Novel genetic and molecular properties of meiotic recombination protein PRDM9

Nicolas Altemose

New College
University of Oxford

A thesis submitted for the degree of Doctor of Philosophy

Michaelmas 2015

Abstract

Meiotic recombination is a fundamental biological process in sexually reproducing organisms, enabling offspring to inherit novel combinations of mutations, and ensuring even segregation of chromosomes into gametes. Recombination is initiated by programmed Double Strand Breaks (DSBs), the genomic locations of which are determined in most mammals by PRDM9, a rapidly evolving DNA-binding protein. In crosses between different mouse subspecies, certain *Prdm9* alleles cause infertility in hybrid males, implying a critical role in fertility and speciation. Upon binding to DNA, PRDM9 deposits a histone modification (H3K4me3) typically found in the promoters of expressed genes, suggesting that binding might alter the expression of nearby genes. Many other questions have remained about how PRDM9 initiates recombination, how it causes speciation, and why it evolves so rapidly. This body of work investigates these questions using complementary experimental and analytical methodologies. By generating a map of human PRDM9 binding sites and applying novel sequence analysis methods, I uncovered new DNA-binding modalities of PRDM9 and identified sequence-independent factors that predict binding and recombination outcomes. I also confirmed that PRDM9 can affect gene expression by binding to promoters, identifying candidate regulatory targets in meiosis. Furthermore, I showed that PRDM9's DNA-binding domain also mediates strong protein-protein interactions that produce PRDM9 multimers, which may play an important functional role. Finally, by generating high-resolution maps of PRDM9 binding in hybrid mice, I provide evidence for a mechanism to explain PRDM9-mediated speciation as a consequence of the joint evolution of PRDM9 and its binding targets. This work reveals that PRDM9 binding on one chromosome strongly impacts DSB formation and/or repair on the homologue, suggesting a novel role for PRDM9 in promoting efficient homology search and DSB repair, both critical for meiotic progression and fertility. One consequence is that PRDM9 may play a wider role in mammalian speciation.

Novel genetic and molecular properties of meiotic recombination protein PRDM9

Nicolas Altemose

New College

University of Oxford

A thesis submitted for the degree of

Doctor of Philosophy

Michaelmas 2015

Acknowledgements

This DPhil would not have been possible without the support and guidance of so many people around me, both inside and out of the lab. I began this programme with very little experience in experimental biology, and I have Nudrat Noor, Ben Davies, Ross Chapman, and Emmanuelle Bitoun to thank for showing me the ropes and being so generous with their time. I must also thank Jonathan Flint and Amarjit Bhomra for the use of their lab space and equipment, as well as all of the animal technicians who went above and beyond to make sure our notoriously wild mice were well cared for. Ana Teixeira and the staff of the Oxford Genomics Centre were incredibly helpful in guiding some of our experimental decisions, and in providing our data in a timely manner. Peter Donnelly has provided constant guidance during my DPhil, and it has been a pleasure navigating our collaborative projects to unlock the mechanisms of recombination and speciation. It has been a pleasure working with Edouard Hatton, Julie Hussin, and Anjali Hinch as part of this collaboration, in which we have teamed up to produce a body of work of which I am very proud. Robbie Davies, Afi Tumian, and Ran Li have been very generous in providing data and coding advice when I have needed it, and Anna Frangou, Thaddeus Aid, Yunli Song, and Mike Salter-Townshend have all made our group meetings and activities interesting and enjoyable.

Simon Myers has been a great mentor and a great friend. He has taught me that careful, even obsessive thought can go a long way in both experimental and analytical endeavours, and he has always treated me like a colleague. I will never forget the time we met for over eight hours, focussing on one particular pattern in our data, until in a moment of insight we cracked the problem and had made a new discovery about the binding patterns of PRDM9. Simon's enthusiasm for science and for the design of experiments and analyses is contagious, and it has made my DPhil all the better for it.

I owe a huge debt in the typesetting of this thesis to one person in particular, Garreth McCrudden, who has been by my side during the ups and downs of this DPhil, as have my family and friends. I must also thank my dear friend Laurie Nevay for his feedback and assistance, and John McManigle for making this very helpful thesis template available.

My time at Oxford was made possible by a gift from the British government in the form of a Marshall Scholarship, which funded the first two years of my

DPhil and spurred me to consider studying in the UK in the first place. I have the Marshall Aid Commemoration Committee to thank for many incredible experiences in the UK, and my fellow Marshall Scholars to thank for their friendship and support. The remainder of my DPhil was funded by a Howard Hughes Medical Institute Gilliam Fellowship for Advanced Study, and the research costs for most of this work were supported in part by the Wellcome Trust Investigator Award 098387/Z/12/Z (to Simon Myers).

I come away from this DPhil, and from the UK, knowing more about mouse testes than one really ever ought to, and with a remarkable new ability to talk about the weather (at length) in all manner of situations.

Abstract

Meiotic recombination is a fundamental biological process in sexually reproducing organisms, enabling offspring to inherit novel combinations of mutations, and ensuring even segregation of chromosomes into gametes. Recombination is initiated by programmed Double Strand Breaks (DSBs), the genomic locations of which are determined in most mammals by PRDM9, a rapidly evolving DNA-binding protein. In crosses between different mouse subspecies, certain *Prdm9* alleles cause infertility in hybrid males, implying a critical role in fertility and speciation. Upon binding to DNA, PRDM9 deposits a histone modification (H3K4me3) typically found in the promoters of expressed genes, suggesting that binding might alter the expression of nearby genes. Many other questions have remained about how PRDM9 initiates recombination, how it causes speciation, and why it evolves so rapidly. This body of work investigates these questions using complementary experimental and analytical methodologies. By generating a map of human PRDM9 binding sites and applying novel sequence analysis methods, I uncovered new DNA-binding modalities of PRDM9 and identified sequence-independent factors that predict binding and recombination outcomes. I also confirmed that PRDM9 can affect gene expression by binding to promoters, identifying candidate regulatory targets in meiosis. Furthermore, I showed that PRDM9's DNA-binding domain also mediates strong protein-protein interactions that produce PRDM9 multimers, which may play an important functional role. Finally, by generating high-resolution maps of PRDM9 binding in hybrid mice, I provide evidence for a mechanism to explain PRDM9-mediated speciation as a consequence of the joint evolution of PRDM9 and its binding targets. This work reveals that PRDM9 binding on one chromosome strongly impacts DSB formation and/or repair on the homologue, suggesting a novel role for PRDM9 in promoting efficient homology search and DSB repair, both critical for meiotic progression and fertility. One consequence is that PRDM9 may play a wider role in mammalian speciation.

A note on J.B.S. Haldane:

John Burdon Sanderson Haldane was also a student and later a fellow at New College, Oxford, when he published his 1922 paper describing the eponymous "Haldane's rule": an observation that speciation tends to begin with hybrid infertility in the heterogametic sex, which is a phenomenon of particular interest in this thesis. Haldane became a founder of mathematical genetics and a great populariser of science. As an homage to him, at the opening of each chapter I include some of his more colourful quotations.

Contents

I have no doubt that in reality the future will be vastly more surprising than anything I can imagine. Now my own suspicion is that the Universe is not only queerer than we suppose, but queerer than we can *suppose.*

— J.B.S. Haldane, *Possible Worlds and Other Papers* (1927)

1 Introduction

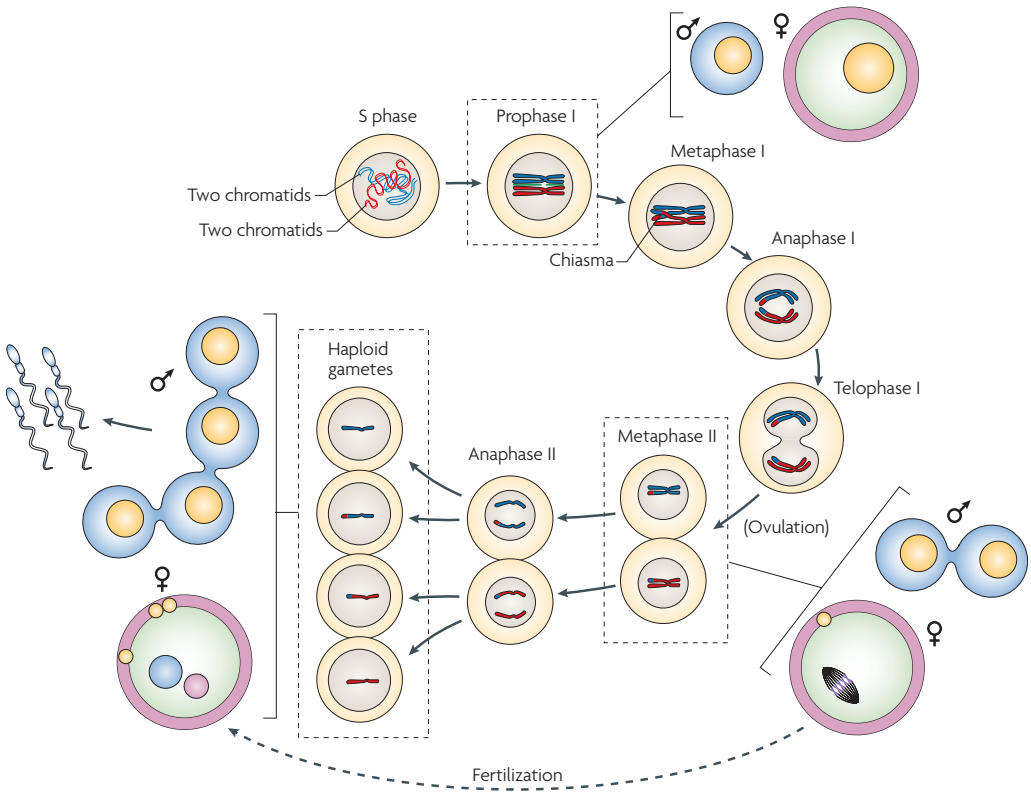
The evolution of sexual reproduction has intrigued biologists for over a century. Fundamentally, sex makes the challenge of reproduction a second-order problem, forcing many organisms to expend great amounts of energy to find, attract, and successfully copulate with their mates. And yet, sexual reproduction evolved very early on along the eukaryotic lineage and has persisted for aeons, with few exceptions in the animal kingdom. What benefit could sexual reproduction confer that would justify the peacock's plumes, the clockwork blooms of coral reefs, or the great upstream migrations of salmon? Why not revert to an asexual mode of reproduction which, at a molecular level, could avoid the complexities and costs of phenomena like genomic imprinting, mis-segregation, sex-linked genes, meiotic drive, polyspermy, and x-inactivation (to name a few). One answer, proposed early on by theory [1], continues to be reinforced by emerging evidence [2, 3]: sexual reproduction is, at its core, an engine for genetic recombination.

Although random mutation generates the substrates for genetic diversity and natural selection, natural selection does not act on single mutations in isolation—additive or complex interactions between mutations can affect the propensity for survival and reproduction. Without recombination, all mutations along a chromosome would be completely linked and inseparable, causing slightly deleterious mutations to piggyback on the selective forces driving strongly beneficial mutations

to fixation [4, 5]. In eukaryotes, mutations on different genes can become unlinked by translocation to different chromosomes, which segregate independently, but chromosome length and number are biophysically constrained, and random translocations are rare and often deleterious. Instead, meiotic recombination emerged as a designated process for unlinking mutations along the same chromosome, by repurposing the ancient cellular machinery that performs DNA damage repair [6]. Theories for the specific evolutionary origins of DNA repair, diploidy, sex, and meiosis remain complex, intertwined, and highly speculative [6], and are beyond the scope of this thesis. Despite its somewhat murky origins, meiotic recombination is now well understood to be an essential and fundamental component of sexual reproduction, facilitating the evolution of complex life and continuing to shape the genetic diversity of all sexually reproducing organisms.

1.1 Mammalian Prophase I and the role of recombination

In mammals, meiosis proceeds with two successive cell divisions that yield a final set of 4 haploid cells, some of which then undergo differentiation into gametes (**Figure 1.1**). Unlike mitosis, the first division in meiosis is presented with the unique challenge of aligning and segregating homologous chromosome pairs, and recombination plays an indispensable role in this process. The second meiotic division then separates sister chromatids, closely resembling normal mitosis. This overview will discuss the timing of key events in Meiotic Prophase I, when homologous chromosomes pair and undergo recombination. In human males, meiosis occurs in waves beginning at puberty, and takes roughly 22 days to complete Prophase I [7]; in females, meiosis begins in fetal ooctyes and only completes after puberty and fertilization (**Figure 1.1**). Prophase I is classically divided into four sub-stages based on the visible appearance of chromatin under the microscope: Leptotene, Zygotene, Pachytene, and Diplotene (**Figure 1.2**).

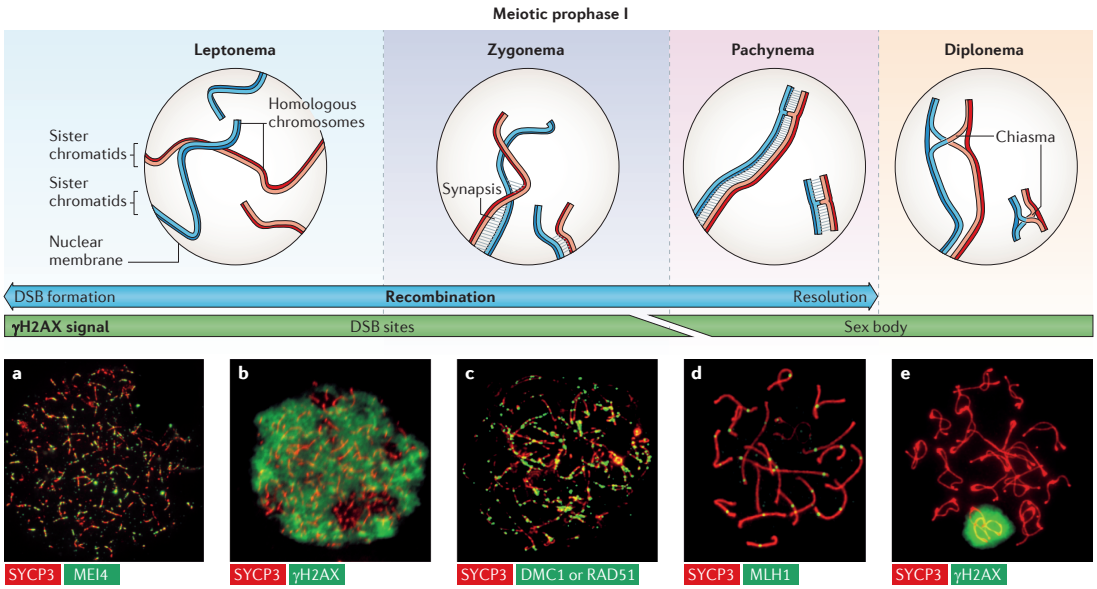


Reproduced from Handel and Schimenti (2010) *Nat. Rev. Genet.*

Figure 1.1: An overview of the stages of mammalian meiosis. Meiosis occurs as a double division of an oocyte or spermatocyte containing $4n$ chromosomes into 4 daughter cells, each with a haploid set of n chromosomes. Recombination occurs entirely in Prophase I, when crossover events become visible as chiasmata, which are essential for the proper alignment and segregation of homologous chromosomes in Meiosis I. Female meiosis begins *in utero* and is arrested at the end of Prophase I until puberty, fully completing only after fertilization.

1.1.1 Leptotene

After the completion of DNA replication in the Pre-Leptotene S-phase, the nucleus contains $4n$ chromosomes, with sister chromatids tightly bound together by cohesins. Homologous chromosomes undergo some degree of presynaptic pairing as they enter Prophase I, mediated at least in part by the meiosis-specific cohesin RAD21L [8]. Each chromosome pair also begins the first steps in assembling the Synaptonemal Complex (SC), a large proteinaceous "zipper" that aligns each homologous chromosome pair along their entire length throughout Prophase I and is required for proper

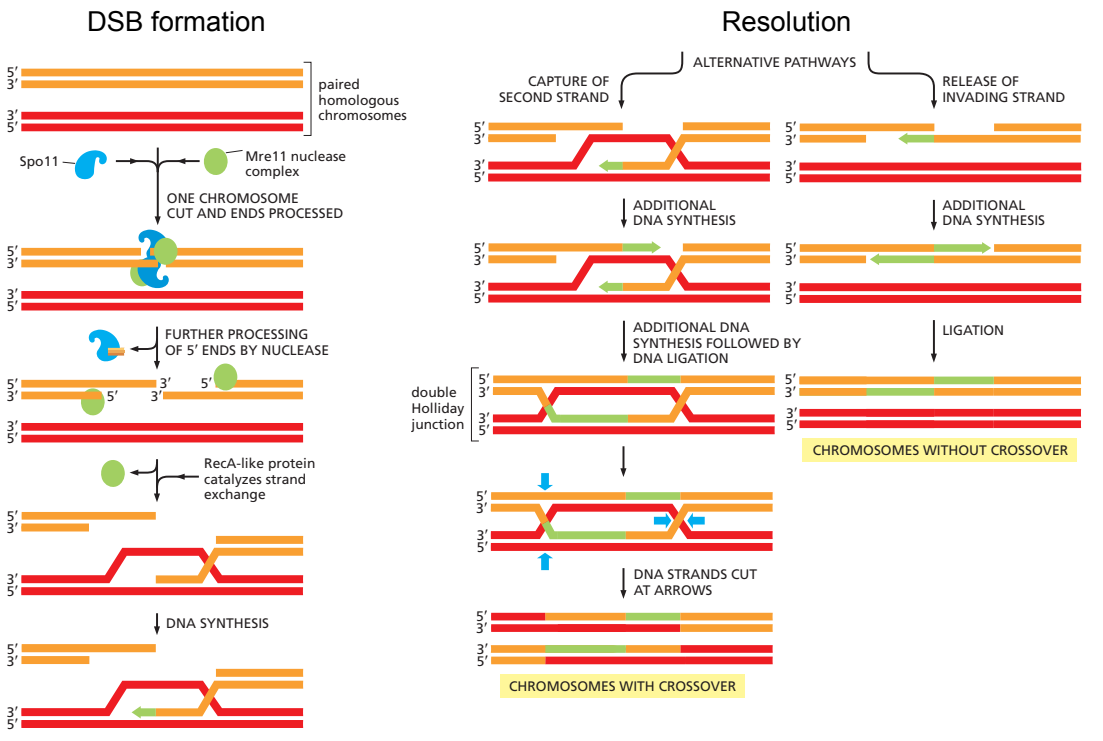


Reproduced from Baudat *et al.* (2013) *Nat. Rev. Genet.*

Figure 1.2: The four stages of Prophase I. Top: an illustration of the appearance of the Synaptonemal Complex throughout Prophase I. Middle: a timeline of key events in recombination and sex body formation. Bottom: Immunofluorescence images of mouse spermatocytes labelled for SC Axial Element protein SYCP3 (red) and various other markers illustrating meiotic progression. a: MEI4 co-localizes with AE proteins and is required for Double Strand Break formation, b: gammaH2AX is a phosphorylated histone variant that spreads around DSB sites, c: DMC1/RAD51 are loaded onto ssDNA after end resection around DSBs and help facilitate homology search, d: MLH1 specifically marks Double Holliday Junctions at crossover sites, e: DSBs on the sex chromosomes remain unrepaired until late in meiosis, and they are pushed to the periphery of the nucleus into the sex body.

crossover processing on each chromosome. Specifically, several Axial Element (AE) proteins (including the SC-specific SYCP2 and SYCP3) begin to bind alongside the cohesins that hold each sister chromatid pair together (reviewed in [9]).

During Leptotene, the cell also begins to induce 200-400 Double-Strand Breaks (DSBs) using the meiosis-specific SPO11 endonuclease in complex with many other proteins [10, 11]. SPO11 dimerises at specific sites on the DNA, and each monomer cleaves the phosphodiester backbone of one strand of DNA and becomes attached at the 5' end of the cleavage site (**Figure 1.3**). MRE11 (as part of the MRN complex) then catalyses a second endonucleolytic cleavage downstream of the DSB on each strand, releasing two SPO11-oligonucleotide complexes (one with 12-26



© 2014 from *Molecular Biology of the Cell*, Sixth Edition by Alberts *et al.*
Reproduced by permission of Garland Science/Taylor & Francis Group LLC.

Figure 1.3: Formation and repair of Double Strand Breaks in meiosis. SPO11 homodimerizes to form DSBs. Each monomer cleaves one strand and becomes covalently attached to the DNA at the 5' end. MRE11 nicks the downstream DNA, freeing SPO11 attached to a short oligo. Ends are then resected by a 5'-3' exonuclease (likely EXO1), and DMC1/RAD51 are loaded onto the single-stranded overhangs, helping to facilitate strand invasion of the homologous chromosome. Synthesis proceeds, copying the homologous chromosome. 90% of the time, the newly synthesised strand is released and anneals back to its native homologue, forming a non-crossover gene conversion event by Synthesis-Dependent Strand Annealing, which results in a small stretch of DNA being copied from one homologue to the other. Alternatively, a Double Holliday Junction can form and then resolve as a crossover, resulting in the reciprocal exchange of chromosome arms.

nucleotides and one with 28-36) [12].

The resulting single-stranded ends are then further resected, likely by the 5′ to 3′ exonuclease EXO1 [13], and the resulting ∼1-kb single-stranded overhangs are then loaded with RAD51 and the meiosis-specific DMC1 [14], a process requiring several other proteins [15]. Gamma-H2AX, a phosphorylated histone variant, also begins to spread along the chromatin surrounding each DSB [16]. RAD51 and DMC1 stabilise and protect the ssDNA ends and help to facilitate strand invasion and homology search along the sister homologue.

In human cells, the task of homology search at one DSB is comparable to searching for a specific 10 cm string in a genome that stretches 2500 km, while excluding the sister chromatid and tolerating a degree of polymorphism [14]. Furthermore, this task must be completed at hundreds of sites simultaneously. It is likely that presynaptic homologue pairing helps to narrow down this search space. In humans and mice, DSBs tend to avoid centromeres and heterochromatic acrocentric short arms, although they are heavily enriched near telomeres in human males [17]. Homology search appears to be required for proper synapsis, since mice with reduced or no SPO11 activity show defects in synapsis, yielding complete infertility in males and females [18].

1.1.2 Zygotene

During Zygotene, the Axial Elements become completely loaded along each sister chromatid pair, with homologous chromosomes still separated by about 400 nm. As DSBs begin to repair, transverse filaments containing proteins including SYCP1, SYCE1, and SYCE2 begin to zip the AEs together until they are only 100 nm apart, a process called synapsis (**Figure 1.2**; [19]). Synapsis tends to proceed from telomeres, which are tethered to one pole of the nuclear envelope by the protein SUN1, causing the maturing synaptonemal complexes to resemble a bouquet inside the nucleus [15].

After homology search is complete, the resulting DNA heteroduplex is processed by a multitude of recombination proteins, which determine its fate as either a crossover (CO) or a noncrossover (NCO; **Figure 1.3**). Roughly 90% of DSBs are processed as NCO gene conversions, which repair quickly by copying a short stretch of DNA from the homologue at each DSB, a process called Synthesis-Dependent Strand Annealing (SDSA) [15]. Most of the remaining 10% will each form a Double Holliday Junction (DHJ), a structure with two mutual heteroduplex stretches binding the homologues together at the site of strand invasion, which will eventually be processed as a CO. Some COs may form *via* another, less well characterised pathway, which may involve single Holliday Junctions [15, 20]. Unlike NCOs, COs involve the reciprocal exchange of entire chromosome arms,

although nonreciprocal gene conversion tracts can occur near the CO breakpoint (**Figure 1.3**). Each chromosome pair requires at least one crossover site to ensure proper pairing at Metaphase I. Failure to form an obligate crossover can lead to missegregation and aneuploidy [15].

1.1.3 Pachytene

At the onset of the Pachytene phase, every chromosome has completely synapsed and formed a complete SC, with the exception of the sex chromosomes in males. The number of DMC1 foci decreases dramatically as most DSB sites complete repair as NCOs, while the protein MLH1 specifically marks DHJs at CO sites. A process called crossover interference prevents COs from occurring near each other along the chromosome and may play a role in limiting the total number of crossovers in meiosis [15]. The SC (although its role continues to be debated) may help to provide a platform for the CO machinery to operate unimpeded by the surrounding chromatin, and it might help to enhance homology search and to ensure that COs occur only between homologues and never between sister chromatids [9].

In spermatogenesis, the unsynapsed X and Y chromosomes are pushed to the periphery of the nucleus, where they form the sex body. They pair only at their PseudoAutosomal Regions (PAR), with DSBs along the rest of the chromosome remaining unrepaired and marked by gamma-H2AX until late in Prophase I (when they likely repair from their sister chromatids) [15]. Genes along the X and Y chromosomes undergo Meiotic Sex Chromosome Inactivation, a specific type of Meiotic Silencing of Unsynapsed Chromatin, although some genes such as certain X-chromosome miRNAs can escape this silencing [21, 22].

1.1.4 Diplotene

Once recombination is complete, the transverse filaments of the SC are removed, causing desynapsis, but homologous chromosomes remain joined at their crossover sites, visible as structures called chiasmata. Chromosomes “condense” further and become visible as tetrads. Chiasmata, and thus crossover sites, are essential

structures for the proper alignment of homologues at the Metaphase I plate, providing the tension necessary for correct spindle formation [19]. Chiasmata dissociate when inter-sister cohesins are cleaved, releasing homologues from each other and signalling the beginning of Anaphase I.

In females, meiosis arrests at the end of diplotene, called the dictyate stage, until puberty. Thus, females must maintain their chiasmata for decades. The degradation of these structures over time may explain why oocytes have a ∼25% aneuploidy rate compared with only 2% for sperm, and this may explain why rates of aneuploidy increase with maternal age [9].

1.2 The effects of recombination and the discovery of hotspots

Efforts to precisely measure the genetic effects of recombination date back to Morgan [23]. By tracking the rate of co-transmission of alleles on the same chromosome through pedigrees, one could map the order of their respective genes with scaled distances proportional to the crossover rate between them. Crossover rates are reported in units of centiMorgans per Megabase (cM/Mb), equivalent to a 1% probability of a crossover event occurring within 1 megabase of DNA. In humans, the genome-wide mean crossover rate is around 1.1 cM/Mb [24]. These genetic maps proved essential for genetic research until the recent advent of whole-genome sequencing made physical mapping of genes much more efficient. However, maps of recombination rates continue to prove essential in combination with genome assemblies for studies of population history and for Genome-Wide Association Studies [24–26]. Recombination rates can be inferred not only by examining linkage of markers through pedigrees, but by examining measures of the correlation between alleles in randomly sampled individuals, called linkage disequilibrium (LD) [25]. Where recombination rates are high, flanking markers tend to be less correlated, and thus the accumulated effects of sex-averaged historic recombination can be directly observed in our genomes today.

Early low-resolution genetic maps revealed that large-scale recombination rates vary among sexes, individuals, and chromosome regions [24, 27]. More recent high-resolution genetic maps based on linkage disequilibrium patterns revealed that 80% of human recombination events cluster in only 10% of the genome, primarily in ∼2-kb regions called recombination hotspots [25, 28]. Recombination hotspots have also been described in other organisms such as chimps and mice, although their positions are not conserved [29–31]. Furthermore, the use of certain human recombination hotspots has been shown to be enriched in different populations [26, 32]. The very existence of these hotspots seemed surprising, since they concentrate the activity of recombination within narrow regions and thus limit haplotypic diversity between hotspots. This reduces the ability to efficiently unlink mutations within the same haplotype, which would seem to counter the aim of recombination itself. Theory would suggest that recombination proceeds best when distributed uniformly across the genome [5]. The answer to this paradox may lie in the complex and multifunctional molecular underpinnings of meiotic recombination.

1.3 PRDM9 determines recombination hotspot locations in humans and mice

Several lines of evidence have recently identified the protein responsible for this diversity of hotspot positioning: PR domain containing 9 (PRDM9), a meiosis-specific protein containing a zinc finger (ZF) domain capable of binding DNA at specific sequence motifs. *Ab initio* motif-finding efforts in the sequences of human hotspots revealed that roughly 40% contain a 13-bp sequence motif [33] that closely matches the *in silico* predicted binding site of PRDM9 [34], and binding of PRDM9 to this motif has been confirmed *in vitro* [35]. Furthermore, *PRDM9* alleles that bind different motifs have been associated with polymorphic hotspot usage in humans and mice [26, 32, 36, 37]. Finally, knockout of PRDM9 in mice has been shown to relocate recombination initiation events to entirely different sites [37]. However, the exact mechanisms by which PRDM9 promotes hotspot formation remain poorly understood. For example, in humans, only a small subset of regions

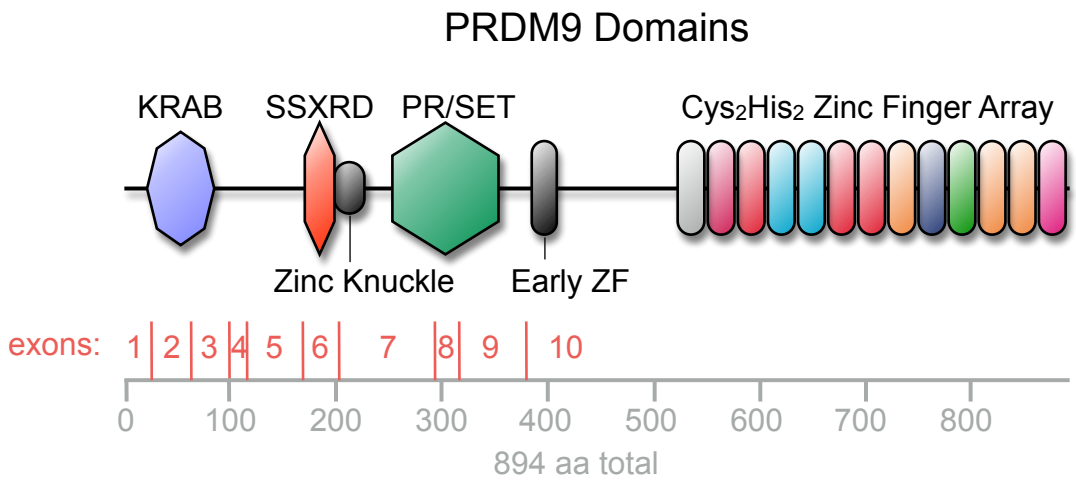


Figure 1.4: Domain annotation of PRDM9. This illustrates the scaled sizes and positions of the annotated domains of the Human B allele of PRDM9. The ZF array binds DNA [38], the PR/SET domain trimethylates Histone H3 at lysine 4 (H3K4me3) [38], the Zinc Knuckle and Early Zinc Finger interact to inhibit the PR/SET domain [39], and the KRAB and SSXRD domains have unknown functions. Exon boundary positions are also shown. The entire ZF array as well as the Early Zinc Finger are encoded by the final exon.

matching the PRDM9 binding motif yield hotspots, and not all hotspots contain PRDM9 binding motifs, suggesting that the presence of a binding motif is neither necessary nor sufficient for hotspot formation [34]. However, experiments in mice show that PRDM9 is responsible for localising nearly all recombination hotspots [37]. Thus, additional factors must strongly influence where PRDM9 binds and whether it promotes hotspot formation when bound.

1.4 DNA-binding activity of the Zinc Finger domain

The final exon of PRDM9 contains a tandem array of 12 canonical Cys_2-His_2 zinc fingers, each encoded by 28 amino acids (**Figure 1.4**). Each zinc finger contains two structural elements: an anti-parallel beta sheet and an alpha helix, with conserved cysteine and histidine residues that together coordinate a zinc ion [40]. Four particular amino acid residues, within the N-terminal (*i.e.* upstream) region of the alpha helix, contact DNA. These are referred to by their positions relative to the

start of the alpha helix domain: -1, 2, 3, and 6 [41]. Each zinc finger binds a 3-bp footprint in the major groove, with base specificity determined primarily by the residues at positions -1, 3, and 6, respectively (with position 2 interacting with the adjacent zinc finger's target). *In silico* predictions of DNA binding specificities have enabled the rational design of Cys_2-His_2 Zinc Finger arrays to target specific DNA sequence motifs, although complex interactions between zinc fingers remain difficult to account for, and each new combination of zinc fingers typically requires optimization [42]. Crystal structures of ZF-DNA complexes also seem to indicate that no more than three zinc fingers can bind DNA simultaneously, due to differences in the spacing and angles between DNA bases and those of the zinc fingers [41]. Thus, it has remained puzzling how proteins with long, naturally occurring ZF arrays bind DNA. At the N-terminal end of the PRDM9 ZF array is a degenerate zinc finger missing the first zinc-binding cysteine residue, which is conserved across humans, chimps, and mice (**Figure 1.5**). In humans, this first zinc finger (not counted amongst the 12 canonical fingers) is degenerate, because it lacks a canonical terminal cap motif [43]. The function of this first zinc finger remains unknown. In the same exon as the zinc finger array, but 112 residues upstream, is another degenerate zinc finger, called the early zinc finger, which retains the appropriate Cys_2His_2 residues to bind zinc, but lacks a terminal cap and is unlikely to be able to bind DNA [43]. This early zinc finger is strongly conserved across species, suggesting some important function, but the region linking it to the ZF array is not [44].

Within PRDM9's ZF array, few residues differ between zinc fingers, apart from those at the DNA-contacting positions [34]. Because of its tandem organisation and repetitiveness, the ZF array of *PRDM9* constitutes a minisatellite repeat, with a repeat unit size of 84 bases. This repetitive structure lends itself to rapid mutation by unequal crossover within the array, which can yield insertions or deletions of entire subsets of zinc fingers, dramatically altering its DNA-binding properties [45]. This dynamic ability to evolve may explain why the Cys_2-His_2 ZF domain is the most common protein motif in the human genome, occurring in over 800 human proteins, compared to only 40 in yeast [46, 47]. Cys_2-His_2 zinc fingers have also

been shown to bind RNA, other proteins, or each other [40], although these versatile behaviours have been characterised far less than their DNA-binding properties.

1.5 H3K4 trimethylation activity of the PR/SET domain

In addition to a zinc finger (ZF) domain, the PRDM9 protein contains a PR/SET domain (**Figure 1.4**), which is capable of trimethylating histone variant 3 at lysine 4 (H3K4me3) on its N-terminal tail [38]. This histone mark is enriched in active gene promoters and enhancers [48, 49] and, in meiosis, it serves as a necessary but insufficient mark of recombination hotspots prior to double strand break (DSB) formation by SPO11, one of the first steps in the initiation of recombination [50, 51]. *Prdm9* knockout mice fail to establish the meiosis-specific H3K4me3 marks associated with recombination hotspots, and their meiotic DSBs relocate to other regions already enriched for H3K4me3, such as gene promoters and enhancers [37]. This suggests a direct role for PRDM9 in positioning recombination events away from functional elements.

A recent biochemical and structural study of PRDM9's PR/SET domain revealed that it is capable of mono-, di-, and tri-methylation activity [39]. Furthermore, it demonstrated that the PR/SET domain can undergo an auto-inhibitory conformational change, which is mediated by an interaction between the Early Zinc Finger and another zinc-binding domain upstream of the PR/SET, called the Zinc Knuckle domain [39]. The biological relevance of this auto-inhibition remains unknown, but it may help to prevent promiscuous trimethylation by PRDM9 not bound to DNA.

Notably, although the PR/SET domain is highly conserved, it does not always co-occur with the ZF array. Different isoforms of PRDM9 detected in mice exclude the exon containing the ZF domain [38] and, in *S. cerevisiae*, the only H3K4me3 methyltransferase lacks a DNA-binding domain altogether [52]. Similar to the DSB pattern seen in *Prdm9* knockout mice, *S. cerevisiae* hotspots tend to occur in promoters [53]. Because H3K4me3 is a marker of expressed gene promoters, it has also been suggested that PRDM9 may act as a transcriptional activator in meiosis

[38]. In fact, tethering PRDM9 to a protein that binds the promoter of a reporter gene results in transactivation of that gene, and this transactivation activity is disrupted by engineered loss-of-function mutations in the PR/SET domain [38]. Furthermore, *Prdm9* knockout mice show reduced expression of a gonad-specific gene called *Morc2b* coinciding with reduced H3K4me3 at its promoter [38]. However, the extent of PRDM9's role as a meiotic transcription factor has not been established, and no gene has yet been proven to be directly regulated by PRDM9 binding.

1.6 N-terminal domains of unknown function

Unlike the increasingly well-understood PR/SET and ZF domains, the N-terminal region of PRDM9 contains two conserved domains whose specific functions remain unknown: the KRAB domain and the SSXRD domain (**Figure 1.4**). This region of *PRDM9* is paralogous to a gene called SSX1, suggesting that *PRDM9* descended from an ancient fusion of an *SSX*-like gene containing the KRAB and SSXRD domains and a *PRDM*-like gene containing the PR/SET and ZF domains [44, 54]. However, PRDM9 has since diverged considerably, sharing only 44% identity with SSX1 at the KRAB domain and only 32% at the SSXRD [44]. While most KRAB-containing Cys_2-His_2 ZF proteins are involved in transcriptional repression by directly binding a corepressor called KAP1 [55], both PRDM9 and SSX1 lack a critical KAP1 binding motif [54, 56]. Indeed, SSX1's KRAB domain fails to bind KAP1 or repress reporter genes [54, 56]. By extension, PRDM9's KRAB domain might not be involved in the KAP1-mediated repressive pathway. However, PRDM9's KRAB domain may have evolved to bind a different set of interacting proteins. The other paralogous domain, the SSXRD of SSX1, has been shown to have strong repressive activity in reporter assays [54, 56]. Additionally, SSX1's SSXRD has been shown to bind directly to histone cores [57] and to a DNA-binding transcriptional activator called LHX4 [58]. It remains to be determined which proteins are bound by PRDM9's KRAB and SSXRD domains or if these domains affect local gene expression or chromatin modifications. One possibility

is that these domains are responsible for binding and recruiting recombination-associated proteins to hotspot loci.

1.7 Meiotic arrest phenotypes in *Prdm9* knock-out mice

Interestingly, *Prdm9* knockout mice still form DSBs, but they fail to complete repair of these breaks, and they arrest in the Pachytene stage of Prophase I, resulting in sterility in both sexes [38]. They also fail to synapse chromosomes properly and exhibit abnormal sex-body formation. This suggests that PRDM9 may be responsible for recruiting or promoting the expression of meiotic DSB repair proteins. Alternatively, PRDM9 may simply sequester DSBs away from regions where repair may be impaired, which might include promoters and enhancers [37]. Consistent with this sterility phenotype in knockout mice, missense mutations in human PRDM9 have been suggested to be associated with male infertility due to azoospermia[59, 60]. However, despite its indispensable role in mouse gamete production and its presence in diverse metazoan organisms [61], *PRDM9* appears to be non-functional in dogs due to a fixed loss-of-function mutation present in the dog genome assembly [62]. This suggests that dogs may use a PRDM9-independent mechanism for recombination initiation, and as in yeast and knockout mice, dog recombination hotspots tend to occur at gene promoters [63]. One recent pre-printed study also identified a fertile woman with two pseudogenized copies of *PRDM9*, implying that in this individual, PRDM9-independent mechanisms are able to rescue fertility in its absence [64].

1.8 Rapid evolution and high diversity of *PRDM9*

Given its critical role in fertility, one might also expect *PRDM9* to be highly conserved and under strong purifying selection. However, *PRDM9* shows signatures of rapid evolution specific to zinc finger residues that affect DNA binding, and implicating strong positive selection at these residues [61]. In fact, *PRDM9* is the

most rapidly evolving zinc finger gene in the human genome [34]. One possible driver of this rapid evolution is the "hotspot conversion paradox", which results from the fact that sequences surrounding recombination initiation sites are cut, resected, and replaced using the homologous chromosome as a template [65, 66]. Thus, if one homologue contains a "hot" recombinogenic allele and the other contains a "cold" non-recombinogenic allele, the cold allele will be used as a template far more often, driving the hot allele to extinction [65, 66]. The effects of this phenomenon have been confirmed in human and mouse binding motifs, which have accumulated motif-disrupting mutations rapidly over time [34, 67, 68]. Therefore, *PRDM9*'s rapid evolution may be explained in part by the disappearance of its preferred binding targets, which has been referred to as the Red Queen Model [68, 69]. Alternatively, hotspots may need to turn over quickly in order to avoid promoting deleterious mutations that arise near "cold" binding sites [66]. It has also been suggested that *PRDM9* may evolve to avoid binding to rapidly evolving repetitive sequences [61]. However, as will be explored in Chapter 3 of this thesis, our work suggests a more nuanced mechanism by which hotspot death drives the rapid evolution of *PRDM9*.

Sequencing of *PRDM9* in mouse, human, chimp, and other primate populations has revealed tremendous extant diversity [35, 70–73]. Although many rarer human alleles have been discovered, 80-90% of European alleles are of the A type [32, 34]. The human reference sequence contains the 12-ZF B allele (shown in **Figure 1.5**), which is predicted to bind essentially the same sites as the A allele but occurs at lower frequency [17, 32]. In African populations, however, different "C-type" alleles occur at a frequency of up to 34% [32], and they bind a motif completely different than that of the A allele [26]. A chimp recombination map revealed that chimps also have hotspots, but they are much weaker than human hotspots and no single motif could be found within them, perhaps owing to the diversity of PRDM9 alleles in chimp populations [31]. One of the more common chimp alleles, the W11a allele [31], contains 18 zinc fingers and occurs at a frequency of roughly 13.4% (alleles Pan.t.-4,8,12,16 from [73]).

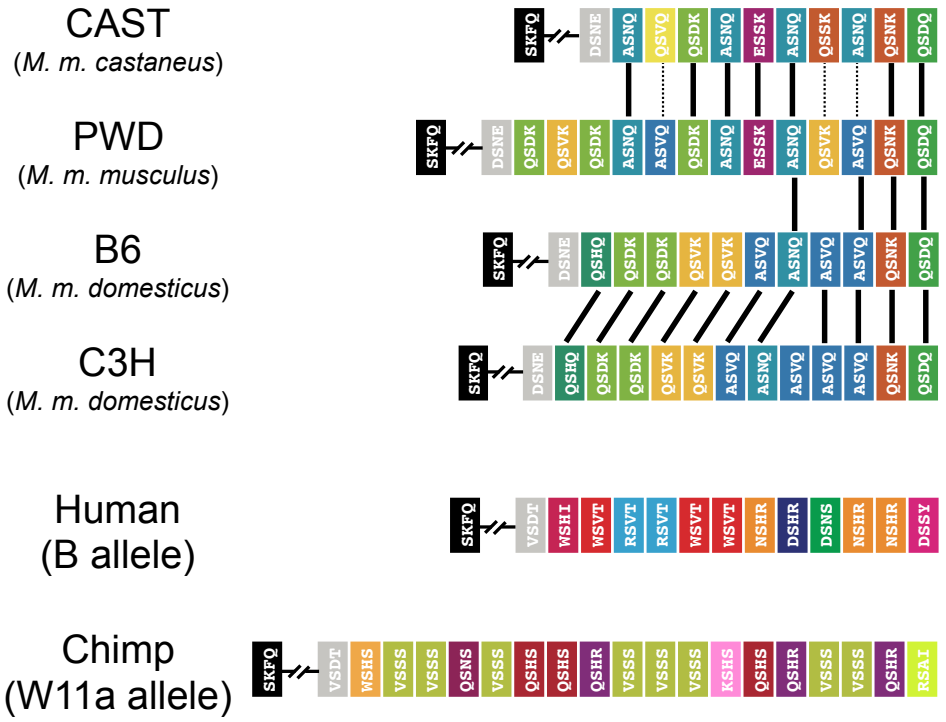


Figure 1.5: Variation in the ZF array between species and subspecies. Each coloured block represents a canonical Zinc Finger and is labelled with the residues found at DNA-contacting positions -1, 2, 3, and 6. Colours distinguish ZFs with different amino acid combinations at these four residues. Black blocks represent the conserved early zinc finger, and grey blocks represent the first, degenerate zinc finger in each array. The first four arrays correspond to four different mouse strains, with their primary subspecies identity listed below each name. Vertical solid black lines between arrays highlight identical aligned zinc fingers, while dotted lines show one-off identity at the DNA-contacting residues. B6 and C3H differ only by one tandemly duplicated zinc finger, and the CAST/PWD arrays are much more similar to each other than to B6. There appears to be no ZF array conservation between species, except at the early zinc finger.

Classical strains derived from the three subspecies of *Mus musculus*, which diverged roughly 350 kya [74], contain three distinct PRDM9 alleles [70, 71] (**Figure 1.5**). The 10-ZF PRDM9 allele present in the *Mus musculus castaneus* strain CAST/Eij (hereafter, CAST) appears to be the most ancestral [67], with a similar structure to the 13 ZF-allele found in the *Mus musculus musculus* strain PWD/Ph [71]. The *Mus musculus domesticus* strain C57BL/6 (hereafter, B6) has a very different 11-ZF allele predicted to have diverged shortly after isolation from the

other subspecies [67, 71]. A different *M. m. domesticus* strain, C3H, has a 12-ZF array very similar to that of the B6 allele, differing only by the tandem duplication of the eighth zinc finger, probably arising from a recent unequal crossover event within the array [75]. Strong erosion of each allele's binding targets has been demonstrated on the B6 and CAST lineages [67].

1.9 The role of PRDM9 in hybrid infertility

Before the discovery that PRDM9 controls the locations of recombination hotspots, an earlier landmark study had already identified it as the first mammalian speciation gene [75]. Crosses between certain strains of *M. musculus musculus* and *M. musculus domesticus* yield sterile male offspring [76], indicating that they are currently undergoing speciation in accordance with Haldane's Rule, which observes that hybrid infertility almost universally arises first in the heterogametic sex [77]. Classic mapping experiments identified the causal locus, called *Hst1*, to be in a 255-kb candidate region on chromosome 17, which contains *Prdm9* (previously named *Meisetz*) and five other genes [78–81]. Mihola *et al.* then created transgenic mice containing randomly integrated Bacterial Artificial Chromosomes (BACs, ∼150-kb cloned DNA regions) spanning the Hst1 region from a different *domesticus* strain known to produce fertile hybrid offspring. They showed that only mice with BACs containing *Prdm9* produced fertile offspring, while other possible candidates were ruled out by exclusion or by previous studies, thereby confirming PRDM9 as the causal hybrid infertility gene [75].

Sterile male hybrids show similar phenotypes to *Prdm9* knockout mice, in that they fail to repair DSBs and become arrested at pachytene [75, 76]. They also show reduced synapsis and impaired sex body formation [75, 82]. However, the complete hybrid sterility phenotype is present only in males and only in one direction of the cross, requiring the presence of multiple possibly interacting loci [83], implying a complex mechanism by which *Prdm9* divergence yields hybrid infertility.

Specifically, PRDM9 has been shown to play an essential role in the complete infertility of F1 hybrid males generated by crossing a *Mus musculus musculus*

Class	Sterile	Semisterile	Semifertile	Fertile	Fertile
TW (mg)	45 to 70	70 to 90	90 to 140	above 140	above 180
SC ($\times 10^6$)	0.00	0.01 to 0.2	0.2 to 1.1	above 1.1	above 3.5
OFM	0.00	0.1 to 1	3 to 4	above 4	above 5
Background:	*Prdm9*[a]:	*Prdm9*[a]:	*Prdm9*[a]:	*Prdm9*[a]:	*Prdm9*[a]:
PWD×B6	PWD/B6	PWD/−		PWD/B6+2B6	PWD/PWD
PWD×B6			PWD/C3H	PWD/−+2C3H	PWD/−+6C3H
PWD×B6				PWD/B6+2C3H	PWD/B6+6C3H
B6×PWD			B6/PWD	−/PWD	PWD/PWD
B6×PWD				C3H/PWD	B6+2C3H/PWD
STUS×B6	STUS/B6			STUS/B6+2B6	
B6×STUS	B6/STUS			B6+2B6/STUS	
B6×B6	−/−				B6/−
B6×B6					B6/B6
B6×B6					PWD/B6
B6×B6					B6+2C3H/B6

TW, mean testicular weight; SC, mean sperm count in paired caput epididymides, OFM, offspring per female per month;
[a]genotype at *Prdm9* (maternal/paternal); −, null; +, added transgenic copies of *Prdm9*; Background: maternal×paternal background. Note that the two fertile classes display overlapping parameters.

Reproduced from Flachs *et al.* (2012) *PLoS Genet.*

Table 1.1: Fertility measures across hybrid mice with varying alleles and dosages of PRDM9

PWD female with a *Mus musculus domesticus* B6 male. Hereafter, I will refer to these as (PWD×B6)F1 hybrids (with the maternal strain always preceding the paternal strain), or less formally as the "Infertile Hybrid" mice (and I use the terms "sterility" and "infertility" interchangeably throughout). The reciprocal cross of (B6×PWD)F1 yields semi-fertile male offspring, able to sire pups, but with reduced fertility parameters such as lower sperm count and smaller testes [76, 84]. Efforts to dissect the genetics of this hybrid infertility phenotype have revealed three necessary genetic conditions for complete F1 sterility: 1) *Prdm9* must be heterozygous, with exactly one copy of the PWD allele and one copy of the B6 allele—changing the copy number of the B6 allele or adding copies of other alleles can partially or fully restore fertility (see **Table 1.1**, from [84]); 2) a 4.7-Mb region on the X chromosome called *Hstx2* must originate from the PWD background [85]; and 3) the autosomes must be heterosubspecific—introducing long homozygous stretches can reduce asynapsis of that chromosome [82]. Without additional data, it has remained difficult to postulate a mechanistic model to marry these three conditions, which I review below:

1.9.1 *Prdm9* heterozygosity and dosage

The classic Infertile Hybrid mice have one maternally inherited copy of the PWD allele of *Prdm9* and one paternally inherited copy of the B6 allele, which differ significantly in zinc finger composition and thus in their predicted binding motifs (see **Figure 1.5**). The original experiments demonstrating *Prdm9* as a causal hybrid sterility gene showed that insertion of BACs containing *Prdm9* from C3H, a different *domesticus* strain that produces fertile offspring when crossed with PWD, could rescue fertility in hybrid mice [75]. Interestingly, the C3H *Prdm9* ZF array differs from the B6 allele only by the insertion of one duplicated zinc finger (see **Figure 1.5**), although they also differ at other sites outside the ZF array [75]. Although this experiment definitively identified *Prdm9* as the causal *Hst1* gene, it remained possible that fertility was restored merely by the resulting increase in *Prdm9* dosage. Indeed, they showed that (PWD×B6)F1 hybrid mice with six integrated copies of the C3H allele showed even larger measures of fertility than hybrid mice with only two integrated copies, but no such effect of *Prdm9* copy number on fertility was evident in the parental strains [75].

Subsequent experiments showed that varying degrees of fertility could be rescued by inactivating the B6 *Prdm9* allele, by replacing it with a consomic Chr17 containing one copy of the C3H allele or the PWD allele, or by adding in extra copies of the B6 or C3H alleles (see Table 1.1) [84]. Other experiments showed that increasing the copy number of $PRDM9^{B6}$ could rescue fertility in otherwise sterile F1 hybrids of STUS (a different *M. m. m.* strain) and B6 [84]. Furthermore, they showed that reducing or increasing the $PRDM9^{B6}$ dosage in the semi-fertile reciprocal cross of B6×PWD made these mice fully fertile (see Table 1.1) [84]. Altogether, these elegant experiments demonstrated a perplexing dosage and allele dependence for PRDM9, with complete infertility arising only from mice with $PRDM9^{PWD/B6}$ on a PWD×B6 background.

1.9.2 *Hstx2*

The fact that (PWD×B6)F1 hybrids are completely sterile while (B6×PWD)F1 hybrids are semi-fertile implies the requirement of an additional sex-linked hybrid sterility locus, called *Hstx2*. Interestingly, it had previously been shown that introgression of a consomic PWD X chromosome into a B6 background was sufficient to cause sterility in males (although spermatogenic arrest occurs post-meiotically, unlike the Infertile Hybrids). Efforts to map this trait, called *Hstx1*, narrowed it to a large region in the middle of ChrX, although other regions on ChrX were required for the full sterility phenotype [86, 87]. Subsequent efforts to fine-map *Hstx1*, as well as the sex-linked hybrid sterility locus, *Hstx2*, narrowed them both to the same 4.7-Mb region on the PWD X chromosome [83, 85]. This region contains 11 protein coding genes (7 with high testis expression) and 20 micro RNAs (which escape meiotic sex chromosome inactivation) [85]. Efforts to characterize these candidates have not yet identified the causal locus (or loci). Finally, the Y chromosome and mitochondrial genome have been formally ruled out as contributors to the Hstx2 phenotype [83].

1.9.3 Heterosubspecificity

Experiments crossing subconsomic C57BL/6J-ChrX.2$^{\mathrm{PWD}}$ (here, abbreviated B6-X.2$^{\mathrm{PWD}}$) with consomic B6-17$^{\mathrm{PWD/B6}}$ showed that the combination of *PRDM9*$^{\mathrm{PWD/B6}}$ and *Hstx2*$^{\mathrm{PWD}}$ in an otherwise B6 background is not sufficient to cause meiotic arrest [82, 83]. This implies that other interacting loci may be required for full infertility, but Quantitative Trait Locus (QTL) mapping efforts could only identify weak associations on Chr13 and Chr14 [83]. However, not even introgressing Chr2$^{\mathrm{PWD/B6}}$ and Chr14$^{\mathrm{PWD/B6}}$ with B6-X.2$^{\mathrm{PWD}}$ and B6-17$^{\mathrm{PWD/B6}}$ was sufficient to reconstitute the hybrid infertility phenotype, implying the possibility that many weak hybrid loci are required to produce the full hybrid sterility phenotype [83]. Alternatively, the F1 hybrid background itself could be the missing requirement; that is, full infertility might require high chromosome-wide sequence divergence between heterosubspecific homologous chromosomes, which is found in the F1 hybrid background but not in the various consomic crosses. Indeed, elegant cytological

experiments showed that chromosome-specific asynapsis rates decreased dramatically on consomic PWD chromosomes 17 or 19 in an otherwise hybrid background in spermatocytes. Additionally, asynapsis rates in Infertile Hybrids were shown to vary considerably across chromosomes, with smaller chromosomes failing to synapse more frequently [82]. Furthermore, the effect of reduced asynapsis at consubspecific chromosomes relative to heterosubspecific chromosomes was also demonstrated in oocytes, although this effect did not depend on *Hstx2* [85]. This implies that infertility might result from the combined effects of asynapsis at particularly sensitive heterosubspecific chromosomes, and that this asynapsis (and the resulting infertility) can be reversed by making those sensitive chromosomes consubspecific. This may explain why introgression of Chr19 rescues a small number of Infertile Hybrid spermatocytes from complete pachytene arrest (although they still fail to produce functional sperm) [83].

1.10 A note on *M. m. castaneus* and other hybrid crosses

The three subspecies of *Mus musculus* emerged from what was essentially a trifurcation event roughly 350 kya [74, 88]. *M. m. domesticus* exists primarily in western Europe and *M. m. musculus* exists primarily in eastern Europe, with only a narrow hybrid zone with limited gene flow between the two subspecies [89], and *M. m. castaneus* is found primarily in southeast Asia. The classical inbred lab strain C57BL/6 (B6) is of 95% *domesticus* origin; the wild-derived PWD/Ph (PWD) inbred strain is of 93% *musculus* origin; and the wild-derived CAST/Eij (CAST) inbred strain is of 88% *castaneus* origin [90]. Deviations from 100% are potentially due to subspecies admixture prior to capture and inbreeding. For example, PWD contains a ∼20-Mb haplotype of *domesticus* origin on Chr10 [90].

Sequencing of these strains revealed that PWD and CAST have nearly identical sequence divergence levels compared with the B6 reference sequence, with roughly one SNP occurring every 135 bases, which is nearly an order of magnitude higher than the sequence divergence between any two humans [88, 91]. However, while

crosses between PWD and B6 yield completely sterile F1 hybrid offspring, crosses between CAST and B6 do not. Close examination of meiotic cells from F1 hybrids of CAST and WSB (another *domesticus* strain) showed that they exhibit higher rates of apoptosis and improper sex body formation [92].

Although PWD×B6 hybrids are the classic model of F1 hybrid infertility in mice, other combinations of strains can also produce infertile offspring. For example, crosses of the *musculus* strain STUS produce infertile offspring when crossed with B6 in both directions. PWK is another *musculus* strain derived from wild mice caught near the origin of PWD outside Prague [93]. (PWKxB6)F1 hybrids are semisterile, but not totally infertile, and this phenotypic difference from (PWD×B6)F1 hybrids has been mapped to the *Hstx2* region of ChrX [94].

1.11 *In vivo* mapping of H3K4me3 and DMC1 in humans and mice

In vivo experiments to date have mapped the locations of intermediate events in recombination by performing Chromatin ImmunoPrecipitation with high-throughput sequencing (ChIP-seq) against the H3K4me3 mark and the DMC1 mark in testis tissue from mice and humans [17, 37, 50, 95]. The DMC1 ChIP-seq method leverages the fact that DMC1-associated DNA is single-stranded to greatly enhance its signal relative to background [96]. Studies in yeast have directly sequenced SPO11-bound DNA oligomers to map DSB sites [53], but this has not yet been published in mice. Recent studies have also published direct PRDM9 ChIP-seq results using a custom anti-PRDM9 antibody in mice, but they obtained very limited signal due to the transience of PRDM9 activity in testes [67, 97]. A study of DMC1 ChIP-seq in knockout mice showed that in the absence of PRDM9, DSBs tend to occur in regions with pre-existing H3K4me3, including active promoter regions, suggesting that PRDM9 plays a role in directing recombination away from these functional regions in mice [37]. ChIP-seq experiments with micrococcal nuclease digestion have further shown that PRDM9 positions nearby nucleosomes, in addition to

marking them with H3K4me3 [95], and *in vitro* work has suggested that PRDM9 is also capable of forming the H3K36me3 mark [95].

These studies and others have identified short DNA sequence motifs enriched in recombination hotspots and in PRDM9-dependent ChIP-seq peaks, but each motif is consistently much shorter than the predicted size of the DNA-binding footprint of its respective PRDM9 allele [34, 35]. One study showed *in vitro* that two mouse PRDM9 alleles appear to bind with all 11 or 12 of their zinc fingers and are sensitive to base substitutions outside the short predicted binding motif [98]. This suggests that long ZF domains can bind DNA with more than three zinc fingers at a time, which was previously thought impossible by the physical constraints of the Cys_2-His_2 ZF structure [41]. These published binding motifs are neither sufficient nor necessary to predict PRDM9 binding, and it has been suggested that binding is influenced by sequence-independent chromatin features in cis.

1.12 Objectives and aims

Given the difficulty of acquiring and manipulating human meiotic tissue samples, *in vivo* studies of the human PRDM9 protein remain difficult with current experimental methods. On the other hand, *in vitro* studies of isolated PRDM9 suffer from limitations of interpretation due to a lack of cellular and epigenetic context. In light of these issues, we developed an experimental system for studying PRDM9 by transfecting an engineered copy of the *PRDM9* gene into a mitotic human cell line. This approach provides several important advantages over *in vitro* and *in vivo* studies: 1) unlike *in vitro* conditions, mitotic human cell lines provide an approximation of PRDM9's native cellular environment; 2) unlike meiotic cells, mitotic cells can be grown abundantly in controlled tissue culture conditions, providing sufficient material for multiple queries into the effects of PRDM9 activity on the epigenome and on the transcriptome; and, 3) transfecting engineered *PRDM9* permits cost-effective manipulation of the protein in ways that would be infeasible when studying the endogenous copy *in vivo*. I have utilised this system to produce the first map

of direct PRDM9 binding sites across the human genome, revealing novel DNA-binding properties of the ZF array. Then, by comparing this map with existing genome annotations and with *in vivo* H3K4me3 and DMC1 data, I have identified factors influencing the cascade from PRDM9 binding to chromatin marking to DSB formation. Additionally, by analysing these data in conjunction with RNA-seq data, I have present new compelling evidence that PRDM9 can activate endogenous genes by binding to their promoters. Finally, by transfecting different engineered truncations and manipulations of PRDM9, I demonstrate that PRDM9 can form multimers, mediated unexpectedly by the ZF array, and that cells transfected with multiple alleles tend to form homo-multimers more efficiently than hetero-multimers. Using some of the analytical approaches that I developed in these cell line experiments, I built similar maps of PRDM9 activity in hybrid and transgenic mice *in vivo* to investigate the mechanism by which *Prdm9* incompatibility causes mouse hybrid sterility. This work suggests a new role for PRDM9 in meiosis and proposes a mechanism by which lineage-specific hotspot death itself can lead to hybrid infertility. Overall, this body of work sheds new light on key events in the initiation and resolution of meiotic recombination in mammals, a complex process resulting from aeons of evolutionary engineering.

If you are faced by a difficulty or a controversy in science, an ounce of algebra is worth a ton of verbal argument.

— J.B.S. Haldane, attributed by J.M. Smith in *Nature* (1965)

2

Artificial expression of PRDM9 reveals novel DNA-binding modes and gene-regulating capabilities

2.1 Introduction

PRDM9 is expressed early in meiotic prophase [99], during which its C2H2 Zinc Finger (ZF) domain binds DNA and its PR/SET domain marks the surrounding chromatin in *cis* with H3K4me3 [38], a mark found at the promoters of transcribed genes [48]. One recent *in vitro* study has suggested that PRDM9 is also capable of forming the H3K36me3 mark [100]. The functions of PRDM9's other domains, a KRAB domain and an SSXRD domain, remain unknown, but PRDM9 has been hypothesised to play a role in transcriptional regulation or protein-protein interactions [38, 75]. By an unknown mechanism, perhaps mediated by these domains and/or the H3K4me3 mark directly or indirectly, PRDM9 recruits SPO11 to form Double Strand Breaks (DSBs) near a subset of its binding sites [14, 50]. These DSBs undergo end resection and the resulting single-stranded DNA ends are coated with RAD51 and the meiosis-specific DMC1 [14].

In vivo experiments to date have mapped the locations of intermediate events in recombination by performing Chromatin ImmunoPrecipitation with high-throughput

sequencing (ChIP-seq) against the H3K4me3 mark and the DMC1 mark in testis tissue from mice and humans [17, 37, 50, 95]. Recent studies have also published direct PRDM9 ChIP-seq results using a custom antibody in mice, but they obtained very limited signal due to the transience of PRDM9 expression and binding in testes [67, 97]. To study the DNA-binding properties of mouse PRDM9, one study sequenced genomic DNA fragments bound *in vitro* by recombinant proteins containing only the PRDM9 ZF array [97].

Previous studies have identified short DNA sequence motifs enriched in human recombination hotspots and in PRDM9-dependent ChIP-seq peaks, but each motif is typically much shorter than the predicted size of the DNA-binding footprint of human PRDM9, with little or no sequence specificity corresponding to the N-terminal zinc fingers [17, 26, 34]. Other long zinc finger proteins, such as CTCF, have been shown only to bind DNA with a small subset of their zinc fingers, with little specificity contributed by the remaining zinc fingers [101, 102]. One study showed *in vitro* that two mouse PRDM9 alleles appear to bind with all 11 or 12 of their zinc fingers and are sensitive to base substitutions outside the short predicted binding motif [98]. This suggests that long ZF domains can bind DNA with more than three zinc fingers at a time, which was previously thought impossible by the physical constraints of the C2H2 ZF structure [41]. It has remained unknown whether human PRDM9 also binds with its full ZF array (as suggested in [33]), as in the mouse alleles, or whether it only binds with a subset of its zinc fingers, as in CTCF. Furthermore, published PRDM9 binding motifs are neither sufficient nor necessary to predict genome-wide PRDM9 binding, DSBs, or recombination [17, 34, 97], and it has been suggested that binding can be influenced by sequence-independent chromatin features in *cis* [97].

Molecular studies of mammalian meiosis lag far behind studies of mitosis, owing to the fact that mammalian meiosis occurs only in the context of complex intercellular interactions within gonadal tissue (reviewed in [103]), while mitosis can be perturbed and measured rapidly in transformed cell lines that constantly divide *in vitro*. After decades of effort to induce meiosis in mammalian cells *ex vivo*,

new methods have begun to emerge, but they require complicated organ culture techniques [103]. *In vivo* perturbation experiments in meiotic cells have been made possible by genome-editing technologies to produce transgenic mice, but at high costs and over long experimental timescales. Furthermore, such experiments are intractable and unethical in humans, and studies of wild-type human testes and ovaries are limited by tissue availability. *In vitro* experiments of purified proteins, in turn, fail to account for behaviours found only in the context of chromatin and the nucleus, such as histone modification and gene expression.

In light of this, we sought to explore the properties of human PRDM9 by expressing various engineered versions of it in a highly transfectable mitotic human cell line, HEK293T, which can be cultured abundantly and at low cost. Whilst this approach will fail to reproduce certain cell-type-specific phenomena found only in spermatocytes and oocytes, it enables us to learn some of the fundamental rules governing the behaviour of PRDM9 in the nucleus. In these cells, we performed ChIP-seq against human PRDM9, H3K4me3, H3K36me3, and chimp PRDM9, as well as ATAC-seq (Assay for Transposase-Accessible Chromatin with high-throughput sequencing) to examine nucleosome positioning and DNA accessibility, and RNA-seq to examine gene expression. Importantly, this system allows us to compare data from transfected and untransfected cells (in which there is weak to no endogenous *PRDM9* expression), allowing us to compare the same genomic sites with and without PRDM9. This approach also allows us to rapidly engineer and test various different alleles and truncations of PRDM9 to explore the properties of its individual domains.

2.2 Results

2.2.1 A map of direct PRDM9 binding in the human genome

With Nudrat Noor, I first performed ChIP-seq in HEK293T cells transfected with the human PRDM9 B-allele containing an N-terminal GFP tag (see **Figure 2.1**). We sequenced two technical ChIP replicates and one sample of input chromatin to serve as a background control for genome-wide variability in the efficiency of sonication and sequencing. To identify regions bound by PRDM9, I developed

a novel peak-calling algorithm that estimates binding enrichment relative to a measure of local background coverage at each position in the genome, then performs a likelihood ratio test to provide a p-value for evidence of binding. This algorithm then identifies peak centres according to a specified p-value threshold and minimum peak separation value, as well as a confidence interval for the centre of each peak (detailed in Methods). This approach utilises information jointly from both ChIP replicates and from the input chromatin sample when calling peaks, which is more sensitive and better suited to our particular experimental design than more commonly used peak-calling algorithms such as MACS, which processes replicates independently [104]. Our highly localised background correction also enables us to provide more accurate estimates of binding enrichment in regions such as DNAse-hypersensitive sites, which are more prone to shearing during sonication and thus become over-represented in ChIP-seq data [105]. We also provide a genome-wide estimate of the fraction of sequenced DNA fragments representing true binding signal in each ChIP replicate, providing a quality control estimate for the relative purity of each sample (listed in **Table 2.1**).

Using this approach, I called 170,198 PRDM9 binding peaks across the genome ($p<10^{-6}$, minimum peak separation of 250 bp). This demonstrates that PRDM9 can bind with some affinity to many more sites in the genome than the ∼30,000 annotated recombination hotspots first identified by LD breakdown [106]. I compared our ChIP-seq data with a set of 18,343 published *in vivo* human DSB hotspot peaks [17] and found evidence for binding at up to 74% of DSB hotspots (at $p<10^{-3}$) after correcting for chance overlaps (see **Figure 2.2**, and Methods). This demonstrates that even in a completely different cell type and expression system, PRDM9 binds the majority of hotspots. The proportion bound in our system is greater (up to 82%) at DSB hotspots not subject to the telomere effect, which substantially increases the probability of DSB formation within roughly 15 Mb of each telomere in human male meiosis [17]. The probability of overlapping DSB hotspots and testis H3K4me3 ChIP-seq peaks also increases with the strength of PRDM9 binding in

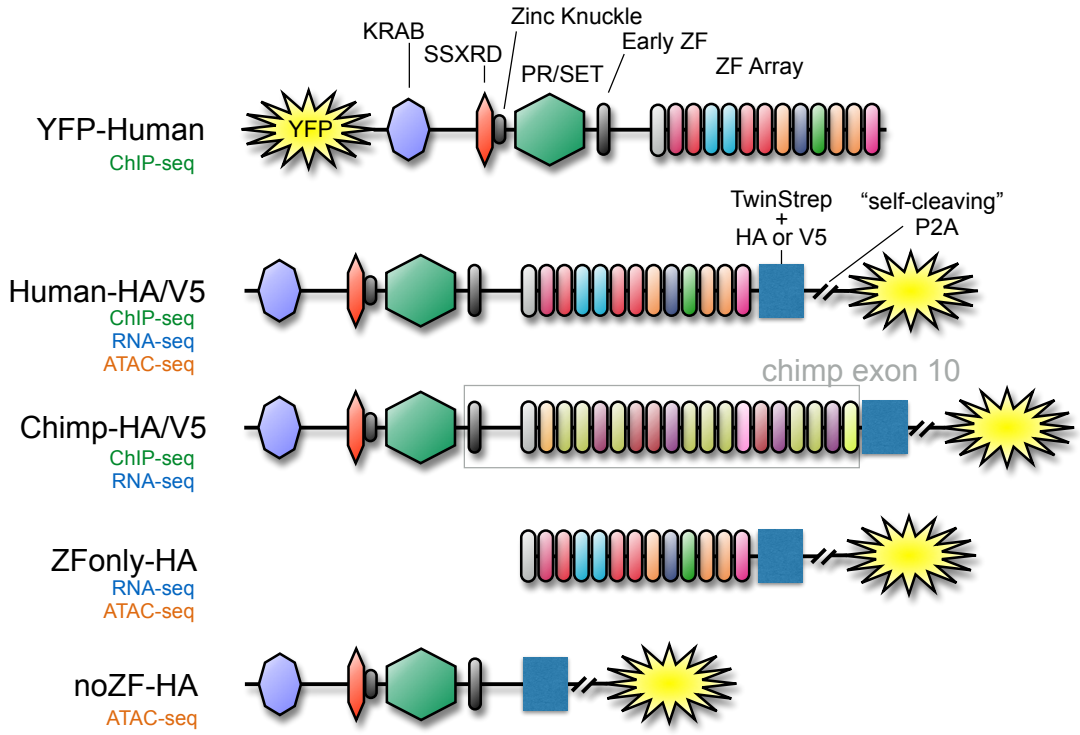


Figure 2.1: Summary of PRDM9 constructs used. Five different tagged PRDM9 cDNAs were generated. Domains are annotated and roughly scaled with their size in the primary amino acid sequence, and the classes of experiments utilising each construct are listed below their names. YFP-Human: Initial experiments used a synthesised cDNA of the human reference ("B") allele with an N-terminal YFP fusion tag, and ChIP-seq was performed using an anti-GFP antibody. Human-HA/V5: subsequent experiments used a different set of tags placed at the C-terminus, all separated by flexible linkers, including a TwinStrep affinity purification tag, a high-specificity HA or V5 epitope tag, and a "self-cleaving" P2A domain that causes the C-terminal YFP tag to fail to attach to the rest of the protein during translation. ChIP-seq was performed using anti-HA and anti-V5 antibodies. Chimp-HA/V5: the same as Human-HA/V5, with the region corresponding to Exon 10 replaced with a synthesised copy of Exon 10 from the chimp reference allele ("W11a"). ZFonly-HA: contains only the human ZF array, beginning at the degenerate first zinc finger. noZF-HA: the exact complement of ZFonly-HA, containing everything upstream of the human ZF array.

	Transfection with:	ChIP antibody	Proportion signal, rep1	Proportion signal, rep2	Fragment number, rep1	Fragment number, rep2	Peak number	Fraction of genome enriched
N-terminal tag	YFP-hPRDM9	GFP	0.225	0.374	80,844,035	79,526,431	170,198	0.023
	YFP-hPRDM9	H3K4me3	0.750	n/a	77,625,060	n/a	470,314	0.096
	YFP-hPRDM9	none (Input)	n/a	n/a	100,861,414	n/a		
	Untransfected	H3K4me3	0.823	0.794	59,156,993	72,839,266	45,758	0.015
	Untransfected	none (Input)	n/a	n/a	98,664,592	n/a		
C-terminal tags	cPRDM9-HAorV5	HA/V5	0.443	0.394	36,662,728	44,594,666	247,717	0.030
	hPRDM9-HAorV5	HA/V5	0.374	0.510	39,385,214	38,717,735	213,885	0.038
	hPRDM9-HA	H3K4me3	0.522	0.544	52,439,279	54,451,480	221,446	0.048
	hPRDM9-HA	H3K36me3	0.334	n/a	59,690,192	n/a	33,625	0.004
	hPRDM9-HA	none (Input)	n/a	n/a	53,219,513	n/a		
	Untransfected	H3K4me3	0.680	0.669	57,205,316	60,883,503	37,932	0.014
	Untransfected	H3K36me3	0.502	n/a	52,368,417	n/a	263,983	0.051
	Untransfected	none (Input)	n/a	n/a	56,445,392	n/a		

Table 2.1: Summary of ChIP-seq datasets. The datasets utilised in this analysis include the initial N-terminal GFP-tagged construct used for most of the analysis as well as the newer C-terminal tagged constructs used in subsequent experiments. Columns 3 and 4 list the proportion of fragments estimated to arise from true signal genome-wide, as computed by our peak calling algorithm. Replicate 2 is assigned "n/a" when only one replicate was performed. Total peak numbers on the autosomes and on the X chromosome are listed in the second-to-last column (HEK293T cells lack a Y chromosome). The final column is an estimate of the proportion of 100-bp bins in the genome with evidence of enrichment at $p<10^{-5}$.

our system, and conversely the probability of overlap increases for hotter DMC1 peaks, especially in non-telomeric regions (see **Figure 2.3**).

To investigate the histone trimethylation activity of PRDM9 and to provide an additional marker of PRDM9 binding, Nudrat Noor and I also performed ChIP-seq against the H3K4me3 mark in both transfected and untransfected cells by the same method. After subtracting sites overlapping "pre-existing" H3K4me3 peaks (those present in untransfected cells), I found that 95% of PRDM9 binding peaks show evidence of local H3K4me3 enrichment ($p<0.01$), and this proportion increases

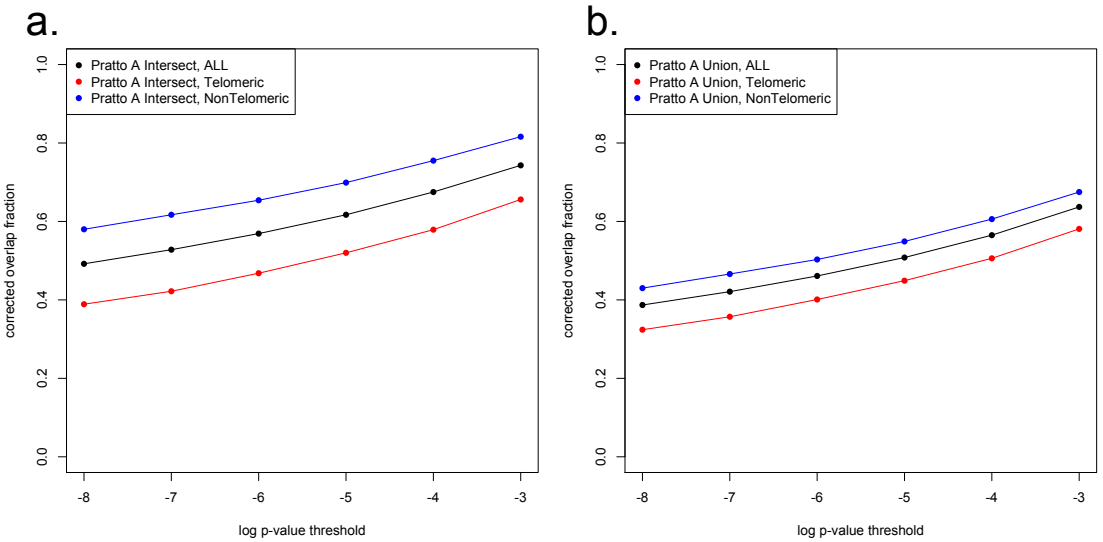


Figure 2.2: PRDM9 binding peaks overlap reported DSB hotspots. I compared our autosomal peaks, called at various p-value thresholds ranging from 10^{-8} to 10^{-3} (minimum peak separation 250 bp), to two sets of published DSB hotspots corresponding to the human A allele (predicted to bind a similar if not identical motif to the B allele [17]): a stringent "Intersect" set of 18,343 hotspots found in multiple individuals (**a**), and a "Union" set of 35,996 hotspots found in any of the samples assayed (after filtering out hotspots wider than 3 kb; **b**). I further split these hotspots into subsets occurring within 15 Mb of a telomere (red lines) or not (blue lines). "Overlap" requires a PRDM9 peak centre to fall within a reported hotspot interval, and overlap fractions were corrected downward to account for chance overlaps (see Methods).

to 100% with increasing PRDM9 binding enrichment (see **Figure 2.3**). That is, PRDM9 makes the H3K4me3 mark essentially everywhere it binds, regardless of the pre-existing chromatin substrate, and the strength of the H3K4me3 signal correlates with the strength of PRDM9 binding ($r = 0.48$) but appears to saturate (see **Figure 2.4**), consistent with the H3K4me3 mark being more persistent (or simply easier to measure) than bound PRDM9.

Next, I compared enrichment values for PRDM9 and H3K4me3 in our cells with *in vivo* testis H3K4me3 and DMC1 enrichment values that I computed from published raw data [17] (see Methods). Both PRDM9 and H3K4me3 enrichment in our HEK293T cells show a nearly identical raw correlation with testis H3K4me3 enrichment ($r = 0.50$), but a much lower raw correlation with testis DMC1 enrichment ($r = 0.21$), consistent with a layer of DSB regulation occurring downstream of PRDM9 binding and H3K4me3 marking (see **Figure 2.4**). However, it should

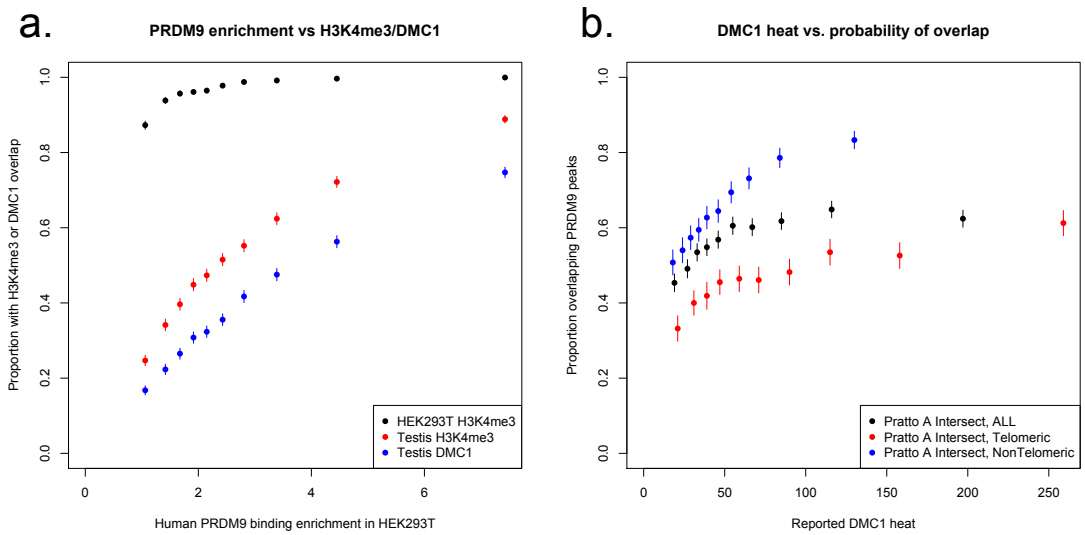


Figure 2.3: Overlap fractions increase with PRDM9/DMC1 enrichment. **a**: H3K4me3 ChIP-seq data from transfected HEK293T cells (this study) and H3K4me3/DMC1 data from testes [17] were force-called in a 1-kb window centred on each PRDM9 binding peak centre ($p<10^{-6}$, minimum peak separation 1000 bp) to provide a p-value for enrichment of each H3K4me3/DMC1 sample at each PRDM9 peak. Peak windows with fewer than 5 input reads from cells or testes were filtered out, to improve enrichment estimates, and windows with excessive genomic coverage (in the top 0.1%ile) or IP coverage (>500 combined fragments) were removed to avoid outliers due to mapping errors or strong, PRDM9-independent H3K4me3 peaks. PRDM9 peaks overlapping H3K4me3 peaks from untransfected cells were removed, leaving 37,188 peaks passing all filters. Peaks were split into deciles according to their PRDM9 enrichment values, and the proportion of peaks with a force-called p-value <0.05 is plotted within each decile. **b**: Corrected overlap fractions with our PRDM9 peaks ($p<10^{-6}$; minimum separation, 250 bp) were computed as in **Figure 2.2a** for DSB hotspots in the "Intersect" set, split into deciles based on their reported heat in the AB_1 individual heterozygous for the A and B alleles [17]. In both plots, points are positioned at the median value of each decile and error bars represent two standard errors of the proportion in each direction.

be noted that the testis H3K4me3 enrichment values only outperform our PRDM9 enrichment estimates *after* we have identified PRDM9 peaks. Taken alone, the testis H3K4me3 data are a poor predictor of testis DMC1 heat, due to low signal in the dataset and a large number of peaks not overlapping DMC1 hotspots. Correlation values between our cell-line data and the testis data consistently improve after filtering out telomeric sites, consistent with the fact that the telomere effect operates downstream of PRDM9 binding and trimethylation activity. Finally, we showed that LD-based recombination rates [107] peak around our PRDM9 binding peak centres, and the local recombination rate increases with PRDM9 binding strength and testis

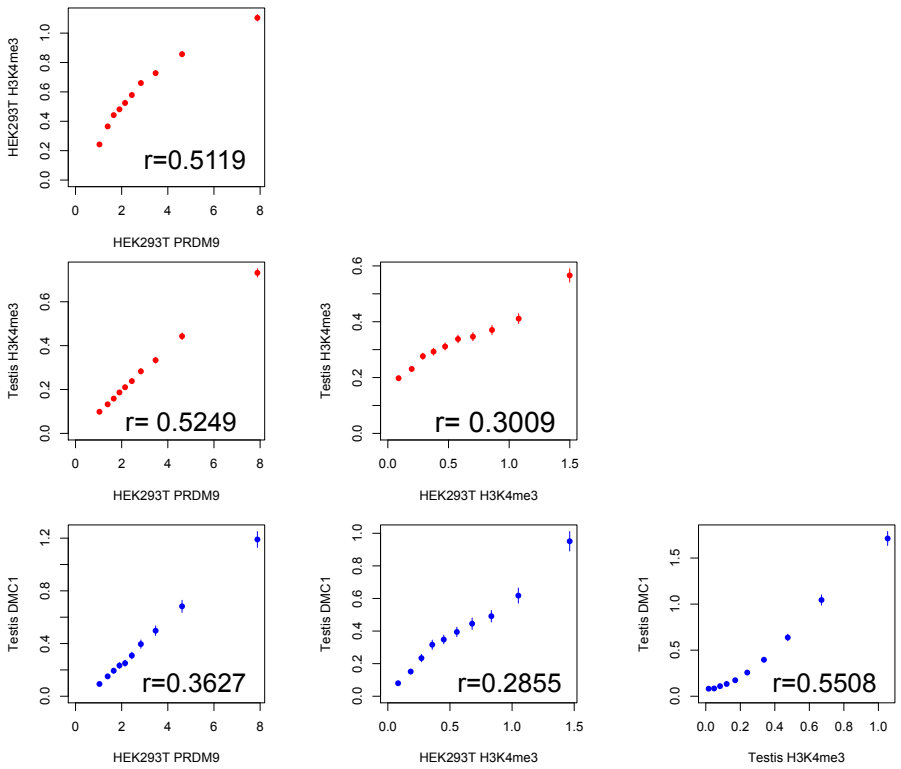


Figure 2.4: Comparison of enrichment values across samples. For the same set of peaks passing all filters in **Figure 2.3**, and further filtered to remove peaks within 15 Mb of a telomere, I plot the mean force-called enrichment value of each sample (y axis) in each enrichment decile bin of each other sample (x axis), with error bars representing ± 2 standard errors of the mean in each bin. Raw correlation values are printed on each plot. All comparisons show a significant positive correlation ($p < 2 \times 10^{-16}$). Testis DMC1/H3K4me3 data are from [17].

H3K4me3/DMC1 enrichment (see **Figure 2.5**). Thus, despite cell-type differences between our HEK293T overexpression system and the chromatin environment of early spermatocytes, our binding peaks capture the majority of biologically relevant recombination hotspots and help to refine their true PRDM9 binding sites.

2.2.2 Binding motifs reveal multiple modes of PRDM9 binding

Next, we identified sequence motifs occurring near PRDM9 peak centres using a novel, Bayesian, *de novo* motif finding algorithm (designed and implemented by Simon Myers, as described in [108]). The algorithm takes as input a set of m 300-bp

sequences extracted from the centres of strong PRDM9 peaks with narrow 99% confidence intervals (<150 bp), after removing repeats. It begins with a heuristic seeding step by first counting the frequency of each unique 10-bp DNA sequence ("10-mer") found in the central 100 bp of each peak, across all peaks. The 10-mer with the greatest enrichment in the central 100-bp region relative to the flanking 100-bp regions becomes the first seed. This seeding 10-mer is then written as a Positional Weight Matrix (PWM) with a probability of 1 for each base in its sequence and passed as input into an iterative motif refinement algorithm. In the first iteration, this algorithm estimates posterior probabilities for matches to the given PWM in each peak sequence relative to a second-order background model initialised on all m peak sequences, with a uniform prior distribution describing the positions of motif matches relative to the peak centre. With each subsequent iteration (up to a user-specified number) it updates the PWM, the background model, and the prior distribution according to the sampled matches. After the final iteration, it removes any peak sequences with a posterior probability >0.75 of containing a motif match, and the remaining peak sequences are iterated through the entire algorithm again, beginning with the seeding step. Iterations stop when fewer than 10 peak sequences match the current seeding motif. Finally, the k resulting PWMs are input jointly into the iterative refinement algorithm, and the best motif match with a posterior probability >0.75 for each peak sequence is reported. This approach assumes that each input peak sequence has exactly one match to one motif as the basis for motif updating. One strong advantage is that it effectively allows detection of multiple motifs that differ by internal insertions and deletions, a feature not present in existing *de novo* motif finding algorithms such as MEME [109].

After applying this approach to our top 5,000 human PRDM9 ChIP-seq peaks ranked by enrichment, we identified seven non-degenerate motifs that are highly enriched in the central 100 bp of each peak (see **Figure 2.6**; detailed in Methods). Together, these seven motifs explain 67% of the top 5,000 binding peaks. Each motif has a close internal match to the canonical PRDM9 13-mer reported previously [33], but they differ in two important respects. Firstly, the motifs we identify are

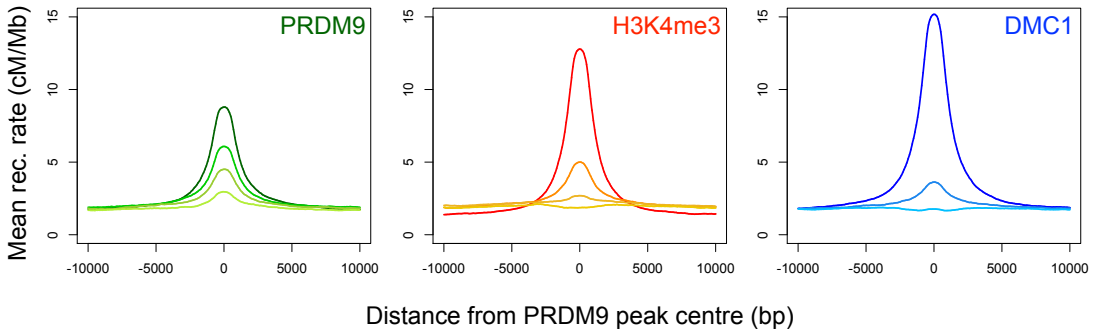


Figure 2.5: Mean recombination rates surrounding subsets of binding peaks. The same peaks from **Figure 2.4** were stratified into quartiles based on increasing PRDM9 enrichment (light green to dark green), force-called testis H3K4me3 enrichment (light orange to red), or force-called DMC1 enrichment (light blue to dark blue). Mean recombination rates (from the HapMap LD-based recombination map [107]) at each base in the 20-kb region centred on each peak centre are plotted for these subsets, with smoothing (ksmooth, bandwidth 25). The lowest two DMC1 quartiles have zero DMC1 enrichment and are merged into one line. Testis DMC1/H3K4me3 data are from [17].

much longer, consistent with the ∼36 bp expected binding footprint of PRDM9's 12 canonical zinc fingers. This suggests that the zinc fingers predicted to bind upstream of the canonical 13-mer are indeed important for binding and show high sequence specificity; Motif 1 resembles an extended motif reported in [33]. By aligning these motifs to each other and to the published *in-silico* motif prediction [34], it becomes apparent that some of these motifs differ with regard to internal spacings within the motif (see **Figure 2.6**). A middle region corresponding to ZF5 and ZF6 is predicted to span 6 bp, but in our motifs these zinc fingers appear to span only 1, 4, or 5 bp without strong DNA binding specificity at these positions. This alternative spacing may explain why the upstream zinc fingers have shown weak to no sequence specificity in previously published hotspot motifs [17, 26, 34]—in a model that disallows indels, the different motifs would be unalignable in the upstream region and their base specificities would be diluted.

Alternative spacing within motifs could explain how long zinc finger arrays like PRDM9's are able to bind DNA despite physical constraints on the number of zinc fingers that can bind DNA consecutively [41]. Our data are consistent with PRDM9 binding discontinuous submotifs with different subsets of zinc fingers

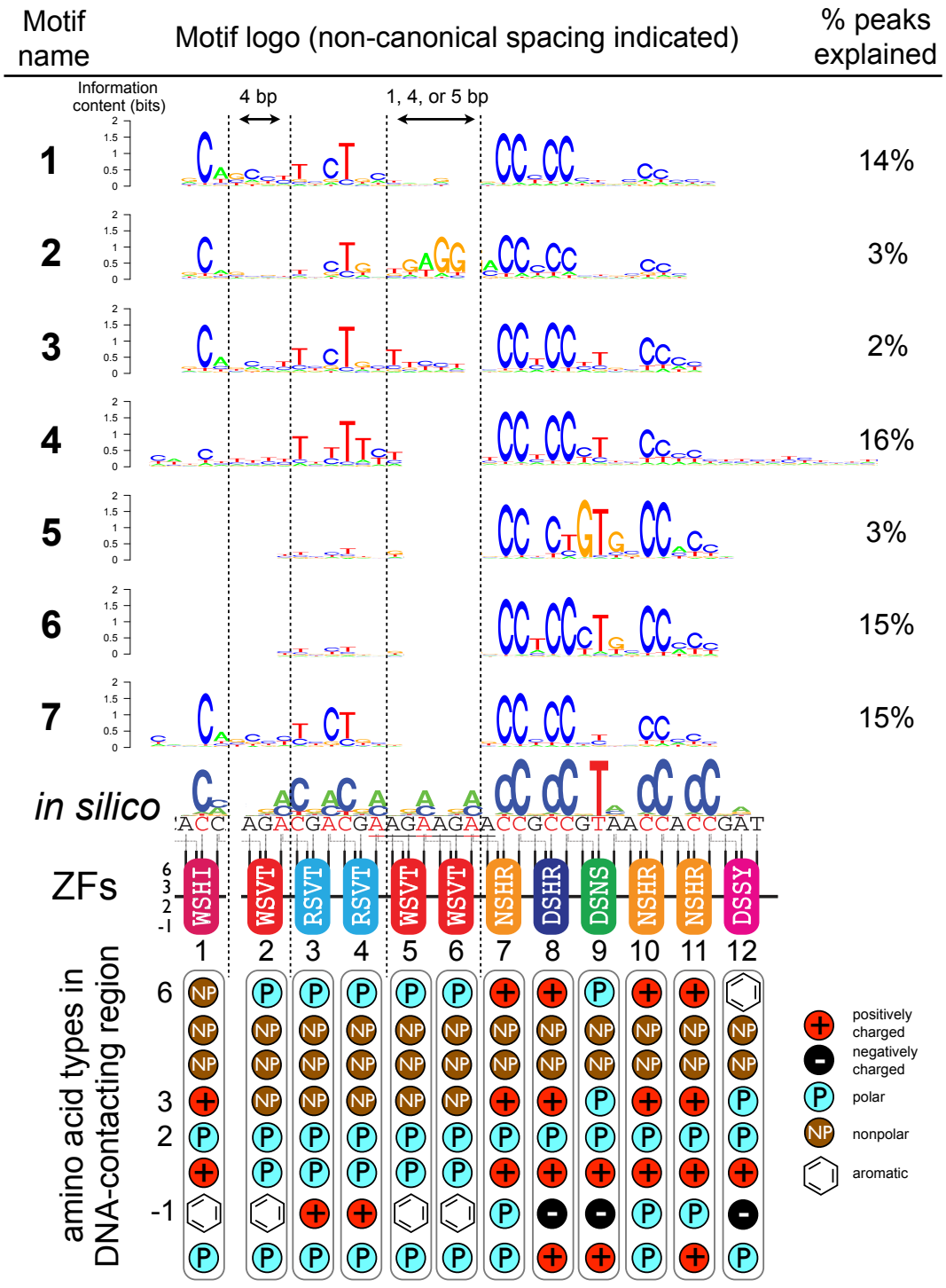


Figure 2.6: Comparison of seven distinct motifs bound by human PRDM9. The seven motif logos produced by our motif-finding algorithm (applied to the top 5,000 PRDM9 binding peaks ranked by enrichment) were aligned manually to each other and to the published *in silico* binding prediction [34], to maximise alignment of the most information-rich bases. Dotted vertical lines show regions with variable or non-canonical spacing. Below: zinc finger residues are illustrated below each zinc finger position, classified by polarity, charge, and presence of aromatic side chains, to show the prevalence of positively charged amino acids at DNA-contacting positions (labelled -1, 2, 3, 6) in the most sequence-specific zinc fingers. This contrasts with the zinc fingers aligning with the variably spaced regions, all of the "WSVT" type, which lack charged residues and have a large aromatic tryptophan residue at the -1 position.

simultaneously, with different allowable spacings between submotifs. Intriguingly, ZF5 and ZF6, which overlap the alternative spacing region, have large aromatic tryptophan residues at the -1 position, which contacts the DNA and is critical for zinc-finger binding specificity (see **Figure 2.6**). These ZFs also lack the positively charged DNA-contacting residues found in the most sequence-specific zinc fingers in the array (perhaps consistent with an electrostatic attraction to the negatively charged DNA). We speculate that these bulky, uncharged middle zinc fingers fail to bind DNA strongly and act more like a linker between the more strongly binding zinc fingers found upstream and downstream. The second zinc finger is of the exact same type as ZFs 5 and 6, and it too appears to occupy a non-canonical 4-bp footprint with low binding specificity in all of our motifs.

2.2.3 PRDM9 binds promoters, though weakly

Although these motifs explain a majority of our binding peaks, the presence of a strong PRDM9 motif match is known to be insufficient to predict hotspot formation [17, 34]. To examine how the primary DNA sequence affects the probability of PRDM9 binding in our cells, I scanned the genome for matches to each of our motifs using FIMO [109] and showed that although the probability of overlapping a PRDM9 binding peak tends to increase with an increasing motif match score, even the strongest 0.1% of motif matches have only a 50% chance of overlapping a peak (see **Figure 2.7**). Thus, binding cannot be reliably predicted by a motif PWM, suggesting that binding can be influenced by the context of each motif match in the chromatin within the nucleus.

One intriguing context for PRDM9 binding is at active gene promoters, which tend to have a more accessible chromatin conformation and tend to be marked with H3K4me3 in all cell types [48]. A study in mice has shown that in the absence of PRDM9, DSBs localise to active promoters marked with H3K4me3, suggesting that PRDM9 may serve to provide alternative H3K4me3 sites to compete with and direct recombination away from promoters [37]. However, efforts to directly detect PRDM9 binding *in vivo* in mice yielded poor enrichment [67], and no such

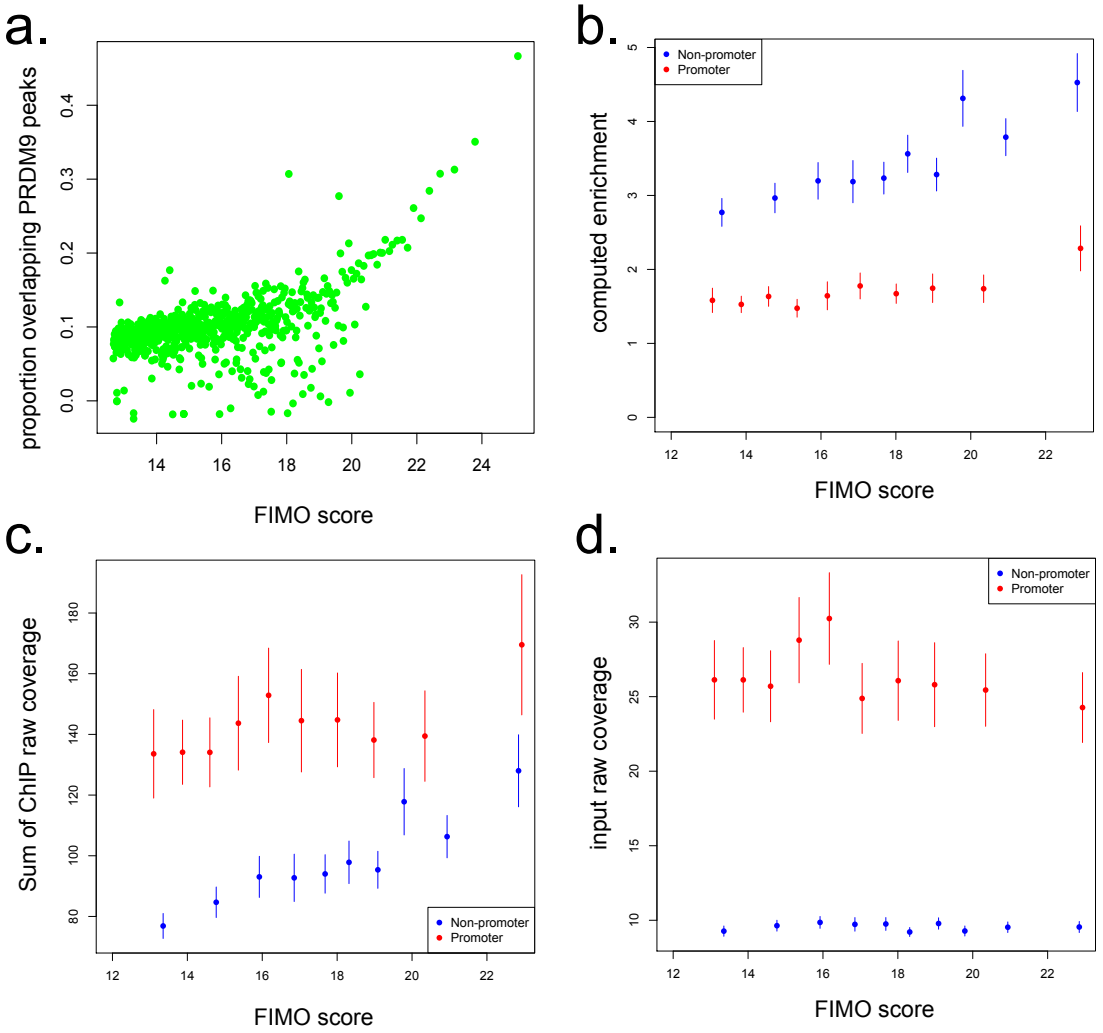


Figure 2.7: Effects of motif score on binding probability and enrichment. All plots correspond to Motif 1. **a**: Given our motif PWMs, I reported the top 1 million matches for each motif in hg19 using FIMO [109]. For 0.1 percentile bins of increasing FIMO score I reported the proportion of motif matches occurring within 150 bp of a PRDM9 peak centre ($p<10^{-6}$, minsep 250). **b**: PRDM9 peaks overlapping Motif 1 (and having more than 5 input reads overlapping the peak centre) were divided into those overlapping promoters (stringently, those within 1 kb of a TSS, overlapping an H3K4me3 peak in untransfected cells, and overlapping a DNAse HS site; red), and non-promoters (failing those criteria and further not overlapping an H3K4me3 peak reported by any ENCODE data [110]—see Methods; blue). Mean enrichment values are plotted in decile bins of FIMO score, with error bars representing ± 2 s.e.m. **c,d**: Same as **b**, but with mean sum of raw ChIP fragment coverage values in each bin (**c**) or mean raw input coverage values in each bin (**d**).

results have been published in humans. The only *in vivo* proxy for PRDM9 binding upstream of DSB formation in human is in the form of H3K4me3 ChIP-seq data [17]. However, because H3K4me3 is already present at active promoters, we cannot reliably determine whether PRDM9 binds these promoters *in vivo*. Thus, our experimental approach allows us to determine for the first time whether human PRDM9 can bind to promoter regions.

Strikingly, of the 10,899 protein coding genes with H3K4me3 surrounding their Transcription Start Site (TSS) in our untransfected cells ($p<0.001$), 83% have a PRDM9 binding peak within 500 bp of the TSS, compared to only 5% expected by chance overlap (yielding 82% corrected overlap—see Methods). However, a more careful analysis showed that this signal of promoter binding is most likely explained by increased power to detect binding at promoters due to their overrepresentation among ChIP-seq reads. Indeed, as shown in (see **Figure 2.7**), peaks that overlap promoters tend to have lower enrichment estimates, even across a range of FIMO scores for Motif 1. By plotting raw ChIP and Input coverage values for these peaks, it becomes clear that promoters have over 2.5 times the input coverage of non-promoters, meaning simply by virtue of being in a promoter one can expect 2.5 times as many reads compared to non-promoters (from both signal and background). After correcting for this fact, the promoter peaks appear weaker than their non-promoter counterparts (see **Figure 2.7**). By virtue of having more coverage, weaker promoter peaks are more likely to be called as significant. Thus, it appears that PRDM9 can bind to promoters, but *for a given motif score* it appears to do so more weakly than in non-promoter regions, perhaps because it must compete for binding in these regions with other proteins such as transcription factors, polymerases, and highly phased nucleosomes.

2.2.4 Recombination outcomes depend on motif types and genomic context

I next explored whether PRDM9 binding peaks containing different motifs might associate with different recombination outcomes. I further grouped peaks over-

lapping each motif by whether they overlap a promoter, and I filtered out any peaks overlapping repeats for this analysis. Interestingly, I observed a lower mean recombination rate [107] around Motif 7 peaks relative to all others (see **Figure 2.8**). Peaks overlapping different motifs have nearly identical PRDM9 enrichment spectra, so their differences in recombination rate are unlikely to be explained by differences in PRDM9 binding. To confirm this, I created matched sets of Motif 4 and Motif 7 binding sites, matched for PRDM9 binding strength and context (non-repeat, non-promoter, non-DNAse-HS). As shown in **Figure 2.8b**, Motif 4 has a higher mean recombination rate across all quartiles of PRDM9 enrichment. Thus, the rate decrease at Motif 7 likely results from an effect that is independent of PRDM9 binding or repeat/promoter context. Alternatively, it may represent a B-allele-specific motif not well represented in LD-based recombination maps, which are dominated by historical recombination events from the predominant A allele of PRDM9.

Across all motifs, peaks overlapping promoters show little to no increase in recombination rate above the background rate of 1.1 cM/Mb [24] (see **Figure 2.8**). However, this effect could potentially be explained by the weaker PRDM9 enrichment that we observe at promoter peaks. To investigate this, I plotted mean recombination rate across levels of PRDM9 enrichment for promoter and non-promoter peaks matching Motif 1, showing that even at overlapping enrichment values (strongly bound promoters versus weakly bound non-promoters), promoter peaks have lower recombination rates (see **Figure 2.8c**). To further rule out any possible confounders, I created matched sets of promoter and non-promoter Motif 1 binding sites with PRDM9 binding enrichment values between 1 and 2. I then plotted the mean recombination rate in the surrounding 20 kb centred on each bound motif site, showing a peak only for non-promoter peaks (see **Figure 2.8d**). Thus, independent of PRDM9 binding strength, binding sites at promoters appear far less likely to become recombination hotspots.

Certain classes of repeats such as THE1 elements have been reported to be enriched in recombination hotspots [33]. In light of this, I examined the joint

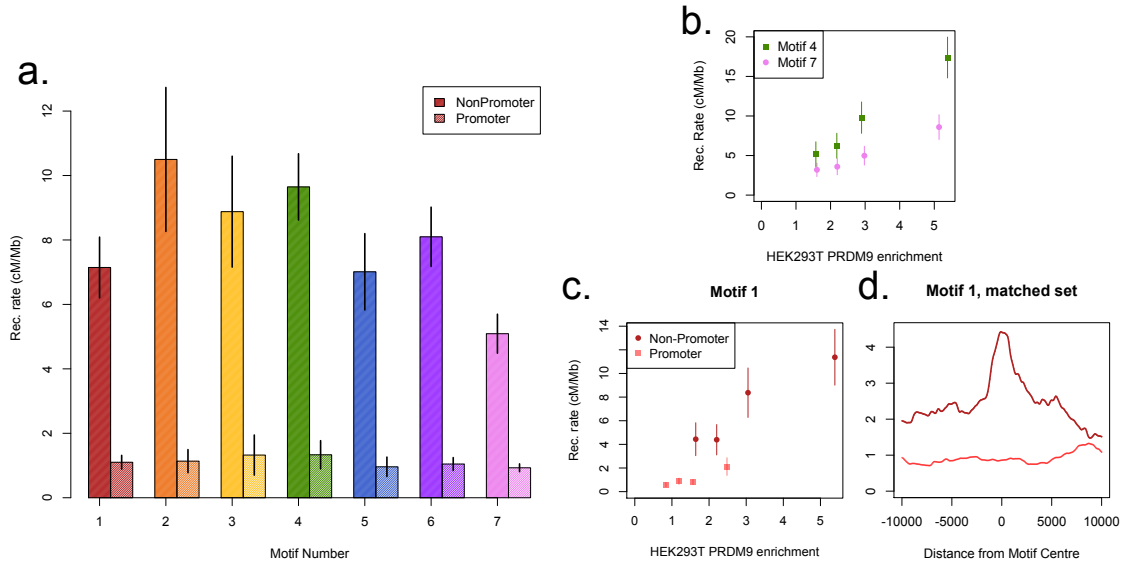


Figure 2.8: Effects of motif type and promoter context on recombination rate. **a**: PRDM9 peaks with motif matches were filtered and divided into promoter and non-promoter sets by the same criteria as **Figure 2.7**, with further requirements that they do not overlap annotated repeats and that they lie more than 15 Mb from a telomere, to avoid confounding results from repeat context and the telomere effect. The mean HapMap recombination rate [107] is reported for the central base of the assigned motif match within each peak (solid bars: non-promoter, shaded bars: promoter, error bars: ±2 s.e.m.). **b**: Mean recombination rates are reported for non-promoter peaks matching Motif 4 (green squares) and Motif 7 (pink circles) split into quartiles of PRDM9 enrichment, showing nearly identical median PRDM9 enrichment values within each quartile but consistently higher mean recombination rates for Motif 4 peaks (with non-overlapping 95% confidence intervals, shown by error bars, except in the first quartile). **c**: Mean recombination rates are reported for promoter (pink squares) and non-promoter (red circles) Motif 1 peaks split into quartiles of PRDM9 enrichment. Both median enrichment values and recombination rates are greater for non-promoter peaks, even in overlapping ranges of enrichment. **d**: Mean recombination rate profile plot surrounding Motif 1 peaks (centred on motif centres) with PRDM9 enrichment values between 1 and 2, split into promoter (228 peaks) and non-promoter (686 peaks) peaks, smoothed (ksmooth, bandwidth=300).

distribution of motif types and repeat types (classified as LINE-1, LINE-2, Alu, THE1, or Other) across PRDM9 binding peaks, showing a strong enrichment of THE1 elements in peaks overlapping Motifs 1, 6, and 7 relative to the genome-wide frequency, as well as a depletion in Motif 5 peaks (see **Figure 2.9**). Alu elements also appear to be highly enriched in Motif 1 peaks, and L1 elements appear depleted at all peaks. Speculatively, were PRDM9 to bind L1 elements, which are among the youngest transposable elements in the genome and thus have higher

sequence similarity, the risk of Non-Allelic Homologous Recombination between elements might increase [111]. Furthermore, if PRDM9 can enhance transcription near its binding sites, it might increase the propensity for an active L1 element to replicate and transpose [111].

Next, for the motif classes with large numbers of peaks in each repeat class (Motifs 1, 6, and 7), I examined the mean recombination rate in each subset of motif type and repeat class (see **Figure 2.9**). The ranking of recombination rates across repeat types within each motif class remain similar, with the highest rates occurring in L2 elements and the lowest rates in Alu elements. Interestingly, Motif 7 continues to show lower mean recombination rates across nearly all repeat classes.

Finally, to examine the effect of local chromatin marks on recombination outcomes, I annotated our binding peaks with whether they overlap ChIP-seq peaks reported for 9 histone variants or modifications reported by the ENCODE project: H3K9me1, H3K9me3, H3K9ac, H2az, H3K27ac, H3K27me3, H3K36me3, H3K79me2, and H4K20me1 [110]. Because these data were collected in a different human cell line (K562), we use them only as a proxy for true chromatin states in HEK293T cells and in spermatocytes, relying on the fact that in comparisons across cell types, many or most chromatin mark locations are similar [110]. Most of these chromatin marks are associated with active enhancers, promoters, and gene bodies, with the exception of H3K9me3, which marks constitutive heterochromatin, and H3K27me3, which marks facultative heterochromatin [110]. Interestingly, mean recombination rate decreases significantly across all chromatin marks tested (95% C.I. ranges -6% to -63%), with the exception of H3K27me3, whose peaks shows a 28% increase above the mean rate for all peaks (95% C.I. 17-40%; see **Figure 2.10**). That is, conditional on binding strength, PRDM9 binding sites overlapping facultative heterochromatin regions, which are typically transcriptionally repressed, appear to be more likely to become recombination hotspots.

Thus, the processing of a motif into a binding site, a binding site into a DSB, and a DSB into a crossover appears to be heavily influenced by genomic context. With our data we have observed effects of promoter context on PRDM9 binding (see

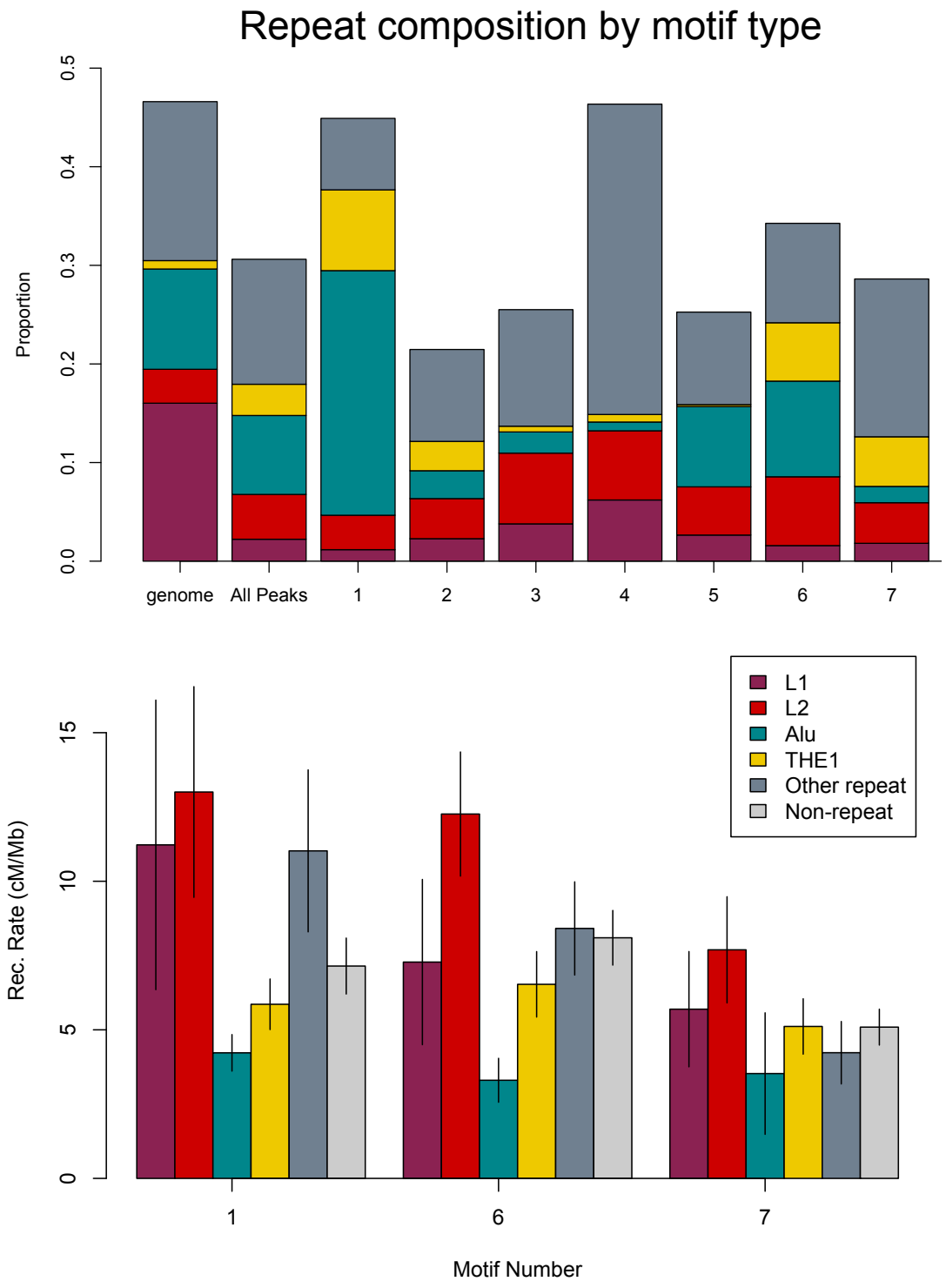


Figure 2.9: Effects of motif type and repeat context on recombination rate. PRDM9 peaks were filtered and non-promoter, non-telomere peaks were selected by the same criteria as **Figure 2.8**, then annotated with overlaps between the peak centre and annotated RepeatMasker repeats, by repeat class (mutually exclusive categories). Top: the frequency of each repeat class in peaks matching each motif type, compared to all peaks and compared to the genome-wide frequency of each repeat class. Bottom: The mean recombination rate for peaks falling into each repeat class and motif type, for Motifs 1, 6, and 7 (error bars = ±2 s.e.m.).

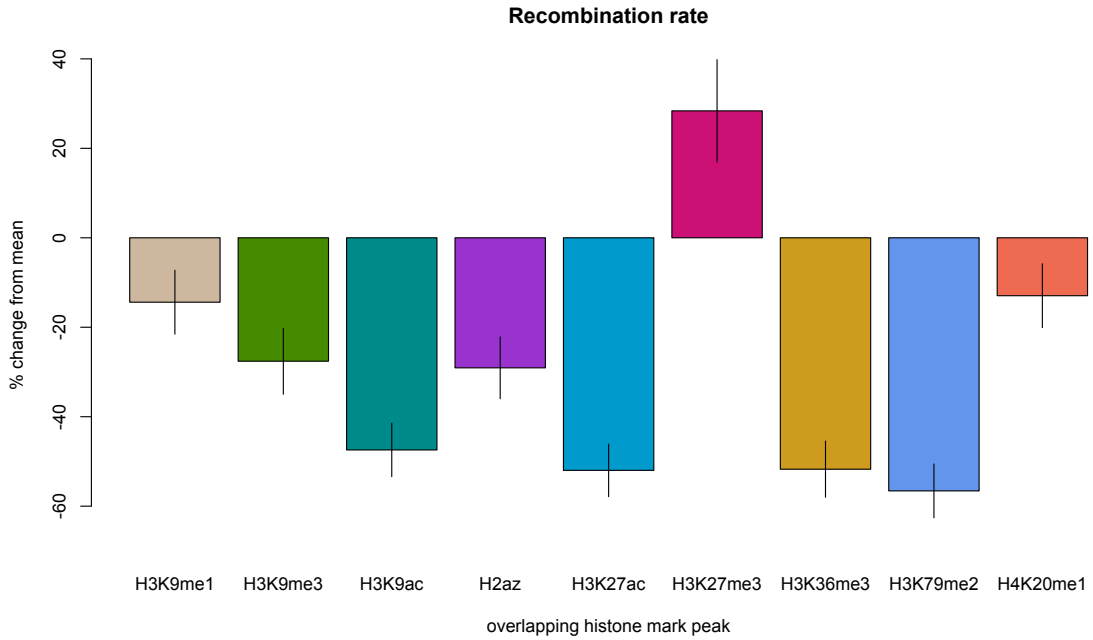


Figure 2.10: Associations of chromatin annotations and recombination rate. PRDM9 peaks were filtered as in **Figure 2.8**, while also removing peaks matching Motif 7 and requiring peaks to have PRDM9 enrichment values in the range [1,2], then annotated with whether they overlap each of 9 reported histone variant peak sets reported for K562 cells [110]. The marginal mean recombination rate is reported for peaks overlapping each histone variant type (categories are not mutually exclusive; error bars = ±2 s.e.m.; scale= % change relative to mean rate for all peaks: 2.62 cM/Mb).

Figure 2.7); the effects of promoter context and telomere proximity in predicting DSB outcomes (see **Figure 2.4**); and the effects of Motif type, repeat type, promoter context, and chromatin context on recombination rates.

2.2.5 PRDM9 also deposits H3K36me3 in *cis*

Next, I explored changes induced by PRDM9 in *cis* to the local chromatin landscape. Among PRDM9 peaks not overlapping pre-existing H3K4me3 peaks, I observed a 4-fold increase in mean H3K4me3 enrichment extending 1 kb to either side of each binding site in transfected cells, consistent with PRDM9 binding and trimethylating H3K4 on up to 3 nucleosomes on either side, as has been shown *in vivo* [95] (see **Figure 2.11**). It has also been observed that PRDM9 can trimethylate the H3K36 residue *in vitro* with similar efficiency to H3K4, and cell-wide H3K36me3 immunofluorescence was observed to increase in cells overexpressing PRDM9 [100].

Other work by our group [108] recently showed a weak enrichment of the H3K36me3 signal surrounding DSB hotspots *in vivo* in mouse testes, but much weaker than the observed enrichment of H3K4me3. We sought to explore this signal locally in our overexpression system by performing low-coverage H3K36me3 ChIP-seq in transfected and untransfected cells. At PRDM9 binding peaks not overlapping pre-existing H3K36me3 peaks, we see a significant local increase of H3K36me3 enrichment surrounding the peak centre (∼50% over flanking enrichment values), but much weaker in magnitude than the H3K4me3 signal at the same sites (see **Figure 2.11**). Speculatively, these observations may be consistent with the H3K36me3 mark having a faster demethylation rate than the H3K4me3 mark. That is, perhaps PRDM9 adds both marks when it binds, but the K36 mark is removed more quickly by ubiquitous H3K36 demethylases [112].

2.2.6 PRDM9 binds preferentially to accessible DNA and phases nearby nucleosomes

Finally, to investigate the effects of PRDM9 on local chromatin accessibility and nucleosome positioning, Emmanuelle Bitoun performed low-coverage ATAC-seq (Assay for Transposase-Accessible Chromatin with high-throughput sequencing) in our HEK293T cells [113]. ATAC-seq utilises a transposase that preferentially cleaves accessible fragments of DNA (such as inter-nucleosome linker regions) and then directly attaches sequencing adapters to the ends of the resulting fragments. The fragment size distribution has peaks corresponding to mono-, di-, tri- and inter-nucleosome fragments, and by partitioning these fragments by size we can plot the positions of nucleosome-free and nucleosome-occupied regions (see **Figure 2.12**). We included samples that were either untransfected or transfected with human PRDM9 or two shorter constructs to serve as controls: the "noZF" construct (which truncates the PRDM9 cDNA before the ZF array) and the "ZFonly" construct (which includes only the ZF array; as illustrated in **Figure 2.1**). The ZFonly construct can presumably bind DNA, based on similar constructs used in published DNA-binding assays [97], but it lacks the PR/SET domain so should be incapable

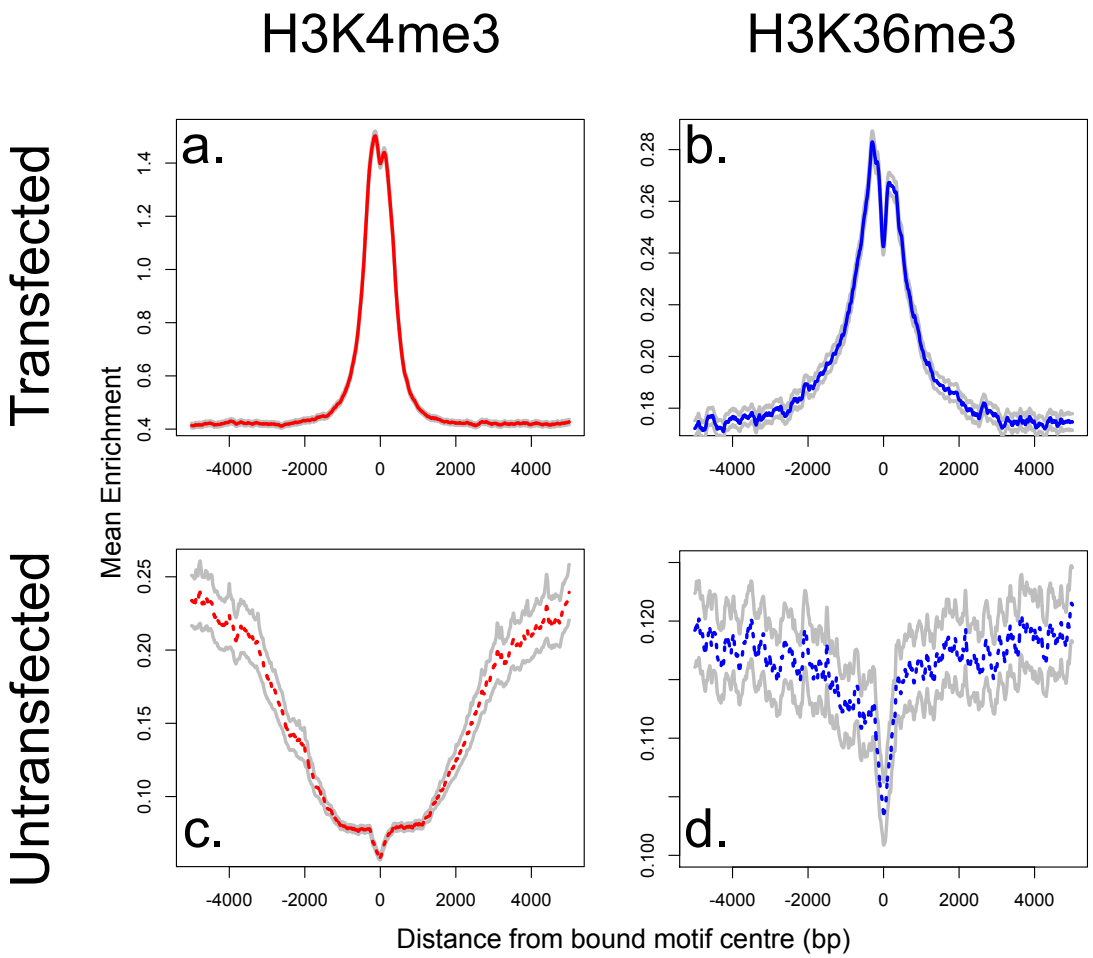


Figure 2.11: Trimethylation of both H3K4me3 and H3K36me3 in *cis*. PRDM9 peaks with motif matches were filtered to remove any overlaps with H3K4me3 or H3K36me3 peaks in untransfected cells. Plots show the the mean enrichment at each base surrounding the motif centre in transfected (**a**,**b**) and untransfected cells (**c**,**d**), with grey lines indicating ±2 s.e.m. at each position. H3K4me3 results are shown in red (**a**,**c**) and H3K36me3 in blue (**b**,**d**). The dips seen in untransfected cells owe to the filter eliminating sites overlapping untransfected peaks, but they help to illustrate the change at these sites before and after transfection with the human-HA construct. Lines are smoothed with ksmooth (bandwidth=25). Note: absolute enrichment values (y axes) cannot be compared across samples.

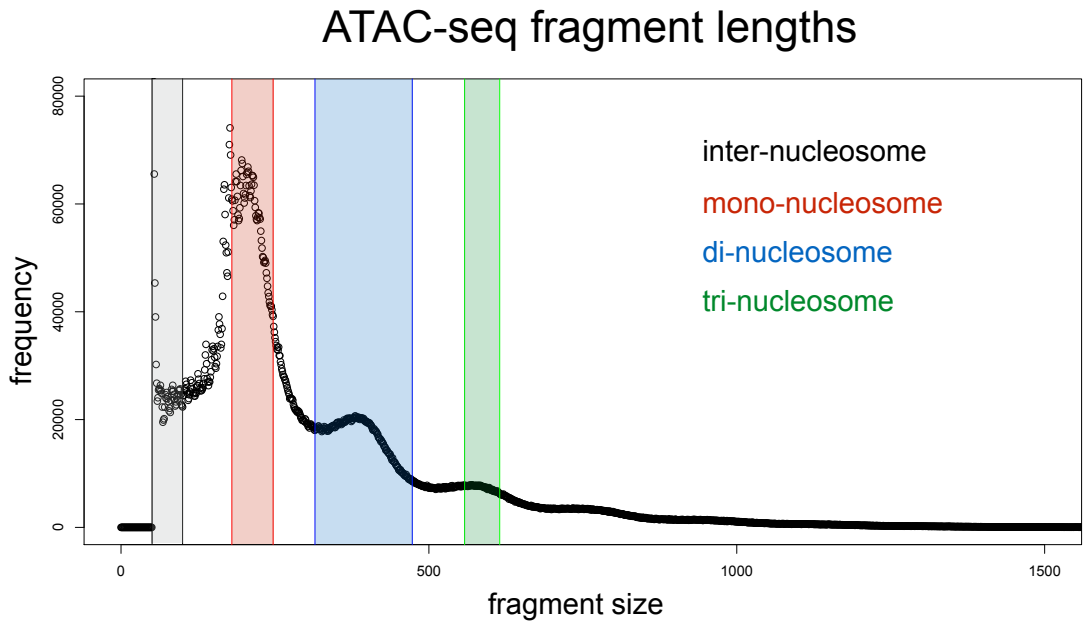


Figure 2.12: Partitioning of ATAC-seq fragments by size. The ATAC-seq fragment size distribution from untransfected cells is shown after removing duplicates and mtDNA reads, highlighting the reported intervals corresponding to nucleosome numbers in each fragment [113].

of forming the H3K4me3 mark. The noZF construct presumably cannot bind DNA and thus serves as a control for effects of transfection apart from the DNA-binding activity of PRDM9.

Unfortunately, in this pilot experiment we obtained a high fraction of reads deriving from mtDNA (58%), which further reduced our effective coverage. However, from these low-coverage pilot data I was still able to observe aggregated signals of nucleosome positioning. As a positive control, I plotted ATAC-seq signals around ∼36,000 Transcription Start Sites (TSSs) and around ∼12,000 known CTCF binding sites, both of which are known to have strong phased nucleosome signals. Our data, even with very low coverage, still have the ability to detect the expected nucleosome occupancy signals for these regions genome-wide (see **Figure 2.13**).

Next, I plotted profiles of mean ATAC-seq coverage at each base surrounding our PRDM9 binding sites in untransfected cells and in cells transfected with our three different constructs (see **Figure 2.14**). I selected only peaks containing a motif match (to better centre the profile plots), and I filtered out peaks overlapping

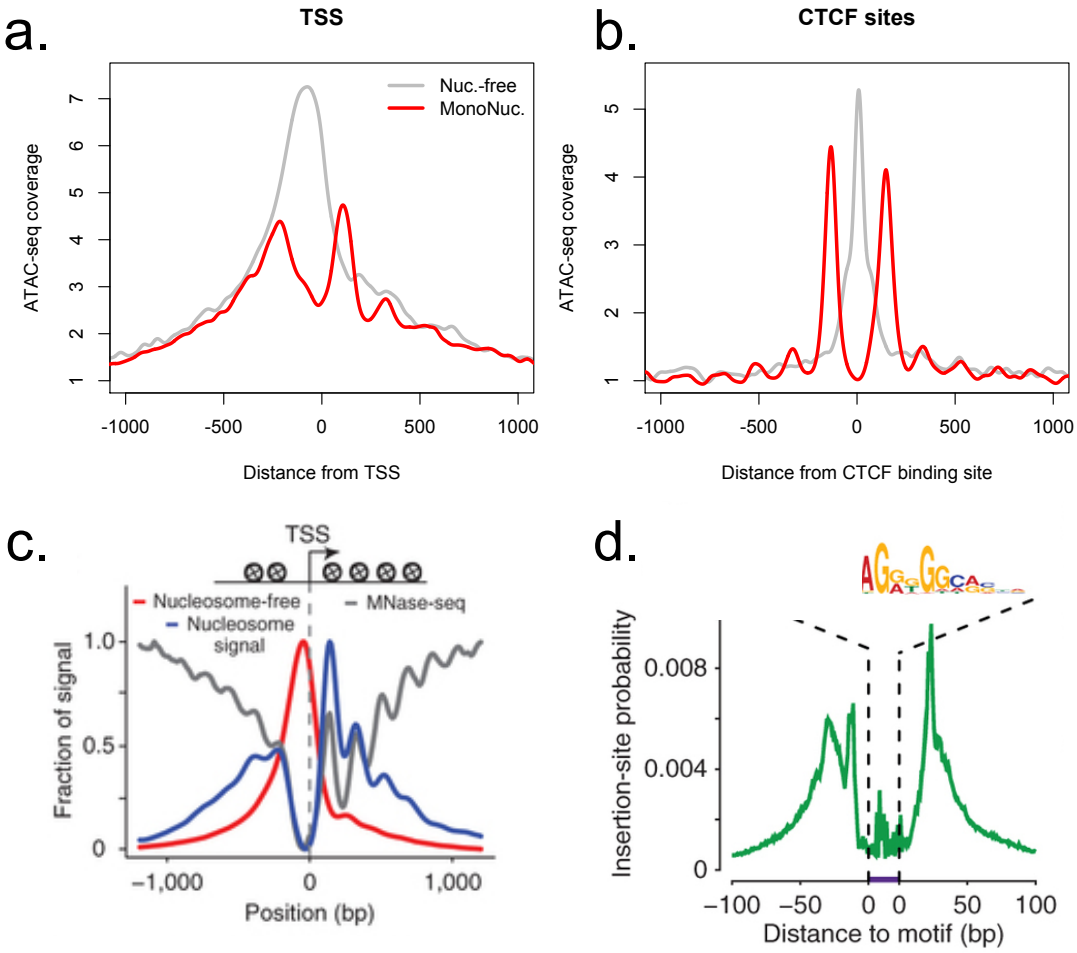


Figure 2.13: ATAC-seq coverage at Transcription Start Sites and CTCF binding sites. Positive control plots showing the mean ATAC-seq inter-nucleosome and mono-nucleosome coverage profiles around TSS and CTCF sites in our untransfected cells (**a**,**b**), compared with those from a previous study (**c** and **d**, reproduced from [113]). "Coverage" here refers to the frequency with which an ATAC fragment centre occurs at each position, such that inter-nucleosome coverage tracks the centres of nucleosome-depleted regions, and mono-nucleosome coverage tracks the centres of phased nucleosomes. Coverage values are normalised to the mean values observed between 1500 and 3000 bases away from each TSS or CTCF site, as a measure of background, and smoothed (ksmooth bandwidth = 50).

annotated DNAse-HS sites to avoid most pre-existing phased nucleosome signals (*e.g.* from TSS and CTCF sites), then I selected the ∼15,000 remaining peaks with enrichment values greater than 5, to ensure they are bound in the majority of cells. I plotted separately the coverage from short fragments corresponding to inter-nucleosome regions and the coverage from longer fragments corresponding to mono-nucleosomes (di- and tri-nucleosome fragments were excluded due to insufficient coverage). In both the untransfected sample and the noZF and ZFonly samples, I observed a ∼2-fold increase in inter-nucleosome fragment coverage at the PRDM9 binding site, consistent with PRDM9 preferentially binding motifs found in inter-nucleosome regions, but I observed no strong nucleosome positioning in the surrounding mono-nucleosome fragment coverage profile (see **Figure 2.14**). By contrast, the human-transfected sample shows a 2.4-fold increase in inter-nucleosome coverage at the binding site and a 1.8-fold increase in mono-nucleosome coverage peaking at 100 bp to either side. This strong phasing signal extends to the surrounding 2-3 nucleosomes in either direction, consistent with *in vivo* observations [95]. The magnitude of these signals is smaller than for CTCF, but this may result from the fact that not every cell will have been transfected successfully with PRDM9, whereas all cells contain endogenous CTCF. Interestingly, our negative ZFonly results indicate that the upstream region of PRDM9 is necessary for phasing the surrounding nucleosomes and for increasing the accessibility of the binding site. This may result either from the bulk of the upstream region or to the trimethylation activity of the PR/SET domain, but not to the binding of the ZF array alone. However, we cannot yet rule out that the ZF construct simply fails to localise to the nucleus, which will have to be confirmed with additional immunofluorescence experiments.

2.2.7 Chimp PRDM9 and human PRDM9 preferentially bind different genomic regions

In order to better understand the epigenetic predictors of binding, I next sought to explore the properties of a PRDM9 allele very different from the human B allele. We chose the chimpanzee W11 allele (present in the chimpanzee reference assembly),

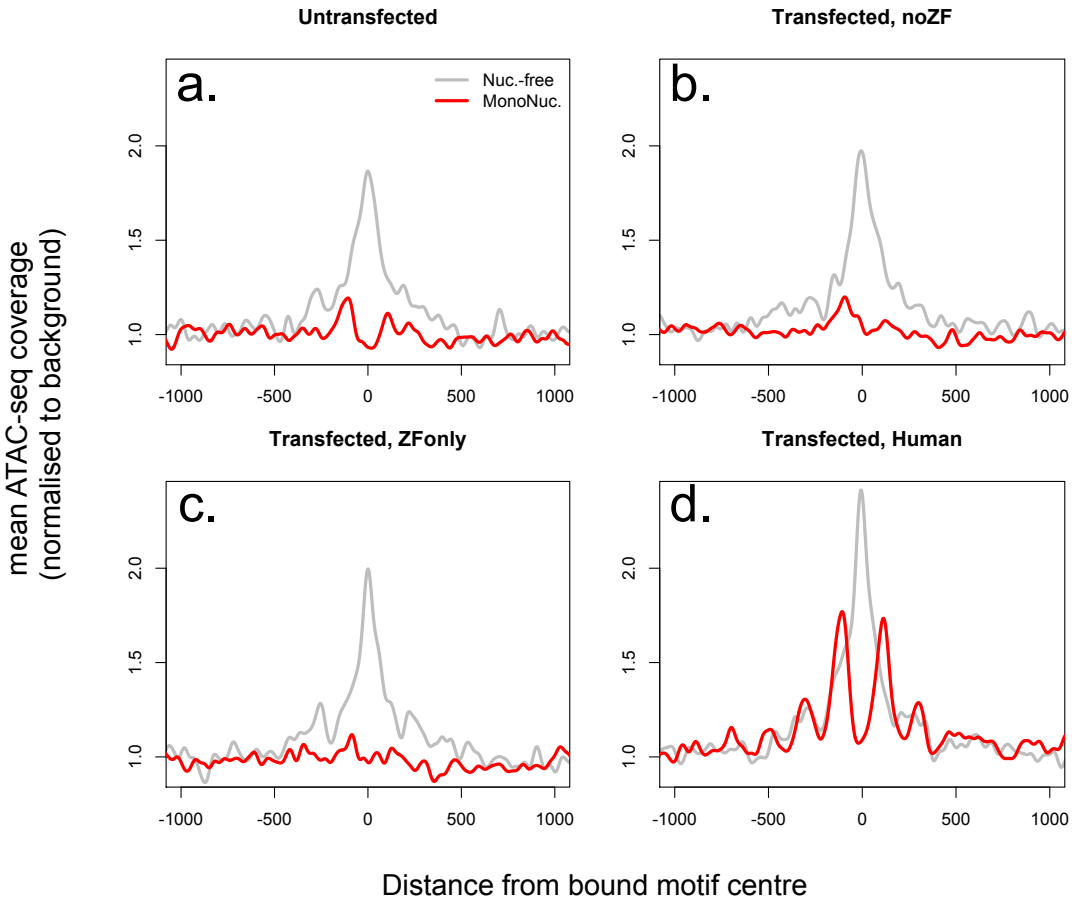


Figure 2.14: ATAC-seq coverage at PRDM9 binding sites. ATAC-seq profile plots as in **Figure 2.13**, but surrounding a set of the ∼15,000 strongest PRDM9 peaks (filtered to require a motif match and to not overlap an annotated DNAse HS site), across 4 different transfection samples. The human-transfected cells show strongly phased nucleosomes centred at ∼100 bp to either side of the motif and an elevated signature of nucleosome depletion at the centre (**d**), when compared to the three controls (**a**,**b**,**c**). The ZFonly result (**c**) suggests that the ZF domain alone is insufficient to produce this nucleosome phasing.

which was initially measured to be at roughly 25% frequency in wild chimpanzees [31], making it one of the most common alleles. Subsequent work measured the frequency of this allele type (Pan.t-4,8,12,16 in their nomenclature) to be closer to 13.4% [73]. An LD-based genetic map of chimp recombination failed to identify definitive motifs at recombination hotspots, which tend to be weaker than those found in humans [31]. This dilution of recombination hotspots could result either from the historic diversity of PRDM9 alleles in chimps or from a reduced degree of specific binding by the predominant chimp alleles relative to the predominant

human alleles. By measuring binding of the chimp allele in human cells, we sought to determine whether this chimp allele also has a definitive binding motif and/or whether it binds more promiscuously across the genome. This chimp allele differs from the human B allele in other important ways as well: the chimp allele has 18 zinc fingers, as opposed to 12, and its zinc fingers are predicted to bind an AT-rich motif as opposed to a GC-rich motif [31, 73].

To examine the binding differences between these two alleles, we synthesised chimp PRDM9 exon 10 containing the W11a zinc finger array and cloned it into our same expression vectors, replacing human exon 10 (the N-terminal, non-ZF region of this exon is nearly identical between human and chimp) (see **Figure 2.1**). I confirmed expression by YFP fluorescence and by western blot, and in both cases I detected a qualitatively fainter signal than for the human allele, perhaps due to a lower transfection or expression efficiency, as might be expected due to the larger size of the chimp allele. I processed the chimp-transfected cells in the same manner as the human-transfected cells for ChIP-seq and achieved a similar proportion of reads from signal versus background.

De novo peak calling at the same thresholds ($p<10^{-6}$); minimum peak separation, 250 bp) yielded 247,717 total peaks, a 50% increase over the number for the human allele. Only 2% of chimp peak centres occurred within 1 kb of a human peak centre, consistent with their ZF arrays having very different binding preferences. At broad scales, peaks for the human allele tend to be overrepresented in GC-rich and DNAse hypersensitive regions (including promoters), but peaks for the chimp allele do not (see **Figure 2.16**). Because we have increased power to detect binding in these regions and have shown that the magnitude of human PRDM9 binding enrichment is lower at promoters, the lack of chimp PRDM9 binding sites in these regions is consistent with chimp PRDM9 failing to bind even weakly. I also examined whether the increased number of chimp PRDM9 peaks implies a more promiscuous binding preference than for human PRDM9. If this were the case, we would expect the enrichment spectrum for chimp PRDM9 peaks to be more concentrated at lower enrichment values. However, a comparison of the enrichment distributions of the

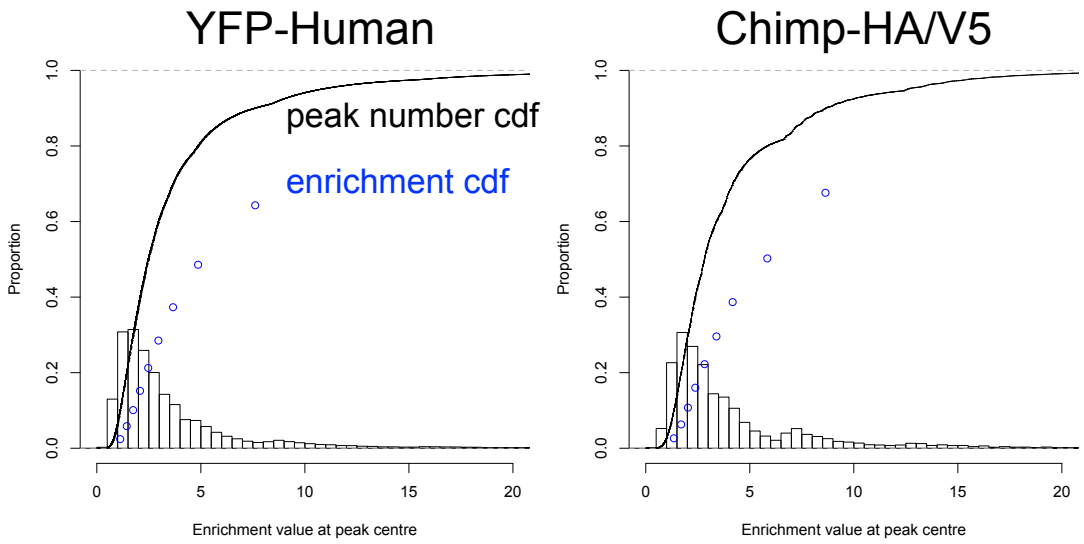


Figure 2.15: Chimp and human PRDM9 binding enrichment distributions. Histograms show the distribution of enrichment values at all Human and Chimp PRDM9 peaks ($p<10^{-6}$, minsep 250), with cdfs of the number of peaks overlaid as black curves. Blue points are placed at the maximum of each enrichment decile bin and show the cumulative proportion of total sum enrichment with each increasing decile. The final decile point resides at the maximum enrichment value (off the x axis) with a value of 1.0. These plots show that for both Human and Chimp peaks, the hottest 20% of peaks contain roughly 50% of the genome-wide enrichment.

two alleles shows little difference between them, with the hottest ∼20% of peaks containing ∼50% of the total enrichment signal in both cases (see **Figure 2.15**). The increased number of chimp peaks could owe to an ascertainment issue such as the fact that chimp binding sites are more spread out across the genome, making them less likely to overlap each other and to be merged into the same peak call.

2.2.8 A novel chimp PRDM9 binding motif

I took the strongest 5,000 chimp peaks and ran their flanking DNA sequences through the same motif-finding pipeline used for the human allele. A single, 17-bp, CT-rich motif was returned, found at 60% of peaks and highly centrally enriched within peaks (see **Figure 2.17**). We compared this motif with published *in silico* binding predictions for this allele [31, 73] and found a close match in the central region of the predicted motifs. Interestingly, our motif almost exactly overlaps a

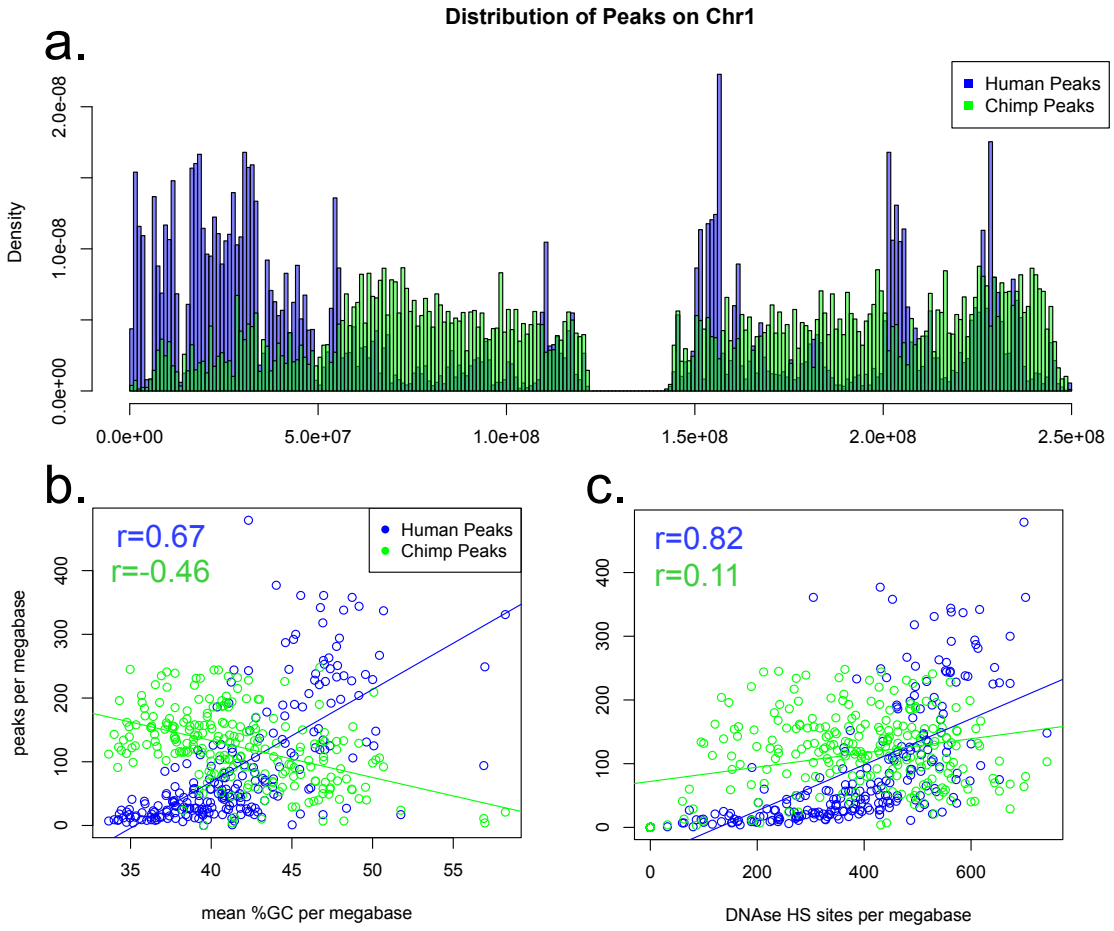


Figure 2.16: Chimp and human PRDM9 binding at broad scales. **a**: bar plot showing the number of chimp (green) and human (blue) PRDM9 peaks ($p<10^{-6}$, minsep 250) in each megabase on chromosome 1 (normalised to the total number of each on chromosome 1). The two alleles tend to peak in opposing regions of the chromosome. **b**,**c**: scatter plots showing the number of peaks against the mean GC content (**b**) or the number of DNAse-HS sites (**c**) in each megabase on Chr1, with correlation values for human peaks printed inside each plot and regression lines added ($p<10^{-6}$ in all cases).

subregion of the *in silico* binding motif that was identified as being common to many different chimp alleles [73]. It remains unclear why there is only a single binding motif and why that motif is shorter than the predicted footprint of the chimp allele, but its predicted overlap among different alleles raises the intriguing hypothesis that chimp alleles may be positively selected to bind a particular submotif, with less pressure on the binding specificity of the surrounding zinc fingers.

Using FIMO [109] I searched for strong matches to the chimp motif in the human genome and showed a weak association between motif score and probability

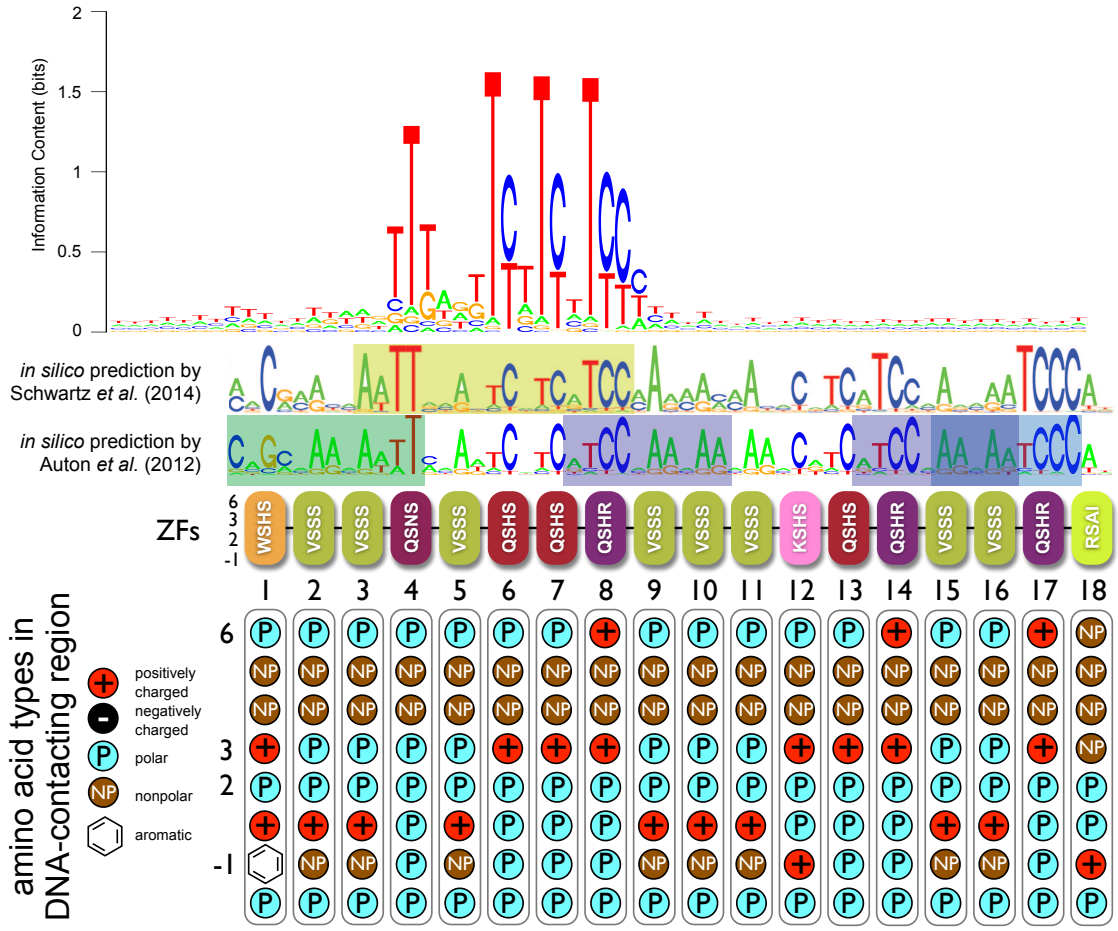


Figure 2.17: A novel chimp binding motif. The chimp binding motif produced by our motif finding algorithm applied to the top 5,000 chimp peaks, ranked by enrichment, aligned with the predicted positions of zinc fingers and with published *in silico* motif predictions [31, 73]. Although the motif is long, only a 16-bp region corresponding to ZFs 4-8 shows any strong sequence specificity, and this almost exactly overlaps a submotif shown to be shared among many different chimp PRDM9 alleles (highlighted in yellow on the Schwartz *et al.* motif, [73]). Zinc-finger residues are illustrated below, as in **Figure 2.6**, showing that chimp lacks tryptophan residues at internal zinc fingers, which are associated with regions of alternative spacing in the human binding footprint.

of overlapping a chimp binding peak in our cells (see **Figure 2.18**). Next, I plotted the chimp recombination rate [31] around the strongest ~40,000 matches as well as at the subset of 5,584 of these sites bound in our transfected cells. A small local increase in recombination rate is visible for the strong motif matches, with a much larger increase for the bound strong motif matches (a 50% increase over the background recombination rate). This confirms that our binding sites likely overlap true chimp hotspots, but the association between binding and recombination

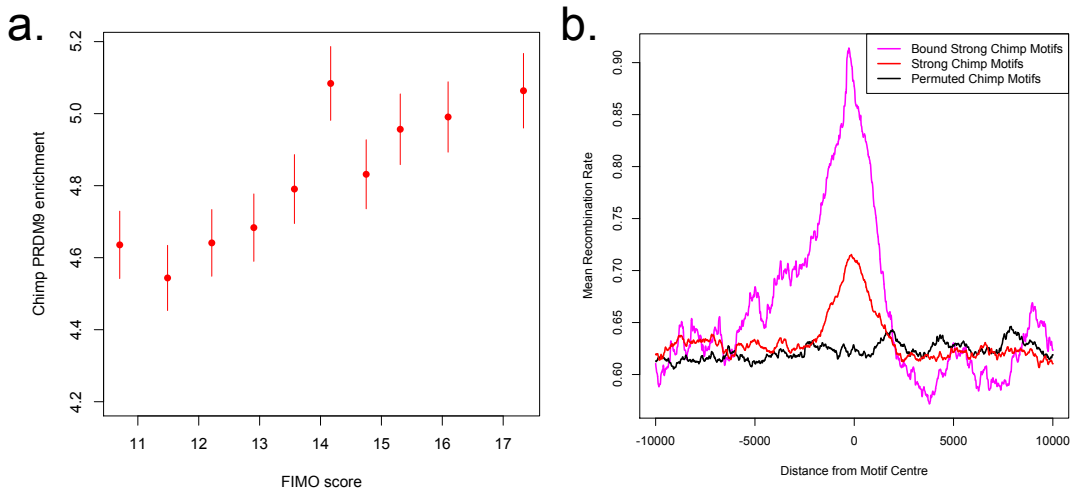


Figure 2.18: Chimp motifs predict binding enrichment and recombination rate. Given our motif PWMs, I identified the top 9 million matches for the chimp motif in hg19 using FIMO [109] then identified peaks containing the motif. **a**: the mean chimp PRDM9 enrichment value in each decile bin of assigned FIMO score (error bars ±2 s.e.m., $p<2x10^{-16}$ for positive correlation). **b**: profile plot showing the mean chimp recombination rate centred on either the strongest ∼40,000 chimp motif matches in the genome (red), the subset of those matches that are among our binding peaks (magenta), or a set of positions shifted randomly in the range [-60000,60000] from the motif match locations (black), as a measure of background (smoothing: ksmooth, bandwidth 25).

rate is much smaller than for the human allele. This may owe to the fact that chimp recombination is less concentrated in hotspots, perhaps owing to a greater historical degree of chimp PRDM9 diversity.

2.2.9 PRDM9 can activate transcription of some genes, including *CTCFL*

Given that the H3K4me3 mark is found at the promoters of actively transcribed genes, it has been hypothesised that PRDM9 may act as a transcription factor in meiosis by adding this mark to promoters [38]. Indeed, previous work has shown that PRDM9 binding can activate transcription of a transgene reporter, and PRDM9 has been proposed as a regulator of the *Morc2b* gene in mice [38]. Our finding that the human B allele binds promoters further raises the possibility that PRDM9's H3K4me3 mark may play a role apart from simply specifying the locations of meiotic DSB breakpoints. To address whether PRDM9 affects the transcription of endogenous genes when it binds to their promoters, we performed

low-coverage RNA-seq after transfecting our HEK293T cells with human PRDM9. Specifically, we sequenced mRNA by enriching for poly-A tagged transcripts. As controls, we also performed RNA-seq in untransfected cells and in cells transfected with the chimp allele or with a construct containing only the human ZF domain (thus incapable of forming the H3K4me3 mark). These latter controls allow us to identify genes activated by transfection itself, as opposed to specific PRDM9 alleles. Emmanuelle Bitoun performed these transfections and the RNA extraction.

I processed the RNA-seq data using the Tuxedo software family [114] and tested for pairwise differential gene expression among our samples using the default Cufflinks-DGE thresholds (p-value <0.05 after Benjamini-Hochberg correction). I began by searching for a very stringent subset of transcripts that were differentially expressed between the Human sample and every other sample (Chimp, ZFonly, Untransfected). Seven transcripts satisfied these criteria, with all seven being upregulated in human-transfected cells relative to all other samples.

Five of the seven human-activated transcripts overlap annotated genes: *MEG3*, *ONECUT3*, *LGALS1*, *VCX*, and *CTCFL*. The latter two genes are specific to spermatogenesis. *VCX* encodes a small highly charged protein of unknown function and has been previously characterized for its involvement in PRDM9-related non-homologous recombination events [33]. *CTCFL* is a variant of *CTCF* expressed exclusively in preleptotene spermatocytes, and male knockout mice show greatly reduced fertility due to meiotic arrest [115]. CTCFL binds a smaller and partially overlapping set of sites compared with CTCF, and its binding motif is subtly different [115]. It may be involved in organising the meiotic chromatin landscape and regulating the transcription of meiotic genes [115]. In our HEK293T cells, CTCFL RNA levels increase 28-fold after transfection with the human allele, from a nearly undetectable baseline transcription level. I plotted PRDM9 and H3K4me3 ChIP-seq enrichment data surrounding the CTCFL locus and confirmed that in transfected cells, PRDM9 binds strongly to a GC-rich repeat near the transcription start site and forms the H3K4me3 mark, which is absent in untransfected cells (see **Figure 2.20**). Elevated RNA-seq coverage can then be seen in coding regions

across the full length of CTCFL. Of course, this does not establish whether PRDM9 is necessary or sufficient for CTCFL expression *in vivo*, but it is clearly able to trigger the transcription of CTCFL in our cells. Because we only performed one replicate of RNA-seq, Emmanuelle Bitoun validated the CTCFL and VCX results using qPCR and confirmed their elevated expression in human-transfected cells (see **Figure 2.19**).

Recent work has shown that *Prdm9* expression begins in pre-leptotene cells in mice [99], which is concurrent with *Ctcfl* expression [115] and thus supports the possibility that PRDM9 may promote *CTCFL* transcription *in vivo*. I examined published H3K4me3 coverage from human testes [17] at the *CTCFL* promoter and found that the *in vivo* H3K4me3 peak occurs roughly 700 bp downstream from the H3K4me3 peak in our transfected cells, consistent with usage of an annotated alternative promoter (see **Figure 2.20**). This suggests that PRDM9 trimethylation may not be the primary activator of *CTCFL* expression *in vivo*, but may serve only to enhance its expression or to increase the frequency of a longer isoform with a different promoter. The failure of the chimp allele to bind to or activate the expression of human *CTCFL* further suggests that this behaviour may not be essential across organisms, although we cannot rule out the possibility that the chimp allele could potentially still bind the *CTCFL* promoter in chimp spermatocytes. However, it also remains unlikely that other human alleles with very different binding preferences, such as the C allele, would bind the same promoter. I also examined the *VCX* locus and found that PRDM9 does not bind near its annotated TSS, but rather in the middle of the gene and in a G-rich repeat near the terminus of the gene (see **Figure 2.21**). Testis H3K4me3 data also show strong peaks in the middle of the gene, suggesting that the promoter may lie away from the TSS, or the TSS may be annotated improperly.

Next, I relaxed the initial stringent filters to identify other protein-coding genes activated by PRDM9 binding near their annotated transcription start sites, similar to *CTCFL*. I searched for all genes with evidence of H3K4me3 within 500 bp of a TSS in the human-transfected sample ($p<0.05$, force-calling, requiring >5

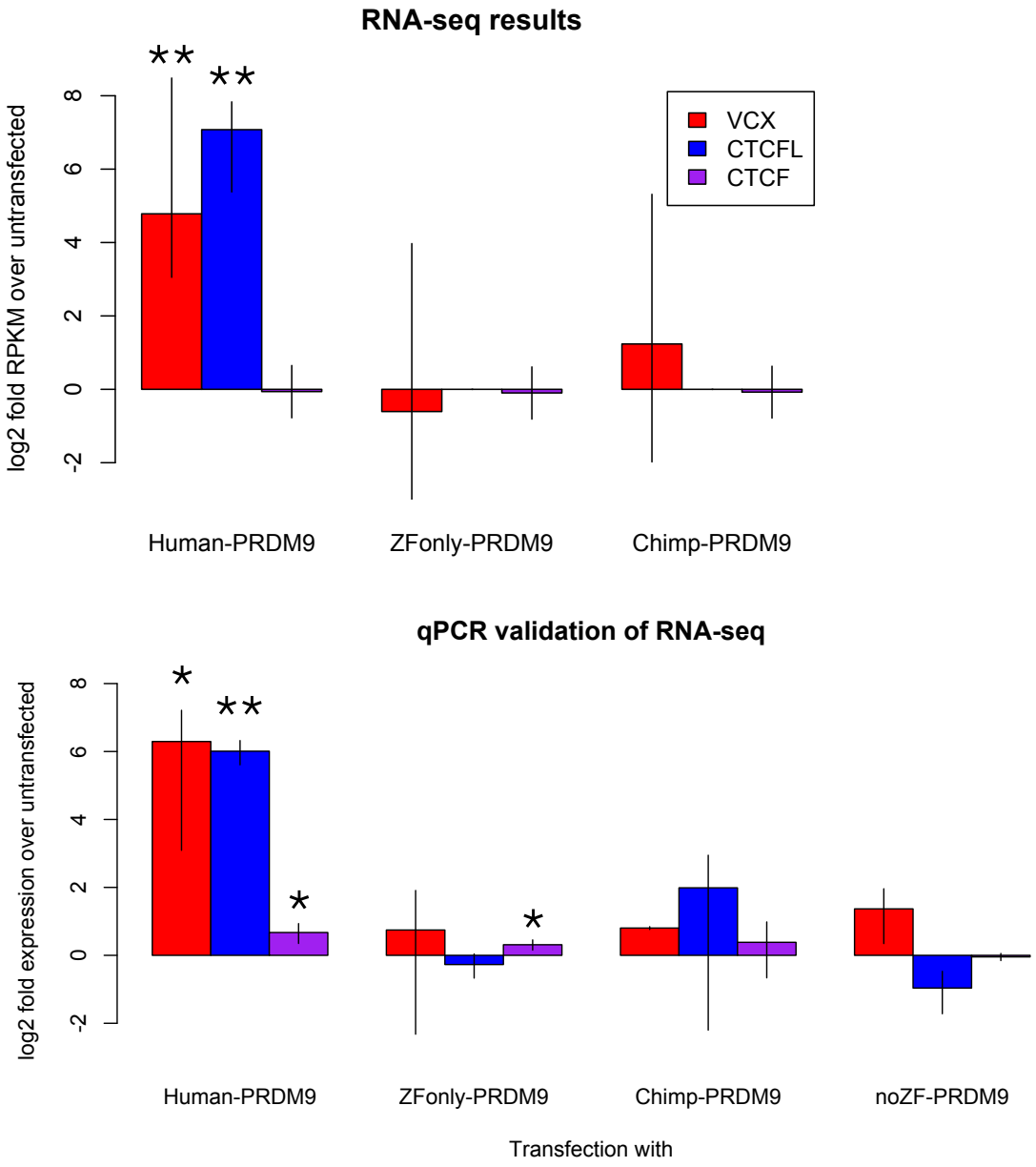


Figure 2.19: Human PRDM9 expression induces expression of *CTCFL* and *VCX*. Upper: bar plots showing the $\log_2$ fold change in computed FPKM (fragments per kilobase of transcript per million mapped reads) values relative to the untransfected sample for *CTCFL* and *VCX*, with *CTCF* as a negative control. Error bars represent maximum ranges of the ratios given confidence intervals for FPKM values computed by cufflinks (*i.e.* if untransfected FPKM values have a confidence interval of $[U_{low},U_{high}]$ and transfected samples have $[T_{low},T_{high}]$, then the ratio interval is reported as $[(T_{low}/U_{high}), (T_{high}/U_{low})]$). Double asterisks indicate significant differential gene expression, as reported by CuffDiff. Lower: qPCR validation results for the same genes from independent biological replicates. Y-axis values are $\log_2$ ratios of $\Delta\Delta$ C_t values for each gene relative to the *TBP* housekeeping gene (see Methods). Error bars represent 2 standard deviations from three technical replicates, and asterisks indicate $p<0.05$ (one asterisk) or $p<0.01$ (two asterisks; one-tailed t test).

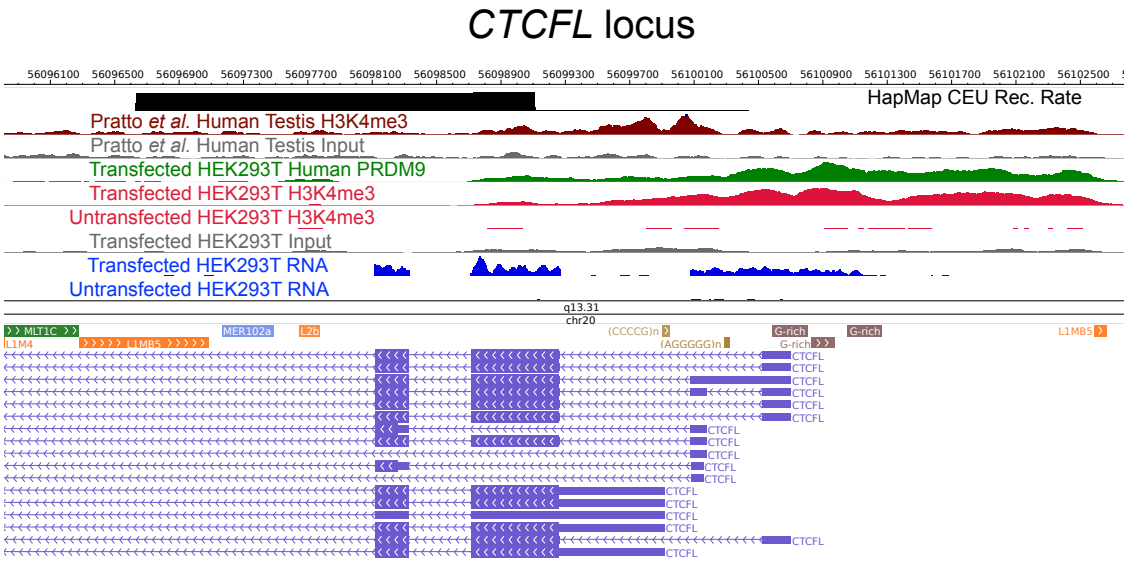


Figure 2.20: Raw coverage values surrounding the *CTCFL* promoter. A browser screenshot from Chr20 near the promoter region of *CTCFL* with custom tracks indicating ChIP-seq and RNA-seq raw coverage data. Human PRDM9 (green) binds a G-rich repeat near the TSS, adding an H3K4me3 mark (light red) where none is present in untransfected cells. RNA-seq coverage (blue) spikes in the coding regions in transfected cells, while it is nearly flat in untransfected cells. Testis H3K4me3 coverage (dark red, from [17]) peaks at a slightly different locus, corresponding to an alternative TSS. An LD-based recombination hotspot is visible in the HapMap CEU Recombination Rate track (top, black) near the promoter region.

input reads) and with defined FPKM values in the untransfected sample. Of the 14,667 genes passing these filters, 10,652 (73%) have a human PRDM9 binding peak within 500 bp of the TSS. Of these, 873 showed at least some evidence of differential expression between the human-transfected and untransfected samples ($p<0.05$), and of these 76 are significant after correction for multiple testing, with 46 significant only in the human-transfected sample ($p<0.05$ after Benjamini-Hochberg correction) (see **Figure 2.19** and **Tables 2.3** and **2.4**; note: *VCX* does not pass these filters due to its unusual H3K4me3 placement away from the TSS). 44/46

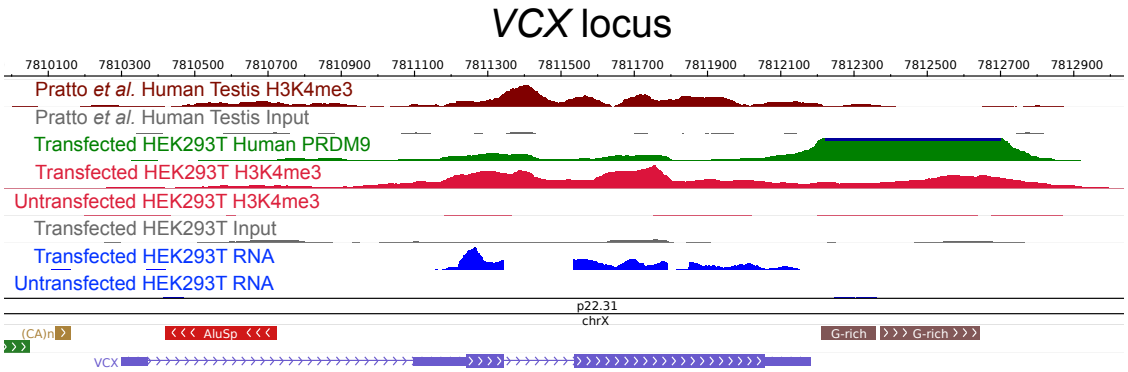


Figure 2.21: Raw coverage values surrounding the *VCX* gene. A browser screenshot from ChrX containin the *VCX* gene with custom tracks indicating ChIP-seq and RNA-seq raw coverage data. Human PRDM9 (green) binds a G-rich repeat near the terminus of *VCX* as well as two loci in the middle of the gene, adding H3K4me3 marks (light red) where none were present in untransfected cells. RNA-seq coverage (blue) spikes in the coding regions in transfected cells, while it is nearly flat in untransfected cells. Testis H3K4me3 coverage (dark red, from [17]) also increases in the gene body, instead of near the annotated TSS.

of these genes show increases, as opposed to decreases, in expression. Because of the low coverage of our RNA-seq data, we are relatively underpowered to detect small changes in gene expression, especially decreases in expression, and especially at short genes or genes with complex alternative splicing patterns [114]. Thus, it remains difficult to assign a precise estimate of what fraction of genes bound and marked by PRDM9 experience changes in expression levels, but it is likely that effects of similar magnitude to *CTCFL* are quite rare. However, our data do make it clear that PRDM9 binding and trimethylation near a promoter can trigger or enhance gene expression in some cases, although it is not sufficient to activate strong expression. Our discovery of differential *CTCFL* expression provides a promising candidate for an *in vivo* target of PRDM9.

2.3 Discussion

Here I have presented the first direct binding map of human PRDM9 in human cells, and through careful analysis of these data I have identified several novel properties of this fascinating protein. Firstly, by applying a novel motif-finding algorithm to the sequences bound by PRDM9 in our cells, I was able to leverage the narrow widths

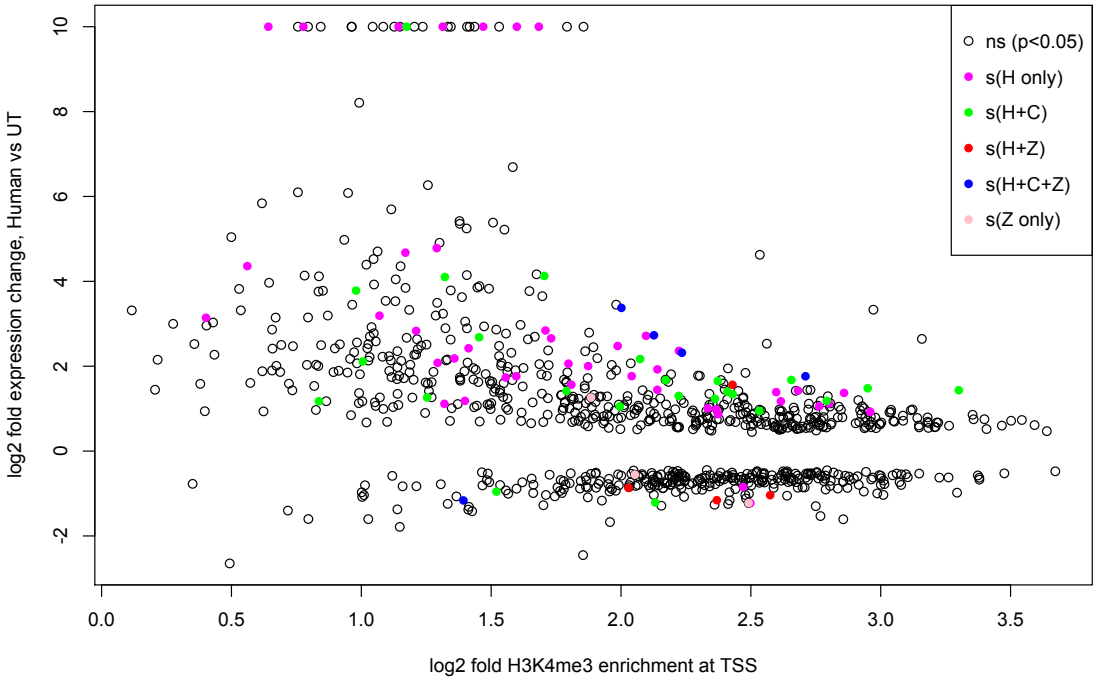


Figure 2.22: Other genes possibly activated by PRDM9. Each point represents a gene with at least some evidence ($p<0.05$) of differential expression in human-transfected cells, with colours indicating genes significant after correction for multiple testing in various different combinations of transfected samples relative to untransfected cells. Genes with infinite fold change values (due to zero untransfected expression) are capped at a $\log_2$ FPKM ratio of 10.

and large number of our ChIP-seq peaks to uncover several unexpected properties of its DNA-binding ZF array. I demonstrated for the first time that several of the upstream zinc fingers have high binding specificity, implying that human PRDM9 can likely bind across the full length of its zinc finger array. By iteratively searching for multiple motifs I effectively allowed for insertions and deletions (indels) within the binding footprint, revealing 7 different modes of PRDM9 binding with different internal spacings between DNA-contacting zinc fingers (**Figure 2.6**). This may explain why published motifs lack sequence specificity corresponding to the upstream zinc fingers—without allowing for indels, the various modes of binding would have been superimposed in this region, diluting the signal at each position. The zinc fingers corresponding to regions of alternative spacing show generally low sequence specificity and contain at their DNA-contacting positions neutrally charged amino acids and a large, aromatic tryptophan, which we speculate may occupy a larger

footprint when contacting the DNA (or may prevent DNA binding altogether). We propose that these ZFs (numbers 5 and 6, of type WSVT) act as linkers of sorts, capable of bending away from the DNA so that PRDM9 can assume a more stable conformation in which more strongly binding zinc fingers are able to bind upstream and downstream. This model is consistent with studies of zinc finger proteins suggesting that there is a limit to the number of contiguous zinc fingers that may bind simultaneously without strain (reviewed in [116]). Indeed, modern techniques for engineering custom zinc finger proteins to bind long motifs involve inserting linker sequences between modules of 2 or 3 zinc fingers to reduce this strain and improve binding affinity and specificity (reviewed in [116]). Thus, the behaviour of the human ZF array likely falls somewhere between a model in which every zinc finger contacts the DNA, and one in which only a few zinc fingers bind (as for CTCF).

Interestingly, I found that the different spacings within motifs may affect their recombination outcomes. PRDM9 peaks containing Motif 7 have lower mean recombination rates and lower signals of meiotic DSBs in the surrounding region relative to the other motifs, even after controlling for potential confounders such as binding enrichment and repeat/promoter context. The cause of this highly significant signal remains an open question, but I hypothesise that it might reflect a meiosis-specific difference at loci containing this motif, such as competitive binding of another transcription factor or a cell-type-specific repressive chromatin mark. It could also reflect a conformational difference when PRDM9 binds in this mode, which might render it less able to recruit the recombination machinery. Alternatively, it might correspond to a B-allele-specific binding mode not represented in LD-based recombination maps, which predominantly represent the A-allele. I also show that motif strength correlates with but does not guarantee PRDM9 binding probability, confirming that factors apart from the primary DNA sequence play a large role in regulating PRDM9 binding, and since not all binding sites overlap hotspots, another layer of regulation must affect recombination outcomes.

One strongly negative predictor of recombination outcomes is presence within an active gene promoter, an effect which has been previously observed in mice [37].

However, I show for the first time that human PRDM9 can bind to most active promoters (82%), perhaps owing to the GC-richness of its motif, although this binding enrichment tends to be weaker for a given motif score than in non-promoter regions (**Figure 2.7**). This fact is surprising and unexpected, given that DSB formation and recombination are known to be suppressed at promoters and other PRDM9-independent H3K4me3 peaks *in vivo* [37]. This demonstrates that the mechanism responsible for suppressing recombination at promoters likely operates downstream of PRDM9 binding but upstream of DSB formation, and the direction of recombination away from promoters does not simply owe to PRDM9's binding preferences or creation of competitive H3K4me3 peaks, as has been suggested in mice with AT-rich PRDM9 binding motifs [37]. If PRDM9 were responsible for directing recombination away from promoters, one would expect any PRDM9 alleles with promoter-enriched binding to be heavily selected against, but instead the human A-allele (which is nearly identical to the B-allele) has reached near-fixation in European humans. Although PRDM9 ChIP-seq against the human A/B-allele has not been performed directly *in vivo*, given the similarity of promoter composition and organisation across cell types, this phenomenon likely occurs *in vivo* as well. This promoter enrichment signal is simply unobserved in existing *in vivo* studies of H3K4me3 and DMC1 ChIP-seq data due to filtering of PRDM9-independent H3K4me3 peaks and the suppression of DSB formation at these sites. It could be the case that PRDM9 initiates recombination only when both its ZF domain binds to DNA and its PR/SET domain forms H3K4me3 *de novo*. One might hypothesise that catalytic activity of the PR/SET domain coupled with zinc finger binding could induce a conformational change in PRDM9 that exposes a protein-protein interaction domain responsible for recruiting recombination machinery. Under such a model, when PRDM9 binds to promoters already saturated for the H3K4me3 mark, it is unable to catalyse H3K4 trimethylation and is thus unable to recruit the recombination machinery. Alternatively, the joint presence of H3K4me3 and H3K36me3, which I show for the first time is enriched in *cis* around PRDM9 binding

sites, might be the true signature of PRDM9 binding and catalysis that signals the cell to localise and activate SPO11 [100].

The ability of human PRDM9 to bind and trimethylate promoters also raises the intriguing possibility that it might also act as a transcription factor, regulating the expression of downstream genes, as was originally hypothesised before the discovery of its role in recombination [38]. Here using RNA-seq data I have shown that PRDM9 can alter the expression of a subset of the genes whose promoters it binds, including the spermatogenesis-specific CTCFL, which is expressed in the same meiotic stages as PRDM9. This alteration of expression may be due to the H3K4me3 mark, but could also be due to the local nucleosome phasing pattern induced by PRDM9 binding. The ability of PRDM9 to affect the transcription of bound promoters may simply add another dimension to its pleiotropic effects across the genome, and this may even help to explain why a single PRDM9 allele predominates in humans. That is, while a multitude of alleles may function equally well in specifying sites of meiotic recombination initiation, perhaps a subset can positively affect fertility by enhancing the expression of key meiotic genes such as CTCFL, and these alleles are driven to high frequency by positive selection. A similar mechanism may also explain why many chimp PRDM9 alleles share a predicted submotif [73]. Here, we have experimentally identified this exact group of zinc fingers as a dominating influence on binding targets, for the first time **Figure 2.17**. Because DSBs are suppressed at promoters, PRDM9 binding sites at promoters might be immune from the effects of hotspot death, which would otherwise eventually abolish its motifs and drive potentially deleterious mutations to fixation in these regions. Indeed, speculatively, this may even explain why recombination is actively suppressed at promoters in certain organisms.

2.4 Methods

2.4.1 Cloning

A cDNA was custom synthesised to contain the full-length (2,685 bp) *PRDM9* transcript from the human reference genome (GRCh37), which is the B allele of

PRDM9. 218 synonymous base changes were engineered into the exon containing the zinc finger domain in order to distinguish the synthetic copy of *PRDM9* from the endogenous copy and to facilitate proper synthesis of this highly repetitive region. Nudrat Noor cloned this cDNA into the pLEXm transient expression vector [117] by ligation with a Venus (YFP) tag at its N-terminus, fused using an AgeI restriction site. A similar synthesised construct was designed to match exon 10 of the chimp PRDM9 reference allele (the "W11a" allele, 2022 bp, codon optimised for human expression and non-repetitiveness). Exons 1-9 were amplified from the human construct, and the chimp allele was fused at the N-terminus with an XbaI site. The ZFonly and noZF alleles were amplified using internal primers designed inside the full-length human construct. For the C-terminally tagged constructs, a 198-bp HA and 213-bp V5 linker were synthesised (having the sequence linker-TwinStrep-linker-HA/V5-linker-P2A) and cloned between each respective PRDM9 allele and a YFP tag using KpnI and AgeI sites, respectively. All construct sequences are listed in Appendix A. C-terminally tagged constructs were cloned into the pLENTI CMV/TO Puro DEST vector (Addgene plasmid # 17293 [118]), owing to its higher transient expression efficiency and to test the possibility of stable lentiviral transduction. Cloning into this vector was performed using the Gateway recombinase-based cloning system (Thermo Fisher Scientific). Constructs were cloned, amplified, and isolated using an Qiagen EndoFree Plasmid Giga Kit to yield transfection-quality DNA, which was verified by restriction digestion and Sanger sequencing.

2.4.2 Transfection

HEK293T cells (ATCC CRL-3216) were chosen owing to their high transfection efficiency, rapid growth rate, and low-cost media requirements. Large-scale transfections of the N-terminal GFP-tagged Human PRDM9 construct were performed as described [117] by Yuguang Zhao in a collaboration with Radu Aricescu. Cells were grown in DMEM media (10% FCS, 1X NEAA, 2 mM L-Glut, Sigma D6546) in 200 ml roller bottles at 37°C/5% CO_2. A transfection cocktail was prepared for each bottle by adding 0.5 mg of chloroform-purified construct DNA to 50 ml of

serum-free DMEM (1X NEAA, 2 mM L-glut) and 1 mg polyethylenimine, followed by a 10-minute incubation, and then addition of 375 μg of kifunensine. After the cells reached 75% confluence, the growth medium was removed from each roller bottle and replaced with 200 ml low-serum DMEM (2% FCS, 1X NEAA, 2 mM L-Glut) and 50 ml transfection cocktail. Cells were then incubated for 72 hours to enable expression of the transfected construct. Expression was verified by placing a small aliquot of detached cells on a glass slide with DAPI and viewing them under a confocal fluorescence microscope at 20× magnification.

I performed all subsequent smaller-scale transfections of the C-terminally tagged constructs in the pLENTI vector using the FuGENE-HD transfection reagent according to manufacturer instructions. HEK293T cells (ATCC CRL-3216) were thawed and incubated at 37°C with 5% CO_2 in DMEM (Sigma D6546) supplemented with 10% foetal bovine serum (Sigma F7524), 1X L-Glutamine (Sigma G7513), and 1X penicillin/streptomycin (Sigma P0781). Confluent cells were split 1:10 and passaged for no longer than one month before transfection. The night before transfection, confluent cells were trypsinised (Sigma T3924), diluted in growth medium, and counted on an automatic haemocytometer (BioRad TC20). For each replicate, 15 million cells were seeded in 30 ml growth medium in a T175 cell culture flask. The following morning, cells were transfected by mixing 30 μg total construct DNA into 800 μl OPTI-MEM (Life Technologies 31985062), then carefully adding 90 μl FuGENE-HD Transfection Reagent and flicking to mix, incubating at room temperature for 15 minutes, and then adding the mixture dropwise to each dish while swirling gently to mix. After 48 hours, cells were imaged briefly with a fluorescent microscope to confirm expression, and were subsequently harvested. As negative controls, additional cells were seeded at the same time but were not transfected.

2.4.3 ChIP (N-terminal YFP-Human)

ChIP-seq was performed according to an online protocol produced by Rick Myers's laboratory [105], which was used to produce much of the ENCODE Project's ChIP-seq data [110], with several optimising modifications.

Crosslinking. Bottles were removed from the incubator and shaken vigorously to detach cells. Fresh formaldehyde was added to a final concentration of 0.75% and cells were incubated at room temperature for 15 minutes. The crosslinking reaction was stopped by adding glycine to a final concentration of 125 mM. Cells were aliquoted to 50 ml conical tubes, centrifuged (2000g, 5 minutes), resuspended in cold 1X PBS, and centrifuged again. Pellets were snap frozen with dry ice, and then stored at -80°C.

Lysis and Sonication. Frozen pellets were thawed and resuspended in cold Farnham Lysis Buffer (5 mM PIPES pH 8.0, 85 mM KCl, 0.5% NP-40, 1 tablet Roche Complete protease inhibitor per 50 ml) to a concentration of 20 million cells per ml, then passed through a 22G needle 20 times to further lyse and homogenise them. Technical replicates were processed in parallel from this point forward (with only one replicate performed for transfected H3K4me3). Lysates were centrifuged and resuspended in 300 μl cold RIPA lysis buffer (1X PBS, 1% NP-40, 0.5% sodium deoxycholate, 0.1% SDS, 1 tablet Roche Complete protease inhibitor per 50 ml) per 20 million cells to lyse nuclei. 300 μl samples were sonicated in a Bioruptor Twin sonication bath in 1.5 ml Eppendorf tubes at 4°C for two 10-minute periods of 30 seconds on, 30 seconds off at high power. Cell debris was removed by centrifugation (14,000 rpm, 15 minutes, 4°C), and supernatants were isolated and brought to a final volume of 1 ml with RIPA. These chromatin preps were snap-frozen in dry ice then stored at -80°C.

Immunoprecipitation. Magnetic beads were washed by adding 200 μl Invitrogen Sheep Anti-Rabbit Dynabeads per sample to 800 μl cold PBS/BSA (1X PBS, 5 mg/ml BSA, 1 tablet Roche Complete protease inhibitor per 50 ml, filtered with 0.45 micron filter). Solutions were placed on a magnetic rack and resuspended in 1 ml PBS/BSA four times. 5 μl Abcam rabbit polyclonal ChIP-grade anti-GFP antibody (ab290) or rabbit polyclonal ChIP-grade anti-H3K4me3 antibody (ab8580) was added and solutions were incubated overnight at 4°C on a rotator. Antibody-coupled beads were washed three times with cold PBS/BSA and resuspended in 100 μl PBS/BSA, then added to 1 ml chromatin preps thawed on ice. One tube

was prepared in parallel without adding beads, to yield a genomic background control sample from total chromatin. Tubes were incubated for 12 hours on a rotator at 4°C, then washed 5 times for 3 minutes each with cold LiCl Wash Buffer (100 mM Tris pH 7.5, 500 mM LiCl, 1% NP-40, 1% sodium deoxycholate, filtered with a 0.45 micron filter unit), then washed once with cold 1X TE (10 mM Tris-HCl pH 7.5, 0.1 mM Na_2-EDTA). Bead pellets were resuspended in 200 μl room-temperature IP elution buffer (1% SDS, 0.1 M $NaHCO_3$, filtered with a 0.45 micron filter unit) and vortexed to mix.

Reverse crosslinking and DNA purification. Samples were incubated in a 65°C water bath for 1 hour with mixing at 15-minute intervals to uncouple beads from protein-DNA complexes. Samples were centrifuged (14,000 rpm, 3 mins) and placed on a magnet to pellet beads, and supernatants were isolated and then incubated in a 65°C water bath overnight to reverse crosslinks. DNA was purified using a Qiagen MinElute reaction cleanup kit and quantified using a Qubit High Sensitivity DNA kit.

2.4.4 ChIP (C-terminal-tagged constructs)

Slight modifications were made for the smaller-scale transfection experiments with C-terminally tagged constructs. Crosslinking was performed in 1% formaldehyde for 5 minutes. Input chromatin was "pre-cleared" to remove chromatin bound non-specifically by the beads. For each sample, 50 μl of equilibrated magnetic beads were resuspended in 100 μl PBS/BSA and added to the chromatin samples for pre-clearing for two hours at 4°C with rotation. Beads were removed, and 100 μl of pre-cleared chromatin was set aside for the input control. 5 μl ChIP-grade rabbit polyclonal antibody (Abcam anti-HA ab9110, anti-V5 ab9116, anti-H3K4me3 ab8580, or anti-H3K36me3 ab9050) was added to the remaining pre-cleared chromatin and incubated overnight at 4°C with rotation. 50 μl beads were washed and resuspended as before, then incubated with the chromatin samples for two hours at 4°C with rotation. After washing and decrosslinking, samples were further incubated with 80 μg RNAse A at 37°C for 60 minutes and then with 80 μg Proteinase K at 55°C for 90 minutes.

2.4.5 ChIP sequencing, mapping, and filtering

DNA was submitted to the Oxford Genomics Centre for library preparation, sequencing, and mapping. For the N-terminal YFP-Human experiments, ChIP and input chromatin DNA samples from transfected and untransfected cells were sequenced in multiplexed paired-end Illumina HiSeq1000 libraries, yielding 51-bp reads. Samples from transfected cells were multiplexed across 3 lanes, yielding roughly 77-101 million properly mapped read pairs (*i.e.* fragments) per replicate (see **Figure 2.1**). Samples from untransfected cells (processed independently) were multiplexed across 2 lanes, yielding roughly 60-99 million properly mapped fragments per sample. For the C-terminal tag experiments, ChIP and input chromatin DNA samples from transfected and untransfected cells were sequenced all together in 6 lanes of paired-end Illumina HiSeq2500 libraries (rapid mode), yielding 51-bp reads with 37 to 64 million reads per replicate (see **Table 2.1**).

Sequencing reads were aligned to hg19 using BWA [119] (option -q 10) followed by Stampy [120] (option –bamkeepgoodreads), and reads not mapped in a proper pair or with an insert size larger than 10 kb were removed. Read pairs representing likely PCR duplicates were also removed by samtools rmdup [121]. Pairs for which neither read had a mapping quality score greater than 0 were removed. For samples with only one replicate, fragments were split at random into two equally-sized pseudo-replicates. Fragment coverage from each replicate was then computed at each position in the genome using in-house code (Appendix 2) and the samtools and bedtools packages [121, 122].

2.4.6 Calling PRDM9 binding peaks

We developed a maximum-likelihood-based peak calling algorithm that takes as input the number of fragments overlapping a bin (a single base position or an interval) from two ChIP replicates and a genomic background control, as well as three constants describing the coverage ratios between these three inputs, which are estimated genome-wide in an initialisation step. The Poisson distribution was chosen as a model of sequencing coverage given its support on all non-negative

integers and simple parameterisation. As specified, this model assumes that the coverage due to signal is proportional between the two ChIP-seq replicates across the genome and that the coverage due to background is proportional among all 3 lanes across the genome. We allow for local estimates of background and signal to account for sequence coverage biases and mappability differences across the genome. *Ab initio* single-base peak calling proceeds in three stages: 1) estimation of constants given coverage values in 100-bp non-overlapping bins genome-wide, 2) single-base maximum likelihood estimation given constants and single-base coverage values, 3) calling of peak centres in the likelihood landscape given a p-value threshold and a threshold on the minimum separation between peak centres (code available in Appendix 2).

Definitions

Let $D_1(i)$, $D_2(i)$ and $G(i)$ be random variables representing the fragment coverage in bin i from the two ChIP-seq replicates and the genomic control, respectively (and let $d_1(i)$, $d_2(i)$ and $g(i)$ represent the observed coverage in bin i). We model the coverage of each sequencing replicate j at bin i as a sample from a Poisson distribution with mean $\lambda_j(i)$,

$$\begin{aligned} D_1(i) &\sim Poisson(\lambda_1(i)), \\ D_2(i) &\sim Poisson(\lambda_2(i)), \\ G(i) &\sim Poisson(\lambda_g(i)), \end{aligned}$$

$$\begin{aligned} \lambda_1(i) &= \alpha_1 b(i) + c(i), \\ \lambda_2(i) &= \alpha_2 b(i) + \beta c(i), \\ \lambda_g(i) &= b(i), \end{aligned}$$

where α_1 and α_2 are constants defining how coverage due to background in the ChIP replicates compares to $b(i)$, a parameter representing the mean coverage in

the genomic control lane at bin i; and β is a constant defining how coverage due to binding enrichment in ChIP replicate 2 compares to $c(i)$, a parameter representing the coverage due to binding enrichment in ChIP replicate 1 at bin i. We wish to test the hypothesis that $c(i) \geq 0$ for each bin i.

Estimating constants

To speed up this step and to provide smoother coverage estimates, I first computed coverage values in 100-bp bins across the autosomes. One can estimate α_j by assuming (conservatively) that when $d_1(i) = 0$ or $d_2(i) = 0$, $c(i) = 0$. That is, one can assume that if ChIP replicate j has coverage 0 at bin i, then any coverage in the other replicate (j') arises purely from background. Thus for all i such that $d_j(i) = 0$

$$\lambda_{j'}(i) = \alpha_{j'} b(i),$$
$$\mathbb{E}_{genome}[\lambda_{j'}(i)] = \alpha_{j'} \mathbb{E}_{genome}[b(i)],$$

and thus one can estimate $\alpha_{j'}$ as

$$\hat{\alpha}_{j'} = \frac{\sum_{i:d_j(i)=0} d_{j'}(i)}{\sum_{i:d_j(i)=0} g(i)}. \tag{2.1}$$

Now an initial estimate of β can be computed using genome-wide coverage means $\bar{d}_1,\ \bar{d}_2,\ \bar{g}$ as follows:

$$\bar{d}_1 \approx \hat{\alpha_1}\bar{g} + \mathbb{E}_{genome}[c(i)],$$
$$\bar{d}_2 \approx \hat{\alpha_2}\bar{g} + \beta\mathbb{E}_{genome}[c(i)],$$

$$\hat{\beta} \approx \frac{\bar{d}_2 - \hat{\alpha_2}\bar{g}}{\bar{d}_1 - \hat{\alpha_1}\bar{g}}. \tag{2.2}$$

Next, maximum likelihood estimation and hypothesis testing are performed across all bins (see below), and $\hat{\beta}$ is re-computed as above, using coverage means from the subset of bins with p$< 10^{-10}$, for which the ratio of coverage between the two replicates will be less affected by noise.

Finally, using the MLEs $\hat{b}(i)$ and $\hat{c}(i)$ for each bin (see subsection below), a genome-wide estimate of the proportion of reads from signal is computed as

$$\frac{\sum_i \hat{c}(i)}{\sum_i (\hat{\alpha_1}\hat{b}(i) + \hat{c}(i))} \tag{2.3}$$

for replicate 1 and as

$$\frac{\sum_i \hat{\beta}\hat{c}(i)}{\sum_i (\hat{\alpha_2}\hat{b}(i) + \hat{\beta}\hat{c}(i))} \tag{2.4}$$

for replicate 2.

Hypothesis Testing

With these estimates of α_j and β, one can compute Maximum Likelihood Estimators for the unknown parameters $b(i)$ and $c(i)$ at each bin i from the coverage data $d_1(i)$, $d_2(i)$ and $g(i)$ (see below for derivation). Then, using these MLEs one can compute a log-likelihood ratio test statistic against a null model in which $c(i) = 0$:

$$\Lambda(i) = 2\log\frac{\max_{b(i),c(i)\geq 0}[L(D_1(i) = d_1(i), D_2(i) = d_2(i), G(i) = g(i))]}{\max_{b(i),c(i)=0}[L(D_1(i) = d_1(i), D_2(i) = d_2(i), G(i) = g(i))]}. \tag{2.5}$$

Under the null hypothesis, the test statistic $\Lambda(i)$ is distributed approximately as a χ^2 distribution (with 1 degree of freedom due to the parameter $c(i)$ and an atom of probability at 0), yielding a p-value at each bin i indicating the probability that the observed likelihood ratio could arise from background alone.

Calculation of Maximum Likelihood Estimators

Recall that at each position the Poisson means for coverage in each lane are (dropping the i notation for succinctness)

$$\lambda_1 = \hat{\alpha_1} b + c,$$
$$\lambda_2 = \hat{\alpha_2} b + \hat{\beta} c,$$
$$\lambda_g = b,$$

where $\hat{\alpha_1}$, $\hat{\alpha_2}$, and $\hat{\beta}$ are constants estimated for the whole genome. To simplify calculations, we reparameterise using a new variable $y = c/b$ and rewrite the above equations as

$$\lambda_1 = \hat{\alpha_1} b + yb,$$
$$\lambda_2 = \hat{\alpha_2} b + \hat{\beta} yb,$$
$$\lambda_g = b.$$

Given the observed coverage values d_1, d_2, and g, the Poisson log likelihood function can be written as

$$\begin{aligned}
\ell &\propto -\lambda_1 + d_1 \log(\lambda_1) - \lambda_2 + d_2 \log(\lambda_2) - \lambda_g + g\log(\lambda_g) \\
&= -\hat{\alpha_1} b - yb + d_1 \log(\hat{\alpha_1} b + yb) - \hat{\alpha_2} b - \hat{\beta} yb + d_2 \log(\hat{\alpha_2} b + \hat{\beta} yb) - b + g\log(b) \\
&= -b(\hat{\alpha_1} + \hat{\alpha_2} + 1) - yb(1 + \hat{\beta}) + d_1 \log(\hat{\alpha_1} b + yb) + d_2 \log(\hat{\alpha_2} b + \hat{\beta} yb) + g\log(b).
\end{aligned} \tag{2.6}$$

Now to maximise ℓ we first obtain the partial derivatives for b and y

$$\begin{aligned}
\frac{\partial \ell}{\partial b} &= -(\hat{\alpha_1} + \hat{\alpha_2} + 1) - y(1 + \hat{\beta}) + \frac{d_1(\hat{\alpha_1} + y)}{b(\hat{\alpha_1} + y)} + \frac{d_2(\hat{\alpha_2} + \hat{\beta} y)}{b(\hat{\alpha_2} + \hat{\beta} y)} + \frac{g}{b} \\
&= -(\hat{\alpha_1} + \hat{\alpha_2} + 1) - y(1 + \hat{\beta}) + \frac{1}{b}(d_1 + d_2 + g),
\end{aligned} \tag{2.7}$$

$$\begin{aligned}
\frac{\partial \ell}{\partial y} &= -b(1 + \hat{\beta}) + \frac{d_1 b}{b(\hat{\alpha_1} + y)} + \frac{d_2 \hat{\beta} b}{b(\hat{\alpha_2} + \hat{\beta} y)} \\
&= -b(1 + \hat{\beta}) + \frac{d_1}{(\hat{\alpha_1} + y)} + \frac{d_2 \hat{\beta}}{(\hat{\alpha_2} + \hat{\beta} y)}.
\end{aligned} \tag{2.8}$$

Next, we set the partials to 0 and solve them as a system to obtain any potential local maxima. We start by solving for b in Equation 2.7 as follows:

$$\begin{aligned} 0 &= -(\hat{\alpha_1} + \hat{\alpha_2} + 1) - y(1 + \hat{\beta}) + \frac{1}{b}(d_1 + d_2 + g); \\ b &= \frac{d_1 + d_2 + g}{\hat{\alpha_1} + \hat{\alpha_2} + 1 + y(1 + \hat{\beta})}. \end{aligned} \tag{2.9}$$

Then, we substitute it into Equation 2.8 and rewrite it as follows, with the aim of simplifying it into quadratic form:

$$\begin{aligned} 0 = -\frac{d_1 + d_2 + g}{\hat{\alpha_1} + \hat{\alpha_2} + 1 + y(1 + \hat{\beta})}(1 + \hat{\beta}) + \frac{d_1}{(\hat{\alpha_1} + y)} + \frac{d_2\hat{\beta}}{(\hat{\alpha_2} + \hat{\beta}y)}; \\ \frac{(d_1 + d_2 + g)(1 + \hat{\beta})}{\hat{\alpha_1} + \hat{\alpha_2} + 1 + y(1 + \hat{\beta})} &= \frac{d_1(\hat{\alpha_2} + \hat{\beta}y) + d_2\hat{\beta}(\hat{\alpha_1} + y)}{(\hat{\alpha_1} + y)(\hat{\alpha_2} + \hat{\beta}y)} \\ &= \frac{d_1\hat{\alpha_2} + d_1\hat{\beta}y + d_2\hat{\beta}\hat{\alpha_1} + d_2\hat{\beta}y}{\hat{\alpha_1}\hat{\alpha_2} + \hat{\alpha_1}\hat{\beta}y + \hat{\alpha_2}y + \hat{\beta}y^2} \\ &= \frac{y(d_1\hat{\beta} + d_2\hat{\beta}) + d_1\hat{\alpha_2} + d_2\hat{\beta}\hat{\alpha_1}}{\hat{\alpha_1}\hat{\alpha_2} + y(\hat{\alpha_1}\hat{\beta} + \hat{\alpha_2}) + \hat{\beta}y^2}. \end{aligned} \tag{2.10}$$

To shorten notation, we substitute in the following variables for constant terms in Equation 2.10:

$$t_1 = (g + d_1 + d_2)(1 + \hat{\beta}),$$

$$t_2 = \hat{\alpha_1} + \hat{\alpha_2} + 1,$$

$$t_3 = 1 + \hat{\beta},$$

$$t_4 = d_1\hat{\alpha_2} + d_2\hat{\beta}\hat{\alpha_1},$$

$$t_5 = d_1\hat{\beta} + d_2\hat{\beta},$$

$$t_6 = \hat{\alpha_1}\hat{\alpha_2},$$

$$t_7 = \hat{\alpha_1}\hat{\beta} + \hat{\alpha_2},$$

yielding

$$\begin{aligned}\frac{t_1}{t_2+yt_3} &= \frac{yt_5+t_4}{t_6+yt_7+\hat{\beta}y^2};\\ 0 &= t_1(t_6+yt_7+\hat{\beta}y^2)-(t_2+yt_3)(yt_5+t_4);\\ 0 &= t_1t_6+yt_1t_7+t_1\hat{\beta}y^2-yt_2t_5-t_2t_4-y^2t_3t_5-yt_3t_4;\\ 0 &= y^2(t_1\hat{\beta}-t_3t_5)+y(t_1t_7-t_2t_5-t_3t_4)+(t_1t_6-t_2t_4). \qquad (2.11)\end{aligned}$$

Now we can solve for y in Equation 2.11 using the quadratic formula, taking the positive root to be $\hat{y}$, the MLE for y, which we report as the "enrichment" value for that bin. To obtain $\hat{b}$, we simply substitute $\hat{y}$ into Equation 2.9 and, to return to the original paramaterisation, $\hat{c}$ is simply computed as $\hat{y}\hat{b}$. Finally, to obtain $\hat{b}_0$, the MLE for b under the background model, we can simply set y to 0 in Equation 2.9, yielding

$$\hat{b}_0 = \frac{d_1+d_2+g}{\hat{\alpha_1}+\hat{\alpha_2}+1}. \qquad (2.12)$$

Peak calling and centring

Given a likelihood ratio value $\Lambda(i)$ for each base i along a chromosome, along with a p-value threshold (which is converted to a lower bound on the likelihood ratio, l) and m, a threshold on the minimum separation between peak centres, initial peak centres are found by identifying all significant bases (bases for which $\Lambda(i) > l$) that are local maxima. Specifically, each significant base is scanned to test if

$$[\Lambda(i) > \max_{i-m<j<i-1} \Lambda(j)] \text{ and } [\Lambda(i) \geq \max_{i+1<j<i+m} \Lambda(j)].$$

At each initial peak centre satisfying these criteria, a confidence interval is computed by identifying the nearest position j to the left and to the right (by a maximum of 1000 bp) where $(\Lambda(i)-\Lambda(j)) > 9.12$, which defines a 99% confidence interval for the peak centre (using χ^2_2, with one degree of freedom for the enrichment factor and one for the peak centre position). All confidence intervals along a chromosome are then sorted from narrowest to widest, and in this order each confidence interval is added one at a time to the final peak set, provided it does not overlap any of the confidence intervals already included in the final peak set. This produces a final peak set with

non-overlapping confidence intervals, favouring inclusion of stronger peaks with narrower confidence intervals. Finally, to refine peak centres in confidence intervals with multiple tied bases, the rounded mean position of all maximal bases is reported as the peak centre. The resulting final peak set reports $\hat{y}$ and the p-value for Λ at the peak centre as the enrichment and p-value for that peak.

Force-calling

This algorithm enables maximum likelihood estimation and hypothesis testing at any arbitrary bin in the genome, when provided with coverage values and estimates of α_1, α_2, and β. This enables us to "force-call" enrichment and p-values at pre-specified locations in the genome, for example to determine what fraction of gene promoters show evidence of H3K4me3 enrichment in a 1-kb window centred on the transcription start site.

Overlap correction

When comparing peak sets to determine overlap proportions, one must account for chance overlaps owing to the width and number of peaks being compared. For comparisons between single-base peak centres and DSB hotspot intervals, for example, I computed the expected number of chance overlaps c between the n peak centres and the t hotspot intervals, each with width w_i, in a genome of size g as

$$c = \sum_{i \in t} \left(1 - \left(\frac{g - w_i}{g} \right)^n \right). \tag{2.13}$$

For more complicated comparisons, for example between two sets of intervals, I computed chance overlaps by randomly shifting the positions of one set of intervals uniformly in the interval [-60000, 60000], then counted the resulting overlaps to estimate c.

Given f observed overlaps between the sets of n and t peaks, we can compute the corrected overlap fraction, o/t as follows. Let o/t be the proportion of systematic overlaps, c/t be the fraction of chance overlaps, and f/t be the proportion of total

overlaps. The probability of no overlap is simply the product of the complements of chance and systematic overlaps, as follows:

$$(1 - f/t) = (1 - o/t)(1 - c/t).$$

Solving for o/t then yields:

$$o/t = 1 - \frac{1 - f/t}{1 - c/t}. \tag{2.14}$$

2.4.7 Motif finding

An implementation of the motif-finding algorithm created by Simon Myers was obtained from Edouard Hatton (fully specified in [108]). For each peak, a 300-bp sequence (centred on the peak centre) was extracted from the reference sequence. *Ab initio* motif calling was performed on sequences from the top 5,000 peaks (ranked by enrichment) that passed a set of stringent filters ($p<10^{-10}$, enrichment >2, C.I. width ≤50, no bases overlapping annotated repeats, number of input reads between 10%ile and 90%ile, and ≥30 reads from ChIP rep1 + ChIP rep2). Up to 20 seeding motifs were allowed, and seeds were refined for 200 iterations. Three separate runs were performed for each sample to verify consensus. For the YFP-Human peaks, a run producing 15 final motifs was chosen, and of these the 7 motifs with >80% of matches occurring in the central 100 bp of each peak sequence were chosen as the final set in order to remove degenerate motifs (*i.e.* those with little base specificity at any position) as well as likely false positives (such as a match to the motif for the AP1 transcription factor). For the Chimp-HA/V5 peaks, only two motifs were produced, one of which was a degenerate CT-rich motif found in only 10% of peaks (but not centrally enriched), so it was filtered out. These final motifs were then force-called on the full set of peaks (without any peak filtering) by rerunning the refinement algorithm with the option to not update the motifs with each iteration. The motif with the greatest posterior probability (of at least 0.75) of a match was reported for each peak, along with position and strand. For identifying motif matches genome wide, I used FIMO (version 4.10.0) [109].

2.4.8 Comparing sequencing datasets

I compared the newer Human-HA/V5 data with the original YFP-Human data and found strong overlap between the peak sets (60%) but a poor correlation in raw coverage values or in our computed enrichment values (r = 0.3). I explored this further and noticed that the newer sequencing run had a strong increase in coverage of GC-rich regions (nearly two-fold higher input coverage in regions with >60% GC), perhaps owing to differences in the ChIP protocol or to downstream differences in the library prep and sequencing steps (Illumina HiSeq 1000 versus Illumina HiSeq 2500). I also cannot exclude any effects due to the different placement of the tags. The sequencing bias in the newer data changes the motifs found by our algorithm: with the Human-HA/V5 data, only one non-degenerate motif is found, and it resembles the 13-bp motif reported previously [33]. Because the original YFP-Human data provide better motif-finding power given their less biased coverage, I utilised these data exclusively for most of the analyses of the human allele. This sequencing bias may explain why only one non-degenerate Chimp motif was found.

2.4.9 ATAC-seq

ATAC libraries were prepared by Emmanuelle Bitoun as described [113]. Briefly, 50,000 cells were lysed in 10 mM Tris-HCl pH 7.4, 10 mM NaCl, 3 mM $MgCl_2$, 0.1% IGEPAL CA-630 and the nuclei were pelleted at 500g for 10 minutes. The transposition reaction was carried out for 30 minutes at 37°C using the Nextera DNA Sample Preparation Kit (Illumina) according to the manufacturer's instructions. The libraries were purified using the MinElute PCR Purification Kit (Qiagen), PCR amplified, multiplexed, and sequenced by the Oxford Genomics Centre on an Illumina HiSeq2500 (rapid mode) to produce 60-77 million sequenced fragments (51-bp, paired-end reads) per sample. Reads were mapped to hg19 using BWA [119] followed by Stampy [120]. PCR duplicates, mtDNA-mapped reads, reads not mapped in a proper pair, reads with mapping quality equal to 0, and pairs with an insert size larger than 2 kb were removed using samtools [121], leaving ∼11 million fragments per sample. Using in-house code, fragments were split by size into into

Gene	Forward primer $5' - 3'$	Reverse primer $5' - 3'$	Reference
TBP	CACGAACCACGG CACTGATT	TTTTCTTGCTGC CAGTCTGGAC	[123]
CTCFL	ACCTGCACAGAC ATTCGGAGAAGT	CTGCACAAACTG CACTGAAACGGA	[123]
CTCF	TCGTCGTTACAA ACACACCCACGA	CTGCACAAACTG CACTGAAACGGA	[123]
VCX	GGCCAAGGCCAC GGAGG	TGGTGAGATCTC TGAGGTCT	[124]
YFP	ACTTCAAGATCC GCCACAAC	GCTTGGACTGGT AGCTCAGG	
PRDM9	GCCACGGCATGC GCCACCATGGTC AAGTATGGAGAG TGTGGACAAGG	ATCGGGTACCTC AATGTGTCCTTT GGTGTGTAATAAC	

Table 2.2: Primers used for qPCR

inter-nucleosome (51-100 bp) and mono-nucleosome fragments (180-247 bp), and the position of the central base in each fragment was reported, as described [113]. This yielded ∼1 million internucleosome and ∼3 million mononucleosome fragments per sample. Fragment centre coverage was computed genome-wide using bedtools [122].

2.4.10 RNA extraction and RT-qPCR

Total RNA was extracted by Emmanuelle Bitoun using the RNeasy kit (Qiagen). For quantitative PCR analysis, RNA was reverse-transcribed using Expand Reverse Transcriptase (Roche), according to the manufacturer's instructions. qPCR reactions were carried out in triplicate using Fast SYBR Green Master Mix (Applied Biosystems) on a CFX real-time C1000 thermal cycler (Bio-Rad), following the manufacturer's guidelines. Data were analysed using the CFX 2.1 Manager software (Bio-Rad) and normalised to the Tata binding protein (*TBP*) gene. Relative gene expression levels were calculated using the $\Delta\Delta C_t$ method. Statistical analysis was carried using a one-tailed t test. Primer sequences are given in **Table 2.2**.

2.4.11 RNA-seq

Total RNA was submitted to the Oxford Genomics Centre for mRNA enrichment, library preparation, and sequencing. Samples were multiplexed and sequenced on an Illumina Hi-Seq2500 (rapid mode), yielding 71-98 million 51-bp paired-end reads per sample. I created a custom reference sequence by merging the hs37d5 reference (used by the 1000 Genomes Project to improve mapping quality [125]) with the construct and vector sequences transfected into our cells. Data were analysed using the Tuxedo software package [114]. Reads were mapped and processed using TopHat (version 2.0.13, options –mate-inner-dist=250 –mate-std-dev 80 –transcriptome-index=Ensembl.GRCh37.genes.gtf); followed by Cufflinks, CuffQuant, and CuffDiff (version 2.2.1); then analysed using CummeRbund.

gene	PRDM9 enrich	H3K4me3 enrich.	fpkm UT	fpkm Human	fpkm Zfonly	fpkm Chimp	delta UH	delta UZ	delta UC	pval UH	pval UZ	pval UC	Chr	TSS position
KRT5	9.369	2.029	0.000	0.260	0.000	0.141	Inf	0.000	Inf	1.00E-04	1.00E+00	2.00E-04	chr12	52914314
KRT9	9.276	1.209	0.000	0.212	0.000	0.005	Inf	0.000	Inf	5.00E-05	1.00E+00	1.00E+00	chr17	39728305
LGALS7	3.990	1.485	0.000	1.000	0.000	0.191	Inf	0.000	Inf	5.00E-05	1.00E+00	3.02E-02	chr19	39264072
RNASE1	8.486	0.561	0.000	0.282	0.000	0.224	Inf	0.000	Inf	1.50E-04	1.00E+00	3.75E-03	chr14	21271437
LGALS9C	1.060	0.714	0.000	0.319	0.000	0.056	Inf	0.000	Inf	5.00E-05	1.00E+00	1.00E+00	chr17	18380112
SH3TC1	11.877	1.769	0.000	0.387	0.000	0.057	Inf	0.000	Inf	5.00E-05	1.00E+00	1.00E+00	chr4	8242571
TH	6.345	2.212	0.000	0.198	0.023	0.079	Inf	Inf	Inf	1.00E-04	1.00E+00	1.00E+00	chr11	2189336
CTCFL	4.575	1.446	0.095	2.625	0.063	0.225	4.782	-0.606	1.235	5.00E-05	1.00E+00	4.60E-02	chr20	56100635
CPNE6	2.156	1.250	0.007	0.178	0.059	0.030	4.676	3.071	2.085	5.00E-05	1.00E+00	1.00E+00	chr14	24540106
CAPN8	1.112	0.476	0.026	0.530	0.157	0.291	4.359	2.603	3.494	1.50E-04	1.23E-02	1.20E-03	chr1	223816407
PAX5	7.200	1.099	0.038	0.351	0.233	0.268	3.190	2.602	2.804	1.00E-04	1.95E-03	1.25E-03	chr9	37002672
C1orf116	1.622	0.321	0.088	0.778	0.351	0.518	3.141	1.992	2.554	5.00E-05	3.80E-03	4.00E-04	chr1	207206101
ONECUT3	2.992	2.270	0.152	1.088	0.154	0.150	2.842	0.019	-0.021	5.00E-05	9.79E-01	9.76E-01	chr19	1752372
LGALS1	2.830	1.315	11.394	81.139	30.436	31.125	2.832	1.417	1.450	5.00E-05	1.65E-03	1.20E-03	chr22	38071615
PDGFB	1.522	3.276	0.233	1.532	0.503	0.556	2.715	1.109	1.255	5.00E-05	8.35E-02	5.23E-02	chr22	39640756
P2RX2	5.616	2.319	1.244	7.843	3.417	3.677	2.656	1.458	1.564	5.00E-05	3.05E-03	1.45E-03	chr12	133195427
NGFR	2.048	2.964	0.626	3.485	1.583	2.088	2.476	1.338	1.738	5.00E-05	2.04E-02	3.60E-03	chr17	47573986
SYT11	0.957	1.663	0.456	2.446	0.957	0.850	2.423	1.069	0.898	5.00E-05	2.54E-02	5.84E-02	chr1	155829300
PALM3	1.890	3.669	1.545	7.929	4.077	4.770	2.359	1.400	1.626	5.00E-05	2.50E-03	4.50E-04	chr19	14168411
HMOX1	0.936	1.564	6.751	30.662	10.291	16.534	2.183	0.608	1.292	5.00E-05	6.99E-02	3.00E-04	chr22	35776828
GAL3ST1	7.307	1.452	1.499	6.332	1.479	2.620	2.079	-0.019	0.806	5.00E-05	9.7E-01	1.29E-01	chr22	30970498
ATP8B3	1.049	2.477	1.130	4.712	2.177	3.128	2.060	0.946	1.469	5.00E-05	5.22E-02	4.20E-03	chr19	1811623

Table 2.3: Genes with significant expression differences in Human PRDM9 samples only ***(Continued on following page).*** 46 protein coding genes with significant differential expression between human-transfected versus untransfected cells (but no significant expression change in the control transfections) are listed along with the enrichment value of the strongest PRDM9 peak within 500 bp of a TSS, the force-called H3K4me3 enrichment value around the TSS, and the RNA-seq values output by Cufflinks and CuffDiff. Genes are listed in reverse order of the fold expression change.

gene	PRDM9 enrich	H3K4me3 enrich.	fpkm UT	fpkm Human	fpkm Zfonly	fpkm Chimp	delta UH	delta UZ	delta UC	pval UH	pval UZ	pval UC	Chr	TSS position
FOSL2	1.967	2.664	1.532	6.134	2.536	3.684	2.002	0.728	1.266	5.00E-05	5.69E-02	1.20E-02	chr2	28615725
SH2D3C	2.872	3.408	1.533	5.824	2.221	3.512	1.926	0.535	1.196	5.00E-05	2.16E-01	5.85E-03	chr9	130517309
CDKN2D	0.771	3.116	5.058	17.190	11.401	10.892	1.765	1.172	1.107	5.00E-05	3.85E-03	6.75E-03	chr19	10679654
MAFK	3.050	2.024	6.186	20.995	11.719	17.739	1.763	0.922	1.520	1.00E-04	5.54E-02	1.00E-03	chr7	1570350
LIF	3.003	1.941	1.503	4.996	4.003	2.896	1.733	1.413	0.946	5.00E-05	5.00E-04	1.22E-02	chr22	30642728
IL6R	1.624	2.503	1.611	4.759	1.916	3.506	1.563	0.251	1.122	5.00E-05	5.09E-01	4.00E-04	chr1	154378091
EPHA2	2.157	3.407	5.909	16.078	9.761	10.661	1.444	0.724	0.851	5.00E-05	1.25E-02	3.60E-03	chr1	16482582
SMAD7	2.888	5.415	4.531	12.164	5.158	7.486	1.425	0.187	0.724	5.00E-05	5.88E-01	3.05E-02	chr18	46475703
NOTCH1	1.855	5.053	6.800	17.804	6.679	11.735	1.389	-0.026	0.787	5.00E-05	9.2E-01	1.10E-03	chr9	139440314
FGFR3	5.369	6.254	11.744	30.364	18.676	23.851	1.370	0.669	1.022	5.00E-05	5.06E-02	3.50E-04	chr4	1795560
SEMA6B	1.287	1.637	6.606	15.000	9.978	12.848	1.183	0.595	0.960	1.50E-04	5.66E-02	2.30E-03	chr19	4558507
PHRF1	0.589	5.129	8.779	19.800	11.884	12.134	1.173	0.437	0.467	5.00E-05	8.50E-02	6.58E-02	chr11	576521
IER2	1.380	5.972	11.676	25.411	17.648	22.595	1.122	0.596	0.952	1.00E-04	5.96E-02	1.60E-03	chr19	13261247
DNAJB2	2.494	1.497	17.410	37.652	25.993	35.953	1.113	0.578	1.046	1.00E-04	4.16E-02	5.50E-04	chr2	220144238
CREBRF	1.839	5.788	2.783	5.778	5.949	4.231	1.054	1.096	0.604	1.00E-04	3.00E-04	2.78E-02	chr5	172483371
KDM6B	1.259	4.042	12.168	24.441	16.253	21.784	1.006	0.418	0.840	5.00E-05	8.44E-02	5.50E-04	chr17	7748233
PPM1D	0.938	4.159	12.296	24.289	20.058	22.153	0.982	0.706	0.849	5.00E-05	5.40E-03	5.50E-04	chr17	58677544
AGRN	1.807	6.779	22.112	42.043	30.938	39.378	0.927	0.485	0.833	1.50E-04	4.51E-02	6.00E-04	chr1	955503
EEF1A2	1.010	4.184	75.109	137.246	105.601	108.222	0.870	0.492	0.527	1.00E-04	2.94E-02	1.98E-02	chr20	62130505
ATXN7L3B	1.602	4.540	48.860	27.135	33.217	27.907	-0.848	-0.557	-0.808	1.00E-04	1.42E-02	7.00E-04	chr12	74931551
PIGM	1.057	4.649	19.365	8.270	11.237	9.532	-1.227	-0.785	-1.023	5.00E-05	8.35E-03	5.50E-04	chr1	160001783

Table 2.4: Genes with significant expression differences in Human PRDM9 samples only *(Continued from previous page).* 46 protein coding genes with significant differential expression between human-transfected versus untransfected cells (but no significant expression change in the control transfections) are listed along with the enrichment value of the strongest PRDM9 peak within 500 bp of a TSS, the force-called H3K4me3 enrichment value around the TSS, and the RNA-seq values output by Cufflinks and CuffDiff. Genes are listed in reverse order of the fold expression change.

The Creator would appear as endowed with a passion for stars, on the one hand, and for beetles on the other, for the simple reason that there are nearly 300,000 species of beetle known.

— J.B.S. Haldane, *What is Life? The Layman's View of Nature* (1949)

3

PRDM9 binding symmetry in hybrid mice is associated with differences in DSB processing, synapsis, and fertility

3.1 Introduction

Prdm9 remains the first and only mammalian gene to be associated with speciation, but the mechanism by which it causes hybrid infertility in mice has remained puzzling. As reviewed in Chapter 1, *Prdm9* has been shown to play an essential role in the complete infertility of F1 hybrid males generated by crossing a *Mus musculus musculus* PWD female with a *Mus musculus domesticus* B6 male. Efforts to dissect the genetics of this hybrid infertility phenotype have revealed three necessary genetic conditions for complete sterility: 1) *Prdm9* must be heterozygous, with exactly one copy of the PWD allele and one copy of the B6 allele—changing the copy number of the B6 allele or adding copies of other alleles can partially or fully restore fertility; 2) a 4.7-Mb region on the X chromosome called *Hstx2* must originate from the PWD background; and 3) the autosomes must be heterosubspecific—introducing long homozygous stretches can reduce asynapsis of that chromosome. Without additional data, it has remained difficult to postulate a mechanistic model to marry these three conditions.

To better understand the role of PRDM9 in recombination and in hybrid infertility, we generated a transgenic line of "Humanized" B6 mice, in which we have replaced the DNA-binding Zinc Finger array of the endogenous B6 *Prdm9* gene with the ZF array from the human reference allele (the B allele). This manipulation effectively reprograms the location of recombination hotspots and completely rescues fertility in (PWD×B6)F1 mice. To explore potential mechanisms for this hybrid rescue, we generated ChIP-seq datasets from adult testes using antibodies against H3K4me3, which is introduced in *cis* by PRDM9, and against DMC1, which marks Double Strand Break sites. In doing so, we present evidence that F1 hybrid infertility arises as a result of accumulated hotspot death, which causes PRDM9 to bind asymmetrically to each homologue in an F1 hybrid, with strong effects on synapsis and fertility.

This work represents a collaborative effort by many people and was recently accepted as an article in *Nature* [108]. Here I describe my major contributions to this work, and summarise our overall major findings for completeness. Specifically, I carried out the H3K4me3 ChIP-seq experiments and developed and applied statistical analysis techniques relating to H3K4me3 peak calling and comparisons with DMC1 at individual hotspots and between mice. Furthermore, I played an important role in the original design of our hybrid fertility rescue experiments and in the design of the subsequent experiments and analyses, and I bred most of the hybrid mice.

3.2 Results

3.2.1 Humanized PRDM9 rescues hybrid fertility

Humanized B6 mice were generated by Ben Davies and his group by first manipulating a mouse embryonic stem cell line to seamlessly replace the endogenous copy of mouse *Prdm9* exon 10 (containing the full ZF array) with a synthesised construct containing the ZF array from the human "B" allele (illustrated in **Figure 3.1**; see [108] for full specification).

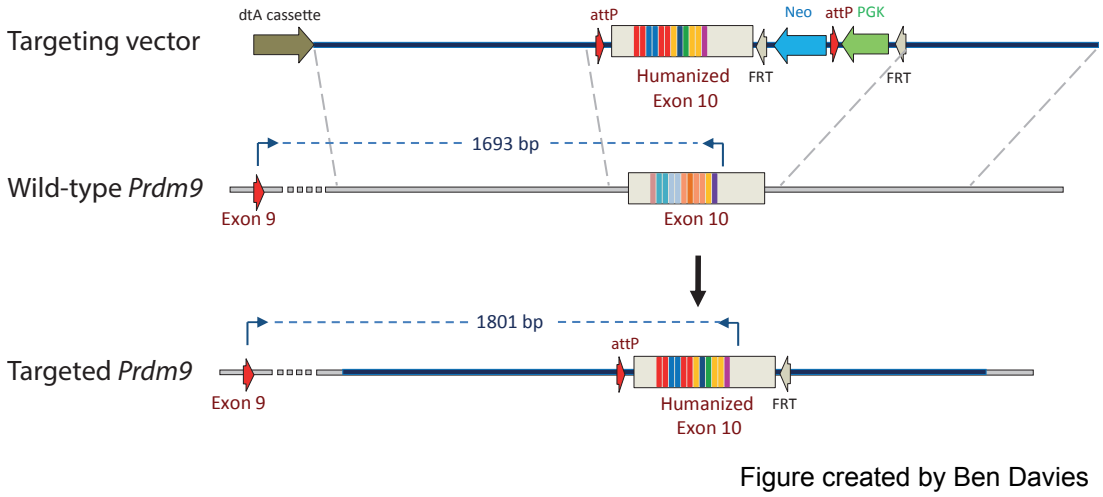


Figure 3.1: Humanization of the *Prdm9* ZF array in B6 mice. Ben Davies and colleagues replaced the entirety of *PRDM9* exon 10 in B6 mice with a "Humanized" exon in which the mouse ZF array was replaced with a synthesised copy of the human *PRDM9* B allele. Controls confirmed equal expression of this allele in mouse testes (see [108]).

The resulting Humanized mice show completely normal fertility in crosses with each other and with wild-type B6 mice, as determined by paired testes weight, sperm count, the ability to sire pups, the rate of forming the sex body, and the rate of autosomal synapsis (as measured by Ben Davies and Daniela Moralli). To explore the effects of this manipulation on hybrid infertility I crossed PWD females with heterozygous $B6^{B6/H}$ males, and fertility of the resulting offspring was measured by Ben Davies and colleagues. As expected, male offspring inheriting the B6 allele showed complete infertility, with an undetectable sperm count, an inability to sire pups, low testis weight, a high rate of failure to produce the sex body, and a high rate of autosomal asynapsis. Intriguingly, their brothers inheriting the Humanized allele showed a complete rescue of fertility by all measures, demonstrating in a more controlled setting than any previous experiment that the DNA-binding domain, and thus likely the DNA-binding activity, of PRDM9 plays a causal role in this hybrid infertility phenotype. I also tested whether humanizing PRDM9 might have any effect on the semi-fertile reciprocal cross of B6×PWD, which shows intermediate fertility metrics. These mice differ from the fully infertile PWD×B6 mice at the sex chromosomes, and this difference in fertility has been mapped to the 4.7-Mb *Hstx2*

locus on ChrX. Despite this difference, in B6×PWD mice inheriting the Humanized PRDM9 allele, full fertility was restored from semi-fertility.

3.2.2 Overlaps between H3K4me3 and DMC1 peaks

To explore the effects of PRDM9 humanization at a molecular level across the genome, I performed ChIP-seq against the H3K4me3 mark in five mice: $B6^{B6/B6}$, $B6^{H/H}$, (PWD×B6)$F1^{PWD/B6}$ (the Infertile mouse), (PWD×B6)$F1^{PWD/H}$(the Humanized Rescue mouse), and (B6×PWD)$F1^{B6/PWD}$ (the Reciprocal semi-fertile mouse). Superscripts indicate *Prdm9* genotypes, and the female strain is written first in crosses. Of this list, Ben Davies generated and bred the transgenic B6 mice, and I performed/led breeding of the hybrid mice. The H3K4me3 mark is deposited by PRDM9 at essentially all of its binding sites (Chapter 2) and so serves as a natural measure of PRDM9 binding. I modified the ChIP-seq protocol used in our HEK293T experiments (Chapter 2) to allow the use of testis tissue as input (see Methods). Complementary data were generated by Gang Zhang, who performed ChIP-seq experiments against DMC1, a meiosis-specific marker of DSBs, in the same mice. Edouard Hatton analysed the DMC1 ChIP-seq data and called DSB hotspots in each mouse. The DMC1 mark occurs several steps downstream of PRDM9 binding and H3K4me3 deposition. Its measured level depends on both the rate of DSB formation and repair, making comparison with the H3K4me3 signal natural.

To build a map of PRDM9 binding from our H3K4me3 ChIP-seq data, I utilised the peak calling algorithm defined in Chapter 2 to compute H3K4me3 enrichment and p-values in 100-bp non-overlapping bins throughout the genome for each mouse. I identified strongly significant bins ($p<1x10^{-5}$) and merged any significant bins occurring within 500 bp of each other, to produce a set of "H3K4me3-enriched regions" for each mouse (summarised in **Table 3.1**).

To produce more refined peak centres, I identified the base in each enriched region with the greatest ChIP coverage and re-computed enrichment in a 1-kb bin centred on that base (see Methods). Our peak-calling algorithm also estimates the fraction of reads originating from enrichment signal, rather than background, in the

Mouse	Prop. H3K4me3 reads from signal rep1 rep2	H3K4me3 fragment number rep1 rep2	H3K4me3 peak number on autosomes	Autosomal DMC1 hotspot number after filtering	Prop. hotspots with H3K4me3 (p<0.05)	Enrichment correlation H3K4me3 vs DMC1	FractionB6 correlation H3K4me3 vs DMC1
(PWDxB6)F1$^{PWD/H}$ Rescue	0.691 0.689	48,354,443 53,373,304	55,073	13,611	0.872	0.598	0.908
(PWDxB6)F1$^{PWD/B6}$ Infertile	0.690 0.615	39,428,180 43,496,423	51,317	19,703	0.811	0.699	0.933
(B6xPWD)F1$^{B6/PWD}$ Reciprocal	0.417 0.357	56,841,041 53,014,525	27,967	16,184	0.458	0.516	0.920
B6$^{B6/B6}$	0.711 n/a	77,837,110	46,950	15,868	0.645	0.566	n/a
B6$^{H/H}$	0.633 n/a	81,362,082	47,710	14,138	0.888	0.629	n/a

Table 3.1: Summary of H3K4me3 ChIP-seq datasets generated. H3K4me3 ChIP-seq was performed in adult male testes from five different mice. Mice are listed in the first column with *Prdm9* genotypes in superscript. Hybrid mice are listed with the maternal strain first. The second column lists estimates of the fraction of sequenced fragments originating from signal as opposed to background (as output by the peak calling algorithm detailed in Chapter 2). The fourth column lists the total number of "H3K4me3-enriched regions" in each mouse, after removing likely PRDM9-independent peaks (see Methods). The fifth column lists the number of DSB hotspots (also called DMC1 peaks) after subtracting those overlapping likely Prdm9-independent H3K4me3 regions. The sixth column lists the proportion of these DSB hotspots with evidence of H3K4me3 enrichment ($p<0.05$ from force-calling in a 1-kb window around each hotspot centre—see Methods). The seventh column shows the raw correlation values between H3K4me3 enrichment and DMC1 heat at these hotspots. Finally, the last column reports the correlation between the fraction of DMC1 heat and H3K4me3 enrichment corresponding to the B6 chromosome in hybrid mice (at hotspots where both are defined—see Methods).

resulting datasets (see **Table 3.1**): estimates ranged up to 71% across different mice, although one mouse – the Reciprocal mouse – showed noticeably higher background, with a value of only 36% for this fraction. To identify H3K4me3-enriched regions likely to occur independently of PRDM9 (for example, at promoters), I identified those H3K4me3-enriched regions that overlap between mice which do not share any PRDM9 allele (see Methods), and then I filtered out any regions overlapping this set for downstream analyses. This approach is conservative, as it assumes there should be no real binding site overlap between any pair of different PRDM9 alleles studied here. However, only 2.6% of DMC1 peaks overlap between the Humanized and B6 alleles in homozygous B6 mice, and so we might expect to lose only a small fraction of truly PRDM9-dependent H3K4me3 peaks using this stringent filter. Notably though, this filter makes us unable to investigate those regions with pre-existing

H3K4me3, which represents a limitation of using the H3K4me3 mark as a measure of PRDM9 activity *in vivo* and underlines the complementary value of having direct PRDM9 binding maps like those produced in Chapter 2.

To examine H3K4me3 enrichment at DSB hotspots, I "force-called" H3K4me3 enrichment values in 1-kb windows centred on each DMC1 peak centre (after filtering out those DMC1 peaks overlapping PRDM9-independent H3K4me3 peaks). I tested a range of window sizes from 100 bp to 3 kb, and selected the 1-kb window size because this yielded the highest correlation between H3K4me3 enrichment and reported DMC1 heats genome-wide (*e.g.* r = 0.7 in the Infertile mouse; **Table 3.1**). This force-calling approach is superior to comparing sets of *de novo* peak calls, as it allows one to estimate H3K4me3 enrichment and a p-value quantifying evidence of PRDM9 binding, for each and every DSB hotspot in a uniform way. **Table 3.1** summarises these values across all five mice and shows that in mice with high proportions of reads from signal, we find evidence for H3K4me3 enrichment at up to 89% of DSB hotspots ($p<0.05$), demonstrating good power in our H3K4me3 data and confirming the dominant role played by PRDM9 in positioning DSB hotspots across mice.

I adapted our peak calling framework to compute the expected fraction of H3K4me3 signal originating from the B6 chromosome for each peak (after correcting for background; see Methods) and compared this with a similar metric computed by Edouard Hatton using DMC1 reads, allowing for joint comparison of PRDM9 binding and DSB formation on an individual homologue for the first time, a key tool in this study that enabled the discoveries presented below. This "B6 fraction" estimate can serve as a measure of how symmetrically PRDM9 binds to the two homologues. For simplicity, in most calculations we use a somewhat arbitrary and broad range of 0.2-0.8 to define "symmetric" hotspots, with hotspots outside this range tending to be bound and broken only one homologue or the other, and thus "asymmetric". Comparing the H3K4me3 and DMC1 B6 fraction estimates across all DSB hotspots, I observed highly significant correlations ($p<10^{-16}$ by a rank-sum test): over 0.9 for each mouse (**Table 3.1** and **Figure 3.2b**), showing for the first

time that at the overwhelming majority of hotspots, strong DMC1 asymmetry results directly from asymmetry in PRDM9 binding and marking.

3.2.3 Comparisons of hybrid mice

Next, I compared properties of autosomal H3K4me3 and DMC1 hotspot asymmetry across mice. One of the most striking differences between the Infertile and Humanized Rescue mice is the proportion of autosomal DSB hotspots that are asymmetric. In the Infertile mouse, 57% of the total H3K4me3 signal at DSB hotspots occurs at asymmetric sites (here, defined as DSB hotspots where >80% of H3K4me3 signal originates from only the B6 or PWD chromosome), compared to only 38% for the Humanized Rescue mouse. **Figure 3.2** illustrates the full spectrum of binding symmetry in both mice, separating H3K4me3 signals by the *Prdm9* allele inferred to be controlling each hotspot.

In the Infertile mouse, both the B6 and PWD *Prdm9* alleles bind primarily in an asymmetric manner, with the B6 allele binding the PWD chromosome preferentially, and the PWD allele binding the B6 chromosome preferentially. In the Humanized Rescue mouse, the PWD allele continues to preferentially bind the B6 chromosome, but the Humanized allele binds much more symmetrically on both chromosomes, changing the overall symmetry profile in this mouse. We reasoned that this asymmetric binding pattern in the Infertile hybrid might be explained by differential motif erosion in each mouse subspecies, as has been recently described for humans versus chimpanzees [68], and more recently for (B6×CAST)F1$^{\mathrm{B6/CAST}}$ hybrids [67].

Edouard Hatton confirmed this, showing that sequence differences directly disrupting PRDM9 binding motifs explain almost all asymmetric DSB hotspots (83.4% of PWD hotspots; 91.3% of B6 hotspots), and result from rapid mutational accumulation along the separate lineages from the common ancestor of B6 and PWD [108]. Motifs bound by the Humanized allele, on the other hand, have not experienced hotspot erosion in either mouse lineage, and any asymmetric binding sites for this allele arise solely from coincidental mutations in human binding

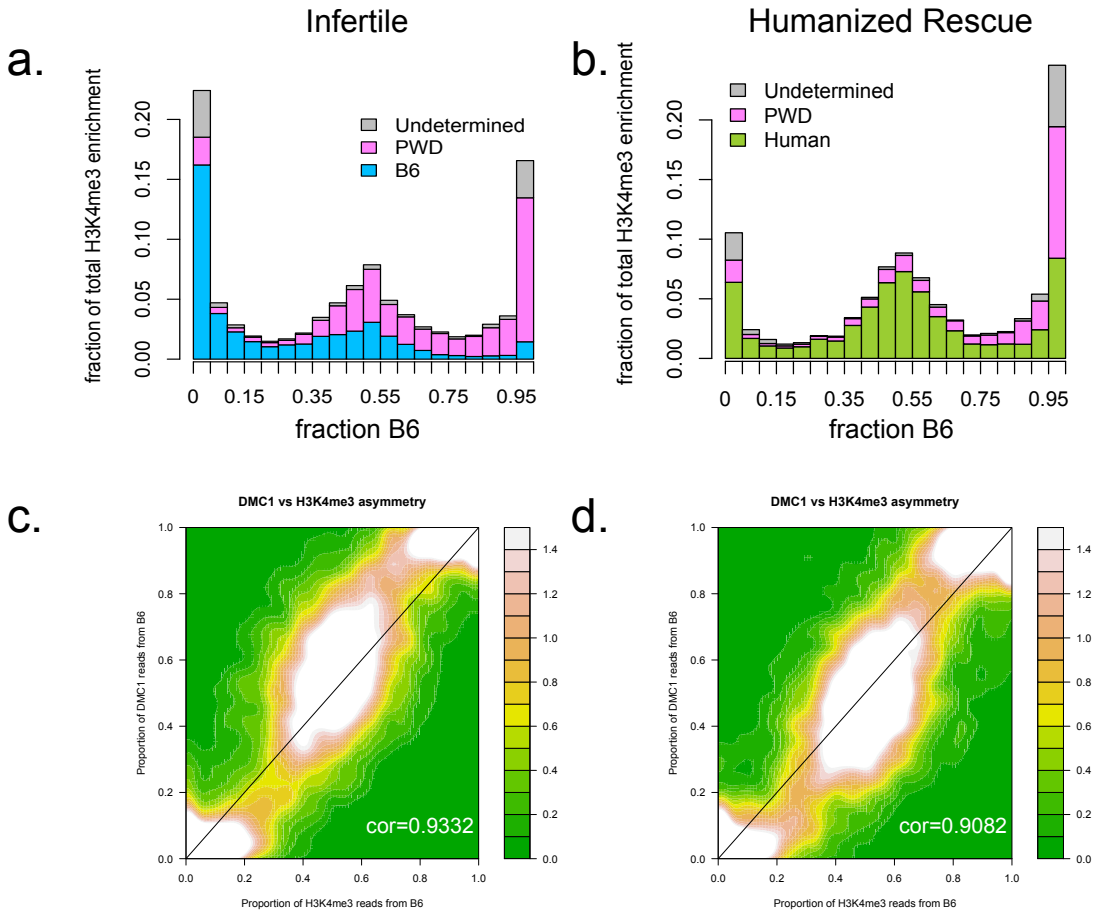


Figure 3.2: Comparisons of H3K4me3 symmetry in hybrid mice. The fraction of H3K4me3 enrichment arising from the B6 homologue ("fractionB6") was computed for each DSB hotspot (after filtering out those overlapping PRDM9-independent H3K4me3 peaks—see Methods). **a**,**b**: Histograms display the fraction of total H3K4me3 enrichment (genome-wide at hotspots) within bins of fractionB6, coloured by the *Prdm9* allele controlling each hotspot. Higher "asymmetry" (defined as fractionB6<0.2 or >0.8) is observed for the B6 and PWD alleles, with each tending to bind the homologue of the other strain. The Humanized allele in the Humanized Rescue mouse, by contrast, tends to bind much more symmetrically (**b**). **c**,**d**: Heat maps of fractionB6 estimates from DMC1 ChIP-seq data (y-axis) compared to H3K4me3 ChIP-seq data (x-axis) at the same DSB hotspots, illustrating high correlations (printed on each plot). S-shaped curves are consistent with the fact that DMC1 signal is elevated at asymmetric binding sites.

motifs between the two mouse lineages. Indeed, 73% of H3K4me3 signal from the Humanized allele in the Humanized Rescue mouse is attributable to symmetric hotspots (those with B6 fractions between 0.2 and 0.8), compared to only 40% for the B6 allele in the Infertile mouse. When comparing the symmetry spectra for DMC1 and H3K4me3 signal, we observed that the DMC1 spectrum concentrates even more heat toward asymmetric hotspots than does H3K4me3. For example, in the Infertile mouse 70% of the total DMC1 signal at DSB hotspots occurs at asymmetric sites, compared with only 57% of the total H3K4me3 signal. We expected that asymmetric sites would be hotter on average than symmetric sites, because they have experienced the strongest differential motif erosion in each mouse subspecies, and thus we can infer that they were among the hottest binding sites for each motif in the ancestral mouse genome. In the F1 hybrid, these strong, heavily eroded ancestral binding motifs are essentially resurrected on the chromosomes of the opposing strain, and this explains the prominence of strong, asymmetric binding sites. Indeed, the H3K4me3 enrichment distribution is significantly right-shifted for asymmetric hotspots relative to symmetric hotspots, likely reflecting this difference in PRDM9 affinity (**Figure 3.3**).

However, we had no reason to believe that this effect should be larger for DMC1 than for H3K4me3 if it were caused by PRDM9 binding affinity alone. We performed statistical comparisons of different groups of hotspots, to test the hypothesis that something downstream of PRDM9 binding and H3K4 trimethylation must be causing this additional excess of DMC1 heat at asymmetric hotspots. To examine whether this global DMC1 excess could also be detected at individual hotspots, I compared the ratio of DMC1 to H3K4me3 across hotspots controlled by the same allele, and I computed the heat attributable to only one chromosome or the other (B6 or PWD). **Figure 3.3** illustrates the results for the B6 allele acting on the PWD chromosome in the Infertile mouse, demonstrating that although DMC1 heat increases (on average) approximately linearly with H3K4me3, asymmetric sites show elevated mean DMC1 heat for a given level of H3K4me3 enrichment, relative to symmetric sites. Across all deciles of H3K4me3 strength, the difference is

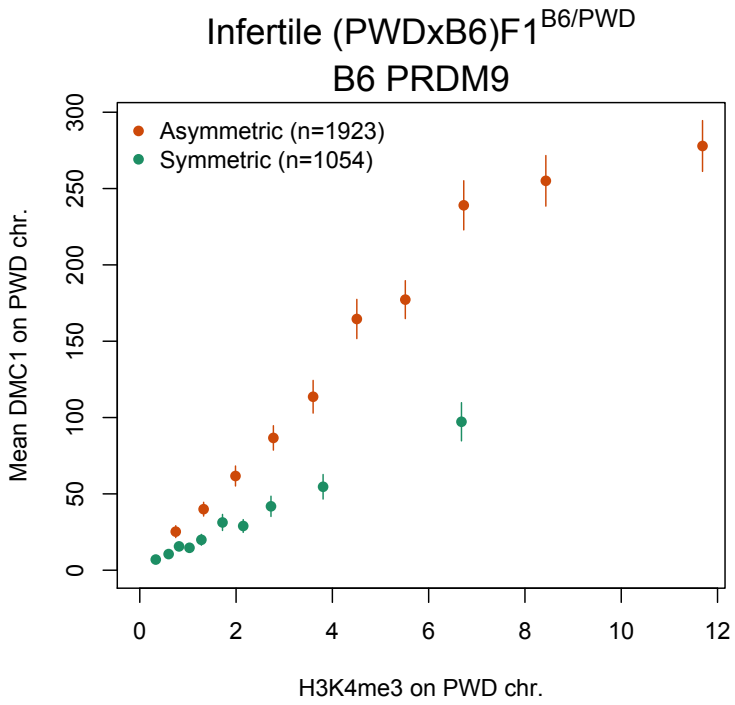


Figure 3.3: Comparisons of mean DMC1 heat and H3K4me3 signal in symmetric and asymmetric hotspots. B6-controlled hotspots in the Infertile mouse were split into subsets of asymmetric hotspots (fraction of H3K4me3 signal from B6 <0.1 or >0.9, orange), or symmetric hotspots (>0.4 and <0.6, green). Mean DMC1 values across hotspots in each decile bin of H3K4me3 enrichment (x-axis) are plotted with error bars representing ±2 s.e.m (x coordinates represent the median H3K4me3 enrichment in each decile bin). Asymmetric hotspots show both greater H3K4me3 enrichment on average, owing to greater binding strength, and greater mean DMC1 enrichment for any given level of H3K4me3 enrichment.

consistent, and far greater than underlying uncertainty due to variability in DMC1 signal across hotspots. Although we do observe an excess of H3K4me3 signal at asymmetric sites relative to symmetric sites (**Figure 3.3**), likely due to greater B6 PRDM9 binding affinity at asymmetric sites for reasons discussed above, the excess of DMC1 signal is *even larger still*. Globally, the ratio of mean DMC1 heat to mean H3K4me3 enrichment at asymmetric hotspots is 2.03-times higher (95% bootstrap C.I. 1.97-2.09) than at symmetric sites in the Infertile mouse. Critically, a similar effect occurs in every combination of mouse, PRDM9 alleles, and backgrounds tested, with estimated fold increases ranging from 1.7 to 2.4 (see **Figure 3.4**), and overwhelming evidence of an increase >1 in every case.

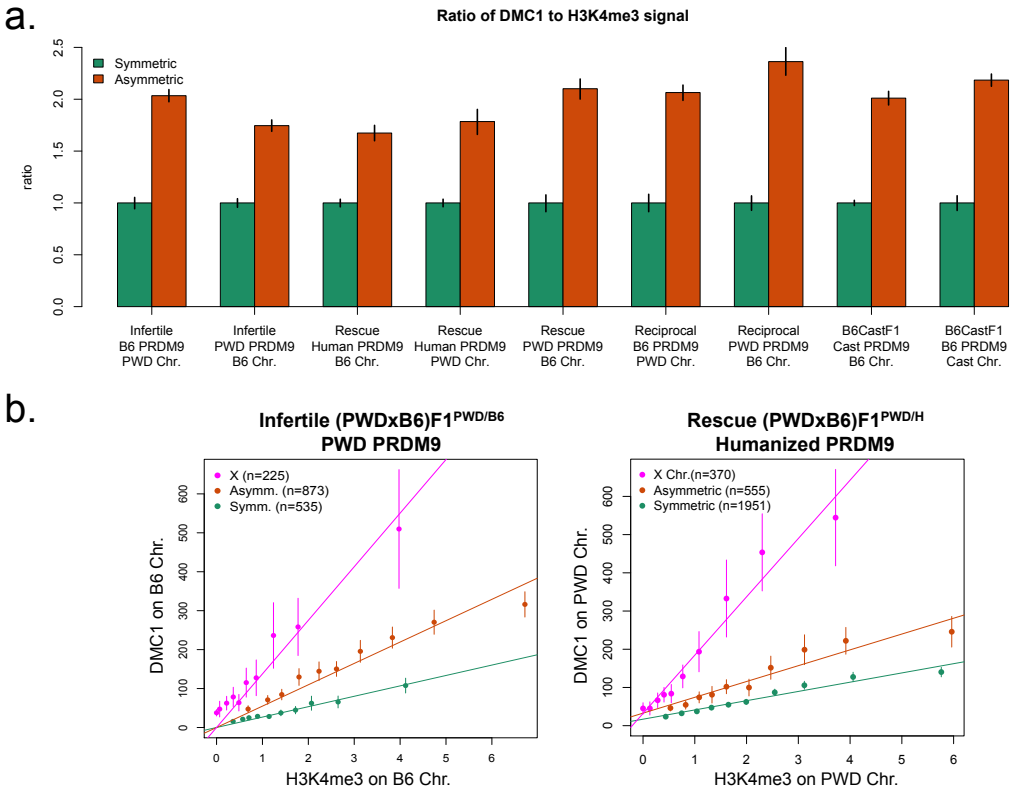


Figure 3.4: DMC1:H3K4me3 ratios across samples and on ChrX. a: Ratios of total DMC1 heat to total H3K4me3 heat at hotspots (filtered not to overlap PRDM9-independent H3K4me3 peaks) across 9 combinations of mouse, PRDM9 allele, and homologue background, split into subsets of asymmetric hotspots (fraction of H3K4me3 signal from B6 <0.1 or >0.9, orange), or symmetric hotspots (>0.4 and <0.6, green). Error bars represent 95% bootstrap C.I.'s. **b**: Decile mean plots as in **Figure 3.3**, but with hotspots from ChrX added for comparison, showing even greater elevation of DMC1 heat.

Importantly, this trend holds true for the Humanized allele as well (which does have a minority of sites that are coincidentally asymmetric due to overlapping pre-existing SNPs between PWD and B6). This shows that this signal cannot be a strange artefact of drive, whereby hotspots with naturally high DMC1 (relative to H3K4me3) are simply those that have eroded the most, and so get a higher ratio. Drive has not impacted the human allele, so this must be a general property of asymmetry, however it has arisen. I next examined whether the same DMC1 excess appears at hotspots on the non-PseudoAutosomal Region (PAR) of the X chromosome, which represents an extreme case of hotspot asymmetry as there

is no homologous chromosome for PRDM9 to bind in male meiosis. Indeed, the X chromosome shows even larger absolute DMC1 heats and even higher DMC1:H3K4me3 ratios than asymmetric autosomal hotspots (see **Figure 3.4b**), although its smaller number of hotspots relative to the total across all autosomes increases the width of the corresponding confidence intervals.

The strength of DMC1 signal on the X chromosome likely owes to the fact that DSBs on the non-PAR portion of the X persist and repair later in meiosis [126], owing to the lack of a homologue from which to repair and a suppressed ability to repair from the sister chromatid during early meiosis (reviewed in [15]). This led us to hypothesise that the DMC1 excess at asymmetric autosomal hotspots might also imply DSB persistence and late repair in meiosis, although it might at this point also be explained by a greater number of breaks occurring at asymmetric hotspots. To rule out certain artefactual explanations for this DMC1 excess at asymmetric autosomal hotspots, we leveraged the fact that hotspots shared between mice typically have highly correlated DMC1 heats (r=0.95 between the Infertile and Reciprocal mice, for example). Thus, by comparing DMC1 heat at an individual hotspot that is symmetric in one mouse and asymmetric in another, we can independently test whether hotspot asymmetry yields larger DMC1 heats. Specifically, I compared the DMC1 heat caused by the Humanized allele on the B6 chromosome in the Humanized Rescue mouse to the heat of the same hotspot – and on the same B6 local genetic background - in the $B6^{H/H}$ mouse. Since the $B6^{H/H}$ mouse is homozygous across the genome, all of its hotspots are necessarily symmetric. Furthermore, the Humanized allele has not eroded in either mouse subspecies, and thus any asymmetric sites found in the hybrid arise purely from coincidental mutations in the Humanized binding motif. This allows us to control for any confounders specific to either mouse allele or their evolutionary history in the mouse genome. As illustrated in **Figure 3.5a**, I found a similar, striking excess of DMC1 heat in the hybrid at asymmetric hotspots relative to symmetric hotspots. In effect, this shows that by transferring a hotspot from a context where

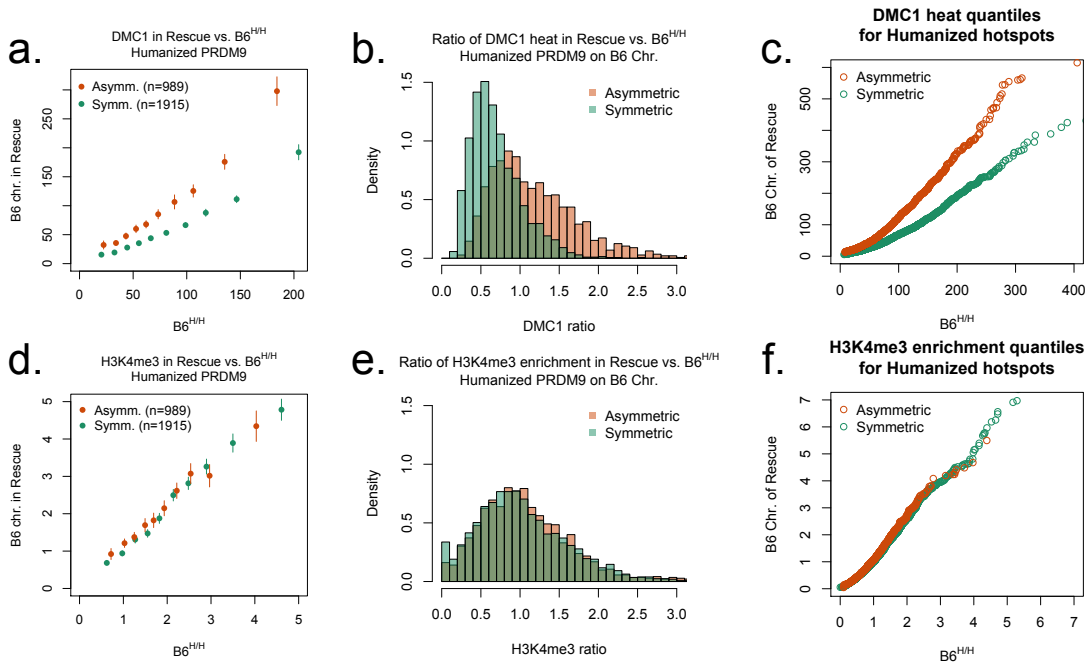


Figure 3.5: Comparisons between mice confirm DMC1 heat elevation at asymmetric hotspots. **a**: Decile mean plot as in **Figure 3.3**, but comparing DMC1 heats at shared hotspots (controlled by the Humanized allele) between the $B6^{H/H}$ mouse (x-axis) and the Humanized Rescue mouse. **b**: For the same hotspots in **a**, a histogram showing a full distributional shift in DMC1 ratio across all hotspots. **c**: For the same hotspots in **a**, a q-q plot showing consistent DMC1 elevation across all percentile bins of DMC1 enrichment, making outlier effects unlikely. **d**,**e**,**f**: As in **a**,**b**,**c**, but for H3K4me3 enrichment normalised to genome-wide total H3K4me3 in each mouse, showing no significant difference between asymmetric and symmetric hotspots.

it is symmetric to a context where it is asymmetric, its relative chromosome-specific DMC1 heat increases significantly.

To rule out the possibility that the signal of excess DMC1 signal at asymmetric hotspots is driven by outliers, I further examined this effect across the entire distribution of DMC1 heats. **Figure 3.5b** illustrates a complete distributional shift of DMC1 heat at asymmetric sites in the Humanized Rescue relative to symmetric sites, when normalised to heat in the $B6^{H/H}$ mouse, although notably with considerable spread, which we explore further below. A quantile-quantile plot comparing the two mice confirms that the excess of DMC1 heat is found consistently across the full range of hotspot heats (**Figure 3.5c**). Importantly, by performing an identical set of analyses comparing H3K4me3 enrichment at

shared hotspots between mice, I confirmed that there is no significant difference between asymmetric and symmetric sites for H3K4me3 between mice (see **Figure 3.5d-f**). That is, transferring a hotspot from a context where it is symmetric to a context where it is asymmetric has no appreciable effect on its probability of being trimethylated, which is predicated on PRDM9 binding. Thus, PRDM9's binding affinity and trimethylation activity seem, at least on average, unaffected by properties of PRDM9 binding to the homologue, whereas DMC1 signal is consistently elevated when the homologue fails to be bound by PRDM9.

3.2.4 Excess heat is not explained by additional factors such as heterozygosity

After gathering these initial data, I was concerned that the signal of excess DSB heat at asymmetric sites might owe to an important potential confounder: heterozygosity. Indeed, asymmetric binding sites contain 55% more SNPs between B6 and PWD in the 120 bases surrounding a DSB hotspot centre (roughly the "nucleosome depleted region" where most if not all DSB breakpoints are likely to occur [95]). These SNPs often occur in the binding motif and explain why PRDM9 fails to bind one homologue, making the site asymmetric. For example, then, the excess of DMC1 heat might in principle be due to the fact that asymmetric sites have more SNPs, and the cell might take longer to repair DSBs at these sites by copying from a mismatching homologue. If so, regions with more SNPs would show an elevation of DMC1 signal regardless of PRDM9 binding asymmetry. To test whether local SNP number can affect the DMC1/H3K4me3 ratio independently of PRDM9 binding, I selected sites trimethylated symmetrically by the Humanized allele in the Humanized Rescue mouse. I then separated these DSB hotspots into two groups: those with 0 SNPs in the central 120 bp (46% of hotspots) and those with $\geq$2 SNPs (25%). I then compared these two groups and found no systematic difference between the two (see **Figure 3.6**), with heavily overlapping CI's for mean DMC1, across H3K4me3 enrichment bins. The same held true with a larger threshold of 3 SNPs, or after repeating a similar analysis using the 1-kb region surrounding

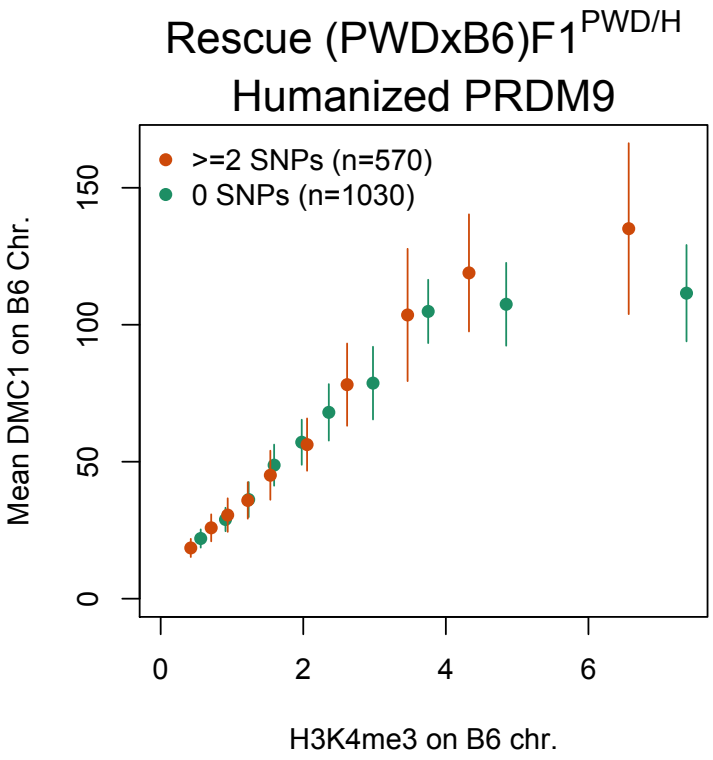


Figure 3.6: Local heterozygosity does not predict DMC1 heat. Decile mean plot as in **Figure 3.3**, but for symmetric hotspots controlled by the Humanized allele in the Humanized Rescue mouse, split into subsets of hotspots with ≥2 SNPs in the central 120-bp region of each hotspot (top quartile, orange), or 0 SNPs in this region (green). No effect of local heterozygosity is evident, suggesting that the elevation of DMC1 heat at asymmetric hotspots **Figure 3.3** cannot be explained by their increased local heterozygosity.

each DSB hotspot centre (not shown). Thus, conditional on PRDM9 binding being symmetric, the number of local SNPs appears to have no effect on DMC1 heat. Edouard Hatton found a similar lack of effect of SNPs or even indels found *within* a PRDM9 binding motif (conditional on symmetric binding, implying these SNPs do not affect PRDM9 binding affinity [108]). Edouard Hatton further showed that other factors such as GC content or DMC1 heat could not explain this effect [108]. These analyses corroborate the finding that PRDM9 binding to the homologue has downstream effects on DSB processing, as measured by DMC1 heats.

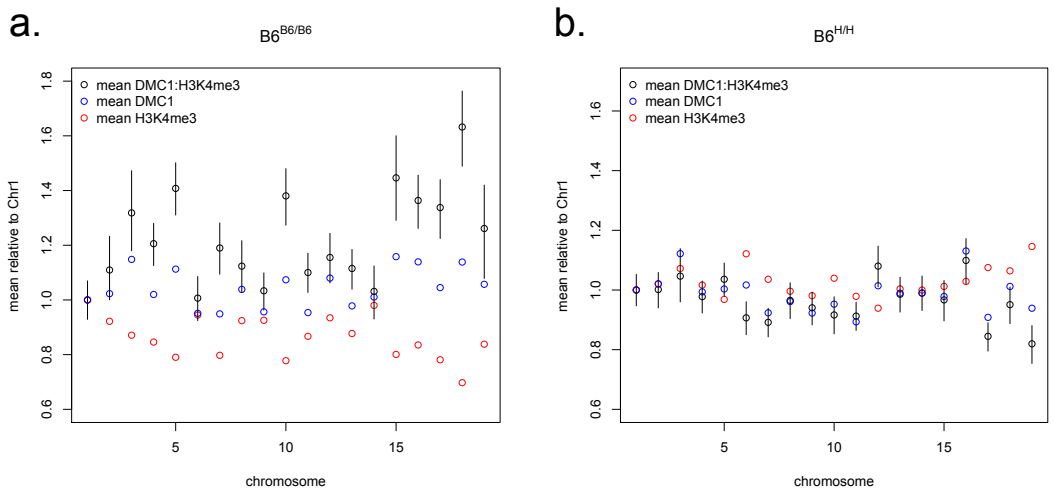


Figure 3.7: Chromosome-wide DMC1:H3K4me3 ratios in homozygous mice. **a**: Mean DMC1 heat (blue), mean H3K4me3 enrichment (red), and the ratio of these two values (black, with 95% bootstrap C.I.'s) are plotted for all hotspots on each autosome in the wild-type B6 mouse, normalised to the values on Chr1. Hotspots overlapping PRDM9-independent H3K4me3 peaks (and hotspots without defined DMC1 and H3K4me3 heats due to low coverage in all lanes) were removed. **b**: As in **a**, but for the $B6^{H/H}$ mouse, showing little variability among chromosomes.

3.2.5 Comparisons of individual chromosomes

Given this strong global signal of elevated DMC1 relative to H3K4me3 at individual asymmetric hotspots, we next asked whether the level of DMC1 relative to H3K4me3 varies at much broader (chromosomal) scales, across the genome. In the simplest case, the B6 wild-type mouse, I computed the ratio of summed DMC1 heat to summed H3K4me3 enrichment across all DSB hotspots passing certain filters (not overlapping PRDM9-independendent H3K4me3 peaks, having defined DMC1 heats and H3K4me3 enrichment values) on each chromosome, and I normalised this ratio to Chromosome 1 (and provided 95% bootstrap confidence intervals). Remarkably, 15 out of 19 chromosomes showed significant differences relative to Chromosome 1 ($p<0.05$), and no single ratio lay within the 95% C.I. for >7 chromosomes, illustrating the high variability in this measure among chromosomes (**Figure 3.7a**).

Partially explaining this signal, shorter chromosomes tend to have higher DMC1:H3K4me3 ratios (r=-0.46 with length; p=0.047, Pearson's test), implying that individual PRDM9 binding sites are either more likely to have a DSB event

on small chromosomes, or their DSBs repair more slowly. Smaller chromosomes also show slightly lower H3K4me3 enrichments (p=0.035, Pearson's test). This H3K4me3 depletion effect could be caused by an ascertainment bias: hotter DMC1 peaks are more likely to be called as hotspots regardless of H3K4me3 enrichment, and as a result of higher DMC1:H3K4me3 ratios on short chromosomes, hotspots with low H3K4me3 enrichment are more likely to be detected on these chromosomes. Given that each chromosome must experience an obligate crossover, it is plausible that the cell does so by enforcing a minimum number of DSBs per chromosome regardless of length, thus raising the probability of a break given PRDM9 binding on smaller chromosomes.

However, there may be another partial explanation for this chromosome effect: if shorter chromosomes have a higher probability of recombination at each PRDM9 binding site, then motif erosion will have had a stronger effect than on longer chromosomes. This would result in weaker mean H3K4me3 signal on shorter chromosomes and would raise the DMC1:H3K4me3 ratio even if there were no strong absolute DMC1 elevation. One way to both examine the effects of erosion, and test for the influence of factors such as imposed constraints on the number of DSBs chromosome-wide, is to compare the wild-type B6 mouse to the B6 mouse homozygous for Humanized *Prdm9*. Strikingly, in the Humanized mouse the chromosome effects largely disappear, with few large differences in mean DMC1 or H3K4me3 levels between chromosomes (**Figure 3.7b**). This may represent the distribution of PRDM9 binding and DMC1 signal prior to any accumulated motif erosion. Speculatively, it may be the case that PRDM9 binding sites weakened by erosion may be more prone to asymmetric binding within a single spermatocyte even in a fully homozygous mouse, and thus the elevated DMC1:H3K4me3 ratio at smaller chromosomes in the $B6^{B6/B6}$ mouse may imply a greater amount of asymmetric binding on these chromosomes within individual spermatocytes.

I next examined PWD-allele controlled hotspots in the Infertile mouse in a similar manner, but by summing haplotype-specific DMC1 and H3K4me3 signals on the B6 chromosome (after additionally filtering out peaks without defined B6

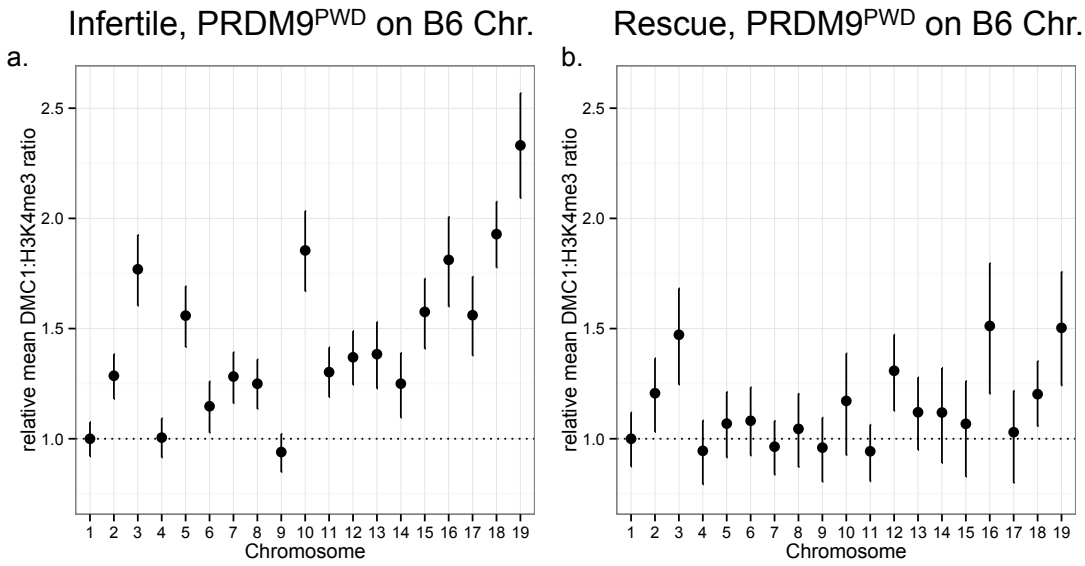


Figure 3.8: Chromosome-wide DMC1:H3K4me3 ratios in hybrid mice. a: For each chromosome, the ratio of mean DMC1 heat to mean H3K4me3 enrichment on the B6 homologue across all DSB hotspots controlled by the PWD allele in the Infertile mouse (with 95% bootstrap C.I.'s). Hotspots overlapping PRDM9-independent H3K4me3 peaks, hotspots without defined DMC1 and H3K4me3 heats (due to low coverage in all lanes), and hotspots without defined fractionB6 values for both DMC1 and H3K4me3 were removed. **b**: As in **a**, but for the Humanized Rescue mouse.

enrichment ratios for both DMC1 and H3K4me3). Again I observed chromosome effects – and these are even more pronounced (>2-fold for Chromosome 19 relative to Chromosome 1) than in the Humanized Rescue or in the B6 wild-type mice (**Figure 3.8a**). Because the PWD allele is also present in the (fertile) Humanized Rescue mouse, which also has the identical PWD/B6 genetic background (except at PRDM9), we can compare behaviour at the same PWD hotspots in the Humanized Rescue mouse. As in the $B6^{B6/B6}$ mouse, the Humanized Rescue mouse shows fewer, and weaker, chromosome effects (**Figure 3.8b**).

The H3K4me3 signal, in contrast, shows no significant chromosomal differences between mice (**Figure 3.9**), and all 19 chromosome effect confidence intervals overlap 1.0. Recalling that the Infertile and Humanized Rescue mice are identical except for the ZF array of their PRDM9 allele inherited from the B6 father, this implies that one PRDM9 allele can exert *trans* effects on the DMC1 signal at hotspots controlled by the other allele, while the H3K4me3 signal remains unchanged at

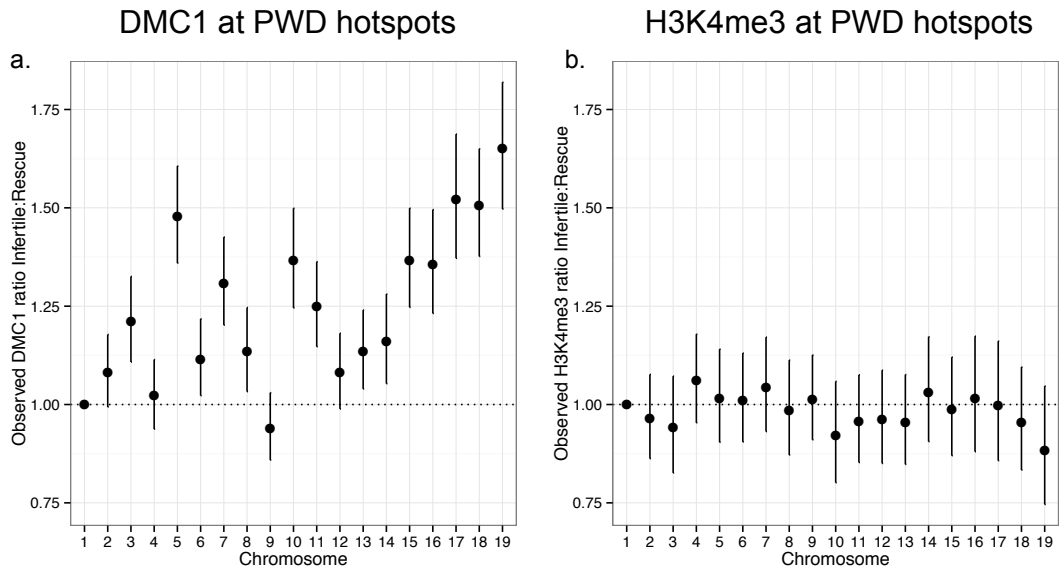


Figure 3.9: Ratios of DMC1 and H3K4me3 at shared hotspots between mice. **a**: For each chromosome, the mean DMC1 heat on the B6 homologue across all DSB hotspots controlled by the PWD allele was computed for the Infertile mouse and for the Humanized Rescue mouse, and the ratio of these two means is plotted (with 95% bootstrap C.I.'s, using identical filters to **Figure 3.8**). **b**: As in **a**, but for H3K4me3 enrichment values normalised to the genome-wide total enrichment in each mouse.

the chromosome-wide level. Similar *trans* effects were observed in comparisons of the B6 PRDM9-controlled hotspots in $B6^{B6/H}$ and $B6^{B6/B6}$ mice (see [108]), ruling out the possibility that these effects are merely due to heterozygosity or due to downstream effects of infertility, such as meiotic arrest.

Because they might reflect delayed DSB repair on certain chromosomes, I hypothesised that these identified *trans* effects might be associated with chromosomal differences in asynapsis rates, which a previous study had shown to be highly variable among five chromosomes in the Infertile mouse [82]. I compared the reported asynapsis rates for these five chromosomes with the chromosome-specific DMC1 heat effects described above and found an identical ranking (p=0.017 by rank correlation permutation test). That is, chromosomes with a greater excess of DMC1 heat in the Infertile mouse also have higher asynapsis rates. Given that DMC1 heat is highly predicted by H3K4me3 symmetry on each chromosome, this implies that chromosomes with lower PRDM9 binding symmetry are at greater risk of asynapsis.

Genome-wide, if PRDM9 binding symmetry is required for proper synapsis, this could explain why mice with lower overall hotspot symmetry have reduced fertility.

Leading naturally on from these observations at both individual hotspots and chromosome-wide, work by Simon Myers tested possible predictors of the magnitudes of the *trans* effects identified here, at individual chromosomes in the Infertile and Humanized Rescue mice (see [108]). After fitting many possible models with predictors including chromosome length, DMC1/H3K4me3 signals, and hotspot symmetry, the best-fitting model was highly predictive (r^2=0.88) and included only *symmetric* hotspot measures – the total H3K4me3 signal from PRDM9 binding on both homologues (*i.e.* symmetrically) at the same hotspots, summed over the entire chromosome – for each of the three *Prdm9* alleles ($p<0.002$ in each case). That is, chromosomes with less hotspot symmetry overall show greater DMC1 elevation. Taken together then, our data imply a possible relationship between PRDM9 binding site erosion, PRDM9 binding (a)symmetry, delayed DSB repair and resulting elevation of DMC1 signals, asynapsis, and infertility, which I explore further in the Discussion below.

3.3 Discussion

Even after decades of research into the mechanism of mouse hybrid infertility, its precise molecular aetiology has remained an enigma. Previous studies have shown that multiple manipulations can rescue hybrid fertility to varying degrees: introducing extra copies of B6 PRDM9, or deleting B6 PRDM9, or adding in a different allele of PRDM9, can all restore at least some fertility [75, 84]. Previous experiments also seemed to imply a direct, necessary interaction between PRDM9 and the *Hstx2* locus on ChrX to yield full infertility [83, 85], as well as a requirement that individual chromosomes be heterosubspecific [82]. Without data at a much finer resolution, it was difficult even to propose a hypothesis that could explain all of these disparate results.

Here we have rescued hybrid fertility with one of the smallest possible manipulations, replacing only the DNA-binding ZF array of the B6 *Prdm9* allele

with the ZF array from the human B allele. Given that the human allele did not co-evolve with the genome of either mouse subspecies but is still capable of rescuing fertility, it is unlikely that the mechanism underlying hybrid infertility involves Dobzhansky-Muller incompatibilities between PRDM9 and other genomic loci. Furthermore, we showed that humanization rescues complete fertility from semi-fertility in reciprocal (B6×PWD)F1 males, suggesting that PRDM9 acts at least somewhat independently of the *Hstx2* locus.

To investigate this rescue mechanism further, we performed ChIP-seq against H3K4me3 and DMC1 to measure fine-scale differences between the Humanized Rescue mice and their Infertile counterparts. By computing the amount of DMC1 and H3K4me3 on each homologue at each hotspot, we came to the important realisation that PRDM9 binding asymmetry is strikingly greater for either the B6 or PWD alleles relative to the Humanized allele. We showed that this binding symmetry owes to the effects of accumulated motif erosion in each subspecies. Because the newly introduced Humanized allele has not caused motif erosion in either subspecies, it is analogous to a naturally occurring PRDM9 allele arising through mutation in a mouse population. Thus, it would seem that hybrid infertility is a transient phenomenon, capable of being reversed by mutations in the PRDM9 ZF array. It may be the case that this transient speciation event can, or will, be reinforced by other reproductive barriers to prevent the fitness cost of rearing infertile male offspring, leading to full speciation before a newly arising PRDM9 allele can rise to fixation.

Furthermore, we discovered a very surprising phenomenon at asymmetric hotspots: DMC1 signal is roughly twofold enriched relative to symmetric hotspots, while H3K4me3 signal remains the same. This DMC1 elevation could result either from a greater frequency of DSB formation at these hotspots or from a longer persistence of the DMC1 mark after break formation. We think the former is less plausible, given that the number of DSBs per cell is tightly controlled, and the cell would require a complex mechanism for monitoring binding symmetry prior to DSB formation. Rather, the more parsimonious explanation is that asymmetric sites experience a delay in repair and DMC1 removal after DSB formation (similar to sites

on the sex chromosomes in male meiosis, which repair late from the sister chromatid). This would suggest a new role for PRDM9 in recombination—not only guiding the initiation of DSB formation, but affecting the downstream processing of DSBs as well.

One potential explanation for such a delay is that symmetric PRDM9 binding and/or H3K4me3 deposition aids the efficiency of homology search by bringing homologous binding sites together before or after DSB formation, thereby reducing the homology search space by orders of magnitude. Indeed, this may explain why recombination hotspots evolved in the first place. Without a critical amount of symmetric binding, homology search is inefficient and, as a result, certain sensitive chromosomes fail to synapse properly, leading to meiotic arrest and apoptosis. Symmetric binding can in principle be restored by increasing the dosage of PRDM9, by inserting an allele with less motif erosion, or by removing one of the asymmetrically binding alleles (to a lesser degree). Indeed, this symmetry-dependent model would predict the fertility outcomes of all hybrid mice in this and previous studies [84] (reproduced in Chapter 1, **Table 1.1**; see [108] for a more formal treatment of this model). Recent unpublished data from Ben Davies have also revealed fertility rescue in a $B6^{H/H}$×STUS cross (in both directions), further supporting this model.

Speculatively, this mode of speciation may be somewhat generalised. The ZF-array of PRDM9, and its target binding sites, are among the fastest evolving regions of the mammalian genome. Given enough time, isolated populations will acquire distinct PRDM9 alleles and these will act to erode distinct binding sites in the genome at the population level. Upon subsequent hybridization, their F1 meiotic cells will be faced with asymmetric binding sites and might experience reductions in fertility, if not complete sterility. Even small reductions in fertility could drive reinforcement of genes that prevent hybridization, leading to a cascade of further reproductive incompatibilities.

3.4 Methods

3.4.1 Animal husbandry

Mice were housed in individually ventilated cages and received food and water *ad libitum*. All studies received local ethical review approval and were performed in accordance with UK Home Office Animals (Scientific Procedures) Act 1986. Experimental groups were determined by genotype and were therefore not randomized, with no animals excluded from the analysis.

3.4.2 ChIP-seq

ChIP-seq was performed as previously described (see Chapter 2) with several modifications to accommodate differences in processing testis tissue instead of cells (noted here). 8-week-old adult males were culled and testes were obtained immediately. The testis tunica was removed, and the tubules were disassociated with tweezers and fixed in 1% formaldehyde in PBS for 5 minutes followed by glycine quenching (125 mM final concentration) for 5 minutes at room temperature. Following washing steps, pellets were resuspended in 900 μl cold RIPA lysis buffer, dounced 20 times and sonicated in 300-μl aliquots in a Bioruptor Twin sonication bath at 4°C for three 10-minute periods of 30 s on, 30 s off at high power, then cell debris was pelleted and removed and aliquots were pooled. For each sample, 50 μl of equilibrated magnetic beads were resuspended in 100 μl PBS/BSA and added to the chromatin samples for pre-clearing for two hours at 4°C with rotation. Beads were removed, and 100 μl of pre-cleared chromatin was set aside for the input control. 5 μl rabbit polyclonal anti-H3K4me3 antibody (Abcam ab8580) was added to the remaining pre-cleared chromatin and incubated overnight at 4°C with rotation. 50 μl of beads were washed and resuspended as before, then incubated with the chromatin samples for two hours at 4°C with rotation. Beads were then washed and decrosslinked, and for input controls, 50 μl of pre-cleared chromatin was used. After descrosslinking, samples were further incubated with 80 μg RNAse

A at 37°C for 60 minutes and then with 80 μg Proteinase K at 55°C for 90 minutes. DNA was purified using a Qiagen MinElute reaction cleanup kit.

ChIP and total chromatin DNA samples were submitted to the Oxford Genomics Centre for library preparation, sequencing, and mapping. Samples were multiplexed and sequenced on an Illumina HiSeq2500 machine (rapid mode), yielding 51-bp, paired-end reads. I prepared two biological replicates plus one genomic input control each for the Infertile (PWD×B6)F1$^{\mathrm{PWD/B6}}$, Reciprocal (B6×PWD)F1$^{\mathrm{B6/PWD}}$, and Humanized Rescue (PWD×B6)F1$^{\mathrm{PWD/H}}$ mice, yielding roughly 40-50 million usable fragments (read pairs) per replicate. For the B6$^{\mathrm{B6/B6}}$ and B6$^{\mathrm{H/H}}$ mice, I prepared one biological replicate each (yielding 70-80 million usable read pairs per sample) and later split read pairs into pseudo-replicates. Sequencing reads were aligned to mm10 using BWA [119] (options -q 10 -t 8) followed by Stampy [120] (options -t 8 –bamkeepgoodreads), and reads not mapped in a proper pair or with insert sizes larger than 10 kb were removed. Read pairs representing likely PCR duplicates were also removed using samtools [119]. Pairs for which neither read had a mapping quality score greater than 0 were removed. Fragment coverage was computed at each position in the genome and in 100-bp non-overlapping bins using in-house code (Appendix 2) and the samtools and bedtools packages [119, 122].

3.4.3 Peak calling

Peak calling was performed using a maximum-likelihood-based peak calling algorithm that uses fragment coverage information from both sequencing replicates and the total chromatin control (specified in Chapter 2, code in Appendix 2). For each bin in the genome the algorithm estimates a ChIP enrichment value relative to local background, and it also provides genome-wide estimates of the proportion of reads originating from signal versus background, giving an estimate of the purity of each replicate. For *de novo* identification of enriched regions I merged adjacent 100-bp non-overlapping bins with $p<10^{-5}$, a threshold expected to produce roughly 250 false positives genome-wide, yielding a set of "enriched regions". Within each of these enriched regions, I identified the base with the maximum read coverage and

called this as the peak centre. Then, having fixed each peak centre I obtained a single enrichment value and p-value for the region, by performing likelihood ratio testing in a fixed 1-kb bin surrounding the peak centre.

To enable filtering of H3K4me3 peaks corresponding to promoters and other PRDM9-independent sources, I identified H3K4me3-enriched regions shared among any one of three possible pairs of mice with different *Prdm9* alleles, for which I obtained H3K4me3 data. I conservatively assumed that any two mice with different PRDM9 alleles should not have any overlapping PRDM9-dependent H3K4me3 peaks, and thus any overlaps are likely to be PRDM9-independent. To enhance our sensitivity to find these PRDM9-independent regions, I took the union of all pairwise intersections of enriched regions between mice with distinct *Prdm9* alleles, as follows:

$$[(\text{PWD} \times \text{B6})\text{F1}^{\text{PWD/H}} \cap \text{B6}^{\text{B6/B6}}]$$
$$\cup[(\text{PWD} \times \text{B6})\text{F1}^{\text{PWD/B6}} \cap \text{B6}^{\text{H/H}}]$$
$$\cup[(\text{B6} \times \text{PWD})\text{F1}^{\text{B6/PWD}} \cap \text{B6}^{\text{H/H}}]$$

For comparisons of H3K4me3 and DMC1 signals at DSB hotspots, I used the same approach to estimate H3K4me3 enrichment. However, testing was performed by "force-calling" in a fixed 1-kb bin, centred on the midpoint of each specified DSB hotspot (as described in Chapter 2). I then removed DSB hotspots overlapping any of the PRDM9-independent H3K4me3 sites defined above. When directly comparing H3K4me3 enrichment between different mice (as in **Figure 3.5**), I normalised H3K4me3 enrichment to the sum across all DSB hotspots being compared, because ChIP-seq data only allow *relative* comparisons among loci, not comparison of *absolute* values. For peak calling with published single-end H3K4me3 ChIP-seq data from a $(\text{B6}\times\text{CAST})\text{F1}^{\text{B6/CAST}}$ mouse [67], I performed all steps identically, but I computed read coverage instead of fragment coverage across the genome (due to the reads being single-end). In heterozygous mice, Edouard Hatton determined which PRDM9 allele controls each hotspot by comparing DSB hotspots in heterozygotes with those in homozygous mice containing each PRDM9 allele (fully specified in [108]).

3.4.4 Haplotype calling

H3K4me3 ChIP-seq reads overlapping the 1-kb region surrounding each DSB hotspot centre were compared with a list of biallelic SNPs distinguishing the PWD genome from the B6 genome (described in [108]). Reads matching one or more PWD alleles and no B6 alleles at these sites (with base quality 20) were assigned to the PWD haplotype, and *vice versa*. Reads not overlapping any SNPs and reads matching alleles from both haplotypes were excluded. I then subtracted the expected background coverage at each site using information from the input lane and from our peak-calling algorithm, as described below.

At a particular hotspot, I denote the number of B6-assigned read pairs from ChIP replicates 1 and 2 by d^{B6}_{rep1} and d^{B6}_{rep2}, respectively. Similarly, d^{B6}_{input} represents the number of B6-assigned read pairs from the genomic input lane, and d^{PWD}_{input} is the corresponding number for PWD. The peak calling algorithm provides estimates $\alpha 1$ and $\alpha 2$ of the ratio of background coverage of each ChIP replicate relative to the input lane. Then, under the assumptions of Chapter 2, the expected number of background reads for this hotspot is given by $0.5(\alpha_1 + \alpha_2)(d^{B6}_{input} + d^{PWD}_{input})$, further assuming that 50% of background reads come from the each of the homologous chromosomes. Then, the number of reads mapping to the B6 chromosome after adjustment for background is given by

$$d^{B6}_{rep1} + d^{B6}_{rep2} - 0.5(\alpha_1 + \alpha_2)(d^{B6}_{input} + d^{PWD}_{input}).$$

A similar adjustment is performed for the PWD chromosome; values below zero were zero-truncated, and then background-subtracted B6 coverage was divided by total background-subtracted B6 plus PWD coverage to estimate the proportion of H3K4me3 signal from the B6 chromosome, and assigned undetermined if there were fewer than 10 haplotype-informative reads per hotspot. Finally, this proportion was then multiplied by the total H3K4me3 enrichment estimate at each hotspot to yield a haplotype-specific enrichment estimate.

Teleology is like a mistress to a biologist: he cannot live without her but he's unwilling to be seen with her in public.

— J.B.S. Haldane, attributed by E. Mayr in *Studies in the Philosophy of Science* (1974)

4

PRDM9 forms homo-multimers, mediated by its zinc finger array

4.1 Introduction and experimental design

The only known intermolecular protein-protein interaction involving PRDM9 is between its PR/SET domain and the histone H3 tail, where PRDM9 adds a trimethyl group at Lysine 4 [38]. One structural study also showed that the Zinc Knuckle and Early Zinc Finger domains interact intramolecularly to autoinactivate the PR/SET domain [39]. However, the functions of the N-terminal SSXRD and KRAB domains remain entirely elusive, although they are hypothesised to take part in some sort of protein-protein interaction [54]. PRDM9's Zinc Finger (ZF) array has been regarded primarily as a DNA-binding domain with no other demonstrated functions, although studies of other zinc finger proteins have shown that ZF domains can participate in highly specific protein-protein interactions, including with each other [40]. Prior to beginning this work, it had also remained unknown whether PRDM9 exists as a free-floating single unit or as part of a larger complex of proteins within the cell. Such a complex could include PRDM9 dimers (or other multimers), but self-binding ability had not been previously examined for PRDM9.

Given the unusual sensitivity of the infertility phenotype in (PWDxB6)F1 hybrid mice to equal doses of two specific PRDM9 alleles (reviewed in Chapter 3), we

initially hypothesised that this dosage sensitivity could be explained if PRDM9 formed heterodimers that were somehow detrimental to fertility in a way that homodimers were not. After discovering the differences between symmetric and asymmetric PRDM9 binding sites in hybrids (described in Chapter 3), we then hypothesised that PRDM9 may mediate homologue pairing by binding symmetrically to homologous sites and then binding to itself, greatly reducing the space and time in which homology search occurs during DSB repair. We initially hypothesised that self-binding could be mediated by the Early ZF and Zinc Knuckle domains interacting inter-molecularly instead of intra-molecularly. Although highly speculative, these hypotheses nevertheless strongly suggested that we should test the ability of PRDM9 to bind to itself and form multimers. Conveniently, we were able to investigate these questions using an experimental approach within the same human cell line transient expression system that we used previously for ChIP-seq experiments (described in Chapter 2).

To examine whether PRDM9 can bind to itself under physiological conditions, I generated cDNA constructs of full-length human PRDM9 (the "B" allele, found in the hg19 reference sequence) tagged at the C-terminus with either HA or V5 epitope "tags" (see **Figure 4.1**).

These short tags can be bound with high specificity by different commercially available antibodies, allowing one to isolate protein complexes containing either tag by immunoprecipitation (IP). Then by the standard experimental technique of western blotting, one can visualise the size and relative quantities of the tagged proteins present in those immunoprecipitated complexes. That is, if PRDM9 does interact with itself in cells expressing both tagged constructs, one would expect that protein complexes immunoprecipitated using the HA antibody should yield a band in a western blot using the V5 antibody, and *vice versa*.

For both constructs, next to the HA/V5 tag I also included a TwinStrep tag (for additional affinity purification capabilities) and Yellow Fluorescent Protein (YFP) separated by a "self-cleaving" P2A sequence. During translation, the YFP tag does not become attached to the PRDM9 protein, leaving only PRDM9 with

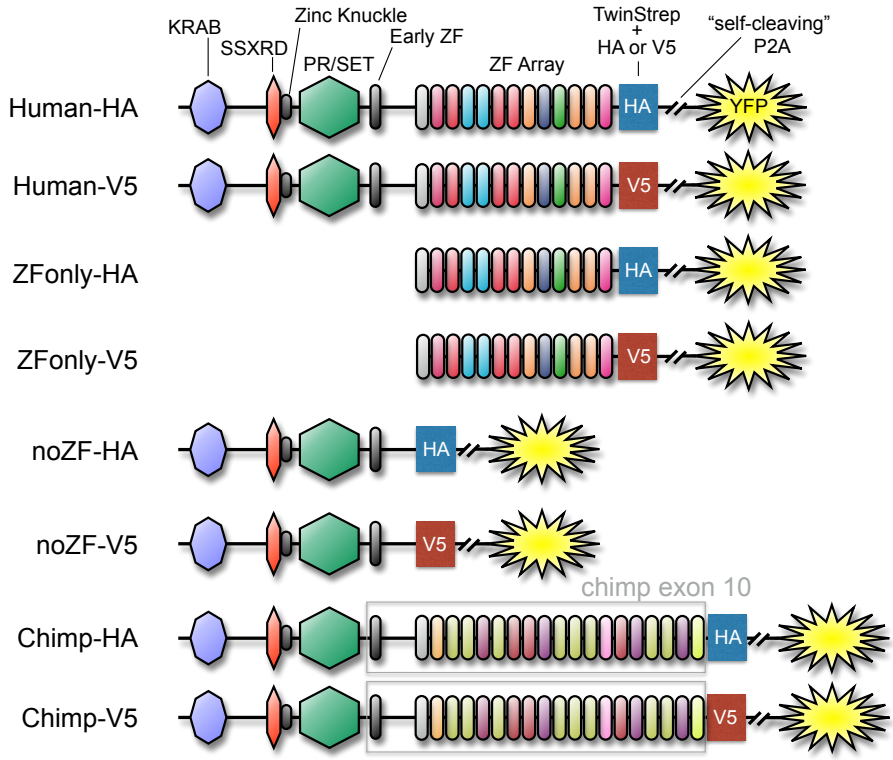


Figure 4.1: Summary of PRDM9 constructs used. Eight different tagged PRDM9 cDNAs were generated. Domains are annotated and roughly scaled with their size in the primary amino acid sequence. Human-HA/V5: a synthesised cDNA of the human reference ("B") allele with tags placed at the C-terminus, all separated by flexible linkers, including a TwinStrep affinity purification tag, a high-specificity HA or V5 epitope tag, and a "self-cleaving" P2A domain that causes the C-terminal YFP tag to fail to attach to the rest of the protein during translation. ZFonly-HA: contains only the human ZF array, beginning at the degenerate first zinc finger. noZF-HA: the exact complement of ZFonly-HA, containing everything upstream of the human ZF array. Chimp-HA/V5: the same as Human-HA/V5, with the region corresponding to Exon 10 replaced with a synthesised copy of Exon 10 from the chimp reference ("W11a") allele.

TwinStrep and HA or V5 tags along with part of the P2A sequence, which together add only 66 or 71 amino acids to the C-terminus of the protein, respectively. This design allows one to confirm full-length protein expression easily by looking for YFP fluorescence under a microscope (see **Figure 4.2**), but it prevents the large YFP tag (238 amino acids) from interfering with the protein's folding or activity.

To express these tagged cDNA constructs, I cloned them into a commercially available mammalian expression vector containing a very strong CMV promoter, a WPRE 3′ element that increases expression efficiency, and an SV40 replication origin

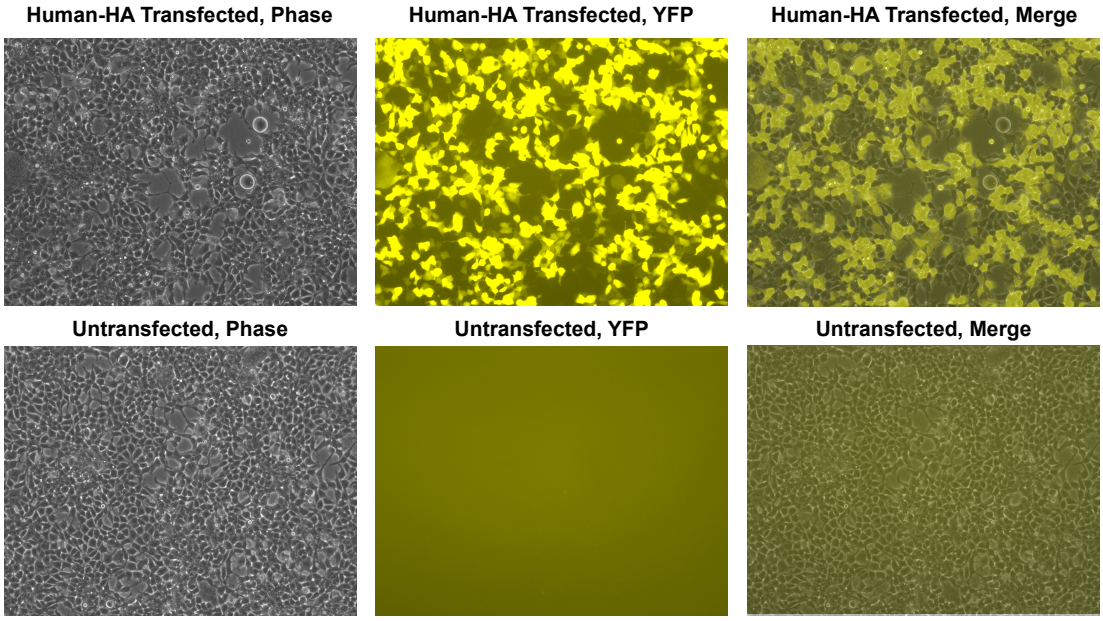


Figure 4.2: Confirmation of expression by fluorescence microscopy. The YFP tag on each construct allows one to quickly check expression levels and transfection efficiency under a fluorescence microscope. Here, Human-HA transfected (upper images) and untransfected cells (lower images) are compared. The left-most images are from phase-contrast light microscopy to show the locations of cells. The middle images are from the YFP channel, false-coloured yellow, and the right-most images overlay these two channels. High expression levels appear visible in the majority of cells.

that allows the plasmid to be replicated in cells containing the SV40 T-antigen. I showed that this vector yielded higher protein expression levels than other vectors that were in use in our group. To transport these expression vectors into cell nuclei where they could be expressed, a process called transfection, I used FuGENE-HD, a highly efficient transfection reagent that forms tight complexes with the plasmid DNA and increases the rate at which the DNA is actively taken into the cell from the surrounding solution. To express multiple constructs simultaneously, I simply mixed the expression construct DNA in equimolar proportions prior to transfection. However, because relative expression levels were not always equal, I included input lanes in each experiment to normalise out any deviation from expression parity.

I chose to perform these experiments in HEK293T cells, a highly transfectable human cell line that contains an SV40 T antigen, allowing it to replicate and amplify our expression vectors once transfected. Because the transfected DNA is not integrated into the genome, it does not typically segregate with the chromosomes

during mitosis and so becomes diluted or lost with each cell cycle. Thus, the expression is transient and diminishes after a few days, which is why one must transfect millions of cells at once and harvest them after 48 hours. To break apart the cells and their nuclei in order to release the protein complexes, I lysed the cells in a gentle detergent with a physiological salt concentration to help preserve protein-protein interactions, and we performed washing steps to remove non-specifically bound proteins using this same buffer (see Methods).

4.2 Results

4.2.1 PRDM9 can bind to itself

For the first iteration of this experiment, I included several controls to demonstrate the specificity of the HA and V5 antibodies (see **Figure 4.3**). Neither antibody yielded any bands when used for IP in untransfected cells. The only bands visible in the western lanes for these cells correspond to the light (∼25 kDa) and heavy (∼50 kDa) chains of the primary antibody, which are present in all IP lanes. I also performed IP with each antibody in cells that were transfected with only one construct. The V5 antibody detects no bands in the input lysate from cells transfected with the HA construct only, or in the HA-immunoprecipitated complexes from these cells, and *vice versa*, demonstrating that each antibody does not bind the other tag. However, when the transfected construct, the IP antibody, and the western antibody all correspond to the same tag, one can see a clear signal of a triplet of bands around ∼100 kDa visible in both the input lysate and in the IP lane. These correspond to the tagged PRDM9 construct (predicted size ∼110 kDA) along with several smaller products thought to result from protein degradation or cleavage near the N-terminus, which have been reported previously [51, 127]. This shows that full-length tagged PRDM9 was successfully expressed in these cells and then successfully isolated by IP, and that the antibodies are highly specific, showing no affinity for the wrong tags or for endogenous proteins.

The last three lanes of this western blot (**Figure 4.1**) demonstrate the relevant co-IP result from cells transfected with both tagged constructs. As is clearly visible

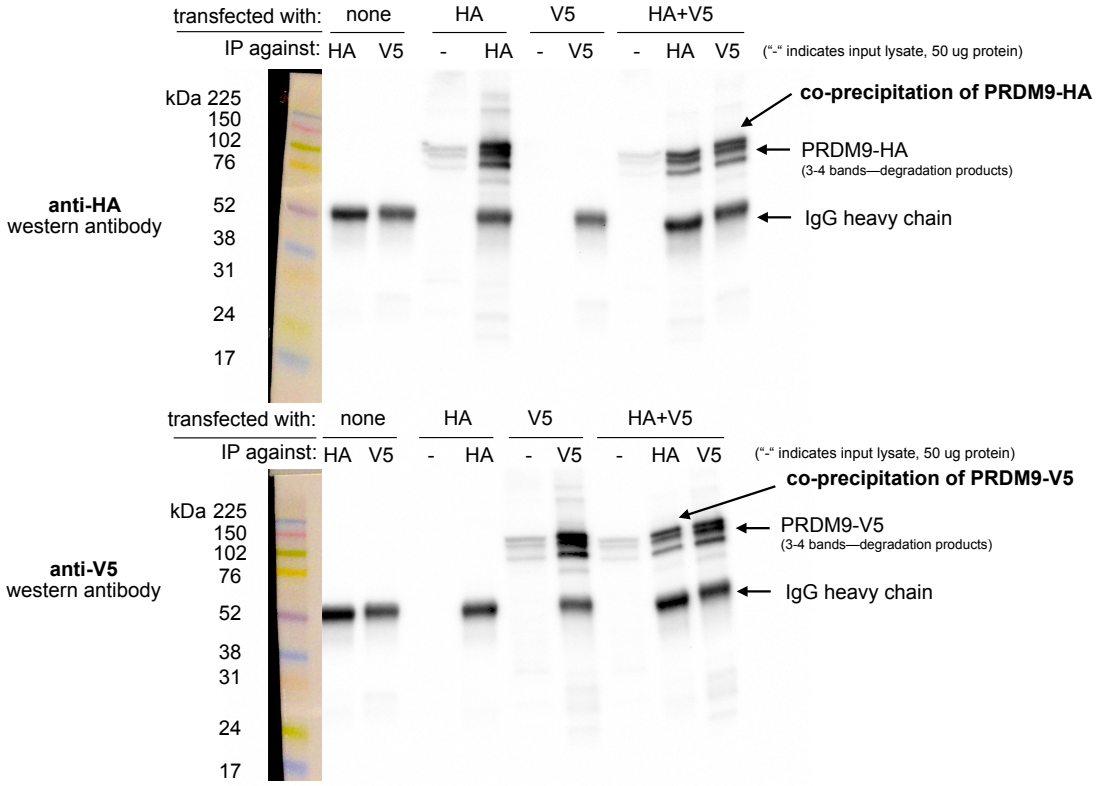


Figure 4.3: PRDM9 can form multimers. Western blot illustrating controls and experimental results. Samples were split and run on two blots separately, one imaged using an anti-HA antibody (upper) and one using an anti-V5 antibody (lower). Exposure time was 4 minutes. Ladder lanes are overlaid on the left, with approximate sizes in kiloDaltons noted. Lanes are labelled according to which full-length Human construct (HA or V5) was used, as well as which antibody was used for immunoprecipitation. IgG heavy chains are visible around ∼50 kDa, while the Human allele is visible as a band around ∼100 kDA with two or three smaller bands beneath it, likely representing degradation products [51, 127]. "-" is a short-hand label for input lanes, for which 50 μg of input chromatin was loaded in each well. The first six lanes demonstrate the specificity of the antibodies and their lack of cross-reactivity. The last two lanes show the co-IP experimental results confirming multimerisation.

from the input lanes and from the IP lanes matching the western antibody, both tags are successfully expressed in these cells and are successfully pulled down by IP. The experimental results then appear very clearly in the lanes where the western antibody and the IP antibody do not match (the rightmost lane in the upper blot and the one immediately to its left in the lower blot). If PRDM9 were unable to bind itself, we would expect these lanes to be empty, as in the negative controls. However, we instead observe striking bands matching the triplet pattern of PRDM9 at exactly the same molecular weight. That is, when we IP against the HA-tagged

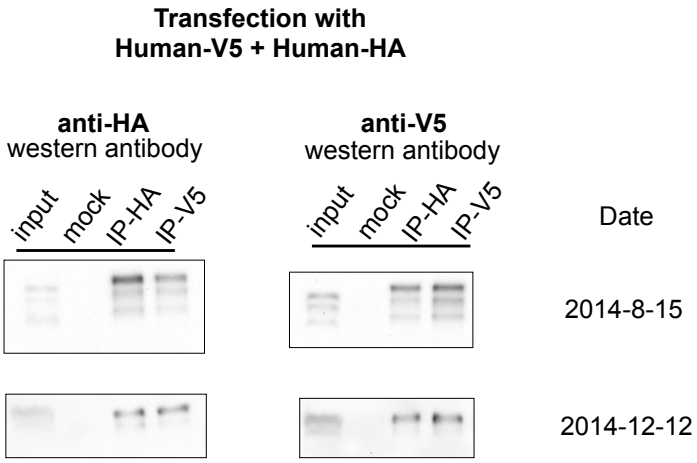


Figure 4.4: Replicate experiments with mock controls. Two independent replicates were performed to confirm the formation of multimers with the full-length human constructs, using IgG mock control lanes to rule out nonspecific co-precipitation. Images were cropped to include only the PRDM9 bands. Input lane bands appear to have run lower on the upper blot due to the use of a higher concentration of loading buffer in the IP lanes, an issue which was avoided in subsequent experiments.

construct, we detect the V5-tagged construct very robustly; and when we IP against the V5-tagged construct, we detect the HA-tagged construct very robustly. This is consistent with human PRDM9 binding strongly to itself.

However, from this experimental result alone we could not rule out the possibility that the co-IP phenomenon was simply due to protein carryover from insufficient washing. That is, perhaps the other tag was observed only because it stuck nonspecifically to the magnetic beads or to the tubes during the IP step. To rule out this possibility, I repeated the same experiment but included a "Mock" control, in which I performed all IP steps exactly the same but used a negative control primary antibody, which lacks affinity for any protein. If the observed co-IP bands still appeared in the Mock lanes, it could only be due to nonspecific interactions with the control antibody or the magnetic beads or the tubes, and not due to a specific interaction with PRDM9 immunocomplexes. As shown in **Figure 4.4**, I replicated this experiment twice more with a mock control and demonstrated that the mock lane is completely blank. This soundly confirms that our tagged human PRDM9 proteins are capable of binding each other when jointly overexpressed in human cells.

4.2.2 Multimerisation is mediated primarily by the ZF array

Next, to dissect PRDM9 and search for the domain responsible for its self-binding behaviour, I essentially split the full-length human PRDM9 cDNA into two pieces: one containing only the C-terminal Zinc Finger domain (the "ZFonly" construct), and one containing everything else (the "noZF" construct; see **Figure 4.1**). For each piece, I generated two C-terminally tagged constructs as I had done with the full-length protein: one with a TwinStrep-HA-P2A-YFP tag, and one with a TwinStrep-V5-P2A-YFP tag. I cloned these new constructs into the same expression vector used for the full-length constructs and transfected them in various combinations with each other and with full-length PRDM9. Because these new constructs are smaller, they tended to transfect and be expressed more efficiently than the full-length construct.

Figure 4.5a illustrates very clearly that the ZF domain alone is responsible for nearly all self-binding activity. In the first experiment, I co-transfected noZF-HA and noZF-V5 then used the V5 antibody for IP and the HA antibody for western detection. Despite very high expression levels visible in the input, only a very faint co-IP band is visible in the IP lane. Because the mock lane is clean, this band is likely to reflect a real but very weak self-binding capability mediated by the non-ZF portion of PRDM9. However, this contrasts greatly with the intense co-IP band visible when co-transfecting ZFonly-HA with ZFonly-V5. This suggests that the zinc finger domain of one PRDM9 protein can bind strongly to the zinc finger domain of another, while the rest of the protein interacts only weakly at best.

To further confirm this, I co-transfected full-length, V5-tagged human PRDM9 with either noZF-HA or ZFonly-HA then performed IP with anti-V5 and western detection with anti-HA. Once again, only a very faint co-IP band is visible with the noZF construct, and a very intense band is visible with the ZFonly construct. To replicate this finding, I repeated the co-transfection of Human-V5 with ZFonly-HA and did the IP-western experiment in both directions, further demonstrating that the ZFonly construct is sufficient to bind and pull down the full-length construct (see **Figure 4.5b**).

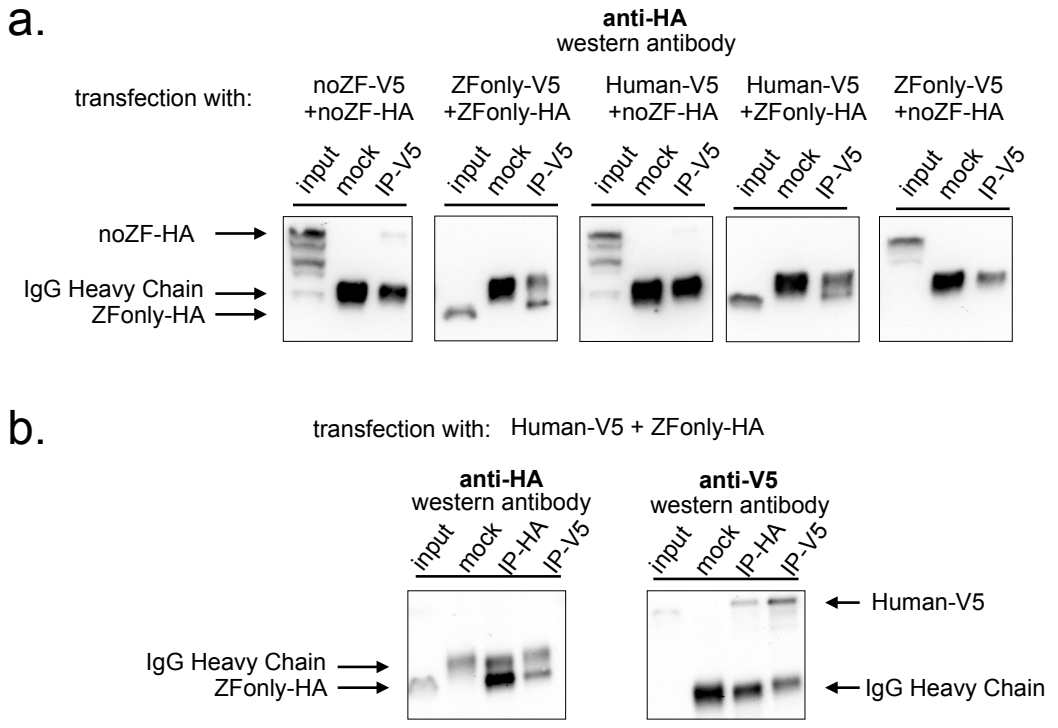


Figure 4.5: Multimerisation is mediated primarily by ZF-ZF binding. Western blots illustrating co-IP results for various combinations of full-length human, noZF, and ZFonly constructs. **a**: The first and third blots show only a very faint co-IP signal despite strong input expression of the noZF construct, indicating that the non-ZF portion of PRDM9 cannot form multimers with itself or full-length PRDM9. The second and fourth blots show strong co-IP signals for the ZFonly construct, indicating that the ZF domain binds itself and binds the full-length Human construct. The fifth plot shows that the ZFonly and noZF constructs do not bind each other and confirms that multimerisation is not mediated by the C-terminal tags. **b**: A replication of the experiment shown in the fourth blot above, but performing the IPs and western blots in both directions. This confirms that the full-length Human construct can pull down the ZFonly construct, and the ZFonly construct is sufficient to pull down the full-length Human construct.

As a negative control, I further co-transfected the noZF construct with the ZFonly construct and demonstrated that no co-IP band is visible at all, ruling out any sort of interaction between the ZF domain and the rest of PRDM9. These results further confirm that any observed self-interaction properties are not simply due to the presence of the C-terminal tags on these constructs but must be mediated by PRDM9 itself.

4.2.3 Heteromultimers of different ZF arrays form less efficiently

Finally, to examine the specificity of ZF array binding, I generated a construct in which I replaced the final exon containing the human ZF array with a synthesised cDNA matching the final exon of the chimpanzee reference PRDM9 allele (W11a). This chimp allele differs from the human allele in both the types and number of zinc fingers present (18 zinc fingers versus 12, respectively). There are also several nonsynonymous changes in the region just upstream of the ZF array, but as demonstrated with the noZF constructs, this region contributes at most only weakly to PRDM9's self-binding ability. Again, using the same C-terminal tags used to generate the full-length Human-HA and Human-V5 constructs, I generated two tagged constructs for the chimp allele and cloned them into the same mammalian expression vector, yielding constructs that I will refer to as Chimp-HA and Chimp-V5.

First, as a positive control, I demonstrated that Chimp-HA and Chimp-V5 can bind each other (see **Figure 4.6a**). Chimp-V5 plus Human-HA, on the other hand, did not yield a visible co-IP band. However, because the input expression level of the Human-HA construct seemed unusually faint in this experiment, we designed additional co-transfection experiments to examine the relative affinities of Human and Chimp zinc fingers for themselves versus each other. First, I performed a *triple* transfection of either Human-V5 or Chimp-V5 as "bait" with both the noZF-HA and ZFonly-HA constructs as "prey". I included the noZF construct because it is expected to have identical weak affinity for the Human and Chimp constructs. That is, any co-IP of the noZF construct should occur independently of the ZF array, allowing it to serve as a reference band to which I could normalise other co-IP band intensities. As shown in **Figure 4.6b**, the human ZFonly-HA construct is pulled down much more efficiently by the Human-V5 bait than by the Chimp-V5 bait. As the input lanes show, prey expression levels were highly similar between the two experiments. Looking at the same samples with V5 western detection (see **Figure 4.6c**), it appears that in this experiment, the Chimp-V5

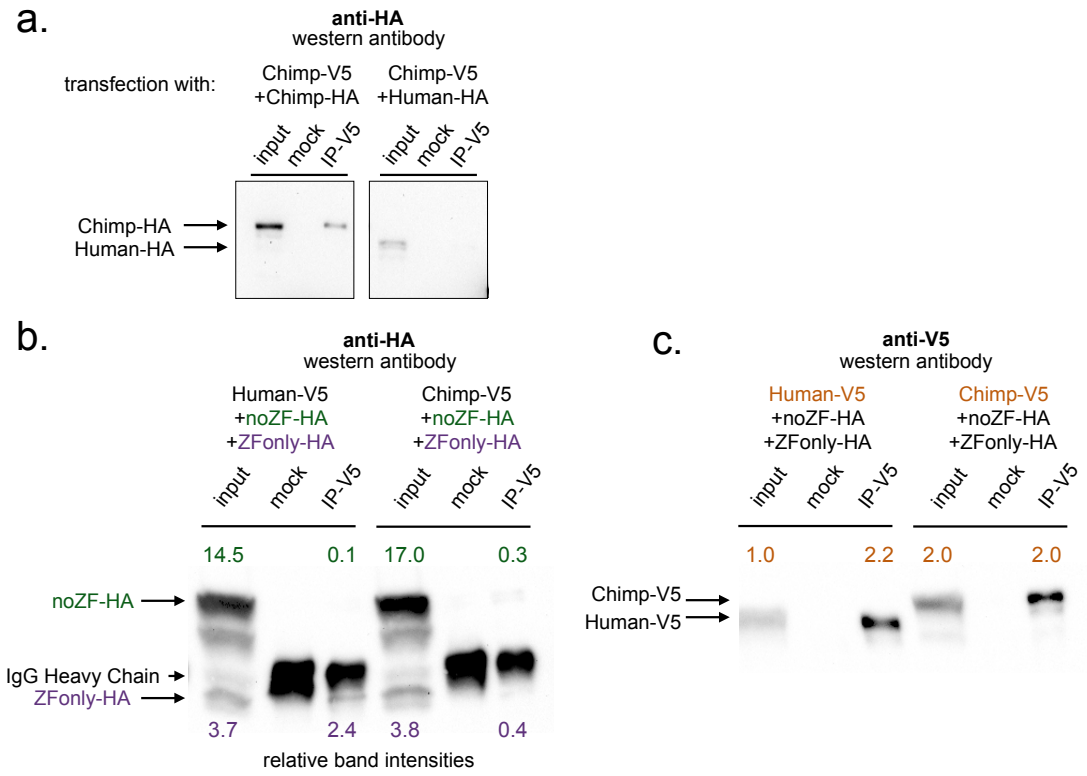


Figure 4.6: Homo-multimers form more frequently. **a**: Western blot illustrating multimerisation of the Chimp construct but weak to no multimerisation of the Chimp and Human constructs. **b**: A triple transfection of the Human or Chimp constructs as bait, with the noZF and ZFonly constructs as prey, to allow normalisation of ZFonly bands to noZF bands (which should be faint and constant between the two experiments). Co-IP relative band intensities are printed above (green, noZF) and below (purple, ZFonly). **c**: The samples as in **b**, but developed with an anti-V5 antibody to show relative expression levels of the Chimp and Human baits.

bait construct had two-fold higher expression than the Human-V5 bait construct (based on a densitometry count of band intensity), consistent with the fact that the noZF-HA co-IP band is more intense for Chimp than for Human. However, despite its greater expression, the Chimp bait pulled down six-fold less ZFonly-HA prey than the Human bait did. By comparing measured relative band intensities normalised to the respective noZF-HA co-IP band, I estimated a 13-fold difference in co-IP efficiency between the Human and Chimp constructs. While this serves only as an approximation of the true binding affinities, which should in the future be measured more precisely by alternative biochemical methods, it seems clear that homo-multimerisation is preferred over hetero-multimerisation.

To confirm this preference for homo-multimerisation in mixtures of Human and Chimp PRDM9, I also performed direct competition experiments between the two alleles. To do so, I transfected cells with equimolar mixtures of DNA for three constructs, for example Chimp-V5 plus Chimp-HA plus Human-HA. In this case Chimp-V5 would be the "bait" pulled down by IP with anti-V5, and Chimp-HA and Human-HA would be the co-IP "prey" detected by western blotting with anti-HA. **Figure 4.7** illustrates that Chimp targets are more efficiently pulled down by a Chimp bait, despite the fact that the Chimp construct appears to be expressed at lower levels in the input lysate in this experiment. For example, as shown in (see **Figure 4.7a**), the Chimp-V5 target is expressed at roughly half the level of the Human-V5 target in the input, but the co-IP band using a Chimp-HA bait is 66% more intense for the Chimp target versus the Human target, suggesting a simple effect size of approximately 3.2 for Chimp-Chimp versus Chimp-Human binding. For the same experiment but using Chimp-HA as bait, this effect size is estimated as 4.0.

Finally, to replicate this result we repeated the experiment but lysed the cells in a high-salt (500 mM NaCl) buffer, which we then diluted back to 150 mM. This should have the effect of disrupting weaker protein-protein interactions formed *in vivo* then allowing proteins complexes to re-form *in vitro* in the lysis buffer. Input lysates were not saved from these experiments, but we can reasonably expect different transfections with the same combinations of prey constructs to produce similar relative expression levels. By comparing the relative intensities of target bands between experiments that simply swap the baits, we are then still able to infer a direction of binding preference. As illustrated in **Figure 4.7b**, once again Chimp shows a strong binding preference for Chimp, and Human shows a strong binding preference for Human, in all combinations of bait and target tags.

Although we did not directly measure relative expression levels of the two targets in the input lysate from each transfection, we can infer these values from the observed band intensities if we assume a constant effect size (odds ratio of forming various multimers) between Chimp-Chimp and Human-Human binding preferences and a constant Chimp:Human input ratio in transfections with identical

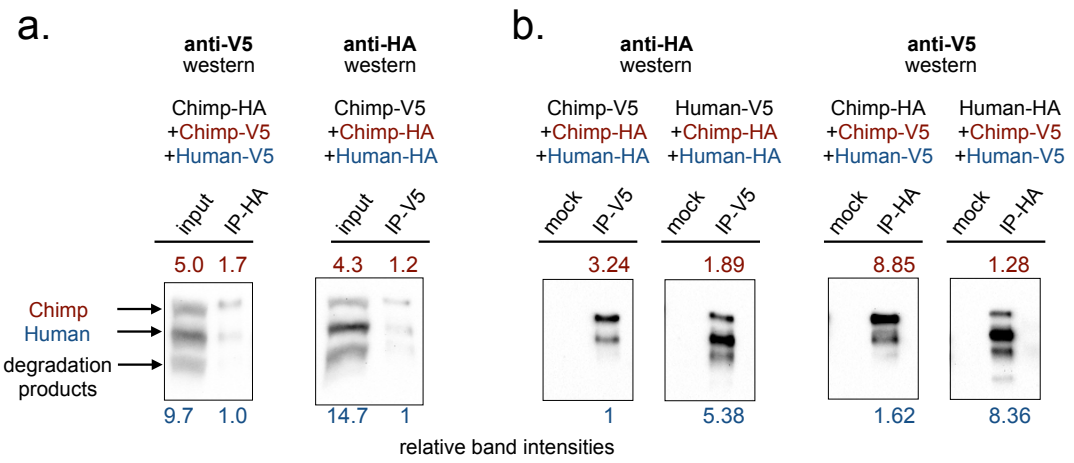


Figure 4.7: Homo-multimers form preferentially in competition assays. Western blots illustrating the results from triple transfections with either Human or Chimp bait and a mixture of Chimp and Human prey. Relative band intensities are plotted above and below (note these values cannot be compared across blots with different western antibodies). **a**: Experiments with chimp as bait including input samples for normalisation. Despite higher Human prey expression in the input, more Chimp prey is pulled down, and this result is consistent in a replicate with the tags and antibodies switched. **b**: Experiments with either Chimp or Human as bait, replicated with both antibody directions, using higher salt lysis conditions. Although no input lane is shown, if we assume prey concentrations are similar between samples with identical prey, then the direction of the effect is consistent: Human bait pulls down more Human prey, and Chimp bait pulls down more Chimp prey.

prey. The effect size can then be solved from a simple system of two equations. From the V5-tagged baits we estimate an effect size of 3.03 and from the HA-tagged baits we estimate an effect size of 5.97. Again, these serve only as rough indicators of relative binding affinities and are subject to high experimental and measurement noise. However, these results consistently show that under multiple different transfection and lysis conditions, homo-multimer affinities are stronger than hetero-multimer affinities by several fold.

It may be the case that this self-binding preference is driven primarily by array length. If, for example, longer zinc finger arrays simply tend to form more stable multimers, owing to simple additivity of ZF affinity, then it might be the case that the Chimp-Chimp preference we observe simply owes to its longer ZF array length and not to the identity of its zinc fingers. This would also have the effect of reducing

the amount of Chimp PRDM9 available to form hetero-multimers, which would in turn drive up the concentration of Human homo-multimers. Further experiments would be needed to deconvolve ZF array length from ZF composition in driving self-binding specificity. However, a composition-based binding preference would seem to be a necessary quality *a priori*, otherwise PRDM9 could promiscuously bind to the many other Cys_2His_2 zinc finger proteins expressed during meiosis. Our data suggest that PRDM9's zinc fingers may not only provide high DNA-binding specificity but also high protein-binding specificity.

4.3 Discussion

Here, I have demonstrated with an *in vitro* experimental approach that PRDM9 can bind to itself and form multimers, a behaviour mediated primarily by its zinc finger array in an allele-specific fashion. When PRDM9 constructs with two different epitope tags are transiently overexpressed in a human cell line, both tags are detected in immunocomplexes isolated using either tag. The region of PRDM9 upstream of the C-terminal ZF domain can only form multimers very weakly by comparison with the ZF domain alone, demonstrating that PRDM9 multimerisation is caused primarily by its DNA-binding ZF domain. Because the ZF domain is highly variable within and between species, I tested the hetero-multimerisation ability of two very different ZF arrays derived from human and chimp and showed that hetero-multimers form much less efficiently than homo-multimers.

This work opens up many new and exciting questions about the biochemical properties and cellular functions of PRDM9 multimers. Emmanuelle Bitoun is now beginning a suite of experiments to further explore this activity *in vitro* and to confirm the existence of PRDM9 multimers *in vivo* in a mouse model. Firstly, it will be important to determine the copy number present in each complex: does PRDM9 simply form dimers or can it form larger complexes with itself? Ultracentrifugation sedimentation experiments with large quantities of recombinant PRDM9 should resolve this question, provided that PRDM9 multimers can form outside of a cellular context. Secondly, it will be important to test whether PRDM9's zinc fingers can

simultaneously bind to DNA and to each other, or whether these two processes compete. This can be ascertained by competition experiments using differentially tagged PRDM9 constructs and biotinylated DNA oligos. It will also be important to test the limits of homo- and hetero-multimerisation between various different types and combinations of zinc finger arrays. Can an allele with only one zinc finger still bind to itself? Do the incompatible hybrid mouse alleles (PWD and B6) show any unusual multimerisation properties? Which amino acids are responsible for mediating PRDM9's ZF-ZF interactions? Finally, *in vivo* experiments will be possible using transgenic mice that I helped design, which contain PRDM9 alleles differentially tagged using the same HA/V5 tags designed for these cell line experiments. Various IP-western, IP-Mass Spectrometry, and Förster Resonance Energy Transfer experiments will then be possible to confirm the existence of PRDM9 multimers in mammalian meiosis.

Currently, we can only speculate about what function PRDM9 multimerisation might serve in meiosis. If ZF-ZF interactions are incompatible with DNA binding, it may be the case that self-binding serves as a means of controlling the number of copies of PRDM9 free to bind to DNA. It could also act as a means of inactivating unbound PRDM9 and preventing it from binding and methylating promiscuously. More intriguing is the possibility that PRDM9 can bind itself and DNA simultaneously, which might fit into our hypothesis about PRDM9-mediated homologue pairing. The inefficiency of hetero-multimer formation suggests that the PRDM9-mediated pairing of homologous binding sites could still occur in an allele-specific way in PRDM9 heterozygotes. Additional experiments will begin to narrow down these hypotheses, but the work presented here represents a critical first step in opening up a new avenue of inquiry about PRDM9.

Subsequent to the writing of this chapter, another group published results confirming the ability of PRDM9 to form multimers [128]. They further show that in cells expressing two different alleles (the human A and C alleles) for which one has an inactive PR/SET domain, binding sites for the mutated allele still show elevated H3K4me3 signal. This indicates that DNA-binding and trimethylation

behaviours remain active in PRDM9 multimers. However, our experiments present several findings not investigated in the previous study, namely that the ZF array is almost entirely responsible for multimer formation and the human and chimp alleles exhibit a strong preference for homomultimer formation. It would be interesting to investigate homo- and hetero-multimer formation for the human A and C alleles and for the B6 and CAST mouse alleles examined in their study. Reduced heteromultimer formation would significantly impact their conclusions regarding the effects of multimer formation on PRDM9 allelic dominance.

4.4 Methods

4.4.1 Cell culture and transfection

HEK293T cells (ATCC CRL-3216) were thawed and incubated at 37°C with 5% CO_2 in DMEM (Sigma D6546) supplemented with 10% foetal bovine serum (Sigma F7524), 1X L-Glutamine (Sigma G7513), and 1X penicillin/streptomycin (Sigma P0781). Confluent cells were split 1:10 and passaged for no longer than a month before transfection. The night before transfection, confluent cells were trypsinised (Sigma T3924), diluted in growth medium, and counted on an automatic haemocytometer (BioRad TC20). For each experiment, 10 million cells were seeded in 20 ml growth medium in a 15 cm round cell culture dish. The following morning, cells were transfected by mixing 30 μg total DNA into 800 μl OPTI-MEM (Life Technologies 31985062), then carefully adding 90 μl FuGENE-HD Transfection Reagent and flicking to mix, incubating at room temperature for 15 minutes, and then adding the mixture dropwise to each dish while swirling gently to mix. After 48 hours, cells were imaged briefly with a fluorescence microscope to confirm expression and were subsequently harvested. As negative controls, additional cells were seeded at the same time but were not transfected.

4.4.2 Cell lysis and immunoprecipitation

Dishes were aspirated to remove media and cells were washed with cold PBS. 2 ml of cold lysis buffer (1% Triton X-100, 150 mM NaCl, 50 mM Tris pH 8.0 plus 2X

final concentration of Roche cOmplete Protease Inhibitor Cocktail Tablets) were added and cells were collected into 2 ml Eppendorf tubes using a cell scraper. Tubes were incubated on ice for 30 minutes and lysates were dounced 20 times in a 2 ml dounce homogeniser with a tight pestle to help shear nuclear membranes. Cells were spun at 2000g for 5 minutes to remove chromatin and cell debris. 100 μl of lysate was set aside as an input control, and the remainder was split evenly among experimental and mock IP conditions. 2 μg of primary antibody (Abcam ChIP-grade rabbit polyclonal anti-HA ab9110 or anti-V5 ab9116, or rabbit polyclonal IgG isotype control ab171870) was added and lysates were incubated for 1 hour at 4°C with rotation. For each sample, 25 μl of magnetic beads (Invitrogen M-280 Sheep Anti-Rabbit Dynabeads) was equilibrated by washing 3 times in 1 ml cold PBS/BSA (1X PBS, 5 mg/ml BSA, filtered with 0.45 micron filter), then resuspending in 25 μl PBS/BSA. Beads were added to the lysates and incubated for an additional hour at 4°C. Tubes were spun down and placed on a magnetic rack for 1 minute. Beads were pipetted up and down in 1 ml cold lysis buffer and rotated for 3 minutes at 4°C. Washing steps were repeated 4 more times, with all steps taking place in a cold room at 4°C.

4.4.3 Western Blotting

Beads were resuspended in 20 μl 2X Laemmli western loading buffer and boiled for 5 minutes at 100°C. Beads were removed on a magnetic stand and supernatants were diluted two-fold. The total protein concentrations of input lysates were estimated using a Pierce BCA Protein Assay Kit (Life Technologies 23227) and a NanoDrop spectrophotometer. 4X Laemmli buffer was added to 50 μg of input protein to a final concentration of 1X then boiled for 5 minutes at 100°C. Samples were run on 10-well 7.5% BioRad mini-Protean TGX pre-cast gels at 150 Volts in standard TGX running buffer for approximately 1 hour, using 5 μl of Full-Range Rainbow Ladder (VWR 95040-114) in one well. Gels were then assembled onto a BioRad mini Trans-Blot transfer pack (with PVDF membrane) according to manufacturer instructions and run on a Trans-Blot Turbo machine on the Mixed MW setting

(2.5A, up to 25V, 7 mins). Membranes were quickly removed and transferred to 50 ml conical tubes, then blocked for 5 minutes with rotation in 10 ml Blocking Buffer (5% milk in PBS with 0.1% Tween-20), which was then poured off. Primary antibodies were diluted 1:5000 in 5 ml blocking buffer and added to the membranes and incubated for 1 hour at room temperature with rotation. Membranes were washed 3 times for 5 minutes each in PBST (PBS with 0.1% Tween). Secondary antibody (Amersham ECL Donkey anti-Rabbit IgG, HRP-linked, NA934) was diluted 1:30,000 in blocking buffer, then 5 ml was added to each membrane and they were incubated for 1 hour at room temperature with rotation. Membranes were washed an additional 3 times in PBST and one final time in PBS. Blots were imaged using a BioRad Clarity ECL kit according to manufacturer instructions and placed between sheets of transparency film to prevent drying during imaging. Imaging was performed using a BioRad ChemiDoc MP Instrument using chemiluminescence hi-sensitivity settings and signal accumulation mode for various exposure times. Image processing was performed in the BioRad ImageLab software, in which relative bands intensities were quantified by densitometry.

I suppose the process of acceptance will pass through the usual four stages: (i) this is worthless nonsense; (ii) this is an interesting, but perverse, point of view; (iii) this is true, but quite unimportant; (iv) I always said so.

— J.B.S. Haldane, *Journal of Genetics* (1963)

5

Conclusions and future directions

Just over a decade has passed since Hayashi *et al.* first discovered the indispensable role of PRDM9 in mouse meiosis, suggesting it might be a transcription factor given its histone H3K4 trimethylation activity [38]. Four years later, PRDM9 became of great interest to a diverse community of biologists who converged on it from independent lines of inquiry that unveiled its fascinating and wide-reaching properties. In 2009, years of careful mapping studies and experiments in mice culminated in the landmark Mihola *et al.* paper, identifying PRDM9 as the first mammalian speciation gene [75]. In that same year Oliver *et al.* showed *PRDM9* to be among the most rapidly evolving mammalian genes, subject to strong positive selection at its DNA-binding residues [61]. Less than a year later, three papers, from statistical genetics approaches [34], sperm typing experiments [36], and a mouse mapping study [35], revealed that PRDM9 is responsible for localising mammalian recombination hotspots. Since then, hundreds of papers have continued to develop and retell the story of PRDM9, but much has remained unknown about how it performs its most basic functions in initiating recombination, and about how (and why) these functions contribute to its rapid evolution and role in speciation. The body of work presented here contributes substantial evidence toward answering each of these big questions, and in turn proposes many new testable hypotheses to address the multitude of new questions following from this work. Here I will

review the major contributions of this work in the context of these big questions, and I will summarise ongoing studies that I have helped design to address some of the open questions that remain.

5.1 PRDM9 as one of several factors controlling recombination outcomes

After overexpressing PRDM9 in a mitotic human cell line and performing ChIP-seq, I applied novel statistical approaches to this well-powered dataset to discover new binding modalities of PRDM9's long zinc finger array. I show that PRDM9 likely binds with different subsets of its zinc fingers simultaneously across the entire array, with one class of low-affinity internal zinc fingers potentially serving as linkers permitting different spacings between the flanking high-affinity zinc fingers. By comparing our binding peaks and motifs to genome annotations and to known recombination hotspots, I then identified several factors that associate with different recombination outcomes. Strikingly, I found that the human reference ("B") allele of PRDM9 can bind to most gene promoters, and by carefully controlling for sequencing biases in these regions with our novel peak-calling algorithm, I showed that motifs within promoters are bound with ∼2-fold lower enrichment than their non-promoter counterparts but are associated with ∼7-fold lower recombination rates. This suggests a complex mechanism by which PRDM9 fails to initiate recombination in promoters, potentially downstream of PRDM9 binding and trimethylation, and not simply owing to a strong binding affinity away from promoters, as has been suggested in mice [37]. Interestingly, I found that one of the seven motifs is associated with consistently lower recombination rates, even after controlling for binding enrichment and repeat/promoter/chromatin context. This could potentially represent a difference between the more common A allele (responsible for most of the signal in LD-based recombination maps) and the reference B allele (used in this study), or some sort of meiosis-specific chromatin mark or transcription factor that associates with this motif and prevents PRDM9 binding or DSB formation in meiosis. I also confirmed that PRDM9 trimethylates H3K36me3 *in cis* and that

it strongly phases the surrounding nucleosomes after binding. Finally, I produced the first experimental data revealing the DNA-binding properties of a chimpanzee PRDM9 allele, showing that its binding targets are most strongly specified by only a small subset of its zinc fingers, which are conserved among many chimpanzee alleles. This raises the intriguing possibility that selection acts in part to favour PRDM9 alleles binding specific motifs, rather than merely favouring alleles with a greater number of available binding sites (as has been suggested by the "Red Queen" hypothesis). Such selective forces could explain why specific classes of PRDM9 alleles predominate in human populations.

Several important questions about the molecular properties of PRDM9 remain unanswered. The functions of its KRAB and SSXRD domains remain entirely unknown, and could be investigated by mutation and truncation experiments in transgenic mice followed by extensive phenotyping. Alternatively, specific hypotheses could be tested rapidly in an experimental system similar to our transfection experiments. For example, to test whether these domains directly bind SPO11 or other recombination initiation proteins, one could co-transfect HEK293T cells with tagged cDNAs for these recombination proteins plus one of various tagged PRDM9 constructs with engineered domain deletions, then perform co-IP western experiments similar to those performed in Chapter 4. This would help to resolve one of the biggest remaining open questions: how does PRDM9 binding recruit DSB proteins to initiate recombination?

To address these and other questions, Ben Davies and colleagues have produced transgenic mice containing PRDM9 alleles tagged at the C-terminus with the same tags used in our HEK293T cell experiments, which I designed to have several important properties: 1) a "self-cleaving" P2A-YFP tag to fluorescently label cells expressing PRDM9, 2) a TwinStrep tag to allow for affinity purification of PRDM9 protein complexes for proteomic experiments such as Mass Spectrometry, and 3) an HA or V5 epitope tag to allow for ImmunoFluorescence or ImmunoPrecipitation experiments *in vivo*, allowing us to potentially visualise single-molecule behaviours or to perform ChIP-seq with a high-affinity, high-specificity antibody. Emmanuelle

Bitoun, Ben Davies, and Daniela Moralli are currently breeding and characterising V5-tagged humanized mice for these purposes. Because the constructs I designed are already cloned into lentiviral vectors, it is also possible to introduce and express these PRDM9 constructs in a variety of cell types, at least transiently. For example, one could attempt fertility rescue in PRDM9 knockout mice or in infertile hybrid mice by injecting PRDM9-containing lentivirus directly into their seminiferous tubules.

5.2 New evidence for PRDM9 as a transcription factor

Given my finding that the human PRDM9 B allele can bind with some affinity to most promoters and deposit the H3K4me3 mark in our cells (and by extension, can likely do so in meiosis as well), I sought to test whether PRDM9 can influence the expression of bound genes. We leveraged the tractability of our experimental system to perform RNA-seq with careful controls, comparing human-PRDM9-transfected cells to untransfected cells and to cells transfected with the chimp allele or with only the human ZF domain. By scanning for stringent subsets of genes with differential expression only in the human-transfected sample, two strong spermatogenesis-specific candidate genes emerged: *CTCFL* and *VCX*, whose expression change was validated by qPCR. I confirmed that PRDM9 binds in the promoter region of *CTCFL* and deposits the H3K4me3 mark *de novo*, and that expression increases from a vanishingly small level in untransfected cells more than 28-fold, to an absolute level at least one-fifth of that measured for the ubiquitously expressed *CTCF* gene (*n.b.*: transfection efficiency is not 100%, and the *CTCFL* promoter may not be bound in every transfected cell, so the true absolute expression level following promoter binding is likely much higher). By relaxing these stringent filters, I identified an additional 45 protein-coding genes whose expression may be affected by PRDM9 binding at their promoters.

These data make it clear that PRDM9 binding is not sufficient to predict a change in gene expression, but it may activate transcription at a subset of its promoter binding sites. This suggests that PRDM9 may have a secondary role as

a partial regulator of certain meiotic genes, and this could explain, for example, why "A and B"-type alleles predominate in non-African human populations or why most chimp alleles are predicted to bind a similar sub-motif. Perhaps specific PRDM9 alleles can marginally improve fertility by enhancing the expression of certain meiotic genes such as *CTCFL*, and over time these become subject to strong positive selection and driven to high frequency. Because DSB formation is suppressed at promoters, these binding sites may not be subject to hotspot death, allowing them to persist (and to provide a selective advantage when bound) over long periods of time. PRDM9's gene-regulatory behaviour might also explain its interaction with the *Hstx2* locus in infertile hybrid (PWD×B6)F1 mice.

To validate the effect of the human B allele on *CTCFL* expression *in vivo*, one could perform RNA-seq or at least qPCR on cDNAs derived from early prophase spermatocytes isolated from human male testes with different *PRDM9* alleles (although fresh samples of suitable genotypes could prove very difficult to acquire). To more generally examine the effect of promoter binding on transcription *in vivo*, we could leverage the fact that the human allele has affinity for many promoters along with the fact that we have transgenic mice containing a tagged, humanized *Prdm9* allele. By measuring PRDM9 binding directly with *in vivo* ChIP-seq in our V5-tagged humanized mice, we could compare expression of bound genes in spermatocytes derived from humanized mice relative to mice with other *Prdm9* genotypes. Alternatively, transgenic mice humanized at the *Ctcfl* promoter region could be generated and crossed with humanized *Prdm9* mice to examine changes in *Ctcfl* expression and fertility phenotypes as they relate to genotypes at these two loci.

5.3 PRDM9 binding symmetry suggests a new role in meiosis

In collaboration with Ben Davies, Peter Donnelly and colleagues, I helped to design and implement transgenic mouse breeding experiments to investigate the mechanism by which PRDM9 causes speciation. We showed that humanizing the B6 *Prdm9* zinc finger exon completely rescued fertility in the infertile (PWD×B6)F1 hybrid

cross, and it also rescued full fertility in subfertile (B6×PWD)F1 hybrid mice, showing in the most controlled manner to date that the ZF array, and thus likely the DNA-binding preference, of *PRDM9* is responsible for its role in speciation. The human *PRDM9* allele evolved independently of either mouse genome and binds a completely different set of sites than either the PWD or B6 alleles, so the fact that it is capable of rescuing fertility suggests that the mechanism underlying hybrid infertility is unlikely to involve classical Dobzhansky-Muller Incompatibilities between *Prdm9* and other genomic loci. This also suggests that *PRDM9*-related speciation is potentially transient, since naturally occurring *Prdm9* mutations within mouse populations could reverse hybrid sterility in a manner similar to the humanized allele. However, given the fitness costs of producing and rearing infertile hybrid offspring, mutations in other loci could potentially arise to reinforce the reproductive isolation of populations with transiently incompatible *PRDM9* alleles.

To investigate potential mechanisms of *Prdm9* incompatibility, we generated H3K4me3 and DMC1 ChIP-seq data in testes from infertile and humanized rescue mice to produce high-resolution maps of recombination intermediates in these hybrid mice. By computing ChIP enrichment values specific to the B6 homologue or the PWD homologue at each hotspot, we discovered that each *Prdm9* allele in the infertile mouse tends to trimethylate histones and promote DSB formation only on the homologue of the opposing strain, producing a large fraction of "asymmetric" hotspots. We showed that this effect is explained by hotspot death within each subspecies (similar to recent findings in a different hybrid cross [67], but with even greater magnitude). Introducing the humanized allele vastly increases the number of symmetric hotspots, because binding targets of the humanized allele have not been subject to hotspot death in either subspecies.

Surprisingly, when I compared the relative enrichments of H3K4me3 and DMC1 at asymmetric and symmetric hotspots, I showed that asymmetric hotspots have 2-fold greater DMC1 heat for a given level of H3K4me3 enrichment when compared to symmetric hotspots, and I validated this increase in DMC1 heat in several different ways, controlling for potential confounders. This increase in DMC1 heat

implies that asymmetric binding sites are either more likely to form DSBs, or their DSBs take longer to complete homology search and repair (with the latter being more likely given the precise control of DSB number in meiosis [99]). I also found that this DMC1:H3K4me3 ratio varies by chromosome, with a stronger magnitude on chromosomes that have been shown to be more prone to asynapsis in infertile hybrids [82].

We propose a potential mechanism to parsimoniously explain these observations as well as the complex and surprising fertility phenotypes reported previously for mice with varying dosages and alleles of *Prdm9* [84]. Our data imply a novel role for PRDM9 in meiosis: by binding to the homologous chromosome at a given DSB site, PRDM9 may enhance the process of homology search and repair, promoting more rapid DMC1 removal and enhancing synapsis, which is thought to spread out from DSB sites [129, 130]. DSBs without PRDM9 binding on the homologous chromosome, on the other hand, experience inefficient homology search and repair, yielding elevated DMC1 levels and elevated levels of asynapsis. Above a certain threshold, this asynapsis can lead to meiotic arrest, apoptosis, azoospermia, and infertility. Anything that increases the total amount of symmetric PRDM9 binding on each chromosome, such as increasing the dosage of PRDM9 or introducing symmetrically binding alleles like the humanized allele, can reduce the probability of asynapsis and thus restore fertility.

This model suggests that the asynapsis of heterosubspecific chromosomes owes not to their sequence divergence broadly, but specifically to their divergence at PRDM9 binding sites, driven by hotspot death. To test this claim, we have designed and begun to implement experiments using 3-way hybrid mice, in which we can theoretically restore PRDM9 binding symmetry without altering total sequence divergence or dosage/alleles of *Prdm9*. Specifically, I have helped cross PWD females with (B6/CAST)F2 and F3 males. The resulting hybrids have pure PWD ancestry on their maternal chromosomes and a mixture of B6 and CAST ancestry on their paternal chromosomes (and have a $Prdm9^{\mathrm{PWD/B6}}$ genotype, the same as the infertile hybrid). Because B6, CAST, and PWD all have similar overall sequence

divergences but three different patterns of hotspot erosion, the paternal CAST segments effectively act to restore PRDM9 binding symmetry for the B6 PRDM9 allele, since it will not have eroded on either the maternal PWD homologue or the paternal CAST homologue in these regions. Fertility will be expected to vary among littermates, which will inherit distinct combinations and amounts of CAST ancestry on their paternal chromosomes. In initial litters of mice, Ben Davies has reported a range of fertility outcomes, from complete azoospermia to a modest rescue of sperm count. DMC1 ChIP-seq data are currently being generated for these mice, and I am helping to design cytological experiments that will interrogate asynapsis frequencies across chromosomes and across mice inheriting varying levels of CAST ancestry.

I also helped to design and breed large pedigrees between CAST and humanized B6 mice, and I generated H3K4me3 ChIP-seq data from F1 hybrid testes, complemented by DMC1 ChIP-seq data generated by Gang Zhang. We also sequenced genomic DNA from 11 F2 offspring and, by leveraging the high sequence divergence between B6 and CAST mice, Ran Li has processed these data to produce an initial map of several hundred recombination events, including crossovers and non-crossover gene conversions. By jointly analysing these datasets, we have shown that nearly all recombination events (which derive from both maternal and paternal meiosis) occur within testis DMC1/H3K4me3 peaks, confirming that PRDM9 controls nearly all hotspots and implying a lack of maternal-specific hotspots. Emmanuelle Bitoun has continued breeding these mice from the F2 generation to generate hundreds of (B6/Cast)F5 mice from which genomic DNA will be sequenced and processed in a similar manner to the F2 mice, generating the largest ever map of non-crossover events in any mammalian genome. From this dataset, we will be able to infer precise non-crossover tract lengths and learn about factors affecting the propensity for DSBs to resolve as crossovers versus non-crossovers.

5.4 A new function for the PRDM9 ZF array

Lastly, I have performed a set of experiments in our HEK293T expression system demonstrating that PRDM9 can bind to itself, forming multimers detected by co-

immunoprecipitation. Using various deletion constructs I discovered unexpectedly that this behaviour is mediated almost entirely by the zinc finger array, which is sufficient to produce multimers. Furthermore, by co-expressing the human and chimp alleles, to resemble a heterozygote, I show that homo-multimers form with several-fold greater efficiency than hetero-multimers, implying that PRDM9's ZF array may not only be responsible for mediating specific protein-DNA interactions but also specific protein-protein interactions. PRDM9's ability to form multimers was discovered independently and recently published by another group [128], but they did not reveal that this phenomenon is mediated by the ZF array or that heterodimers form less efficiently. They suggest that heterodimers in heterozygous mice may serve to exacerbate the effects of allelic dominance: when one allele binds DNA more strongly than the other, the weaker allele will tend to be sequestered away in heterodimers bound to the stronger allele's sites, and thus the weaker allele's sites will be bound less often. If heterodimers form less efficiently than homodimers, then this effect will be greatly weakened. Furthermore, given that multimerisation is mediated by the ZF array, it seems logical *a priori* that PRDM9 should only multimerise in a sequence-specific manner—otherwise, it could bind any of the many other Cys_2His_2 zinc finger proteins expressed in meiosis. Emmanuelle Bitoun is currently dissecting the ZF-ZF binding behaviour of PRDM9 with a series of *in vitro* experiments, to determine which zinc fingers and residues are responsible for mediating multimer formation. Furthermore, she is applying several biochemical techniques to investigate the size of PRDM9 complexes, to determine if they are dimers or larger multimers.

Speculatively, in light of our discoveries of the role of PRDM9 binding symmetry in DSB repair, synapsis, and fertility, PRDM9's multimerisation ability may provide the missing mechanistic link. If PRDM9 multimers can bind homologous chromosomes simultaneously, then they might be able to bring symmetric binding sites together prior to DSB formation, providing a degree of presynaptic homologue pairing and vastly reducing the homology search space. In an alternative model, the recombination machinery could hold onto PRDM9 after DSB formation, then utilise

its multimerisation ability to home in on the homologous binding site by binding to a PRDM9 protein already bound there. This ability would resemble the chromatin-looping behaviour of CTCF, another long Cys_2His_2 zinc finger protein that uses multiple ZF arrays to bind multiple sites simultaneously, looping the intervening chromatin and physically pairing domains such as enhancers and promoters (reviewed in [131]). If PRDM9 multimers can accomplish a similar effect, but between homologues instead of on the same homologue, this could explain how PRDM9 binding symmetry improves homology search, DSB repair, and synapsis.

Emmanuelle Bitoun is currently designing chromatin conformation capture experiments to test whether PRDM9 binding sites can physically associate with each other in transfected HEK293T cells. *In vivo* validation of this phenomenon will remain challenging, but it might be possible to visualise PRDM9-mediated pairing of binding sites using single-molecule imaging techniques in spermatocytes from transgenic mice with differentially tagged PRDM9 alleles.

5.5 Final remarks

The potential to reduce the search space for homology search may help to explain why recombination hotspots exist in the first place, despite the fact that they limit haplotypic diversity between hotspots and thus seem to contradict the fundamental function of recombination. Recombination hotspots have been observed in many diverse species, including species which lack PRDM9, such as yeast and dogs, in which hotspots occur at promoters [53, 63]. In PRDM9 knockout mice, DSBs localise to promoters [37], and synapsis and DSB repair are severely impaired, although, notably, some chromosomes are still able to synapse properly [99]. A recent pre-published paper has also identified a fertile human female with two non-functional copies of *PRDM9* [64]. In light of these findings, and in light of my observation of human PRDM9 binding at promoters, I propose that the role of PRDM9 is not simply to direct recombination away from promoters [37]. Rather, our observations in hybrid mice suggest that PRDM9 plays a direct role in improving the efficiency of homology search, DSB repair, and synapsis—making

hotspots even better at serving their biological purpose. Here, in collaboration with many others, I have presented a wide array of new evidence for this and other novel functions of PRDM9, adding to our growing understanding of this fascinating and important protein. Furthermore, by releasing our high-throughput sequencing data publicly, I will provide new tools and resources for others to contribute to the growing body of knowledge regarding PRDM9. Ongoing efforts to understand PRDM9 and recombination will continue to require diverse modalities of inquiry, from the fields of biochemistry, structural biology, cell and molecular biology, and experimental and statistical genetics. Because of its effects on fertility and genome evolution, recombination is subject to strong natural selection, and although many components of this process are highly conserved, divergent processes like hotspot localisation will require careful study in a variety of model species. To borrow the words of J.B.S. Haldane once again:

> *"Intense selection favours a variable response to the environment... Were this not so, the world would be much duller than is actually the case."* [132]

Appendices

A

Final PRDM9 construct sequences

1. Chimp allele (W11a_Annaclara_18ZNF) with N-terminal YFP (3903 bp)

   KozakSequence(6bp)-start(3bp)-YFP(714bp)-AgeI(6bp)-PRDM9_HumanB allele_first9exons(1140bp)-XbaI(6bp)-Chimp_W11a_ZFexon(2022bp)-stops (6bp)

```
GCCACCATGGTGAGCAAGGGCGAGGAGCTGTTCACCGGGGTGGTGCCCATCCTGGTCGAG
CTGGACGGCGACGTAAACGGCCACAAGTTCAGCGTGTCCGGCGAGGGCGAGGGCGATGCC
ACCTACGGCAAGCTGACCCTGAAGCTGATCTGCACCACCGGCAAGCTGCCCGTGCCCTGG
CCCACCCTCGTGACCACCCTGGGCTACGGCCTGCAGTGCTTCGCCCGCTACCCCGACCAC
ATGAAGCAGCACGACTTCTTCAAGTCCGCCATGCCCGAAGGCTACGTCCAGGAGCGCACC
ATCTTCTTCAAGGACGACGGCAACTACAAGACCCGCGCCGAGGTGAAGTTCGAGGGCGAC
ACCCTGGTGAACCGCATCGAGCTGAAGGGCATCGACTTCAAGGAGGACGGCAACATCCTG
GGGCACAAGCTGGAGTACAACTACAACAGCCACAACGTCTATATCACCGCCGACAAGCAG
AAGAACGGCATCAAGGCCAACTTCAAGATCCGCCACAACATCGAGGACGGCGGCGTGCAG
CTCGCCGACCACTACCAGCAGAACACCCCCATCGGCGACGGCCCCGTGCTGCTGCCCGAC
AACCACTACCTGAGCTACCAGTCCAAGCTGAGCAAAGACCCCAACGAGAAGCGCGATCAC
ATGGTCCTGCTGGAGTTCGTGACCGCCGCCGGGATCACTCTCGGCATGGACGAGCTGTAC
AAGACCGGTAGCCCTGAAAAGTCCCAAGAGGAGAGCCCAGAAGAAGACACAGAGAGAACA
GAGCGGAAGCCCATGGTCAAAGATGCCTTCAAAGACATTTCCATATACTTCACCAAGGAA
GAATGGGCAGAGATGGGAGACTGGGAGAAAACTCGCTATAGGAATGTGAAAAGGAACTAT
AATGCACTGATTACTATAGGTCTCAGAGCCACTCGACCAGCTTTCATGTGTCACCGAAGG
CAGGCCATCAAACTCCAGGTGGATGACACAGAAGATTCTGATGAAGAATGGACCCCTAGG
CAGCAAGTCAAACCTCCTTGGATGGCCTTAAGAGTGGAACAGCGTAAACACCAGAAGGGA
ATGCCCAAGGCGTCATTCAGTAATGAATCTAGTTTGAAAGAATTGTCAAGAACAGCAAAT
TTACTGAATGCAAGTGGATCAGAGCAGGCTCAGAAACCAGTGTCCCCTTCTGGAGAAGCA
AGTACCTCTGGACAGCACTCAAGACTAAAACTGGAACTCAGGAAGAAGGAGACTGAAAGA
```

AAGATGTATAGCCTGCGAGAAAGAAAGGGTCATGCATACAAAGAGGTCAGCGAGCCGCAG
GATGATGATTACCTCTATTGTGAGATGTGTCAGAACTTCTTCATTGACAGCTGTGCTGCC
CATGGGCCCCCTACATTTGTAAAGGACAGTGCAGTGGACAAGGGGCACCCCAACCGTTCA
GCCCTCAGTCTGCCCCCAGGGCTGAGAATTGGGCCATCAGGCATCCCTCAGGCTGGGCTT
GGAGTATGGAATGAGGCATCTGATCTGCCGCTGGGTCTGCACTTTGGCCCTTATGAGGGC
CGAATTACAGAAGACGAAGAGGCAGCCAACAATGGATACTCCTGGCTGATCACCAAGGGG
AGAAACTGCTATGAGTATGTGGATGGAAAAGATAAATCCTGGGCCAACTGGATGAGGTAT
GTGAACTGTGCCCGGGATGATGAAGAGCAGAACCTGGTGGCCTTCCAGTACCACAGGCAG
ATCTTCTATAGAACCTGCCGAGTCATTAGGCCAGGCTGTGAACTGCTGGTCTGGTATGGG
GATGAATACGGCCAGGAACTGGGCATCAAGTGGGGCAGCAAGTGGAAGAAAGAGCTCATG
GCAGGGAGATCTAGAGAGCCCAAGCCTGAAATCCACCCCTGTCCAAGTTGTTGCCTGGCC
TTTAGCAGCCAGAAGTTCCTGTCCCAGCACGTCGAGAGAAATCACAGCTCCCAGAACTTC
CCAGGACCCAGCGCAAGAAAGCTGCTGCAGCCCGAGAACCCTTGCCCAGGCGACCAGAAT
CAGGAGCAGCAGTACCCCGACCCTCGGTCCAGAAACGATAAGACCAAAGGCCAGGAAATC
AAGGAGAGGTCTAAACTGCTGAATAAGCGCACATGGCAGCGAGAGATTAGCCGGGCCTTC
TCTAGTCCCCCTAAAGGACAGATGGGCAGCTGCAGAGTGGGCAAGAGGATCATGGAGGAA
GAGTCCAGAACCGGGCAGAAAGTCAACCCCGGAAATACAGCCAAGCTGTTCGTGGGCGTC
GGGATCAGCAGGATTGCTAAGGTGAAATACGGGGAGTGCGGACAGGGCTTTAGCGTGAAA
TCCGATGTCATTACCCACCAGAGAACACATACTGGAGAAAAGCCATACGTGTGCCGCGAG
TGTGGGCGAGGATTCTCTTGGAAAAGTCACCTGCTGTCCCATCAGCGCACCCACACAGGC
GAAAAGCCCTACGTGTGCCGGGAGTGTGGCAGAGGGTTTAGCGTCAAATCAAGCCTGCTG
TCCCATCGGACCACACACACCGGGGAAAAGCCTTACGTGTGCAGGGAGTGTGGACGCGGC
TTCTCTGTCAAATCCTCTCTGCTGAGTCATCAGCGCACTCACACCGGAGAGAAACCATAC
GTGTGCCGAGAGTGTGGGCGGGGATTTTCACAGCAGAGCAATCTGCTGAGCCATCAGCGG
ACACACACTGGCGAAAAACCCTACGTGTGCAGAGAGTGTGGCAGGGGGTTCTCAGTCAAG
AGTTCACTGCTGAGCCATCAGAGAACCCACACAGGGGAAAAACCTTACGTGTGCCGAGAA
TGCGGACGAGGCTTTTCCCAGCAGTCTCATCTGCTGTCTCACCAGAGGACTCATACCGGA
GAAAAACCATATGTCTGCCGGGAGTGTGGGAGAGGATTCAGTCAGCAGTCACACCTGCTG
AGCCATCAGCGCACACACACTGGCGAGAAGCCTTACGTGTGCAGAGAATGCGGACGAGGG
TTTAGCCAGCAGTCCCATCTGCTGAGGCACCAGCGCACCCATACAGGGGAGAAACCCTAC
GTGTGCAGGGAATGCGGACGGGGCTTCTCCGTCAAAAGCTCCCTGCTGTCTCACCAGAGA
ACTCATACCGGAGAGAAGCCTTACGTGTGCCGCGAATGCGGGAGGGGCTTCAGCGTCAAA
TCTAGTCTGCTGTCACACAGGACTACCCATACCGGCGAGAAACCTTACGTGTGCAGAGAG
TGCGGACGAGGGTTCAGCGTGAAGTCAAGCCTGCTGTCCCACCAGCGGACACATACTGGG
GAGAAGCCATACGTGTGCAGGGAATGTGGCAGAGGCTTTTCTAAGCAGTCCCACCTGCTG
AGCCACCAGAGAACTCATACAGGCGAGAAACCGTACGTGTGCCGGGAATGTGGGCGCGGA
TTCTCACAGCAGAGCCATCTGCTGAGCCATCAGAGGACCCATACCGGCGAAAAACCATAC
GTGTGCAGAGAGTGTGGTCGAGGGTTTTCCCAGCAGTCTCACCTGCTGCGACACCAGAGA
ACACACACTGGCGAAAAGCCGTACGTGTGCAGAGAGTGTGGAAGGGGCTTCTCTGTGAAG
AGCTCCCTGCTGAGTCACCAGAGAACTCACACAGGCGAGAAGCCGTACGTGTGCAGAGAG
TGTGGGCGAGGATTTTCTGTCAAAAGTTCACTGCTGAGCCACCAGCGGACTCACACCGGC
GAGAAGCCCTACGTGTGCAGAGAATGTGAGAGGGGGTTCAGTCAGCAGTCCCACCTGCTG
CGCCACCAGAGGACCCATACTGGCGAAAAACCGTACGTGTGCAGAGAGTGTGGTAGAGGG
TTTAGTAGACAGAGCGCACTGCTGATTCACCAGAGGACCCATACAGGCGAAAAGCCATAA
TGA

2. YFP only (729 bp)

   KozakSequence(6bp)-start(3bp)-YFP(714bp)-stops(6bp)

   GCCACCATGGTGAGCAAGGGCGAGGAGCTGTTCACCGGGGTGGTGCCCATCCTGGTCGAG
   CTGGACGGCGACGTAAACGGCCACAAGTTCAGCGTGTCCGGCGAGGGCGAGGGCGATGCC
   ACCTACGGCAAGCTGACCCTGAAGCTGATCTGCACCACCGGCAAGCTGCCCGTGCCCTGG
   CCCACCCTCGTGACCACCCTGGGCTACGGCCTGCAGTGCTTCGCCCGCTACCCCGACCAC
   ATGAAGCAGCACGACTTCTTCAAGTCCGCCATGCCCGAAGGCTACGTCCAGGAGCGCACC
   ATCTTCTTCAAGGACGACGGCAACTACAAGACCCGCGCCGAGGTGAAGTTCGAGGGCGAC
   ACCCTGGTGAACCGCATCGAGCTGAAGGGCATCGACTTCAAGGAGGACGGCAACATCCTG
   GGGCACAAGCTGGAGTACAACTACAACAGCCACAACGTCTATATCACCGCCGACAAGCAG
   AAGAACGGCATCAAGGCCAACTTCAAGATCCGCCACAACATCGAGGACGGCGGCGTGCAG
   CTCGCCGACCACTACCAGCAGAACACCCCCATCGGCGACGGCCCCGTGCTGCTGCCCGAC
   AACCACTACCTGAGCTACCAGTCCAAGCTGAGCAAAGACCCCAACGAGAAGCGCGATCAC
   ATGGTCCTGCTGGAGTTCGTGACCGCCGCCGGGATCACTCTCGGCATGGACGAGCTGTAC
   AAGTAATGA

3. noZF with N-terminal YFP (2301 bp)

   KozakSequence(6bp)-start(3bp)-YFP(714bp)-AgeI(6bp) PRDM9_HumanB allele_toZFarray(1566bp)-stops(6bp)

   GCCACCATGGTGAGCAAGGGCGAGGAGCTGTTCACCGGGGTGGTGCCCATCCTGGTCGAG
   CTGGACGGCGACGTAAACGGCCACAAGTTCAGCGTGTCCGGCGAGGGCGAGGGCGATGCC
   ACCTACGGCAAGCTGACCCTGAAGCTGATCTGCACCACCGGCAAGCTGCCCGTGCCCTGG
   CCCACCCTCGTGACCACCCTGGGCTACGGCCTGCAGTGCTTCGCCCGCTACCCCGACCAC
   ATGAAGCAGCACGACTTCTTCAAGTCCGCCATGCCCGAAGGCTACGTCCAGGAGCGCACC
   ATCTTCTTCAAGGACGACGGCAACTACAAGACCCGCGCCGAGGTGAAGTTCGAGGGCGAC
   ACCCTGGTGAACCGCATCGAGCTGAAGGGCATCGACTTCAAGGAGGACGGCAACATCCTG
   GGGCACAAGCTGGAGTACAACTACAACAGCCACAACGTCTATATCACCGCCGACAAGCAG
   AAGAACGGCATCAAGGCCAACTTCAAGATCCGCCACAACATCGAGGACGGCGGCGTGCAG
   CTCGCCGACCACTACCAGCAGAACACCCCCATCGGCGACGGCCCCGTGCTGCTGCCCGAC
   AACCACTACCTGAGCTACCAGTCCAAGCTGAGCAAAGACCCCAACGAGAAGCGCGATCAC
   ATGGTCCTGCTGGAGTTCGTGACCGCCGCCGGGATCACTCTCGGCATGGACGAGCTGTAC
   AAGACCGGTAGCCCTGAAAAGTCCCAAGAGGAGAGCCCAGAAGAAGACACAGAGAGAACA
   GAGCGGAAGCCCATGGTCAAAGATGCCTTCAAAGACATTTCCATATACTTCACCAAGGAA
   GAATGGGCAGAGATGGGAGACTGGGAGAAAACTCGCTATAGGAATGTGAAAAGGAACTAT
   AATGCACTGATTACTATAGGTCTCAGAGCCACTCGACCAGCTTTCATGTGTCACCGAAGG
   CAGGCCATCAAACTCCAGGTGGATGACACAGAAGATTCTGATGAAGAATGGACCCCTAGG
   CAGCAAGTCAAACCTCCTTGGATGGCCTTAAGAGTGGAACAGCGTAAACACCAGAAGGGA
   ATGCCCAAGGCGTCATTCAGTAATGAATCTAGTTTGAAAGAATTGTCAAGAACAGCAAAT
   TTACTGAATGCAAGTGGATCAGAGCAGGCTCAGAAACCAGTGTCCCCTTCTGGAGAAGCA
   AGTACCTCTGGACAGCACTCAAGACTAAAACTGGAACTCAGGAAGAAGGAGACTGAAAGA
   AAGATGTATAGCCTGCGAGAAAGAAAGGGTCATGCATACAAAGAGGTCAGCGAGCCGCAG
   GATGATGATTACCTCTATTGTGAGATGTGTCAGAACTTCTTCATTGACAGCTGTGCTGCC
   CATGGGCCCCCTACATTTGTAAAGGACAGTGCAGTGGACAAGGGGCACCCCAACCGTTCA

GCCCTCAGTCTGCCCCCAGGGCTGAGAATTGGGCCATCAGGCATCCCTCAGGCTGGGCTT
GGAGTATGGAATGAGGCATCTGATCTGCCGCTGGGTCTGCACTTTGGCCCTTATGAGGGC
CGAATTACAGAAGACGAAGAGGCAGCCAACAATGGATACTCCTGGCTGATCACCAAGGGG
AGAAACTGCTATGAGTATGTGGATGGAAAAGATAAATCCTGGGCCAACTGGATGAGGTAT
GTGAACTGTGCCCGGGATGATGAAGAGCAGAACCTGGTGGCCTTCCAGTACCACAGGCAG
ATCTTCTATAGAACCTGCCGAGTCATTAGGCCAGGCTGTGAACTGCTGGTCTGGTATGGG
GATGAATACGGCCAGGAACTGGGCATCAAGTGGGGCAGCAAGTGGAAGAAAGAGCTCATG
GCAGGGAGAGAACCAAAGCCAGAGATCCATCCATGTCCCTCATGCTGTCTGGCCTTTTCA
AGTCAGAAATTTCTCAGTCAACATGTAGAACGCAATCACTCCTCTCAGAACTTCCCAGGA
CCATCTGCAAGAAAACTCCTCCAACCAGAGAATCCCTGCCCAGGGGATCAGAATCAGGAG
CAGCAATATCCAGATCCACACAGCCGTAATGACAAAACCAAAGGTCAAGAGATCAAAGAA
AGGTCCAAACTCTTGAATAAAAGGACATGGCAGAGGGAGATTTCAAGGGCCTTTTCTAGC
CCACCCAAAGGACAAATGGGGAGCTGTAGAGTGGGAAAAAGAATAATGGAAGAAGAGTCC
AGAACAGGCCAGAAAGTGAATCCAGGGAACACAGGCAAATTATTTGTGGGGGTAGGAATC
TCAAGAATTGCAAAATAATGA

4. Full-length human with N-terminal YFP (3414 bp)

   KozakSequence(6bp)-start(3bp)-YFP(714bp)-AgeI(6bp)-PRDM9_HumanB allele_full(2679bp)-stops(6bp)

GCCACCATGGTGAGCAAGGGCGAGGAGCTGTTCACCGGGGTGGTGCCCATCCTGGTCGAG
CTGGACGGCGACGTAAACGGCCACAAGTTCAGCGTGTCCGGCGAGGGCGAGGGCGATGCC
ACCTACGGCAAGCTGACCCTGAAGCTGATCTGCACCACCGGCAAGCTGCCCGTGCCCTGG
CCCACCCTCGTGACCACCCTGGGCTACGGCCTGCAGTGCTTCGCCCGCTACCCCGACCAC
ATGAAGCAGCACGACTTCTTCAAGTCCGCCATGCCCGAAGGCTACGTCCAGGAGCGCACC
ATCTTCTTCAAGGACGACGGCAACTACAAGACCCGCGCCGAGGTGAAGTTCGAGGGCGAC
ACCCTGGTGAACCGCATCGAGCTGAAGGGCATCGACTTCAAGGAGGACGGCAACATCCTG
GGGCACAAGCTGGAGTACAACTACAACAGCCACAACGTCTATATCACCGCCGACAAGCAG
AAGAACGGCATCAAGGCCAACTTCAAGATCCGCCACAACATCGAGGACGGCGGCGTGCAG
CTCGCCGACCACTACCAGCAGAACACCCCCATCGGCGACGGCCCCGTGCTGCTGCCCGAC
AACCACTACCTGAGCTACCAGTCCAAGCTGAGCAAAGACCCCAACGAGAAGCGCGATCAC
ATGGTCCTGCTGGAGTTCGTGACCGCCGCCGGGATCACTCTCGGCATGGACGAGCTGTAC
AAGACCGGTAGCCCTGAAAAGTCCCAAGAGGAGAGCCCAGAAGAAGACACAGAGAGAACA
GAGCGGAAGCCCATGGTCAAAGATGCCTTCAAAGACATTTCCATATACTTCACCAAGGAA
GAATGGGCAGAGATGGGAGACTGGGAGAAAACTCGCTATAGGAATGTGAAAAGGAACTAT
AATGCACTGATTACTATAGGTCTCAGAGCCACTCGACCAGCTTTCATGTGTCACCGAAGG
CAGGCCATCAAACTCCAGGTGGATGACACAGAAGATTCTGATGAAGAATGGACCCCTAGG
CAGCAAGTCAAACCTCCTTGGATGGCCTTAAGAGTGGAACAGCGTAAACACCAGAAGGGA
ATGCCCAAGGCGTCATTCAGTAATGAATCTAGTTTGAAAGAATTGTCAAGAACAGCAAAT
TTACTGAATGCAAGTGGATCAGAGCAGGCTCAGAAACCAGTGTCCCCTTCTGGAGAAGCA
AGTACCTCTGGACAGCACTCAAGACTAAAACTGGAACTCAGGAAGAAGGAGACTGAAAGA
AAGATGTATAGCCTGCGAGAAAGAAAGGGTCATGCATACAAAGAGGTCAGCGAGCCGCAG
GATGATGATTACCTCTATTGTGAGATGTGTCAGAACTTCTTCATTGACAGCTGTGCTGCC
CATGGGCCCCCTACATTTGTAAAGGACAGTGCAGTGGACAAGGGGCACCCCAACCGTTCA
GCCCTCAGTCTGCCCCCAGGGCTGAGAATTGGGCCATCAGGCATCCCTCAGGCTGGGCTT
GGAGTATGGAATGAGGCATCTGATCTGCCGCTGGGTCTGCACTTTGGCCCTTATGAGGGC

CGAATTACAGAAGACGAAGAGGCAGCCAACAATGGATACTCCTGGCTGATCACCAAGGGG
AGAAACTGCTATGAGTATGTGGATGGAAAAGATAAATCCTGGGCCAACTGGATGAGGTAT
GTGAACTGTGCCCGGGATGATGAAGAGCAGAACCTGGTGGCCTTCCAGTACCACAGGCAG
ATCTTCTATAGAACCTGCCGAGTCATTAGGCCAGGCTGTGAACTGCTGGTCTGGTATGGG
GATGAATACGGCCAGGAACTGGGCATCAAGTGGGGCAGCAAGTGGAAGAAAGAGCTCATG
GCAGGGAGAGAACCAAAGCCAGAGATCCATCCATGTCCCTCATGCTGTCTGGCCTTTTCA
AGTCAGAAATTTCTCAGTCAACATGTAGAACGCAATCACTCCTCTCAGAACTTCCCAGGA
CCATCTGCAAGAAAACTCCTCCAACCAGAGAATCCCTGCCCAGGGGATCAGAATCAGGAG
CAGCAATATCCAGATCCACACAGCCGTAATGACAAAACCAAAGGTCAAGAGATCAAAGAA
AGGTCCAAACTCTTGAATAAAAGGACATGGCAGAGGGAGATTTCAAGGGCCTTTTCTAGC
CCACCCAAAGGACAAATGGGGAGCTGTAGAGTGGGAAAAAGAATAATGGAAGAAGAGTCC
AGAACAGGCCAGAAAGTGAATCCAGGGAACACAGGCAAATTATTTGTGGGGGTAGGAATC
TCAAGAATTGCAAAAGTCAAGTATGGAGAGTGTGGACAAGGTTTCAGTGTTAAATCAGAT
GTTATTACACACCAAAGGACACATACAGGGGAGAAGCTCTACGTCTGCAGGGAGTGTGGG
CGGGGCTTCAGCTGGAAGTCCCACCTGCTGATCCACCAGAGAATTCACACCGGCGAGAAA
CCCTACGTGTGCAGAGAATGTGGGAGGGGATTTTCTTGGCAGTCTGTGCTGCTGACACAC
CAGCGCACACACACCGGCGAGAAGCCTTATGTCTGTAGAGAATGTGGGCGGGGATTCTCA
AGACAGAGCGTCCTGCTGACCCACCAGAGGCGCCACACAGGCGAGAAACCATATGTCTGC
AGGGAGTGTGGGCGGGGATTTTCCAGACAGTCTGTGCTGCTGACTCATCAGAGAAGACAC
ACAGGCGAGAAGCCCTATGTCTGCAGGGAATGCGGCCGGGGATTCAGTTGGCAGTCCGTC
CTGCTGACACACCAGAGAACTCATACAGGCGAGAAACCTTACGTGTGCAGGGAATGTGGG
CGCGGCTTCAGCTGGCAGTCCGTGCTGCTGACCCATCAGAGAACCCACACAGGCGAGAAG
CCATATGTGTGCCGGGAATGTGGGAGAGGATTCTCTAACAAATCTCATCTGCTGAGGCAT
CAGAGAACCCATACCGGCGAGAAGCCTTACGTCTGCAGGGAATGCGGGCGGGGATTCAGA
GACAAATCTCACCTGCTGAGGCACCAGCGCACTCATACAGGCGAGAAGCCTTATGTGTGT
AGGGAATGTGGGAGGGGATTCCGCGATAAAAGCAATCTGCTGTCCCACCAGCGGACTCAT
ACCGGCGAGAAGCCATACGTCTGTAGGGAGTGCGGCAGGGGATTTTCTAACAAATCCCAT
CTGCTGCGCCACCAGCGGACCCATACTGGCGAGAAGCCCTATGTGTGTCGCGAGTGCGGG
AGGGGATTCCGCAACAAGAGCCACCTGCTGCGGCACCAGAGAACTCACACCGGAGAAAAG
CCTTACGTGTGCAGGGAGTGTGGGCGCGGCTTTAGCGACAGGTCCTCTCTGTGCTATCAT
CAGAGGACTCACACCGGAGAGAAGCCCTACGTGTGCCGGGAGGATGAGTAATGA

5. Full-length human with C-terminal YFP (3465 bp)

   KozakSequence(6bp)-start(3bp)-PRDM9_HumanBallele_full(2679bp) -KpnI(6bp)-linker(45bp)-AgeI(6bp)-YFP(714bp)-stops(6bp)

GCCACCATGAGCCCTGAAAAGTCCCAAGAGGAGAGCCCAGAAGAAGACACAGAGAGAACA
GAGCGGAAGCCCATGGTCAAAGATGCCTTCAAAGACATTTCCATATACTTCACCAAGGAA
GAATGGGCAGAGATGGGAGACTGGGAGAAAACTCGCTATAGGAATGTGAAAAGGAACTAT
AATGCACTGATTACTATAGGTCTCAGAGCCACTCGACCAGCTTTCATGTGTCACCGAAGG
CAGGCCATCAAACTCCAGGTGGATGACACAGAAGATTCTGATGAAGAATGGACCCCTAGG
CAGCAAGTCAAACCTCCTTGGATGGCCTTAAGAGTGGAACAGCGTAAACACCAGAAGGGA
ATGCCCAAGGCGTCATTCAGTAATGAATCTAGTTTGAAAGAATTGTCAAGAACAGCAAAT
TTACTGAATGCAAGTGGCTCAGAGCAGGCTCAGAAACCAGTGTCCCCTTCTGGAGAAGCA
AGTACCTCTGGACAGCACTCAAGACTAAAACTGGAACTCAGGAAGAAGGAGACTGAAAGA
AAGATGTATAGCCTGCGAGAAAGAAAGGGTCATGCATACAAAGAGGTCAGCGAGCCGCAG

GATGATGATTACCTCTATTGTGAGATGTGTCAGAACTTCTTCATTGACAGCTGTGCTGCC
CATGGGCCCCCTACATTTGTAAAGGACAGTGCAGTGGACAAGGGGCACCCCAACCGTTCA
GCCCTCAGTCTGCCCCCAGGGCTGAGAATTGGGCCATCAGGCATCCCTCAGGCTGGGCTT
GGAGTATGGAATGAGGCATCTGATCTGCCGCTGGGTCTGCACTTTGGCCCTTATGAGGGC
CGAATTACAGAAGACGAAGAGGCAGCCAACAATGGATACTCCTGGCTGATCACCAAGGGG
AGAAACTGCTATGAGTATGTGGATGGAAAAGATAAATCCTGGGCCAACTGGATGAGGTAT
GTGAACTGTGCCCGGGATGATGAAGAGCAGAACCTGGTGGCCTTCCAGTACCACAGGCAG
ATCTTCTATAGAACCTGCCGAGTCATTAGGCCAGGCTGTGAACTGCTGGTCTGGTATGGG
GATGAATACGGCCAGGAACTGGGCATCAAGTGGGGCAGCAAGTGGAAGAAAGAGCTCATG
GCAGGGAGAGAACCAAAGCCAGAGATCCATCCATGTCCCTCATGCTGTCTGGCCTTTTCA
AGTCAGAAATTTCTCAGTCAACATGTAGAACGCAATCACTCCTCTCAGAACTTCCCAGGA
CCATCTGCAAGAAAACTCCTCCAACCAGAGAATCCCTGCCCAGGGGATCAGAATCAGGAG
CAGCAATATCCAGATCCACACAGCCGTAATGACAAAACCAAAGGTCAAGAGATCAAAGAA
AGGTCCAAACTCTTGAATAAAAGGACATGGCAGAGGGAGATTTCAAGGGCCTTTTCTAGC
CCACCCAAAGGACAAATGGGGAGCTGTAGAGTGGGAAAAAGAATAATGGAAGAAGAGTCC
AGAACAGGCCAGAAAGTGAATCCAGGGAACACAGGCAAATTATTTGTGGGGGTAGGAATC
TCAAGAATTGCAAAAGTCAAGTATGGAGAGTGTGGACAAGGTTTCAGTGTTAAATCAGAT
GTTATTACACACCAAAGGACACATACAGGGGAGAAGCTCTACGTCTGCAGGGAGTGTGGG
CGGGGCTTCAGCTGGAAGTCCCACCTGCTGATCCACCAGAGAATTCACACCGGCGAGAAA
CCCTACGTGTGCAGAGAATGTGGGAGGGGATTTTCTTGGCAGTCTGTGCTGCTGACACAC
CAGCGCACACACACCGGCGAGAAGCCTTATGTCTGTAGAGAATGTGGGCGGGGATTCTCA
AGACAGAGCGTCCTGCTGACCCACCAGAGGCGCCACACAGGCGAGAAACCATATGTCTGC
AGGGAGTGTGGGCGGGGATTTTCCAGACAGTCTGTGCTGCTGACTCATCAGAGAAGACAC
ACAGGCGAGAAGCCCTATGTCTGCAGGGAATGCGGCCGGGGATTCAGTTGGCAGTCCGTC
CTGCTGACACACCAGAGAACTCATACAGGCGAGAAACCTTACGTGTGCAGGGAATGTGGG
CGCGGCTTCAGCTGGCAGTCCGTGCTGCTGACCCATCAGAGAACCCACACAGGCGAGAAG
CCATATGTGTGCCGGGAATGTGGGAGAGGATTCTCTAACAAATCTCATCTGCTGAGGCAT
CAGAGAACCCATACCGGCGAGAAGCCTTACGTCTGCAGGGAATGCGGGCGGGGATTCAGA
GACAAATCTCACCTGCTGAGGCACCAGCGCACTCATACAGGCGAGAAGCCTTATGTGTGT
AGGGAATGTGGGAGGGGATTCCGCGATAAAAGCAATCTGCTGTCCCACCAGCGGACTCAT
ACCGGCGAGAAGCCATACGTCTGTAGGGAGTGCGGCAGGGGATTTTCTAACAAATCCCAT
CTGCTGCGCCACCAGCGGACCCATACTGGCGAGAAGCCCTATGTGTGTCGCGAGTGCGGG
AGGGGATTCCGCAACAAGAGCCACCTGCTGCGGCACCAGAGAACTCACACCGGAGAAAAG
CCTTACGTGTGCAGGGAGTGTGGGCGCGGCTTTAGCGACAGGTCCTCTCTGTGCTATCAT
CAGAGGACTCACACCGGAGAGAAGCCCTACGTGTGCCGGGAGGATGAGGGTACCGGAGGT
GGAGGTAGTGGTGGAGGTGGAAGCGGAGGTGGAGGTAGTACCGGTGTGAGCAAGGGCGAG
GAGCTGTTCACCGGGGTGGTGCCCATCCTGGTCGAGCTGGACGGCGACGTAAACGGCCAC
AAGTTCAGCGTGTCCGGCGAGGGCGAGGGCGATGCCACCTACGGCAAGCTGACCCTGAAG
CTGATCTGCACCACCGGCAAGCTGCCCGTGCCCTGGCCCACCCTCGTGACCACCCTGGGC
TACGGCCTGCAGTGCTTCGCCCGCTACCCCGACCACATGAAGCAGCACGACTTCTTCAAG
TCCGCCATGCCCGAAGGCTACGTCCAGGAGCGCACCATCTTCTTCAAGGACGACGGCAAC
TACAAGACCCGCGCCGAGGTGAAGTTCGAGGGCGACACCCTGGTGAACCGCATCGAGCTG
AAGGGCATCGACTTCAAGGAGGACGGCAACATCCTGGGGCACAAGCTGGAGTACAACTAC
AACAGCCACAACGTCTATATCACCGCCGACAAGCAGAAGAACGGCATCAAGGCCAACTTC
AAGATCCGCCACAACATCGAGGACGGCGGCGTGCAGCTCGCCGACCACTACCAGCAGAAC

ACCCCCATCGGCGACGGCCCCGTGCTGCTGCCCGACAACCACTACCTGAGCTACCAGTCC
AAGCTGAGCAAAGACCCCAACGAGAAGCGCGATCACATGGTCCTGCTGGAGTTCGTGACC
GCCGCCGGGATCACTCTCGGCATGGACGAGCTGTACAAGTAATGA

6. Full-length human with C-terminal TwinStrep-HA-P2A-YFP tag (3618 bp)

KozakSequence(6bp)-start(3bp)-PRDM9_HumanBallele_full(2679bp)
-KpnI(6bp)-HAlinker(198bp)-AgeI(6bp)-YFP(714bp)-stops(6bp)
HAlinker is: -shortlinker(6bp)-StrepII(24bp)-linker(36bp)-StrepII(24bp)-
shortlinker(15bp)-HA(27bp)-shortlinker(9bp)-P2A(57bp)-

GCCACCATGAGCCCTGAAAAGTCCCAAGAGGAGAGCCCAGAAGAAGACACAGAGAGAACA
GAGCGGAAGCCCATGGTCAAAGATGCCTTCAAAGACATTTCCATATACTTCACCAAGGAA
GAATGGGCAGAGATGGGAGACTGGGAGAAAACTCGCTATAGGAATGTGAAAAGGAACTAT
AATGCACTGATTACTATAGGTCTCAGAGCCACTCGACCAGCTTTCATGTGTCACCGAAGG
CAGGCCATCAAACTCCAGGTGGATGACACAGAAGATTCTGATGAAGAATGGACCCCTAGG
CAGCAAGTCAAACCTCCTTGGATGGCCTTAAGAGTGGAACAGCGTAAACACCAGAAGGGA
ATGCCCAAGGCGTCATTCAGTAATGAATCTAGTTTGAAAGAATTGTCAAGAACAGCAAAT
TTACTGAATGCAAGTGGCTCAGAGCAGGCTCAGAAACCAGTGTCCCCTTCTGGAGAAGCA
AGTACCTCTGGACAGCACTCAAGACTAAAACTGGAACTCAGGAAGAAGGAGACTGAAAGA
AAGATGTATAGCCTGCGAGAAAGAAAGGGTCATGCATACAAAGAGGTCAGCGAGCCGCAG
GATGATGATTACCTCTATTGTGAGATGTGTCAGAACTTCTTCATTGACAGCTGTGCTGCC
CATGGGCCCCCTACATTTGTAAAGGACAGTGCAGTGGACAAGGGGCACCCCAACCGTTCA
GCCCTCAGTCTGCCCCCAGGGCTGAGAATTGGGCCATCAGGCATCCCTCAGGCTGGGCTT
GGAGTATGGAATGAGGCATCTGATCTGCCGCTGGGTCTGCACTTTGGCCCTTATGAGGGC
CGAATTACAGAAGACGAAGAGGCAGCCAACAATGGATACTCCTGGCTGATCACCAAGGGG
AGAAACTGCTATGAGTATGTGGATGGAAAAGATAAATCCTGGGCCAACTGGATGAGGTAT
GTGAACTGTGCCCGGGATGATGAAGAGCAGAACCTGGTGGCCTTCCAGTACCACAGGCAG
ATCTTCTATAGAACCTGCCGAGTCATTAGGCCAGGCTGTGAACTGCTGGTCTGGTATGGG
GATGAATACGGCCAGGAACTGGGCATCAAGTGGGGCAGCAAGTGGAAGAAAGAGCTCATG
GCAGGGAGAGAACCAAAGCCAGAGATCCATCCATGTCCCTCATGCTGTCTGGCCTTTTCA
AGTCAGAAATTTCTCAGTCAACATGTAGAACGCAATCACTCCTCTCAGAACTTCCCAGGA
CCATCTGCAAGAAAACTCCTCCAACCAGAGAATCCCTGCCCAGGGGATCAGAATCAGGAG
CAGCAATATCCAGATCCACACAGCCGTAATGACAAAACCAAAGGTCAAGAGATCAAAGAA
AGGTCCAAACTCTTGAATAAAAGGACATGGCAGAGGGAGATTTCAAGGGCCTTTTCTAGC
CCACCCAAAGGACAAATGGGGAGCTGTAGAGTGGGAAAAAGAATAATGGAAGAAGAGTCC
AGAACAGGCCAGAAAGTGAATCCAGGGAACACAGGCAAATTATTTGTGGGGGTAGGAATC
TCAAGAATTGCAAAAGTCAAGTATGGAGAGTGTGGACAAGGTTTCAGTGTTAAATCAGAT
GTTATTACACACCAAAGGACACATACAGGGGAGAAGCTCTACGTCTGCAGGGAGTGTGGG
CGGGGCTTCAGCTGGAAGTCCCACCTGCTGATCCACCAGAGAATTCACACCGGCGAGAAA
CCCTACGTGTGCAGAGAATGTGGGAGGGGATTTTCTTGGCAGTCTGTGCTGCTGACACAC
CAGCGCACACACACCGGCGAGAAGCCTTATGTCTGTAGAGAATGTGGGCGGGGATTCTCA
AGACAGAGCGTCCTGCTGACCCACCAGAGGCGCCACACAGGCGAGAAACCATATGTCTGC
AGGGAGTGTGGGCGGGGATTTTCCAGACAGTCTGTGCTGCTGACTCATCAGAGAAGACAC
ACAGGCGAGAAGCCCTATGTCTGCAGGGAATGCGGCCGGGGATTCAGTTGGCAGTCCGTC
CTGCTGACACACCAGAGAACTCATACAGGCGAGAAACCTTACGTGTGCAGGGAATGTGGG
CGCGGCTTCAGCTGGCAGTCCGTGCTGCTGACCCATCAGAGAACCCACACAGGCGAGAAG

```
CCATATGTGTGCCGGGAATGTGGGAGAGGATTCTCTAACAAATCTCATCTGCTGAGGCAT
CAGAGAACCCATACCGGCGAGAAGCCTTACGTCTGCAGGGAATGCGGGCGGGGATTCAGA
GACAAATCTCACCTGCTGAGGCACCAGCGCACTCATACAGGCGAGAAGCCTTATGTGTGT
AGGGAATGTGGGAGGGGATTCCGCGATAAAAGCAATCTGCTGTCCCACCAGCGGACTCAT
ACCGGCGAGAAGCCATACGTCTGTAGGGAGTGCGGCAGGGGATTTTCTAACAAATCCCAT
CTGCTGCGCCACCAGCGGACCCATACTGGCGAGAAGCCCTATGTGTGTCGCGAGTGCGGG
AGGGGATTCCGCAACAAGAGCCACCTGCTGCGGCACCAGAGAACTCACACCGGAGAAAAG
CCTTACGTGTGCAGGGAGTGTGGGCGCGGCTTTAGCGACAGGTCCTCTCTGTGCTATCAT
CAGAGGACTCACACCGGAGAGAAGCCCTACGTGTGCCGGGAGGATGAGGGTACCAGCGCC
TGGTCCCACCCCCAGTTTGAAAAGGGCGGAGGATCTGGCGGCGGAAGTGGCGGATCTGCT
TGGAGCCACCCTCAGTTTGAAAAGGGGGCCAGCGGCGAGTACCCCTACGACGTGCCAGAT
TACGCTGGCAGCGGCGCCACCAACTTCAGCCTGCTGAAACAGGCCGGCGACGTGGAAGAG
AACCCTGGCCCTACCGGTGTGAGCAAGGGCGAGGAGCTGTTCACCGGGGTGGTGCCCATC
CTGGTCGAGCTGGACGGCGACGTAAACGGCCACAAGTTCAGCGTGTCCGGCGAGGGCGAG
GGCGATGCCACCTACGGCAAGCTGACCCTGAAGCTGATCTGCACCACCGGCAAGCTGCCC
GTGCCCTGGCCCACCCTCGTGACCACCCTGGGCTACGGCCTGCAGTGCTTCGCCCGCTAC
CCCGACCACATGAAGCAGCACGACTTCTTCAAGTCCGCCATGCCCGAAGGCTACGTCCAG
GAGCGCACCATCTTCTTCAAGGACGACGGCAACTACAAGACCCGCGCCGAGGTGAAGTTC
GAGGGCGACACCCTGGTGAACCGCATCGAGCTGAAGGGCATCGACTTCAAGGAGGACGGC
AACATCCTGGGGCACAAGCTGGAGTACAACTACAACAGCCACAACGTCTATATCACCGCC
GACAAGCAGAAGAACGGCATCAAGGCCAACTTCAAGATCCGCCACAACATCGAGGACGGC
GGCGTGCAGCTCGCCGACCACTACCAGCAGAACACCCCCATCGGCGACGGCCCCGTGCTG
CTGCCCGACAACCACTACCTGAGCTACCAGTCCAAGCTGAGCAAAGACCCCAACGAGAAG
CGCGATCACATGGTCCTGCTGGAGTTCGTGACCGCCGCCGGGATCACTCTCGGCATGGAC
GAGCTGTACAAGTAATGA
```

7. Full-length human with C-terminal TwinStrep-V5-P2A-YFP tag (3633 bp)

KozakSequence(6bp)-start(3bp)-PRDM9_HumanBallele_full(2679bp)
-KpnI(6bp)-V5linker(213bp)-AgeI(6bp)-YFP(714bp)-stops(6bp)
V5linker is: -shortlinker(6bp)-StrepII(24bp)-linker(36bp)-StrepII(24bp)
-shortlinker(15bp)-V5(42bp)-shortlinker(9bp)-P2A(57bp)-

```
GCCACCATGAGCCCTGAAAAGTCCCAAGAGGAGAGCCCAGAAGAAGACACAGAGAGAACA
GAGCGGAAGCCCATGGTCAAAGATGCCTTCAAAGACATTTCCATATACTTCACCAAGGAA
GAATGGGCAGAGATGGGAGACTGGGAGAAAACTCGCTATAGGAATGTGAAAAGGAACTAT
AATGCACTGATTACTATAGGTCTCAGAGCCACTCGACCAGCTTTCATGTGTCACCGAAGG
CAGGCCATCAAACTCCAGGTGGATGACACAGAAGATTCTGATGAAGAATGGACCCCTAGG
CAGCAAGTCAAACCTCCTTGGATGGCCTTAAGAGTGGAACAGCGTAAACACCAGAAGGGA
ATGCCCAAGGCGTCATTCAGTAATGAATCTAGTTTGAAAGAATTGTCAAGAACAGCAAAT
TTACTGAATGCAAGTGGCTCAGAGCAGGCTCAGAAACCAGTGTCCCCTTCTGGAGAAGCA
AGTACCTCTGGACAGCACTCAAGACTAAAACTGGAACTCAGGAAGAAGGAGACTGAAAGA
AAGATGTATAGCCTGCGAGAAAGAAAGGGTCATGCATACAAAGAGGTCAGCGAGCCGCAG
GATGATGATTACCTCTATTGTGAGATGTGTCAGAACTTCTTCATTGACAGCTGTGCTGCC
CATGGGCCCCCTACATTTGTAAAGGACAGTGCAGTGGACAAGGGGCACCCCAACCGTTCA
GCCCTCAGTCTGCCCCCAGGGCTGAGAATTGGGCCATCAGGCATCCCTCAGGCTGGGCTT
GGAGTATGGAATGAGGCATCTGATCTGCCGCTGGGTCTGCACTTTGGCCCTTATGAGGGC
```

CGAATTACAGAAGACGAAGAGGCAGCCAACAATGGATACTCCTGGCTGATCACCAAGGGG
AGAAACTGCTATGAGTATGTGGATGGAAAAGATAAATCCTGGGCCAACTGGATGAGGTAT
GTGAACTGTGCCCGGGATGATGAAGAGCAGAACCTGGTGGCCTTCCAGTACCACAGGCAG
ATCTTCTATAGAACCTGCCGAGTCATTAGGCCAGGCTGTGAACTGCTGGTCTGGTATGGG
GATGAATACGGCCAGGAACTGGGCATCAAGTGGGGCAGCAAGTGGAAGAAAGAGCTCATG
GCAGGGAGAGAACCAAAGCCAGAGATCCATCCATGTCCCTCATGCTGTCTGGCCTTTTCA
AGTCAGAAATTTCTCAGTCAACATGTAGAACGCAATCACTCCTCTCAGAACTTCCCAGGA
CCATCTGCAAGAAAACTCCTCCAACCAGAGAATCCCTGCCCAGGGGATCAGAATCAGGAG
CAGCAATATCCAGATCCACACAGCCGTAATGACAAAACCAAAGGTCAAGAGATCAAAGAA
AGGTCCAAACTCTTGAATAAAAGGACATGGCAGAGGGAGATTTCAAGGGCCTTTTCTAGC
CCACCCAAAGGACAAATGGGGAGCTGTAGAGTGGGAAAAAGAATAATGGAAGAAGAGTCC
AGAACAGGCCAGAAAGTGAATCCAGGGAACACAGGCAAATTATTTGTGGGGGTAGGAATC
TCAAGAATTGCAAAAGTCAAGTATGGAGAGTGTGGACAAGGTTTCAGTGTTAAATCAGAT
GTTATTACACACCAAAGGACACATACAGGGGAGAAGCTCTACGTCTGCAGGGAGTGTGGG
CGGGGCTTCAGCTGGAAGTCCCACCTGCTGATCCACCAGAGAATTCACACCGGCGAGAAA
CCCTACGTGTGCAGAGAATGTGGGAGGGGATTTTCTTGGCAGTCTGTGCTGCTGACACAC
CAGCGCACACACACCGGCGAGAAGCCTTATGTCTGTAGAGAATGTGGGCGGGGATTCTCA
AGACAGAGCGTCCTGCTGACCCACCAGAGGCGCCACACAGGCGAGAAACCATATGTCTGC
AGGGAGTGTGGGCGGGGATTTTCCAGACAGTCTGTGCTGCTGACTCATCAGAGAAGACAC
ACAGGCGAGAAGCCCTATGTCTGCAGGGAATGCGGCCGGGGATTCAGTTGGCAGTCCGTC
CTGCTGACACACCAGAGAACTCATACAGGCGAGAAACCTTACGTGTGCAGGGAATGTGGG
CGCGGCTTCAGCTGGCAGTCCGTGCTGCTGACCCATCAGAGAACCCACACAGGCGAGAAG
CCATATGTGTGCCGGGAATGTGGGAGAGGATTCTCTAACAAATCTCATCTGCTGAGGCAT
CAGAGAACCCATACCGGCGAGAAGCCTTACGTCTGCAGGGAATGCGGGCGGGGATTCAGA
GACAAATCTCACCTGCTGAGGCACCAGCGCACTCATACAGGCGAGAAGCCTTATGTGTGT
AGGGAATGTGGGAGGGGATTCCGCGATAAAAGCAATCTGCTGTCCCACCAGCGGACTCAT
ACCGGCGAGAAGCCATACGTCTGTAGGGAGTGCGGCAGGGGATTTTCTAACAAATCCCAT
CTGCTGCGCCACCAGCGGACCCATACTGGCGAGAAGCCCTATGTGTGTCGCGAGTGCGGG
AGGGGATTCCGCAACAAGAGCCACCTGCTGCGGCACCAGAGAACTCACACCGGAGAAAAG
CCTTACGTGTGCAGGGAGTGTGGGCGCGGCTTTAGCGACAGGTCCTCTCTGTGCTATCAT
CAGAGGACTCACACCGGAGAGAAGCCCTACGTGTGCCGGGAGGATGAGGGTACCAGCGCC
TGGTCCCACCCCCAGTTCGAGAAGGGCGGAGGATCTGGCGGCGGAAGTGGCGGATCTGCT
TGGAGCCACCCTCAGTTTGAAAAGGGGGCCAGCGGCGAGGGCAAGCCCATCCCTAATCCT
CTGCTGGGCCTGGACAGCACAGGCAGCGGCGCTACAAACTTCAGCCTGCTGAAGCAGGCC
GGCGACGTGGAAGAGAACCCTGGACCTACCGGTGTGAGCAAGGGCGAGGAGCTGTTCACC
GGGGTGGTGCCCATCCTGGTCGAGCTGGACGGCGACGTAAACGGCCACAAGTTCAGCGTG
TCCGGCGAGGGCGAGGGCGATGCCACCTACGGCAAGCTGACCCTGAAGCTGATCTGCACC
ACCGGCAAGCTGCCCGTGCCCTGGCCCACCCTCGTGACCACCCTGGGCTACGGCCTGCAG
TGCTTCGCCCGCTACCCCGACCACATGAAGCAGCACGACTTCTTCAAGTCCGCCATGCCC
GAAGGCTACGTCCAGGAGCGCACCATCTTCTTCAAGGACGACGGCAACTACAAGACCCGC
GCCGAGGTGAAGTTCGAGGGCGACACCCTGGTGAACCGCATCGAGCTGAAGGGCATCGAC
TTCAAGGAGGACGGCAACATCCTGGGGCACAAGCTGGAGTACAACTACAACAGCCACAAC
GTCTATATCACCGCCGACAAGCAGAAGAACGGCATCAAGGCCAACTTCAAGATCCGCCAC
AACATCGAGGACGGCGGCGTGCAGCTCGCCGACCACTACCAGCAGAACACCCCCATCGGC
GACGGCCCCGTGCTGCTGCCCGACAACCACTACCTGAGCTACCAGTCCAAGCTGAGCAAA

GACCCCAACGAGAAGCGCGATCACATGGTCCTGCTGGAGTTCGTGACCGCCGCCGGGATC
ACTCTCGGCATGGACGAGCTGTACAAGTAATGA

8. hZFonly with C-terminal TwinStrep-HA-P2A-YFP tag (2052 bp)

KozakSequence(6bp)-start(3bp)-PRDM9_HumanBallele_ZFarrayOnly (1113bp)-KpnI(6bp)-HAlinker(198bp)-AgeI(6bp)-YFP(714bp)-stops(6bp) HAlinker is: -shortlinker(6bp)-StrepII(24bp)-linker(36bp)-StrepII(24bp) -shortlinker(15bp)-HA(27bp)-shortlinker(9bp)-P2A(57bp)-

GCCACCATGGTCAAGTATGGAGAGTGTGGACAAGGTTTCAGTGTTAAATCAGATGTTATT
ACACACCAAAGGACACATACAGGGGAGAAGCTCTACGTCTGCAGGGAGTGTGGGCGGGGC
TTCAGCTGGAAGTCCCACCTGCTGATCCACCAGAGAATTCACACCGGCGAGAAACCCTAC
GTGTGCAGAGAATGTGGGAGGGGATTTTCTTGGCAGTCTGTGCTGCTGACACACCAGCGC
ACACACACCGGCGAGAAGCCTTATGTCTGTAGAGAATGTGGGCGGGGATTCTCAAGACAG
AGCGTCCTGCTGACCCACCAGAGGCGCCACACAGGCGAGAAACCATATGTCTGCAGGGAG
TGTGGGCGGGGATTTTCCAGACAGTCTGTGCTGCTGACTCATCAGAGAAGACACACAGGC
GAGAAGCCCTATGTCTGCAGGGAATGCGGCCGGGGATTCAGTTGGCAGTCCGTCCTGCTG
ACACACCAGAGAACTCATACAGGCGAGAAACCTTACGTGTGCAGGGAATGTGGGCGCGGC
TTCAGCTGGCAGTCCGTGCTGCTGACCCATCAGAGAACCCACACAGGCGAGAAGCCATAT
GTGTGCCGGGAATGTGGGAGAGGATTCTCTAACAAATCTCATCTGCTGAGGCATCAGAGA
ACCCATACCGGCGAGAAGCCTTACGTCTGCAGGGAATGCGGGCGGGGATTCAGAGACAAA
TCTCACCTGCTGAGGCACCAGCGCACTCATACAGGCGAGAAGCCTTATGTGTGTAGGGAA
TGTGGGAGGGGATTCCGCGATAAAAGCAATCTGCTGTCCCACCAGCGGACTCATACCGGC
GAGAAGCCATACGTCTGTAGGGAGTGCGGCAGGGGATTTTCTAACAAATCCCATCTGCTG
CGCCACCAGCGGACCCATACTGGCGAGAAGCCCTATGTGTGTCGCGAGTGCGGGAGGGGA
TTCCGCAACAAGAGCCACCTGCTGCGGCACCAGAGAACTCACACCGGAGAAAAGCCTTAC
GTGTGCAGGGAGTGTGGGCGCGGCTTTAGCGACAGGTCCTCTCTGTGCTATCATCAGAGG
ACTCACACCGGAGAGAAGCCCTACGTGTGCCGGGAGGATGAGGGTACCAGCGCCTGGTCC
CACCCCCAGTTTGAAAAGGGCGGAGGATCTGGCGGCGGAAGTGGCGGATCTGCTTGGAGC
CACCCTCAGTTTGAAAAGGGGGCCAGCGGCGAGTACCCCTACGACGTGCCAGATTACGCT
GGCAGCGGCGCCACCAACTTCAGCCTGCTGAAACAGGCCGGCGACGTGGAAGAGAACCCT
GGCCCTACCGGTGTGAGCAAGGGCGAGGAGCTGTTCACCGGGGTGGTGCCCATCCTGGTC
GAGCTGGACGGCGACGTAAACGGCCACAAGTTCAGCGTGTCCGGCGAGGGCGAGGGCGAT
GCCACCTACGGCAAGCTGACCCTGAAGCTGATCTGCACCACCGGCAAGCTGCCCGTGCCC
TGGCCCACCCTCGTGACCACCCTGGGCTACGGCCTGCAGTGCTTCGCCCGCTACCCCGAC
CACATGAAGCAGCACGACTTCTTCAAGTCCGCCATGCCCGAAGGCTACGTCCAGGAGCGC
ACCATCTTCTTCAAGGACGACGGCAACTACAAGACCCGCGCCGAGGTGAAGTTCGAGGGC
GACACCCTGGTGAACCGCATCGAGCTGAAGGGCATCGACTTCAAGGAGGACGGCAACATC
CTGGGGCACAAGCTGGAGTACAACTACAACAGCCACAACGTCTATATCACCGCCGACAAG
CAGAAGAACGGCATCAAGGCCAACTTCAAGATCCGCCACAACATCGAGGACGGCGGCGTG
CAGCTCGCCGACCACTACCAGCAGAACACCCCCATCGGCGACGGCCCCGTGCTGCTGCCC
GACAACCACTACCTGAGCTACCAGTCCAAGCTGAGCAAAGACCCCAACGAGAAGCGCGAT
CACATGGTCCTGCTGGAGTTCGTGACCGCCGCCGGGATCACTCTCGGCATGGACGAGCTG
TACAAGTAATGA

9. hZFonly RECODED to be less repetitive DNA sequence (same AA sequence) with C-terminal TwinStrep-HA-P2A-YFP tag (2052 bp)

   KozakSequence(6bp)-start(3bp)-PRDM9_HumanBallele_ZFarrayOnly RECODED(1113bp)-KpnI(6bp)-HAlinker(198bp)-AgeI(6bp)-YFP(714bp) -stops(6bp) HAlinker is: -shortlinker(6bp)-StrepII(24bp)-linker(36bp)- StrepII(24bp)-shortlinker(15bp)-HA(27bp)-shortlinker(9bp)-P2A(57bp)-

   GCCACCATGGTCAAGTATGGAGAGTGTGGACAAGGTTTCAGTGTTAAATCAGATGTTATT
   ACACACCAAAGGACACATACAGGGGAGAAGCTCTACGTCTGCAGGGAGTGTGGGCGGGGC
   TTCAGCTGGAAGTCCCACCTGCTGATCCACCAGAGAATTCACACAGGAGAAAAGCCTTAT
   GTGTGCAGAGAGTGCGGGAGAGGGTTCAGTTGGCAGTCTGTTCTCCTGACACACCAGAGA
   ACTCATACCGGAGAGAAGCCCTATGTTTGTAGAGAGTGTGGCCGCGGTTTTAGTAGACAA
   TCTGTTCTGTTGACCCACCAAAGAAGACATACAGGCGAAAAACCATATGTGTGCCGCGAG
   TGTGGAAGGGGATTTTCAAGACAATCAGTTCTGCTCACTCACCAAAGGAGACACACCGGC
   GAGAAGCCATATGTCTGTCGCGAATGTGGTCGCGGATTTTCATGGCAATCTGTCCTGTTG
   ACTCACCAGCGCACACACACTGGTGAAAAACCCTACGTTTGTCGCGAGTGCGGAAGGGGT
   TTTAGCTGGCAAAGCGTGCTGCTCACCCACCAGCGGACTCACACTGGAGAAAAACCTTAC
   GTTTGCCGGGAATGTGGCAGGGGATTTAGCAATAAGAGCCATTTGCTCAGACATCAGCGC
   ACCCATACTGGTGAGAAACCTTATGTTTGTCGGGAATGCGGACGCGGCTTTAGGGATAAA
   TCACATCTGTTGAGACACCAACGCACCCACACCGGGGAGAAGCCTTACGTGTGTAGAGAA
   TGTGGACGGGGATTTAGGGACAAAAGCAACCTGTTGTCACATCAAAGGACCCACACAGGG
   GAAAAGCCCTACGTGTGCAGGGAATGTGGAAGAGGATTCAGTAATAAAAGTCACCTCCTG
   CGCCATCAGAGGACCCATACAGGAGAGAAACCCTATGTGTGTAGGGAGTGCGGCAGAGGT
   TTTAGAAACAAGTCCCATTTGCTGAGACACCAGAGGACACACACAGGCGAGAAACCATAC
   GTTTGCAGAGAATGCGGAAGAGGCTTTAGCGATAGGTCAAGTTTGTGTTATCATCAAAGA
   ACCCACACTGGCGAAAAGCCATACGTGTGCCGGGAGGATGAGGGTACCAGCGCCTGGTCC
   CACCCCCAGTTTGAAAAGGGCGGAGGATCTGGCGGCGGAAGTGGCGGATCTGCTTGGAGC
   CACCCTCAGTTTGAAAAGGGGGCCAGCGGCGAGTACCCCTACGACGTGCCAGATTACGCT
   GGCAGCGGCGCCACCAACTTCAGCCTGCTGAAACAGGCCGGCGACGTGGAAGAGAACCCT
   GGCCCTACCGGTGTGAGCAAGGGCGAGGAGCTGTTCACCGGGGTGGTGCCCATCCTGGTC
   GAGCTGGACGGCGACGTAAACGGCCACAAGTTCAGCGTGTCCGGCGAGGGCGAGGGCGAT
   GCCACCTACGGCAAGCTGACCCTGAAGCTGATCTGCACCACCGGCAAGCTGCCCGTGCCC
   TGGCCCACCCTCGTGACCACCCTGGGCTACGGCCTGCAGTGCTTCGCCCGCTACCCCGAC
   CACATGAAGCAGCACGACTTCTTCAAGTCCGCCATGCCCGAAGGCTACGTCCAGGAGCGC
   ACCATCTTCTTCAAGGACGACGGCAACTACAAGACCCGCGCCGAGGTGAAGTTCGAGGGC
   GACACCCTGGTGAACCGCATCGAGCTGAAGGGCATCGACTTCAAGGAGGACGGCAACATC
   CTGGGGCACAAGCTGGAGTACAACTACAACAGCCACAACGTCTATATCACCGCCGACAAG
   CAGAAGAACGGCATCAAGGCCAACTTCAAGATCCGCCACAACATCGAGGACGGCGGCGTG
   CAGCTCGCCGACCACTACCAGCAGAACACCCCCATCGGCGACGGCCCCGTGCTGCTGCCC
   GACAACCACTACCTGAGCTACCAGTCCAAGCTGAGCAAAGACCCCAACGAGAAGCGCGAT
   CACATGGTCCTGCTGGAGTTCGTGACCGCCGCCGGGATCACTCTCGGCATGGACGAGCTG
   TACAAGTAATGA

10. hZFonly with C-terminal TwinStrep-V5-P2A-YFP tag (2052 bp)

    KozakSequence(6bp)-start(3bp)-PRDM9_HumanBallele_ZFarrayOnly

(1113bp)-KpnI(6bp)-V5linker(213bp)-AgeI(6bp)-YFP(714bp)-stops(6bp)
V5linker is: -shortlinker(6bp)-StrepII(24bp)-linker(36bp)
-StrepII(24bp)-shortlinker(15bp)-V5(42bp)-shortlinker(9bp)-P2A(57bp)-

GCCACCATGGTCAAGTATGGAGAGTGTGGACAAGGTTTCAGTGTTAAATCAGATGTTATT
ACACACCAAAGGACACATACAGGGGAGAAGCTCTACGTCTGCAGGGAGTGTGGGCGGGGC
TTCAGCTGGAAGTCCCACCTGCTGATCCACCAGAGAATTCACACCGGCGAGAAACCCTAC
GTGTGCAGAGAATGTGGGAGGGGATTTTCTTGGCAGTCTGTGCTGCTGACACACCAGCGC
ACACACACCGGCGAGAAGCCTTATGTCTGTAGAGAATGTGGGCGGGGATTCTCAAGACAG
AGCGTCCTGCTGACCCACCAGAGGCGCCACACAGGCGAGAAACCATATGTCTGCAGGGAG
TGTGGGCGGGGATTTTCCAGACAGTCTGTGCTGCTGACTCATCAGAGAAGACACACAGGC
GAGAAGCCCTATGTCTGCAGGGAATGCGGCCGGGGATTCAGTTGGCAGTCCGTCCTGCTG
ACACACCAGAGAACTCATACAGGCGAGAAACCTTACGTGTGCAGGGAATGTGGGCGCGGC
TTCAGCTGGCAGTCCGTGCTGCTGACCCATCAGAGAACCCACACAGGCGAGAAGCCATAT
GTGTGCCGGGAATGTGGGAGAGGATTCTCTAACAAATCTCATCTGCTGAGGCATCAGAGA
ACCCATACCGGCGAGAAGCCTTACGTCTGCAGGGAATGCGGGCGGGGATTCAGAGACAAA
TCTCACCTGCTGAGGCACCAGCGCACTCATACAGGCGAGAAGCCTTATGTGTGTAGGGAA
TGTGGGAGGGGATTCCGCGATAAAAGCAATCTGCTGTCCCACCAGCGGACTCATACCGGC
GAGAAGCCATACGTCTGTAGGGAGTGCGGCAGGGGATTTTCTAACAAATCCCATCTGCTG
CGCCACCAGCGGACCCATACTGGCGAGAAGCCCTATGTGTGTCGCGAGTGCGGGAGGGGA
TTCCGCAACAAGAGCCACCTGCTGCGGCACCAGAGAACTCACACCGGAGAAAAGCCTTAC
GTGTGCAGGGAGTGTGGGCGCGGCTTTAGCGACAGGTCCTCTCTGTGCTATCATCAGAGG
ACTCACACCGGAGAGAAGCCCTACGTGTGCCGGGAGGATGAGGGTACCAGCGCCTGGTCC
CACCCCCAGTTCGAGAAGGGCGGAGGATCTGGCGGCGGAAGTGGCGGATCTGCTTGGAGC
CACCCTCAGTTTGAAAAGGGGGCCAGCGGCGAGGGCAAGCCCATCCCTAATCCTCTGCTG
GGCCTGGACAGCACAGGCAGCGGCGCTACAAACTTCAGCCTGCTGAAGCAGGCCGGCGAC
GTGGAAGAGAACCCTGGACCTACCGGTGTGAGCAAGGGCGAGGAGCTGTTCACCGGGGTG
GTGCCCATCCTGGTCGAGCTGGACGGCGACGTAAACGGCCACAAGTTCAGCGTGTCCGGC
GAGGGCGAGGGCGATGCCACCTACGGCAAGCTGACCCTGAAGCTGATCTGCACCACCGGC
AAGCTGCCCGTGCCCTGGCCCACCCTCGTGACCACCCTGGGCTACGGCCTGCAGTGCTTC
GCCCGCTACCCCGACCACATGAAGCAGCACGACTTCTTCAAGTCCGCCATGCCCGAAGGC
TACGTCCAGGAGCGCACCATCTTCTTCAAGGACGACGGCAACTACAAGACCCGCGCCGAG
GTGAAGTTCGAGGGCGACACCCTGGTGAACCGCATCGAGCTGAAGGGCATCGACTTCAAG
GAGGACGGCAACATCCTGGGGCACAAGCTGGAGTACAACTACAACAGCCACAACGTCTAT
ATCACCGCCGACAAGCAGAAGAACGGCATCAAGGCCAACTTCAAGATCCGCCACAACATC
GAGGACGGCGGCGTGCAGCTCGCCGACCACTACCAGCAGAACACCCCCATCGGCGACGGC
CCCGTGCTGCTGCCCGACAACCACTACCTGAGCTACCAGTCCAAGCTGAGCAAAGACCCC
AACGAGAAGCGCGATCACATGGTCCTGCTGGAGTTCGTGACCGCCGCCGGGATCACTCTC
GGCATGGACGAGCTGTACAAGTAATGA

11. Chimp allele (W11a_Annaclara_18ZNF) with C-terminal TwinStrep-HA-P2A-YFP tag (4119 bp)

*SphI(6bp)-KozakSequence(6bp)-start(3bp)-PRDM9_HumanBallele_*
first9exons(1140bp)-XbaI(6bp)-Chimp_W11a_ZFexon(2022bp)-KpnI(6bp)
-HAlinker(198bp)-AgeI(6bp)-YFP(714bp)-stops(6bp)-XhoI(6bp)
HAlinker is: -shortlinker(6bp)-StrepII(24bp)-linker(36bp)-StrepII(24bp)

-shortlinker(15bp)-HA(27bp)-shortlinker(9bp)-P2A(57bp)-

GCATGCGCCACCATGAGCCCTGAAAAGTCCCAAGAGGAGAGCCCAGAAGAAGACACAGAG
AGAACAGAGCGGAAGCCCATGGTCAAAGATGCCTTCAAAGACATTTCCATATACTTCACC
AAGGAAGAATGGGCAGAGATGGGAGACTGGGAGAAAACTCGCTATAGGAATGTGAAAAGG
AACTATAATGCACTGATTACTATAGGTCTCAGAGCCACTCGACCAGCTTTCATGTGTCAC
CGAAGGCAGGCCATCAAACTCCAGGTGGATGACACAGAAGATTCTGATGAAGAATGGACC
CCTAGGCAGCAAGTCAAACCTCCTTGGATGGCCTTAAGAGTGGAACAGCGTAAACACCAG
AAGGGAATGCCCAAGGCGTCATTCAGTAATGAATCTAGTTTGAAAGAATTGTCAAGAACA
GCAAATTTACTGAATGCAAGTGGATCAGAGCAGGCTCAGAAACCAGTGTCCCCTTCTGGA
GAAGCAAGTACCTCTGGACAGCACTCAAGACTAAAACTGGAACTCAGGAAGAAGGAGACT
GAAAGAAAGATGTATAGCCTGCGAGAAAGAAAGGGTCATGCATACAAAGAGGTCAGCGAG
CCGCAGGATGATGATTACCTCTATTGTGAGATGTGTCAGAACTTCTTCATTGACAGCTGT
GCTGCCCATGGGCCCCCTACATTTGTAAAGGACAGTGCAGTGGACAAGGGGCACCCCAAC
CGTTCAGCCCTCAGTCTGCCCCCAGGGCTGAGAATTGGGCCATCAGGCATCCCTCAGGCT
GGGCTTGGAGTATGGAATGAGGCATCTGATCTGCCGCTGGGTCTGCACTTTGGCCCTTAT
GAGGGCCGAATTACAGAAGACGAAGAGGCAGCCAACAATGGATACTCCTGGCTGATCACC
AAGGGGAGAAACTGCTATGAGTATGTGGATGGAAAAGATAAATCCTGGGCCAACTGGATG
AGGTATGTGAACTGTGCCCGGGATGATGAAGAGCAGAACCTGGTGGCCTTCCAGTACCAC
AGGCAGATCTTCTATAGAACCTGCCGAGTCATTAGGCCAGGCTGTGAACTGCTGGTCTGG
TATGGGGATGAATACGGCCAGGAACTGGGCATCAAGTGGGGCAGCAAGTGGAAGAAAGAG
CTCATGGCAGGGAGATCTAGAGAGCCCAAGCCTGAAATCCACCCCTGTCCAAGTTGTTGC
CTGGCCTTTAGCAGCCAGAAGTTCCTGTCCCAGCACGTCGAGAGAAATCACAGCTCCCAG
AACTTCCCAGGACCCAGCGCAAGAAAGCTGCTGCAGCCCGAGAACCCTTGCCCAGGCGAC
CAGAATCAGGAGCAGCAGTACCCCGACCCTCGGTCCAGAAACGATAAGACCAAAGGCCAG
GAAATCAAGGAGAGGTCTAAACTGCTGAATAAGCGCACATGGCAGCGAGAGATTAGCCGG
GCCTTCTCTAGTCCCCCTAAAGGACAGATGGGCAGCTGCAGAGTGGGCAAGAGGATCATG
GAGGAAGAGTCCAGAACCGGGCAGAAAGTCAACCCCGGAAATACAGCCAAGCTGTTCGTG
GGCGTCGGGATCAGCAGGATTGCTAAGGTGAAATACGGGGAGTGCGGACAGGGCTTTAGC
GTGAAATCCGATGTCATTACCCACCAGAGAACACATACTGGAGAAAAGCCATACGTGTGC
CGCGAGTGTGGGCGAGGATTCTCTTGGAAAAGTCACCTGCTGTCCCATCAGCGCACCCAC
ACAGGCGAAAAGCCCTACGTGTGCCGGGAGTGTGGCAGAGGGTTTAGCGTCAAATCAAGC
CTGCTGTCCCATCGGACCACACACACCGGGGAAAAGCCTTACGTGTGCAGGGAGTGTGGA
CGCGGCTTCTCTGTCAAATCCTCTCTGCTGAGTCATCAGCGCACTCACACCGGAGAGAAA
CCATACGTGTGCCGAGAGTGTGGGCGGGGATTTTCACAGCAGAGCAATCTGCTGAGCCAT
CAGCGGACACACACTGGCGAAAAACCCTACGTGTGCAGAGAGTGTGGCAGGGGGTTCTCA
GTCAAGAGTTCACTGCTGAGCCATCAGAGAACCCACACAGGGGAAAAACCTTACGTGTGC
CGAGAATGCGGACGAGGCTTTTCCCAGCAGTCTCATCTGCTGTCTCACCAGAGGACTCAT
ACCGGAGAAAAACCATATGTCTGCCGGGAGTGTGGGAGAGGATTCAGTCAGCAGTCACAC
CTGCTGAGCCATCAGCGCACACACACTGGCGAGAAGCCTTACGTGTGCAGAGAATGCGGA
CGAGGGTTTAGCCAGCAGTCCCATCTGCTGAGGCACCAGCGCACCCATACAGGGGAGAAA
CCCTACGTGTGCAGGGAATGCGGACGGGGCTTCTCCGTCAAAAGCTCCCTGCTGTCTCAC
CAGAGAACTCATACCGGAGAGAAGCCTTACGTGTGCCGCGAATGCGGGAGGGGCTTCAGC
GTCAAATCTAGTCTGCTGTCACACAGGACTACCCATACCGGCGAGAAACCTTACGTGTGC
AGAGAGTGCGGACGAGGGTTCAGCGTGAAGTCAAGCCTGCTGTCCCACCAGCGGACACAT
ACTGGGGAGAAGCCATACGTGTGCAGGGAATGTGGCAGAGGCTTTTCTAAGCAGTCCCAC

```
CTGCTGAGCCACCAGAGAACTCATACAGGCGAGAAACCGTACGTGTGCCGGGAATGTGGG
CGCGGATTCTCACAGCAGAGCCATCTGCTGAGCCATCAGAGGACCCATACCGGCGAAAAA
CCATACGTGTGCAGAGAGTGTGGTCGAGGGTTTTCCCAGCAGTCTCACCTGCTGCGACAC
CAGAGAACACACACTGGCGAAAAGCCGTACGTGTGCAGAGAGTGTGGAAGGGGCTTCTCT
GTGAAGAGCTCCCTGCTGAGTCACCAGAGAACTCACACAGGCGAGAAGCCGTACGTGTGC
AGAGAGTGTGGGCGAGGATTTTCTGTCAAAAGTTCACTGCTGAGCCACCAGCGGACTCAC
ACCGGCGAGAAGCCCTACGTGTGCAGAGAATGTGAGAGGGGGTTCAGTCAGCAGTCCCAC
CTGCTGCGCCACCAGAGGACCCATACTGGCGAAAAACCGTACGTGTGCAGAGAGTGTGGT
AGAGGGTTTAGTAGACAGAGCGCACTGCTGATTCACCAGAGGACCCATACAGGCGAAAAG
CCAGGTACCAGCGCCTGGTCCCACCCCCAGTTTGAAAAGGGCGGAGGATCTGGCGGCGGA
AGTGGCGGATCTGCTTGGAGCCACCCTCAGTTTGAAAAGGGGGCCAGCGGCGAGTACCCC
TACGACGTGCCAGATTACGCTGGCAGCGGCGCCACCAACTTCAGCCTGCTGAAACAGGCC
GGCGACGTGGAAGAGAACCCTGGCCCTACCGGTGTGAGCAAGGGCGAGGAGCTGTTCACC
GGGGTGGTGCCCATCCTGGTCGAGCTGGACGGCGACGTAAACGGCCACAAGTTCAGCGTG
TCCGGCGAGGGCGAGGGCGATGCCACCTACGGCAAGCTGACCCTGAAGCTGATCTGCACC
ACCGGCAAGCTGCCCGTGCCCTGGCCCACCCTCGTGACCACCCTGGGCTACGGCCTGCAG
TGCTTCGCCCGCTACCCCGACCACATGAAGCAGCACGACTTCTTCAAGTCCGCCATGCCC
GAAGGCTACGTCCAGGAGCGCACCATCTTCTTCAAGGACGACGGCAACTACAAGACCCGC
GCCGAGGTGAAGTTCGAGGGCGACACCCTGGTGAACCGCATCGAGCTGAAGGGCATCGAC
TTCAAGGAGGACGGCAACATCCTGGGGCACAAGCTGGAGTACAACTACAACAGCCACAAC
GTCTATATCACCGCCGACAAGCAGAAGAACGGCATCAAGGCCAACTTCAAGATCCGCCAC
AACATCGAGGACGGCGGCGTGCAGCTCGCCGACCACTACCAGCAGAACACCCCCATCGGC
GACGGCCCCGTGCTGCTGCCCGACAACCACTACCTGAGCTACCAGTCCAAGCTGAGCAAA
GACCCCAACGAGAAGCGCGATCACATGGTCCTGCTGGAGTTCGTGACCGCCGCCGGGATC
ACTCTCGGCATGGACGAGCTGTACAAGTAATGACTCGAG
```

12. Chimp allele (W11a_Annaclara_18ZNF) with C-terminal TwinStrep-V5-P2A-YFP tag (4134 bp)

    SphI(6bp)-KozakSequence(6bp)-start(3bp)-PRDM9_HumanBallele_ first9exons(1140bp)-XbaI(6bp)-Chimp_W11a_ZFexon(2022bp)-KpnI(6bp) -V5linker(213bp)-AgeI(6bp)-YFP(714bp)-stops(6bp)-XhoI(6bp) V5linker is: -shortlinker(6bp)-StrepII(24bp)-linker(36bp)-StrepII(24bp) -shortlinker(15bp)-V5(42bp)-shortlinker(9bp)-P2A(57bp)-

```
CATGCGCCACCATGAGCCCTGAAAAGTCCCAAGAGGAGAGCCCAGAAGAAGACACAGAGA
GAACAGAGCGGAAGCCCATGGTCAAAGATGCCTTCAAAGACATTTCCATATACTTCACCA
AGGAAGAATGGGCAGAGATGGGAGACTGGGAGAAAACTCGCTATAGGAATGTGAAAAGGA
ACTATAATGCACTGATTACTATAGGTCTCAGAGCCACTCGACCAGCTTTCATGTGTCACC
GAAGGCAGGCCATCAAACTCCAGGTGGATGACACAGAAGATTCTGATGAAGAATGGACCC
CTAGGCAGCAAGTCAAACCTCCTTGGATGGCCTTAAGAGTGGAACAGCGTAAACACCAGA
AGGGAATGCCCAAGGCGTCATTCAGTAATGAATCTAGTTTGAAAGAATTGTCAAGAACAG
CAAATTTACTGAATGCAAGTGGATCAGAGCAGGCTCAGAAACCAGTGTCCCCTTCTGGAG
AAGCAAGTACCTCTGGACAGCACTCAAGACTAAAACTGGAACTCAGGAAGAAGGAGACTG
AAAGAAAGATGTATAGCCTGCGAGAAAGAAAGGGTCATGCATACAAAGAGGTCAGCGAGC
CGCAGGATGATGATTACCTCTATTGTGAGATGTGTCAGAACTTCTTCATTGACAGCTGTG
CTGCCCATGGGCCCCCTACATTTGTAAAGGACAGTGCAGTGGACAAGGGGCACCCCAACC
```

GTTCAGCCCTCAGTCTGCCCCCAGGGCTGAGAATTGGGCCATCAGGCATCCCTCAGGCTG
GGCTTGGAGTATGGAATGAGGCATCTGATCTGCCGCTGGGTCTGCACTTTGGCCCTTATG
AGGGCCGAATTACAGAAGACGAAGAGGCAGCCAACAATGGATACTCCTGGCTGATCACCA
AGGGGAGAAACTGCTATGAGTATGTGGATGGAAAAGATAAATCCTGGGCCAACTGGATGA
GGTATGTGAACTGTGCCCGGGATGATGAAGAGCAGAACCTGGTGGCCTTCCAGTACCACA
GGCAGATCTTCTATAGAACCTGCCGAGTCATTAGGCCAGGCTGTGAACTGCTGGTCTGGT
ATGGGGATGAATACGGCCAGGAACTGGGCATCAAGTGGGGCAGCAAGTGGAAGAAAGAGC
TCATGGCAGGGAGATCTAGAGAGCCCAAGCCTGAAATCCACCCCTGTCCAAGTTGTTGCC
TGGCCTTTAGCAGCCAGAAGTTCCTGTCCCAGCACGTCGAGAGAAATCACAGCTCCCAGA
ACTTCCCAGGACCCAGCGCAAGAAAGCTGCTGCAGCCCGAGAACCCTTGCCCAGGCGACC
AGAATCAGGAGCAGCAGTACCCCGACCCTCGGTCCAGAAACGATAAGACCAAAGGCCAGG
AAATCAAGGAGAGGTCTAAACTGCTGAATAAGCGCACATGGCAGCGAGAGATTAGCCGGG
CCTTCTCTAGTCCCCCTAAAGGACAGATGGGCAGCTGCAGAGTGGGCAAGAGGATCATGG
AGGAAGAGTCCAGAACCGGGCAGAAAGTCAACCCCGGAAATACAGCCAAGCTGTTCGTGG
GCGTCGGGATCAGCAGGATTGCTAAGGTGAAATACGGGGAGTGCGGACAGGGCTTTAGCG
TGAAATCCGATGTCATTACCCACCAGAGAACACATACTGGAGAAAAGCCATACGTGTGCC
GCGAGTGTGGGCGAGGATTCTCTTGGAAAAGTCACCTGCTGTCCCATCAGCGCACCCACA
CAGGCGAAAAGCCCTACGTGTGCCGGGAGTGTGGCAGAGGGTTTAGCGTCAAATCAAGCC
TGCTGTCCCATCGGACCACACACACCGGGGAAAAGCCTTACGTGTGCAGGGAGTGTGGAC
GCGGCTTCTCTGTCAAATCCTCTCTGCTGAGTCATCAGCGCACTCACACCGGAGAGAAAC
CATACGTGTGCCGAGAGTGTGGGCGGGGATTTTCACAGCAGAGCAATCTGCTGAGCCATC
AGCGGACACACACTGGCGAAAAACCCTACGTGTGCAGAGAGTGTGGCAGGGGGTTCTCAG
TCAAGAGTTCACTGCTGAGCCATCAGAGAACCCACACAGGGGAAAAACCTTACGTGTGCC
GAGAATGCGGACGAGGCTTTTCCCAGCAGTCTCATCTGCTGTCTCACCAGAGGACTCATA
CCGGAGAAAAACCATATGTCTGCCGGGAGTGTGGGAGAGGATTCAGTCAGCAGTCACACC
TGCTGAGCCATCAGCGCACACACACTGGCGAGAAGCCTTACGTGTGCAGAGAATGCGGAC
GAGGGTTTAGCCAGCAGTCCCATCTGCTGAGGCACCAGCGCACCCATACAGGGGAGAAAC
CCTACGTGTGCAGGGAATGCGGACGGGGCTTCTCCGTCAAAAGCTCCCTGCTGTCTCACC
AGAGAACTCATACCGGAGAGAAGCCTTACGTGTGCCGCGAATGCGGGAGGGGCTTCAGCG
TCAAATCTAGTCTGCTGTCACACAGGACTACCCATACCGGCGAGAAACCTTACGTGTGCA
GAGAGTGCGGACGAGGGTTCAGCGTGAAGTCAAGCCTGCTGTCCCACCAGCGGACACATA
CTGGGGAGAAGCCATACGTGTGCAGGGAATGTGGCAGAGGCTTTTCTAAGCAGTCCCACC
TGCTGAGCCACCAGAGAACTCATACAGGCGAGAAACCGTACGTGTGCCGGGAATGTGGGC
GCGGATTCTCACAGCAGAGCCATCTGCTGAGCCATCAGAGGACCCATACCGGCGAAAAAC
CATACGTGTGCAGAGAGTGTGGTCGAGGGTTTTCCCAGCAGTCTCACCTGCTGCGACACC
AGAGAACACACACTGGCGAAAAGCCGTACGTGTGCAGAGAGTGTGGAAGGGGCTTCTCTG
TGAAGAGCTCCCTGCTGAGTCACCAGAGAACTCACACAGGCGAGAAGCCGTACGTGTGCA
GAGAGTGTGGGCGAGGATTTTCTGTCAAAAGTTCACTGCTGAGCCACCAGCGGACTCACA
CCGGCGAGAAGCCCTACGTGTGCAGAGAATGTGAGAGGGGGTTCAGTCAGCAGTCCCACC
TGCTGCGCCACCAGAGGACCCATACTGGCGAAAAACCGTACGTGTGCAGAGAGTGTGGTA
GAGGGTTTAGTAGACAGAGCGCACTGCTGATTCACCAGAGGACCCATACAGGCGAAAAGC
CAGGTACCAGCGCCTGGTCCCACCCCCAGTTCGAGAAGGGCGGAGGATCTGGCGGCGGAA
GTGGCGGATCTGCTTGGAGCCACCCTCAGTTTGAAAAGGGGGCCAGCGGCGAGGGCAAGC
CCATCCCTAATCCTCTGCTGGGCCTGGACAGCACAGGCAGCGGCGCTACAAACTTCAGCC
TGCTGAAGCAGGCCGGCGACGTGGAAGAGAACCCTGGACCTACCGGTGTGAGCAAGGGCG

AGGAGCTGTTCACCGGGGTGGTGCCCATCCTGGTCGAGCTGGACGGCGACGTAAACGGCC
ACAAGTTCAGCGTGTCCGGCGAGGGCGAGGGCGATGCCACCTACGGCAAGCTGACCCTGA
AGCTGATCTGCACCACCGGCAAGCTGCCCGTGCCCTGGCCCACCCTCGTGACCACCCTGG
GCTACGGCCTGCAGTGCTTCGCCCGCTACCCCGACCACATGAAGCAGCACGACTTCTTCA
AGTCCGCCATGCCCGAAGGCTACGTCCAGGAGCGCACCATCTTCTTCAAGGACGACGGCA
ACTACAAGACCCGCGCCGAGGTGAAGTTCGAGGGCGACACCCTGGTGAACCGCATCGAGC
TGAAGGGCATCGACTTCAAGGAGGACGGCAACATCCTGGGGCACAAGCTGGAGTACAACT
ACAACAGCCACAACGTCTATATCACCGCCGACAAGCAGAAGAACGGCATCAAGGCCAACT
TCAAGATCCGCCACAACATCGAGGACGGCGGCGTGCAGCTCGCCGACCACTACCAGCAGA
ACACCCCCATCGGCGACGGCCCCGTGCTGCTGCCCGACAACCACTACCTGAGCTACCAGT
CCAAGCTGAGCAAAGACCCCAACGAGAAGCGCGATCACATGGTCCTGCTGGAGTTCGTGA
CCGCCGCCGGGATCACTCTCGGCATGGACGAGCTGTACAAGTAATGACTCGAG

13. noZF with C-terminal TwinStrep-HA-P2A-YFP tag (2517 bp)

SphI(6bp)-KozakSequence(6bp)-start(3bp)-PRDM9_HumanBallele_ toZFarray(1566bp)-KpnI(6bp)-HAlinker(198bp)-AgeI(6bp)-YFP(714bp) -stops(6bp)-XhoI(6bp)
HAlinker is: -shortlinker(6bp)-StrepII(24bp)-linker(36bp)-StrepII(24bp) -shortlinker(15bp)-HA(27bp)-shortlinker(9bp)-P2A(57bp)-

GCATGCGCCACCATGAGCCCTGAAAAGTCCCAAGAGGAGAGCCCAGAAGAAGACACAGAG
AGAACAGAGCGGAAGCCCATGGTCAAAGATGCCTTCAAAGACATTTCCATATACTTCACC
AAGGAAGAATGGGCAGAGATGGGAGACTGGGAGAAAACTCGCTATAGGAATGTGAAAAGG
AACTATAATGCACTGATTACTATAGGTCTCAGAGCCACTCGACCAGCTTTCATGTGTCAC
CGAAGGCAGGCCATCAAACTCCAGGTGGATGACACAGAAGATTCTGATGAAGAATGGACC
CCTAGGCAGCAAGTCAAACCTCCTTGGATGGCCTTAAGAGTGGAACAGCGTAAACACCAG
AAGGGAATGCCCAAGGCGTCATTCAGTAATGAATCTAGTTTGAAAGAATTGTCAAGAACA
GCAAATTTACTGAATGCAAGTGGATCAGAGCAGGCTCAGAAACCAGTGTCCCCTTCTGGA
GAAGCAAGTACCTCTGGACAGCACTCAAGACTAAAACTGGAACTCAGGAAGAAGGAGACT
GAAAGAAAGATGTATAGCCTGCGAGAAAGAAAGGGTCATGCATACAAAGAGGTCAGCGAG
CCGCAGGATGATGATTACCTCTATTGTGAGATGTGTCAGAACTTCTTCATTGACAGCTGT
GCTGCCCATGGGCCCCCTACATTTGTAAAGGACAGTGCAGTGGACAAGGGGCACCCCAAC
CGTTCAGCCCTCAGTCTGCCCCCAGGGCTGAGAATTGGGCCATCAGGCATCCCTCAGGCT
GGGCTTGGAGTATGGAATGAGGCATCTGATCTGCCGCTGGGTCTGCACTTTGGCCCTTAT
GAGGGCCGAATTACAGAAGACGAAGAGGCAGCCAACAATGGATACTCCTGGCTGATCACC
AAGGGGAGAAACTGCTATGAGTATGTGGATGGAAAAGATAAATCCTGGGCCAACTGGATG
AGGTATGTGAACTGTGCCCGGGATGATGAAGAGCAGAACCTGGTGGCCTTCCAGTACCAC
AGGCAGATCTTCTATAGAACCTGCCGAGTCATTAGGCCAGGCTGTGAACTGCTGGTCTGG
TATGGGGATGAATACGGCCAGGAACTGGGCATCAAGTGGGGCAGCAAGTGGAAGAAAGAG
CTCATGGCAGGGAGAGAACCAAAGCCAGAGATCCATCCATGTCCCTCATGCTGTCTGGCC
TTTTCAAGTCAGAAATTTCTCAGTCAACATGTAGAACGCAATCACTCCTCTCAGAACTTC
CCAGGACCATCTGCAAGAAAACTCCTCCAACCAGAGAATCCCTGCCCAGGGGATCAGAAT
CAGGAGCAGCAATATCCAGATCCACACAGCCGTAATGACAAAACCAAAGGTCAAGAGATC
AAAGAAAGGTCCAAACTCTTGAATAAAAGGACATGGCAGAGGGAGATTTCAAGGGCCTTT
TCTAGCCCACCCAAAGGACAAATGGGGAGCTGTAGAGTGGGAAAAAGAATAATGGAAGAA
GAGTCCAGAACAGGCCAGAAAGTGAATCCAGGGAACACAGGCAAATTATTTGTGGGGGTA

GGAATCTCAAGAATTGCAAAAGGTACCAGCGCCTGGTCCCACCCCCAGTTTGAAAAGGGC
GGAGGATCTGGCGGCGGAAGTGGCGGATCTGCTTGGAGCCACCCTCAGTTTGAAAAGGGG
GCCAGCGGCGAGTACCCCTACGACGTGCCAGATTACGCTGGCAGCGGCGCCACCAACTTC
AGCCTGCTGAAACAGGCCGGCGACGTGGAAGAGAACCCTGGCCCTACCGGTGTGAGCAAG
GGCGAGGAGCTGTTCACCGGGGTGGTGCCCATCCTGGTCGAGCTGGACGGCGACGTAAAC
GGCCACAAGTTCAGCGTGTCCGGCGAGGGCGAGGGCGATGCCACCTACGGCAAGCTGACC
CTGAAGCTGATCTGCACCACCGGCAAGCTGCCCGTGCCCTGGCCCACCCTCGTGACCACC
CTGGGCTACGGCCTGCAGTGCTTCGCCCGCTACCCCGACCACATGAAGCAGCACGACTTC
TTCAAGTCCGCCATGCCCGAAGGCTACGTCCAGGAGCGCACCATCTTCTTCAAGGACGAC
GGCAACTACAAGACCCGCGCCGAGGTGAAGTTCGAGGGCGACACCCTGGTGAACCGCATC
GAGCTGAAGGGCATCGACTTCAAGGAGGACGGCAACATCCTGGGGCACAAGCTGGAGTAC
AACTACAACAGCCACAACGTCTATATCACCGCCGACAAGCAGAAGAACGGCATCAAGGCC
AACTTCAAGATCCGCCACAACATCGAGGACGGCGGCGTGCAGCTCGCCGACCACTACCAG
CAGAACACCCCCATCGGCGACGGCCCCGTGCTGCTGCCCGACAACCACTACCTGAGCTAC
CAGTCCAAGCTGAGCAAAGACCCCAACGAGAAGCGCGATCACATGGTCCTGCTGGAGTTC
GTGACCGCCGCCGGGATCACTCTCGGCATGGACGAGCTGTACAAGTAATGACTCGAG

14. noZF with C-terminal TwinStrep-V5-P2A-YFP tag (2532 bp)

SphI(6bp)-KozakSequence(6bp)-start(3bp)-PRDM9_HumanBallele_ toZFarray(1566bp)-KpnI(6bp)-V5linker(213bp)-AgeI(6bp)-YFP(714bp) -stops(6bp)-XhoI(6bp)
V5linker is: -shortlinker(6bp)-StrepII(24bp)-linker(36bp)-StrepII(24bp) -shortlinker(15bp)-V5(42bp)-shortlinker(9bp)-P2A(57bp)-

GCATGCGCCACCATGAGCCCTGAAAAGTCCCAAGAGGAGAGCCCAGAAGAAGACACAGAG
AGAACAGAGCGGAAGCCCATGGTCAAAGATGCCTTCAAAGACATTTCCATATACTTCACC
AAGGAAGAATGGGCAGAGATGGGAGACTGGGAGAAAACTCGCTATAGGAATGTGAAAAGG
AACTATAATGCACTGATTACTATAGGTCTCAGAGCCACTCGACCAGCTTTCATGTGTCAC
CGAAGGCAGGCCATCAAACTCCAGGTGGATGACACAGAAGATTCTGATGAAGAATGGACC
CCTAGGCAGCAAGTCAAACCTCCTTGGATGGCCTTAAGAGTGGAACAGCGTAAACACCAG
AAGGGAATGCCCAAGGCGTCATTCAGTAATGAATCTAGTTTGAAAGAATTGTCAAGAACA
GCAAATTTACTGAATGCAAGTGGATCAGAGCAGGCTCAGAAACCAGTGTCCCCTTCTGGA
GAAGCAAGTACCTCTGGACAGCACTCAAGACTAAAACTGGAACTCAGGAAGAAGGAGACT
GAAAGAAAGATGTATAGCCTGCGAGAAAGAAAGGGTCATGCATACAAAGAGGTCAGCGAG
CCGCAGGATGATGATTACCTCTATTGTGAGATGTGTCAGAACTTCTTCATTGACAGCTGT
GCTGCCCATGGGCCCCCTACATTTGTAAAGGACAGTGCAGTGGACAAGGGGCACCCCAAC
CGTTCAGCCCTCAGTCTGCCCCCAGGGCTGAGAATTGGGCCATCAGGCATCCCTCAGGCT
GGGCTTGGAGTATGGAATGAGGCATCTGATCTGCCGCTGGGTCTGCACTTTGGCCCTTAT
GAGGGCCGAATTACAGAAGACGAAGAGGCAGCCAACAATGGATACTCCTGGCTGATCACC
AAGGGGAGAAACTGCTATGAGTATGTGGATGGAAAAGATAAATCCTGGGCCAACTGGATG
AGGTATGTGAACTGTGCCCGGGATGATGAAGAGCAGAACCTGGTGGCCTTCCAGTACCAC
AGGCAGATCTTCTATAGAACCTGCCGAGTCATTAGGCCAGGCTGTGAACTGCTGGTCTGG
TATGGGGATGAATACGGCCAGGAACTGGGCATCAAGTGGGGCAGCAAGTGGAAGAAAGAG
CTCATGGCAGGGAGAGAACCAAAGCCAGAGATCCATCCATGTCCCTCATGCTGTCTGGCC
TTTTCAAGTCAGAAATTTCTCAGTCAACATGTAGAACGCAATCACTCCTCTCAGAACTTC
CCAGGACCATCTGCAAGAAAACTCCTCCAACCAGAGAATCCCTGCCCAGGGGATCAGAAT

CAGGAGCAGCAATATCCAGATCCACACAGCCGTAATGACAAAACCAAAGGTCAAGAGATC
AAAGAAAGGTCCAAACTCTTGAATAAAAGGACATGGCAGAGGGAGATTTCAAGGGCCTTT
TCTAGCCCACCCAAAGGACAAATGGGGAGCTGTAGAGTGGGAAAAAGAATAATGGAAGAA
GAGTCCAGAACAGGCCAGAAAGTGAATCCAGGGAACACAGGCAAATTATTTGTGGGGGTA
GGAATCTCAAGAATTGCAAAAGGTACCAGCGCCTGGTCCCACCCCCAGTTCGAGAAGGGC
GGAGGATCTGGCGGCGGAAGTGGCGGATCTGCTTGGAGCCACCCTCAGTTTGAAAAGGGG
GCCAGCGGCGAGGGCAAGCCCATCCCTAATCCTCTGCTGGGCCTGGACAGCACAGGCAGC
GGCGCTACAAACTTCAGCCTGCTGAAGCAGGCCGGCGACGTGGAAGAGAACCCTGGACCT
ACCGGTGTGAGCAAGGGCGAGGAGCTGTTCACCGGGGTGGTGCCCATCCTGGTCGAGCTG
GACGGCGACGTAAACGGCCACAAGTTCAGCGTGTCCGGCGAGGGCGAGGGCGATGCCACC
TACGGCAAGCTGACCCTGAAGCTGATCTGCACCACCGGCAAGCTGCCCGTGCCCTGGCCC
ACCCTCGTGACCACCCTGGGCTACGGCCTGCAGTGCTTCGCCCGCTACCCCGACCACATG
AAGCAGCACGACTTCTTCAAGTCCGCCATGCCCGAAGGCTACGTCCAGGAGCGCACCATC
TTCTTCAAGGACGACGGCAACTACAAGACCCGCGCCGAGGTGAAGTTCGAGGGCGACACC
CTGGTGAACCGCATCGAGCTGAAGGGCATCGACTTCAAGGAGGACGGCAACATCCTGGGG
CACAAGCTGGAGTACAACTACAACAGCCACAACGTCTATATCACCGCCGACAAGCAGAAG
AACGGCATCAAGGCCAACTTCAAGATCCGCCACAACATCGAGGACGGCGGCGTGCAGCTC
GCCGACCACTACCAGCAGAACACCCCCATCGGCGACGGCCCCGTGCTGCTGCCCGACAAC
CACTACCTGAGCTACCAGTCCAAGCTGAGCAAAGACCCCAACGAGAAGCGCGATCACATG
GTCCTGCTGGAGTTCGTGACCGCCGCCGGGATCACTCTCGGCATGGACGAGCTGTACAAG
TAATGACTCGAG

B

R and Perl code

```
#DeNovoPeakCalling.SingleBase.R
#A peak calling algorithm that computes single-base enrichment and likelihood values
#by Nick Altemose
#13 Jul 2015
######

##step1: initialise inputs and outputs
library("parallel")
library("IRanges")
options("scipen"=8)
btpath = "/data/smew2/altemose/software/bedtools2/bin/bedtools"
batchsize=1000000
cithresh=0.99

#set these all to 1 if preprocessing steps are needed, and to 0 if this has already been done for a given sample
computecoverage="0,0,0" #preprocess read files to produce single-base coverage for r1,r2,genomic (0 or 1 for each)
createbin=0 #process coverage files into binary format for easy reading
computelikelihoods=0 #compute likelihood and enrichment values at each base


searchwidth = 1000 #maximum CI width
minsep = 250 #minimum separation between peak centres
pvalthresh = 0.000001 #maximum p-value at peak centre

args=commandArgs(TRUE)
datapath = args[2] #path to read pair position bed files
sample = args[3] #a name for the sample
rep1suffix = args[4] #the index of the files containing r1 reads
rep2suffix = args[5] #the index of the files containing r2 reads
```

```
genomicsuffix = args[6] #the index of the files containing genomic input reads
qual = as.integer(args[7]) #the mapq threshold used
pvalthresh = as.numeric(args[8])
minsep = as.numeric(args[9])
computecoverage = as.character(args[10])
createbin = as.integer(args[11])
computelikelihoods = as.integer(args[12])


#example hardwired input
#datapath="/data/smew2/altemose/hekchip/0_InputData"
#sample = "noZFK4"
#rep1suffix = 243
#rep2suffix = 244
#genomicsuffix = "HEKinput"
#qual=1
#chr=1
#computecoverage=0
#createbin=0
#computelikelihoods=0


lthresh = qchisq(pvalthresh,df=1,lower.tail=F)
cilhood=qchisq(cithresh,df=2)
chrs=seq(1,22,1) #change if chromosome number differs
chrs=c(chrs,"X")
computecoverages=as.integer(strsplit(computecoverage,',')[[1]])


##step2: read in constants estimated previously by EstimateConstants.R, calculate constant terms used in likelihood calculations
constfile=paste("Constants/Constants.q",qual,".",sample,".100wide.100slide.bed",sep="")

constdata=read.table(constfile,header=TRUE)

alpha1=constdata[26,2]
alpha2=constdata[26,3]
beta=constdata[26,4]

term2=1+alpha1+alpha2
term3=beta+1
term6=alpha1*alpha2
term7=alpha1*beta+alpha2

outfileBase = paste("FinalOutput/SingleBasePeaks.",sample,".p",pvalthresh,".sep",minsep,sep="")
outfileALL = paste("SingleBasePeaks.",sample,".p",pvalthresh,".sep",minsep,".ALL.bed",sep="")
outfileALLa = paste("SingleBasePeaks.",sample,".p",pvalthresh,".sep",minsep,".autosomal.bed",sep="")


##step3: declare a function to perform calculations on one chromosome
```

```
79
80 getEnrichments=function(chr){
81
82 ##step4: define input and output file names
83
84 posfileA = paste(datapath,"/bychr/FragPos.q",qual,".",rep1suffix,".chr
       .chr",chr,".bed",sep="")
85 posfileB = paste(datapath,"/bychr/FragPos.q",qual,".",rep2suffix,".chr
       .chr",chr,".bed",sep="")
86 posfileG= paste(datapath,"/bychr/FragPos.q",qual,".",genomicsuffix,".
       chr.chr",chr,".bed",sep="")
87
88 covfileA = paste(datapath,"/bychr/SingleBaseFragDepth.q",qual,".",
       rep1suffix,".chr",chr,".bed.gz",sep="")
89 covfileB = paste(datapath,"/bychr/SingleBaseFragDepth.q",qual,".",
       rep2suffix,".chr",chr,".bed.gz",sep="")
90 covfileG = paste(datapath,"/bychr/SingleBaseFragDepth.q",qual,".",
       genomicsuffix,".chr",chr,".bed.gz",sep="")
91
92 covfileAb = paste(datapath,"/bychr/SingleBaseFragDepth.q",qual,".",
       rep1suffix,".chr",chr,".binary.gz",sep="")
93 covfileBb = paste(datapath,"/bychr/SingleBaseFragDepth.q",qual,".",
       rep2suffix,".chr",chr,".binary.gz",sep="")
94 covfileGb = paste(datapath,"/bychr/SingleBaseFragDepth.q",qual,".",
       genomicsuffix,".chr",chr,".binary.gz",sep="")
95
96 outfileLhood= paste("bychr/SingleBaseLikelihood.",sample,".chr",chr,".
       binary.r",sep="")
97 outfileEnrich= paste("bychr/SingleBaseEnrichment.",sample,".chr",chr,"
       .binary.r",sep="")
98 outfilePeaks= paste("FinalOutput/SingleBasePeaks.",sample,".p",
       pvalthresh,".sep",minsep,".chr",chr,".bed",sep="")
99
100
101 chrsizefile = paste("/data/smew2/altemose/hekchip/0_InputData/bychr/
       chromsize.chr",chr,".bed",sep="")
102 chrlen = read.table(chrsizefile,header=F,colClasses=c('character','
       integer'))[1,2]
103
104 ##step5: if not done already, compute single-base coverage values
       across chromosome and compress
105
106 if(computecoverages[1]==1){
107   system(paste(btpath," genomecov -d -i ",posfileA," -g ",chrsizefile,
         " | awk '{print $3}' | gzip >",covfileA,sep=" "))
108   conA=gzfile(covfileA,open="r")
109   conAb=gzfile(covfileAb,open="wb")
110   writeBin(scan(conA,what=integer(1),nlines=chrlen,quiet=TRUE),conAb)
111   close(conA)
112   close(conAb)
113   print(paste("done computing coverage A for",chr,date()))
114 }
115 if(computecoverages[2]==1){
116   system(paste(btpath," genomecov -d -i ",posfileB," -g ",chrsizefile,
         " | awk '{print $3}' | gzip >",covfileB,sep=" "))
```

```
  conB=gzfile(covfileB,open="r")
  conBb=gzfile(covfileBb,open="wb")
  writeBin(scan(conB,what=integer(1),nlines=chrlen,quiet=TRUE),conBb)
  close(conB)
  close(conBb)
  print(paste("done computing coverage B for",chr,date()))
}
if(computecoverages[3]==1){
  system(paste(btpath," genomecov -d -i ",posfileG," -g ",chrsizefile,
      " | awk '{print $3}' | gzip >",covfileG,sep=" "))
  conG=gzfile(covfileG,open="r")
  conGb=gzfile(covfileGb,open="wb")
  writeBin(scan(conG,what=integer(1),nlines=chrlen,quiet=TRUE),conGb)
  close(conG)
  close(conGb)
  print(paste("done computing coverage G for",chr,date()))
}

if(sum(computecoverages)==0 && createbin==1){

  conA=gzfile(covfileA,open="r")
  conAb=gzfile(covfileAb,open="wb")
  writeBin(scan(conA,what=integer(1),nlines=chrlen,quiet=TRUE),conAb)
  close(conA)
  close(conAb)

  conB=gzfile(covfileB,open="r")
  conBb=gzfile(covfileBb,open="wb")
  writeBin(scan(conB,what=integer(1),nlines=chrlen,quiet=TRUE),conBb)
  close(conB)
  close(conBb)

  conG=gzfile(covfileG,open="r")
  conGb=gzfile(covfileGb,open="wb")
  writeBin(scan(conG,what=integer(1),nlines=chrlen,quiet=TRUE),conGb)
  close(conG)
  close(conGb)

  print(paste("done converting coverage files to binary",chr,date()))

}


##step6: if not done already, compute single-base likelihood and
    enrichment values across chromosome, in batches

if(computelikelihoods==1){


conAb=gzfile(covfileAb,open="rb")
conBb=gzfile(covfileBb,open="rb")
conGb=gzfile(covfileGb,open="rb")
conOUTL=gzfile(outfileLhood,open="wb")
conOUTE=gzfile(outfileEnrich,open="wb")
startpos=0
```

```
while(startpos<chrlen){
  basenum=batchsize
  if((startpos+batchsize)>chrlen){
    basenum=chrlen-startpos
  }
  covA = readBin(conAb,what=integer(1),n=basenum)
  covB = readBin(conBb,what=integer(1),n=basenum)
  covG = readBin(conGb,what=integer(1),n=basenum)

  basepeaks = list(rep(0,basenum),rep(0,basenum))
  if(sum(covA+covB+covG)>0){
    basepeaks=makepeaks.singlebase(r1=covA,r2=covB,g=covG)
  }

  writeBin(basepeaks[[1]],conOUTL)
  writeBin(basepeaks[[2]],conOUTE)

  startpos = startpos+batchsize
}
close(conAb)
close(conBb)
close(conGb)
close(conOUTL)
close(conOUTE)
print(paste("done computing likelihoods",chr,date()))


}

##step7: read in all single-base log likelihood values across
    chromosome
con1=gzfile(outfileLhood,open="rb")
vec=readBin(con1,what=numeric(1),n=chrlen)
close(con1)

##step8: find bases above pvalue threshold that are local maxima (>
    minsep bases left and >= minsep bases to the right)
q0=which(vec>=lthresh)
q=q0[q0>searchwidth & q0<(chrlen-searchwidth)]
rm(q0)
qtest = vapply(q,function(x) (vec[x]>max(vec[(x-minsep):(x-1)]) && vec
    [x]>=max(vec[(x+1):(x+minsep)])),USE.NAMES=F,FUN.VALUE=logical(1))
newq=q[qtest]
rm(q)

##step9: around these bases find confidence intervals
getci=function(i){
  left = i-searchwidth+max(0,which(vec[(i-searchwidth):(i-1)]<=(vec[i
      ]-cilhood)))
  right = i + min(searchwidth,which(vec[(i+1):(i+searchwidth)]<=(vec[i
      ]-cilhood)))
  return(c(i,vec[i],left,right))
}

confints = t(sapply(newq,getci,USE.NAMES=F,simplify="array"))
```

```
rm(newq)

##step10: sort by CI width, find overlaps, indicate those not overlapping shorter interval

confints=confints[order(confints[,4]-confints[,3]),]
query = IRanges(confints[,3],confints[,4])
overlaps = as.matrix(findOverlaps(query,ignoreSelf=TRUE))

includevec = rep(0,dim(confints)[1])
includevec[1]=1
for(i in 1:dim(confints)[1]){
  includevec[i]=1
  if(sum(includevec[overlaps[overlaps[,1]==i & overlaps[,2]<i,2]])>0){
    includevec[i]=0
  }
}
confints=cbind(confints,includevec)


##step11: refine centres within each CI (take mean position of all bases with llhood=max)
refinecentres=function(i){
  floor(mean(range(confints[i,3]-1+which(vec[(confints[i,3]):(confints[i,4])]==confints[i,2]))))
}
newcentres=vapply(1:dim(confints)[1],refinecentres,USE.NAMES=F,FUN.VALUE=numeric(1))
confints[,1]=newcentres
confints=confints[order(confints[,1]),]


##step12: add in coverage and enrichment values and pvalues
localsigcounts=vapply(confints[,1],function(x) sum(vec[(x-minsep):(x+minsep)]>=lthresh),USE.NAMES=F,FUN.VALUE=integer(1))
rm(vec)

conAb=gzfile(covfileAb,open="rb")
covA=readBin(conAb,what=integer(1),n=chrlen)
confints=cbind(confints,covA[confints[,1]])
rm(covA)
close(conAb)

conBb=gzfile(covfileBb,open="rb")
covB=readBin(conBb,what=integer(1),n=chrlen)
confints=cbind(confints,covB[confints[,1]])
rm(covB)
close(conBb)

conGb=gzfile(covfileGb,open="rb")
covG=readBin(conGb,what=integer(1),n=chrlen)
confints=cbind(confints,covG[confints[,1]])
rm(covG)
close(conGb)

```

```
270
271 con2=gzfile(outfileEnrich,open="rb")
272 enrich=readBin(con2,what=numeric(1),n=chrlen)
273 close(con2)
274 confints=cbind(confints,enrich[confints[,1]])
275 rm(enrich)
276
277 signif=pchisq(confints[,2],df=1,lower.tail=F)
278 confints=cbind(confints,signif)
279 confints=cbind(confints,localsigcounts-1)
280
281
282 ##step12: print final peaks to text file
283
284 sub=confints[confints[,5]==1,]
285
286 write.table(cbind(paste("chr",chr,sep=""),sub[,1]-1,sub[,1],sub[,3]-1,
      sub[,c(4,6,7,8,9,2,10,11)]),file=outfilePeaks,quote=F,row.names=F,
      col.names=c("chr","centrestart","centrestop","CIstart","CIstop","
      covr1","covr2","covg","enrich","lhood","pval","SigLocalBases"),sep
      ="\t")
287
288 print(paste("done getting peak intervals",chr,date()))
289
290 return(1)
291
292 }
293
294
295
296 ####declare function used for calculating likelihoods and enrichments
297 makepeaks.singlebase=function(r1,r2,g){
298
299   g[g==0]=0.5
300   sumcov = r1+r2+g
301
302   term1=sumcov*term3
303   term4=beta*alpha1*r2+alpha2*r1
304   term5=beta*(r1+r2)
305
306   bterm=term1*term7-term2*term5-term3*term4
307   cterm=term1*term6-term2*term4
308   rm(term4)
309   aterm=term1*beta-term3*term5
310   rm(term5,term1)
311
312
313   yvals=(-bterm+sqrt(bterm^2-4*aterm*cterm))/2/aterm
314   rm(aterm,bterm,cterm)
315   yvals[yvals<0]=0
316   yvals[yvals>1e9]=1e9
317
318
319   bvals=sumcov/(term2+(term3*yvals))
320
```

```
    bvalsnull=sumcov/term2
    yvalsnull=rep(0,length(bvalsnull))

    lhooddiff=2*( (sumcov*(log(bvals)-1)+r1*log(alpha1+yvals)+r2*log(
        alpha2+yvals*beta)) - (sumcov*(log(bvalsnull)-1)+r1*log(alpha1+
        yvalsnull)+r2*log(alpha2+yvalsnull*beta)) )
    rm(bvals, bvalsnull, yvalsnull)

    lhooddiff[is.na(lhooddiff)]=0
    lhooddiff[sumcov==0.5]=0
    yvals[sumcov==0.5]=0
    rm(sumcov)

    return(list(lhooddiff,yvals))
}









print(date())
funfunc = mclapply(chrs,getEnrichments,mc.preschedule=TRUE,mc.cores=
    length(chrs))

system(paste("perl MergeFinalOutputs.pl ",outfileBase," ",outfileALL,
    sep=""))
system(paste("perl MergeFinalOutputs.autosomal.pl ",outfileBase," ",
    outfileALLa,sep=""))


print(paste("done!:",date()))

quit(save="no",runLast=FALSE)
```

text/DeNovoPeakCalling–SingleBase.R

```
#EstimateConstants.R
#estimates genome-wide constants alpha1, alpha2, and beta (and prop.
    reads from signal)
#by Nick Altemose
#20 Feb 2013
######

##step1: initialise inputs and outputs
library("parallel")
options(scipen=20)
btpath = "/data/smew2/altemose/software/bedtools2/bin/bedtools"
rep1=3
rep2=4
genomic=5

args=commandArgs(TRUE)
datapath = args[2]
```

```
sample = args[3]
wide=args[4]
slide=args[5]
rep1suffix = args[6]
rep2suffix = args[7]
genomicsuffix = args[8]
qual = as.integer(args[9])



#example hardwired input
#datapath="/data/smew2/altemose/MouseH3K4me3/PWDB6F1"
#sample = "Infertile"
#wide=1000
#slide=100
#rep1suffix = 243
#rep2suffix = 245
#genomicsuffix = 241
#qual=1







chrs=seq(1,19,1) #change if necessary
chrs=c(chrs,"X")

outfile1 = paste("FinalOutput/Constants.q",qual,".",sample,".",wide,"wide.",slide,"slide.bed",sep="")




getConstants=function(chr){

posfileA = paste(datapath,"/bychr/FragPos.q",qual,".",rep1suffix,".chr.chr",chr,".bed",sep="")
posfileB = paste(datapath,"/bychr/FragPos.q",qual,".",rep2suffix,".chr.chr",chr,".bed",sep="")
posfileG= paste(datapath,"/bychr/FragPos.q",qual,".",genomicsuffix,".chr.chr",chr,".bed",sep="")

windowfile = paste("bychr/chr",chr,".windows.",wide,"wide.",slide,"slide.bed",sep="")

infile1=paste("tmp/Temp0.FragDepth.",sample,".q",qual,".",rep1suffix,".chr",chr,".",wide,"wide.",slide,"slide.bed",sep="")
infile2=paste("tmp/Temp0.FragDepth.",sample,".q",qual,".",rep2suffix,".chr",chr,".",wide,"wide.",slide,"slide.bed",sep="")
infile3=paste("tmp/Temp0.FragDepth.",sample,".q",qual,".",genomicsuffix,".chr",chr,".",wide,"wide.",slide,"slide.bed",sep="")


##step2: load data, estimate constants alpha1, alpha2, and beta
```

```
63
64
65 system(paste(btpath," coverage -a ",posfileA," -b ",windowfile," -
       counts >",infile1,sep=""))
66 system(paste(btpath," coverage -a ",posfileB," -b ",windowfile," -
       counts >",infile2,sep=""))
67 system(paste(btpath," coverage -a ",posfileG," -b ",windowfile," -
       counts >",infile3,sep=""))
68
69
70 counts = read.table(infile1,header=FALSE,colClasses=c('NULL','integer'
       ,'integer','integer'))
71
72 counts=counts[order(counts[,1]),]
73 countstemp = read.table(infile2,header=FALSE,colClasses=c('NULL','
       integer','NULL','integer'))
74 countstemp=countstemp[order(countstemp[,1]),]
75 counts[,4]=countstemp[,2]
76 countstemp = read.table(infile3,header=FALSE,colClasses=c('NULL','
       integer','NULL','integer'))
77 countstemp=countstemp[order(countstemp[,1]),]
78 counts[,5]=countstemp[,2]
79 rm(countstemp)
80
81 alpha1.est = sum(counts[(counts[,rep2]==0),rep1])/sum(counts[(counts[,
       rep2]==0),genomic])
82 alpha2.est = sum(counts[(counts[,rep1]==0),rep2])/sum(counts[(counts[,
       rep1]==0),genomic])
83 beta.est0 = (mean(counts[,rep2])-alpha2.est*mean(counts[,genomic]))/(
       mean(counts[,rep1])-alpha1.est*mean(counts[,genomic]))
84
85 zeroregions1=sum((counts[,rep2]==0) & (counts[,rep1] + counts[,genomic
       ])>0)
86 zeroregions2=sum((counts[,rep1]==0) & (counts[,rep2] + counts[,genomic
       ])>0)
87
88
89 #counts=counts[counts[,genomic]>0,] #remove regions with 0 genomic
       coverage
90 counts[(counts[,genomic]==0 & (counts[,rep1]+counts[,rep2])>0),genomic
       ]=0.5 #pseudocount regions with 0 genomic coverage and >0 IP
       coverage to have genomic coverage of 0.5
91 counts=counts[counts[,genomic]>0,]
92
93
94 ##step3: find initial set of p-values, find confident set of peaks, re
       -do estimate of beta and redo p-value calls
95
96
97 peaks=makepeaks(counts,alpha1=alpha1.est,alpha2=alpha2.est,beta=beta.
       est0,r1=rep1,r2=rep2,g=genomic)
98
99 q=which(!is.na(peaks[,"p-value"]) & peaks[,"p-value"]<1e-10 & peaks[,"
       yhat_alt"]>5)
100 if(length(q)<100){
```

```
  q=which(!is.na(peaks[,"p-value"]) & peaks[,"p-value"]<1e-5 & peaks[,"yhat_alt"]>1)
}
rm(peaks)

beta.est = (mean(counts[q,rep2])-alpha2.est*mean(counts[q,genomic]))/(mean(counts[q,rep1])-alpha1.est*mean(counts[q,genomic]))


peaks=makepeaks(counts,alpha1=alpha1.est,alpha2=alpha2.est,beta=beta.est,r1=rep1,r2=rep2,g=genomic)
gthresh = quantile(peaks[,"cov_g"],0.999)


#Step4: return constant values

r1comb=mean(peaks[,"bhat_alt"]*(peaks[,"yhat_alt"]+alpha1.est))
r1sig=mean(peaks[,"bhat_alt"]*peaks[,"yhat_alt"])
r2comb=mean(peaks[,"bhat_alt"]*(beta.est*peaks[,"yhat_alt"]+alpha2.est))
r2sig=mean(peaks[,"bhat_alt"]*peaks[,"yhat_alt"]*beta.est)


return(c(chr,alpha1.est,alpha2.est,beta.est,mean(counts[,genomic]),mean(counts[,rep1]),mean(counts[,rep2]),as.integer(dim(counts)[1]),zeroregions1,zeroregions2,length(q),r1sig/r1comb,r2sig/r2comb,gthresh))

}


##step5: declare functions to find MLE values for each window

makepeaks=function(test=counts,alpha1,alpha2,beta,r1=rep1,r2=rep2,g=genomic){

  sumcov = test[,r1]+test[,r2]+test[,g]

  term1=(sumcov)*(beta+1)
  term2=1+alpha1+alpha2
  term3=beta+1
  term4=beta*alpha1*test[,r2]+alpha2*test[,r1]
  term5=beta*(test[,r1]+test[,r2])
  term6=alpha1*alpha2
  term7=alpha1*beta+alpha2

  aterm=term1*beta-term3*term5
  bterm=term1*term7-term2*term5-term3*term4
  cterm=term1*term6-term2*term4

  rm(term1,term2,term3,term4,term5,term6,term7)

  yvals=(-bterm+sqrt(bterm^2-4*aterm*cterm))/2/aterm
  rm(aterm,bterm,cterm)

```

```
148    yvals[yvals<0]=0
149    yvals[yvals>1e9]=1e9
150
151    bvals=(sumcov)/(1+alpha1+alpha2+(beta+1)*yvals)
152
153    bvalsnull=(sumcov)/(1+alpha1+alpha2)
154    yvalsnull=rep(0,length(bvalsnull))
155
156    lhooddiff=2*(lhood(yvals,bvals,test,alpha1,alpha2,beta,r1,r2,g)-
           lhood(yvalsnull,bvalsnull,test,alpha1,alpha2,beta,r1,r2,g))
157
158   rm(bvalsnull,yvalsnull)
159
160    signif=pchisq(lhooddiff,df=1,lower.tail=F)
161
162    results=cbind(test[,1],test[,2],yvals,signif,lhooddiff,test[,r1],
           test[,r2],test[,g],bvals)
163    colnames(results)=c("start","stop","yhat_alt","p-value","Lhood_diff"
           ,"cov_r1","cov_r2","cov_g","bhat_alt")
164    return(results)
165
166 }
167 lhood=function(yhat,bhat,test,alpha1,alpha2,beta,r1=rep1,r2=rep2,g=
       genomic){
168    sumcov = test[,r1]+test[,r2]+test[,g]
169    ourterm=sumcov*(log(bhat)-1)+test[,r1]*log(alpha1+yhat)+test[,r2]*
           log(alpha2+yhat*beta)
170    return(ourterm)
171 }
172
173
174 coln= c("chr","alpha1","alpha2","beta","meancovgenomic","meancovrep1",
       "meancovrep2","totalnonzerobins","alpha1trainingregions","
       alpha2trainingregions","betatrainingregions","rep1signal","
       rep2signal","genomiccov999thpctile")
175
176 print(date())
177 data=mclapply(chrs,getConstants,mc.preschedule=TRUE,mc.cores=20)
178 data2=t(simplify2array(data))
179 data2[,1]=unlist(lapply(data,function(x) as.character(x[[1]])))
180
181 data2=rbind(data2,c("autosomal",rep("NA",13)))
182 for(m in c(2,3,4,5,6,7,12,13,14)){
183    data2[21,m]=weighted.mean(as.numeric(data2[1:19,m]),as.numeric(data2
           [1:19,8]))
184 }
185 for(m in c(8,9,10,11)){
186    data2[21,m]=sum(as.numeric(data2[1:19,m]))
187 }
188
189 write.table(data2,file=outfile1,quote=FALSE,sep="\t",row.names=F,col.
       names=coln)
190
191 print(paste("printed results to",outfile1))
192 print(date())
```

```
quit(save="no",runLast=FALSE)
```

text/EstimateConstants.R

```
#ForceCall.R
#implementation of an algorithm to perform LR testing of ChIP-seq data at a fixed set of sites
#by Nicolas Altemose
#10 November 2015
######

##Initialise inputs and outputs

library("parallel")
options(scipen=20)
btpath = "/data/smew2/altemose/software/bedtools2/bin/bedtools" #path to bedtools executable
rep1=3 #column in which fragment depth for IP replicate 1 is reported
rep2=4 #column in which fragment depth for IP replicate 2 is reported
genomic=5 #column in which fragment depth for genomic input is reported
chrs=c(seq(1,19,1),"X")
system("mkdir tmp",ignore.stdout = T, ignore.stderr = T)

args=commandArgs(TRUE)
constfile = args[2] #path to a file containing genome-wide estimates for constants alpha1/2 & beta (output of GetConstants.r)
bedfilebase = args[3] #path and base name for a 3-column bed file listing positions of windows in which to do force-calling, split by chromosome
posfileIP1base = args[4] #path and base name for files listing fragment positions for IP replicate 1 (ending in e.g. "chr1.bed")
posfileIP2base = args[5] #path and base name for files listing fragment positions for IP replicate 2 (ending in e.g. "chr1.bed")
posfileGbase = args[6] #path and base name for files listing fragment positions for genomic input (ending in e.g. "chr1.bed")
outfile = args[7] #name of outputfile

##example hardwired input
#constfile = "Constants.q1.Infertile.1000wide.100slide.bed"
#bedfilebase = "tmp/dmc1hotspots_PWDB6F1.PRDM9pb.txt.1kbslop"
#posfileIP1base = "/data/smew2/altemose/MouseH3K4me3/PWDB6F1/bychr/FragDepth.q0.243.chr"
#posfileIP2base = "/data/smew2/altemose/MouseH3K4me3/PWDB6F1/bychr/FragDepth.q0.245.chr"
#posfileGbase = "/data/smew2/altemose/MouseH3K4me3/PWDB6F1/bychr/FragDepth.q0.241.chr"
#outfile = "ForceCall.H3K4me3.1kb.PWDB6F1.PRDM9pb.Infertile.txt"

#read in constants
constdata=read.table(constfile,header=TRUE)

alpha1.est=constdata[dim(constdata)[1],2]
alpha2.est=constdata[dim(constdata)[1],3]
beta.est=constdata[dim(constdata)[1],4]
```

```
40
41
42 #declare function for each chromosome
43 getEnrichments=function(chr){
44
45 posfileIP1 = paste(posfileIP1base,".chr",chr,".bed",sep="")
46 posfileIP2 = paste(posfileIP2base,".chr",chr,".bed",sep="")
47 posfileG= paste(posfileGbase,".chr",chr,".bed",sep="")
48
49 bedfile= paste(bedfilebase,".chr",chr,".bed",sep="")
50
51 tempfile1=paste("/tmp/",outfile,".chr",chr,".temp1.bed",sep="")
52 tempfile2=paste("/tmp/",outfile,".chr",chr,".temp2.bed",sep="")
53 tempfileG=paste("/tmp/",outfile,".chr",chr,".tempG.bed",sep="")
54
55 outfiletemp = paste("tmp/",outfile,".chr",chr,".bed",sep="")
56
57
58 ## now get coverage at windows centred on DSB midpoints
59
60 system(paste(btpath," coverage -a ",posfileIP1," -b ",bedfile," -
    counts >",tempfile1,sep=""))
61 system(paste(btpath," coverage -a ",posfileIP2," -b ",bedfile," -
    counts >",tempfile2,sep=""))
62 system(paste(btpath," coverage -a ",posfileG," -b ",bedfile," -counts
    >",tempfileG,sep=""))
63
64
65 #print(paste("reading in results and computing constants for DSB
    regions.",date()))
66 counts = read.table(tempfile1,header=FALSE,colClasses=c('character','
    integer','integer','integer'))
67
68 counts=counts[order(counts[,2]),]
69 countstemp = read.table(tempfile2,header=FALSE,colClasses=c('NULL','
    integer','NULL','integer'))
70 countstemp=countstemp[order(countstemp[,1]),]
71 counts[,5]=countstemp[,2]
72 countstemp = read.table(tempfileG,header=FALSE,colClasses=c('NULL','
    integer','NULL','integer'))
73 countstemp=countstemp[order(countstemp[,1]),]
74 counts[,6]=countstemp[,2]
75 rm(countstemp)
76
77 countsfilt=counts[,2:6]
78 countsfilt[(countsfilt[,genomic]==0 & (countsfilt[,rep1]+countsfilt[,
    rep2])>0),genomic]=0.5 #pseudocount regions with 0 genomic
    coverage and >0 IP coverage to have genomic coverage of 0.5
79 countsfilt=countsfilt[countsfilt[,genomic]>0,]
80
81 peaks=makepeaks(countsfilt,alpha1=alpha1.est,alpha2=alpha2.est,beta=
    beta.est,r1=rep1,r2=rep2,g=genomic)
82
83 validcount = 0
84 for (i in 1:dim(counts)[1]){
```

```
85    startpos = counts[i,2]
86    peakinfo = peaks[peaks[,1]==startpos,]
87    if(length(peakinfo)>0){
88      counts[i,7]=peakinfo[3]
89      counts[i,8]=peakinfo[5]
90      counts[i,9]=peakinfo[4]
91      validcount=validcount+1
92    }else{
93      counts[i,7]="NA"
94      counts[i,8]="NA"
95      counts[i,9]="NA"
96    }
97  }
98
99  colnames(counts)=c("chr","start","stop","cov_r1","cov_r2","cov_g","
      enrichment","Lhood_diff","p-value")
100
101
102 ##write final output file
103
104 options(scipen=8)
105 write.table(counts,file=outfiletemp,quote=FALSE,sep="\t",row.names=F,
      col.names=T)
106
107 return(1)
108
109 }
110
111
112 makepeaks=function(test=counts,alpha1,alpha2,beta,r1=rep1,r2=rep2,g=
      genomic){
113
114   sumcov = test[,r1]+test[,r2]+test[,g]
115
116   term1=(sumcov)*(beta+1)
117   term2=1+alpha1+alpha2
118   term3=beta+1
119   term4=beta*alpha1*test[,r2]+alpha2*test[,r1]
120   term5=beta*(test[,r1]+test[,r2])
121   term6=alpha1*alpha2
122   term7=alpha1*beta+alpha2
123
124   aterm=term1*beta-term3*term5
125   bterm=term1*term7-term2*term5-term3*term4
126   cterm=term1*term6-term2*term4
127
128   rm(term1,term4,term5,term6,term7)
129
130   yvals=(-bterm+sqrt(bterm^2-4*aterm*cterm))/2/aterm
131   rm(aterm,bterm,cterm)
132
133   yvals[yvals<0]=0
134   yvals[yvals>1e9]=1e9
135
136   bvals=(sumcov)/(term2+term3*yvals)
```

```
137
138   bvalsnull=(sumcov)/(term2)
139   yvalsnull=rep(0,length(bvalsnull))
140
141   lhooddiff=2*(lhood(yvals,bvals,test,alpha1,alpha2,beta,r1,r2,g)-
          lhood(yvalsnull,bvalsnull,test,alpha1,alpha2,beta,r1,r2,g))
142
143   rm(bvalsnull,yvalsnull)
144
145   signif=pchisq(lhooddiff,df=1,lower.tail=F)
146
147   results=cbind(test[,1],test[,2],yvals,signif,lhooddiff,test[,r1],
          test[,r2],test[,g])
148   colnames(results)=c("start","stop","yhat_alt","p-value","Lhood_diff"
          ,"cov_r1","cov_r2","cov_g")
149   return(results)
150
151 }
152 lhood=function(yhat,bhat,test,alpha1,alpha2,beta,r1=rep1,r2=rep2,g=
      genomic){
153   sumcov = test[,r1]+test[,r2]+test[,g]
154   ourterm=sumcov*(log(bhat)-1)+test[,r1]*log(alpha1+yhat)+test[,r2]*
          log(alpha2+yhat*beta)
155   return(ourterm)
156 }
157
158
159 #####run all chromosomes in parallel
160
161 print(date())
162 funfunc = mclapply(chrs,getEnrichments,mc.preschedule=TRUE,mc.cores
      =20)
163 print(paste("done!:",date()))
164
165
166 finaltable=read.table(paste("tmp/",outfile,".chr1.bed",sep=""), header
      =T)
167 for(chr in chrs[2:length(chrs)]){
168   outfiletemp = paste("tmp/",outfile,".chr",chr,".bed",sep="")
169   tempdata1=read.table(outfiletemp, header=T)
170   finaltable=rbind(finaltable,tempdata1)
171 }
172
173 write.table(finaltable,file=paste(outfile,".bed",sep=""),sep="\t",col.
      names=T,row.names=F,quote=F)
174
175 colnames(finaltable)=c("chr","start","stop","cov_r1","cov_r2","cov_g",
      "enrichment","Lhood_diff","p-value")
176
177 quit(save="no",runLast=FALSE)
```

text/ForceCall.R

```
1 #GetHaplotypeReadCounts.R
2 #implementation of an algorithm to identify fractionB6 at each hotspot
```

```
    in hybrid mice
#by Nicolas Altemose
#10 November 2015
######

library("Rsamtools")
library("inline")
library("Rcpp")
source("SNPExtractionFunctions.R")
library("parallel")
rm(reformatReads);reformatReads <- cxxfunction(signature(variablesIn="
    list"),cppReformatReads,plugin="RcppArmadillo")

options(scipen=20)
btpath = "/data/smew2/altemose/software/bedtools2/bin/bedtools"
bqFilter=20 # only use bases with bq greater than this number
iSizeUpperLimit=10000 # don't use reads with mate pairs this far apart
pthresh=0.01
mapqthresh=1

args=commandArgs(TRUE)
datapath = args[2]
sample = args[3]
wide = args[4]
slide = args[5]
rep1suffix = args[6]
rep2suffix = args[7]
genomicsuffix = args[8]
qual = args[9]


#datapath="/data/smew2/altemose/MouseH3K4me3/B6", datapath="/data/
    smew2/altemose/MouseH3K4me3/PWDB6F1"
#sample = "Reciprocal"
#wide=1000
#slide=100
#rep1suffix=285
#rep2suffix=285
#genomicsuffix=281
#qual=1




chrs=seq(1,19,1)
chrs=c(chrs,"X")
chrin=1


for(ch in 1:length(chrs)){

chrin=chrs[ch]

print(chrin)
```

```
55 print(date())
56
57 rep1file = paste(datapath,"/merged.rmdup.",rep1suffix,".bam",sep="")
58 rep2file = paste(datapath,"/merged.rmdup.",rep2suffix,".bam",sep="")
59 genomicfile = paste(datapath,"/merged.rmdup.",genomicsuffix,".bam",sep
       ="")
60 vcffile = paste("bychr/PWDsnps.",sample,"DSBs.chr",chrin,".vcf",sep=""
       )
61 regionfile = paste("FinalOutput/FinalDSBPeakRegions.q",qual,".",sample
       ,".chr",chrin,".",wide,"wide.",slide,"slide.bed",sep="")
62 outfile = paste("FinalOutput/HaplotypesPlusEnrichmentDSBs.ndr.q",qual,
       ".",sample,".chr",chrin,".",wide,"wide.",slide,"slide.bed",sep="")
63
64 chr=paste("chr",chrin,sep="") # chromosome of interest
65 regions = read.table(regionfile,header=TRUE)
66 vcf=read.table(vcffile,colClasses=c('factor','integer','NULL','factor'
       ,'factor','NULL','NULL','NULL','NULL','character'),sep="\t")
67 vcf = subset(vcf,grepl("1/1",vcf[,5])==TRUE)
68
69 #chr=chr,regions=regions,vcf=vcf,genomicfile=genomicfile,rep1file=
       rep1file,rep2file=rep2file,mapqthresh=mapqthresh
70
71 combineCounts=function(m){
72   result1=c(regions[m,],rep(NA,14))
73   if(regions[m,9]<300){
74     regionStart = regions[m,2]
75     regionEnd = regions[m,3]
76
77     #print(regionStart)
78     L=as.integer(vcf[,2])
79     which=L>regionStart & L<regionEnd & nchar(as.character(vcf[,3]))
         ==1 # also only bi-allelic SNPs - c++ code doesn't work
         otherwise
80     L=L[which]
81     ref=as.character(vcf[which,3])
82     alt=as.character(vcf[which,4])
83     T=as.integer(length(L))
84
85     midpoint=floor(regionStart+(regionEnd-regionStart)/2)
86     ndrstart=midpoint-60
87     ndrend=midpoint+60
88     ndrL=as.integer(vcf[,2])
89     ndrwhich=ndrL>ndrstart & ndrL<ndrend & nchar(as.character(vcf[,3])
         )==1 # also only bi-allelic SNPs - c++ code doesn't work
         otherwise
90     ndrL=ndrL[ndrwhich]
91     ndrnum=as.integer(length(ndrL))
92
93     inputcounts=c(0,0,0,0)
94     if(regions[m,9]>0){
95       inputData=getReadInformation(bamName=genomicfile,ref=ref,alt=alt
           ,L=L,T=T,mapqthresh=mapqthresh,chr=chr,regionStart=
           regionStart,regionEnd=regionEnd)
96       inputcounts = getRegionCounts(inputData)
97     }
```

```
98
99      IP1counts=c(0,0,0,0)
100     if(regions[m,7]>0){
101       IP1Data=getReadInformation(bamName=rep1file,ref=ref,alt=alt,L=L,
              T=T,mapqthresh=mapqthresh,chr=chr,regionStart=regionStart,
              regionEnd=regionEnd)
102       IP1counts = getRegionCounts(IP1Data)
103     }
104
105     IP2counts=c(0,0,0,0)
106     if(regions[m,8]>0){
107       IP2Data=getReadInformation(bamName=rep2file,ref=ref,alt=alt,L=L,
              T=T,mapqthresh=mapqthresh,chr=chr,regionStart=regionStart,
              regionEnd=regionEnd)
108       IP2counts = getRegionCounts(IP2Data)
109     }
110
111     result1=c(regions[m,],IP1counts,IP2counts,inputcounts,T,ndrnum)
112   }
113
114   return(result1)
115 }
116
117 pos=as.list(seq(1,dim(regions)[1],1))
118
119 data=mclapply(pos,combineCounts,mc.preschedule=TRUE,mc.cores=16)
120 data2=t(simplify2array(data))
121 data2[,1]=unlist(lapply(data,function(x) as.character(x[[1]])))
122 print(date())
123
124 options(scipen=8)
125 write.table(data2,file=outfile,quote=FALSE,sep="\t",row.names=F,col.
      names=c(names(regions),"IP1.B6reads","IP1.PWDreads","IP1.
      noSNPreads","IP1.conflictreads","IP2.B6reads","IP2.PWDreads","IP2.
      noSNPreads","IP2.conflictreads","input.B6reads","input.PWDreads","
      input.noSNPreads","input.conflictreads","snps.region","snps.ndr"))
126 #write.table(data2,file=outfile,quote=FALSE,sep="\t",row.names=F,col.
      names=F)
127
128 }
129
130
131
132
133
134 quit(save="no",runLast=FALSE)
```

text/GetHaplotypeReadCounts.R

```
1 #SNPExtractionFunctions.R
2 #code written by Robert Davies to parse read information, to be used
      with GetHaplotypeReadCounts.R
3
4 ###
5 ### c++ function to be called within R
```

```
###

library("Rsamtools")
library("inline")
library("Rcpp")

###
### c++ function to be called within R
###
cppReformatReads <- '
#include <cmath>
#include <vector>
#include <iostream>
#include <algorithm>
#include <time.h>
//
// only a single list with everything comes in - declare off this list
//
  Rcpp::List cvariablesIn(variablesIn);
//
// now subdivide into components in c++
//
  arma::ivec firstReadSpot = as<arma::ivec>(cvariablesIn["
      firstReadSpot"]);
  arma::ivec secondReadSpot = as<arma::ivec>(cvariablesIn["
      secondReadSpot"]);
  arma::ivec posRead = as<arma::ivec>(cvariablesIn["posRead"]);
  const int numberOfReads = as<int>(cvariablesIn["numberOfReads"]);
  const int T = as<int>(cvariablesIn["T"]);
  arma::ivec L = as<arma::ivec>(cvariablesIn["L"]);
  Rcpp::CharacterVector seqRead = as<Rcpp::CharacterVector>(
      cvariablesIn["seqRead"]);
  Rcpp::CharacterVector qualRead = as<Rcpp::CharacterVector>(
      cvariablesIn["qualRead"]);
  Rcpp::CharacterVector ref = as<Rcpp::CharacterVector>(cvariablesIn["
      ref"]);
  Rcpp::CharacterVector alt = as<Rcpp::CharacterVector>(cvariablesIn["
      alt"]);
  Rcpp::List splitCigarRead = as<Rcpp::List>(cvariablesIn["
      splitCigarRead"]);
  arma::ivec lengthOfSplitCigarRead = as<arma::ivec>(cvariablesIn["
      lengthOfSplitCigarRead"]);
  arma::ivec iSizeTooBigRead = as<arma::ivec>(cvariablesIn["
      iSizeTooBigRead"]);
  arma::ivec mapq = as<arma::ivec>(cvariablesIn["mapq"]);
  const int bqFilter = as<int>(cvariablesIn["bqFilter"]);
//
// new variables
//
  int x1, x2, y, i, iPair, iM, iRead, t;
  int nSNPInRead = 0;
  int nReadSpanningSNPs = 0;
  char s;
  double d, phiTemp, phiRef, phiAlt, phi;
  Rcpp::List sampleReads;
```

```
52    arma::ivec seqLocal(200);
53    arma::ivec qualLocal(200);
54    arma::ivec posLocal(200); // there shouldnt be this many SNPs
55    int refPosition, refOffset, strandOffset;
56    int iNumOfMs, readPosOverall, loopEnd, cigarLength;
57    Rcpp::CharacterVector cigarType(1);
58    Rcpp::CharacterVector cigar2(1);
59    int iU, nU, iAll, curPos, whileVar, localbq;
60    int tMin=0;
61    int tMax=0; // left and right boundaries of what SNPs to look at
62  //
63  // loop over each pair of reads
64  //
65    for(iRead=0; iRead<=numberOfReads-1; iRead++)
66  //  for(iRead=16909; iRead<=16909; iRead++)
67    {
68  //std::cout << "iRead " << iRead <<  "\\n";
69      nSNPInRead=-1; // reset to 0
70      // only use if the insert size is within acceptable margins
71      if(iSizeTooBigRead(firstReadSpot(iRead))==0)
72      {
73        // determine whether there are 2 or 1 read
74        loopEnd=1;
75        if(secondReadSpot(iRead)==-1) // no second read
76          loopEnd=0; // ergo only loop over one read
77        // loop twice over the two parts of the read
78        for(iPair=0; iPair<=loopEnd; iPair++)
79        {
80          // set position in reads of current read
81          if(iPair==0)
82            readPosOverall=firstReadSpot[iRead];
83          if(iPair==1)
84            readPosOverall=secondReadSpot[iRead];
85          //
86          // also as the GATK bounds BQ my MQ, and only uses BQ>17, only
                use mapq>17
87          //
88          if(mapq(readPosOverall)>=bqFilter)
89          {
90            //
91            // for this read, find eligible SNPs (tStart, tEnd)
92            //
93            double curPos;
94            if(loopEnd==0) // only one read
95              curPos=posRead(firstReadSpot(iRead));
96            if(loopEnd==1)
97              curPos=(posRead(firstReadSpot(iRead))+posRead(
                    secondReadSpot(iRead)))/2;
98            // now move tMin forward until A) first within 10000 and B)
                  not in front of SNP
99            whileVar=0;
100           while(whileVar==0)
101           {
102             // dont continue if too far
103             if(tMin<(T-1))
```

```
104             {
105               // continue while no more than 10000 bp before
106               if((2000+ L(tMin))<curPos)
107               {
108                 tMin++;
109               } else {
110                 whileVar=1; // break loop - done!
111               }
112             } else {
113               whileVar=1; // break loop
114             }
115           }
116           // now - go right until >1000 bp away
117           whileVar=0;
118           while(whileVar==0)
119           {
120             // dont continue if too far
121             if(tMax<(T-1))
122             {
123               // continue while no more than 1000 bp after
124               if(( L(tMax)-2000)<curPos)
125               {
126                 tMax++;
127               } else {
128                 whileVar=1; // break loop - done!
129               }
130             } else {
131               whileVar=1; // break loop
132             }
133           }
134           //  if(iRead % 10 == 0 && iRead<200)
135           //
136           //
137           // now, for this read, calculate whether there are SNPs
138           //
139           refPosition=posRead(readPosOverall);
140           Rcpp::List cigarReadInfo = as<Rcpp::List>(splitCigarRead(
                readPosOverall));
141           iNumOfMs = lengthOfSplitCigarRead(readPosOverall);
142           // set some things
143           refOffset=0; // offset against the reference sequence
144           strandOffset=0; // offset in the strand
145           // get cigar info from the read
146           arma::ivec cigarLengthVec = as<arma::ivec>(cigarReadInfo(0))
                ;
147           Rcpp::CharacterVector cigarTypeVec = as<Rcpp::
                CharacterVector>(cigarReadInfo(1));
148           //
149           // now, loop over each part of the read (M, D=del, I=ins)
150           //
151           for(iM=0;iM<=iNumOfMs;iM++)
152           {
153             cigarLength=cigarLengthVec(iM);
154             cigarType(0)=cigarTypeVec(iM);
155             // if its an M - scan
```

```
156             if(cigarType(0)=="M")
157             {
158               x1 = refPosition + refOffset; // left part of M
159               x2 = refPosition + refOffset + cigarLength-1; // right
                      part of M
160               for(t=tMin; t<=tMax; t++) // determine whether that snps
                      is spanned by the read
161               {
162                 y = L[t];
163                 if(x1 <= y && y <= x2) // if this is true - have a SNP
                        !
164                 {
165                   s = seqRead[readPosOverall][y-refPosition-refOffset+
                          strandOffset];
166                   // check if ref or ALT - only keep if true
167                   // also only use if BQ at least bqFilter (17) (as in
                          17 or greater)
168                   localbq=int(qualRead[readPosOverall][y-refPosition-
                          refOffset+strandOffset])-33;
169                   // also bound BQ above by MQ
170                   if(localbq>mapq(readPosOverall)) // if greater, than
                          reduce
171                     localbq=mapq(readPosOverall);
172                   if((s==ref[t][0] || s==alt[t][0]) && (localbq>=
                          bqFilter))
173                   {
174                     // is this the reference or alternate?
175                     nSNPInRead = nSNPInRead+1;
176                     if(s==ref[t][0])
177                       seqLocal[nSNPInRead] = 0;
178                     if(s==alt[t][0])
179                       seqLocal[nSNPInRead] = 1;
180                     qualLocal[nSNPInRead] = localbq;
181                     posLocal[nSNPInRead] = t;
182                   } // end of check if ref or alt
183                 } // end of whether this SNP intersects read
184               } // end of loop on SNP
185               // now, bump up ref and pos offset by x1
186               refOffset=refOffset + cigarLength;
187               strandOffset=strandOffset + cigarLength;
188             } // end of if statement on whether cigar type is M
189             // if it is an insertion - bump the strand offset
190             if(cigarType(0)=="I")
191               strandOffset=strandOffset+cigarLength;
192             // if it is a deletion - bump the reference position
193             if(cigarType(0)=="D")
194               refOffset=refOffset+cigarLength;
195           } // close for loop on each M type within read
196         } // end of check on mapping quality of this read
197       } // end of loop on 1 or two reads
198       if(nSNPInRead > -1) // save result!
199       {
200         arma::ivec pR = posLocal.subvec(0,nSNPInRead); // position (R
              means Read)
201         arma::ivec sR = seqLocal.subvec(0,nSNPInRead); // sequence
```

```
202         arma::ivec qR = qualLocal.subvec(0,nSNPInRead); // quality
203         // get average physical location
204         //
205         // turn this into one unique value per SNP
206         //
207         arma::ivec pRU = arma::unique(pR); // the U means unique
208         nU = pRU.n_elem-1; // 0-based length of pRU - ie (n)umber of (
                U)nique SNPs in read
209         arma::vec phiU(nU+1); // keep phis here
210         for(iU=0; iU<=nU; iU++) // for each unique entry
211         {
212           // reset phis
213           phiAlt=1;
214           phiRef=1;
215           // go through all elements looking for that SNP
216           for(iAll=0; iAll<=nSNPInRead; iAll++)
217           {
218             if(pR(iAll)==pRU(iU)) // there is a match - consider
219             {
220               // turn into phi
221               //   calculate probability from phred scale
222               phiTemp= 1 - pow(10,-(double(qR(iAll))/10));
223               if(qR(iAll)<=0)
224                 phiTemp=0.5; // BQ = 0 so no probability so make it
                      0.5
225               // scale to appropriate base
226               phi = (1-phiTemp) * ( 1-sR(iAll)) + phiTemp * sR(iAll);
227               phiAlt=phiAlt * phi;
228               phiRef=phiRef * (1-phi);
229             } // end if statement on whether there is a match
230           } // end of for loop going through all SNPs in the read
231           // now calculate probability - calculate numerator and
                denominator
232           // where for a= product_{i=0} P(alt,i)
233           // where for b= product_{i=0} (1-P(alt,i))
234           // phi = P(alt) = a/(a+b)
235           phiU(iU)=phiAlt/(phiAlt+phiRef); // done!
236         } // end of for loop for each unique element
237         //
238         // get physical position for the read
239         // hmmm - right now in effect weighted by occurence
240         //
241         d=0;
242         for(i=0; i<=nSNPInRead; i++)
243           d = d + L(pR(i));
244         d = d/(1+nSNPInRead);
245         //
246         // save results but dont label list elements to save space
247         //
248         // save a smaller version unless need to debug
249         sampleReads.push_back(Rcpp::List::create(
250           Rcpp::Named("nSNPs")= nU,
251           Rcpp::Named("avPos")= d,
252           Rcpp::Named("snpInL")= pRU,
253           Rcpp::Named("snpProb")= phiU,
```

```
254             Rcpp::Named("iRead")= iRead,
255             Rcpp::Named("frs")= firstReadSpot[iRead],
256             Rcpp::Named("srs")= secondReadSpot[iRead]));
257           //sampleReads.push_back(Rcpp::List::create(iRead,nU,d,phiU,pRU
                ,pR,sR,qR));
258           //sampleReads.push_back(Rcpp::List::create(iRead,nSNPInRead,d,
                phiU,pRU,pR,sR,qR));
259           // save number of SNPs as well
260           nReadSpanningSNPs = nReadSpanningSNPs +1;
261         }  // end of save result
262       } // close if statement on whether or not the insert size is too
            big
263     } // close for loop on read pair
264     //
265     // done
266     //
267     return(wrap(sampleReads));
268 '
269
270
271
272 ###
273 ### for a bam, for a set of SNPs, get information
274 ###
275 getReadInformation=function(bamName,ref,alt,L,T,mapqthresh,chr=chr,
        regionStart=regionStart,regionEnd=regionEnd)
276 {
277   ###
278   ### set some flags and load in a single result
279   ###
280   flag=      scanBamFlag(isPaired = TRUE, isProperPair = NA,
          isUnmappedQuery = FALSE, hasUnmappedMate = FALSE, isMinusStrand
          = NA, isMateMinusStrand = NA, isFirstMateRead = NA,
          isSecondMateRead = NA, isSecondaryAlignment = NA,
          isNotPassingQualityControls = FALSE, isDuplicate = FALSE)
281   ### okay actually start loading here
282   what=c("qname","strand","pos","seq","qual","cigar","isize","mapq")
283   ### set which region of the genome to interrogate
284   eval(parse(text=  (paste("which = RangesList(\"",chr,"\"=IRanges(",
          regionStart,",",regionEnd,"))",sep=""))))
285   #idx=paste(substr(bamName,1,nchar(bamName)-3),"bam.bai",sep="")
286   idx1=paste(substr(bamName,1,nchar(bamName)-3),"bai",sep="")
287   idx2=paste(bamName,".bai",sep="")
288   if(file.exists(idx1)) idx=idx1
289   if(file.exists(idx2)) idx=idx2
290   param=ScanBamParam(flag=flag,which=which,what=what) # define
          parameters
291   sampleData=scanBam(file=bamName,index=idx,param=param) # load the
          data
292   ###
293   ### reformat some things
294   ###
295   # super basic - read or pair
296   qname=sampleData[[1]]$qname
297   qnameUnique=unique(qname)
```

```
298   if(length(qnameUnique)==0) {
299     return(NA)
300   }
301   else{
302   qnameInteger=match(qname,qnameUnique)
303   # get positions - NOTE - THEY ARE 0 BASED
304   firstReadSpot=as.integer(match(1:max(qnameInteger),qnameInteger)-1)
        # first instance
305   y=qnameInteger
306   y[firstReadSpot+1]=NA
307   secondReadSpot=as.integer(match(1:max(qnameInteger),y)-1) # second
        read - MAY BE NA
308   secondReadSpot[is.na(secondReadSpot)  ]=as.integer(-1) # Switch to
        -1 - skip over if -1 in c++ code
309   numberOfReads=as.integer(length(firstReadSpot))
310   # get more info read as well
311   mapq=as.integer(sampleData[[1]]$mapq)
312   posRead=as.integer(sampleData[[1]]$pos)
313   cigarRead=sampleData[[1]]$cigar
314   strandRead=sampleData[[1]]$strand
315   seqRead=as.character(sampleData[[1]]$seq)
316   qualRead=as.character(sampleData[[1]]$qual) # hmm
317   qualRead=as.character(sampleData[[1]]$qual) # hmm
318   iSizeTooBigRead=as.integer(abs(sampleData[[1]]$isize)>
        iSizeUpperLimit | as.integer(sampleData[[1]]$mapq)<mapqthresh)
319   iSizeTooBigRead[is.na(iSizeTooBigRead)==TRUE]=0
320
321   ###
322   ### also, need to reconfigure cigar properly
323   ###
324   splitCigarRead=lapply(1:length(cigarRead), function(x) list(as.
        integer(100),"M"))
325   # also lost of 51M - skip these
326   which= cigarRead=="51M"
327   splitCigarRead[which]=lapply(1:sum(which), function(x) list(as.
        integer(51),"M"))
328   which=cigarRead!="51M" & cigarRead!="100M"
329   splitCigarRead[which]=  lapply(cigarRead[which],function(c) {
330   y=unlist(strsplit(c,""))
331   t1=is.na(as.numeric(y))
332   t2=(1:length(t1))[t1]
333   t3=c(1,t2[-length(t2)]+1)
334   n=length(t2)
335   t5=array(0,n)
336   t6=array("",n)
337   for(i in 1:n)
338   {
339     t5[i]=substr(c,t3[i],t2[i]-1)
340     t6[i]=substr(c,t2[i],t2[i])
341   }
342       return(list(as.integer(t5),t6))
343   })
344   lengthOfSplitCigarRead=as.integer(unlist(lapply(splitCigarRead,
        function(x) length(x[[1]])-1))) # 0 BASED
345   ###
```

```
346   ### push through c++ function
347   ###
348   sampleReads=reformatReads(list(
349     firstReadSpot=firstReadSpot,
350     secondReadSpot=secondReadSpot,
351     posRead=posRead,
352     numberOfReads=numberOfReads,
353     T=T,
354     L=L,
355     seqRead=seqRead,
356     qualRead=qualRead,
357     ref=ref,
358     alt=alt,
359     splitCigarRead=splitCigarRead,
360     lengthOfSplitCigarRead=lengthOfSplitCigarRead,
361     iSizeTooBigRead=iSizeTooBigRead,
362     mapq=mapq,
363     bqFilter=bqFilter))
364   ###
365   ### return something clean
366   ###
367   # quick check
368   a=unlist(lapply(sampleReads,function(x) qname[x$frs+1]))
369   b=unlist(lapply(sampleReads,function(x) {
370     if(x$srs==-1) return(NA) else return(qname[x$srs+1]) }))
371   if(sum(a!=b,na.rm=TRUE)>0) {
372     print("ERROR - problem associating reads");      return(NA)
373   }
374   sampleReads=lapply(sampleReads,function(x) return(x[names(x)!="frs"
        & names(x)!="srs" & names(x)!="iRead"]))
375   names(sampleReads)=a
376   return(sampleReads)
377   }
378 }
379
380
381
382
383
384 getRegionCounts=function(nData)
385 {
386   b6count=0
387   pwdcount=0
388   nosnpcount=0
389   contracount=0
390   if(length(nData)>0){
391     counts=sapply(nData, getSNPcounts,simplify="matrix")
392     nosnpcount=sum(counts[1,]==0 & counts[2,]==0)
393     b6count= sum(counts[1,]>0 & counts[2,]==0)
394     pwdcount= sum(counts[1,]==0 & counts[2,]>0)
395     contracount = sum(counts[1,]>0 & counts[2,]>0)
396   }
397   return(c(b6count,pwdcount,nosnpcount,contracount))
398 }
399
```

```
getSNPcounts = function(nlist){
  bcount=as.integer(sum(nlist$snpProb<=pthresh))
  pcount=as.integer(sum(nlist$snpProb>=(1-pthresh)))
  return (c(bcount,pcount))
}
```

```
rm(reformatReads);reformatReads <- cxxfunction(signature(variablesIn="
    list"),cppReformatReads,plugin="RcppArmadillo")
```

text/SNPExtractionFunctions.R

```
#GetFragDepthSingleBase.pl
#given paired-end BAM input, reports start and end positions of
    fragments and computes coverage depth genome-wide
#by Nick Altemose
#2012

use warnings;
use strict;
my $tic = time;
print "\n\n";

my $usage = "USAGE: samtools view -F12 -q1 <Input.PositionSorted.bam>
    | perl GetFragDepth.SingleBase.pl <FragDepthOutputPrefix> <
    FragPosOutputPrefix> <read length, default 51> <max frag length,
    default 10000> <bedtools path>";


my $outfile1;
my $outfile2;
my $readlen = 51;
my $maxlen = 10000;

my $bedtoolspath = '/data/smew2/altemose/software/bedtools2/bin';


if(defined $ARGV[0] && defined $ARGV[1]){
  $outfile1 = $ARGV[0];
  $outfile2 = $ARGV[1];
  chomp($outfile1);
}
else{
  die "$usage\n";
}
if(defined $ARGV[2]){
  $readlen = $ARGV[2];
  chomp($readlen);
}
if(defined $ARGV[3]){
  $maxlen = $ARGV[3];
```

```
  chomp($maxlen);
}
if(defined $ARGV[4]){
  $bedtoolspath = $ARGV[4];
  chomp($bedtoolspath);
}

#my $bigbedfile = 'bedToBigBed';
#unless(-e $bigbedfile){
# die "ERROR: Could not find $bigbedfile in this directory\n";
#}

my $chrsizefile = 'hg19.chromsizes.tbl';
unless(-e $chrsizefile){
  die "ERROR: Could not find $chrsizefile in this directory\n";
}



open(OUT2, '>'.$outfile2.'.bedtemp');
my @memorylist;
my %memorynames;

my $prevchr = '';
my $prevpos=0;
my $chrnum=0;
my $sort=0;

print "chr\n";
while(my $line = <STDIN>){
  chomp($line);
  if($line=~/^\S+\t\S+\t(\S+)\t(\S+)\t\S+\t\S+\t(\S+)\t(\S+)\t/){
    my $rchr = $1;
    next if($rchr eq '*');
    my $rpos = $2;
    my $pchr = $3;
    my $ppos = $4;

    if($rchr ne $prevchr){
      $chrnum++;
      @memorylist=();
      %memorynames=();
      $prevpos=0;
      print "reading $rchr\n";
    }

    if($rchr eq $pchr || $pchr eq '='){
      if(((abs($ppos-$rpos)+$readlen+1)<=$maxlen)){
        my $start=$rpos;
        if($ppos>$rpos){
          my $stop = $ppos+$readlen-1;
          my $start0 = $start-1;
          print OUT2 "$rchr\t$start0\t$stop\t1\t0\t+\n";
          push(@memorylist,$rpos);
          $memorynames{$rpos}=0;
```

```
91          }
92          else{
93            unless(exists $memorynames{$ppos}){
94              $start = $ppos;
95              my $stop = $rpos+$readlen-1;
96              my $start0 = $start-1;
97              print OUT2 "$rchr\t$start0\t$stop\t1\t0\t+\n";
98            }
99          }
100         $prevpos = $start;
101       }
102     }
103     foreach my $pos(@memorylist){
104       if($pos<($rpos-2*$maxlen)){
105         shift(@memorylist);
106         delete $memorynames{$pos};
107       }
108       else{
109         last;
110       }
111     }
112     $prevchr = $rchr;
113   }
114   elsif($line!~/^\@/){
115     print "ERROR parsing line: $line\n";
116   }
117 }
118 close OUT2;
119
120 print "\nsorting...\n";
121
122 my $sortedoutfile = "$outfile2.sorted.bed";
123 system("sort -k1,1V -k2,2n $outfile2.bedtemp >$sortedoutfile");
124 system("rm -f $outfile2.bedtemp");
125 # system("./bedToBigBed $outfile2.bedtemp.sorted $chrsizefile $
    outfile2.bigbed");
126 # system("tar -chzf $outfile2.bed.tar.gz $outfile2.bedtemp.sorted");
127
128
129 system("$bedtoolspath/bedtools genomecov -bg -i $sortedoutfile -g $
    chrsizefile >$outfile1.bed");
130 #system("$bedtoolspath/bedtools genomecov -bg -5 -i $sortedoutfile -g
    $chrsizefile >$outfile1.5prime.bed");
131 #system("$bedtoolspath/bedtools genomecov -bg -3 -i $sortedoutfile -g
    $chrsizefile >$outfile1.3prime.bed");
132
133
134 #calculate and display runtime
135 my $toc = time;
136 my $elapsed = $toc-$tic;
137 printf("\n\nTotal running time: %02d:%02d:%02d\n\n", int($elapsed /
    3600), int(($elapsed % 3600) / 60), int($elapsed % 60));
```

text/GetFragDepth–SingleBase.pl

References

[1] A Weismann. *Essays on heredity and kindred biological subjects*. 1889.

[2] J G Hussin et al. "Recombination affects accumulation of damaging and disease-associated mutations in human populations". In: *Nature Genetics* 47.4 (2015), pp. 400–404.

[3] A Burt. "Perspective: sex, recombination, and the efficacy of selection—was Weismann right?" In: *Evolution* 54.2 (2000), pp. 337–351.

[4] H J Muller. "The relation of recombination to mutational advance". In: *Mutation Research* 1.1 (1964), pp. 2–9.

[5] J Felsenstein. "The evolutionary advantage of recombination." In: *Genetics* 78.2 (1974), pp. 737–756.

[6] A S Wilkins and R Holliday. "The Evolution of Meiosis From Mitosis". In: *Genetics* 181.1 (2009), pp. 3–12.

[7] I D Adler. "Comparison of the duration of spermatogenesis between male rodents and humans." In: *Mutation Research* 352.1-2 (1996), pp. 169–172.

[8] K I Ishiguro et al. "Meiosis-specific cohesion mediates homolog recognition in mouse spermatocytes". In: *Genes & Development* 28.6 (2014), pp. 594–607.

[9] M A Handel and J C Schimenti. "Genetics of mammalian meiosis: regulation, dynamics and impact on fertility". In: *Nature Reviews Genetics* 11.2 (2010), pp. 124–136.

[10] S Keeney, C N Giroux, and N Kleckner. "Meiosis-specific DNA double-strand breaks are catalyzed by Spo11, a member of a widely conserved protein family". In: *Cell* 88 (1997), pp. 375–384.

[11] F Baudat and B de Massy. "Regulating double-stranded DNA break repair towards crossover or non-crossover during mammalian meiosis". In: *Chromosome Research* 15 (2007), pp. 565–577.

[12] M J Neale, J Pan, and S Keeney. "Endonucleolytic processing of covalent protein-linked DNA double-strand breaks." In: *Nature* 436.7053 (2005), pp. 1053–1057.

[13] K Wei et al. "Inactivation of exonuclease 1 in mice results in DNA mismatch repair defects, increased cancer susceptibility, and male and female sterility". In: *Chromosome Research* 17 (2003), pp. 603–614.

[14] M J Neale and S Keeney. "Clarifying the mechanics of DNA strand exchange in meiotic recombination." In: *Nature* 442.7099 (2006), pp. 153–158.

[15] F Baudat, Y Imai, and B de Massy. "Meiotic recombination in mammals: localization and regulation." In: *Nature Reviews Genetics* 14.11 (2013), pp. 794–806.

[16] S K Mahadevaiah et al. “Recombinational DNA double-strand breaks in mice precede synpasis”. In: *Nature Genetics* 27 (2001), pp. 271–276.

[17] F Pratto et al. “DNA recombination. Recombination initiation maps of individual human genomes.” In: *Science* 346.6211 (2014), p. 1256442.

[18] L Kauppi et al. “Numerical constraints and feedback control of double-strand breaks in mouse meiosis.” In: *Genes & Development* 27.8 (2013), pp. 873–886.

[19] B Alberts et al. *Molecular Biology of the Cell*. New York: Garland Science, 2002.

[20] J K Holloway et al. “MUS81 generates a subset of MLH1-MLH3-independent crossovers in mammalian meiosis”. In: *PLoS Genetics* 4 (2008), e1000186.

[21] F Carofiglio et al. “SPO11-independent DNA repair foci and their role in meiotic silencing.” In: *PLoS Genetics* 9.6 (2013), e1003538.

[22] R Song et al. “Many X-linked microRNAs escape meiotic sex chromosome inactivation.” In: *Nature Genetics* 41.4 (2009), pp. 488–493.

[23] T H Morgan et al. *The Mechanism of Mendelian Heredity*. Holt & Co., 1915.

[24] A Kong et al. “A high-resolution recombination map of the human genome”. In: *Nature* 31.3 (2002), pp. 241–247.

[25] G A T McVean et al. “The fine-scale structure of recombination rate variation in the human genome”. In: *Science* 304.5670 (2004), pp. 581–584.

[26] A G Hinch et al. “The landscape of recombination in African Americans”. In: *Nature* 476.7359 (2011), pp. 170–175.

[27] K W Broman et al. “Comprehensive human genetic maps: individual and sex-speific variation in recombination”. In: *American Journal of Human Genetics* 63.3 (1998), pp. 861–869.

[28] 1000 Genomes Project Consortium. “A map of human genome variation from population-scale sequencing.” In: *Nature* 467.7319 (2010), pp. 1061–1073.

[29] W Winckler et al. “Comparison of fine-scale recombination rates in humans and chimpanzees”. In: *Science* 308.5718 (2005), pp. 107–111.

[30] K Paigen et al. “The recombinational anatomy of a mouse chromosome.” In: *PLoS Genetics* 4.7 (2008), e1000119–e1000119.

[31] A Auton et al. “A Fine-Scale Chimpanzee Genetic Map from Population Sequencing”. In: *Science* 336.6078 (2012), pp. 193–198.

[32] I L Berg et al. “Variants of the protein PRDM9 differentially regulate a set of human meiotic recombination hotspots highly active in African populations”. In: *Proceedings of the National Academy of Sciences of the United States of America* 108.30 (2011), p. 12378.

[33] S Myers et al. “A common sequence motif associated with recombination hot spots and genome instability in humans”. In: *Nature Genetics* 40.9 (2008), pp. 1124–1129.

[34] S Myers et al. “Drive against hotspot motifs in primates implicates the PRDM9 gene in meiotic recombination.” In: *Science* 327.5967 (2010), pp. 876–879.

[35] F Baudat et al. “PRDM9 is a major determinant of meiotic recombination hotspots in humans and mice.” In: *Science* 327.5967 (2010), pp. 836–840.

[36] E D Parvanov, P M Petkov, and K Paigen. “Prdm9 controls activation of mammalian recombination hotspots.” In: *Science* 327.5967 (2010), p. 835.

[37] K Brick et al. “Genetic recombination is directed away from functional genomic elements in mice.” In: *Nature* 485.7400 (2012), pp. 642–645.

[38] K Hayashi, K Yoshida, and Y Matsui. “A histone H3 methyltransferase controls epigenetic events required for meiotic prophase”. In: *Nature* 438.7066 (2005), pp. 374–378.

[39] H Wu et al. “Molecular Basis for the Regulation of the H3K4 Methyltransferase Activity of PRDM9”. In: *Cell Reports* 5.1 (2013), pp. 13–20.

[40] A S McCarty et al. “Selective dimerization of a C2H2 zinc finger subfamily”. In: *Molecular Cell* 11.2 (2003), pp. 459–470.

[41] A V Persikov and M Singh. “An expanded binding model for Cys2His2 zinc finger protein–DNA interfaces”. In: *Physical Biology* 8.3 (2011), p. 035010.

[42] A V Persikov et al. “A systematic survey of the Cys2His2 zinc finger DNA-binding landscape”. In: *Nucleic Acids Research* 43.3 (2015), pp. 1965–1984.

[43] S Iuchi. “C2H2 Zinc Fingers as DNA binding domains”. In: *Zinc Finger Proteins: From Atomic Contact to Cellular Function*. Ed. by S Iuchi and N Kuldell. Springer US, 2005. Chap. 2, pp. 7–13.

[44] Thomas Pringle. “PRDM9: meiosis and recombination”. In: *UCSC Genome Wiki* (August 2011).

[45] A J Jeffreys et al. “Recombination regulator PRDM9 influences the instability of its own coding sequence in humans”. In: *Proceedings of the National Academy of Sciences of the United States of America* 110.2 (2013), pp. 600–605.

[46] International Human Genome Sequencing Consortium. “Initial sequencing and analysis of the human genome”. In: *Nature* 409.6822 (2001), pp. 860–921.

[47] J C Venter et al. “The sequence of the human genome”. In: *Science* 291.5507 (2001), pp. 1304–1351.

[48] H Santos-Rosa et al. “Active genes are tri-methylated at K4 of histone H3.” In: *Nature* 419.6905 (2002), pp. 407–411.

[49] A Pekowska et al. “H3K4 tri-methylation provides an epigenetic signature of active enhancers.” In: *The EMBO Journal* 30.20 (2011), pp. 4198–4210.

[50] F Smagulova et al. “Genome-wide analysis reveals novel molecular features of mouse recombination hotspots”. In: *Nature* 472.7343 (2011), pp. 375–378.

[51] C Grey et al. “Mouse PRDM9 DNA-Binding Specificity Determines Sites of Histone H3 Lysine 4 Trimethylation for Initiation of Meiotic Recombination”. In: *PLoS Biology* 9.10 (2011), e1001176.

[52] J Sollier et al. “Set1 is required for meiotic S-phase onset, double-strand break formation and middle gene expression”. In: *The EMBO Journal* 23.9 (2004), pp. 1957–1967.

[53] J Pan et al. “A hierarchical combination of factors shapes the genome-wide topography of yeast meiotic recombination initiation.” In: *Cell* 144.5 (2011), pp. 719–731.

[54] Z Birtle and C P Ponting. “Meisetz and the birth of the KRAB motif”. In: *Bioinformatics* 22.23 (2006), pp. 2841–2845.

[55] J R Friedman et al. “KAP-1, a novel corepressor for the highly conserved KRAB repression domain.” In: *Genes and Development* 10.16 (1996), pp. 2067–2078.

[56] F L Lim et al. “A KRAB-related domain and a novel transcription repression domain in proteins encoded by SSX genes that are disrupted in human sarcomas.” In: *Oncogene* 17.15 (1998), pp. 2013–2018.

[57] H Kato et al. “SYT associates with human SNF/SWI complexes and the C-terminal region of its fusion partner SSX1 targets histones.” In: *The Journal of Biological Chemistry* 277.7 (2002), pp. 5498–5505.

[58] D R H de Bruijn et al. “The C terminus of the synovial sarcoma-associated SSX proteins interacts with the LIM homeobox protein LHX4.” In: *Oncogene* 27.5 (2008), pp. 653–662.

[59] S Irie et al. “Single-nucleotide polymorphisms of the PRDM9 (MEISETZ) gene in patients with nonobstructive azoospermia.” In: *Journal of Andrology* 30.4 (2009), pp. 426–431.

[60] T Miyamoto et al. “Two single nucleotide polymorphisms in PRDM9 (MEISETZ) gene may be a genetic risk factor for Japanese patients with azoospermia by meiotic arrest.” In: *Journal of Assisted Reproduction and Genetics* 25.11-12 (2008), pp. 553–557.

[61] P L Oliver et al. “Accelerated evolution of the Prdm9 speciation gene across diverse metazoan taxa.” In: *PLoS Genetics* 5.12 (2009), e1000753.

[62] E Axelsson et al. “Death of PRDM9 coincides with stabilization of the recombination landscape in the dog genome”. In: *Genome Research* 22.1 (2011), pp. 51–63.

[63] A Auton et al. “Genetic recombination is targeted towards gene promoter regions in dogs.” In: *PLoS Genetics* 9.12 (2013), e1003984–e1003984.

[64] V Narasimhan et al. “Health and population effects of rare gene knockouts in adult humans with related parents”. In: *bioRxiv* (2015), p. 031641.

[65] G Coop and S Myers. “Live hot, die young: transmission distortion in recombination hotspots”. In: *PLoS Genetics* 3.3 (2007), e35.

[66] A Boulton, R S Myers, and R J Redfield. “The hotspot conversion paradox and the evolution of meiotic recombination.” In: *Proceedings of the National Academy of Sciences of the United States of America* 94.15 (1997), pp. 8058–8063.

[67] C L Baker et al. “PRDM9 Drives Evolutionary Erosion of Hotspots in Mus musculus through Haplotype-Specific Initiation of Meiotic Recombination.” In: *PLoS Genetics* 11.1 (2015), e1004916–e1004916.

[68] Y Lesecque et al. “The red queen model of recombination hotspots evolution in the light of archaic and modern human genomes.” In: *PLoS Genetics* 10.11 (2014), e1004790–e1004790.

[69] F Ubeda and J F Wilkins. “The Red Queen theory of recombination hotspots”. In: *Journal of Evolutionary Biology* 24.3 (2011), pp. 541–553.

[70] J Buard et al. “Diversity of Prdm9 zinc finger array in wild mice unravels new facets of the evolutionary turnover of this coding minisatellite.” In: *PLoS ONE* 9.1 (2014), e85021–e85021.

[71] H Kono et al. “Prdm9 Polymorphism Unveils Mouse Evolutionary Tracks”. In: *DNA Research* 21.3 (2014), pp. 315–326.

[72] I L Berg et al. “PRDM9 variation strongly influences recombination hot-spot acitivity and meiotic instability in humans”. In: *Nature Genetics* 42 (2010), pp. 859–863.

[73] J J Schwartz et al. “Primate evolution of the recombination regulator PRDM9.” In: *Nature Communications* 5 (2014), p. 4370.

[74] A Geraldes et al. “Higher differentiation among subspecies of the house mouse (Mus musculus) in genomic regions with low recombination.” In: *Molecular Ecology* 20.22 (2011), pp. 4722–4736.

[75] O Mihola et al. “A mouse speciation gene encodes a meiotic histone H3 methyltransferase.” In: *Science* 323.5912 (2009), pp. 373–375.

[76] J Forejt and P Ivanyi. “Genetic studies on male sterility of hybrids between laboratory and wild mice (Mus musculus L.)” In: *Genetics Research* 24.02 (1974), pp. 189–206.

[77] J B S Haldane. “Sex ratio and unisexual sterility in hybrid animals”. In: *Journal of Genetics* 12.2 (1922), pp. 101–109.

[78] Z Trachtulec et al. “Physical Map of Mouse Chromosome 17 in the Region Relevant for Positional Cloning of the Hybrid sterility 1 Gene”. In: *Genomics* 23.1 (1994), pp. 132–137.

[79] S Gregorová et al. “Sub-milliMorgan map of the proximal part of mouse chromosome 17 including the hybrid sterility 1 gene”. In: *Mammalian genome : official journal of the International Mammalian Genome Society* 7.2 (1996), pp. 107–113.

[80] Z Trachtulec et al. “Isolation of candidate hybrid sterility 1 genes by cDNA selection in a 1.1 megabase pair region on mouse chromosome 17”. In: *Mammalian Genome* 8.5 (1997), pp. 312–316.

[81] Z Trachtulec et al. “Fine Haplotype Structure of a Chromosome 17 Region in the Laboratory and Wild Mouse”. In: *Genetics* 178.3 (2008), pp. 1777–1784.

[82] T Bhattacharyya et al. “Mechanistic basis of infertility of mouse intersubspecific hybrids.” In: *Proceedings of the National Academy of Sciences of the United States of America* 110.6 (2013), E468–E477.

[83] M Dzur-Gejdosova et al. “Dissecting the genetic architecture of F1 hybrid sterility in house mice”. In: *Evolution* 66.11 (2012), pp. 3321–3335.

[84] P Flachs et al. “Interallelic and intergenic incompatibilities of the prdm9 (hst1) gene in mouse hybrid sterility.” In: *PLoS Genetics* 8.11 (2012), e1003044–e1003044.

[85] T Bhattacharyya et al. “X Chromosome Control of Meiotic Chromosome Synapsis in Mouse Inter-Subspecific Hybrids”. In: *PLoS Genetics* 10.2 (2014), e1004088.

[86] R Storchová et al. “Genetic analysis of X-linked hybrid sterility in the house mouse.” In: *Mammalian Genome : official journal of the International Mammalian Genome Society* 15.7 (2004), pp. 515–524.

[87] S Gregorová et al. “Mouse consomic strains: exploiting genetic divergence between Mus m. musculus and Mus m. domesticus subspecies.” In: *Genome Research* 18.3 (2008), pp. 509–515.

[88] T M Keane et al. “Mouse genomic variation and its effect on phenotypes and gene regulation”. In: *Nature* 477.7364 (2011), pp. 289–294.

[89] V Janoušek et al. “Genome-wide architecture of reproductive isolation in a naturally occurring hybrid zone between Mus musculus musculus and M. m. domesticus”. In: *Molecular Ecology* 21.12 (2012), pp. 3032–3047.

[90] H Yang et al. “Subspecific origin and haplotype diversity in the laboratory mouse”. In: *Nature Genetics* 43.7 (2011), pp. 648–655.

[91] B Yalcin et al. “Next-generation sequencing of experimental mouse strains”. In: *Mammalian Genome* 23.9-10 (2012), pp. 490–498.

[92] M A White et al. “Genetics and evolution of hybrid male sterility in house mice.” In: *Genetics* 191.3 (2012), pp. 917–934.

[93] S Gregorová and J Forejt. “PWD/Ph and PWK/Ph inbred mouse strains of Mus m. musculus subspecies–a valuable resource of phenotypic variations and genomic polymorphisms”. In: *Folia Biol* 46.1 (2000), pp. 31–41.

[94] P Flachs et al. “Prdm9 incompatibility controls oligospermia and delayed fertility but no selfish transmission in mouse intersubspecific hybrids.” In: *PLoS ONE* 9.4 (2014), e95806–e95806.

[95] C L Baker et al. “PRDM9 binding organizes hotspot nucleosomes and limits Holliday junction migration”. In: *Genome Research* 24.5 (2014), pp. 724–732.

[96] P P Khil et al. “Sensitive mapping of recombination hotspots using sequencing-based detection of ssDNA”. In: *Genome Research* 22.5 (2012), pp. 957–965.

[97] M Walker et al. “Affinity-seq detects genome-wide PRDM9 binding sites and reveals the impact of prior chromatin modifications on mammalian recombination hotspot usage.” In: *Epigenetics & Chromatin* 8 (2015), p. 31.

[98] T Billings et al. “DNA binding specificities of the long zinc finger recombination protein PRDM9.” In: *Genome Biology* 14.4 (2013), R35–R35.

[99] F Sun et al. “Nuclear localization of PRDM9 and its role in meiotic chromatin modifications and homologous synapsis”. In: *Chromosoma* (2015), pp. 1–19.

[100] M S Eram et al. “Trimethylation of Histone H3 Lysine 36 by Human Methyltransferase PRDM9 Protein”. In: *Journal of Biological Chemistry* 289.17 (2014), pp. 12177–12188.

[101] W W Quitschke et al. “Differential effect of zinc finger deletions on the binding of CTCF to the promoter of the amyloid precursor protein gene”. In: *Nucleic Acids Research* 28.17 (2000), pp. 3370–3378.

[102] L Liu and D W Heermann. “The interaction of DNA with multi-Cys2His2 zinc finger proteins”. In: *Journal of Physics: Condensed Matter* 27.6 (2015), p. 064107.

[103] M D Griswold. “Making male gametes in culture.” In: *Proceedings of the National Academy of Sciences of the United States of America* 109.42 (2012), pp. 16762–16763.

[104] Y Zhang et al. “Model-based Analysis of ChIP-Seq (MACS)”. In: *Genome Biology* 9.9 (2008), R137.

[105] D S Johnson et al. “Genome-wide mapping of in vivo protein-DNA interactions.” In: *Science* 316.5830 (2007), pp. 1497–1502.

[106] S Myers et al. “A fine-scale map of recombination rates and hotspots across the human genome.” In: *Science* 310.5746 (2005), pp. 321–324.

[107] The International HapMap Consortium. “A second generation human haplotype map of over 3.1 million SNPs”. In: *Nature* 449.7164 (2007), pp. 851–861.

[108] B Davies et al. “Re-engineering the zinc-fingers of PRDM9 reverses hybrid sterility in mice”. In: *Nature* (2016), doi:10.1038/nature16931.

[109] T L Bailey et al. “The MEME Suite”. In: *Nucleic Acids Research* 43.W1 (2015), W39–W49.

[110] The ENCODE Project Consortium. “An integrated encyclopedia of DNA elements in the human genome”. In: *Nature* 489.7414 (2012), pp. 57–74.

[111] R Beck et al. “LINE-1 elements in structural variation and disease.” In: *Annual review of genomics and human genetics* 12.1 (2011), pp. 187–215.

[112] H Sein, S Varv, and A Kristjuhan. “Distribution and Maintenance of Histone H3 Lysine 36 Trimethylation in Transcribed Locus”. In: *PLoS ONE* 10.3 (2015).

[113] J D Buenrostro et al. “Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position”. In: *Nature Methods* 10.12 (2013), p. 1213.

[114] C Trapnell et al. “Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks.” In: *Nature Protocols* 7.3 (2012), pp. 562–578.

[115] F Sleutels et al. “The male germ cell gene regulator CTCFL is functionally different from CTCF and binds CTCF-like consensus sites in a nucleosome composition-dependent manner”. In: *Epigenetics & Chromatin* 5.1 (2012), p. 8.

[116] A Klug. “The Discovery of Zinc Fingers and Their Applications in Gene Regulation and Genome Manipulation.” In: *Biochemistry* 79 (2010), pp. 213–31.

[117] A R Aricescu, W Lu, and E Y Jones. “A time- and cost-efficient system for high-level protein production in mammalian cells”. In: *Acta Crystallographica Section D* 62.10 (2006), pp. 1243–1250.

[118] E Campeau et al. “A Versatile Viral System for Expression and Depletion of Proteins in Mammalian Cells”. In: *PLoS ONE* 4.8 (2009), e6529.

[119] H Li and R Durbin. “Fast and accurate short read alignment with Burrows-Wheeler transform”. In: *Bioinformatics* 25.14 (2009), pp. 1754–1760.

[120] G Lunter and M Goodson. “Stampy: a statistical algorithm for sensitive and fast mapping of Illumina sequence reads”. In: *Genome Research* (2011).

[121] H Li et al. “The sequence alignment/map format and SAMtools”. In: *Bioinformatics* 25.16 (2009), pp. 2078–2079.

[122] A R Quinlan and I M Hall. “BEDTools: a flexible suite of utilities for comparing genomic features.” In: *Bioinformatics* 26.6 (2010), pp. 841–842.

[123] W C Hines et al. “BORIS (CTCFL) is not expressed in most human breast cell lines and high grade breast carcinomas.” In: *PLoS ONE* 5.3 (2010), e9738.

[124] B T Lahn and D C Page. “A human sex-chromosomal gene family expressed in male germ cells and encoding variably charged proteins.” In: *Human Molecular Genetics* 9.2 (2000), pp. 311–319.

[125] 1000 Genomes Project Consortium. “An integrated map of genetic variation from 1,092 human genomes”. In: *Nature* 491.7422 (2012), pp. 56–65.

[126] T Ashley et al. “Dynamic changes in Rad51 distribution on chromatin during meiosis in male and female vertebrates”. In: *Chromosoma* 104 (1995), pp. 19–28.

[127] F Cole et al. “Mouse tetrad analysis provides insights into recombination mechanisms and hotspot evolutionary dynamics”. In: *Nature Genetics* (2014), pp. 1–11.

[128] C L Baker et al. “Multimer Formation Explains Allelic Suppression of PRDM9 Recombination Hotspots.” In: *PLoS Genetics* 11.9 (2015), e1005512.

[129] H Qiao et al. “Interplay between Synaptonemal Complex, Homologous Recombination, and Centromeres during Mammalian Meiosis”. In: *PLoS Genetics* 8.6 (2012), e1002790.

[130] K A Henderson and S Keeney. “Synaptonemal complex formation: where does it start?” In: *Bioessays* 27.10 (2005), pp. 995–998.

[131] C-T Ong and V G Corces. “CTCF: an architectural protein bridging genome topology and function”. In: *Nature Reviews Genetics* 15.4 (2014), pp. 234–246.

[132] J B S Haldane. *The Causes of Evolution*. Princeton University Press, 1932.