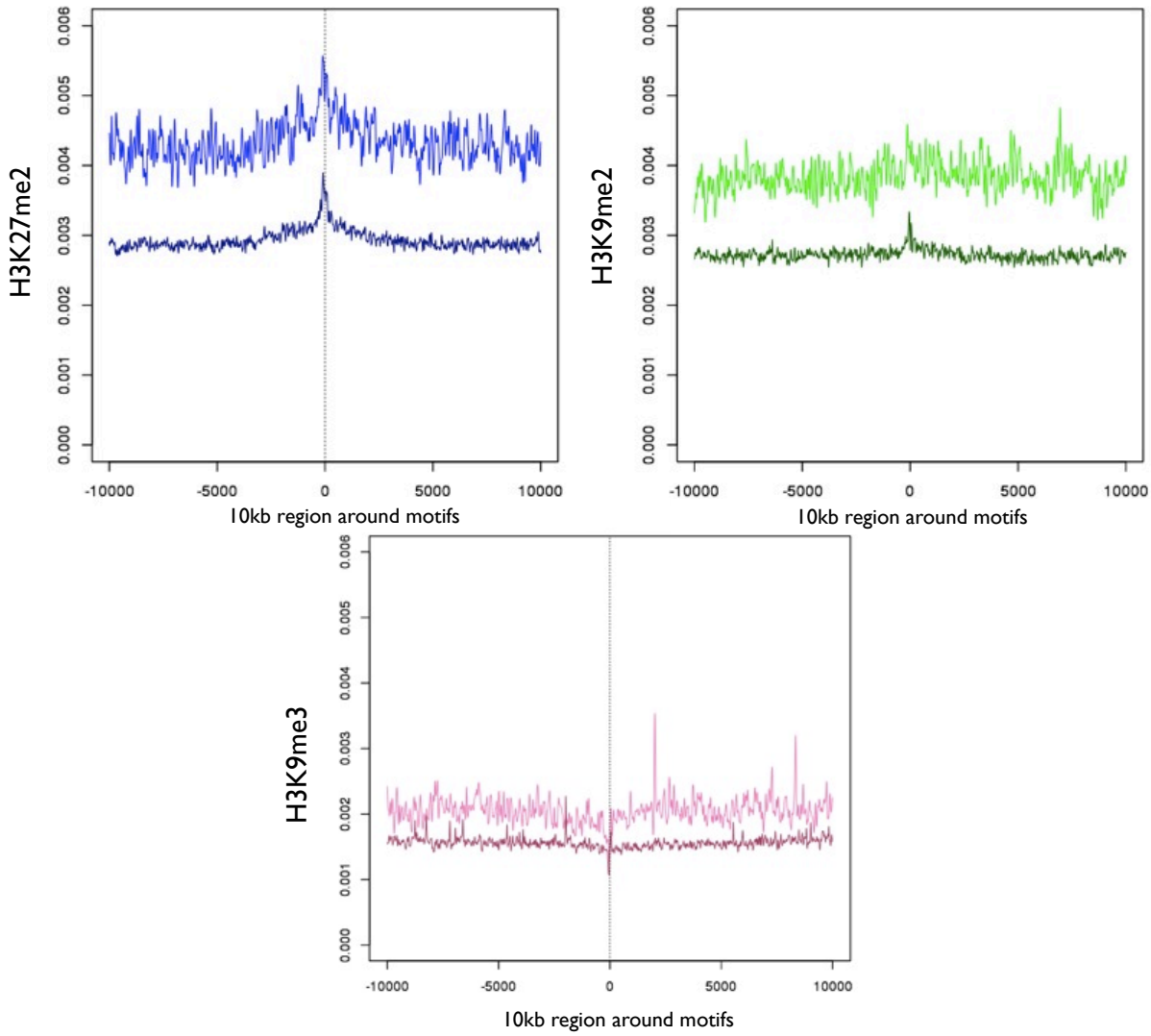


Figure 3.14: ***Histone modifications marking heterochromatin regions enriched around hotspot motifs.*** *Darker shaded lines (dark blue, dark green and maroon) show signals of histone marks around coldspot motifs, and lighter colours (light blue, light green and light pink) show signals of histone marks around hotspot motifs.*

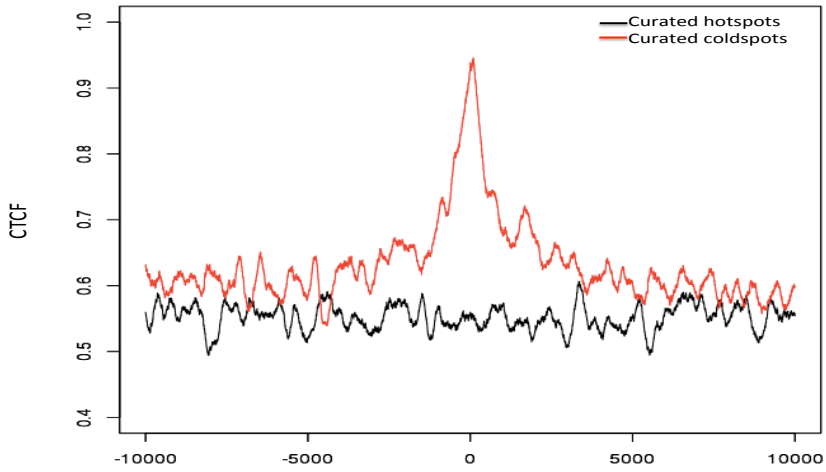


Figure 3.15: ***CTCF signals surrounding the canonical 13-mer motifs.***

because TF peaks in coldspot cases are transcription start sites (Supplementary figures 6 and 7). Overall, we were not able to observe any meaningful biological associations on investigating these factors.

We also explored the level of CTCF signal around 13-mer hotspot and coldspot motifs. CTCF is a DNA binding zinc finger protein which plays a role in organising chromatin structures. Figure 3.15 shows CTCF signals being clearly enriched locally around coldspots, however no such enriched signal is observed around hotspot motifs. We will again revisit this analysis in the next chapter, to determine whether CTCF enrichment around the canonical motif sites has an impact on PRDM9 binding.

3.7 Summary

In conclusion, the motivation of this analysis was to understand which type of the 13-mer motifs are able to stimulate crossover hotspots. We explored various epigenetic marks to inquire if there are certain factors that are able to mark motifs that subsequently become hotspots. Our analysis showed that the motifs are

likely to form hotspots if they are situated in less chromatin accessible regions of the genome. We also observed that hotspot regions containing 13-mer motifs are more nucleosome rich regions, with the motif likely to be situated at or very close to a nucleosome. The positioning of these nucleosomes is likely to be biologically plausible as PRDM9 would need to bind to chromatin wrapped around a nucleosome to be able to confer the H3K4me3 mark, which in turn may trigger the recruitment of recombination machinery. The motifs situated in hotspots are also more likely to be present in transcriptionally repressed regions i.e. heterochromatin, and depleted in transcriptionally active regions. Together, these findings informed us of some of the chromatin features that appear to be necessary for motifs to form hotspots. It was still, however, unclear if all the 13-mer motifs residing in hotspots, are in fact bound by PRDM9, and conversely if all PRDM9 bound motifs are able to form hotspots. This could also enable us to assign a causal direction to the above associations. To further get an understanding of this, we performed a ChIP-seq analysis for PRDM9, which is discussed in the next chapter.

Chapter 4

Mapping PRDM9 Binding Sites in the Genome

4.1 Introduction

We have been able to establish that PRDM9 binds specifically to the 13-mer motif and that motifs which form hotspots are enriched in chromatin inaccessible regions. PRDM9 is suggested to initiate meiotic recombination by interacting with the canonical 13-mer motif. It is suggested to be involved in the recombination pathway by modifying chromatin with the H3K4me3 mark [84]. This modification in turn might stimulate reactions of, or make chromatin accessible to, SPO11, which in turn catalyses double stranded breaks[78]. However, we still do not understand how PRDM9 "selectively binds" to its target sites in the genome to mark hotspot locations. For instance, it may bind at locations with certain unique sequence features, chromatin structure, histone modifications etc. [89]. This question may best be answered by determining binding sites for PRDM9 in the genome. In this chapter, we will discuss the ChIP-seq experiment performed to determine PRDM9 binding in a human cell line.

This experiment was designed to help us answer various key questions about the rules of PRDM9 binding in the genome. We know that PRDM9 has been reported to control all recombination hotspots in humans and mice [79, 20, 75],

however, about 40% of these hotspots contain a match to the 13-mer motif [68], which is also the predicted binding motif for PRDM9 [53, 77]. How then, does it control other hotspots which do not contain the motif? This ChIP-seq assay would help us to test certain hypotheses about possible mechanisms involved. For instance, if a hotspot does not contain an exact match to the canonical motif for PRDM9 to bind, does it then bind to more degenerate versions of the motif, as suggested by our in-vitro experiments? If PRDM9 does need to interact with the 13-mer motifs, then what factors determine which motifs PRDM9 can occupy to initiate recombination? For example, does it need to bind motifs with a specific chromatin conformation, or in nucleosome rich regions, as suggested by our previous analysis, or does binding depend on other cis-related features like primary DNA sequence, or more [135]? On the other hand, if PRDM9 does not necessarily need to bind with the canonical motif to initiate recombination, then could it be that it does not directly bind the motifs and is rather aided by other factors which help in its indirect binding? Further, do all, or only a fraction of the sites that are bound by PRDM9 also translate into hotspots? To answer these questions, we performed the ChIP-seq experiment.

Identifying genome-wide binding sites for PRDM9 by ChIP-seq is challenging to pursue under ideal conditions, as this protein is expressed only in meiotic tissue [87, 91]; in order to perform this investigation, we would need to work with human testis tissue, which is not easily achievable. We therefore designed an alternate approach which involved expressing PRDM9 in non-meiotic HEK293T (Human Embryonic Kidney) cells. HEK293T cells are widely used for expressing recombinant proteins owing to their ability to perform post-translational folding thereby generating functional proteins. This cell line was chosen owing to its ease of maintenance, transfection, efficiency of protein production and translation of proteins [158]. Together, these attributes made HEK293T cells an attractive system to express recombinant PRDM9. Although taking this approach was not ideal, under the existing constraints, this was carried out with the intent that it would be able to provide a good evidence for PRDM9 binding sites, given that HEK293T cells would presumably provide a genomic landscape similar to that of meiotic cells [84, 113, 142]. The key idea was to use this system to try

and understand the binding behaviour of PRDM9 by performing a genome-wide analysis and to understand the strength of overlap between PRDM9 sites and recombination hotspots.

ChIP-seq, an experimental technique applied to understand protein-DNA interactions [159], mainly consists of five steps. These steps are discussed in more detail in subsequent sections, however, here I give a brief description of the protocol employed. Firstly, we expressed the recombinant GFP-tagged PRDM9 protein, in a human HEK293T cell line [158]. After expression, proteins bound to DNA were fixed covalently by treating the cells with formaldehyde; this lead to the cross-linking of proteins and DNA. Once the cells had been cross-linked, the chromatin was then fragmented into small fragments, about 150-250bp in size, by sonication. Sonication conditions were optimized to produce a reproducible size of sheared DNA fragments which is important for library preparation prior to sequencing [160]. Following fragmentation of DNA, immunoprecipitation (IP) was carried out [161], which was done using a specific primary antibody against the tag of our recombinant PRDM9. The choice of antibody is crucial to the success of any ChIP-seq experiment, as the strength and specificity of binding of this antibody to cross-linked proteins will determine subsequent enrichment of protein bound chromatin [162]. Finally, the cross-links were reversed and DNA fragments (which were bound to PRDM9) were purified. There was still some non-specific DNA pulled down in the IP step owing to random cross-links between protein and DNA or some non-specific binding of the primary antibody to other proteins, which was accounted for by comparing the ChIP sample with both untransfected and transfected genomic control samples (See figure 4.1).

The enrichment of IP DNA was confirmed by performing real-time PCR, which was used to measure fold enrichment of certain target DNA sequences in IP samples versus enrichment in non-IP control samples [159]. Also, to confirm if the immunoprecipitation had worked, a western blot was performed on IP DNA samples using primary antibody against our tagged-PRDM9 [163]; the correct sized band indicating that the IP sample is enriched for PRDM9 bound chromatin. After confirmation of enrichment, the samples were sent off for library prepara-

tion and sequencing. The sequenced fragments or tags were finally mapped to the human reference genome (hg19) and with the help of peak calling softwares, regions with enriched tag counts were identified. In parallel, we also performed an H3K4me3 and H3K4me2 ChIP to try and measure the activity of PRDM9 PR/SET domain.

4.2 Methods

4.2.1 PRDM9 expression in Mammalian HEK293T Cells

N- and C-terminal GFP tagged and N- and C-terminal His tagged constructs were prepared by cloning into mammalian expression vector pHLsec. The pHLsec plasmid was used as it is very efficient in DNA production, owing to a strong combination of cytomegalovirus enhancer and chick beta-actin promoter. This ampicillin resistant plasmid is also efficient because of its small size, which allows cloning of constructs with varying lengths [164]. PRDM9 cDNA (B allele; Supplementary figure 1) was cloned into pHLsec vector as follows: Company (GenWay Bio) provided cDNA was amplified with primers carrying restriction sites. The size of this amplified product was approximately 2700bp. PCR product and plasmid were digested with restriction enzymes according to the cloning plan as given in (supplementary figure 8). Following a 1 hour digestion at 37^0, gel purified bands of insert and vector were ligated for 10 minutes with the Quick ligation Kit (New England BioLabs). Transformation was done using cd5 alpha electrocompetent E.coli cells, by incubating cells with ligation mix for 30 minutes on ice, followed by heat shock at 42^0 for 45 sec, and then resting on ice for 2 mins. The cells then had a 1 hr recovery step in s.o.c medium in a 37^0 shaker/incubator, and then finally plated on Ampicillin agar plates overnight. For colony isolation, colonies were selected and grown overnight with LB and Ampicillin in 37^0 shaker.

For plasmid QC, DNA from overnight culture was purified using QIAprep spin kit from QIAGEN. Miniprep purified plasmid was then tested with restriction digestion analysis to confirm that the construct was carrying the insert (i.e frag-

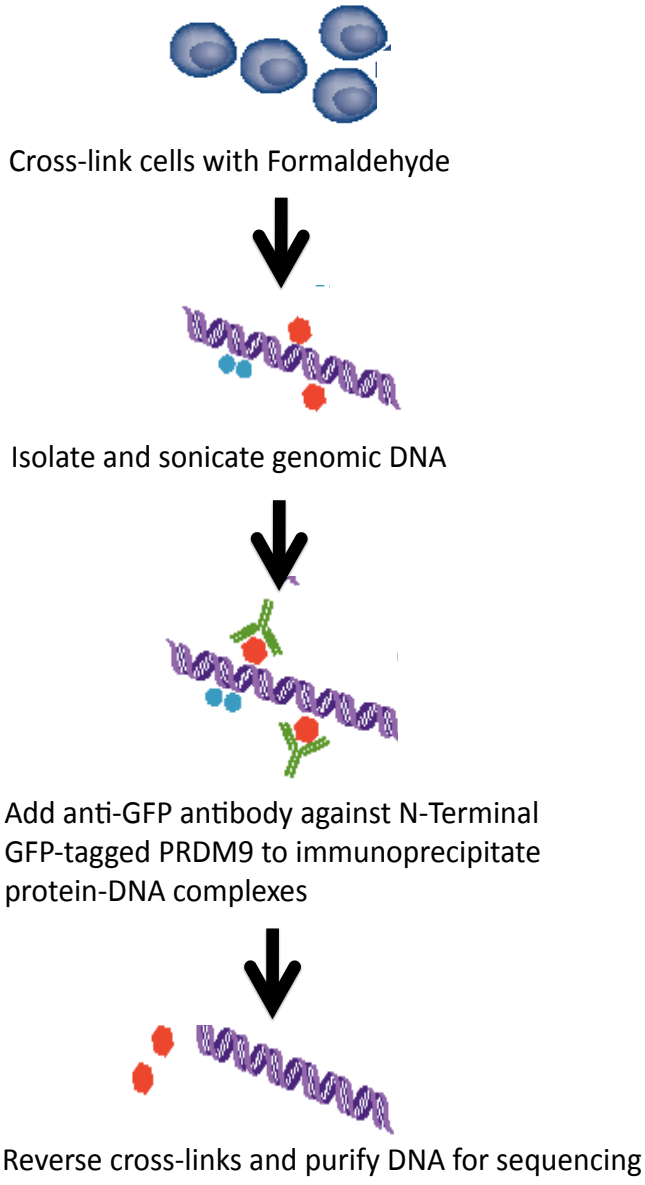


Figure 4.1: ***Chip-Seq Experiment*** *Illustration of the experiment carried out to immunoprecipitate GFP-tagged PRDM9 expressed in HEK293T cells. Transfected cells were cross-linked with formaldehyde, resulting in the expressed PRDM9 (red circles) and other proteins (blue circles) to be tightly bound with DNA. This cross-linked DNA was then sheared to smaller fragments by sonication. DNA fragments were immunoprecipitated by anti-GFP antibody (green) against the GFP tag on the N terminal end of PRDM9. Immunoprecipitated protein-DNA complexes were then treated for reversing cross-links, leading to purified DNA and finally library preparation and sequencing. [Figure adapted from Mardis et al. and Valouev et al. [160, 162]]*

ment of about 2700bp in size on agarose gel). Constructs were then sent for sequencing. Primers were designed tiled across the whole length of the PRDM9 sequence, spanning about 500bp, to check for accuracy of insert sequence in sequenced constructs.

Finally, after sequence confirmation, about 4ug of purified GFP-tagged PRDM9 construct was used for transient transfection in human HEK293T cells, using Lipofectamine (Life Technologies) as the transfection reagent. PRDM9 transfected cells were visualized under the microscope after 48 hours to observe localisation in the nucleus, and to confirm that the protein has maintained functionality (see Figure 4.2). Further, PRDM9 expression was confirmed by performing western blots on the samples. After confirming that PRDM9 is being expressed in HEK293T cells, we proceeded to large scale transfections. For large scale transfections only N-terminal GFP and N-Terminal His tagged constructs were used (as C-Terminal tags might interfere with the binding properties of PRDM9 zinc fingers). Six bottles each (containing about 600M HEK293T cells per bottle) were transfected with the His and GFP tagged PRDM9 constructs.

4.2.2 Chromatin Immunoprecipitation of PRDM9 Transfected Cells

To optimise the ChIP protocol, firstly, a range of different cross-linking and sonication conditions were tested to determine the most favourable conditions, based on DNA intensity and fragment size distribution. The conditions tested for the formaldehyde cross-linking reaction with glycine included cross-linking with 0.75% formaldehyde for 5, 10 and 15 minutes, followed by testing sonication times of 10, 15 or 20 minutes, on a small fraction of the cells [These steps were performed by Nick Altemose]. After determining the best conditions i.e. cross-linking for 10 mins at 0.75% formaldehyde concentration, and sonicating for 20 minutes, the same conditions were applied to all cells.

The IP reaction was optimized by testing different types of primary antibodies

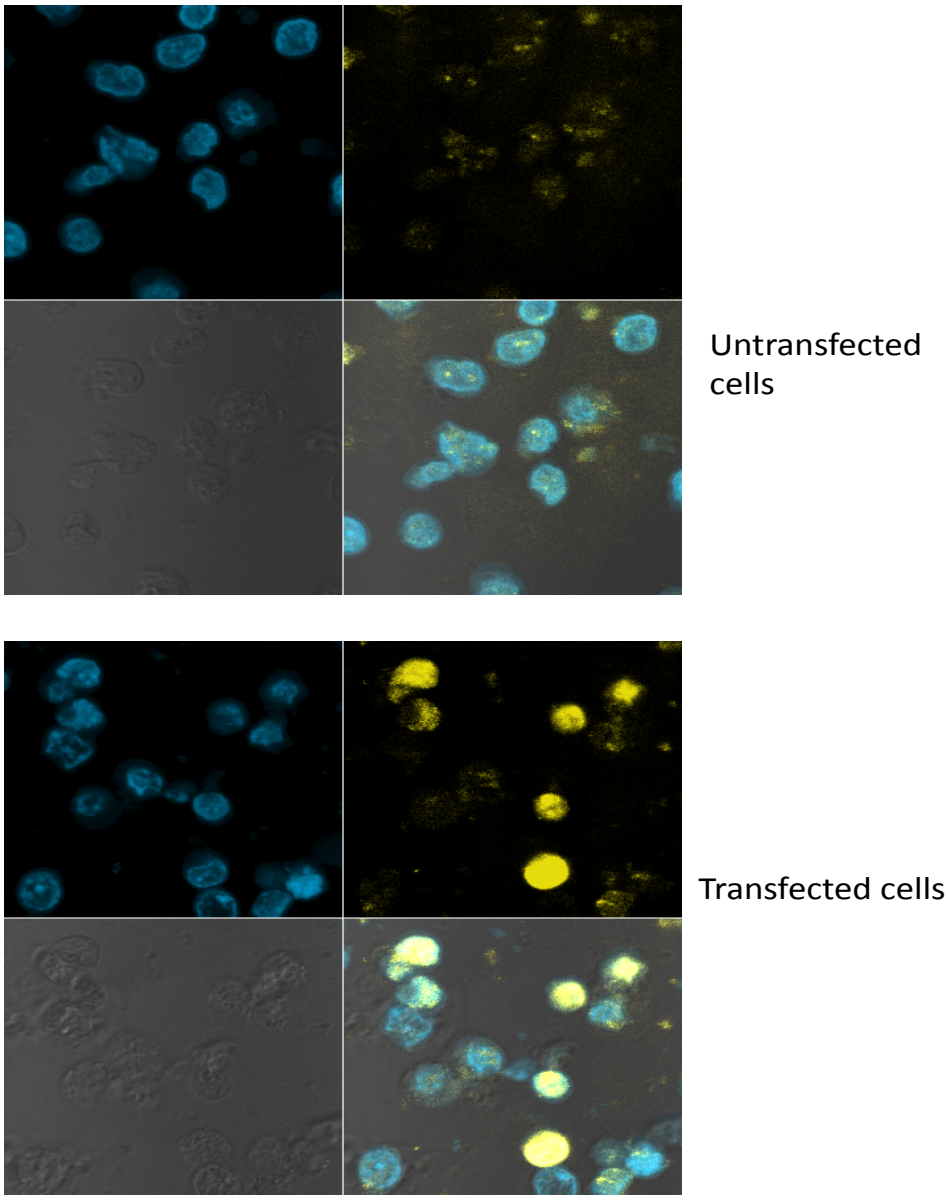


Figure 4.2: ***PRDM9 expression in HEK293T cells.*** *In each of the two (transfected and untransfected cells) images shown, the bottom left square shows HEK293T cells. The top left square shows DNA stained in HEK293T cells; top right square shows GFP tag expressed in these cells and the bottom right square shows a merged image of the three squares. GFP-tagged PRDM9 is clearly expressed in the nucleus of transfected cells.*

against PRDM9 (antibodies tested being: 1) anti-His (Invitrogen), 2) anti-GFP (Invitrogen), 3) anti-GFP (Abcam) and 4) anti-PRDM9 (Abcam)), along with varying concentrations of primary and secondary antibody. To test the performance of each primary antibody, we performed the IP reaction (using ChIP-seq protocol from Rick Myer's lab) and tested the enrichment of PRDM9 bound chromatin by performing western blots. Anti-His and anti-GFP (Invitrogen) were not able to show any positive results on the western blot analysis, suggesting that either these antibodies were not able to efficiently recognise their binding target or the experimental protocol (e.g. wash buffers, incubation times etc.) needed to be further optimized. We then optimised the experimental conditions by performing ChIP for H3K4me3, for which ChIP-grade antibodies are available, and enrichment can easily be quantified by real-time PCR (given standard primers like RPL19 and RA518 [165]). We carefully determined the best conditions for an optimal IP reaction after testing various incubation times to form protein-antibody complex, and wash buffers including the number of washes required to control for non-specific binding. Once real-time PCR showed positive results (between 5-80 fold enrichment) for H3K4me3-IP DNA, we proceeded to use this optimised protocol to perform PRDM9-IP, again using the anti-His and anti-GFP (Invitrogen) antibodies, but were still unable to achieve successful results on western blot. We then tested anti-PRDM9 and a new ChIP-grade anti-GFP (Abcam) antibody, and observed that the western blots performed on IP DNA using the new anti-GFP showed the best results i.e. a clear band detected by both anti-GFP and anti-PRDM9 antibodies against IP samples 4.3. In parallel, we also performed an H3K4me3-IP and an H3K4me2-IP on PRDM9 transfected HEK293T cells.

The optimised IP protocol is described briefly as follows: Firstly, magnetic beads coated with anti-rabbit secondary antibody were washed with PBS/BSA and incubated with primary antibody overnight, to make protein-antibody complex. These antibody-coupled beads were then added to 1ml sheared chromatin preparation (described above) and incubated for 2-3 hours. The immunoprecipitated samples were then washed with wash buffers (5 times with LiCL buffer and once with TE) using magnetic racks. The supernatant from the last wash was discarded

and bead pellet was resuspended in IP elution buffer. Reversal of cross-links was performed by incubating samples at 65^0C overnight to elute immuno-bound chromatin from the beads. Finally, the supernatant containing IP DNA was collected to perform checks with IP-western and qPCR, prior to sending off for sequencing.

Our control samples were non-IP, both transfected and untransfected, DNA. We also attempted to immunoprecipitate untransfected HEK293T cells with anti-GFP antibody, however, this generated DNA at a much lower concentration compared to the transfected samples, hence not enough to be sequenced as controls.

4.2.3 Verification by IP-Western

Western blots were performed on IP samples to confirm if the antibodies that were being tested against PRDM9 are able to pull down the targeted PRDM9 bound DNA complex. Figure 4.3 shows western blot results for two IP samples and controls; IP samples being His-tagged PRDM9-IP, GFP-tagged PRDM9-IP (immunoprecipitated with anti-GFP antibody), and non-IP samples from transfected cells as controls. For the western blot analysis the antibodies anti-His, anti-GFP and anti-PRDM9 were each used against PRDM9-IP samples. There is a distinct band between 120 and 160 KDa in both anti-GFP and anti-PRDM9 lanes for the NV-IP sample (lanes 8 and 13), not present in NH-IP lanes. There are some thick bands at 50KDa in IP samples, which could be the antibody picking up IgG from ChIP DNA. Other non-specific bands in anti-GFP and anti-PRDM9 lanes could either be non-specific binding (from primary antibody incubation overnight) or degradation products. Notably, this result shows that we are able to pull down GFP-tagged PRDM9 with an anti-GFP antibody in our ChIP samples. We had also performed IP with anti-PRDM9, but western blot results showed a very weak signal in this case.

4.2.4 Verification by qPCR

In order to determine that the ChIP experiment worked, qPCR reactions were performed, to help detect and quantify target DNA molecules. This would give us an estimate of fold enrichment of our predicted PRDM9 binding positions. A

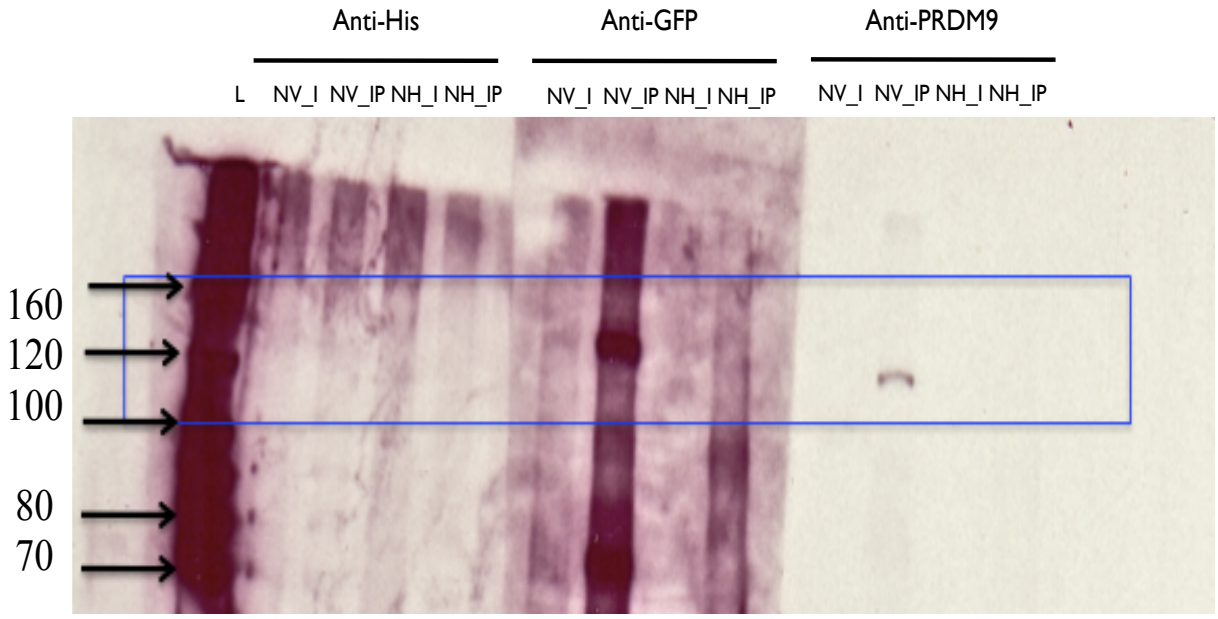


Figure 4.3: ***Validation of the ChIP-seq experiment by IP-western.*** *Four samples are tested in this western blot. NV-I = Non-IP sample from gfp-tagged PRDM9 transfected cells, NV-IP = IP sample from GFP-tagged PRDM9 transfected cells immunoprecipitated with anti-GFP antibody, NH-I = Non-IP sample from His-tagged PRDM9 transfected cells, NH-IP = IP sample from His-tagged PRDM9 transfected cells immunoprecipitated with anti-GFP antibody. Each of these four samples were used to perform WB with anti-His, anti-GFP and anti-PRDM9 antibodies. PRDM9 band is clearly pulled down by both anti-GFP and anti-PRDM9 antibody lanes, confirming that PRDM9 bound chromatin is being pulled down by the ChIP-seq assay.*

total of 24 Positive and 8 negative control primer pairs were constructed for the qPCR. All positive control primers were made from hotspot regions containing the 13-mer motif (16 from non-repeat hotspots regions, 4 from THE1 and 4 from L2 regions). Negative control primers were divided into 4 motif containing and 4 non-motif containing sequences. The negative control motif containing sequences came from coldspot regions with <20% of genome-wide average recombination rate (which could also, in theory, be bound but not active hotspots). From a total of 32 primers, 21 were used for further analyses as these produced single PCR bands with genomic template DNA [Primer designing and qPCR experiments were performed by Nick Altemose].

We aimed to test whether immunoprecipitation by anti-GFP to pull down the GFP-tagged PRDM9 had worked. Of the 15 positive control primers tested, we observed that 6 were more than 2-fold enriched relative to genomic control regions (genomic DNA was diluted to the same concentration as the IP DNA). On normalising relative to the negative controls, it was seen that one positive control region (PR2) was over 5 fold enriched and five other positive control regions (PNR7, PNR10, PNR11, PNR13, and PR7) were over 2-fold enriched over genomic DNA (see Table 4.1). H3K4me3 IP DNA was also tested for enrichment by qPCR using GAPDH, PR719 and RA518 primers and we observed between 5-85 fold enrichment in these regions tested (Table 4.1). In summary, about half of the positive control primers tested exhibited between 2 and 5 fold enrichment across the qPCR experiments performed. Following the above mentioned results from IP western and qPCR analysis, we proceeded with sequencing of IP DNA samples.

4.2.5 Sequencing

DNA samples were prepared for sequencing in ultra pure water, purifying DNA with a mini-prep kit. About 20ng of the sample as measured by qubit was submitted for library preparation, followed by sequencing, which was performed by the core genomics facility at the Wellcome Trust Centre for Human Genetics.

Antibody (IP)	Primers	Type of Positive controls	Fold-Enrichment
PRDM9-IP	PR2	L2 repeat region	5.09
	PR3	L2 repeat region	1.31
	PR8	THE1D repeat region	1.11
	PNR1	Non-repeat region	1.35
	PNR4	Non-repeat region	0.37
	PNR5	Non-repeat region	1.46
	PNR6	Non-repeat region	0.93
	PNR7	Non-repeat region	2.31
	PNR8	Non-repeat region	1.03
	PNR9	Non-repeat region	0.88
	PNR10	Non-repeat region	3.11
	PNR11	Non-repeat region	2.02
	PNR13	Non-repeat region	2.64
	PR4	L2 repeat region	1.28
	PR7	THE1B repeat region	2.34
PRDM9-IP	GAPDH		5.49
	RPL19		3.25
	RA518		0.45
H3K4me3-IP	GAPDH		22.66
	RPL19		85.06
	RA518		5.33

Table 4.1: Validation of ChIP-seq experiment by qPCR. *About half of the positive control regions tested in PRDM9-IP samples exhibit between 2 and 5-fold enrichment. PRDM9-IP samples show an enrichment for H3K4me3 mark in two of the three regions tested, which are known to be positive for the trimethylation mark. H3K4me3-IP samples show between 5 and 85 fold enrichment for the mark in three positive control regions tested.*

Sequencing was carried out on four samples: 2 ChIP cases of PRDM9-IP DNA (to serve as biological replicates), one transfected genomic control sample (non-IP) and one H3K4me3-IP sample. Three lanes of sequencing were used to perform a 4-way multiplex on the DNA samples, to generate 51 nucleotide sequence of paired-end reads on the Solexa sequencing platform. Each of these samples, on sequencing yielded about 180 million reads.

4.3 Results

4.3.1 ChIP-seq data analysis shows enrichment in predicted PRDM9 binding sites

For each of the sequenced samples Bam files were used for initial data visualisation using the Integrative Genomic Visualization (IGV) software, to look at selected regions of the genome, for example, positive and negative control regions used earlier for qPCR checks. IGV is a helpful tool for real-time visualisation of large genome data at base pair level, and also at broader scales, e.g. 100kb or 1Mb, which can help give a sense of the genomic landscape [166].

An initial screening using twenty-five regions containing the 13-mer motif, showed that clear PRDM9 peaks appeared to be centred at or very close to the 13-mer motif sites. We observed that the strongest signals from the PRDM9-IP samples appeared to have around 150 reads corresponding to about 50-fold enrichment relative to the corresponding genomic control lane. Most of the THE1 and L2 repeat regions (explored owing to the strong overrepresentation of these repeat elements in recombination hotspots [55]), showed that the peaks exhibiting PRDM9 binding, contained an exact match to the 13-mer motif. Of the 25 regions explored in the PRDM9-IP sample, binding peaks appeared to be narrow, mostly intronic and almost always at the motif sites. This provided a strong indication of PRDM9 binding in target regions given that there were no such peaks in the corresponding control lane. H3K4me3 peaks appeared to overlap the PRDM9 peaks with some off-set and enrichment being weak and apparently diffused over

about 1-2kb around the PRDM9 binding site, and interestingly, similar to hotspot width.

Figure 4.4 shows broad scale plots for recombination hotspots E, F and CG. These hotspots are three randomly selected cases from the "superhotspots" identified by Jeffreys et al., which are very active hotspots detected by high-resolution sperm crossover assays [63]. Each of these three hotspots exhibits enrichment of PRDM9 binding in case of both biological replicates, whereas no such enrichment is seen in the genomic control lane. Further, in each case, the PRDM9 peaks also contained the 13-bp motif to which PRDM9 is predicted to be bound. However, we also observed cases where a hotspot has no evident 13-mer but still corresponds to a PRDM9 peak. For example, Jeffrey's hotspot H was called by our assay as having a PRDM9 binding peak. but did not contain the 13-mer motif.

In addition to the first set of ChIP-seq experiments, we also performed H3K4me2 and H3K4me3-IP on untransfected HEK293T cells [assay performed by Nick Altemose]. This would help us to answer questions as whether PRDM9 requires a pre-existing H3K4me2 mark to confer its trimethylation mark [75], and whether H3K4me3 in untransfected HEK293T cells is comparable to the H3K4me3 levels in the PRDM9 transfected cells. Figure 4.5 is an IGV snapshot centred at the Hotspot F region. It is evident from this plot that PRDM9 is able to confer the H3K4me3 mark independently, without a pre-existing dimethylation mark in this region, as we see no enrichment in the H3K4me2-IP lane. We can also see that the H3K4me3 mark is absent from this region in the cells that have not been transfected with PRDM9, again supporting that PRDM9 made the trimethylation mark in this hotspot region. No background enrichment is evident in either transfected or untransfected genomic control lanes.

4.3.2 Implementing peak calling algorithms

To process files for peak calling, we first merged all PRDM9-IP and control samples across the three multiplexed lanes using samtools, resulting in two PRDM9-

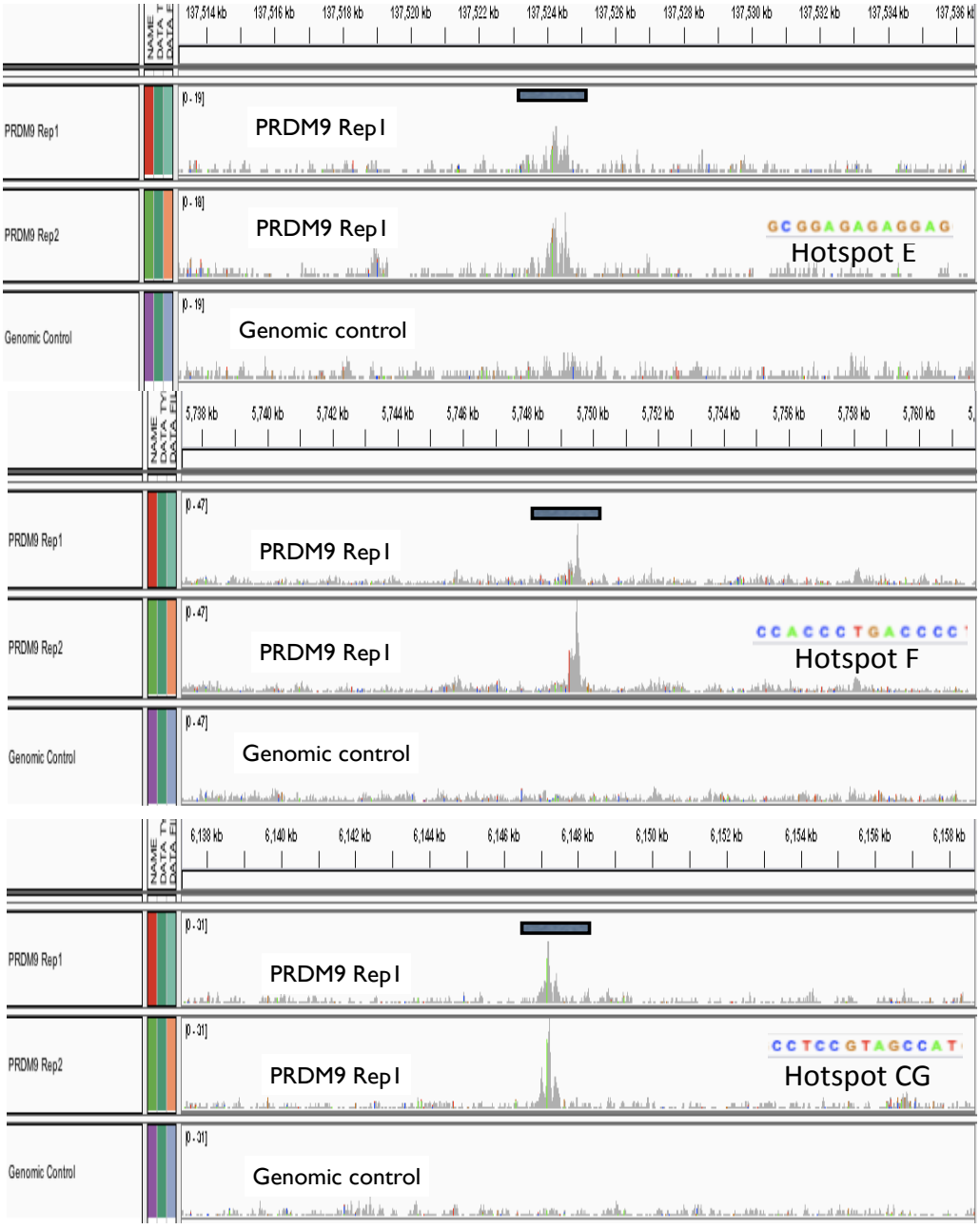


Figure 4.4: ***Visualizing raw read depth in selected hotspot regions.*** *IGV snapshot showing read depth data around hotspots E (Top), F (Middle) and CG (Bottom). For each hotspot case illustrated, the plot is centred at the respective hotspot, with the three lanes corresponding to three samples: two PRDM9-IP replicates (first two lanes) and transfected genomic control (last lane). An enrichment for PRDM9 reads is seen in all hotspot cases, consistently in each of the two IP replicates, relative to the control lane. Motifs on the bottom right are the best matches to the 13-mer found within 100bp of the centre of these hotspots.*

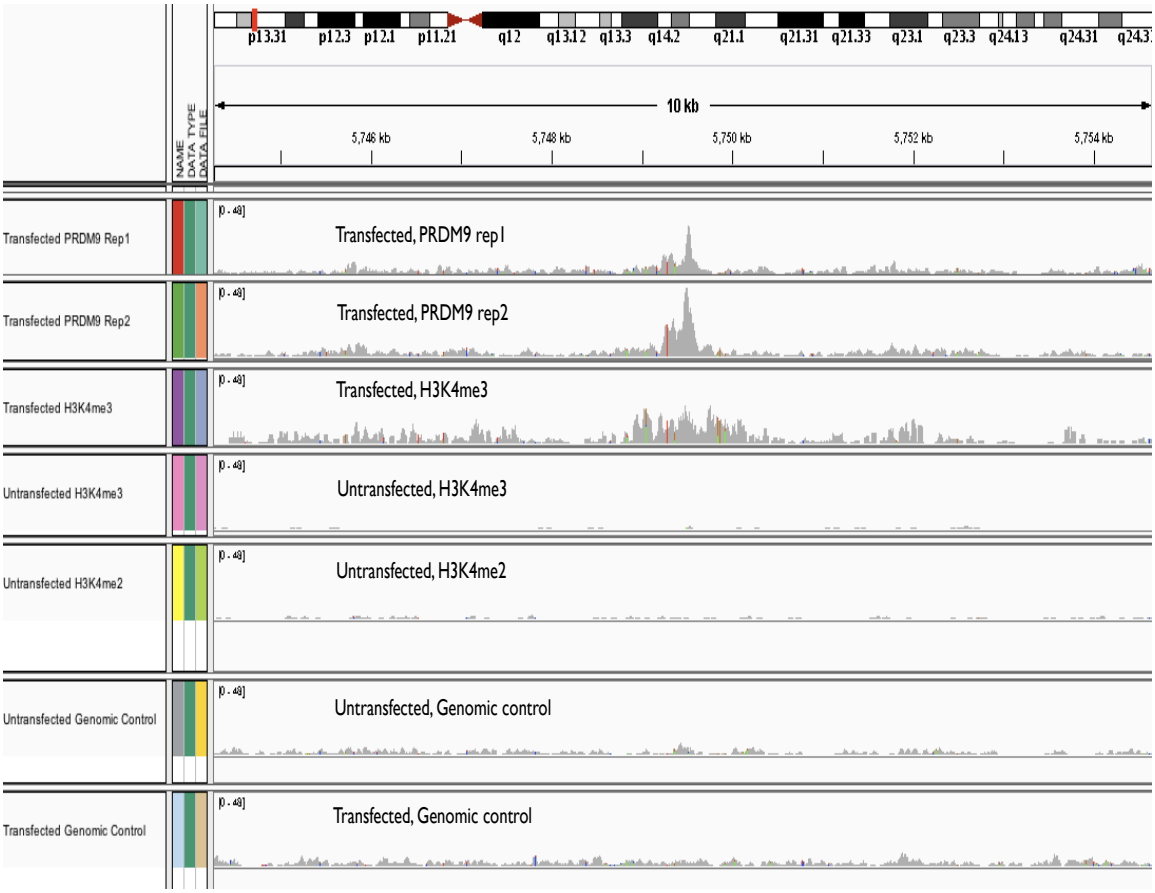


Figure 4.5: ***Visualizing raw read depth in untransfected and transfected IP samples.*** *IGV snapshot showing read depth data around hotspot F. Lanes 1 and 2 show PRDM9-IP samples enriched at this hotspot. Lane 3 shows enrichment of H3K4me3-IP in PRDM9 transfected cells. Lanes 4 and 5 show H3K4me3 and H3K4me2-IP in untransfected cells with no enrichment, and lanes 6 and 7 show untransfected and transfected genomic control with no background enrichment at this hotspot. This illustrates that PRDM9 is responsible for the trimethylation mark at this example hotspot.*

IP lanes and one genomic control lane. We removed all poorly mapped reads with mapping quality below 20 using Picard. Duplicate reads were removed using Samtools. Initial peak calling was performed using **MACS (Model-based Analysis of ChIP-Seq)** [167, 168] to identify PRDM9 binding sites in the genome. MACS is a non-interactive command line tool, which takes two input files i.e. mapped reads from IP samples in a range of formats (we used BAM files for the two PRDM9-IP cases) and a control data set (transfected genomic control). MACS calculates FDR based on the number of peaks in the control sample compared to the IP case sample, which are called at the same p-value threshold. It calls peaks using a two step process: First, it models read shift size and then calls peaks. As IP DNA reads do not give the exact location of protein binding, but rather present the ends of ChIP fragments, using the distribution of reads MACS models shift size. ChIP fragments are sequenced from both ends, the density of reads around the protein binding site is expected to show a bimodal enrichment. i.e. plus strand reads enriched upstream and minus strand reads enriched downstream of a true binding site. MACS uses a parameter (bandwidth) to slide windows across the genome in order to find an enrichment of reads relative to expected values set as default. These peaks are then selected, separated by strand and aligned at the centre point. The distance between the modes of the two peaks (d) is used to shift all reads by d/2 towards the 3' end for precisely locating the protein binding site [169]. This method identified a total of 37,180 PRDM9 binding peaks.

In parallel an **In-house peak caller** was developed by Simon Myers, which jointly calls peak intensities for both the ChIP sample and control lanes. This method was designed to take into account the information on local read depth from the control lane (a property similar to that of MACS), along with the information on paired end reads. The latter property is not accounted for by MACS which may result in a loss of power. Fragment coverage in 100bp bins was used from the two IP case lanes and the genomic control lane to compute two estimates α and β, with α indicating the comparison of background coverage in IP case samples to mean coverage in control lane and β indicating comparison of coverage owing to binding enrichment in one IP case lane to the binding enrichment in the other IP case lane in each bin. Given these two estimates, a likelihood ratio

test was performed, giving a p-value for each bin that indicates the probability of coverage observed in IP case lanes resulting from background alone. For peak calling, all regions with a p-value of $< 10^{-05}$ were taken, and to get peak centres, the likelihood ratio test was repeated for each base in these peak regions, and the position with the largest likelihood ratio was taken as the centre of the peak. The peaks called by the In-house method were classified as "Total peaks" (or the full set of peaks called) and "strong peaks" (which are the very strong peak cases, where the IP case lane fragment depth of over 100 had $>$ 4-fold enrichment in both the case lanes and a p-value of $< 10^{-10}$). This method identified 178,197 PRDM9 bound sites (Total peak set), with strong peaks constituting 6,221 peaks of this set.

Comparing the peak widths of the two peak callers, it was seen that the MACS called peaks which ranged between 300-13,000bp (mean width of peaks= 1292, median width of peaks= 960), whereas the In-house method called peaks ranging from 300-3000bp (mean width= 237, median width=240). We also looked at the overlap between peaks called by each of these methods and 19 "superhotspots" (very active crossover hotspots identified using sperm crossover assays targeting regions with extreme LD breakdown in the HAPMAP genotypes) reported by Jeffreys et al. [63]. MACS was able to call 10 out of the 19 super hotspots, whereas 14 out of 19 superhotspots were marked as peaks by our In-house method. We therefore found the In-house method the better choice to use for further analysis.

4.3.3 Distribution of PRDM9 peaks in recombination hotspots, promoters and genomic background

We analysed how many of the PRDM9 binding peaks called by either of the two peak callers overlapped hotspots and vice versa. About 30% of the 37,180 peaks called by MACS overlapped with HAPMAP recombination hotspots (compared to 7% of the genome), conversely, 26% of these hotspots overlapped with the binding peaks. Looking at the binding peaks called by the In-house algorithm, we observed that about 19% of the 178,000 peaks overlapped with recombination

a)

Peak Callers	Total PRDM9 Peaks called	PRDM9 Peaks overlapping Hotspots	Hotspots Overlapping PRDM9 Peaks
MACS	37180	11239 (30.2%)	9007 (26.3%)
In-House	178197	34573 (19.4%)	17957 (52.6%)

b)

Background	% of Total PRDM9 peaks overlapping	% of Strong PRDM9 peaks overlapping
Non-Repeat	47.5%	38.8%
ALU	14.2%	8.2%
THE1	4.5%	17.4%
L1	5.6%	5.3%
L2	8.3%	8.9%

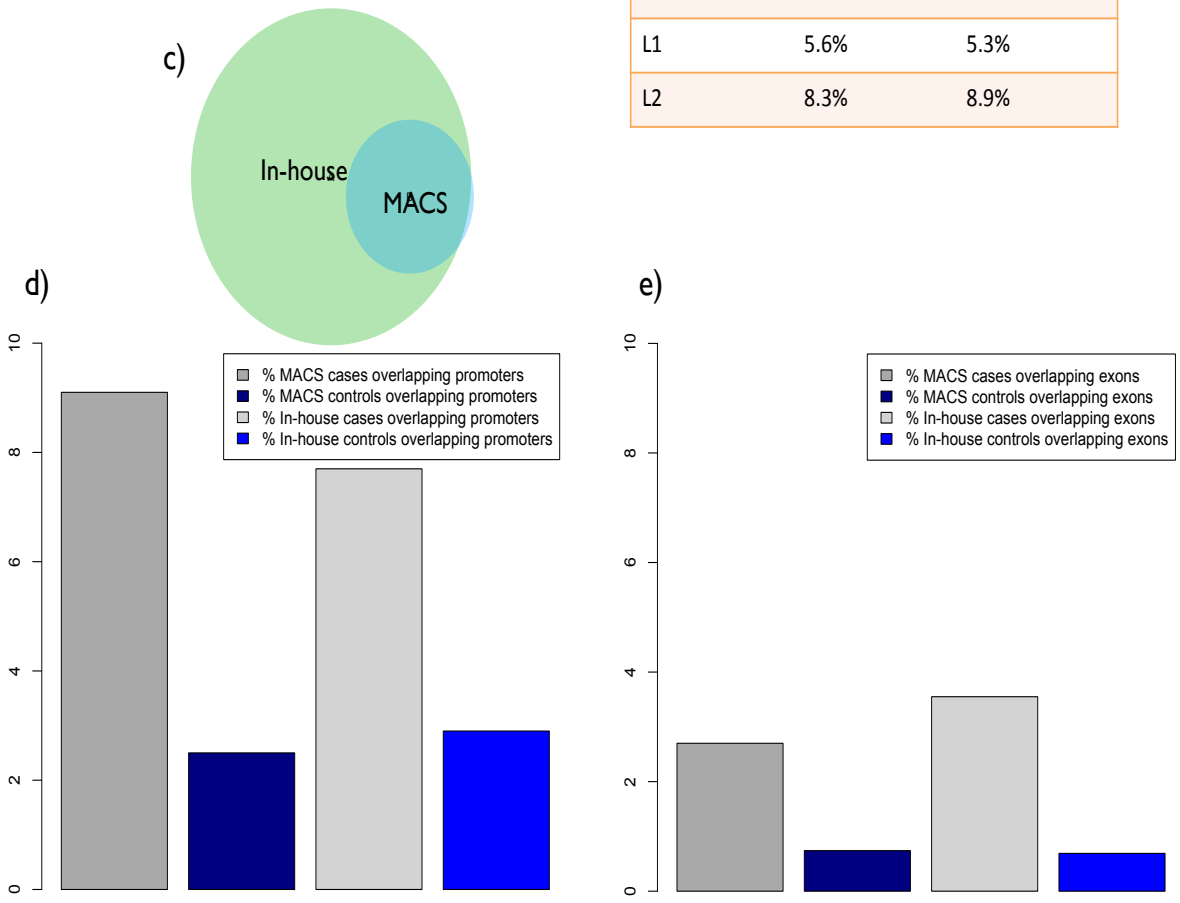


Figure 4.6: ***Overlap of PRDM9 peaks with genomic annotations.*** *a) Overlap of PRDM9 binding peaks identified by MACS and In-house method with HAPMAP recombination hotspots and vice-versa, b) Overlap of PRDM9 peaks identified by In-house method with various genomic backgrounds, c) Venn diagram showing overlap between In-house and MACS peaks, d) Percentage of PRDM9 peaks that overlap with promoter regions and e) Percentage of PRDM9 peaks that overlap exons.*

hotspots (compared to 6% of the genome), and about 53% of the recombination hotspots overlapped with the PRDM9 binding peaks. Given that the In-house peak caller was able to capture double the fraction of hotspots compared to MACS, we therefore opted to use the expanded peak set called by the In-house method for all subsequent analysis. The In-house method was also preferred as this approach made full use of our case and control sample lanes, called narrower peaks and called more superhotspots compared to MACS, which on the contrary generated peaks that were much wider and would increase uncertainty with respect to PRDM9 binding locations within peaks called.

There were 47% hotspots that were not accounted for by PRDM9 binding detected by our assay. Possible reasons for PRDM9 not binding in these regions could be that these hotspots are not active in the HEK293T cell line used for this experiment, but may still bind these hotspots in meiotic cells. This may be because of a difference between the two cells types with respect to chromatin state or other cis-acting features. It could also be that our In-House method was not able to detect these bound sites.

We also investigated the percentage overlap of PRDM9 peaks called by both peak callers with promoters and exons in the genome. Previous findings have shown that PRDM9 binding occurs outside of promoter regions [75]. Interestingly however, we noted that PRDM9 binding sites are about 5-fold enriched in both promoter and exonic regions as opposed to our genomic control regions. The control set for the In-house method was generated using a command line tool (shuffle) from the Bedtools package; this tool, by randomly permuting genomic locations of the reference human genome, created a comparable set of 178,000 genomic control regions that excluded the PRDM9 peak regions called by our method. A control set for MACS peak cases was also generated in exactly the same way. Figure 4.6 shows that more PRDM9 peaks called by the in-house method overlapped with promoters (7.7%) and exons (9.1%), compared to controls, 2.5% and 2.9% respectively. Similar distribution was observed when analysing peaks called by MACS.

Figure 4.6 also shows the distribution of PRDM9 peaks in different genomic backgrounds. On examining peaks in non-repeat regions, we observed that 47% of the 178,297 peaks overlapped the non-repeat region, whereas about 14% and 4% of the PRDM9 peaks identified overlap the ALU and THE1 regions, respectively. Overlapping peaks are defined as PRDM9 peak regions that overlap one more more bases of a repeat or non-repeat background.

4.3.4 Identification of a 14-bp motif enriched in PRDM9 peaks

We next used our PRDM9 peak calls to search for novel motifs in the binding sites. In order to perform an *ab-initio* motif discovery, we used a web interface toolkit called the MEME suite. This suite has multiple tools that facilitate various types of motif analysis, two of which "MEME-ChIP" (or MEME) and "FIMO" we used to perform our analysis [170, 171].

MEME-ChIP takes ChIP-seq peak regions to perform *ab-initio* motif discovery and also performs a motif enrichment analysis using a computational pipeline. FASTA sequences of the peak regions are taken as input, which the software centres and trims to 100bp, thus the peaks must contain the motif within their 100-bp central region. The trimmed sequences are then used by the motif discovery algorithm. A maximum of 600 sequences can be processed by MEME-ChIP at one time owing to computational complexity, therefore if a larger set is given, it randomly samples 600 sequences from the larger number of input sequences to discover novel motifs and also provides position specific weight matrices (PWM) for the motifs. The statistical enrichment of a motif in the given set of input sequences is carried out by a motif enrichment algorithm (AME) [171, 172].

FASTA sequences of strong peak regions (cases) of 150bp length were given as input to MEME-ChIP, which randomly selected 600 sequences for motif discovery. The same number of randomly selected control sequences, 150bp in length, were generated by shuffling genomic regions, excluding the PRDM9 peaks and repeat

regions. The strong peak cases were investigated to find motifs by stratifying into four subsets to account for any sequence biases 1) PRDM9 peak region sequence without any filtration, 2) Peak regions masked for repeats, 3) peak regions masked for promoters (promoters were defined as regions that begin at 2kb from the start of a transcription start site and at the TSS) and 4) peak regions masked for promoter and repeat regions.

Figure 4.7 shows the percentage of sites a motif inferred by MEME is estimated to be present in each of the 4 subsets of peaks tested. Out of the total strong peak regions submitted to MEME-ChIP, 600 were randomly selected, of which 557 contained a close match to the 11-bp motif CCTCCCTCC. Repeat masked peak regions contained an 18-mer in 87% of all regions analysed. A longer 24-bp motif was found in peak regions including promoters 99% of the time, whereas the same regions outside promoters were enriched in a 14-mer motif 97% of the time. Each of these motifs was a close match to the previously identified 13-mer (CCNCCNTNNCCNC) motif [68].

Finally, we also repeated this analysis after conditioning on motif width being 25-50bp, to uncover any longer motifs that may be present in the PRDM9 binding sites. We observed 30-33bp motifs in each of these cases of the full peak set, excluding promoters and repeat masked peaks (data not shown). Figure 4.8 shows a 33bp motif discovered by filtering for promoter regions, and was found to be centrally enriched in 520 (87%) of the 600 randomly selected regions. This motif contains the 14-mer described above and also some upstream sequence which appears to be a close match (i.e. 6 out of 8 of the upstream bases in figure 4.8) to upstream sequence discovered by Myers et al. [68]. The upstream sequence flanks the motif and was found by looking for a difference in base composition at positions around the 7-mer (CCTCCCT), between narrow hotspots and coldspots. Notably, this demonstrates that our peaks are able to perfectly recover the sequence feeling of human hotspots. For subsequent motif analysis we used the 14-mer motif identified in non-promoter, non-repeat PRDM9 peak cases.

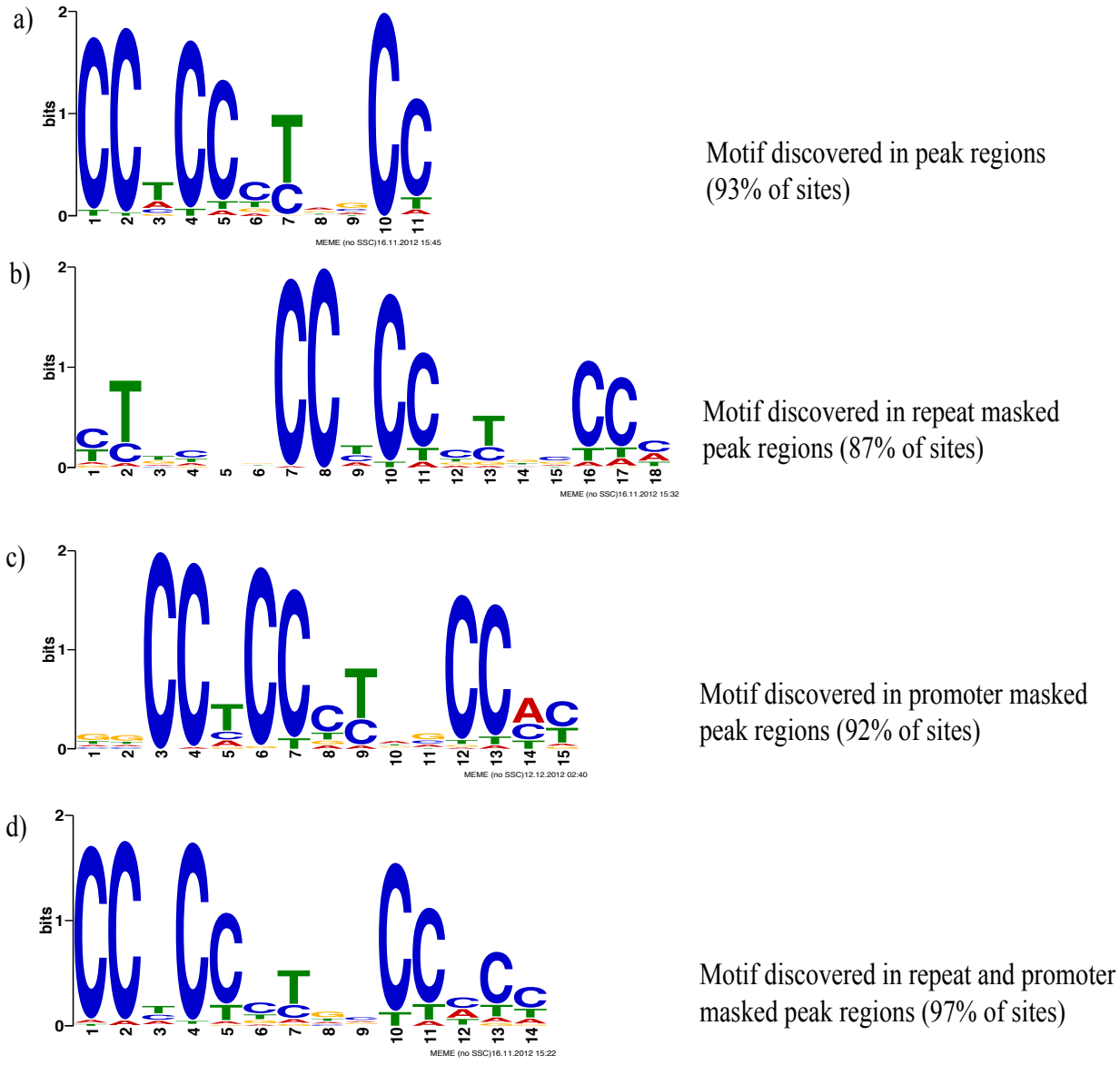


Figure 4.7: ***Motifs discovered from the PRDM9 binding peaks identified from ChIP-seq assay.*** *Motifs discovered by the MEME-ChIP software from 600 strongest PRDM9 binding peak regions, stratified as: a) Full set of strong peaks b) Repeat masked peak regions c) Promoter masked peak regions and d) promoter and repeat masked peak regions. Motifs in all four subsets match closely to the previously identified 13-bp motif and are enriched in PRDM9 peaks.*

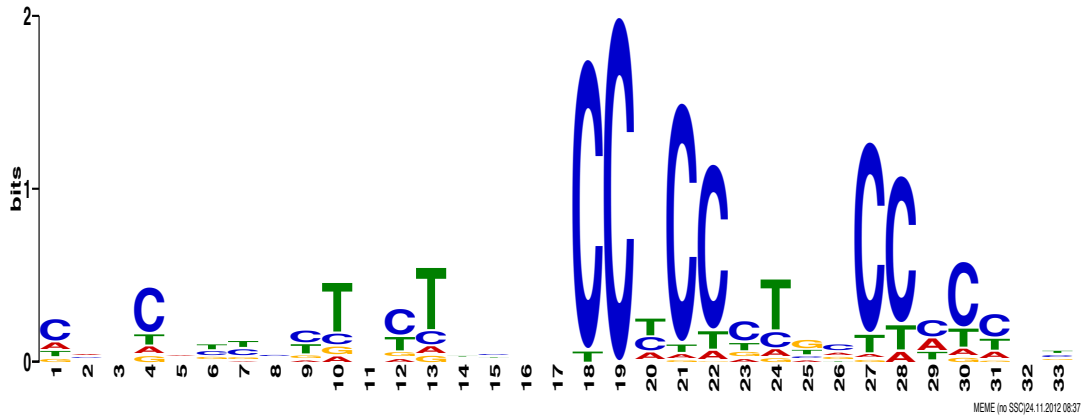


Figure 4.8: ***Search for longer motifs in peak regions.*** *Motif identified in non-promoter regions conditioning MEME to identify motifs longer than 25bp.*

To determine an association between the PRDM9 peaks and the 14-mer motif discovered, we searched for all matches to the 14-mer across the genome. We used the FIMO program from MEME suite to identify all 14-mer motifs in the genome [170]. FIMO used as input, the PWM generated by MEME-ChIP for the 14-mer motif and a p-value threshold. The PWM was slightly adjusted to avoid zero weights. This was done by increasing the number of sites at which motif (n) was found by one ($nPWM + 1$). For example, if the an entry in the 14-mer PWM was x, on adjustment, this would become:

$$(x + 1) * (nPWM)/(nPWM + 1) + 1/(nPWM + 1),$$

where $nPWM$ is the number of motifs found. Therefore if x=0, on adjustment, this value would be replaced with $1/(nPWM+1)$. FIMO also computes log-odds scores for each position of the motif in the genome (as described by its PWM) and converts these scores into p-values. With the inputs above, FIMO identified genome-wide positions of the 14-bp motif with a p-value threshold below 10^{-05}.

4.3.5 Recombination rates around 14-mer motifs

We next investigated whether the PRDM9 bound motifs also correspond to an elevated rate of recombination. Figure 4.9 shows HAPMAP recombination rates around bound motifs in different repeat and non-repeat backgrounds. PRDM9 bound motifs clearly exhibit a strong, localised peak for recombination activity whereas unbound sites do not show a peak. Looking at bound motifs overlapping promoter regions versus those in non-promoter regions, we observed that recombination rate in PRDM9 bound promoters was much lower compared to non-promoter bound regions. This indicates that although PRDM9 appears to show binding preference in promoter regions, these regions clearly exhibit lower crossover recombination rates, which may be a consequence of PRDM9 itself somehow suppressing this activity, or there may be an independent mechanism responsible here. The regions strongly bound by PRDM9 also showed elevated rates of recombination around non-promoter regions as opposed to bound promoters. Further, looking at the rates around PRDM9 bound THE1 and Alu repeat regions, we found an increased rate of recombination around bound sites as opposed to unbound sites.

4.3.6 Strength of motif correlated with PRDM9 binding

FIMO, used to obtain positions of all motif matches in the genome, returned a total of 804,610 14-mer motif positions. Each motif found had a corresponding score indicating strength of motif match. Given this information, we aimed to ask two questions: 1) How predictive is the presence of a motif for PRDM9 binding and 2) How predictive is the motif score for PRDM9 binding. We first investigated overlaps between PRDM9 peaks and motif presence and observed that there were 3424 (55%) strong PRDM9 peaks that overlapped the 14-mer motifs, whereas only 6% of these peaks contained a control 14-mer motif. Control motifs were generated by using the bedtools package to randomly select 14-mers in the genome, while excluding regions that lie within 2kb of the canonical 14-mer, ensuring that there are no overlaps with other control 14-mers. In case of the full set of PRDM9 peaks, we saw that 55443 (32%) of these peaks overlapped with 14bp

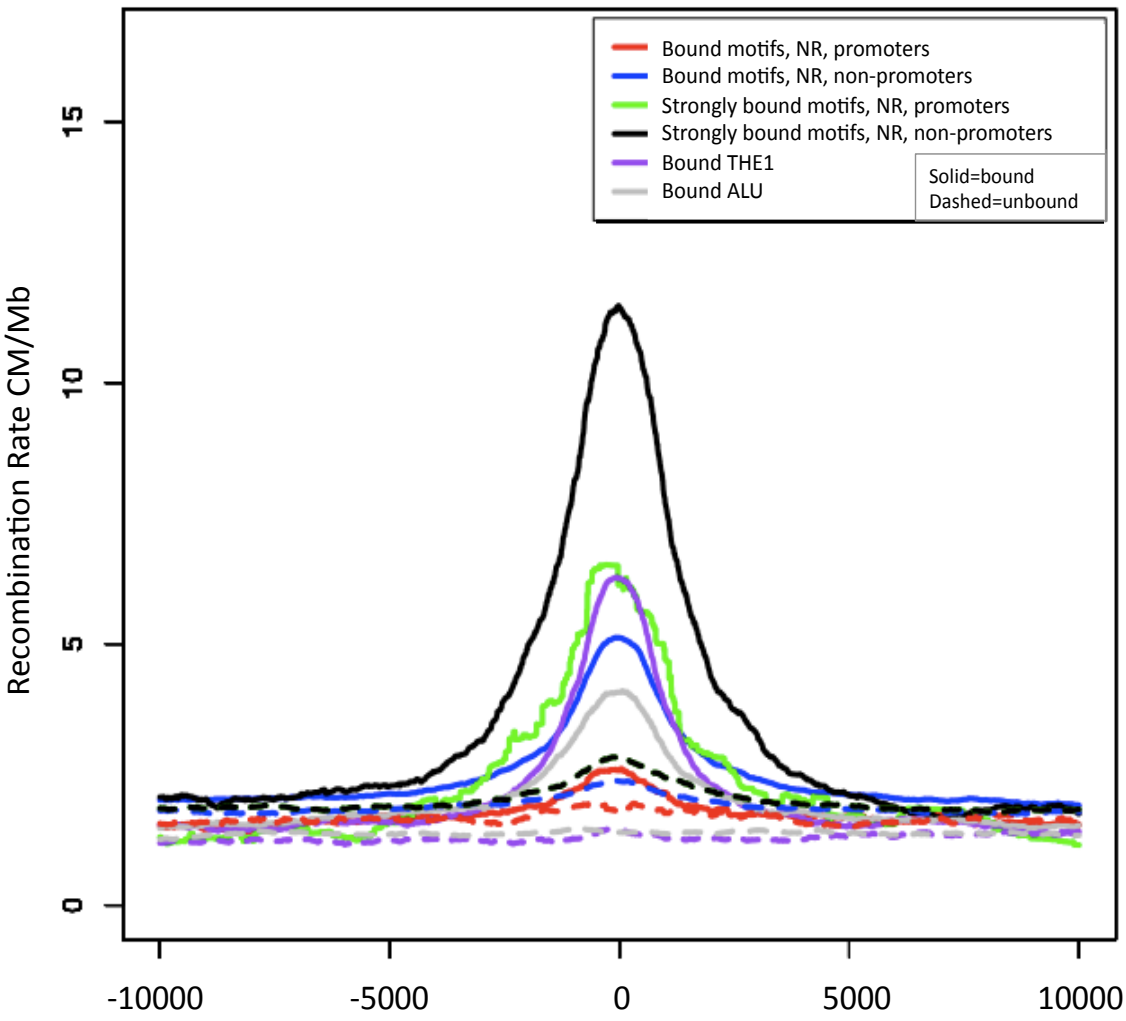


Figure 4.9: ***Recombination rates around PRDM9 bound sites.*** *Recombination rates around non-repeat, THE1 and ALU repeat elements. Dashed lines represent rates around unbound motif sites.*

motifs (conversely, about 10% motifs overlapping same peaks), compared to 17% containing the control motifs. The motifs overlapping or not overlapping PRDM9 peaks were classified as 'Bound' and 'Unbound' motifs, respectively. These findings implied that not all PRDM9 peaks contained the 14-mer motif and not all motif occurrences were bound by PRDM9. In addition, stronger peaks appear to have different characteristics, with a strong motif overlap than weaker peaks. More recent evidence shows more than 35% of remaining strong peaks do still contain the motif (findings from motif work done by Simon Myers indicates that more than 90% of strong peaks contain degenerate versions of the motif), but would have a score lower than 13.5.

One possible reason for the low (32%) overlap between PRDM9 peaks and motifs could be that we are unable to get binding information for hotspot motifs that are enriched in repeat regions as a consequence of low mappability in these elements. Although paired-end sequencing has been shown to increase coverage in the repetitive regions of the genome, these accessibility issues (in repeat or certain chromatin states) are not completely eliminated by paired-end reads, e.g. repeat identification or placement concern would remain if a repeat region is longer than read length (for instance, our read length is about 200bp whereas a THE1 repeat element is 330bp long). Hence the sequenceability of such repeat elements may affect signal detection for hotspot motifs present in these regions. In the process of deleting poorly mapped reads, about 7% of the reads in THE1 repeat elements, 5% of reads in ALU repeats, 3% of L1 and 1% L2 reads were filtered out from our data, owing low mapability scores (<20). These lost reads may contain a fraction of hotspot motifs which we are unable to access for our analysis.

We further investigated whether the FIMO score of 14-mer motifs was indicative of being bound. The score indicates the strength of match to the 14-mer motif identified by MEME; the higher the score, the closer the match. Figure 4.10 shows that as the FIMO score increases the fraction of motifs bound also increases, thereby suggesting that stronger the genomic sequence matches to the 14-mer motif the better chances it has of being bound by PRDM9. In cases where motif presence or strength of motifs do not play a role in PRDM9 binding,

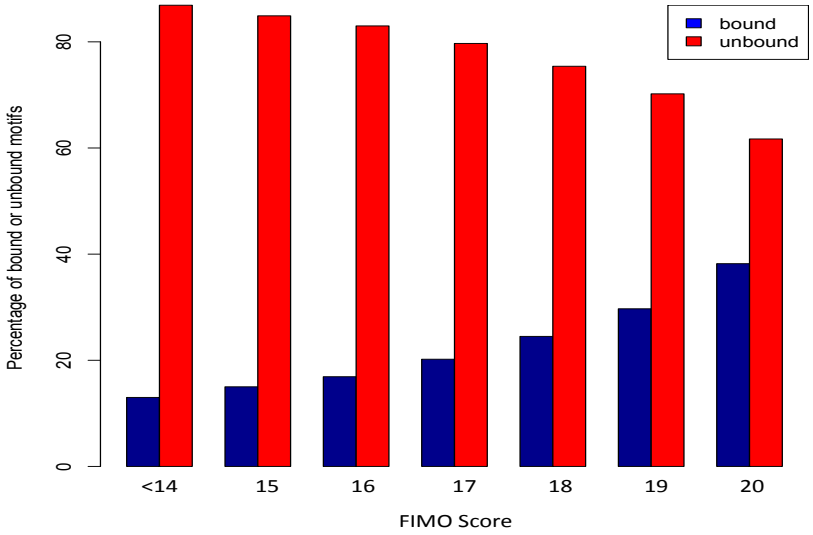


Figure 4.10: ***Stronger motif matches correlated with PRDM9 binding.*** *Motifs are stratified by FIMO scores (13 indicating the poorest match and 20 indicating the best match to the 14-mer motif). The plot shows that the fraction of PRDM9 bound peaks increase with an increase in motif scores.*

it might be that other transcription factors are directly or indirectly playing a role to facilitate binding activity.

4.3.7 Genomic context influences PRDM9 binding and hotspot probability

We next aimed to understand whether the various genomic contexts are able to influence PRDM9 binding and the probability of becoming hotspots. We therefore investigated PRDM9 binding in non-repeat and various repeat contexts (see Table 4.2). We specifically looked at four types repeat elements THE1, ALU, L1 and L2. The choice of the THE1, L2 and ALU repeat contexts, was based on previous observations by Myers et al. showing enrichment of these elements within recombination hotspots [68]. These repeat elements also show a peak of recombination activity around the centre of the canonical 13-mer motif compared to repeats lacking this motif, thereby indicating that the motif has a strong impact

on the enrichment of these repeat elements in recombination hotspots. On the contrary, L1 repeat elements have been shown to be strongly underrepresented in recombination hotspots [68], [55].

About 47% of the total bound motifs were in the non-repeat region, 10% in THE1 repeat regions and 9.4% in ALU repeat elements. In case of strongly bound motifs there were 30.8% in non-repeat regions, 22.2% in THE1 and 2% in ALU repeat regions. Hence, THE1 elements among other repeat and non-repeat regions seem to most influence PRDM9 binding. Out of the total PRDM9 peaks that overlapped non-repeat regions, 15% overlapped the 14-bp motif, with 20% of these bound motifs making a hotspot. Similarly, 36.3% and 2.2% of all peaks found in THE1 and ALU repeat background, also contained the 14-mer motif, and of these motif containing bound regions 24% and 19% overlapped HAPMAP hotspots.

There was a striking difference observed between backgrounds in how often they are bound, or strongly bound by PRDM9, with binding being about 15-fold more likely in THE1 than ALU elements. We also observed more subtle 2-fold differences in binding and hotspot probability in case of L1 and L2. Notably, LINE elements appear to often form hotspots once bound by PRDM9. These results illustrated that genomic background not only very strongly influences binding, but also hotspot probability on binding. This impact of genomic background might be owing to chromatin structure in these regions or it might also be due to the presence of other neighbouring sequences, e.g. local motifs specific to certain backgrounds which have previously been shown to be correlated with recombination activity [68].

4.3.8 Motif search in PRDM9 chromatin accessible regions

Our next objective was to understand the relationship with DNase accessibility i.e. whether a motif in chromatin accessible regions is more likely to be bound

Background	Total motifs	PRDM9 Bound motifs N=76172	Percentage bound	PRDM9 Bound motifs in Hotspots N=15694	Percentage of bound cases which are hotspots
NR	235593 (36.1%)	35526 (46.6%)	15%	7200 (45.8%)	20.2%
THE1	20803 (3.2%)	7563 (9.9%)	36.3%	1844 (11.7%)	24.3%
ALU	324820 (49.8%)	7181 (9.4%)	2.2%	1363 (8.6%)	18.9%
L1	46328 (7.1%)	1473 (2.1%)	3.1%	500 (3.2%)	33.9%
L2	23743 (3.6%)	3239 (4.4%)	13.6%	977 (6.2%)	30.1%

Background	Total motifs	Strongly Bound motifs N=4936	Percentage strongly bound	Strongly Bound motifs in Hotspots N=1968	Percentage of strong bound cases which are hotspots
NR	235593 (36.1%)	1525 (30.8%)	0.64%	663 (33.6%)	43.4%
THE1	20803 (3.2%)	1112 (22.5%)	5.3%	536 (27.2%)	48.2%
ALU	324820 (49.8%)	131 (2.6%)	0.04%	54 (2.7%)	41.2%
L1	46328 (7.1%)	115 (2.3%)	0.24%	70 (3.5%)	60.8%
L2	23743 (3.6%)	265 (5.3%)	1.1%	186 (9.4%)	70.1%

Table 4.2: Distribution of PRDM9 bound and unbound motifs in different genomic backgrounds. *Total PRDM9 bound motifs and strongly bound motifs are enriched in non-repeat and THE1 repeat elements.*

by PRDM9 or not. We looked at the effect of DNase accessibility on the motif using DNase hypersensitivity data on the gm12878 line from ENCODE [135]. In order to perform this analysis, we took PRDM9 peaks, and measured the DNase accessibility signal at the peak centre in non-repeat cases, matched for genomic coverage and enrichment score. We stratified the motifs into those that were in the bottom 10% of the distribution and those that were in the top 10% of the distribution. We then performed a motif search using MEME for the two sets of peaks i.e. motifs falling in more chromatin accessible and less chromatin accessible regions. The initial question we were aiming to address is whether degeneracy depends on chromatin nature.

Surprisingly, MEME results showed that in the top 10% of the most DNase accessible peaks, the top scoring motif was a perfect match to the CTCF motif (present in about 25% of the cases). These regions did not contain our canonical 14-mer motif (see figure 4.11). This shows that we are able to find centrally distributed CTCF binding sites by taking peaks without the PRDM9 motif; hence the CTCF motif, alone, is able to mark some PRDM9 peaks. This might suggest that CTCF is a good marker of chromatin accessibility, which is interesting as CTCF is known to position cohesin. Some other transcription factor binding sites appeared to be enriched in these peaks also, e.g. AP-1 is one of the transcription factors that appears to mark many non-14mer peaks (data not shown).

The bottom 10% least chromatin accessible peaks revealed the 14-bp PRDM9 target motif in about 67% of the sites. Further, on repeating the same analysis with strongly bound PRDM9 peaks, we observed an exact match to the PRDM9 motif in 96% of the least chromatin accessible cases. While in case of strongly bound regions present in the most chromatin accessible regions, we also observed the presence of the 14-mer motif, however, this appeared to be of a more degenerate nature, implying a role of chromatin structure.

Masking peak regions for promoters appeared to infer the same motifs in case of both bound and strongly bound peak regions i.e. the CTCF motif was again found in the most chromatin accessible PRDM9 binding sites, but it appeared

that this signal gets stronger, now found in over half of the most chromatin accessible PRDM9 binding sites. It could be that there are a number of peaks where in our assay PRDM9 pulls CTCF sites in accessible chromatin; this could indicate indirect binding of PRDM9 [135]. It could be that these are just false positive results but it might also indicate a real and different mode of binding associated with cohesin.

4.3.9 Canonical motif match revealed in hotspot cases with low 14-mer motif probability

We were next interested to know how PRDM9 is able to bind hotspots with low motif probability. Information on these hotspots with low motif probability was taken from a previous analysis performed by Myers et al. [68]. Looking at PRDM9 binding peaks in this set of hotspots, we noted the corresponding DNase sensitivity signals to be low. In comparison, binding peaks in hotspots with higher motif probability showed relatively higher DNase accessibility. It was therefore intriguing how these hotspots are bound by PRDM9, while the majority of similar low motif probability hotspots do not have a binding peak.

To investigate this, we looked at 600 PRDM9 peaks in hotspots which were annotated as having low motif probability for the hotspot they are in, and also had low DNase accessibility. These regions were interesting, as they are bound by PRDM9 and make hotspots but it is not understood how binding occurs, and it may be possible that these low motif probability hotspots represent a different mechanism which is why we can't see binding at them. These regions were run through the MEME motif finder to discover whether these contained degenerate matches of the canonical motif or a novel, previously undiscovered, motif.

Interestingly, results from the MEME analysis identified the 14-mer motif in 63% of these cases, even from this set of low motif probability, hotspot cases, thereby explaining most of the binding targets in these regions (figure 4.11). This meant that at least some of the hotspots annotated as unlikely to contain the 14-mer,

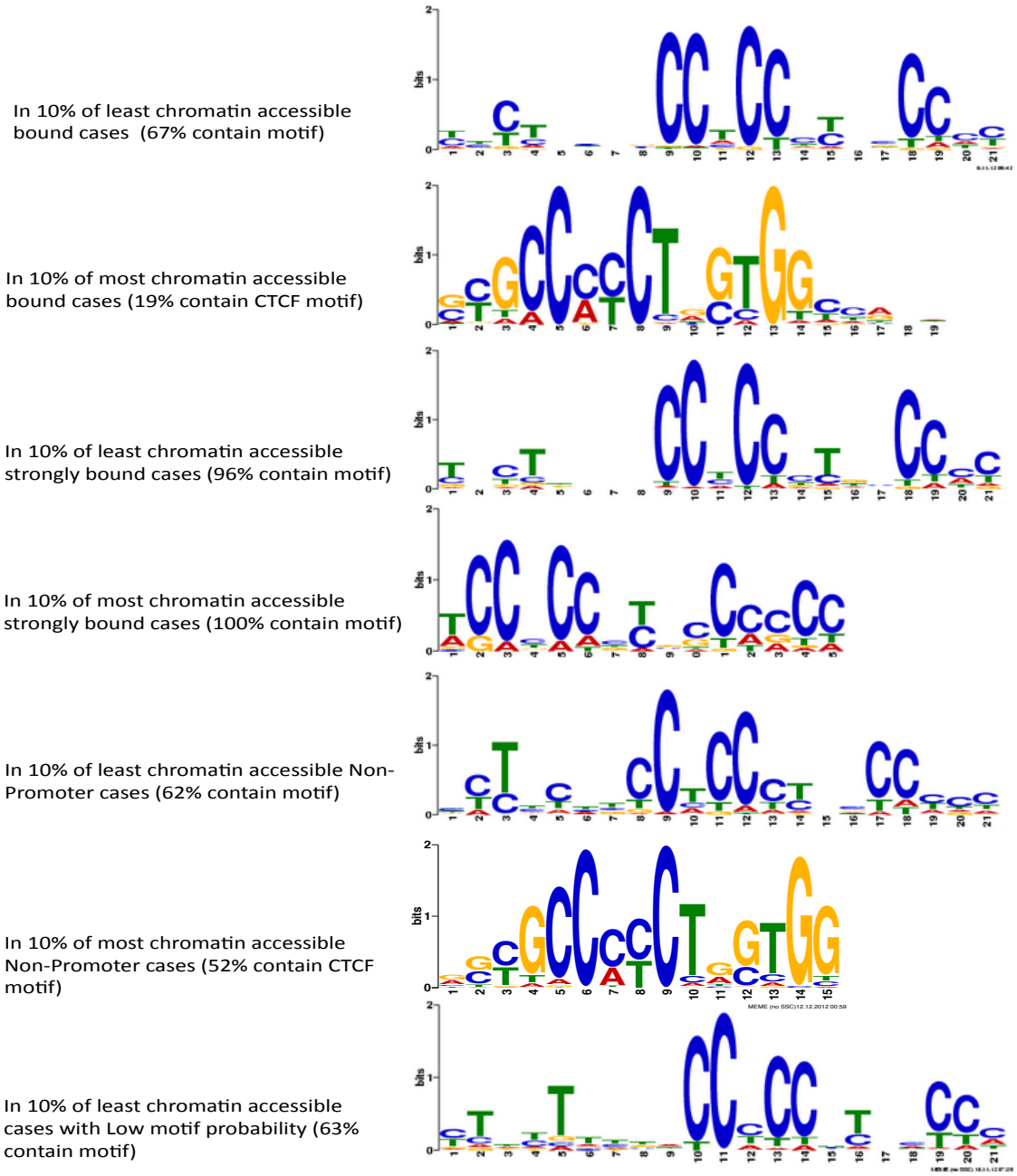


Figure 4.11: ***Motifs in PRDM9 peaks stratified by chromatin accessibility.*** *Least chromatin accessible peak regions (regions in the bottom 10% of DNase hypersensitivity signal distribution) appear to be enriched in motifs that closely match the 14-mer motif. Highly chromatin accessible regions (marked as those with the top 10% of DNase hypersensitivity signals) appear to be enriched in CTCF motifs.*

actually do contain more degenerate matches (given that there are never *exact matches* to the canonical motif to which PRDM9 binds). Also interesting is the upstream sequence which appears to be relatively more important in these cases (e.g. Figure 4.8). Hence, in many PRDM9 bound hotspots regions, we are able to pick up binding to the canonical motif , even if this was not picked up previously. However for the remaining cases where this motif is not present, it remains less clear how these are targeted by PRDM9 to become hotspots, and future work is needed to address this.

4.3.10 PRDM9 binding enriched in open chromatin regions

In Chapter 3 we discussed findings showing the effect of chromatin around the canonical motif in recombination hotspots; noting that the 13-mer motifs present within hotspot regions tend to exhibit lower chromatin accessibility compared to motifs in non-hotspot (coldspot) regions. Chromatin accessibility data was used from a non-meiotic ENCODE cell line (gm12878) [139]. With the PRDM9 binding data now available, we were able to extend our previous analysis. Based on this new information, we were able to stratify PRDM9 bound sites into bound hotspots and bound coldspot regions. We next asked whether chromatin accessibility is able to influence PRDM9 binding, and also whether open chromatin plays a role in hotspot formation once PRDM9 is bound to its target sites.

As done previously, we investigated the distribution of DNase hypersensitive sites around the strand oriented 14-mer motifs, bound or unbound by PRDM9, and also around bound motifs that form hotspots as opposed to those that do not. Figure 4.12 shows chromatin accessible sites surrounding the 14-mer motifs within different repeat and non-repeat backgrounds. Interestingly, bound motifs exhibit a prominent spike of chromatin accessibility compared to the 14-mer motifs that are not bound by PRDM9; the latter, on the contrary showing a slight dip at the motifs. Bound motifs when further stratified into bound hotspots cases and bound coldspots cases show that although there is a spike in chromatin accessibility in case of both hotspot and coldspot cases, the hotspot motif cases tend to be

surrounded by reduced levels of chromatin accessible regions, both towards the centre of the motif and in the background regions, as opposed to bound coldspot cases. This confirms our previous findings where 13-mer motifs in hotspot regions exhibited lower chromatin accessibility, possibly implicating the role of nucleosome rich regions in hotspot cases. Further, these results also extend to our previous findings in that they tell us that PRDM9 prefers to bind motifs in chromatin accessible regions, however, relatively lower accessibility allows hotspot activity. This is consistent across the genomic contexts explored.

The bound 14-mer motifs in THE1 repeat elements showed slightly higher DNase accessibility, localised around both bound and unbound motif cases, however on stratifying into hotspots and coldspots, there did not appear to be much of a difference between the two motif groups i.e. both having comparable chromatin accessibility surrounding the motifs. In case of motifs falling in ALU repeat regions, chromatin accessibility appeared to be much reduced, with a prominent dip towards the centre of the motif, in both bound and unbound cases. Hotspot cases had comparatively lower background compared to bound coldspot motifs. Overall, these results exhibit similar chromatin accessibility patterns in both hotspot and coldspot cases across genomic backgrounds, with almost similar average DNase sensitivity in all hotspot sets.

4.3.11 Nucleosome signals enriched around PRDM9 bound motifs

Further, in search of factors influencing PRDM9 binding and hotspot formation, we next investigated whether the positioning of nucleosomes is able to influence PRDM9 binding. In the previous chapter we saw that 13-mer motifs tend to have positioned nucleosomes surrounding them, in both hotspot and coldspot cases, with nucleosome signals being more enriched around hotspots. Here, we would like to get a sense of whether PRDM9 requires nucleosomes present at or around the motif site prior to binding, or if positioned nucleosomes are rather just a good substrate to form hotspots and do not affect binding affinity of PRDM9.

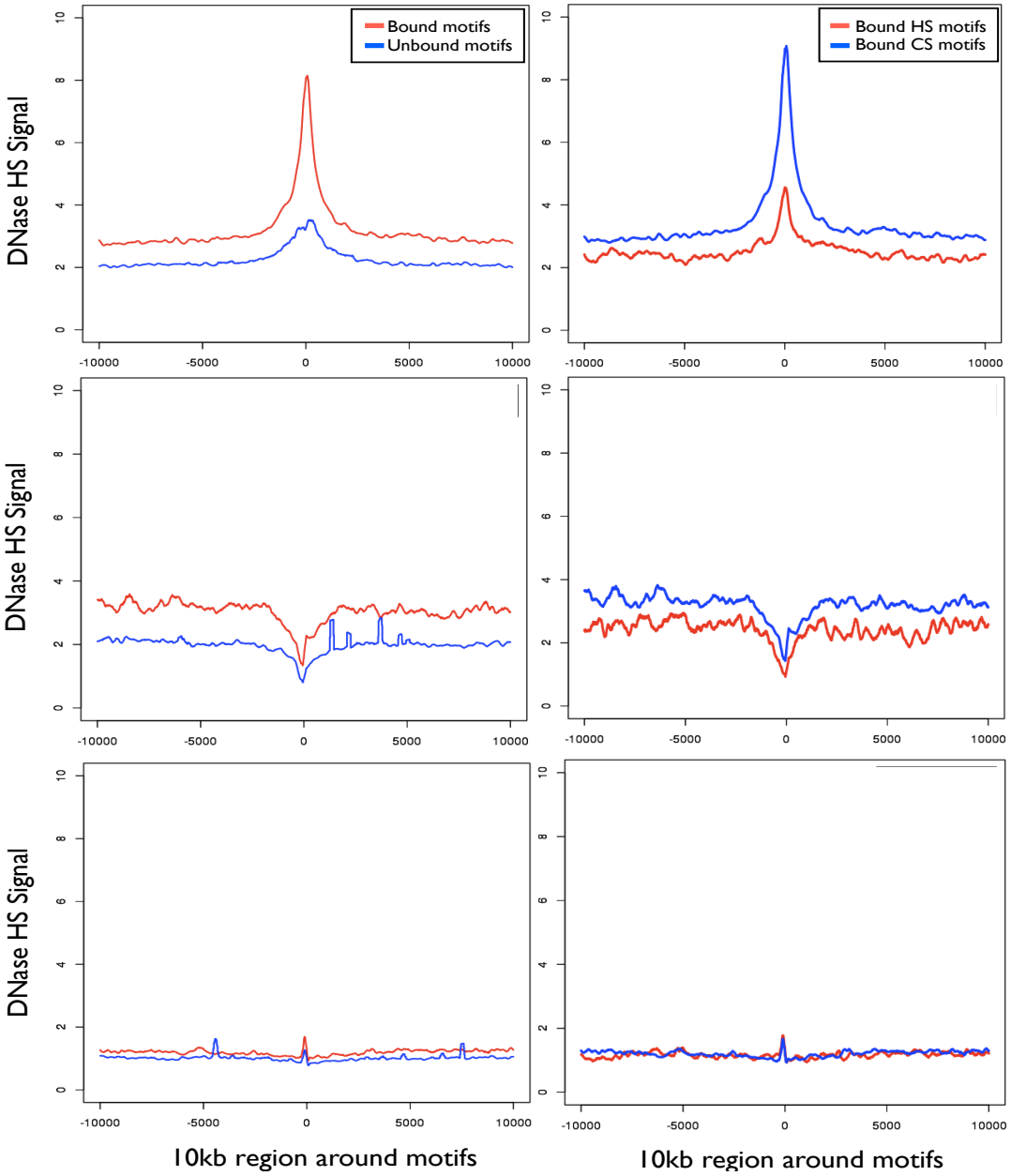


Figure 4.12: *DNase Hypersensitivity signals surrounding PRDM9 bound, unbound (left column), bound hotspot and bound coldspot (right column) 14-mer motifs, in Non-repeat (top row), ALU (middle row) and THE1 (bottom row) repeat backgrounds.*

Using stratified data as in the preceding section, we explored the nucleosome positioning effect on bound, unbound, bound hotspot and bound coldspot motif cases. Figure 4.13 shows plots for MNase data surrounding 2kb from the centre of the motif in the total set of bound and unbound cases. We observed evidence of nucleosomes positioned in both PRDM9 bound and unbound cases. Nucleosomes appear to be positioned 1kb on either side of both bound and unbound cases, however, the background levels of nucleosomes are higher in the PRDM9 bound cases. Interestingly, when stratified by hotspots and coldspots, bound hotspot motifs appear to exhibit higher nucleosome level, particularly towards the centre of the motifs as opposed to bound coldspot motif cases. We also repeated the above analysis with strongly bound motif cases, and observed similar results.

We also examined nucleosome effects on 14-mer motifs in THE1 and ALU repeat elements (See figure 4.14). In both cases, nucleosomes signals at the background level were again higher in bound motifs cases than unbound ones, with the region immediately surrounding the 14-mer motifs showing no difference with respect to nucleosome levels. Both THE1 and ALU cases showed a nucleosome signal positioned at the motif site, while in case of ALUs there is a strong signal of a second motif positioned just upstream of the motif site. Further, both repeat cases showed no apparent broad scale, or local differences between bound hotspot and coldspot motifs.

These results tell us that nucleosomes near the 14-mer motifs tend to be positioned before PRDM9 binding; it could be that nucleosomes are positioned by the high GC content of motifs, as seen previously in case of the CTCF motifs. Although binding appears to be stronger in regions where there is an elevated nucleosome density, with the motifs occurring at, or very near, a positioned nucleosome, it appears that once the motifs are bound by PRDM9 and the H3K4me3 mark is made, the chromatin changes may lead to the surrounding nucleosomes sliding into the region centred at the motif. Hence, the nucleosome signal at the motif sites in bound hotspots relative to bound coldspot cases, could be a downstream event of PRDM9 binding.

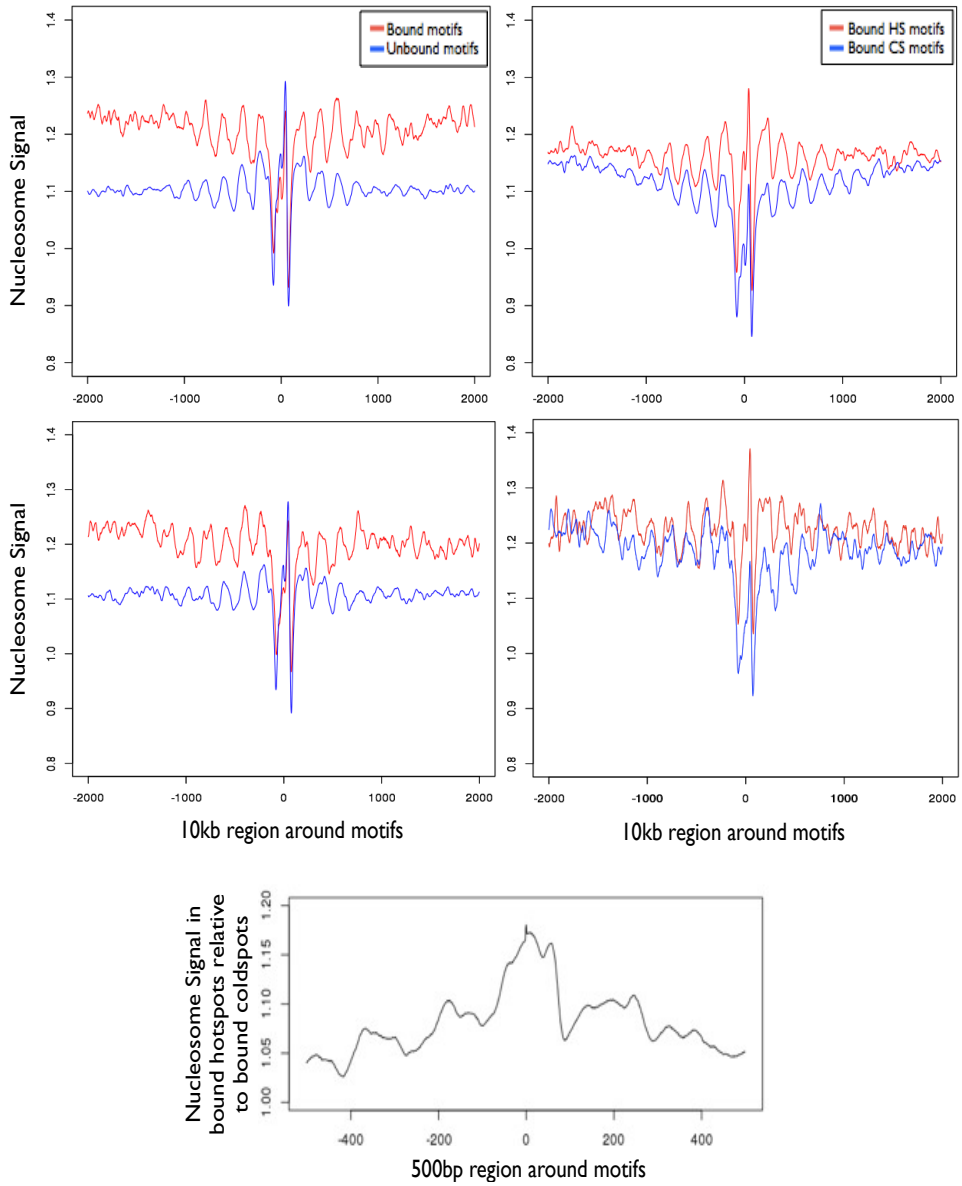


Figure 4.13: *Nucleosome signals surrounding PRDM9 bound, unbound (left), bound hotspot and bound coldspot (right) 14-mer motifs, stratified by all bound cases (top row) and strongly bound cases (middle row). Bottom row plot shows a ratio plot of bound hotspot motifs relative to bound coldspot motifs.*

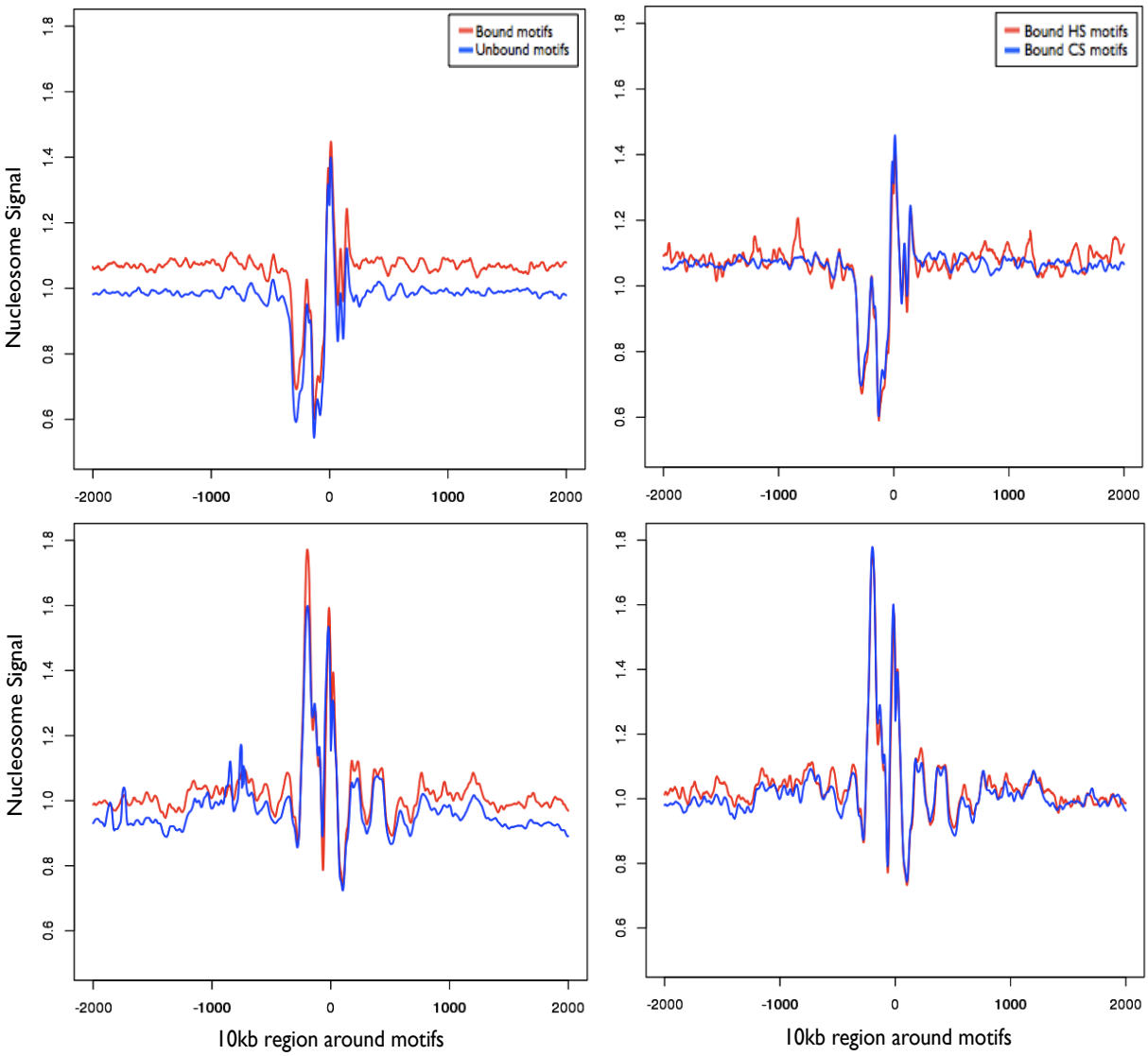


Figure 4.14: *MNase signals surrounding PRDM9 bound, unbound (left), bound hotspot and bound coldspot (right) 14-mer motifs in THE1 (top row) and ALU repeat (bottom row) elements.*

4.3.12 Transcription activating marks depleted around PRDM9 bound hotspot motifs

We were interested to know if transcription regulating marks are able influence PRDM9 binding and hotspots. In the previous chapter we observed that hotspots are more likely to occur in transcriptionally repressed regions of the genome. We extended the previous analysis using updated data and our ChIP-seq binding information. We investigated the effect of 10 transcription activating and repressing histone modifying marks on the binding of 14-mer motifs (See Figure 4.15, 4.17). Data on histone modification marks was taken from the UW ENCODE track on UCSC; the 10 histone modifications used are a total of all marks available for the gm12878 cell line (build 37). Of the 10 marks used, 6 mark promoters (H3K27ac, H3K4me1, H3K4me2, H3K4me3, H3K27me3, H3K9ac), 3 mark gene bodies (H3K36me3, H3K79me2, H4K20me1), and 1 marks constitutive heterochromatin (H3K9me3). All the marks studied, other than H3K9me3 and H3K27me3, are transcription activating.

The levels of each of these marks were explored in the four motif subsets mentioned previously. Interestingly, we noted that all 5 of the transcription activating marks in promoter regions, were enriched around bound motifs, however further dissection illustrated reduced levels of these marks in bound hotspot motif cases compared to bound coldspot cases. In almost all of these cases we observed a bimodal peak, with the peaks flanking the centre of the motif. These results imply that although PRDM9 binding itself is not affected by the presence of gene activating marks, however, most of these bound regions are unlikely to form hotspots.

Similarly, 2 out of 3 modifications marking gene bodies and transcriptionally active regions, were also seen to be enriched in bound motif coldspots and depleted in hotspots. Contrastingly, one mark of active chromatin, H3K36me3, appears to exhibit a lower signal in bound hotspots compared to bound coldspots. In case of the two transcriptionally repressive marks examined (H3K9me3 and H3K27me3) marking promoter and constitutive heterochromatin, respectively, we noted that the bound hotspot motifs were enriched for both these marks compared to bound

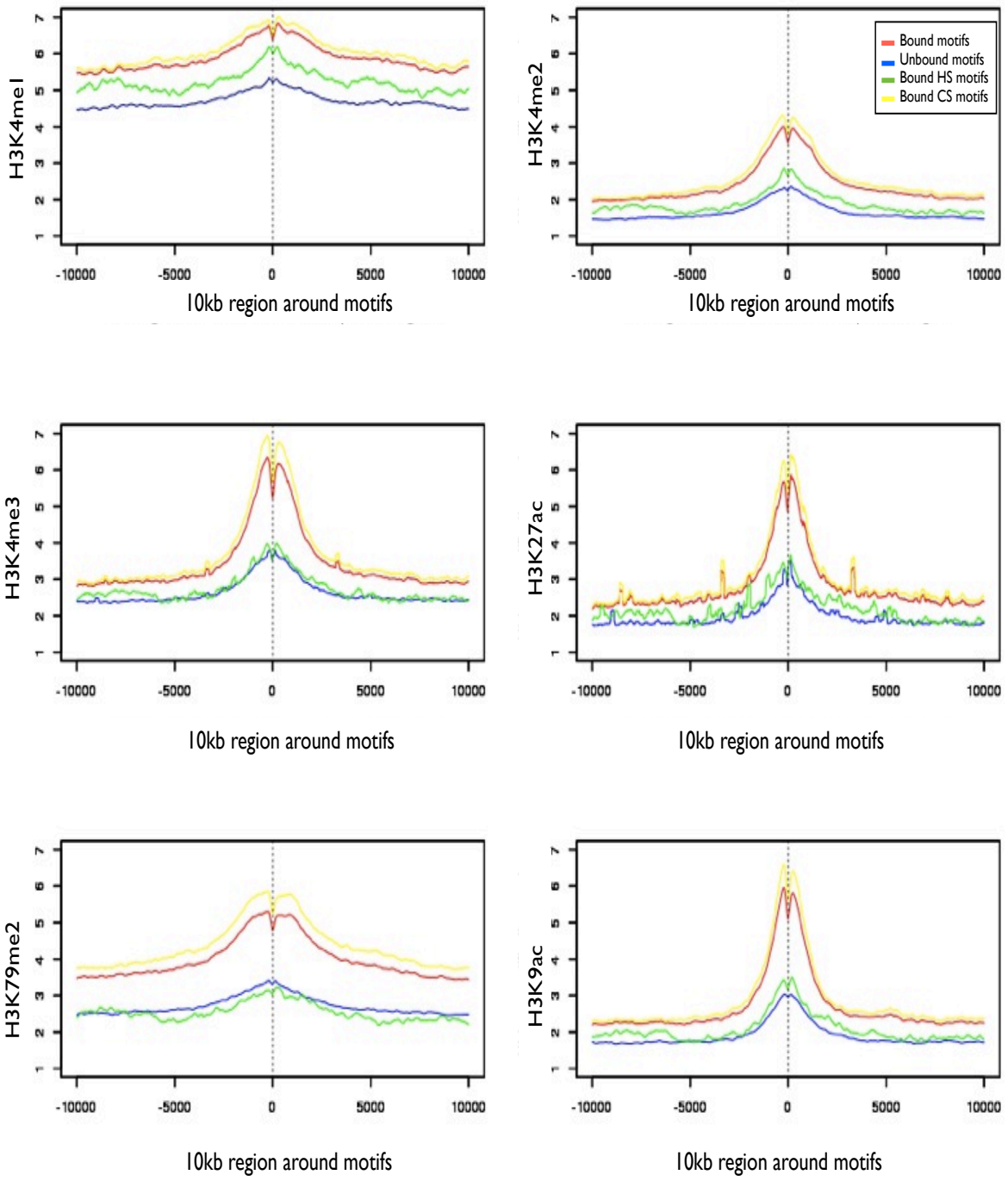


Figure 4.15: *Histone modifications marks at promoters surrounding PRDM9 bound, unbound, bound hotspot and bound coldspot 14-mer motifs. All are transcription activating marks.*

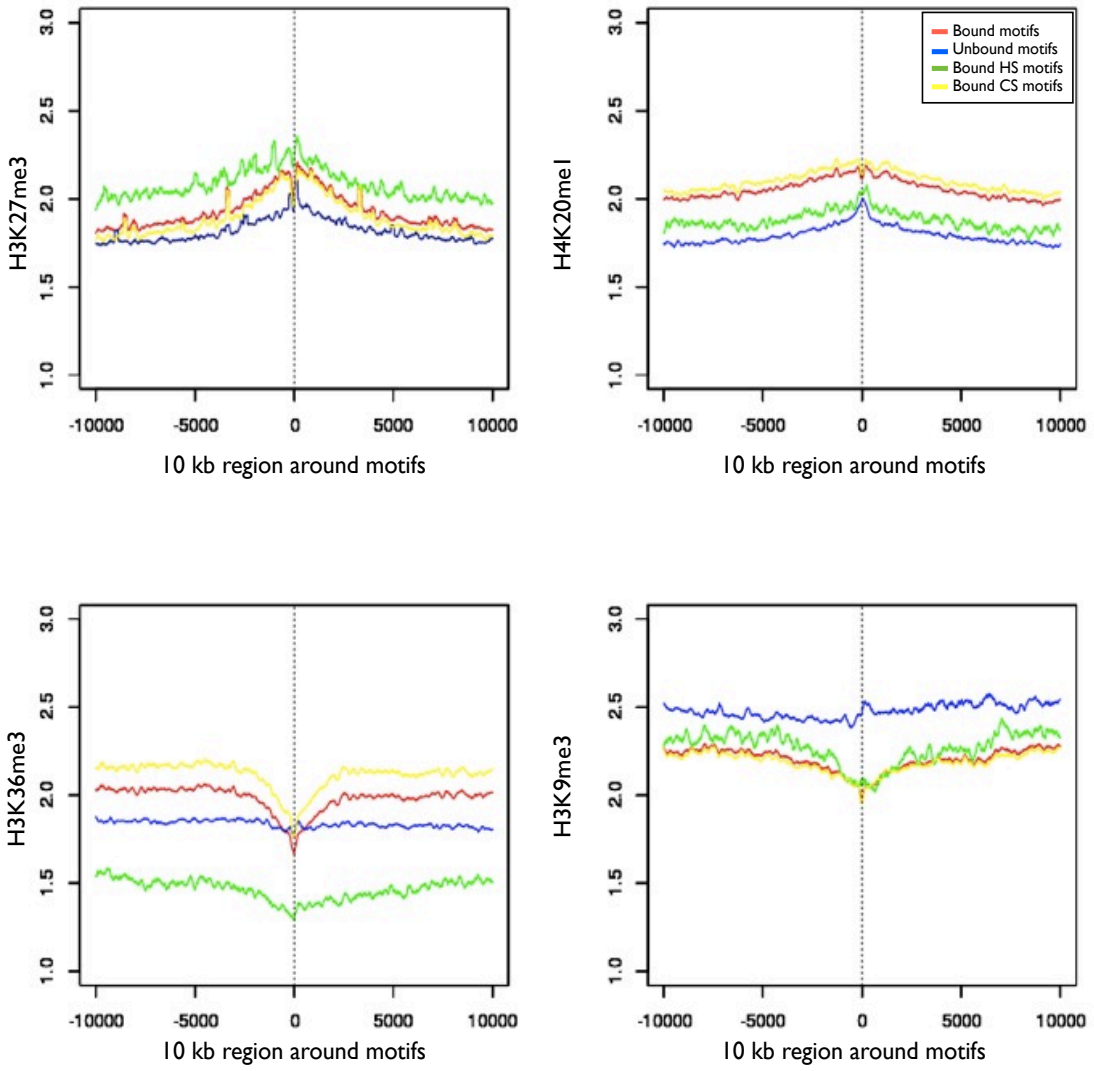


Figure 4.16: *Histone modifications marking gene bodies (H3K27me3, H3K36me3, H4K20me1) and constitutive heterochroamtin (H3K9me3) surrounding PRDM9 bound, unbound, bound hotspot and bound coldspot 14-mer motifs.*

coldspot cases. H3K9me3, a marker for constitutive heterochromatin, shows a less dramatic lowering in bound hotspot cases. Hence, almost all activating or transcribed chromatin marks showed a particular pattern, with bound motifs showing a peak, but bound motifs corresponding to hotspots, showing a much weaker peak, similar to genomic background [148]. In each of these cases, a dip at motif position itself might relate to a nucleosome positioning effect. Together, these results may hint towards a chromatin "code" for hotspot cases, suggesting that hotspots after binding can occur outside transcribed regions, but away from deeply inaccessible heterochromatin, which prevents binding (as seen in the case of the H3K9me3 level, which marks facultative heterochromatin, and is depleted in PRDM9 bound cases).

4.4 Transcription Factors

We also investigated if hotspots tend to be marked by other transcription factors near the motifs. We used all available ChIP-seq data for transcription factors on the gm12878 cell line taken from UCSC (Supplementary figures 9 and 10). Two interesting TFs that we looked at were CTCF and RAD21. Rad21 is one of the proteins that make up the cohesin complex [173, 174]; the cohesion complex is important for chromosome segregation and DNA repair [175]. CTCF is a protein known to interact with the cohesion complex and organises chromatin structure [176, 177, 178]. We observed that Rad21 signals were depleted at the unbound motif sites, but were enriched, locally, around PRDM9 bound motif cases. These signals were also enriched around bound hotspot and bound coldspot motif sites.

Further, contrary to our earlier findings (discussed in Chapter 3), we observed that CTCF associated peaks were enriched around PRDM9 bound hotspot cases. With our extended analysis, we found that CTCF peaks are also enriched around PRDM9 bound cases, and contrastingly, depleted in unbound cases. This implies that PRDM9 might require direct or cooperative binding with CTCF. It might also be that CTCF is simply a marker of chromatin accessible regions near the PRDM9 binding sites which subsequently become hotspots. The association between CTCF peaks and PRDM9 bound motif sites can further be explored by

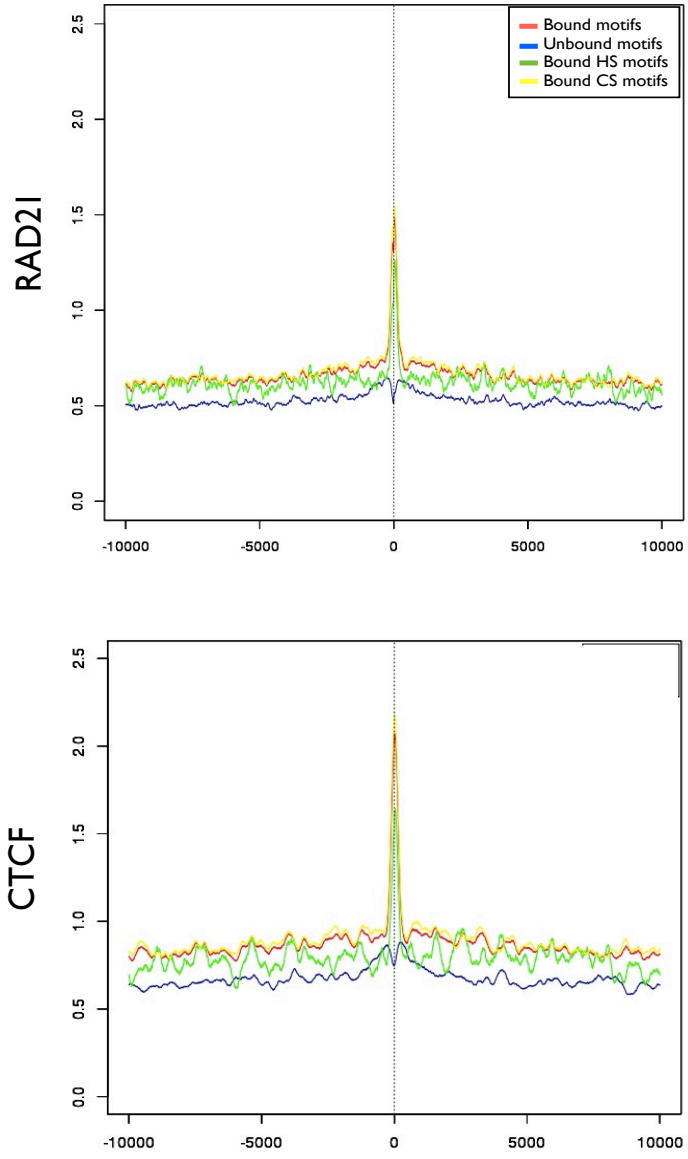


Figure 4.17: *Signals of selected transcription factors, RAD21 and CTCF, surrounding 14-mer motifs.*

looking at H3K4me3 signals in our transfected cell lines around CTCF enriched PRDM9 bound sites, which is planned future work.

4.5 Summary

Our ChIP-seq experiment helped us to uncover some interesting binding properties of PRDM9. We were able to identify over 170,000 PRDM9 binding sites in the genome, enriched for the 14-mer motif. The strength of the motif match, as well as the genomic background it is present in, are good predictors of PRDM9 binding. Our binding peaks overlapped with most known recombination hotspots.

We noted that PRDM9 confers the H3K4me3 mark at bound motif regions, and does not depend on a pre-existing H3K4me2 mark to add its trimethylation mark. The PRDM9 bound sites had elevated rates of recombination, as would be expected. Interestingly, binding was also enriched in promoters and exons, however, binding in these regions showed decreased recombination rates.

We observed that PRDM9 binding depends upon open chromatin and nucleosome positioning around the 14-mer motifs, whereas, subsequent hotspot formation is dependant on nucleosome rich, and relatively chromatin inaccessible regions compared to coldspots. Finally, our results also suggest that PRDM9 is able to bind in actively transcribing regions, however, hotspot formation tends to occur away from transcribed regions.

Chapter 5

Discussion

Our motivation for this research was, broadly, to understand the molecular mechanisms that influence meiotic recombination initiation in humans. Recombination is a crucial event that occurs in most sexually reproducing organisms in meiotic cells, during which DNA from homologous chromosomes is exchanged with DNA from its other homolog, forming a crossover. This crossover process is important for mechanical reasons (alignment during metaphase and proper segregation of chromosomes) and for the reason that this produces novel allelic combinations (given that recombinant chromosomes are a chimera of two parental chromosomes) [179, 180]. Therefore, trying to understand how this crossover process occurs, is key to understanding genetic diversity, and natural selection.

PRDM9, a zinc finger protein, has been identified as a major determinant of recombination hotspots (narrow 1-2kb regions with high crossover activity)[75, 95, 79]. PRDM9 consists of a DNA binding zinc finger domain, within which the number and type of zinc fingers determines the sequence specificity with which it binds to DNA. It also consists of a KRAB domain, which is possibly involved in protein-protein interaction, and a SET domain [87]. The SET domain is responsible for conferring the H3K4me3 mark, a mark which has been shown to be enriched around meiotic recombination hotspots in yeast and mammals [86]. Further, the predicted binding motif for PRDM9 matches the 13-mer motif identified previously, which is reported to be present in over 40% of all recombination hotspots [68]. The structure of PRDM9, therefore, consolidates its role in hotspot

specification, as has been established by multiple studies [53, 79, 80, 77, 81, 82].

We understand that PRDM9 activity is likely to occur upstream of the DSB formation and crossover process, as the distribution of DSBs is affected by the absence of PRDM9, in PRDM9 knockout mice [20, 89]. Given that almost all DSBs (and subsequently crossovers) are affected by the absence of PRDM9 [92, 83, 20, 75, 181, 40], it raises the question of whether this protein is responsible for marking all recombination hotspots. As mentioned previously, matches to the canonical 13-mer motif, predicted to be bound by PRDM9 are present in only a fraction (40%) of the hotspots [68]. Thus if PRDM9 does bind the motif to trigger recombination, what happens in cases of other hotspots which contain no clear evidence for this motif being present? Also, if PRDM9 really is involved in making recombination hotspots by binding the motif, then there are over 300,000 such motifs in the genome, a number which is over 10-fold greater than the reported recombination hotspots; if all these motifs can serve as potential binding sites for PRDM9, how does it preferentially bind to some and not others to make hotspots, and do other features play a role even given binding?

5.1 PRDM9 binding is sequence specific but also dependant on context

Firstly, with the help of our gel-shift assays, we were able to gain a better understanding of how PRDM9 interacts with DNA. We were able to confirm that PRDM9, as predicted, is able to bind specifically to the 13-mer motif. With the help of these assays we established that PRDM9 is able to strongly bind the LD based consensus sequence, a strong determinant of hotspot occurrence, referred to as the core motif (CCTCCCTNNCCAC), as well as the degenerate version of the motif (CCNCCNTNNCCNC) located in the THE1 and L2 repeat backgrounds (both these repeat backgrounds have been shown to exhibit elevated recombination rates in the presence of this motif). On the contrary, no binding was observed with a control sequence containing no motif match, indicating that PRDM9 is able to bind specifically to the canonical motif. The probes tested for

this analysis were about 40 bp in length, thus also containing the expanded motif shown previously [68], and also discussed in Chapter 4.

We were also able to confirm the specificity of PRDM9 binding by creating targeted disruptions in the 13-mer motif sequence. We observed that PRDM9 binding is strongly reduced when the probes designed from motif containing THE1 and L2 regions were changed at certain single positions within the motif. PRDM9 binding was seen to be sensitive to single base pair changes in the probes tested. It was, however, also interesting to note that certain changes in the motif, i.e. at a non-degenerate base in THE1 probe and degenerate base in L2 probe, were able to sustain and abolish PRDM9 binding, respectively. We would expect any changes at non-degenerate positions to disrupt binding and at degenerate positions to not affect binding, however, the digression from expected results may be explained by our more recent ChIP-seq results, where the PRDM9 binding motif shows that the non-degenerate position (base 5 of the motif) is the most degenerate of all non-degenerate positions, whereas the degenerate position (base 12 of the motif), is the least degenerate of all degenerate bases. This may also relate to the use of different zinc fingers and their different binding modes which we discuss later in this chapter. These results demonstrated that although the motif degeneracy is subtle, however, single base pair changes clearly strongly impact the ability of PRDM9 to bind.

Remarkably, on testing a number of different hotspots, previously thought not to contain a clear motif match for PRDM9 binding [79], it was seen that every one of these hotspots did in fact contain more degenerate versions of the 13-mer near the centre of the hotspot. Hence, PRDM9 is able to bind with all hotspots tested in-vitro (those containing exact matches to the 13-mer as well as those with the more degenerate matches), which raises the question that even though PRDM9 is capable of exhibiting binding affinity even for more the degenerate forms of the motif, why are only a fraction of such sequences, otherwise common in the genome, picked out for binding. As mentioned previously, we used probes of about 40bp for this analysis, however, it may be possible that additional bases are required for target specificity. This, along with other interacting chromatin

features may provide a plausible explanation for how PRDM9 selectively choses binding sites in the genome.

Our results have since, also been supported by various reports of sequence specific binding of PRDM9 by DeMassy's and Jeffreys' labs through in-vitro studies. Jeffreys et al. further showed that the type of motif bound by PRDM9 is clearly dependant on the variations in PRDM9 alleles [77, 79, 80]. However, no study has yet directly looked at where a single human PRDM9 allele is able to bind in the genome. It is interesting to note, however, that motif independent mechanisms also exist e.g. in yeast, whose hotspots are although regulated by an H3K4 methyltransferase Set1, do not have a DNA sequence motif associated with recombination activity (as Set1 has no DNA binding domain).

Following these experiments, we next aimed to understand if there are other chromatin related factors that promote PRDM9 activity and hotspot formation: i.e., does PRDM9 need some epigenetic or sequence related marks surrounding the motif site to facilitate or promote PRDM9 binding, or does it require a direct or indirect interaction with other transcription factors to be able to recruit recombination machinery, or possibly, a combination of both? Hence, to get a better understanding of sequence context and other chromatin features involved, we performed ChIP-seq for PRDM9, along with using various publicly available resources to attempt to answer these questions.

5.2 Mapping PRDM9 binding sites in the genome

With our ChIP-seq analysis to determine genome-wide binding sites for PRDM9, we were able to build the first map for PRDM9 binding sites in a human cell line (HEK293T). We identified over 170,000 binding sites for PRDM9 in the genome, and these binding sites overlapped with 53% of the HAPMAP recombination hotspots. It is possible that the hotspots that do not coincide with our binding peaks may not be accessible in the HEK293T cell line. It may also be that these hotspots are made independent of PRDM9, however, given the multiple sources of evidence implicating the role of PRDM9 in all hotspot cases, and our gel-shift

results, this seems very unlikely. Also, given that not all hotspots are strong binding sites, at least in HEK293T cells, and some have only poor motif matches, these are unlikely to be strongly bound even in meiotic tissue.

Interestingly, we noted that PRDM9 binding was enriched in promoter regions and exons. On exploring the recombination rates around promoter and non-promoter PRDM9 bound regions, we observed that the recombination rate in bound regions overlapping promoters appeared to be much lower than the rate in non-promoter, bound regions. This indicates that although PRDM9 binding is enriched in promoter regions, the recombination level at these regions remains low as compared to non-promoter regions, hence PRDM9 is unable to initiate recombination at some of its binding sites. Further, binding appears to be weaker in promoters and might be opportunistic due to open chromatin, thus, this may not be strong enough for recombination to occur. Other reasons could be the chromatin structure at these promoter regions which may promote recruitment of transcription machinery and thereby transcription, which in turn would repress recombination activity, evidence of which we see in our analysis. A previous report showed that when PRDM9 is not expressed, meiotic DSBs tend to occur at functional elements, like gene promoters, suggesting that PRDM9 is required to initiate recombination away from these functional elements [20], where hotspot activity is low [55]. Our results show that if PRDM9 is not only responsible for helping the recombination machinery to keep away from these functional elements, but that it is also able to bind to these elements while suppressing hotspot formation. As promoter regions are mostly marked by H3K4me3, it might be that PRDM9 is not able to confer its own H3K4me3 mark, thereby inhibiting recombination activity.

We used the MEME-ChIP software to find novel motifs in the PRDM9 bound regions and identified a 14-mer motif in the most strongly bound regions after masking for promoters and repeats; this newly found motif is essentially an exact match to the canonical 13-mer motif found, and predicted to bind with PRDM9 previously. On conditioning for a longer motif length extending over 25 bp, we found another longer motif of 33bp in length, containing the 14-bp motif as well

an upstream sequence which matches, again almost exactly, with the upstream sequence reported by Myers et al in 2008 by an LD based analysis of ancient hotspots, hence demonstrating that bound regions are able to capture the sequence feel of meiotic crossover hotspots.

Interestingly, more recent analysis on PRDM9 peak regions using an in-house motif finder has shown that the PRDM9 binding site is in fact represented by 6 motifs that account for deletions or insertions relative to the canonical motif (see figure 5.1). This analysis represents the binding complexity of PRDM9. The motif analysis we performed earlier, using MEME-ChIP, does not account for indels, hence the new motif analysis using the in-house algorithm (work done by Simon Myers) shows that the canonical motif can be refined into 6 motifs (found enriched at the centre of the PRDM9 peaks), where each of these motifs implies binding by a different subset of zinc fingers, and also implying a strong role for the upstream zinc fingers binding upstream of the canonical motif. Accounting for these 6 motifs, we see that the motifs overlapping the strongest PRDM9 peaks increase to about 95%. Another interesting finding from recent work is that about 10% of PRDM9 peaks contain the CTCF binding motif, but no match to the canonical motif. This observation is consistent with our previous analysis, where CTCF binding sites were enriched around bound motifs. This might imply that PRDM9 is able to bind directly or indirectly to CTCF.

Previous work showed that genomic context is important in determining hotspot formation [55, 68]. We know that hotspots range in the activity, being most active in THE1 background, followed by L2 and non-repeat backgrounds, and finally ALU being least active in hotspot activity. We explored the presence of 14-mer motifs bound by PRDM9 in various genomic backgrounds. PRDM9 bound motifs occur more frequently in non-repeat backgrounds, whereas among repeat regions they were enriched in the THE1 repeat elements. Bound motifs in non-repeat and THE1 repeat regions were found likely to form hotspots. We observed the same enrichment patterns in strongly bound PRDM9 motifs. Bound and unbound motifs in hotspots on being stratified by motif scores (representing strength of motif match) and by different genomic backgrounds, showed that

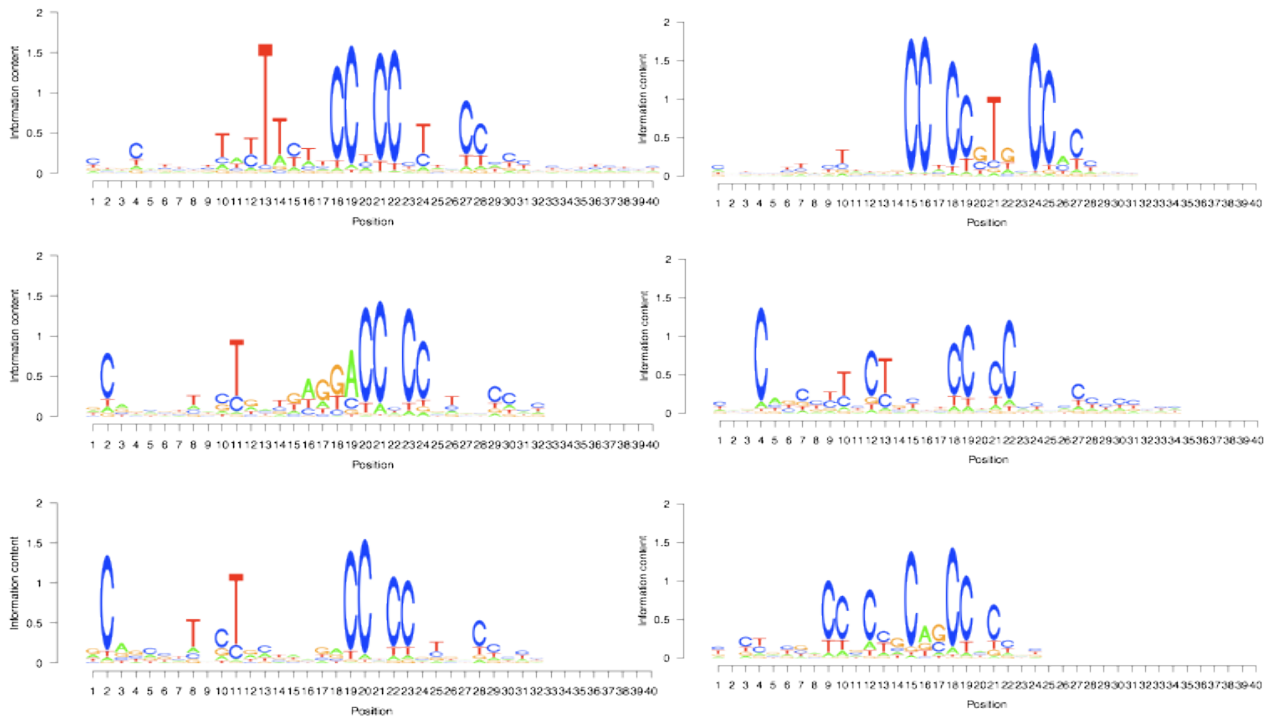


Figure 5.1: ***Six motifs discovered by an In-house motif finder after accounting for internal spacing.*** *Motif analysis and plot made by Simon Myers.*

bound hotspots are more enriched in higher scoring motif matches. This analysis showed that both background context and strength of match are predictors of PRDM9 binding and subsequently making a recombination hotspot. Overall, our ChIP-seq results agree with previous results by explaining that influence of background is due to the preferential binding of PRDM9 in these backgrounds, even after accounting for motif strength. This means that motif strength is predictive but not sufficient to cause hotspots. The extended motif is likely to be part of the reason, along with other chromatin related factors, which we explored next. Although these various features were examined in a non-meiotic cell line, our findings are likely to also hold true for meiotic cells, given that various chromatin features e.g. H3K4me3, nucleosome positioning etc., that are present around recombination hotspots in meiotic cells, have also been found to hold in mitotic cell lines [84, 141, 113, 142].

5.2.1 Chromatin inaccessible regions also contain 14-mer motifs

We wanted to explore if the level of chromatin accessibility is associated with the type of motif present. We hypothesised that in less chromatin accessible regions, PRDM9 is likely to need a strong match to the motif to bind it, however, in more chromatin accessible regions, we might expect PRDM9 to be able to bind to more degenerate versions of the motif. In agreement with this hypothesis we observed a match to the 14-mer in the less chromatin accessible regions. However, interestingly, we observed that peaks in more chromatin accessible regions were enriched for the CTCF motif. This signal seemed to get stronger after masking for promoters, and explains peaks (even after allowing for the 6 additional PRDM9 binding motifs discussed about earlier). We also found binding sites for another transcription factor, AP-1, marking some peaks in accessible chromatin regions.

We also noted that PRDM9 binding peaks in hotspots with low motif probability (taken from a previous analysis) did not correspond to higher chromatin accessibility. To test what sequence PRDM9 is likely to bind in these hotspots

containing motifs with low probability and also less DNase accessibile, we ran this subset of these positions through MEME for motif finding. Remarkably, we again found a close match to the 14-mer motif in 63% of the cases, even from this set of low motif probability cases. This implied that even though some of these motif cases were not considered as a strong match for PRDM9 to bind, they were still in fact good enough matches for PRDM9 to bind with. The upstream sequence discovered in these cases, again appeared to be an important feature promoting PRDM9 binding. Hence, with this analysis we were able to explain hotspots that were previously difficult to understand, by identifying 'hidden motifs' i.e. most of the hotspots can be explained by the more degenerate versions, if not exact matches to the motif, which can serve as PRDM9 target sites.

We also performed ChIP-seq for H3K4me3 (transfected and untransfected) and H3K4me2 on untransfected HEK293T cells, to enable us to understand questions like: which H3K4me3 marks in the PRDM9 transfected cells are made by PRDM9 and which ones are made independent of PRDM9, and also whether an H3K4me2 mark is a pre-requisite substrate for PRDM9 to add the trimethylation mark to. An initial screening of this data showed that at many of the hotspot regions tested, the H3K4me3 mark in transfected cells clearly overlaps the PRDM9 binding peaks centred at these hotspots. As no such enrichment of H3K4me3 was observed in the corresponding regions in untransfected cells, this demonstrates the mark is induced by PRDM9 binding. Further, there was also no enrichment observed in the H3K4me2 mark in most PRDM9 bound regions, which implies that the H3K4me3 mark conferred by PRDM9 does not necessarily need dimethylation as a substrate.

5.3 Local chromatin environment around PRDM9 binding sites

5.3.1 Chromatin accessibility

We initially investigated the affect of chromatin accessibility around 13-mer motifs in recombination hotspots and compared with coldspots. DNase hypersensitivity

data from lymphoblastoid cell lines showed an enrichment of DNase hypersensitive sites surrounding the 13-mer motifs; this signal was found to be enriched around both recombination hotspot and coldspot motifs, hence suggesting that it is a feature of motif containing DNA sequence, more than relating to binding itself and persists after accounting for sequence biases. However, the hotspot to coldspot ratio plot showed that the level of open chromatin is lower in the region immediately surrounding the motifs in hotspots. This reduced levels of open chromatin may be suggestive of nucleosome enriched regions in hotspots. These results are comparable with some other reports suggesting that open chromatin is positively correlated with double strand break hotspots in yeast [141, 182] and mice [111, 180] ; the open chromatin structure is likely to be required for the recruitment of recombination machinery. Interestingly, however, there have been individual occurrences of hotspots reported not to be associated with open chromatin. In mice, for example, the *psmb9* hotspot does not contain a DNase hypersensitive site [111], and in most hotspot motifs, we do not observe an individually strong DNase peak either.

We further extended this analysis using PRDM9 binding information from the ChIP-seq assay to understand if chromatin accessibility is a feature that is a prerequisite for PRDM9 binding prior to hotspot formation, as we were now able to stratify our motifs by PRDM9 bound, unbound, bound hotspot and bound coldspots. This stratification of data showed that DNase accessible sites are found around all PRDM9 bound motifs, at a level much more enriched (over 3.5-fold) compared to coldspots, which on the contrary, appear to show decreased levels at the motif sites. However, bound motifs that subsequently form crossover hotspots were depleted in open chromatin regions compared to coldspots. These results confirm that in case of hotspot motifs, chromatin accessibility is lower than coldspot cases. Together these results demonstrate that PRDM9 binding occurs in chromatin accessible regions, however, it is likely to lead to hotspot formation if bound regions are present in, presumably, nucleosome rich regions of lower chromatin accessibility.

5.4 Nucleosomes

The next follow-up question led us to investigate evidence of nucleosomes around motif sites. We observed that nucleosomes are positioned around both bound and unbound motifs. Given that nucleosomes are positioned even in unbound motifs, it would mean that it is not PRDM9 binding itself that positions nucleosomes; hence, the positioning is independent of binding and is rather a feature of the motifs theme selves, possibly because of their being GC rich. However, looking at nucleosome positioning around bound hotspot and bound coldspot motifs, we found that PRDM9 bound regions were likely to become hotspots if they are in more nucleosome rich regions. These results suggested (and also confirmed previous DNase analysis) that once PRDM9 is bound, hotspot formation is likely to occur in more nucleosome dense regions. Nucleosome positioning around recombination hotspots has previously been reported by Smagulova et al. in mice [75].

5.5 Histone marks

We also investigated histone marks as potential features promoting recombination hotspot activity. On investigating various histone modification marks, we observed that PRDM9 is able to bind in transcriptionally active regions (seen by PRDM9 bound motifs showing an elevation in transcription activating marks), however, binding is inhibited in heterochromatin regions e.g. the H3K9me3 mark is found to be low in PRDM9 bound cases, compared to unbound cases, implying that this mark might play a role in suppressing PRDM9 binding along with subsequent downstream events.

By dissecting PRDM9 bound and unbound cases, we had the ability to understand the downstream events. All modifications marking promoters and gene bodies that function to activate transcription were enriched around bound coldspot motifs but not around bound hotspot motifs. On the other hand, we observed that in case of each of the transcription repressive marks examined, hotspots were enriched for these signals compared to coldspots. These results suggest that in

humans, recombination activity is likely to be repressed around actively transcribing regions.

There have been some reports in yeast that have also shown a negative correlation between transcription activating marks like H3K36me3 and H3k79me3 and recombination. These are two of the histone modifications we looked at, that mark the body of genes, and showed reduced signals around crossover hotspot motif cases compared to coldspot motifs. H3K79 is associated with DNA repair and is marked by a methyltransferse called Dot1p. The deletion of this methyltransferase leads to an inefficient repair of double strand breaks by sister chromatids. This might explain the positive association between H3K79me3 and reduced level of recombination activity, owing to this mark's role in DNA repair [183, 184]. Similarly, H3K36 is suggested to be indirectly involved in affecting DSB frequency, by stimulating other histone marks that are able to directly induce DSBs. One such study on the HIS4 hotspot in yeast, reports that H3K36me3 represses DSB formation by recruiting a histone deacetylase called Rpd3. Hence, an indirect and negative correlation between H3K36me3 and DSB frequency may be observed when the deacetylase reduces acetylation, which in turn may condense chromatin structure thereby reducing DSB frequency [185, 186, 187, 188]. In our case, although we see reduced levels of acetylation marks around crossover hotspot motifs, there is also an enrichment of acetylation marks in coldspot motif cases, which might implicate the role of transcription itself which is antagonistic.

There have not, however, been many reports that demonstrate a positive or negative correlation between recombination and transcription repressing histone marks, however, Buard et al., showed that in one of the mouse strains (b6) with an inactive *psmb9* hotspot, the H3K9me3 and H3K27me3 repressive marks were seen to be depleted at that hotspot [15, 122].

5.6 Transcription Factors

We also investigated the relationship of hotspots with various transcription factors, and found most of the TF binding sites to be enriched around coldspot motifs. This is probably because of the presence of promoters at these sites which make them more likely to become coldspots, hence there is no implication of a causal relationship. An interesting signal was that of CTCF binding sites that were clearly peaked around bound motifs and also in bound hotspot and coldspot cases, but were depleted at the unbound motifs. As CTCF motifs do not mark all hotspots specifically, one explanation for this observation could be that CTCF is a marker for places where the 14-mer motif is found in hotspots and where PRDM9 can access its binding sites. This could be suggestive of CTCF being directly associated with PRDM9 or of cooperative binding, which would explain PRDM9 peaks where there is only a CTCF motif enriched at the centre of the peaks (mentioned earlier); however, these regions (containing only a CTCF motif but no 13-mer motif) do not form hotspots. The association of PRDM9 and CTCF could further be investigated by looking at H3K4me3 marks in peaks enriched for CTCF motif.

5.7 Conclusion

In conclusion, with this research we have been able to identify the characteristics of PRDM9 through our ChIP-seq assay, along with certain genomic features associated with hotspot activity in humans. We were able to prove by in-vitro and in-vivo binding methods that PRDM9 binds sequence specifically to the canonical motif. With our ChIP-seq assay we were able to present the first map of PRDM9 binding sites in humans, enriched for a 14-bp motif. We found that PRDM9 is able to bind gene regulating regions (promoters and exons), however these binding sites exhibit low recombination activity, relative to other PRDM9 bound non-promoter sites in the genome. PRDM9 binding was also enriched in non-repeat and THE1 repeat elements, with the strength of the canonical motif match being important, but not sufficient in predicting binding and subsequent

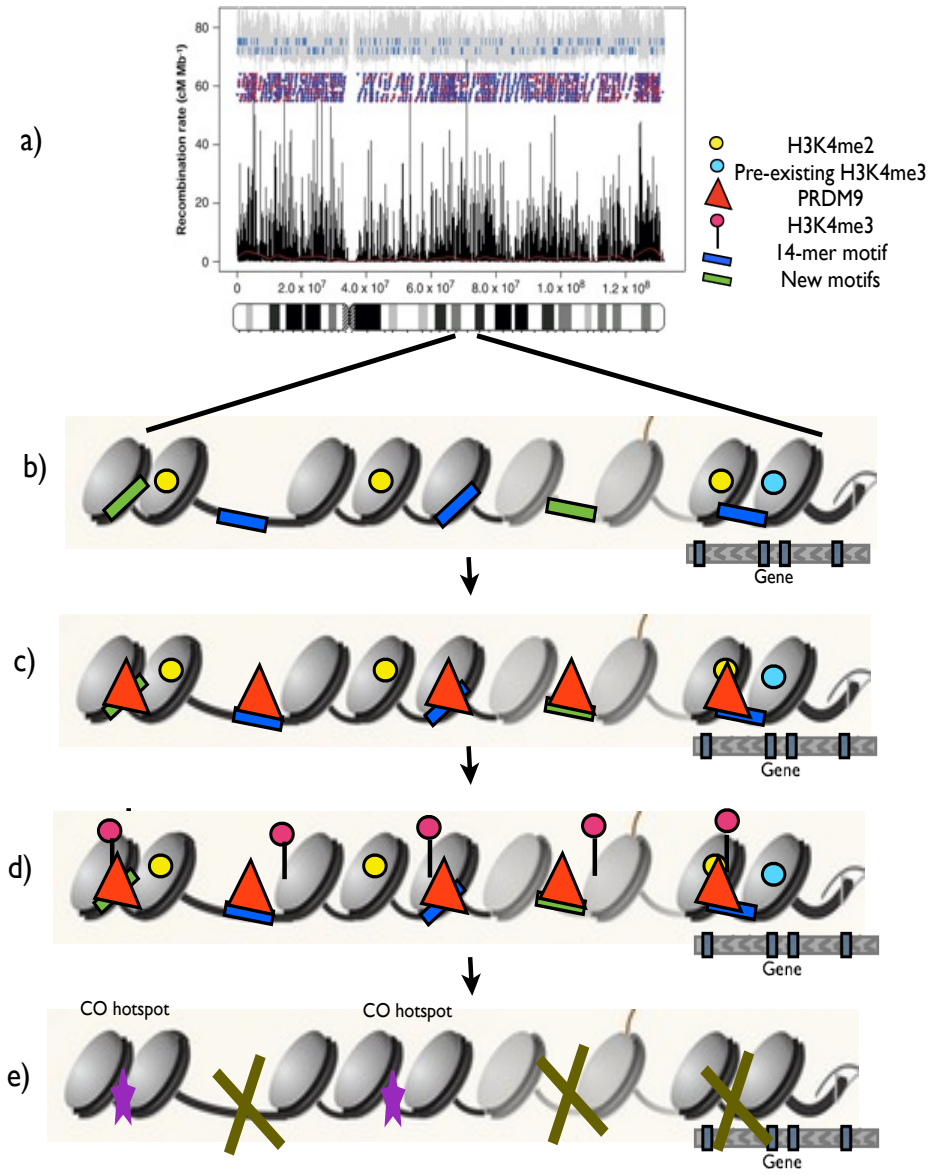


Figure 5.2: ***Model for PRDM9 binding and hotspot formation.*** *PRDM9 binds with 13-mers, and the newer motifs identified by accounting for indels; binding is more likely in nucleosome rich but chromatin accessible regions. On binding, PRDM9 makes an H3K4me3 mark, and does not require a pre-existing H3K4me2 mark as substrate. This in turn, recruits recombination machinery and forms crossover hotspots. Hotspot formation is favoured in nucleosome dense regions and away from transcribed regions.*

hotspot formation.

We found that PRDM9 is responsible for all H3K4me3 marks in the bound regions and does not necessarily depend upon the presence of an H3K4me2 mark to add its trimethylation mark. We investigated the chromatin environment around the PRDM9 bound sites and found that chromatin accessible regions, almost always containing the canonical motif, but within a rich nucleosome environment, away from dense constitutive heterochromatin, are important predictors of PRDM9 binding. We also found that once PRDM9 is bound, crossover hotspot formation is likely to occur in more nucleosome dense and less chromatin accessible sites, that are located away from transcribing regions (see figure 5.2).

Hence, our results have shown that PRDM9 binding in the genome is dependent on both the primary sequence and surrounding epigenetic factors, and together these factors promote binding and the positioning of crossover hotspot locations in the human genome.

Supplementary Material

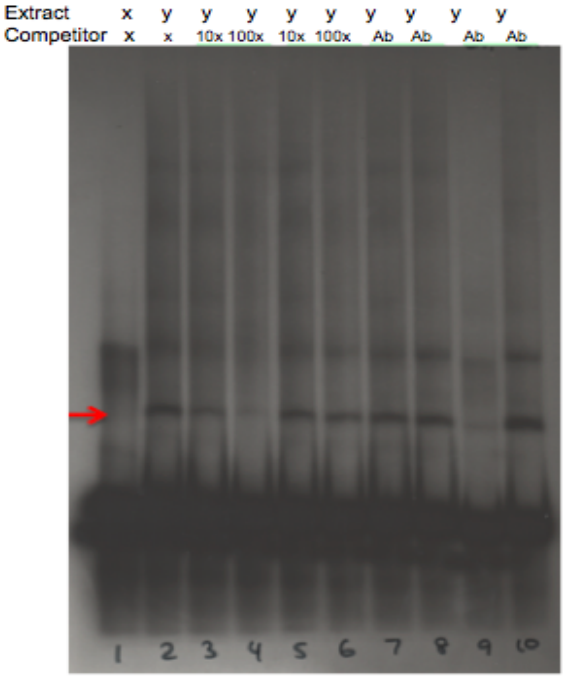


Figure 1: EMSA with testis nuclear extract. Red arrow indicating protein-DNA complex of interest. Lane1: THE1 probe only, Lane 2: THE1 probe with nuclear extract, Lanes 3 and 4: 10 and 100 fold excess THE1 unlabelled competitor, Lanes 5 and 6: 10 and 100 fold excess of Cold-1 unlabelled competitor, Lanes 7 and 8: Antibody lanes; Testing different concentrations of Ab after adding nuclear extract to optimise supershift conditions. Lanes 9 and 10: Antibody lanes; Testing different concentrations of Ab before adding nuclear extract for optimizing conditions.

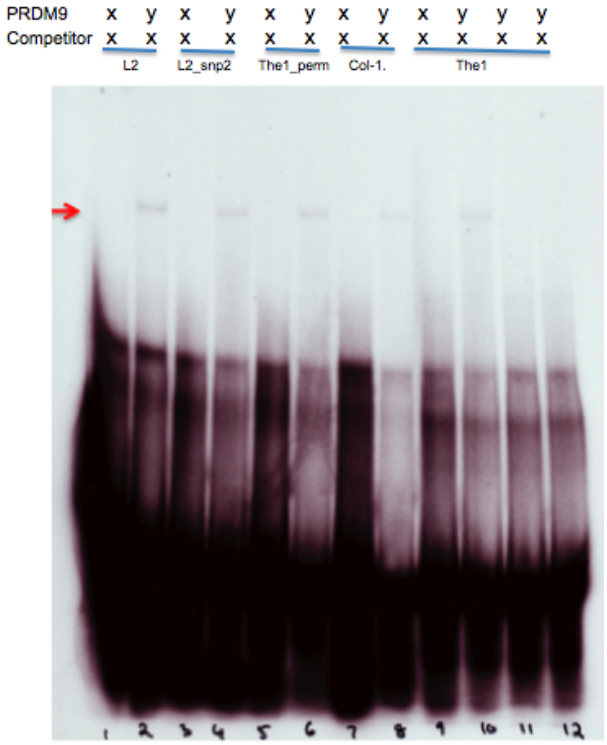


Figure 2: EMSA with company synthesized recombinant PRDM9. Red arrow indicating protein-DNA complex of interest. Lane1: L2 probe only, Lane 2: L2 probe with PRDM9, Lanes 3: L2-snp2 probe only, Lane 4: L2-snp2 probe with PRDM9, Lane 5: THE1-permuted probe only, Lane 6: THE1-permuted probe with PRDM9, Lane 7: Cold-1 probe only, Lane 8: Cold-1 probe with PRDM9, Lane 9: THE1 probe only, Lanes 10-12: Testing THE1 probes with batches 1,2 and 3 of PRDM9 respectively.

Figure 3: cDNA and amino acid sequence for human PRDM9 protein referenced in databases (Ensembl release 56, based on the Genome Reference Consortium GRCh37) containing 13 zinc fingers

No.	Construct	Primer (fwd)/ restriction site	Primer (rev)/ restriction site	Template	Cloning startegy
1	C-terminal His tagged PRDM9	gacacgCAATTGGCCACC atgAGCCCTGAAAAGTCC CAAGAGGAGAG(MfeI)	tgattCTCGAGTCATTA gtggtgatggtgatgatggtggt gcttggtaccCTCATCCTC CCGGCACACGTAGG (kpn1)	PRDM9	-Cut insert with mfe1/xho1 -Clone in pBacPak9 (cut with ecor1/xho1)
2	C-terminal monovenus	-	-	-	Subclone monovenus from cd45mV vector into C-term His tagged construct (cut with xho1/kpn1)
3	N-terminal His tagged PRDM9	gacacgCAATTGGCCACC atggtcaccatcaccac catcacAGCCCTGAAAAG TCCCAAGAGGAGAG	CgcctcgagtcattaCTCAT CCTCCCGGCACACG TAGG	PRDM9	-Cut with mfe1/xho1 -Clone in pBacPac9 (cut with ecor1/xho1)
4a	pBacPac monoVenus	gacacgCAATTGGCCACCat ggtcaccatcaccaccatcacATG GTGAGCAAGGGCGAGG AGCT catcacATGGTGAGCAAG GGCGAGGAGCT (Mfe1)	cgcggtaccCTTGTACA GCTCGTCCATGCCG AGA	Monovenus	-Cut with mfe1/kpn1 -Clone in pBacPac9 (cut with ecor1/kpn1)
4	N-terminal monovenus tagged	CgcggtaccAGCCCTGAAAA GTCCCAAGAGGAGAG	tgattCTCGAGTCATTA CTCATCCTCCCGGC ACACGTAGG	PRDM9	-Cut with xho1/kpn1 -Clone in pBacPac9-monovenus (cut with xho1/kpn1)

Figure 4: PRDM9 cloning plan with pBacPac9 vector into insect cell system

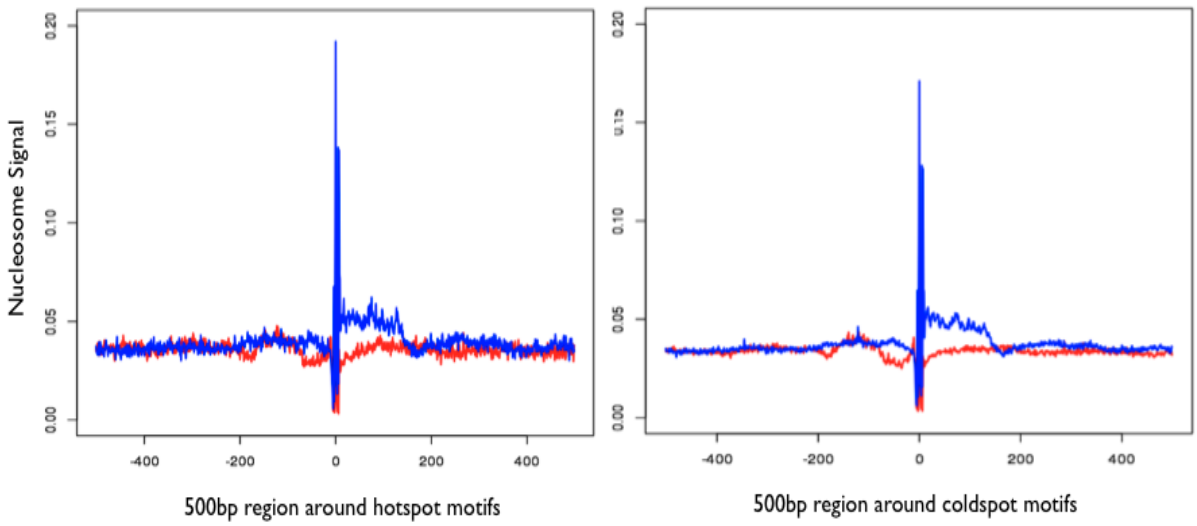


Figure 5: Nucleosome positioning around 13-mer motifs. MNase cuts in a 500bp region around 13-mer motifs in hotspots and coldspots, separated by plus (red line) and minus (blue line) strand.

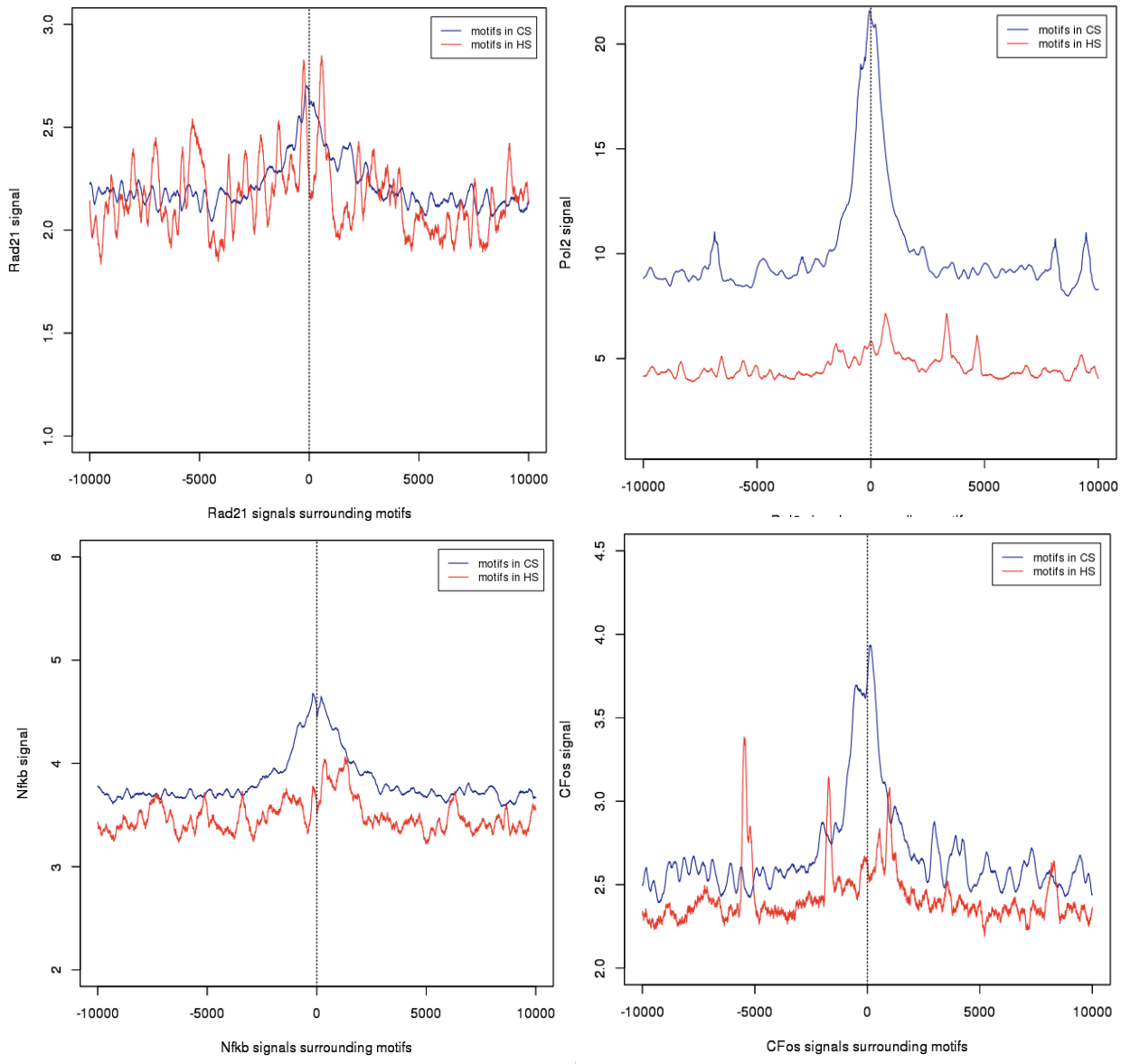


Figure 6: Signals of transcription factors (Rad21, Pol2, Nfkb and CFos), surrounding 13-mer motifs (hg18).

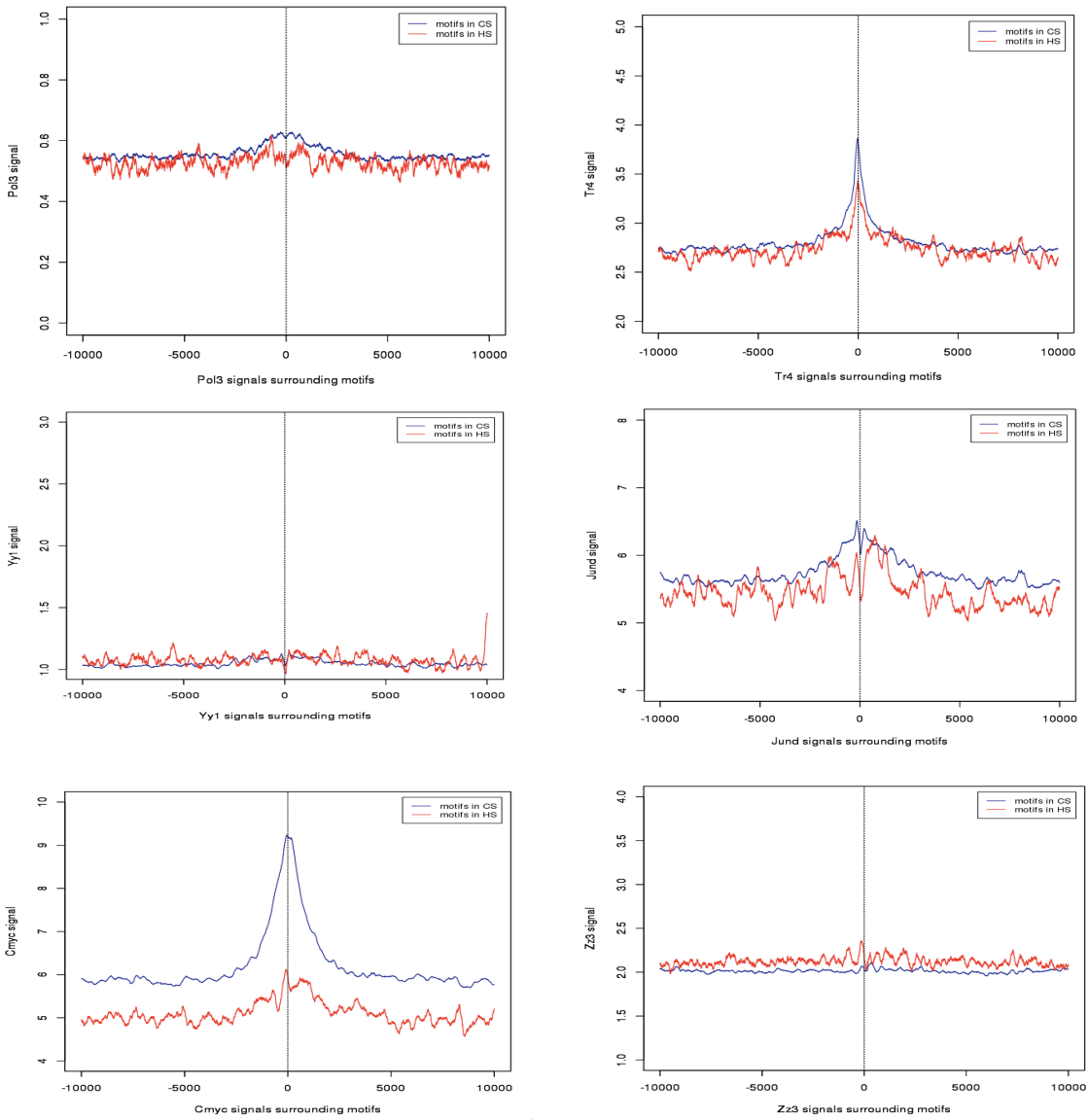


Figure 7: Signals of transcription factors, surrounding 13-mer motifs (hg18).

No.	Construct	Primer (fwd)/ restriction site	Primer (rev)/ restriction site	Template	Cloning strategy
1	C-terminal His	gacacgCAATTGGCCACC atgAGCCCTGAAAAGTCC CAAGAGGAGAG(MfeI)	cgcggtaccCTCATC CTCCCGGCACACGT A GG (kpn1)	PRDM9	-Cut insert with mfe1/kpn1 -Clone in pHLsec (cut with ecor1/kpn1)
2	C-terminal GFP	gacacgCAATTGGCCACC atgAGCCCTGAAAAGTCC CAAGAGGAGAG(MfeI)	cgcggtaccCTCATC CTCCCGGCACACGT A GG (kpn1)	PRDM9	-Cut with mfe1/kpn1 -Clone in pHLsec GFP (cut with ecor1/kpn1)
3	N-terminal His	gacacgCAATTGGCCACC atggtcaccatcaccac catcacAGCCCTGAAAAG TCCCAAGAGGAGAG	Cgcggtacctcat taCTCATCCTCCC GGCACACGTAGG Reverse, stop codons)	PRDM9	-Cut with mfe1/kpn1 -Clone in pHLsec (cut with ecor1/kpn1)
4a	His in GFP	gacacgCAATTGGCCACC atggtcaccatcaccac catcacATGGTGAGCAAG GGCGAGGAGCT (Mfe1)	cgcaccggtCTTGT ACAGCTCGTCCATG CCGAGA (Age1)	Venous	-Cut with ecor1/Age1 -Clone in pHLsec (cut with ecor1/Age1)
4	N-terminal GFP	cgcaccggtAGCCCTGAA AAGTCCCAAGAGGAGAG (Age1)	Cgcggtacctcat taCTCATCCTCCC GGCACACGTAGG (Reverse, stop codons)	PRDM9	-Cut with Age1/kpn1 -Clone in pHLsec GFP (cut with Age1/kpn1)

Shaded in green is start codon, red is stop codon, blue is restriction site, yellow is reverse and grey is His tag

Figure 8: PRDM9 cloning plan with pHLsec vector into mammalian cell system

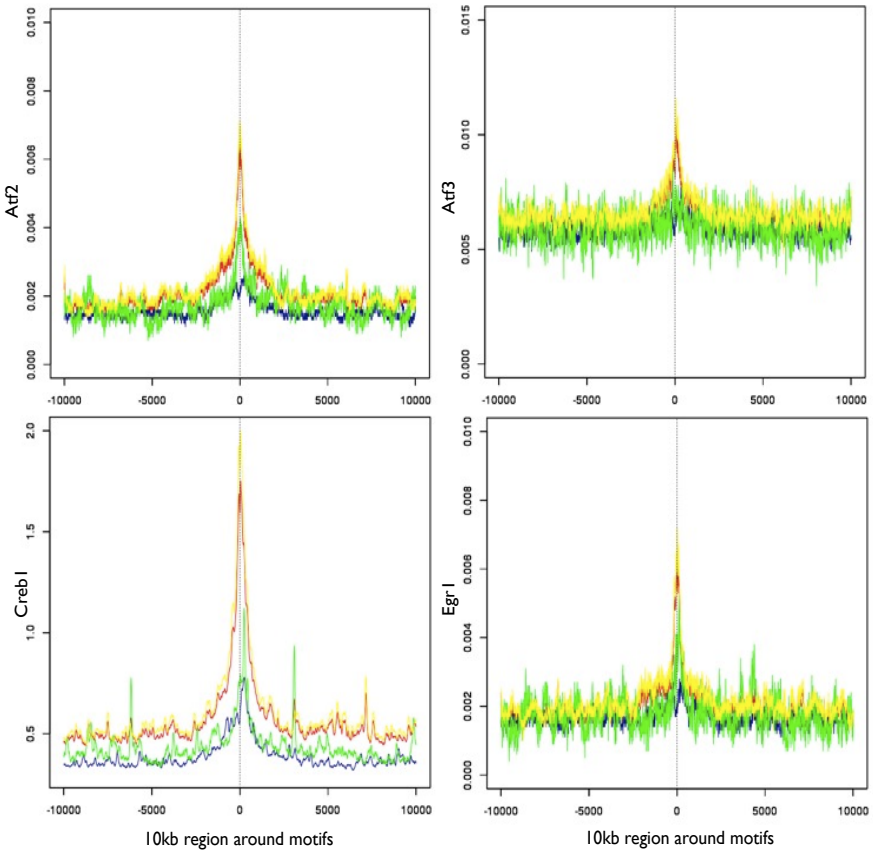


Figure 9: Signals of transcription factors, surrounding motifs (hg19). Red line: Bound motifs, Blue line: Unbound motifs, Green line: Bound hotspots, Yellow line: Bound coldspots.

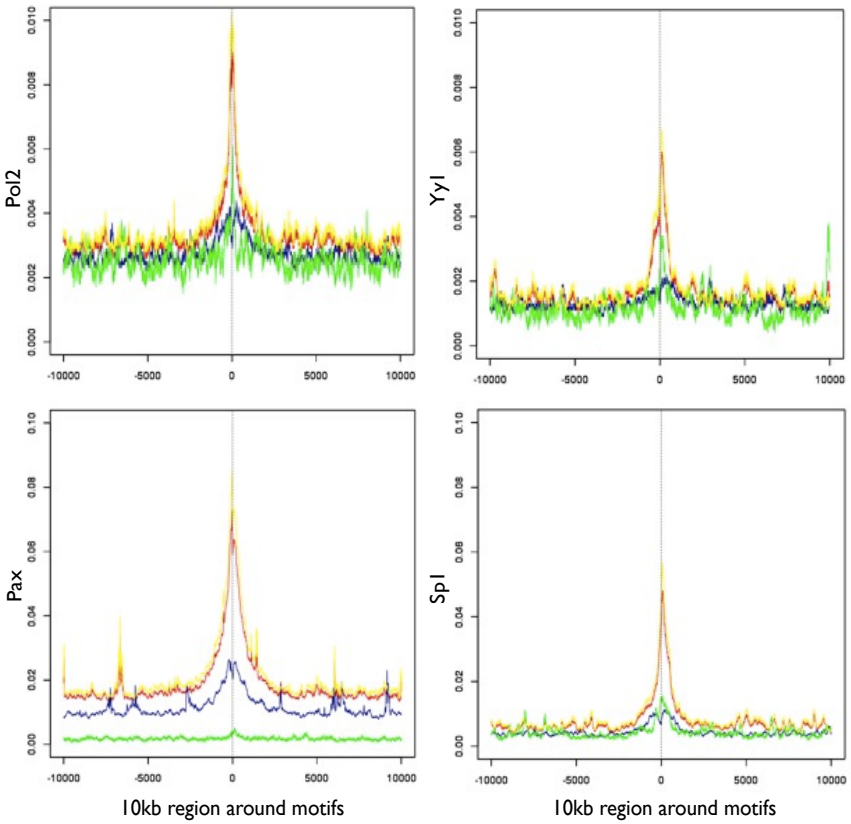


Figure 10: Signals of transcription factors (hg19) surrounding motifs. Red line: Bound motifs, Blue line: Unbound motifs, Green line: Bound hotspots, Yellow line: Bound coldspots.

References

[1] S Henry S Theodor and JS Matthias. *Microscopical researches into the accordance in the structure and growth of animals and plants*. London, The Sydenham Society, 1847. 1

[2] B Kish. Forgotten leaders in modern medicine: Valentin, *Gruby*, *Remak*, *Auerbach*. *Transactions of the American Philosophical Society*, 44:139–317, 1954. 1

[3] GA Silver. Virchow, the heroic model in medicine: health policy by accolade. *American Journal of Public Health*, 77(1):86, 1987. 1

[4] N Paweletz. Walther *Flemming*: pioneer of mitosis research. *Nat Rev Mol Cell Biol*, 2(1):72–5, Jan 2001. 1

[5] V Orel. The "useful questions of heredity" before *Mendel*. *J Hered*, 100(4):421–3, 2009. 1

[6] G Hamoir. The discovery of meiosis by e. *Van Beneden*, a breakthrough in the morphological phase of heredity. *International Journal of Developmental Biology*, 36(1):9–15, 1992. 1

[7] PK Stanford. August *Weismann's* theory of the germ-plasm and the problem of unconceived alternatives. *Hist Philos Life Sci*, 27(2):163–99, 2005. 2

[8] KR Benson. T. *H. Morgan's* resistance to the chromosome theory. *Nat Rev Genet*, 2(6):469–74, Jun 2001. 2

[9] JL Gerton and RS Hawley. Homologous chromosome interactions in meiosis: diversity amidst conservation. *Nat Rev Genet*, 6(6):477–87, Jun 2005. 2

[10] MA Handel and JC Schimenti. Genetics of mammalian meiosis: regulation, dynamics and impact on fertility. *Nat Rev Genet*, 11(2):124–36, Feb 2010. 2

[11] N Kleckner. Meiosis: How could it work? *Proceedings of National Academy of Science*, 93:8167–8174, 1996. 2

[12] M Lichten and B de Massy. The impressionistic landscape of meiotic recombination. *Cell*, 147(2):267–70, Oct 2011. 2, 3, 4, 22

[13] K Nasmyth M Petronczki, MF Siomos. Un ménage à quatre: the molecular biology of chromosome segregation in meiosis. *Cell*, 112(4):423–40, Feb 2003. 2, 3

[14] AS Wilkins and R Holliday. The evolution of meiosis from mitosis. *Genetics*, 181(1):3–12, 2009. 3

[15] R Kumar and B De *Massy*. Inititation of meiotic recombination in mammals. *Genes*, 1:521–549, 2012. 3, 4, 5, 8, 135

[16] F Baudat C Grey and B de Massy. Genome-wide control of the distribution of meiotic recombination. *PLoS Biol*, 7(2):e35, Feb 2009. 3, 16

[17] JP Lao and N Hunter. Trying to avoid your sister. *PLoS Biol*, 8(10):e1000519, 2010. 3

[18] CN Giroux S Keeney and N Kleckner. Meiosis-specific *DNA* double-strand breaks are catalyzed by *Spo11*, a member of a widely conserved protein family. *Cell*, 88(3):375–84, Feb 1997. 4, 23

[19] S Keeney. Spo11 and the formation of *DNA* double-strand breaks in meiosis, 2008. 4

[20] P Khil DR Camerini-Otero K Brick, F Smagulova and GV Petukhova. Genetic recombination is directed away from functional genomic elements in mice. *Nature*, 485(7400):642–5, May 2012. 4, 8, 21, 26, 79, 125, 128

[21] GM Chin M Hayashi and AM Villeneuve. C. elegans germ cells switch between distinct modes of double-strand break repair during meiotic prophase progression. *PLoS Genet*, 3(11):e191, Nov 2007. 5

[22] ML Sanderson. Proteins involved in meiotic recombination: a role in male infertility? *Systems Biology Reproductive Medicine*, 54(2):57–74, 2008. 5, 8

[23] Jing MJ Neale, J Pan and S Keeney. Endonucleolytic processing of covalent protein-linked *DNA* double-strand breaks. *Nature*, 436(7053):1053–7, Aug 2005. 5

[24] Scott F Cole, S Keeney and M Jasin. Comprehensive, fine-scale dissection of homologous recombination outcomes at a hot spot in mouse meiosis. *Mol Cell*, 39(5):700–10, Sep 2010. 5

[25] S Keeney F Cole and M Jasin. Evolutionary conservation of meiotic *DSB* proteins: more than just *Spo11*. *Genes Dev*, 24(12):1201–7, Jun 2010. 5

[26] KJ Schimenti LA Wilson-DM Cooper E Brignull MA Handel DL Pittman, J Cobb and JC Schimenti. Meiotic prophase arrest with failure of chromosome synapsis in mice deficient for *Dmc1*, a germline-specific reca homolog. *Mol Cell*, 1(5):697–705, Apr 1998. 5

[27] G Y Matsuda T Habu Y Nishimune K Yoshida, G Kondoh and T Morita. The mouse *RecA*-like gene *Dmc1* is required for homologous chromosome synapsis during meiosis. *Mol Cell*, 1(5):707–18, Apr 1998. 5

[28] M Alsheimer J Fraune, S Schramm and R Benavente. The mammalian synaptonemal complex: protein components, assembly and role in meiotic recombination. *Exp Cell Res*, 318(12):1340–6, Jul 2012. 5

[29] K Ishiguro and Y Watanabe. Chromosome cohesion in mitosis and meiosis. *J Cell Sci*, 120(Pt 3):367–9, Feb 2007. 5

[30] KA Henderson and S Keeney. Tying synaptonemal complex initiation to the formation and programmed repair of dna double-strand breaks. *Proc Natl Acad Sci U S A*, 101(13):4519–24, Mar 2004. 6

[31] C Jérôme F Baudat, J Buard and B de Massy. [*What* determines the localisation of spots of meiotic recombination?]. *Med Sci (Paris)*, 27(12):1053–5, Dec 2011. 6

[32] JL Jillian and SJ Boulton. The choice in meiosis – defining the factors that influence crossover or non-crossover formation. *Journal of Cell Science*, 124:501–513, 2011. 6

[33] T Allers and M Lichten. Differential timing and control of noncrossover and crossover recombination during meiosis. *Cell*, 106(1):47–57, Jul 2001. 6

[34] L Duret and N Galtier. Biased gene conversion and the evolution of mammalian genomic landscapes. *Annu Rev Genomics Hum Genet*, 10:285–311, 2009. 6, 14

[35] N Chuzhanova C Férec J Chen, DN Cooper and GP Patrinos. Gene conversion: mechanisms, evolution and human disease. *Nature Genetics*, 8:762–775, 2007. 6

[36] JM Burke GG Presting J Ross-Ibarra J Shi, SE Wolf and RK Dawe. Widespread gene conversion in centromere cores. *PLoS Biol*, 8(3):e1000327, Mar 2010. 6

[37] P Kenneth K Paigen and P Petkov. Mammalian recombination hot spots: properties, control and evolution. *Nat Rev Genet*, 11(3):221–33, Mar 2010. 6, 8, 9, 10, 11, 14, 18

[38] J Buard and B de Massy. Playing hide and seek with mammalian meiotic crossover hotspots. *Trends Genet*, 23(6):301–9, Jun 2007. 6

[39] P Sung and H Klein. Mechanism of homologous recombination: mediators and helicases take on regulatory functions. *Nat Rev Mol Cell Biol*, 7(10):739–50, Oct 2006. 7

[40] AJ Jeffreys L Kauppi and S Keeney. Where the crossovers are: recombination distributions in mammals. *Nat Rev Genet*, 5(6):413–24, Jun 2004. 8, 9, 125

[41] I Tiemann-Boege N Arnheim, P Calabrese. Mammalian meiotic recombination hot spots. *Annu Rev Genet*, 41:369–99, 2007. 8, 9

[42] P Hunt T Hassold, S Sherman. Counting cross-overs: characterizing meiotic recombination in mammals. *Hum Mol Genet*, 9(16):2409–19, Oct 2000. 8

[43] V Chow-M Nigro S Ma-Sai KA Ferguson, EC Wong. Abnormal meiotic recombination in infertile men and its association with sperm aneuploidy. *Hum Mol Genet*, 16(23):2870–9, Dec 2007. 8

[44] C Greene-E Ko A Rademaker F Sun, P Turek and HR Martin. Abnormal progression through meiosis in men with nonobstructive azoospermia. *Fertil Steril*, 87(3):565–71, Mar 2007. 8

[45] MJ Farrer-LM Cullen MM Coleman-R Williamson RK Wyse R Palmer AM Kessling CM Howard, GE Davies. Meiotic crossing-over in nondisjoined chromosomes of children with trisomy 21 and a congenital heart defect. *Am J Hum Genet*, 53(2):462–71, Aug 1993. 8

[46] J Navarro M Campillo-F García S Egozcue C Abad J Egozcue M Codina-Pascual, M Oliver-Bonet and J Benet. Synapsis and meiotic recombination analyses: Mlh1 focus in the xy pair as an indicator. *Hum Reprod*, 20(8):2133–9, Aug 2005. 8

[47] AL Barlow and MA Hultén. Crossing over analysis at pachytene in man. *Eur J Hum Genet*, 6(4):350–8, 1998. 8

[48] GM Hartshorne C Tease and MA Hultén. Patterns of meiotic recombination in human fetal oocytes. *Am J Hum Genet*, 70(6):1469–79, Jun 2002. 8

[49] ST Globus and S Keeney. The joy of six: how to control your crossovers. *Cell*, 149(1):11–2, Mar 2012. 9

[50] J Sainz GM Jonsdottir SA Gudjonsson-B Richardsson S Sigurdardottir J Barnard B Hallbeck G Masson A Shlien ST Palsson ML Frigge TE Thorgeirsson JR Gulcher A Kong, DF Gudbjartsson and K Stefansson. A high-resolution recombination map of the human genome. *Nat Genet*, 31(3):241–7, Jul 2002. 9, 10

[51] M Yoshino T Sagai T Shiroishi, T Koide and K Moriwaki. Hotspots of homologous recombination in mouse meiosis. *Adv Biophys*, 31:119–32, 1995. 9

[52] M Jasin L Kauppi and S Keeney. Meiotic crossover hotspots contained in haplotype block boundaries of the mouse genome. *Proc Natl Acad Sci U S A*, 104(33):13396–401, Aug 2007. 9

[53] A Tumian RE Bontrop C Freeman-TS MacFie G McVean P Donnelly S Myers, R Bowden. Drive against hotspot motifs in primates implicates the *PRDM9* gene in meiotic recombination. *Science*, 327(5967):876–9, Feb 2010. 9, 14, 15, 17, 18, 30, 42, 80, 125

[54] S Pfeifer O Venn L Ségurel T Street EM Leffler R Bowden I Aneas J Broxholme P Humburg Z Iqbal G Lunter J Maller RD Hernandez C Melton A Venkat MA Nobrega R Bontrop SR Myers P Donnelly M Przeworski A Auton, A Fledel-Alon and G McVean. A fine-scale chimpanzee genetic map from population sequencing. *Science*, 336(6078):193–8, Apr 2012. 9, 17, 18

[55] C Freeman G McVean S Myers, L Bottolo and P Donnelly. A fine-scale map of recombination rates and hotspots across the human genome. *Science*, 310(5746):321–4, Oct 2005. 9, 11, 12, 91, 107, 128, 129

[56] J Hey. What's so hot about recombination hotspots? *PLoS Biol*, 2(6):e190, Jun 2004. 9

[57] S Horvath J McNicholas J Srelinger C Wake E Long B Mach B M Steinmetz, K Minard and L Hood. A molecular map of the immune response region from the major histocompatibility complex of the mouse. *Nature*, 300(5887):35–42, Nov 1982. 9

[58] R Neumann and AJ Jeffreys. Polymorphism in the activity of human crossover hotspots independent of local *DNA* sequence variation. *Hum Mol Genet*, 15(9):1401–11, May 2006. 9, 11

[59] C Ober JK Pritchard G Coop, X Wen and M Przeworski. High-resolution mapping of crossovers reveals extensive variation in fine-scale recombination patterns among humans. *Science*, 319(5868):1395–8, Mar 2008. 10

[60] G Coop and M Przeworski. An evolutionary view of human recombination. *Nat Rev Genet*, 8(1):23–34, Jan 2007. 10

[61] S Hunt P Deloukas DR Bentley GAT McVean, SR Myers and P Donnelly. The fine-scale structure of recombination rate variation in the human genome. *Science*, 304(5670):581–4, Apr 2004. 10

[62] Ld Kauppi C May R Neumann MT Slingsby AJ Jeffreys, JK Holloway and AJ Webb. Meiotic recombination hot spots and human *DNA* diversity. *Philos Trans R Soc Lond B Biol Sci*, 359(1441):141–52, Jan 2004. 10, 11

[63] IL Berg AJ Webb and A Jeffreys. Sperm cross-over activity in regions of the human genome showing extreme breakdown of marker association. *Proc Natl Acad Sci U S A*, 105(30):10471–6, Jul 2008. 10, 92, 96

[64] N Li G Hellenthal MJ Rieder DA Nickerson DC Crawford, T Bhangale and M Stephens. Evidence for substantial fine-scale variation in recombination rates across the human genome. *Nat Genet*, 36(7):700–6, Jul 2004. 10

[65] B Behr J Wang, HC Fan and SR Quake. Genome-wide single-cell analysis of recombination activity and de novo mutation rates in human sperm. *Cell*, 150(2):402–12, Jul 2012. 11

[66] S Myers G McVean. *PRDM9* marks the spot. *Nat Genet*, 42(10):821–2, Oct 2010. 11

[67] A Auton and G McVean. Recombination rate estimation in the presence of hotspots. *Genome Res*, 17(8):1219–27, Aug 2007. 11

[68] A Auton P Donnelly S Myers, C Freeman and G McVean. A common sequence motif associated with recombination hot spots and genome instability in humans. *Nat Genet*, 40(9):1124–9, Sep 2008. 12, 13, 14, 16, 22, 32, 35, 40, 52, 80, 100, 106, 107, 110, 124, 125, 126, 129

[69] AJ Jeffreys and R Neumann. Reciprocal crossover asymmetry and meiotic drive in a human recombination hot spot. *Nat Genet*, 31(3):267–71, Jul 2002. 13

[70] R Neumann AJ Jeffreys. Factors influencing recombination frequency and distribution in a human meiotic crossover hotspot. *Hum Mol Genet*, 14(15):2277–87, Aug 2005. 13

[71] D Camerini-Otero J Zheng, P Pavel and TM Przytycka. Detecting sequence polymorphisms associated with meiotic recombination hotspots in the human genome. *Genome Biol*, 11(10):R103, 2010. 13

[72] F Baudat and B de Massy. Cis- and trans-acting elements regulate the mouse *Psmb9* meiotic recombination hotspot. *PLoS Genet*, 3(6):e100, Jun 2007. 13

[73] N Hanzawa-H Gotoh T Shiroishi, T Sagai and K Moriwaki. Genetic control of sex-dependent meiotic recombination in the major histocompatibility complex of the mouse. *EMBO J*, 10(3):681–6, Mar 1991. 14

[74] AJ Jeffreys and R Neumann. The rise and fall of a human recombination hot spot. *Nat Genet*, 41(5):625–9, May 2009. 14

[75] K Brick-P Khil D Camerini-Otero F Smagulova, IV Gregoretti and GV Petukhova. Genome-wide analysis reveals novel molecular features of mouse recombination hotspots. *Nature*, 472(7343):375–8, Apr 2011. 14, 21, 22, 24, 26, 66, 79, 92, 98, 124, 125, 134

[76] MP Francino and H Ochman. Strand asymmetries in *DNA* evolution. *Trends Genet*, 13(6):240–5, Jun 1997. 14

[77] C Grey-A Fledel-Alon C Ober M Przeworski G Coop F Baudat, J Buard and B De Massy. Prdm9 is a major determinant of meiotic recombination hotspots in humans and mice. *Science*, 327(5967):836–40, Feb 2010. 14, 16, 17, 18, 19, 30, 80, 125, 127

[78] PM Petkov E Parvanov and K Paigen. Prdm9 controls activation of mammalian recombination hotspots. *Science*, 327(5967):835, Feb 2010. 14, 16, 17, 79

[79] KG Lam-S Sarbajna-L Odenthal-Hesse CA May IL Berg, R Neumann and AJ Jeffreys. Prdm9 variation strongly influences recombination hot-spot activity and meiotic instability in humans. *Nat Genet*, 42(10):859–63, Oct 2010. 14, 17, 18, 19, 20, 21, 30, 44, 79, 124, 125, 126, 127

[80] S Sarbajna-L Odenthal-Hesse NJ Butler IL Berg, R Neumann and AJ Jeffreys. Variants of the protein prdm9 differentially regulate a set of human meiotic recombination hotspots highly active in *African* populations. *Proc Natl Acad Sci U S A*, 108(30):12378–83, Jul 2011. 14, 17, 21, 125, 127

[81] DF Gudbjartsson-G Masson-A Sigurdsson A Jonasdottir GB Walters A Jonasdottir GA Gylfason KT Kristinsson SA Gudjonsson ML Frigge A Helgason U Thorsteinsdottir A Kong, G Thorleifsson and K Stefansson. Fine-scale recombination rate differences between sexes, populations and individuals. *Nature*, 467(7319):1099–103, Oct 2010. 14, 125

[82] AG Hinch et al. The landscape of recombination in *African Americans*. *Nature*, 476(7359):170–5, Aug 2011. 14, 21, 125

[83] PM Petkov ED Parvanov, Ng Siemon and K Paigen. Trans-regulation of mouse meiotic recombination hotspots by *Rcr1*. *PLoS Biol*, 7(2):e36, Feb 2009. 16, 50, 125

[84] W Lin-S Bonfils-V Géli A Nicolas V Borde, N Robine. Histone h3 lysine 4 trimethylation marks meiotic recombination initiation sites. *EMBO J*, 28(2):99–111, Jan 2009. 16, 26, 54, 66, 79, 80, 131

[85] G Chauveau-Le Friec-F Langa F Baudat C Grey, P Barthès and B de Massy. Mouse prdm9 dna-binding specificity determines sites of histone h3 lysine 4 trimethylation for initiation of meiotic recombination. *PLoS Biol*, 9(10):e1001176, Oct 2011. 16, 27, 66

[86] CP Ponting. What are the genomic drivers of the rapid evolution of *PRDM9*? *Trends Genet*, 27(5):165–71, May 2011. 17, 18, 124

[87] J Bayes-Z Birtle-KC Roach N Phadnis A Scott G Lunter HS Malik PL Oliver, L Goodstadt and CP Ponting. Accelerated evolution of the *Prdm9* speciation gene across diverse metazoan taxa. *PLoS Genet*, 5(12):e1000753, Dec 2009. 17, 18, 80, 124

[88] MJ Neale. Prdm9 points the zinc finger at meiotic recombination hotspots. *Genome Biol*, 11(2):104, 2010. 17

[89] A Hochwagen and GAB Marais. Meiosis: a *PRDM9* guide to the hotspots of recombination. *Curr Biol*, 20(6):R271–4, Mar 2010. 17, 79, 125

[90] EM Leffler L Ségurel and M Przeworski. The case of the fickle fingers: how the *PRDM9* zinc finger protein specifies meiotic recombination hotspots in humans. *PLoS Biol*, 9(12):e1001211, Dec 2011. 18, 22

[91] RO Emerson JH Thomas and J Shendure. Extraordinary molecular evolution in the *PRDM9* fertility gene. *PLoS One*, 4(12):e8505, 2009. 18, 80

[92] K Yoshida K Hayashi and Y Matsui. A histone *H3* methyltransferase controls epigenetic events required for meiotic prophase. *Nature*, 438(7066):374–8, Nov 2005. 18, 125

[93] Y Miyagawa-T Ueda-Y Matsuoka Y Matsui A Okuyama Y Nishimune S Irie, A Tsujimura and H Tanaka. Single-nucleotide polymorphisms of the prdm9 (meisetz) gene in patients with nonobstructive azoospermia. *J Androl*, 30(4):426–31, 2009. 18

[94] N Sakugawa-H Sato-H Hayashi M Namiki T Miyamoto, E Koh and K Sengoku. Two single nucleotide polymorphisms in *PRDM9* (meisetz) gene may be a genetic risk factor for japanese patients with azoospermia by meiotic arrest. *J Assist Reprod Genet*, 25(11-12):553–7, 2008. 18

[95] A Di Rienzo V Muñoz-Fuentes and C Vilà. Prdm9, a major determinant of meiotic recombination hotspots, is not functional in dogs and their wild relatives, wolves and coyotes. *PLoS One*, 6(11):e25498, 2011. 18, 124

[96] A Ratnakumari-LUPA Consortium CP Ponting P Chris P E Axelsson, MT Webster and K Lindblad-Toh. Death of *PRDM9* coincides with stabilization of the recombination landscape in the dog genome. *Genome Res*, 22(1):51–63, Jan 2012. 18

[97] IV Getun ZK Wu and PRJ Bois. Anatomy of mouse recombination hot spots. *Nucleic Acids Res*, 38(7):2346–54, Apr 2010. 22

[98] SD Boyd-CL Smith A Fire A Sidow A Valouev, SM Johnson. Determinants of nucleosome organization in primary human cells. *Nature*, 474(7352):516–20, Jun 2011. 22

[99] PN Cockerill. Structure and function of active chromatin and *DNase* i hypersensitive sites. *FEBS J*, 278(13):2182–210, Jul 2011. 23

[100] A Nicolas. Relationship between transcription and initiation of meiotic recombination: toward chromatin accessibility. *Proc Natl Acad Sci U S A*, 95(1):87–9, Jan 1998. 23

[101] T Shibata K Ohta and A Nicolas. Changes in chromatin structure at recombination initiation sites during yeast meiosis. *EMBO J*, 13(23):5754–63, Dec 1994. 23

[102] F Xu Q Fan and TD Petes. Meiosis-specific double-strand dna breaks at the his4 recombination hot spot in the yeast *Saccharomyces cerevisiae*: control in cis and trans. *Mol Cell Biol*, 15(3):1679–88, Mar 1995. 23

[103] S Keeney and N Kleckner. Communication between homologous chromosomes: genetic alterations at a nuclease-hypersensitive site can alter mitotic chromatin structure at that site both in cis and in trans. *Genes Cells*, 1(5):475–89, May 1996. 23

[104] D Gadelle PC Varoutas A Nicolas A Bergerat, B de Massy and P Forterre. An atypical topoisomerase *II* from archaea with implications for meiotic recombination. *Nature*, 386(6623):414–7, Mar 1997. 23

[105] TC Wu and M Lichten. Factors that affect the location and frequency of meiosis-induced double-strand breaks in *Saccharomyces cerevisiae*. *Genetics*, 140(1):55–66, May 1995. 23

[106] MH Shen R Shenkar and N Arnheim. *DNase I*-hypersensitive sites and transcription factor-binding motifs within the mouse *E* beta meiotic recombination hot spot. *Mol Cell Biol*, 11(4):1813–9, Apr 1991. 24

[107] T Tonthat J Stuart S Ranade H Peckham K Zeng JA Malek G Costa K McKernan A Sidow A Fire SM Johnson A Valouev, Jd Ichikawa. A high-resolution, nucleosome position map of *C.* elegans reveals a lack of universal sequence-dictated positioning. *Genome Res*, 18(7):1051–63, Jul 2008. 24

[108] JM Maniar A Valouev A Sidow MA Kay L Gracey, Z Chen and AZ Fire. An in vitro-identified high-affinity nucleosome-positioning signal is capable of transiently positioning a nucleosome in vivo. *Epigenetics Chromatin*, 3(1):13, 2010. 24

[109] RD Kornberg and Y Lorch. Twenty-five years of the nucleosome, fundamental particle of the eukaryote chromosome. *Cell*, 98(3):285–94, Aug 1999. 24

[110] F Fagerström-Billai P Korber A Lantermann, A Strålfors and K Ekwall. Genome-wide mapping of nucleosome positions in schizosaccharomyces pombe. *Methods*, 48(3):218–25, Jul 2009. 24

[111] AM Khalil-M Ahmad IV Getun, ZK Wu and PRJ Bois. Nucleosome occupancy landscape and dynamics at mouse recombination hotspots. *EMBO Rep*, 11(7):555–60, Jul 2010. 24, 50, 133

[112] ZK Wu IV Getun and PRJ Bois. Organization and roles of nucleosomes at mouse meiotic recombination hotspots. *Nucleus*, 3(3):244–50, 2012. 24, 26

[113] L Marín R Serrano L Quintales E de Castro, I Soriano and F Antequera. Nucleosomal organization of replication origins and meiotic recombination hotspots in fission yeast. *EMBO J*, 31(1):124–37, Jan 2012. 24, 54, 80, 131

[114] N Musa-Lempel A Afek, I Sela and DB Lukatsky. Nonspecific transcription-factor-*DNA* binding influences nucleosome occupancy in yeast. *Biophys J*, 101(10):2465–75, Nov 2011. 24

[115] SK Kota and R Feil. Epigenetic transitions in germ cell development and meiosis. *Dev Cell*, 19(5):675–86, Nov 2010. 25

[116] H Cedar and Y Bergman. Linking *DNA* methylation and histone modification: patterns and paradigms. *Nat Rev Genet*, 10(5):295–304, May 2009. 26

[117] JK Choi and LJ Howe. Histone acetylation: truth of consequences? *Biochem Cell Biol*, 87(1):139–50, Feb 2009. 26

[118] A Harnicarová-G Galiová E Bártová, J Krejcí and S Kozubek. Histone modifications and nuclear architecture: a review. *J Histochem Cytochem*, 56(8):711–21, Aug 2008. 26, 70

[119] R Schneider and R Grosschedl. Dynamics and interplay of nuclear architecture, genome organization, and gene expression. *Genes Dev*, 21(23):3027–43, Dec 2007. 26

[120] R Kniewel and S Keeney. Histone methylation sets the stage for meiotic *DNA* breaks. *EMBO J*, 28(2):81–3, Jan 2009. 26

[121] C Soustelle-K Suhre A Nicolas V Géli C de La Roche Saint-André J Sollier, W Lin. Set1 is required for meiotic *S*-phase onset, double-strand break formation and middle gene expression. *EMBO J*, 23(9):1957–67, May 2004. 26

[122] C Grey-B de Massy J Buard, P Barthès. Distinct histone modifications define initiation and repair of meiotic recombination in the mouse. *EMBO J*, 28(17):2616–24, Sep 2009. 27, 135

[123] D G Anderson and S C Kowalczykowski. The recombination hot spot chi is a regulatory element that switches the polarity of dna degradation by the *RecBCD* enzyme. *Genes Dev*, 11(5):571–81, Mar 1997. 27

[124] AK Eggleston-SD Lauder SC Kowalczykowski, DA Dixon and WM Rehrauer. Biochemistry of homologous recombination in *Escherichia Coli*. *Microbiol Rev*, 58(3):401–65, Sep 1994. 27

[125] DP Thomas AW Michael, D Margaret. Transcription factors are required for the meiotic recombination hotspot at the *HIS4* locus in saccharomyces cerevisiae. *Proceedings of National Academy of Sciences*, 90:6621–6625, 1993. 27

[126] Q Fan DT Kirkpatrick and TD Petes. Maximal stimulation of meiotic recombination by a yeast transcription factor requires the transcription activation domain and a *DNA*-binding domain. *Genetics*, 152(1):101–15, May 1999. 27, 28

[127] ER Siegel WP Wahls and MK Davidson. Meiotic recombination hotspots of fission yeast are directed to loci that express non-coding rna. *PLoS One*, 3(8):e2887, 2008. 27

[128] MD Krawchuk N Kon, SC Schroeder and WP Wahls. Regulation of the *Mts1-Mts2*-dependent ade6-*M26* meiotic recombination hot spot and developmental decisions by the *Spc1* mitogen-activated protein kinase of fission yeast. *Mol Cell Biol*, 18(12):7575–83, Dec 1998. 28

[129] T Shibata K Hirota, K Mizuno and K Ohta. Distinct chromatin modulators regulate the formation of accessible and repressive chromatin at the fission yeast recombination hotspot ade6-*M26*. *Mol Biol Cell*, 19(3):1162–73, Mar 2008. 28

[130] TM Przytycka-J Li M Wu, C Kwoh and J Zheng. Epigenetic functions enriched in transcription factors binding to mouse recombination hotspots. *Proteome Sci*, 10 Suppl 1:S11, 2012. 28

[131] ATM Bagshaw WW Steiner, PA Davidow. Important characteristics of sequence-specific recombination hotspots in *Schizosaccharomyces Pombe*. *Genetics*, 187(2):385–96, Feb 2011. 28

[132] B Hwang C Tsai, V Smider and G Chu. Electrophoretic mobility shift assays for protein-*DNA* complexes involved in *DNA* repair. *Methods Mol Biol*, 920:53–78, 2012. 30, 31

[133] V Sommermeyer E Brachet and V Borde. Interplay between modifications of chromatin and meiotic recombination hotspots. *Biol Cell*, 104(2):51–69, Feb 2012. 50

[134] ENCODE Project Consortium and Ewan et al. Birney. Identification and analysis of functional elements in 1% of the human genome by the encode pilot project. *Nature*, 447(7146):799–816, 2007. 51

[135] ENCODE Project Consortium and Ian et al. Dunham. An integrated encyclopedia of *DNA* elements in the human genome. *Nature*, 489:57–74, 2012. 51, 80, 109, 110

[136] S Cuddapah-T Roh A Barski Z Wang G Wei DE Schones, K Cui and K Zhao. Dynamic regulation of nucleosome positioning in the human genome. *Cell*, 132(5):887–98, Mar 2008. 51, 66

[137] R Cui-T Roh D Schones Z Wang G Wei I Chepelev A Barski, S Cuddapah and K Zhao. High-resolution profiling of histone methylations in the human genome. *Cell*, 129(4):823–37, May 2007. 51, 55, 70

[138] PC Scacheri-G Renaud MJ Halawi MR Erdos R Green PS Meltzer TG Wolfsberg GE Crawford, S Davis and FS Collins. *DNase*-chip: a high-resolution method to identify *DNase I* hypersensitive sites using tiled microarrays. *Nat Methods*, 3(7):503–9, Jul 2006. 54

[139] RE Thurman et al. The accessible chromatin landscape of the human genome. *Nature*, 489(7414):75–82, Sep 2012. 54, 112

[140] B Lee-D London D Keefe E Birney VR Iyer GE Crawford AP Boyle, L Song and TS Furey. High-resolution genome-wide in vivo footprinting of diverse transcription factors in human cells. *Genome Res*, 21(3):456–64, Mar 2011. 54, 55

[141] JD Lieb LE Berchowitz, SE Hanlon and GP Copenhaver. A positive but complex association between meiotic double-strand break hotspots and open chromatin in *Saccharomyces Cerevisiae*. *Genome Res*, 19(12):2245–57, Dec 2009. 54, 131, 133

[142] BF Pugh L Zhang, H Ma. Stable and dynamic nucleosome states during a meiotic developmental process. *Genome Res*, 21(6):875–84, Jun 2011. 54, 80, 131

[143] IK Moore-Y Fondufe-Mittendorf AJ Gossett Y Field JD Lieb J Widom E Segal D Tillo, N Kaplan and TR Hughes. High nucleosome occupancy is encoded at human regulatory sequences. *PLoS One*, 5(2):e9129, 2010. 55

[144] FV Roy F Daenen and PJ De Bleser. Low nucleosome occupancy is encoded around functional human transcription factor binding sites. *BMC Genomics*, 9:332, 2008. 55

[145] Mayetri PG Giresi, M Gupta and JD Lieb. Regulation of nucleosome stability as a mediator of chromatin function. *Curr Opin Genet Dev*, 16(2):171–6, Apr 2006. 55

[146] L Gaveglia-S Althammer-J González-Vallinas E Eyras F Le Dily R Zaurin D Soronellas GP Vicent M Beato C Ballaré, G Castellano. Nucleosome-driven transcription factor binding and gene regulation. *Mol Cell*, Nov 2012. 66

[147] CL Peterson Y Fu, M Sinha and Z Weng. The insulator binding protein *CTCF* positions 20 nucleosomes around its binding sites across the human genome. *PLoS Genet*, 4(7):e1000138, 2008. 67

[148] P Green G McVicker. Genomic signatures of germline gene expression. *Genome Res*, 20(11):1503–11, Nov 2010. 68, 72, 121

[149] S Henikoff and A Shilatifard. Histone modification: cause or cog? *Trends Genet*, 27(10):389–96, Oct 2011. 68

[150] P Lefevre and C Bonifer. Analyzing histone modification using crosslinked chromatin treated with micrococcal nuclease. *Methods Mol Biol*, 325:315–25, 2006. 68

[151] A Izzo and R Schneider. Chatting histone modifications in mammals. *Brief Funct Genomics*, 9(5-6):429–43, Dec 2010. 70

[152] N Dillon. Heterochromatin structure and function. *Biol Cell*, 96(8):631–7, Oct 2004. 70

[153] PV Kharchenko et al. Comprehensive analysis of the chromatin landscape in *Drosophila Melanogaster*. *Nature*, 471(7339):480–5, Mar 2011. 70

[154] R Goitein-V Kottusch-H Cedar-M Marcus K Sperling, BS Kerem. *DNase I* sensitivity in facultative and constitutive heterochromatin. *Chromosoma*, 93(1):38–42, 1985. 70

[155] P Héry-S Barral-J Thuret-S Dimitrov M Gérard S Chantalat, A Depaux. Histone *H3* trimethylation at *Lysine* 36 is associated with constitutive and facultative heterochromatin. *Genome Res*, 21(9):1426–37, Sep 2011. 70

[156] SCR Elgin and Shiv I S SIS Grewal. Heterochromatin: silence is golden. *Curr Biol*, 13(23):R895–8, Dec 2003. 70

[157] JR Whetstine JC Black, RV Rechem. Histone lysine methylation dynamics: establishment, regulation, and biological impact. *Mol Cell*, 48(4):491–507, Nov 2012. 70

[158] TG Smart P Thomas. Hek293 cell line: a vehicle for the expression of recombinant proteins. *J Pharmacol Toxicol Methods*, 51(3):187–200, 2005. 80, 81

[159] TS Furey. *ChIP*-seq and beyond: new and improved methodologies to detect and characterize protein-*DNA* interactions. *Nat Rev Genet*, 13(12):840–52, Dec 2012. 81

[160] ER Mardis. *ChIP*-seq: welcome to the new frontier. *Nat Methods*, 4(8):613–4, Aug 2007. 81, 83

[161] A Sundquist-C Medina-E Anton-S Batzoglou RM Myers A Valouev, DS Johnson and A Sidow. Genome-wide analysis of transcription factor binding sites based on chip-seq data. *Nat Methods*, 5(9):829–34, Sep 2008. 81

[162] S Pott ET Liu and M Huss. Q&a: *ChIP-seq* technologies and the study of gene regulation. *BMC Biol*, 8:56, 2010. 81, 83

[163] MJ Fritzler JJ Moser, EKL Chan. Optimization of immunoprecipitation-western blot analysis in detecting *GW182*-associated components of *GW/P* bodies. *Nat Protoc*, 4(5):674–85, 2009. 81

[164] W Lu AR Aricescu and EY Jones. A time- and cost-efficient system for high-level protein production in mammalian cells. *Acta Crystallogr D Biol Crystallogr*, 62(Pt 10):1243–50, Oct 2006. 82

[165] KL Loveland-K Ohbo-L Robb-F Schwenk J Seibler D Roellig A Kranz K Anastassiadis S Glaser, S Lubitz and AF Stewart. The *Histone* 3 *Lysine* 4 methyltransferase, *Mll2*, is only required briefly in development and spermatogenesis. *Epigenetics Chromatin*, 2(1):5, 2009. 86

[166] S Minucci-G Natoli I Barozzi, A Termanini. Fish the chips: a pipeline for automated genomic annotation of *ChIP-Seq* data. *Biol Direct*, 6:51, 2011. 91

[167] Tao J Feng, T Liu and Y Zhang. Using *MACS* to identify peaks from *ChIP-Seq* data. *Curr Protoc Bioinformatics*, Chapter 2:Unit 2.14, Jun 2011. 95

[168] B Qin-Y Zhang-SX Liu J Feng, T Liu. Identifying *ChIP*-seq enrichment using *MACS*. *Nat Protoc*, 7(9):1728–40, Sep 2012. 95

[169] C Meyer-J Eeckhoute-DS Johnson-BE Bernstein C Nusbaum Chad RM Myers M Brown W Li Y Zhang, T Liu and XS Liu. Model-based analysis of *ChIP-Seq (MACS)*. *Genome Biol*, 9(9):R137, 2008. 95

[170] WS Noble CE Grant, TL Bailey. *FIMO*: scanning for occurrences of a given motif. *Bioinformatics*, 27(7):1017–8, Apr 2011. 99, 102

[171] FA Buske-M Frith-CE Grant-L Clementi J Ren WW Li TL Bailey, M Boden and WS Noble. *MEME SUITE*: tools for motif discovery and searching. *Nucleic Acids Res*, 37(Web Server issue):W202–8, Jul 2009. 99

[172] TL Bailey P Machanick. *MEME-ChIP*: motif analysis of large *DNA* datasets. *Bioinformatics*, 27(12):1696–7, Jun 2011. 99

[173] S Verschoor-A Inselman-MA Handel-MJ McKay Michael J H Xu, M Beasley. A new role for the mitotic *RAD21/SCC1* cohesin in meiotic chromosome cohesion and segregation in the mouse. *EMBO Rep*, 5(4):378–84, Apr 2004. 121

[174] S Fujiyama-Nakamura S Kato Y Watanabe K Ishiguro, J Kim. A new meiosis-specific cohesin complex implicated in the cohesin code for homologous pairing. *EMBO Rep*, 12(3):267–75, Mar 2011. 121

[175] M Eccles-E Dickinson J Horsfield M Mönnich, S Banks. Expression of cohesin and condensin genes during zebrafish development supports a non-proliferative role for cohesin. *Gene Expr Patterns*, 9(8):586–94, Dec 2009. 121

[176] PD Kehayova-F Pauli K Newberry R Myers K Monahan, ND Rudnick and T Maniatis. Role of *CCCTC* binding factor *(CTCF)* and cohesin in the generation of single-cell diversity of protocadherin- gene expression. *Proc Natl Acad Sci U S A*, 109(23):9125–30, Jun 2012. 121

[177] B Lee and Vishwanath R Iyer. Genome-wide studies of *CCCTC*-binding factor *(CTCF)* and cohesin provide insight into chromatin structure and regulation. *J Biol Chem*, 287(37):30906–13, Sep 2012. 121

[178] S Watt-PC Schwalie MD Wilson H Xu RG Ramsay DT Odom P AJ Faure, D Schmidt and Flicek. Cohesin regulates tissue-specific expression by stabilizing highly occupied cis-regulatory modules. *Genome Res*, 22(11):2163–75, Nov 2012. 121

[179] WP Wahls and MK Davidson. New paradigms for conserved, multifactorial, cis-acting regulation of meiotic recombination. *Nucleic Acids Res*, 40(20):9983–9, Nov 2012. 124

[180] SL Page and RS Hawley. Chromosome choreography: the meiotic ballet. *Science*, 301(5634):785–9, Aug 2003. 124, 133

[181] B de Massy. Distribution of meiotic recombination sites. *Trends Genet*, 19(9):514–22, Sep 2003. 125

[182] N Kleckner L Xu. Sequence non-specific double-strand breaks and interhomolog interactions prior to double-strand break formation at a meiotic recombination hot spot in yeast. *EMBO J*, 14(20):5115–28, Oct 1995. 133

[183] F van PR Gafken and DE Gottschling. Dot1p modulates silencing in yeast by methylation of the nucleosome core. *Cell*, 109(6):745–56, Jun 2002. 135

[184] V Cordón-Preciado F Cortés-Ledesma L Aragón A Aguilera PA San-Segundo F Conde, E Refolio. The *Dot1* histone methyltransferase and the *Rad9* checkpoint adaptor contribute to cohesin-dependent double-strand break repair by sister chromatid recombination in *Saccharomyces Cerevisiae*. *Genetics*, 182(2):437–46, Jun 2009. 135

[185] L Mariño-Ramírez L Hansen, N Kim and D Landsman. Analysis of biological features associated with meiotic recombination hot and cold spots in *Saccharomyces Cerevisiae*. *PLoS One*, 6(12):e29711, 2011. 135

[186] PW Greenwell-E Rinella DC Bouck Y Shibata BD Strahl P Mieczkowski-TD Petes JD Merker, M Dominska. The histone methylase *Set2p* and the histone deacetylase *Rpd3p* repress meiotic recombination at the *HIS4* meiotic recombination hotspot in *Saccharomyces Cerevisiae*. *DNA Repair (Amst)*, 7(8):1298–308, Aug 2008. 135

[187] L Florens-T Suganuma SK Swanson KK Lee W Shia S Anderson-W John MP Washburn P Michael MJ Carrozza, BF Li and JL Workman. Histone *H3* methylation by *Set2* directs deacetylation of coding regions by *Rpd3S* to

suppress spurious intragenic transcription. *Cell*, 123(4):581–92, Nov 2005. 135

[188] K Hirota-N Kon WP Wahls E Hartsuiker H Murofushi T Shibata T Yamada, K Mizuno and K Ohta. Roles of histone acetylation and chromatin remodeling factor in a meiotic recombination hotspot. *EMBO J*, 23(8):1792–803, Apr 2004. 135