



UNIVERSITY OF
LEICESTER

PRDM9 DIVERSITY, RECOMBINATION LANDSCAPES AND CHILDHOOD LEUKAEMIA



Thesis submitted for the degree of
Doctor of Philosophy
at the University of Leicester

Ihthisham Ali
Department of Genetics and Genome Biology
University of Leicester

May 2020

PRDM9 Diversity, Recombination Landscapes and Childhood Leukaemia

Ihthisham Ali

Abstract

PR/SET domain 9 (PRDM9) protein is a potent selector of meiotic recombination activation sites via DNA sequence recognition and is known to be highly polymorphic in the critical Zinc Finger (ZnF) array. This diversity can influence the recombination landscape and translate into genome-wide differences between major population groups. PRDM9 variation has also been implicated in genomic instability in cancer.

To investigate a potential link between rare PRDM9 alleles and elevated ancestral recombination rates in the Major Histocompatibility Complex II (MHCII) region associated with childhood Acute Lymphoblastic Leukaemia (ALL) in a British cohort, two sub-regions were targeted using sperm-typing methods. This revealed that the DNA3 hotspot is PRDM9 A-regulated, whilst the African-enriched AA hotspot is activated by rare C-type (Ct) alleles containing K-type ZnFs. However, the latter may not be as active as historical population estimates indicate, or unsampled PRDM9 alleles may activate this hotspot more efficiently. Screening for Ct alleles and other K-ZnF containing alleles in the British ALL cohort provided potential links with PRDM9, though not strong support for previous work. Investigation of other candidate genome-wide associated markers indicated a link with FIGNL-1, a protein complex involved in homologous recombination, which was supported by an independent German cohort.

A large Next-Generation Sequencing (NGS) dataset including rarely sampled populations was used for PRDM9 allele discovery, along with a comparison study on the capabilities of NGS platforms to characterise the long

PRDM9 ZnF arrays. The Illumina 100bp paired-end read format was useful for filtering known alleles but de novo assembly was unable to resolve ZnF array structure. Ion Torrent 400bp read data provided only incremental improvement over 200bp reads. Finally, nanopore sequencing showed promising results although improved basecalling and read mapping methods as well as de novo assembly would be required to displace Sanger as the definitive method for ZnF array structure characterisation.

Dedication

"He makes you in the wombs of your mothers in stages, one after another, in three veils of darkness" - Qur'an 39:6

"The capacity to blunder slightly is the real marvel of DNA. Without this special attribute, we would still be anaerobic bacteria and there would be no music." -
Lewis Thomas

"The peach was once a bitter almond; cauliflower is nothing but a cabbage with a college education. Training is everything." - Mark Twain

for my Dad

Acknowledgements

This project was born and completed due to the resilient efforts of Dr Celia A May. Her knowledge and range is immense - quite frankly unfathomable! Yet, what makes her so remarkable is her generosity in sharing her expertise and collaborating in such a way that empowered me to make the right decisions independently. These are all well known facts. So I would like to highlight another facet to this great scientist. Dr May has a strong work ethic that is founded on established principles and guidelines, working continuously to achieve her aims in this way. I learned a lot about how important planning is before execution. This is one of the reasons why the bulk of my labwork went smoothly, apart from the rare hiccups once or twice a year. She has also helped me complete this project despite major upheavals in my life. So for all these reasons and more, I would like Dr May to know that I am forever grateful and indebted to her guidance and assistance.

Rita Neumann comes next. She is the wizard with the hat full of tips and tricks! The strategies that Dr May and I came up with were tested and improved upon with the fine tuning and troubleshooting suggestions that Rita made. I know I can apply these techniques to great effect in my future work. My knowledge was enriched from the hours and hours of discussions in the computer room or the corridor outside G19. Thankyou, Rita. As always, I wish you more brilliant discoveries in the future. I am going to miss you telling me all about them.

I would like to express my sincere appreciation to my collaborators Dr Pamela Thompson and Dr Pille Hallast. Your contribution and assistance allowed me explore a wider range of techniques and theoretical aspects than first anticipated. Hopefully our work together contributes to their respective research areas.

Thanks to my original G19/Leicester crew, especially Maria, Carmen, Enjie, John (Wagstaff), Vicky, Sho and Wu. For the fun times in and away from the lab, for the walks and nights out, sitting in the sun (when it appears) with Starbucks, and the constant buzz that helped the days move along. You guys probably now wish I stayed the reserved person I was in the first 6 months! I would also like to thank Jordan, Orie and all the new staff and students during my 2018/2019 stint in G19.

My sincere thanks to Professor Mark Jobling for your wisdom and your gift to my research, other staff and students in Lab G2 who helped me out with various aspects of data analysis on next-gen sequencing, Dr John Wetton for our work on the MinION developer version and various minor missions here and there, Dr Cas Cramer for giving me the chance to work in GENIE events, Dr Ezio Rosato for giving me the chance to do some teaching, Dr Richard Badge for specific questions in bioinformatics and software tools, Dr Adam Webb for the helpful suggestions and the LDU map tool, Sam, Giordano and their supervisor Dr Sandra Belezza for help with learning to edit code and Dr Yan Huang for giving me the opportunity to present my work and career to visiting fellows and being the most hands-on landlord you can find in Leicester.

A special thankyou for Dr Ed Hollox and Dr Chris Talbot for their one-two punch dressing down of my annual reports. I hope your insight and guidance has worked its way into my thesis.

Thanks to my home group especially Bodey, Yummy, Whosen, Samarey, Naape, Adhuham, Hassan Saeed, Afra, Wishana, Hussain Adam sir, DNA Lab, DCL and all of Forensic Services, the Masroshi group, Panda and Hussain, and many others at home and work who make me feel like I actually do achieve things. Rebel 300! Special thanks to Debs for helping me get through difficult times, coffee and dinner, and learning to leave me alone sometimes.

Last but most importantly, my heartfelt thanks to my tiny little family. My dear sister Ibthi, you are the unsung heroine. Zoie: You are my constant. I love you more than life itself. Your sense of self and diplomacy at such a young age is simply amazing. It is a wondrous experience to watch you grow and develop. Here's to many more eating-out days, afternoon rides over the bridge and standup comedy. Los Angeles better be ready. I would also like to thank Zoie's mummy for making it all work.

The bookends of this entire journey belong to my mom and dad, Hilala Waheed and the late great Ali Yoosuf. This was their dream. I know that no matter where I go from here, I ride the everlasting waves and winds of two great forces of nature.

Table of Contents

<i>Abstract</i>	<i>i</i>
<i>Dedication</i>	<i>iii</i>
<i>Acknowledgements</i>	<i>iv</i>
<i>Table of Contents</i>	<i>vii</i>
<i>Abbreviations</i>	<i>xi</i>
CHAPTER 1: INTRODUCTION	1
1.1 Mutation	2
1.2 Meiosis	3
1.3 Meiotic recombination	6
1.3.1 Analysis of recombination in the human genome	8
1.3.2 Some lessons learned	9
1.4 Hotspot regulation	13
1.5 PRDM9 regulation of hotspots	13
1.5.1 PRDM9-DNA binding	16
1.5.2 Prospects for PRDM9 allele discovery: sequencing technologies	20
1.6 PRDM9 and genomic instability	22
1.7 Childhood ALL	23
1.7.1 Epidemiology	24
1.7.2 Causal factors	25
1.7.3 Potential links to PRDM9 activated hotspot	27
1.8 Overall research aims	28
1.8.1 Research questions and hypotheses	28
1.8.2 Research aims	29
CHAPTER 2: MATERIALS AND METHODS	32
2.1 Materials	32
2.1.1 Chemical reagents	32
2.1.1.1 PCR	32
2.1.1.2 Gel Electrophoresis	33
2.1.1.3 SNP genotyping	33
2.1.1.4 Sanger Sequencing	34

2.1.1.5 Ion Torrent sequencing	35
2.1.1.6 Nanopore sequencing 2015 R7 flow cell developer version	35
2.1.1.7 Nanopore sequencing 2019 R9.4.1 flow cell chemistry	35
2.1.2 Reagent kits	36
2.1.3 Labware	36
2.1.4 Instruments and equipment	37
2.1.5 Online resources	38
2.1.7 Software	39
2.1.7.1 Commercial/Publicly available software	39
2.1.7.2 In-house software	39
2.1.6 Samples	40
2.1.6.1 Human DNA	40
2.1.7.2 Datasets	41
2.2 Methods	42
2.2.1 PCR	42
2.2.2 Sperm crossover assay	44
2.2.3 Gel Electrophoresis	45
2.2.4 SNP genotyping	47
2.2.5 Sanger Sequencing	48
2.2.6 Ion Torrent sequencing	49
2.2.7 Nanopore sequencing	49
2.2.8 Illumina HiSeq data processing	50
CHAPTER 3: CHILDHOOD ALL AND PRDM9 REGULATION OF TWO MHCII HOTSPOTS	52
3.1 Introduction	52
3.1.1 Childhood B-ALL susceptibility	52
3.1.2 DNA3 hotspot	54
3.1.3 An African-enriched hotspot	55
3.1.4 Study aims	56
3.2 Results	57
3.2.1 SNP genotype survey of the DNA3 and AA hotspots	57
3.2.2 Optimisation of ASPs	60
3.2.3 Selection of donors suitable for CO analysis	64

3.2.4 Linkage phasing and sperm CO assay design	66
3.2.5 DNA3 hotspot is activated by PRDM9 A	68
3.2.6 AA hotspot is still active but weak	72
3.3 Discussion	75
3.3.1 PRDM9 A binding at DNA3 hotspot	75
3.3.2 Variation in activity of A/N individuals at DNA3 hotspot	76
3.3.3 Conclusions	77
CHAPTER 4: CHARACTERISATION OF THE AA HOTSPOT	79
4.1 Introduction	79
4.1.1 African-enriched hotspots	79
4.1.2 Crossover:Non-crossover variation and GC-biased gene conversions	80
4.1.3 Study aims	81
4.2 Results	81
4.2.1 Comparison of ZnF array sub-motifs in activating Ct alleles	82
4.2.2 Extended LD survey of the AA hotspot region	83
4.2.3 Hotspot morphology	87
4.3 Discussion	89
CHAPTER 5: CHILDHOOD ALL AND PRDM9 ALLELES RARE TO EUROPEANS	91
5.1 Introduction	91
5.1.1 K-ZnF containing PRDM9 alleles	91
5.1.2 FIGNL1 variants and other GWAS SNPs	93
5.1.3 Study aims	94
5.2 Results	95
5.2.1 Further validation of a PRDM9-SNP Haplotype Network	95
5.2.2 SNP genotyping on the British ALL cohort	101
5.2.3 Testing for association between K-ZnF and B-ALL	104
5.2.4 FIGNL1, CDKN2A and MHCII SNPs	108
5.2.5 Imputation of FIGNL1 SNP genotypes in a German cohort	110
5.2.5.1 Imputation model	111
5.2.5.2 Imputation Results	115
5.3 Discussion	116
5.3.1 Association of D, L20 and Ct alleles with childhood ALL	116

5.3.2 Association of FIGNL1 SNPs with childhood ALL	116
CHAPTER 6: PRDM9 ALLELE DISCOVERY USING SECOND AND THIRD GENERATION SEQUENCING	118
6.1 Introduction	118
6.1.1 Novel PRDM9 alleles	118
6.1.2 Sanger versus NGS platforms	119
6.1.3 Study Aims	120
6.2 Results	120
6.2.1 Validation by Sanger sequencing	120
6.2.2 Illumina dataset remapping	123
6.2.3 Ion Torrent sequencing and mapping	135
6.2.4 Nanopore sequencing and mapping	139
6.2.5 Illumina dataset read depth and variant site analysis	145
6.2.6 De novo assembly of Illumina dataset	151
6.2.8 Overall platform comparison	154
6.3 Discussion	155
CHAPTER 7: DISCUSSION	158
7.1 Future work	159
Appendices	161
Appendix I: PCR design	161
Appendix II: SNP genotyping	161
Appendix III: Crossover activity	162
Appendix IV: PRDM9 ZnF arrays	162
Appendix V: Bioinformatic pipelines/scripts	163
Appendix VI: Sequencing platform comparison	163
Appendix VII: Supplementary figures and table	164
Bibliography	166

Abbreviations

ALL	Acute Lymphoblastic Leukaemia
ASO	Allele-specific Oligonucleotide
ASP	Allele-specific Primer
B-ALL (BCP-ALL)	B-cell Acute Lymphoblastic Leukaemia
bp	Base pairs
CMT1A	Charcot-Marie-Tooth type 1A
CI	Confidence Interval
CO	Crossover
CS	Control Sequence
Ct	C-type (PRDM9 allele)
dHJ	Double Holliday junction
DSB	Double-Strand Break
DSBR	Double-Strand Break Repair
EMSA	Electrophoretic Mobility Shift Assay
GWAS	Genome-Wide Association Studies
HNPP	Hereditary Neuropathy with Pressure Palsies
HR	Homologous Recombination
kb	kilobases
LD	Linkage Disequilibrium
LDU	Linkage Disequilibrium Units
LINE	Long Interspersed Nuclear Element
MHCII	Major Histocompatibility Complex II
Mb	Megabases
NAHR	Non-Allelic Homologous Recombination
NCO	Non-Crossover
ONT	Oxford Nanopore Technologies
OR	Odds Ratio

PacBio	Pacific Biosciences
PAR	Pseudoautosomal region
PCR	Polymerase Chain Reaction
PRDM9	PR/SET domain-containing 9
SC	Synaptonemal Complex
SDSA	Synthesis-Dependent Strand Annealing
SMRT	Single-Molecule Real-Time Sequencing
SNP	Single Nucleotide Polymorphism
T-ALL (TCP-ALL)	T-cell Acute Lymphoblastic Leukaemia
ZnF	Zinc Finger

CHAPTER 1: INTRODUCTION

This thesis is based on the two major themes of 1) examining the effects of PR/SET domain 9 (PRDM9) protein allelic diversity with respect to its ability to select initiation sites for meiotic recombination and by escalation its ability to affect change in chromosomal recombination landscapes of different populations, and 2) investigating the potential role of PRDM9 allelic variants in the development of childhood Acute Lymphoblastic Leukaemia (ALL).

Recombination is a unique and essential feature of meiosis that produces gamete cells for sexual reproduction (Kleckner, 1996; Zickler and Kleckner, 1998; Grelon, 2016). Rates and distribution of meiotic recombination are non-random along the length of chromosomes (Kong et al., 2002, Myers et al., 2005) and concentrated at hotspots (Lichten and Goldman, 1995; McVean et al., 2004; Jeffreys, Kauppi and Neumann, 2001) surrounded by haplotype blocks. Recombination is under control by both cis- (Jeffreys and Neumann, 2005; Myers et al., 2005; Myers et al., 2008) and trans-regulatory factors. In mice and humans, the multi-allelic PRDM9 protein is a potent selector of initiation sites for recombination and therefore a major player in the trans-regulation of recombination hotspots (Baudat et al., 2010). PRDM9-mediated regulation of hotspots can be examined directly in the human male germline using high-resolution batch sperm crossover (CO) detection and resolution mapping by leveraging SNP heterozygosity (Jeffreys, Kauppi and Neumann, 2001). The Zinc Finger (ZnF) array of PRDM9 is highly polymorphic (Berg et al., 2010) and this translates to a wide diversity of PRDM9 alleles amongst populations (Berg et al., 2011). Previous work has revealed how PRDM9-mediated hotspot activation differentiates recombination landscapes in European and African populations (Berg et al., 2011; Hinch et al., 2011). Characterising PRDM9 ZnF arrays is normally achieved via Sanger sequencing yet newer sequencing platforms and associated datasets potentially offer opportunities for characterising novel

PRDM9 alleles and provide insight into the shifts and reshaping of recombination landscapes in hitherto more diverse individuals and populations.

Non-allelic Homologous Recombination (NAHR) plays an underlying role in the development of complex genetic disease (Turner et al., 2008; Zhang et al., 2009; Piazza and Heyer, 2019). PRDM9 A-activated hotspots have been shown to influence meiotic NAHR that result in two related complex genetic disorders, namely Charcot-Marie Tooth Type 1A (CMT1A) and Hereditary Neuropathy with Pressure Palsies (HNPP) (Berg et al. 2010). Hussin et al. (2013) identified an excess of PRDM9 alleles rare to Europeans in a French-Canadian childhood ALL cohort and Thompson et al. (2014) reported increased recombination rates at the DNA3 hotspot in the Major Histocompatibility Complex II (MHCII), one of the disease susceptibility loci for B-cell precursor Acute Lymphoblastic Leukaemia (B-ALL), a major subtype of ALL. The DNA3 hotspot was previously characterised using European males who predominantly carry PRDM9 A alleles (and so has been presumed to be PRDM9 A-regulated but not formally tested). However, revisiting the MHCII region with a more diverse panel of PRDM9 allele carriers may help to paint a clearer picture of the local recombination landscape in the MHCII in different populations.

The following sections of this Chapter further introduces these concepts whilst the last section presents the research questions, hypotheses and aims of this thesis.

1.1 Mutation

Ultimately all variation in the genome is a result of mutation, caused by error prone DNA replication (Tsai, 2014), abnormal DNA replication proteins and enzymes (Johnson et al., 2000), impaired DNA damage response, DNA damage caused by environmental mutagens such as ionising radiation, ultraviolet light, chemicals, endogenous reactive oxygen species, etc. (Liu, Yip

and Zhou, 2012; Price and D'Andrea, 2013). These changes in the genome have wide implications on biochemical processes, phenotype and quality of life.

Genetic diversity can be generated in several ways. Opportunistic forces such as the competition of millions of sperm cells to fertilise a single egg cell is clearly a means whereby a genetically unique individual is formed. However, this competition may not be wholly opportunistic as sperm characteristics in terms of morphology, acrosome and motility (Liu and Baker, 1992), and oocyte quality (e.g. zona pellucida mutations as reported by Zhou et al., 2019) may influence the success of fertilisation. Presumably, choice of cell at ovulation is also largely random but this process can also be affected by other genetic factors (Dunson et al., 2001; Loutradis et al., 2012).

1.2 Meiosis

Meiosis is the specialised mode of cell division that forms the basis of sexual reproduction in eukaryotes. It reduces, or more accurately halves, the ploidy from the parent cell ($2n$, where n stands for the number of copies of any given chromosome or locus or gene) to the daughter cell (n) so that the full set of chromosomes found in the parent cell, the chromosome complement, is restored via the process of fertilisation (Kleckner, 1996). In mammals, events in meiosis are very similar but with differences in timing, regulation, proteins involved and structures formed (de Massy, 2013). In all cases, diploid precursor cells of both sexes undergo meiosis and produces one haploid mature oocyte in females or four haploid spermatozoa in males.

Healthy functioning cells spend most of their life in Interphase where chromosomes are extended and open, existing as DNA-histone based chromatin such that the genes can be accessed for transcription to generate required proteins and non-coding RNAs for the natural functions of each particular cell. The Interphase itself is divided into three phases called G1 (Gap1), S (Synthesis)

and G2 (Gap2) (Zickler and Kleckner, 1998). It is during the middle S-phase that semi-conservative replication of the genome occurs (Hanawalt, 2004) prior to either mitotic or meiotic cell division. Each chromosome is copied, leading to the doubling of the chromosome complement ($4n$) (Kleckner, 1996). Two strands of the DNA double helix separate and become templates for producing two new DNA double helices. Also, the main microtubule organising centre, the centrosome, forms outside the nucleus and replicates itself during this time.

As with mitosis, meiosis can be described as undergoing four stages, namely prophase, metaphase, anaphase and telophase. This is followed by cytokinesis and the formation of new daughter cells. Unlike mitosis however, meiosis includes two rounds of chromosome segregation (Kleckner, 1996). Meiosis I sees reductional chromosome segregation where, following the pairing of homologous chromosomes, the duplicated non-sister chromatids separate from each other. Since the orientation of each of these paired homologues on the metaphase plate is independent of any other homologue pair, the maternal and paternal copies of each chromosome are randomly distributed to the resulting daughter cells, so-called independent assortment. This is a major means by which genetic variation is introduced in each generation. By contrast, in Meiosis II, equational chromosome segregation occurs where sister chromatids separate akin to mitosis.

Another profoundly important feature of meiosis is that in Prophase I of Meiosis I (Kleckner, 1996; Zickler and Kleckner, 1998), non-sister chromatids are involved in crossing over and the generation of novel recombinant chromosomes; this exchange of genetic material between homologues effectively results in new chromosomes consisting of combinations of grandpaternal and grandmaternal DNA being passed into the resulting haploid cells. This Homologous Recombination (HR) therefore has wide implications due to its effect of shuffling haplotypes, introducing further variation into the genome.

In Prophase I, transient inter-homologue interactions occur with non-sister chromatids during this time (Kleckner, 1996). Non-sister chromatids engage in homology searches which ultimately lead to synapsis, the physical tethering of non-sister chromatids and crossing over. These events are made possible with the induction of programmed Double Strand Breaks (DSBs) to allow strand invasion, HR and the formation of a tripartite proteinaceous scaffold between the chromosomes known as the Synaptonemal Complex (SC). The 'goal' of prophase I is to pair homologous chromosomes, first through SC-mediated protein:protein interactions and second via HR-mediated DNA:DNA interactions manifest as so-called chiasmata. Prophase I ends with the nuclear membrane starting to disintegrate and the centrosomes begin to pull away in linear fashion from the mid-plate of the cell towards opposite poles, thereby separating the duplicated non-sister chromatid pairs.

In Metaphase I, crossing over is complete and centrosomes are in the poles perpendicular to the plane of alignment of the chromosomes. The spindle fibres originating from the centrosomes attach to the kinetochores, protein structures assembled on the centromeres of each duplicated non-sister chromatid. Independent orientation of the SC-based chromatid pairs effectively causes the random assortment of each homologue into the new daughter cells, an integral mode of creating new combinations of alleles in each offspring. During Metaphase I, the chiasmata are seen during metaphase moving from the origin of HR towards the telomeres or ends of the chromosomes. Anaphase I, differs from Anaphase of mitosis, in that sister chromatids move to the same pole whilst the corresponding non-sister chromatids are pulled to the opposite pole. In Telophase I, the chromatids have almost completely transited to the poles. Microtubules start to disassemble and cytokinesis initiates. New nuclear envelopes start forming and this phase ends with the complete splitting of the nascent cells. In Prophase II, centrosomes replicate again and pull apart towards opposite poles this time along the plane of the aligned chromosomes. Nuclear envelopes in both new cells disintegrate. In Metaphase II, centrosomes complete

their migration to the poles and the microtubules reach back to the centromeres on the chromosomes and the chromatids line up. In Anaphase II, the kinetochore microtubules pull apart each sister chromatid away from each other towards opposite poles.

Thus two events in meiosis, namely HR and the independent assortment of chromosomes, ensure each gamete is unique.

1.3 Meiotic recombination

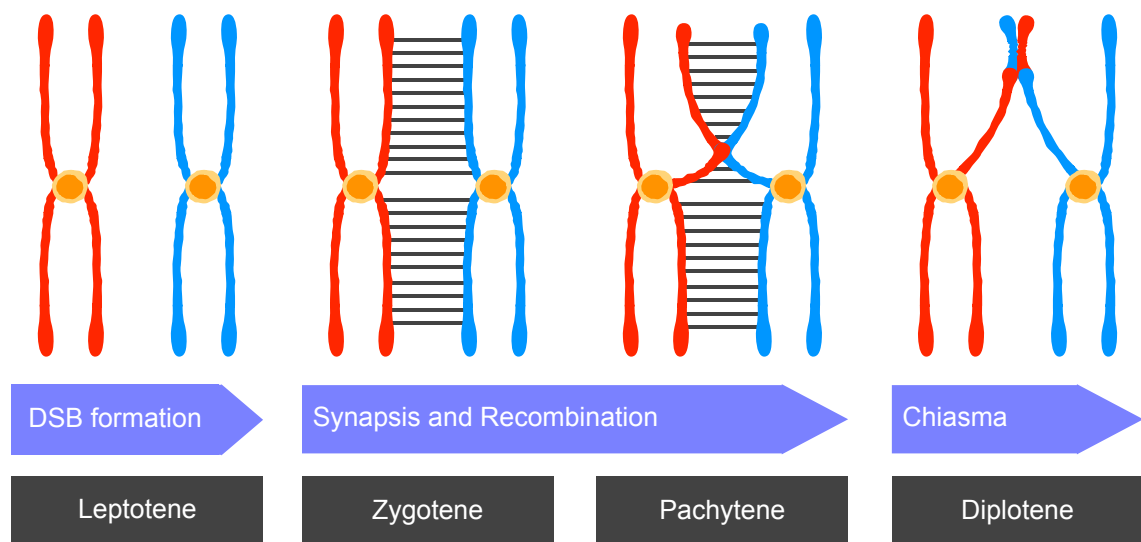


Fig. 1 Prophase I substages in meiosis. Prophase I is divided into Leptotene, Zygotene, Pachytene and Diplotene. Sister chromatids formed from DNA replication come into contact with their homologues. During Leptotene, chromatin condenses further and recombination is initiated with the formation of programmed DSBs throughout Leptotene. These DSBs start disappearing towards the end of Leptotene and during the first part of Zygotene. In Zygotene, homologous chromosomes become paired between non-sister chromatids and synapses are formed. The Pachytene substage is where crossovers between non-sister chromatids occur. During Diplotene, the homologous chromosomes start to separate and chiasmata are observed moving along towards the ends of the chromosomes as the nuclear membrane and nucleoli disintegrate.

HR occurs during Prophase I which itself can be further subdivided into five stages: Leptotene, Zygotene, Pachytene, Diplotene and Diakinesis (Fig. 1). During Leptotene, DSBs are induced in excess across the genome by Sporulation-specific 11 (Spo11), a topoisomerase-like enzyme first identified in *S. cerevisiae* but shown to be highly conserved across species (Keeney, Giroux

and Kleckner, 1997). When comparing organisms, ~160 DSBs are introduced in wild-type *S. cerevisiae*, 230–400 DSBs in male and 250–370 DSBs in female mice (de Massy, 2013) and ~600 DSBs in humans (Grelon, 2016). These DSBs are subsequently repaired by HR leading to two distinct classes of recombinant molecules: the classical COs mentioned previously and gene conversions without exchange of flanking regions or so-called noncrossovers (NCOs) (Baudat and de Massy, 2007). COs involve a reciprocal exchange of flanking regions of DNA between homologous chromosomes whereas NCOs are non-reciprocal transfers of DNA segments from one chromosome to the other as detailed in Fig. 2. It has been estimated from mice and human studies that only 10% and 4% of DSBs respectively, resolve into COs and that the rest result in NCOs (Baudat and de Massy, 2007).

HR has two main roles and consequences for meiosis. Firstly, it has the function of physically assisting in the correct segregation of chromosomes into haploid cells by homology searching between non-sister chromatids, synapsis and formation of the SC and maintenance of chiasmata towards the end of Prophase 1. Secondly, recombination is a unique way of randomly introducing genetic diversity into these haploid cells. Since recombination is crucial for successful meiosis, DSBs are induced in excess across chromosomes to aid in homology search and ensure faithful synapsis. Most of these DSBs are repaired as NCOs but one or two COs per chromosome are required for correct chromosome disjunction. Non-disjunction can result in aneuploidy where gamete cells may form with an abnormal number of chromosomes. For example, a gamete containing three instead of the usual two copies of chromosome 21 can still be fertilised but lead to a child born with Down Syndrome (Antonarakis et al., 1992).

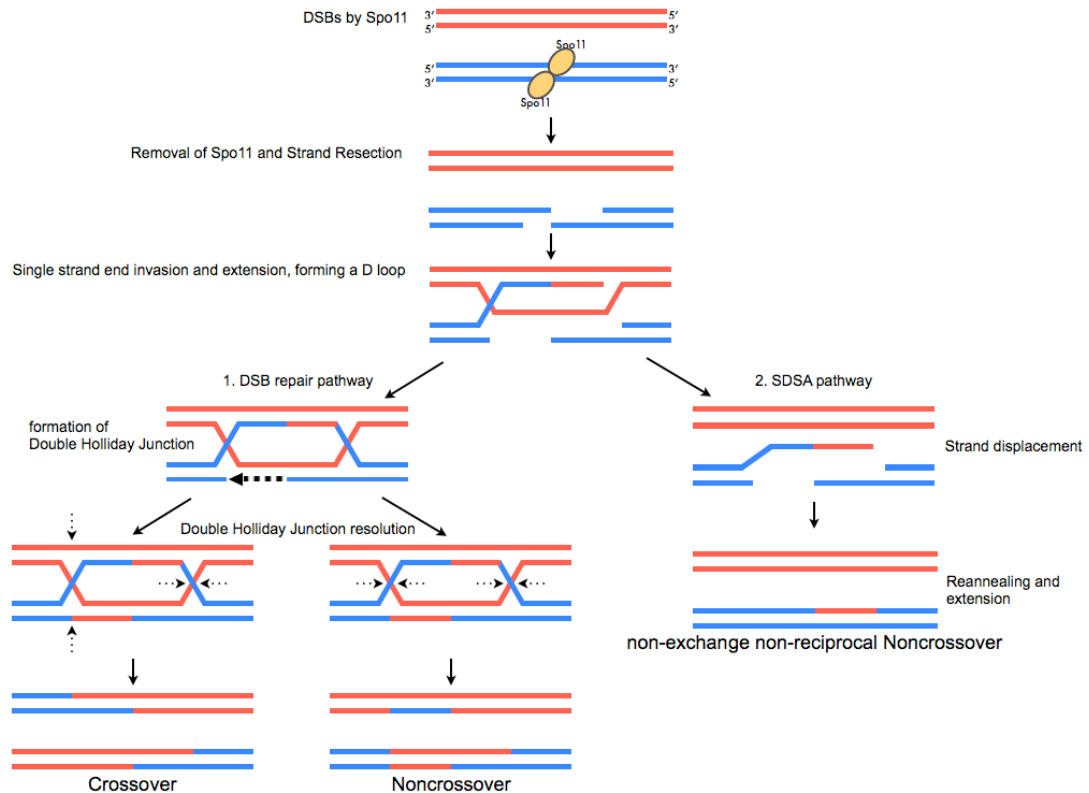


Fig. 2 Simplified diagram showing meiotic recombination between non-sister chromatids in mammals (two non-sister chromatids are distinguished by red and blue colours, their respective sister chromatids are not shown). During leptotene stage of prophase I, DSBs are induced in excess across the genome by Spo11. Spo11 is then endonucleotically removed followed by strand resection with Exo1 exonuclease activity leaving 3' overhangs. One of the single strand ends invades a non-sister chromatid and uses it as a template to extend its own DNA and this results in a D-loop formation. From this point on, recombination can go down two major routes: Double Strand Break Repair (DSBR) pathway or Synthesis-Dependent Strand Annealing (SDSA) pathway. In the DSBR pathway, the second strand is captured by the D-loop followed by DNA synthesis and ligation. The resulting double Holliday junction (dHJ) can in theory be resolved in two possible ways, one leading to a CO event, where flanking regions of DNA have been exchanged between non-sister chromatids, and the other leading to a NCO, with unequal segments of DNA being exchanged between chromatids. In the SDSA pathway, the invading strand is displaced from the D-loop and reanneals back to the sister chromatid following by DNA synthesis and ligation. This results in a non-reciprocal gene conversion without exchange of flanking regions or NCO, where a segment of DNA has been transferred back into the chromatid where the DSB originated. Note that the eventual products of recombination are determined by mismatch repair of the heteroduplex DNA shown here whereby either of the strands may be used as the template for mismatch repair.

1.3.1 Analysis of recombination in the human genome

Our most comprehensive understanding of recombination, its regulation and evolution have been obtained from model organisms such as *S. cerevisiae* and mice (Keeney, Giroux and Kleckner, 1997; Baudat and de Massy, 2007; Paigen and Petkov, 2010; de Massy, 2013). Due to the technical, ethical and availability restraints of directly studying events within gamete cells, meiotic

recombination in humans had been comparatively less understood. However, a variety of direct and indirect approaches have advanced our understanding over the last two decades. A powerful indirect method popular particularly following the advent of genome-wide SNP typing and international efforts like the HapMap project (May, Slingsby and Jeffreys, 2008; Egel and Lankenau, 2008), is to look at population diversity data and infer via levels of Linkage Disequilibrium (LD), or more specifically from the lack thereof, the probable sites of historical recombination (Myers et al., 2008). Of course, other factors such as recurrent mutation, population admixture, natural selection, population bottlenecks and genetic drift can also impact on LD levels (May, Slingsby and Jeffreys, 2007). Initially, direct approaches relied on identifying de novo recombinants from pedigrees but the small size of families and low frequency of CO events for a given interval of the genome imposed limits on this approach, with perhaps the notable exception being the detailed studies of the Icelandic nation as a result of the DeCODE project (Kong et al. 2002, Kong et al. 2004, Kong et al. 2010). Alternative direct approaches include batch screening of thousands of sperm DNA molecules (Kauppi, May and Jeffreys, 2009) which overcomes the issue of modest sample sizes and was employed in this work. An even more recent development has been the creation of whole genome recombination initiation maps, which exploit DNA-protein interactions to isolate HR intermediates from testis samples followed by Next Generation Sequencing (NGS) to reveal the distribution of DSBs (Pratto et al. 2014). These DSB maps clearly reveal sites of both COs and NCOs in the human genome.

1.3.2 Some lessons learned

Linkage mapping of 146 two-generation Icelandic families (equaling 1257 meioses typed for 5136 microsatellites) revealed the sex-averaged genome-wide recombination frequency in humans to be 1.1cM/Mb (Kong et al., 2002), meaning that within any 1 kilobase (kb) interval there is a probability that 1 in every 90,000 gametes will experience a CO. In fact however, linkage maps of

females are longer than those of males, indicating that female meiosis supports $\sim 1.6\times$ more recombination events than male meiosis. In addition to these sex differences, linkage analysis has also revealed that recombination is not uniformly distributed across chromosomes, that it is suppressed around centromeres and elevated towards the telomeres, the latter more so in male than in female meiosis (Fig. 3).

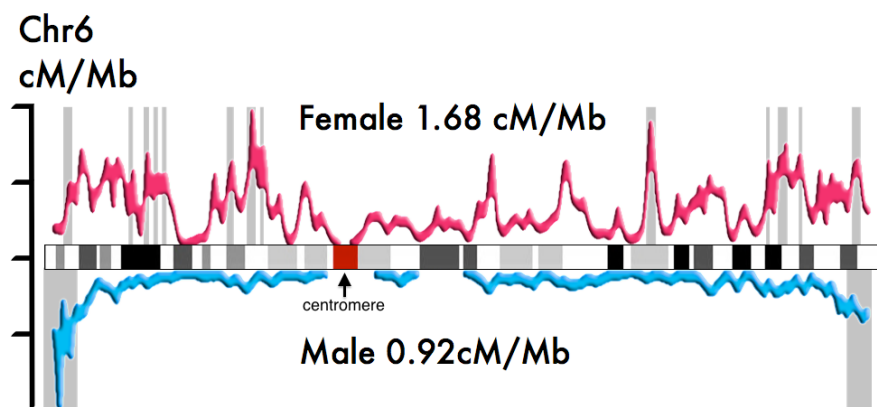


Fig. 3 Recombination profile of Chromosome 6 in humans. Female recombination profile is shown in pink and male recombination profile is shown in blue. Observed peaks of recombination activity are highlighted in grey. Genome average of recombination in females and males are also indicated. Adapted from Chowdhury et al. (2009).

At the megabase level, what is most striking is that there are proximal regions containing peaks of recombination activity, which is supported by genome-wide LD data from population analyses (McVean et al. 2004, Myers et al. 2005). These latter studies have indicated that $\sim 80\%$ of all recombination events occur in just 20% of the genome, indeed there are in excess of 30,000 sites of historical recombination dispersed throughout the human genome. In general, these LD hotspots described using the CEU HapMap dataset were found to be within $\sim 50\text{kb}$ of genes, but outside of the transcribed domain (Myers et al. 2005). Genome-wide, the sheer number of hotspots in humans is relatable to the number of genes, around 32,000 (International Human Genome Sequencing Consortium, 2004), but these hotspots are found in the promoter and intergenic regions and spaced 50-100kb apart (deMassy, 2013).

In males, the two sex chromosomes show very limited homology as they diverged from a pair of ordinary autosomes some 300 million years ago (Ohno, 1967). Nonetheless, there is still a requirement for crossing over to ensure their correct segregation. Therefore in the male germline, recombination occurs at the Pseudo-Autosomal Regions (PARs) at the tips of each end of the X and Y sex chromosomes, the two largest regions that still exhibit complete homology but which comprise <5% of these chromosomes. The larger PAR1 at the ends of the short arms is 2.7Mb in size; family data indicated that there is a single obligatory PAR1-mediated CO in every male meiosis (Rouyer et al. 1986) and diminished recombination has been linked to 24, XY aneuploid sperm (single copy of each autosome, both X and Y chromosome) (Hassold et al., 1990; Shi et al., 2001). Recombination in the shorter PAR2 (~330kb) does occur but does not appear to be obligatory (Freije et al. 1992). Indeed, pairing of the X and Y seems to be initiated at the short arms (Chandley et al., 1984; Chandley et al., 1987; Speed and Chandley, 1990; Armstrong, Kirkham and Hulten, 1994) but there are transient interactions along the entire length of the X and Y chromosomes such that it has been speculated that men who are not so efficient in COs in PAR1 may use COs in PAR2 to compensate (Sarbjana et al. 2012).

Batch sperm analyses have provided most information at the fine scale, i.e. kb to sub-kilobase level. These studies have shown that CO events tend to cluster into very narrow regions of DNA, called recombination hotspots, which are 1-2kb in width, flanked by recombinationally-inert regions. For example, at the DNA3 hotspot in the MHCII region, sperm typing of multiple men by exploiting informative (heterozygous) Single Nucleotide Polymorphisms (SNPs) was used to demonstrate the rapid rise of recombination activity across a 1.2 kb interval (Jeffreys, Kauppi and Neumann, 2001). The distribution of COs is compatible with a normal distribution with activity most intense at the centre and decreasing symmetrically down on either side. This supports the idea that the Spo11-induced DSBs originate at the centre and recombination intermediates migrate outwards before being resolved.

Recombination intensity can vary between hotspots (Table 1), from 0.4cM/Mb in the DNA1 hotspot in the MHCII region of chromosome 6 (Jeffreys, Kauppi and Neumann, 2001) to over 1000cM/Mb at the F hotspot on chromosome 12 (Odenthal-Hesse et al., 2014). Thus, the activity of some recombination hotspots are below the genome average, which is why hotspots are defined by the relative inertness of their respective flanking regions. Based on 132 recombinants amongst 12.3 million meioses, the background recombination intensity in the regions of the genome outside of hotspots averages to 0.37 cM/Mb (0.27, 0.47) (Tiemann-Boege et al., 2006).

Table 1. List of recombination hotspots and their recombination intensities

Hotspot	Intensity, cM/Mb	Location	Reference
DNA1	0.4	MHCII/Chr6	Jeffreys, Kauppi and Neumann, 2001
DNA2	8	MHCII/Chr6	
DNA3	100	MHCII/Chr6	
NID1	38	intron, Chr1	Jeffreys and Neumann, 2005
SHOX	300	Chr Xp/Yp PAR1	May et al., 2002
β-Globin	200	intron, Chr11	Holloway, Lawson and Jeffreys, 2006
SPRY3	685	intron, Chr21	Shriparna Sarbajna (PhD thesis)
F	1100	intron, Chr12	Odenthal-Hesse et al., 2014
K	260	intergenic, Chr8	

Batch sperm CO assays have also shown that recombination activity at specific hotspots can vary significantly between individuals (Jeffreys and Neumann, 2002). Also, since DSBs made by Spo11 are required for recombination initiation, it begs the question as to how Spo11 selects the sites for hotspot activation. These are some of the observations that point to the existence of mechanisms of control for recombination activity.

1.4 Hotspot regulation

Both cis-acting elements and trans-acting regulators have been shown to influence hotspot activity. Where recombination activity at a specific hotspot varies between individuals, in many cases this is due to variation in primary DNA sequence such as the existence of SNP alleles and private mutations unique to an individual. An example of such a cis-acting element was found at the DNA2 hotspot in the MHCII (Jeffreys and Neumann, 2002), where men who were homozygous for the G allele of the central SNP rs416622 A/G (or FG11 A/G) had 2-20 fold lower recombination frequencies compared to those who were homozygous for A or A/G heterozygous at this SNP. Similarly, at the NID1 hotspot in chromosome 1, a man who was homozygous for the C allele of the SNP M-57.8 C/T located 70bp from the hotspot centre had a 3-5 fold higher recombination frequency compared to those who were T/T homozygous and C/T heterozygous for the SNP (Jeffreys and Neumann, 2005).

1.5 PRDM9 regulation of hotspots

In recent years, the system of hotspot regulation that has been most intensely studied is the selection of recombination initiation sites. One protein in particular has become known as a major specifier of hotspot initiation: the Positive Regulating Domain containing 9 protein or PRDM9 (Baudat et al., 2010). PRDM9 belongs to a family of transcription factors that methylate histones or recruit methylation proteins to sites where DSBs subsequently localise. PRDM9 is a meiosis-specific histone H3 methyl transferase which has a KRAB domain, a PR/SET [Su(var)3-9, Enhancer-of-zeste and Trithorax] domain and an array of Cysteine-2 Histidine-2 (C2H2) zinc fingers (ZnFs) (Wolfe, Nekludova and Pabo, 2000; Emerson and Thomas, 2009; Wolf, Greenberg and Macfarlan, 2015) arranged in tandem on the C-terminal which is coded by a 84 base pair (bp) repeat minisatellite (Hayashi, Yoshida and Matsui, 2005) (Fig. 4; Baudat, Imai and de Massy, 2013). PRDM9 is capable of binding to DNA via the

ZnF array and use its SET domain H3K4me3 methyl transferase capability to trimethylate Lysine 4 of H3 histones. This is presumed to initiate chromatin remodelling events that subsequently allow recombination machinery, including Spo11, to access the DNA.

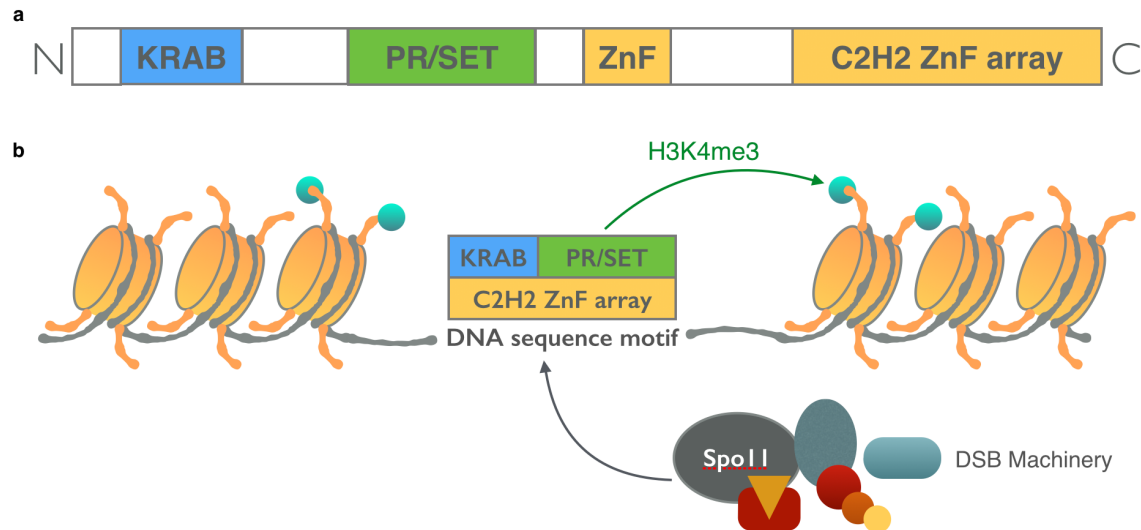


Fig. 4 PRDM9 protein activation of recombination hotspots (Adapted from Baudat, Imai and de Massy, 2013) **a** Major domains of PRDM9 protein including a KRAB protein–protein binding domain (blue), a PR/SET domain (green) which trimethylates Lysine 4 of Histone H3 (H3K4) by transferring active methyl groups bound to Sulphur in S-Adenosyl Methionine (SAM) to Lysine 4 and a polymorphic ZnF array (yellow) on the C-terminal. PRDM9 allelic variants are defined by the variability of the ZnF array. More than 45 different alleles have been characterised with ZnF arrays containing 8-18 84bp-long ZnFs. More than 26 different types of ZnFs have been identified in these arrays. **b** PRDM9 alleles are predicted to bind to specific sequence motifs in the genome followed by very localised methylation that is presumed to remodel the chromatin which then allows recombination machinery including Double-Strand-Break (DSB) inducing Sporulation-specific topoisomerase 11 (Spo11) to access the DNA.

Studies on PRDM9 ZnF arrays found in European and African populations have shown they are highly polymorphic with at least 20 different ZnF repeat types differing over 31 of the 84bp positions, and with arrays ranging from 8-18 ZnFs in length (Berg et al., 2010). Furthermore, studies have shown that hotspot activity is very sensitive to these DNA variations, with some individuals using a substantially different set of hotspots (Baudat et al., 2010; Berg et al., 2010; Berg et al., 2011, Hinch et al., 2011). Searching for sequence motifs enriched in recombination hotspots to which entire PRDM9 ZnF arrays are predicted to bind yielded a 13bp degenerate sequence motif, CCNCCNTNNCCNC, for the PRDM9 allele A. The encoded ZnF array of allele A

is predicted to match all 8 of the non-degenerate positions of the motif, a so-called 8/8 match (Myers et al., 2008). PRDM9 A is the most commonly observed allele accounting for 84% of North European alleles and ~50% of alleles carried by Sub-Saharan Africans (Berg et al., 2010).

In recombination hotspots where this motif has been located, sperm typing has shown that men carrying A alleles have significantly higher recombination frequencies compared to those carrying non-A alleles (Fig. 5; Berg et al., 2010). However, only 41% of the human hotspots carry this 13bp sequence motif (Myers et al., 2008) and hotspots that do not contain this motif can also be activated by PRDM9 A (Berg et al., 2010). Additionally, PRDM9 alleles that are not predicted to bind the motif still influence hotspot activity in motif-containing hotspots. Clearly, the mechanism for hotspot regulation is much more complex than simple binding interactions between PRDM9 and DNA.

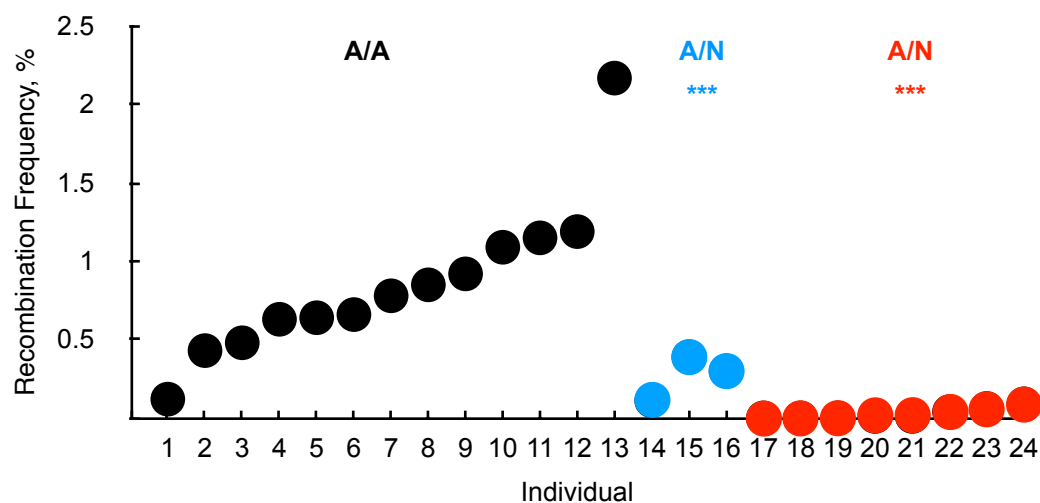


Fig. 5 Variation in sperm CO activity between men at hot spot F containing a central 13bp motif adapted using supplementary data from Berg et al. (2010). Men carrying two PRDM9 A alleles (A/A) are shown in black, men carrying one A allele (A/N) are shown in blue, men carrying two non-A alleles (N/N) are shown in red with men in each group in ascending order of recombination frequencies. Mann-Whitney test results for the significance of differences between the A/A group and the A/N or N/N groups are given at the top right (ns, not significant, $P > 0.05$; * $P < 0.05$; ** $P < 0.01$; *** $P < 0.001$) in their group colours.

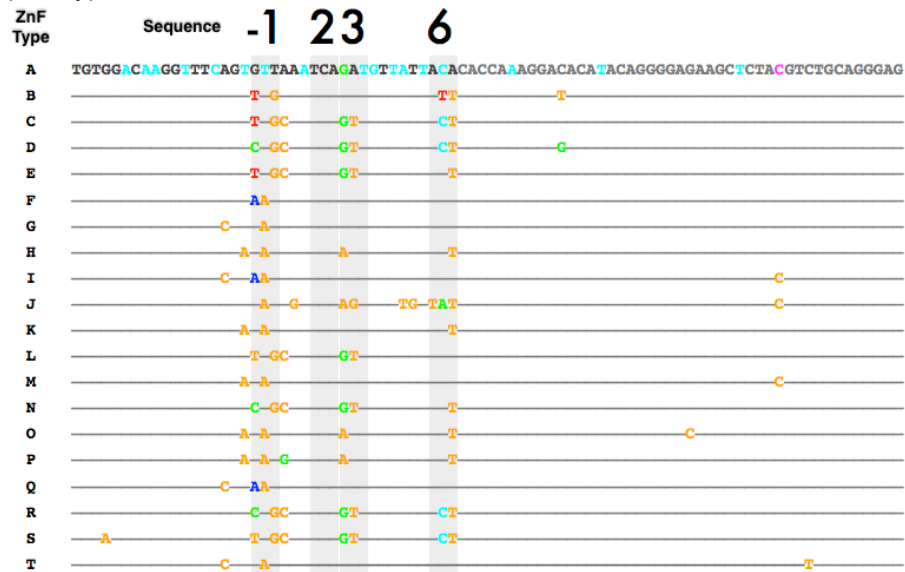
Non-PRDM9 A alleles have the capability to bind to other sequence motifs to activate hotspots. A survey of 156 North Europeans and 74 African

(predominantly Southern Eastern) individuals showed that the latter have a greater diversity of PRDM9 allelic variants with 19 different ZnF arrays (alleles A, B, C, L4-L7, L11-L19 and L21-L23) compared with the Europeans that were found to carry alleles just 13, (A, B, C, D, E, L1-L3, L8-L10, L20 and L24) (Berg et al., 2010). The C-type (Ct) alleles, which consist of alleles C, L4, L6, L8 and L14-L19, have a frequency of 34.5% in Africans (as opposed to just 1.3% in Europeans) and they do not activate hotspots associated with the 13bp motif (Berg et al., 2011). Ct alleles only have a 5/8 match to the non-degenerate positions of the PRDM9 A-associated 13bp motif. These Ct alleles were instead shown to bind to a 16bp sequence motif CCNCNNTNNNCNTNNC (Kong et al., 2010). In Africans, Ct variants such as L6 allele have been shown to activate hotspots to higher recombination frequencies, sometimes 4% per sperm, compared to the PRDM9 A-regulated hotspots and men homozygous for the A allele have been shown to use different sets of hotspots from men homozygous for Ct variants (Berg et al., 2011). PRDM9 A and Ct variants determine 85% of hotspot locations in both populations (Berg et al., 2010). Additionally, the morphology of the hotspots that are regulated by A and Ct variants are very similar, indicating that downstream processing of DSBs is equivalent. However, the overall result is that initiation of different hotspots by A and Ct variants changes the recombination landscapes in these two populations.

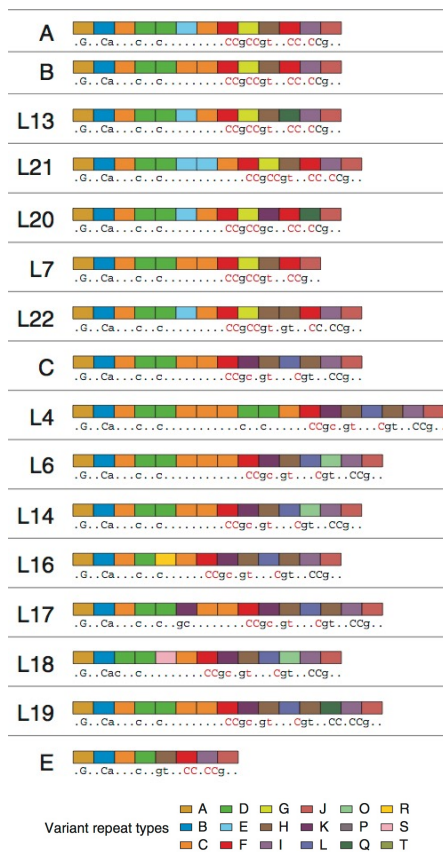
1.5.1 PRDM9-DNA binding

Each C₂H₂ ZnF has the ability to recognise and bind to three adjacent DNA bases (Klug, 2010). Polymorphisms in the PRDM9 ZnF array are largely concentrated in these three DNA-contact residues (Billings et al., 2013) as indicated in Fig. 6. In addition to these coding sequence changes, minisatellite shuffling has added, subtracted and rearranged the order of the ZnFs (Jeffreys et al., 2013). These mutations affect the binding ability of PRDM9 to DNA sequences and so have an impact upon the recombination activity in individuals carrying different PRDM9 alleles.

a ZnF repeat types



b ZnF array types



c DNA-contact residues of the ZnFs

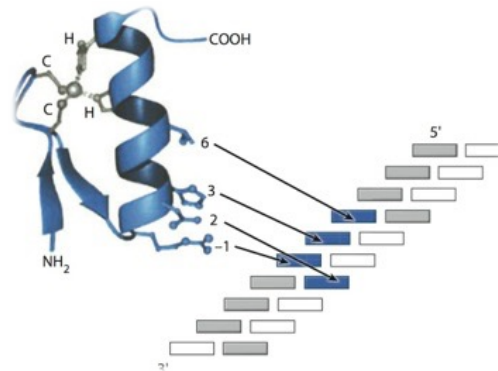


Fig. 6 PRDM9 ZnF array **a** 20 ZnF repeat types (reproduced from Berg et al., 2010). Full sequence of type A is shown. Variations between repeat types are indicated by coloured letters and four DNA contact residues in each are indicated by grey shaded bins with helical positions indicated above (see **c** of this figure). **b** 16 ZnF array types (reproduced from Berg et al., 2010). ZnF repeat types that form the arrays are indicated by colours with colour key shown below. Letters correspond to ZnF repeat types in **a** of this figure. **c** DNA-contact residues for the ZnF (reproduced from Klug, 2010). Contacts are made by amino acid residues with helical positions -1, 3, and 6 are on the coding strand and from position 2 to the noncoding strand.

Patel et al. (2016) used in vitro DNA-binding assays with oligonucleotides specifically representing the 13bp PRDM9 A-associated hotspot sequence (Myers et al., 2008) to bind a partial ZnF array (ABCDDEC**FGHF**IJ, obtained ZnFs in bold) of PRDM9. Actual hotspot sequence derived from the THE1B transposon (Myers et al., 2008) was found to bind with five times higher affinity than the 13bp consensus motif (Patel et al., 2016 Fig.1D). Whilst the I-ZnF was not visible, the FGHF mini-motif was bound exclusively to the major groove of the oligonucleotide sequence with each α -helix of the ZnFs hydrogen-binding with up to four adjacent purine (A/G) bases of the complementary strand. His and Arg residues of the F- and G-ZnFs were hydrogen-bonded with C-G base pairs whilst Asn residue of the H-ZnF hydrogen-bonded with A-T base pairs.

The non-degenerate positions of the 13bp PRDM9 A-associated sequence motif were distinguished using a combination of their ability for hydrogen-bonding and steric complimentary with G bases corresponding to Arg or His residues and A bases corresponding to Asn residues. The degenerate positions of the 13bp PRDM9 A-associated motif also made hydrogen-bonds with ZnFs conferring a level of versatility of PRDM9 A for variations in the hotspot sequence. PRDM9 L13 was also shown to have a high affinity to the A-associated sequence. PRDM9 L9/L24 showed a lower affinity and PRDM9 L20 had a different sequence specification. In agreement with previous predications by Hinch et al. (2011), PRDM9 C bound with a higher affinity to its associated 16bp motif compared to PRDM9 A (Patel et al., 2017). By measuring the difference between binding affinities between FGHF, FGHFI and FGHFIJ, the final J-ZnF was found to be non-interacting. However, dropping the I-ZnF caused a significant loss of binding affinity. Interestingly, the I-ZnF was shown to interact with a CpG site found in the THE1B sequence and a non-specific negative control but more importantly not in the predicted 13bp motif (Myers et al., 2008). CpG sites are an identifier of C-base methylation in eukaryotic cells. However, there was no effect of replacing this C-base with a methylated C-base as binding affinity remained the same.

When FKHLHI of the ZnF array in PRDM9 C allele was assessed for affinity with the 16bp PRDM9 C-associated motif compared to the THE1B motif (Patel et al., 2017), it was found that the affinity of the C allele to the C-specific motif was 10 times higher than that of C allele to the A-specific motif further confirming the proposal by Pratto et al., (2014) that PRDM9 C binding to its 16bp motif is stronger than PRDM9 A to its 13bp motif. Comparing relative binding of PRDM9 A and C alleles to the THE1B motif (Patel et al., 2017) also demonstrated that these PRDM9 alleles activated different sets of hotspots as noted by other approaches (Berg et al., 2011; Hinch et al., 2011; Pratto et al., 2014), although PRDM9 C is more selective compared to PRDM9 A.

Patel et al. (2016) reported that PRDM9 A and L20 differ by a single His to Asn residue change and this lowers binding affinity of L20 to the THE1B motif. Changing the this motif to substitute the relevant C:G base pairs to A:T base pairs such that the His residue binds to the G base on the complementary strand improved binding affinity for L20 but reduced affinity for PRDM9 A. Both alleles have similar binding affinities but the authors speculated a dosage issue. For example, in an individual carrying A/L20, L20 alleles bind preferentially to MSTM1b motif but since the dominant A alleles have high affinities for many hotspots, the proportion of alleles that contribute to MSTM1b hotspot activation is smaller. Hence why L20 is observed as an enhancer of MSTM1b hotspot (Berg et al., 2010).

In summary, both degenerate and non-degenerate positions of the predicated 13bp and 16bp sequence motifs play a role in 'fitting' the ZnF array to the DNA. As such, sequences that differ from these motifs can also be bound by PRDM9 alleles. Whilst predicated sequence motifs for PRDM9 allele binding is a highly useful approach to locating hotspots, comparing the order of ZnFs in PRDM9 ZnF arrays is not a useful way of predicting how well alleles are likely to bind. Instead, direct approaches such as measuring binding affinity, X-ray crystallography and electrophoretic mobility shift assays (EMSA) seem more

likely to be more informative as to the specific means by which PRDM9 binds hotspot DNA. Using these approaches, strong correlations between specific PRDM9 alleles and their best binding motifs may then be identified. However, even here there is some ambiguity as with the case at the MSTM1a hotspot which is activated only by alleles L9 (ABCDDECFGPFQJ) and L24 (ABCDDECFTPFQJ) which differ by one ZnF.

1.5.2 Prospects for PRDM9 allele discovery: sequencing technologies

Traditional dideoxynucleotide terminator or Sanger sequencing has been employed for characterising the majority of PRDM9 alleles known to date (Genbank, 2008; Oliver et al., 2009; Berg et al. 2010; Baudat et al. 2010; Berg et al., 2011; Hussin et al. 2013; Oliver-Bonet, 2013). This characterisation has focused solely on the ZnF arrays as the differentiator. With each ZnF being 84bp long and all currently known ZnF arrays having a range of 8-18 ZnFs, this amounts to some 672-1512bp. With Sanger sequencing, one can achieve up to ~1kb of analysable data and so one read or two overlapping reads is sufficient to characterise the entire ZnF array.

However, processing large sample sets can be time-consuming and costlier than more recent sequencing platforms. Since the mid-2000s, newer sequencing technologies have emerged with higher throughput, lower per sample costs and comparable read quality (Shendure and Lieberman Aiden, 2012). The GS20 pyrosequencing platform produced by 454 Life Sciences/Roche became the first of the 'next generation sequencers' (Mardis, 2008). This was followed by the polony sequencing method, which later became known as the SOLiD system used for deep sequencing (Voelkerding, Dames and Durtschi, 2009). In 2007, the Solexa sequencing method became the basis for the Illumina sequencing platform (Voelkerding, Dames and Durtschi, 2009). All these platforms were similar as they employed sequencing-by-synthesis (SBS) methods. Newer versions of SBS platforms have been produced in the

intervening years, for example the S5 by Thermo Fisher Scientific (Shin et al., 2017) and the HiSeq/MiSeq platform by Illumina (Quail et al., 2012; Nicholas et al., 2012). These so-called second generation sequencing methods all employ a massively-parallel approach of simultaneously producing large numbers of different sequencing reads each at exceptionally high coverage. Another shared feature is that individual read lengths are all relatively short, typically in the range 100-400bp (Quail et al., 2012).

Third generation sequencers have now entered the foray with unique sets of features and capabilities. The two main contenders are the Single-Molecule Real-Time (SMRT) sequencing (Quail et al., 2012) and MinION nanopore (Mikheyev and Tin, 2014) sequencing systems developed by Pacific Biosciences (PacBio) and Oxford Nanopore Technologies (ONT), respectively. Both these approaches provide near-real-time long read data and potentially are more suited to sequence through entire repetitive regions (McGinty et al., 2017) such as PRDM9 ZnF arrays. PacBio sequencing monitors fluorescent light emissions as each nucleotide is incorporated by a DNA polymerase that is immobilised at the bottom of each well of the instrument. The MinION monitors individual changes in the electrical field as long DNA molecules pass through nano-scaled biological pores. In both cases, read length is ultimately dependent on the quality of starting material. For example, using just eleven reads, the MinION was able to sequence and assemble a continuous 2.2Mb sequence with the longest single mapping read being 1.3Mb (Chr5:143515773-144831515) (Payne et al., 2019). However, data from both types of third generation sequencer have tended to suffer from higher error rates than previous technologies (McGinty et al., 2017).

Nanopore sequencing on the ONT MinION device has seen success in characterising short tandem repeats with single base pair resolution (Jain et al., 2015). It has also been used to determine complex genomic rearrangements resulting from aberrant DSBR which forms the underlying genetic basis for many

cancers (McGinty et al., 2017). Since the time of the MinION Access Program (MAP) in which the University of Leicester participated, there has been interest in the capabilities of this device to characterise PRDM9 ZnF arrays.

1.6 PRDM9 and genomic instability

Copy number variation, translocation, ploidy number changes and other structural variants are known to cause complex genetic disease (Zhang et al., 2009; Turner et al., 2008; Piazza and Heyer, 2019). Non-Allelic Homologous Recombination (NAHR) is a key means by which such large-scale variation is thought to arise (Inoue and Lupski, 2002). NAHR can occur when DNA sequences that are >90% identical are found in more than one location in the genome. In such cases, occasionally these paralogous sequences cause mispairing between the homologues and subsequent crossing over generates distinct recombinant chromatids. This can occur between sequences on the same chromosome and on different chromosomes as normal homology searching and pairing is disrupted (Lupski, 2004; Arnheim, Calabrese and Tiemann-Boege, 2007).

More relevantly, Berg et al. (2010) first implicated PRDM9 in meiotic NAHR, specifically in relation to the reciprocal genomic disorders CMT1A and HNPP. Using batch sperm screening, it was demonstrated that carriers of the common PRDM9 A allele have the highest frequency of these events. The genomic origin of these two disorders is located on chromosome 17p11.2p12, and recombination exchanges cluster into two 98% identical 24kb repeats oriented head-to-tail and flanking a ~1.5Mb region containing the Peripheral Myelin Protein 22 (*PMP22*) gene (Lupski, 2004). Misalignment of the repeats from non-sister chromatids can lead to NAHR with the duplication product of this event resulting in a gamete with two copies of the 1.5Mb region which can give rise to children with CMT1A. The reciprocal event involves a deletion of the 1.5Mb region, which ultimately can lead to children with HNPP. The

phenotypes of these two genomic disorders are presumably a consequence of gene dosage problems associated with *PMP22*. Sperm typing of the 1.4kb *CMT1A* hotspot found in the 24kb repeats using men of different *PRDM9* genotypes not only showed that *PRDM9* A has a major effect on this rearrangement, but also indicated that non-A alleles are protective against rearrangement (Berg et al., 2010). Pratto et al. (2014) further investigated the co-localisation of meiotic DSBs and structural variants, finding that homology-based processes such as NAHR were in fact enriched at *PRDM9* A-activated hotspots.

Perfect 8/8 matches to the 13bp *PRDM9* A-associated motif were also found in the highly unstable MS1 and CEB1 minisatellites and a 7/8 match for a third, B6.7 (Berg et al. 2010). These minisatellites mutate to new allele lengths by gene conversion-like events in the germline (Buard et al. 1998, Tamaki et al. 1999, Berg et al. 2003). As with *CMT1A*/HNPP, screening for de novo length changes by small-pool Polymerase Chain Reaction (PCR) (Jeffreys et al. 1994) using DNA from sperm donors with different *PRDM9* genotypes showed a significant effect with non-A alleles reducing activity and therefore being more protective against instability. Alternatively, Ct variants of *PRDM9* may cause pathological genomic rearrangements at alternate hotspot-containing minisatellites and potentially drive instability in them as well. Since *PRDM9* ZnF variants are encoded by a minisatellite which evidently influences its own evolution by generating alleles that cause instability at the *PRDM9* minisatellite, it has the potential to substantially change the recombination landscape through novel *PRDM9* variants (Jeffreys et al., 2013).

1.7 Childhood ALL

Leukaemia refers to cancers of the haematopoietic and lymphatic systems of the body. Acute leukaemia involves immature stem cells and chronic leukaemia involves mature cells. In acute leukaemia, the main underlying

processes of leukaemogenesis is thought to be completed at birth (in utero) (Gale et al, 1997) either through translocation or gain/loss of chromosomes (hyper- and hypo-diploidy). Acute leukaemia is also categorised into myeloid and lymphoid lineages (Greaves, 1999).

1.7.1 Epidemiology

Sporadic leukaemia remains one of the most common diseases in children from developed countries such as Great Britain (Stiller, Allen and Eatock, 1995), USA (Gurney et al., 1995) and Canada (Xie, Onysko and Morrison, 2018). International studies show that ALL occurs in both children and adult demographics but is more prevalent in the paediatric populations of developed countries (Stiller, Allen and Eatock, 1995; Gurney et al., 1995; Siegel, Naishadham and Jemal, 2012; Xie, Steliarova-Foucher, 2017; Onysko and Morrison, 2018).

During the 1980s, 30% of overall childhood cancer incidences were that of leukaemia and 85% of these cases accounted for childhood ALL (Parkin et al., 1988; Linet and Devesa, 1991; Gurney et al., 1995) with 65-70% being of the B-ALL major subtype (Pui, Behm and Crist, 1993). Later international surveys from the 2000s show that incidence rates of childhood ALL in children aged 0-14 years have since increased to 140.6 (95% CI 140.1–141.1) per million person-years by 2010 (Steliarova-Foucher, 2017). However, the gains have been dominated by the peak patient demographic and children under 20 years of age. Incidence of ALL cases are shown to peak at 2-5 years of age and have no overt sex bias (Inaba, Greaves and Mullighan, 2013). The highest incidence rates for leukaemia were seen in Hispanic white and southeast Asian children. Notably, a low incidence in US nationals of African ancestry (Gurney et al., 1995; Taylor et al., 2002) was consistent with earlier reported low frequencies of ALL in ethnically matched groups (Miller and Dalager, 1974; Williams, 1985). Childhood leukaemia has a reported higher incidence rate in children of

European origin compared to admixed African-American children. Through decades of study and clinical management experience, the overall survival for ALL in children of this age group and under 20 years of age has improved from just 10% in the 1970s to 90% in the 2010s. The prognosis and survival rate for infants and adults remain low.

1.7.2 Causal factors

The causal factors influencing the development of ALL appear to be a combination of genetic predisposition (CNVs, translocations, point mutations, SNPs, population post-admixture effects), endogenous and exogenous exposures (infectious pathogens, ionising radiation from nuclear weapons and X-ray sources, non-ionising radiation such as electromagnetic radiation from power lines, chemical agents such as pesticides, etc.) and lifestyle (overstimulation of immune response to viral pathogens after infancy, tobacco, obesity) pertinent to both developed and developing countries (Inaba, Greaves and Mullighan, 2013; Malouf and Ottersbach, 2018).

The genetic basis of ALL is subdivided into B-ALL and T cell precursor ALL (T-ALL) with each category containing both cytogenetically characterised and several undefined subtypes. B-ALL is the more common form, where in the USA, 26% of childhood cancers diagnosed are the various subtypes of B-ALL (Cancer Facts and Figures 2014). The major classified subtypes include translocation-based t(4;11) MLL-AF4, t(12;21) ETV6-RUNX1, t(1;19) E2A-PBX1 and t(9;22) BCR-ABL1 (Malouf and Ottersbach, 2018) and chromosomal rearrangement-based high hyperdiploidy with non-random gain of a minimum 5 chromosomes (including X, 4, 6, 10, 14, 17, 18, and 21) and hypodiploidy with less than 44 chromosomes. ETV6-RUNX1 which constitutes the t(12;21) translocation is found in 25% of childhood B-ALL making it even more common than the Mixed Lineage Leukaemia (MLL) t(4;11) translocations which have been found in only 10% of paediatric patients. MLL-rearrangement is thought to

be the more aggressive form of B-ALL with poor prognosis and the lowest survival rate. Similar to PRDM9, MLL is a H3K4 methyltransferase and it can initiate localised opening up of chromatin to allow DSBs to occur. The *MLL* gene expression is however not meiosis-specific and is expressed in adult haematopoiesis. Due to its higher frequency, the *ETV6-RUNX1* fusion gene and its associated t(12;21) translocation, is a more likely candidate for investigation in relation genomic instability caused by PRDM9 activated hotspot.

Lymphocytes in general originate from pluripotent haematopoietic stem cells that differentiate into myeloid tissues cells. Differentiation commitment takes place as progenitor cells enter the bone marrow (B cells) and thymus (T cells). The prognosis for B-ALL is characterised by B cell precursor cells undergoing differentiation arrest and unregulated production of pre-leukaemic lymphoblasts that are then circulated away from the bone marrow to the spleen, liver, thymus, lymphoid and central nervous system (Malouf and Ottersbach, 2018). This proliferation of lymphoblasts affects the production and activity of other cellular blood components such as red blood cells that transport oxygen and platelets that are required for blood clotting. The resulting immunological and physiological handicaps lead to swelling in the abdomen and lymph nodes, bone tenderness, fatigue and loss of appetite, all of which combine to debilitate patients.

The HapMap project created a haplotype map of the human genome defining narrow hotspot loci interspersed amongst larger haplotype blocks (International HapMap Consortium, 2005). This has greatly aided Genome Wide Association Studies (GWAS) in locating susceptibility loci linked to complex disease with underlying genetic abnormalities. Several GWAS have connected the Major Histocompatibility Complex II (MHCII) locus to childhood ALL (Taylor et al., 2011; Urayama et al., 2013). The MHCII contains a cluster of genes crucial for immunological response (Taylor et al., 2009), including the *HLA-DR*, *DQ* and *DP* genes. B-ALL and T-ALL patients were shown to possess a specific set of

variants of the HLA-DPB1 protein. SNP alleles in the vicinity of HLA-DPB1 were also shown to have a significant association with two independent cohorts of B-ALL (Hosking et al., 2011; Urayama et al., 2013). These studies indicate that SNP alleles identified by GWAS are not fully indicative markers for disease. In fact, it is speculated that a number of low risk changes in the genome must accumulate to result in full blown leukaemia (Malouf and Ottersbach, 2018).

1.7.3 Potential links to PRDM9 activated hotspot

Thompson et al. (2014) reported significantly higher recombination rates at the DNA3 hotspot in the MHCII region in a British B-ALL cohort. Since B-ALL involves the B cell lymphoid lineage and MHCII susceptibility loci and because PRDM9 alleles have been linked to genomic instability (Berg et al. 2010), it is possible that there is an immune aetiology for B-ALL that can be explained by DNA3 hotspot activity and PRDM9 status-dependent abnormal genomic rearrangements, similar to the events leading to the development of CMT1A/HNPP disorders. Furthermore, Hussin et al. (2013) demonstrated an overrepresentation of PRDM9 alleles that are rare in European populations, in parents of children who have developed B-ALL, the most common subtype of childhood ALL, in two independent (French Canadian and American) groups. However, the relatively high frequency of these alleles in African populations (Berg et al. 2010; Hinch et al., 2011) combined with a lower incidence of childhood leukaemia among African Americans suggest that PRDM9 variation alone cannot explain disease aetiology. It has therefore been proposed that the second promotional trigger for B-ALL is some form of environmental exposure. However, the suggested candidate sources of these exposures, such as electromagnetic fields, have been largely irreproducible and often been contested.

1.8 Overall research aims

1.8.1 Research questions and hypotheses

There were several research questions explored in this work:

- 👤 Firstly, whether the DNA3 hotspot is indeed PRDM9-A regulated as it was hypothesised based on the fact that the earlier study by Jeffreys, Kauppi and Neumann (2001) using men of North European origin possessing predominantly A alleles. However, Hussin et al. (2013) and Thompson et al. (2014)'s work would be reinforced if Ct or K-ZnF containing alleles also influenced this hotspot contrary Hinch et al's (2011) predictions.
- 👤 Secondly, a putative African-American (AA) hotspot near to DNA3 hotspot was hypothesised to be PRDM9 Ct-activated based on the fact that the African-enriched genetic map produced by Hinch et al (2011) had a strong association with the SNP rs6889665, a predictive marker for Ct alleles. A study needed to be done to establish whether this putative hotspot was the activator and genuine source of the elevated recombination rates as reported by Thompson et al. (2014) and therefore could have a link with the K-ZnF containing PRDM9 alleles as reported by Hussin et al (2013). Overall, PRDM9 activation of DNA3 and AA hotspots would help establish whether there is a relationship with PRDM9-associated genomic rearrangement in the MHCII susceptibility loci for B-ALL, similar to CMT1A/HNPP as reported by Berg et al (2010), in the process painting a clearer picture of the local recombination landscape in the MHCII in different populations.
- 👤 Thirdly, if the PRDM9 trans-regulation of DNA3 and AA hotspots were established with both or either hotspot pointing to an association with K-ZnF containing alleles, a British ALL cohort was available to screen for these alleles using predictive SNPs. Reproducing an excess of these alleles in this independent cohort would align with Hussain et al (2013)'s data on two independent cohorts. However, as previously hypothesised, it is predicted

that any K-ZnF containing PRDM9 alleles would be linked to AA hotspot activation and not that of DNA3 hotspot.

- 👤 Fourthly, it was predicted that several SNPs identified by Dr Pamela Thompson at the University of Manchester through GWAS would have a strong association with B-ALL as they were in LD with disease susceptibility loci containing genes related to immunological response. To confirm this, SNP genotyping on the British ALL cohort and healthy controls was required.
- 👤 Lastly, it was confirmed that non-North European populations have a greater diversity of PRDM9 alleles as demonstrated in major African subpopulations. Since Jeffreys et al. (2013) found several hundred novel PRDM9 alleles in sperm mutants, PRDM9 diversity is hypothesised to be much greater. A larger sampling of more diverse world populations would help to further uncover this diversity. In turn, studies of non-A, Ct, and non-Ct allele hotspot regulation may reveal further differentiation in recombination landscapes compared to the data presented by Hinch et al (2011). Additionally, Sanger sequencing is not entirely practical for characterising PRDM9 alleles in large sample sets. The main drawback of most NGS platforms is the shorter read lengths that cannot resolve 84bp long ZnF repeats sometimes arranged as homopolymers in large ZnF arrays. Hence, initially it was hypothesised that these methods were not suitable to replace Sanger method. Yet, recent innovations in second generation sequencing and third generation long read platforms aspire to overcome this pitfall. It was worthwhile to assess to these platforms to redetermine their capability to accurately assemble PRMD9 ZnF arrays, the benefit mainly of being able to screen larger sample sets at a time.

1.8.2 Research aims

The aims of this project were to explore the themes of PRDM9 ZnF array variability in diverse populations and the role of rare PRDM9 alleles in the

disease aetiology of B-ALL. To achieve this, multiple studies involving a variety of methods and targets were investigated as shown in Fig. 7.

Batch sperm typing assays for European and African donors carrying a variety of PRDM9 alleles were used to confirm the PRDM9 regulation of two meiotic hotspots, DNA3 located in the MHCII and already associated with ALL via our collaborator Dr Pamela Thompson at the University of Manchester, and a newly described AA hotspot identified by LD-based approaches by Hinch et al. (2011). Through mapping sperm CO resolution points, the AA hotspot was further characterised terms morphology at the sub-kilobase level.

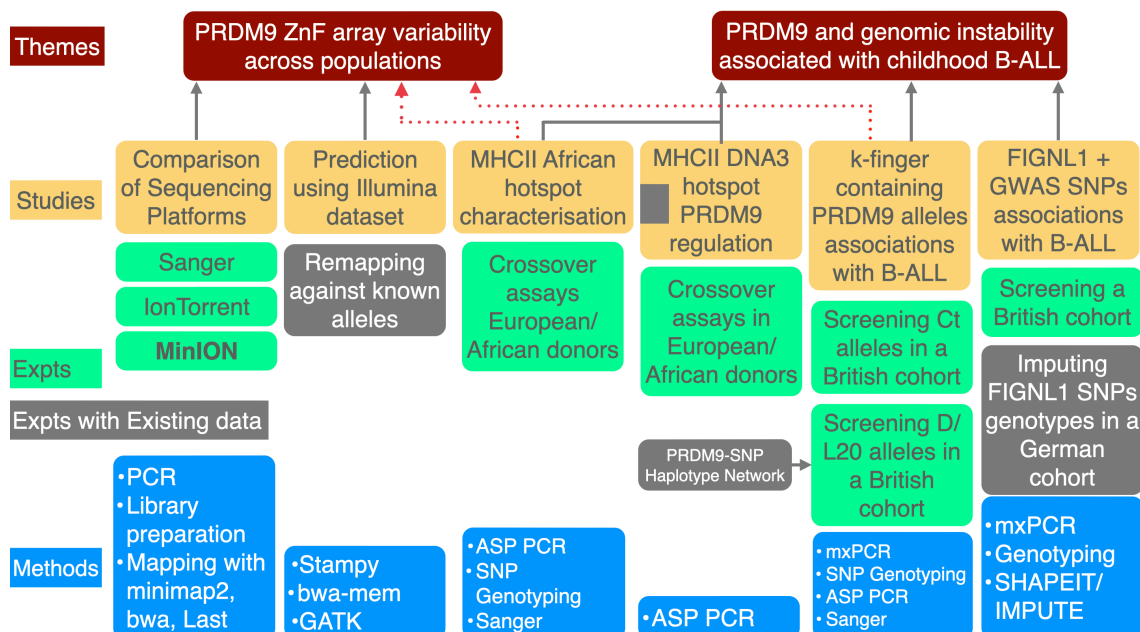


Fig. 7 Project plan

Using predictive SNP markers for K-ZnF containing Ct, D and L20 alleles of PRDM9, multiplex PCR and SNP genotyping methods were used to screen a British B-ALL cohort for potential carriers. Statistical comparisons such as Odds Ratio (OR) were made to determine any differences between patient and control groups. Sanger sequencing was then used to characterise the PRDM9 ZnF arrays of representative sample of potential carriers to confirm whether

they carried K-ZnF containing alleles. This was a collaboration with Dr Pamela Thompson at the University of Manchester.

As part of the same collaboration, additional SNPs linked via GWAS to ALL susceptibility were also genotyped in the British ALL cohort. OR comparisons were made to determine any differences between patient and control groups. To assess the reproducibility of these results, a SNP genotype dataset from a German ALL cohort was used. Whilst these SNPs were genotyped in the British ALL cohort, they were not typed in the German dataset. Therefore, HapMap and the British ALL cohort data were phased bioinformatically and used to construct a reference panel in order to impute the SNP genotypes in the German ALL cohort and ultimately calculate ORs.

Finally, an existing Illumina HiSeq sequencing read dataset corresponding to the PRDM9 gene region derived from a diverse set of populations was remapped to known PRDM9 alleles and as well as mapped de novo to assess whether second generation sequencing data are capable of accurately characterizing PRDM9 ZnF arrays. The approach was validated by Sanger sequencing in a subset of the individuals. PRDM9 ZnF array amplicons recovered from these individuals were also used to sequence the ZnF arrays on the Ion Torrent platform with both 200bp and 400bp read length sequencing chemistries, as well as by MinION nanopore sequencing. In addition to acting as a comparison of sequence platforms and their ability to faithfully characterise PRDM9 ZnF arrays, this study was overall used to search for novel PRDM9 alleles.

CHAPTER 2: MATERIALS AND METHODS

2.1 Materials

2.1.1 Chemical reagents

2.1.1.1 PCR

- 💧 Water (H₂O) - rated at 18.2MΩ (Sigma-Aldrich/Merck KGaA)
- 💧 5 Units/μl BIOTAQ *Taq* DNA polymerase (Bioline/Meridian Bioscience Inc.)
- 💧 5 Units/μl KAPA *Taq* DNA polymerase (Sigma-Aldrich/Merck KGaA)
- 💧 2.5 Units/μl *Pfu* polymerase (Promega Corporation)
- 💧 1M Tris Base
- 💧 3 mg/mL high-molecular-weight herring DNA in dH₂O
- 💧 3 mg/mL high-molecular-weight salmon DNA in dH₂O
- 💧 11.1x PCR reaction buffer prepared according to Kauppi, May and Jeffreys (2009) (Table 2):

Table 2 11.1x PCR reaction buffer composition

Component	Stock Concentration	Volume ratio	PCR Reaction Concentration
Tris-HCL pH8.8	2M	167	45mM
Ammonium Sulphate	1M	83	11mM
MgCl ₂	1M	33.5	4.5mM
2-mercaptoethanol	100%	3.6	6.7mM
EDTA pH8.0	10mM	3.4	4.4μ
dATP	100mM	75	1mM
dCTP	100mM	75	1mM
dGTP	100mM	75	1mM
dTTP	100mM	75	1mM
Bovine serum albumin (Ambion)	10mg/ml	85	113μg/ml
:676 total volume			

- 🧪 1M Tris-HCL pH7.5
- 🧪 Universal and Allele-Specific Primers (ASPs) (Protein Nucleic Acid Chemistry Laboratory, Sigma-Aldrich/Merck KGaA) (Appendix I)

2.1.1.2 Gel Electrophoresis

- 🧪 Water (H₂O) - rated at 18.2MΩ
- 🧪 0.8% w/v Seakem LE Agarose (Lonza)
- 🧪 1 × Tris-Borate EDTA (TBE) with 0.5 µg/ml Ethidium Bromide (EtBr)
- 🧪 100% Glycerol
- 🧪 Loading dye: Bromophenol Blue (Bioline)
- 🧪 ΦX174 DNA/BsuRI (HaeIII digest) (72, 118, 194, 234, 271, 281, 310, 603, 872, 1078, 1353 bp markers) (New England Biolabs Inc.)
- 🧪 λ DNA (HindIII digest) marker (125, 564, 2027, 2322, 4361, 6557, 9416, 23130 bp markers) (New England Biolabs Inc.)

2.1.1.3 SNP genotyping

- 🧪 All purpose loading dye made up of 30% (v/v) glycerol, 0.5x TBE, bromophenol blue
- 🧪 Allele-Specific Oligonucleotides (ASO) (Appendix II)
- 🧪 Denaturing Mix made up of 0.5M NaOH, 2M NaCl, 25mM EDTA
- 🧪 10x Kinase Mix made up of 700mM Tris-HCl, pH 7.5, 100mM MgCl₂, 50mM spermidine trichloride, 20mM dithiothreitol and 60µl dH₂O
- 🧪 T4 polynucleotide kinase (New England Biolabs)
- 🧪 10 mCi/mL γ-³²P ATP (Perkin-Elmer)
- 🧪 Kinase Stop Solution made up of 25mM diNa EDTA, 0.1% SDS, 10µM ATP and 4.7ml dH₂O
- 🧪 50x Denhardt's Solution made up of 5g ficoll 400, 5g polyvinylpyrrolidone, 5g BSA (fraction V, Sigma) and H₂O to 500ml

- 🧪 TMAC Hybridization Solution made up of 3M $(\text{CH}_3)_4\text{N}(\text{Cl})$ [Tetramethyl ammonium chloride], 0.6% SDS, 1mM diNaEDTA, 10mM Na phosphate pH6.8, 5x Denhardt's solution, 4 $\mu\text{g}/\mu\text{l}$ yeast RNA, 27.6 ml dH₂O
- 🧪 1M Sodium Phosphate Buffer pH6.8 made from 1M Disodium Hydrogen Phosphate (Na_2HPO_4) and 1M Sodium Dihydrogen Phosphate (NaH_2PO_4). 1M stock at pH6.8 = 46.3ml 1M Na_2HPO_4 + 53.7ml 1M NaH_2PO_4 . The following protocol is derived from Molecular Cloning: A Laboratory Manual by T. Maniatis et al. (Malke, 1984)
- 🧪 TMAC Wash Solution made up of 3M TMAC, 0.6% SDS, 1mM diNaEDTA, 10mM Na phosphate pH6.8
- 🧪 3x Saline Sodium Citrate (SSC), 2L, made up from 300ml 20x SSC and 1800ml distilled water
- 🧪 2x Saline Sodium Citrate (SSC), 1L, made up from 150ml 20x SSC and 850ml distilled water
- 🧪 Autoradiograph-film developing solutions (developer, stop and fixer)

2.1.1.4 Sanger Sequencing

- 🧪 20U/ μl Exonuclease 1 (New England Biolabs Inc.)
- 🧪 1U/ μl Shrimp Alkaline Phosphatase (New England Biolabs Inc.)
- 🧪 BigDye® Terminator v3.1 5 \times Sequencing Buffer (Thermo Fisher Scientific)
- 🧪 BigDye® Terminator v3.1 Ready Reaction Mix (Thermo Fisher Scientific)
- 🧪 2.3 μM sequencing primer (Protein Nucleic Acid Chemistry Laboratory) (Appendix I)
- 🧪 2.2% w/v SDS. Alternatively, gel electrophoresis, band excision and Zymoclean™ Gel DNA Recovery Kit (Zymo Research) were used to purify PCR products.

2.1.1.5 Ion Torrent sequencing

- 🧫 Agilent DNA 1000 Reagents (Agilent)
- 🧫 Agencourt® AMPure XP beads (Beckman Coulter)
- 🧫 Ion Xpress™ Library kit and barcodes (Thermo Fisher Scientific)
- 🧫 Ion PGM™ HI-Q OT2 Kit (Thermo Fisher Scientific)

2.1.1.6 Nanopore sequencing 2015 R7 flow cell developer version

- 🧫 DNA 'CS' control template
- 🧫 10x NEBNext End Repair buffer (New England Biolabs Inc.)
- 🧫 NEBNext End Repair enzyme mix (New England Biolabs Inc.)
- 🧫 10x NEBNext dA-Tailing buffer (New England Biolabs Inc.)
- 🧫 10x NEBNext dA-Tailing enzyme mix Klenow exo- (New England Biolabs Inc.)
- 🧫 Agencourt® AMPure XP beads (Beckman Coulter)
- 🧫 Genomic004 (SQK-MAP004) kit made up of Wash buffer, Elution Buffer, HP adaptor, Adapter Mix, Fuel Mix, EP buffer (Oxford Nanopore Technologies Limited)
- 🧫 Blunt/TA Master Ligase Mix (New England Biolabs Inc.)
- 🧫 Nuclease-free water
- 🧫 Dynabeads™ His-Tag Isolation and Pulldown beads (Thermo Fisher Scientific)

2.1.1.7 Nanopore sequencing 2019 R9.4.1 flow cell chemistry

- 🧫 Agencourt® AMPure XP beads (Beckman Coulter)
- 🧫 NEBNext FFPE Repair Mix (M6630)
- 🧫 NEBNext End repair / dA-tailing Module (E7546)
- 🧫 NEBNext Quick Ligation Module (E6056)
- 🧫 Nuclease-free water
- 🧫 Freshly prepared 70% ethanol in nuclease-free water

2.1.2 Reagent kits

- 🧪 Zymoclean™ Gel DNA Recovery Kit (Zymo Research)
- 🧪 Performa® Gel Filtration Cartridge (EdgeBio)
- 🧪 Agilent DNA 1000 chip (Agilent)
- 🧪 Ion PGM™ 200 Xpress™ Template Kit (Thermo Fisher Scientific)
- 🧪 Ion PGM™ 400 Xpress™ Template Kit (Thermo Fisher Scientific)
- 🧪 Ion 316™ Chip Kit v2 BC (Thermo Fisher Scientific)
- 🧪 Qubit™ dsDNA HS Assay Kit (Thermo Fisher Scientific)
- 🧪 Genomic004 (SQK-MAP004) kit (Oxford Nanopore Technologies Limited)
- 🧪 Ligation Sequencing Kit (SQK-LSK109) (Oxford Nanopore Technologies Limited)

2.1.3 Labware

- 🧪 DNA LoBind tubes (Eppendorf)
- 🧪 Protein LoBind tubes (Eppendorf)
- 🧪 High profile 0.2ml PCR tubes
- 🧪 96-well non-skirted skirted plates
- 🧪 96-well semi-skirted plates
- 🧪 96-well skirted plates
- 🧪 96-well adhesive tape
- 🧪 0.2ml PCR 8-tube strips
- 🧪 0.2ml PCR 8-tube strip caps
- 🧪 MicroAmp™ Optical Adhesive Film
- 🧪 Screw-top 1.5-mL microcentrifuge tubes
- 🧪 0.1-10, 1-200 and 100-1,000µl pipette tips, sterile (pre-PCR) or autoclaved (post-PCR)
- 🧪 500ml and 1L Duran bottles, autoclaved
- 🧪 Whatman™ 3MM blotting paper
- 🧪 Nylon membrane filter (Bio-Rad)

- 🔧 X-ray films (Fuji)
- 🔧 Whatman® Benchkote® surface protector
- 🔧 SaroGold high quality food-wrap

2.1.4 Instruments and equipment

- 🔧 Biosafety Level II hood
- 🔧 Veriti™ thermal cyclers (Thermo Fisher Scientific)
- 🔧 0.1-2, 4-25, 1-200, 100-1,000µl pipettes (various brands)
- 🔧 Ice bucket with ice
- 🔧 Eppendorf 5415 D centrifuge with 24 tube rotor
- 🔧 Plate centrifuge
- 🔧 Vortex mixer
- 🔧 Agarose gel tray and combs (20, 34 and 47 teeth)
- 🔧 Gel tank and power supply
- 🔧 GeneGenius Gel Imaging System (Syngene)
- 🔧 Genetic Analyzers (3130xl, 3700) for Sanger sequencing (conducted by PNACL/Source Bioscience)
- 🔧 Precision weighing balance, $\pm 0.0001\text{g}$
- 🔧 Water bath
- 🔧 Timer
- 🔧 Cold room at 4-8°C
- 🔧 -20°C freezer
- 🔧 Dotblot manifold
- 🔧 1000ml Büchner flask with rubber stopper and tubing
- 🔧 Suction pump
- 🔧 Heating block
- 🔧 Hand-held UV reader
- 🔧 Hand-held dark reader (non-UV)
- 🔧 Benchtop dark reader with screen and goggles (non-UV)
- 🔧 Rotisserie oven for hybridisation tubes

- 🧪 Hybridisation tubes
- 🧪 Geiger counter
- 🧪 Cin Bin (incineration)
- 🧪 Cassettes for X-ray film
- 🧪 -80°C freezer
- 🧪 Dark room with Developer, Stop and Fixer solution
- 🧪 2100 Bioanalyzer (Agilent)
- 🧪 Ion Torrent™ PGM™ 316 Sequencer (Thermo Fisher Scientific)
- 🧪 Magnetic separator 1.5 ml Eppendorf tubes
- 🧪 Hula/rotator mixer
- 🧪 Nanodrop ND1000 spectrophotometer (Thermo Fisher Scientific)
- 🧪 Qubit™ fluorometer (Invitrogen)
- 🧪 Ion Personal Genome Machine®
- 🧪 MinION Nanopore sequencers (Oxford Nanopore Technologies Limited)

2.1.5 Online resources

- 🧪 UCSC Genome Browser (<https://genome.ucsc.edu>)
- 🧪 dbSNP/NCBI SNP database (<https://www.ncbi.nlm.nih.gov/snp/>)
- 🧪 GenBank for sequences of PRDM9 alleles (<https://www.ncbi.nlm.nih.gov/genbank/>)
- 🧪 1000 Genomes Project
- 🧪 Mathematical Genetics and Bioinformatics Groups: Software for HapMap data used in ShapeIT reference panel (<https://mathgen.stats.ox.ac.uk>)
- 🧪 Galaxy platform (<https://usegalaxy.org>)
- 🧪 Prof. Alec Jeffreys' Mutation and Recombination Research Program for annotated sequences of the MHCII region and SNP genotype data (<https://www.le.ac.uk/ge/ajj/labhome.html>)

2.1.7 Software

2.1.7.1 Commercial/Publicly available software

- 🔗 Factura™ Feature Identification v1.2.0 - Applied Biosystems © 1993
- 🔗 AutoAssembler v1.4.0 - Perkin Elmer © 1995
- 🔗 OligoAnalyzer (Integrated DNA Technologies)
- 🔗 Reverse Complement at bioinformatics.org (Scilico, LLC)
- 🔗 MacVector v17.0.5
- 🔗 ShapeIT v2.17
- 🔗 Impute2 v2.3.2
- 🔗 PHASE
- 🔗 Samtools v1.9
- 🔗 Picard tools v1.119 and v1.124
- 🔗 Torrent Suite™ Software 5.0.2
- 🔗 Integrative Genomics Viewer v2.3.46 and v2.4.19 (Broad Institute and the Regents of the University of California)
- 🔗 Burrows-Wheeler Aligner (BWA) v0.7.10-r789
- 🔗 Minimap2/miniasm
- 🔗 macOS Mojave v10.14.6 (18G87) (Apple Inc.)
- 🔗 OS X v10.3, v10.4, v10.7 and v10.8 (Apple Inc.)
- 🔗 Pages v8.1 (6369) - Apple Inc.
- 🔗 Numbers v6.1 (6369) (Apple Inc.)
- 🔗 Microsoft Excel

2.1.7.2 In-house software

- 🔗 CO Poisson Calculator True Basic® v.4.1 (developed by Prof AJ Jeffreys)
- 🔗 Gel band size Calculator True Basic® v.4.1 (developed by Prof AJ Jeffreys)
- 🔗 Diploid LD |D'| plot program True Basic® v.4.1 (developed by Prof AJ Jeffreys)

- 👤 LDU mapping (program and web interface developed by Dr Adam Webb)
- 👤 Haplotype Extractor Program (developed by Prof AJ Jeffreys, used by Jon H Wetton)
- 👤 ALICE High Performance Computing (HPC) cluster
- 👤 SPECTRE (Special Computational Teaching and Research Environment)

2.1.6 Samples

2.1.6.1 Human DNA


- 👤 The University of Leicester's North European semen donor panel (NE panel) and Sub-Saharan African sperm donor panel (SSA panel) were collected in the early 1990s (Monckton et al., 1994; Jeffreys, Kauppi and Neumann, 2001) and ethical approval has been granted for their use in de novo mutation and recombination studies (Leicester NRES Committee, East Midlands, REC reference 6659 (14/EM/0135) granted to Dr Celia A May). The samples have been anonymised and sperm DNA was prepared according to Kauppi, May and Jeffreys (2009).
- 👤 The University of Manchester's British ALL cohort consisted of blood samples from B-ALL patient and family members collected with informed consent and ethical approval (Taylor et al., 1995; 'The United Kingdom Childhood Cancer Study: objectives, materials and methods. UK Childhood Cancer Study Investigators', 2000; Taylor et al., 2009). Control DNA samples were obtained from umbilical cord blood of healthy newborns from St Mary's Hospital, Manchester, UK between 1991 and 1999.
- 👤 The University of Leicester's Illumina HiSeq 2000 dataset was generated using DNA samples obtained by informed consent and ethical approval granted to Professor Mark A Jobling (University of Leicester Research Ethics Committee reference: maj4-cb66). DNA from these anonymous lymphoblastoid cell lines, peripheral blood or saliva samples was prepared

according to Batini et al. (2015). Table 3 shows the various population groups in the dataset/DNA samples.

Table 3 Population/groups in the Illumina HiSeq2000 dataset

Abbreviation	Population/country of origin	#
him	Himalayan and Bhutan samples	31
eng	England [Herefordshire and Worcestershire]	29
TSI	HapMap	23
gre	Greece	22
CEU	HapMap	21
YRI	HapMap	21
ser	Serbia	20
hun	Hungary	20
bav	Germany [Bavaria]	20
bas	Spanish Basque country	20
spa	central Spain	20
fri	Netherlands [Frisia]	20
den	Denmark	20
nor	Norway	20
saami	Finland [Saami]	20
ork	Orkney	20
tur	Turkey	20
pal	Palestinians	20
CHB	HapMap	20
ire	Ireland	19
baka	Cameroon/Gabon	6
Mbuti	Congo	5
Bakola	Cameroon	3
JPT	Japan	3
mbe	Republic of Congo	3
Aus	Australia	2
LWK	HapMap	2
Biaka	CAR/Congo	1
kung	San people/Southern African	1
Man	Isle of Man	1
melan	Albania	1
MXL	HapMap	1
Ngoumba	Cameroon	1
		<hr/> 456 individuals

2.1.7.2 Datasets

 The Illumina HiSeq 2000 sequence data was kindly provided by Prof Mark A Jobling. The original dataset was generated according to Batini et al. (2015) and the BAM files for the *PRDM9* gene ZnF array region were prepared by Dr Pille Hallast.

- 👤 A German ALL cohort SNP genotype dataset was kindly provided by Prof Martin Stanulla, Hannover Medical School. This SNP genotype dataset was obtained from ETV6-RUNX1+ patients and healthy controls according Ellinghaus et al. (2012)
- 👤 HapMap 3 SNP genotype reference panels for haplotype phasing and imputation of FIGNL SNP genotypes were obtained from the Mathematical Genetics and Bioinformatics Groups: Software at the University of Oxford (https://mathgen.stats.ox.ac.uk/impute/impute_v2.html#reference)

2.2 Methods

2.2.1 PCR

All PCRs were carried out using the custom 11.1x PCR reaction buffer as detailed in Table 2 with the additional components described here. In PCR with universal primers (Appendix I), typically 10µl reactions containing 5-10ng of total DNA were used. All PCRs used a 20:1 mix of *Taq*/*Pfu* polymerase enzymes were used with a final reaction concentration of 0.03U/µl *Taq*, 0.0015U/µl *Pfu*. The *Pfu* enzyme has 3' to 5' exonuclease activity which removes mis-incorporated nucleotides at the 3' end produced by *Taq* polymerase. HMW herring and salmon was used to cocoon the target DNA indirectly assisting the access to amplification assembly and also blocking the pits on the inner surface of the plasticware used.

PCR cycling programs were optimised for all primer pairs, firstly by designing compatible primers using software tools and by eye, and secondly by assessing their performance empirically via annealing temperature titrations. As a result, the main differences between PCR cycling programs were in the annealing temperatures and extension times, the latter of which which depended on the allowance of 1min per 1kb or 1-2min more if a particular DNA sample was deemed to be degraded. Degradation tests were done by amplifying

using primer pairs with increasing amplicon lengths. Typical PCR cycling conditions included an initial denaturation holding stage at 96°C for 1min, 26-27 cycles at 96°C for 20s, variable annealing temperatures (53-61°C) for 30s and amplicon length dependent extension at 61/62/65°C at extension times allowing for 1min per kb. This was followed by a standard 2min final extension at 61/62/65°C. Details of major PCR designs used can be found in Appendix I.

Concentrated unprecipitated primer stock from PNACL were precipitated, as an example, first by adding 200µl of crude primer preparation in a 1.5ml eppendorf tube, then adding 20µl 2M Na Acetate pH7.0 and 600µl 100% EtOH and mixing thoroughly. The volume of 2M Na Acetate was calculated to be 1/10th of the volume of crude primer preparation and the volume of 100% EtOH to be 3 times the volume of crude primer preparation. The mix was kept at -80°C for 1 hour or at -20°C overnight and spun down at full speed (~140000rpm) for at least 20mins. Translucent or off-white coloured pellets, often around 1mm in diameter, formed at this stage. The supernatant was removed and the pellets were washed with 800µl 80% EtOH (or a volume at least 200µl higher than the volume of 100% EtOH used for precipitation). The tube was flicked and inverted a few times, taking care not to disturb the pellet. The mix was then spun down at full speed for 5min. The EtOH was carefully removed by spinning and removing a few times to ensure that the liquid was removed as much as possible. This was followed by vacuum drying for 2min. The mix was then redissolved in 45µl deionised water, mixed vigorously and spun down, repeating three times. A 1 in 5 dilution of concentrated primer in 5mM Tris pH7.5 was prepared. The concentration of primer was measured in ng/µl using a Nanodrop spectrophotometer. Actual concentration of primer was calculated by accounting for the dilution. Working stocks of 10µM primer were then made using 5mM Tris pH7.5 or rather PCR water, when making fresh primer for immediate use in PCR.

2.2.2 Sperm crossover assay

Sperm crossover studies were performed according to Kauppi, May and Jeffreys (2009). The specific type of assay used was batch sperm-typing where recombinant molecules were selectively amplified in two rounds of repulsion-phase allele-specific PCR on pools of sperm DNA from men with suitable SNP heterozygosity. SNP genotyping via radio-labelled ASO hybridisation was done over large interval overlapping amplicons (6-9kb amplicons, ~15kb total length) surrounding hotspots.

ASPs were designed such as that SNPs were located at the 3' end of the primers, specifically on the last base in the sequence. Software such as OligoAnalyzer and BLAT search, visual inspection of sequence context and evaluation of properties such as melting temperature, G-C content, runs of homopolymers, etc. were used to examine the specificity of the primers and the products that could potentially generated. Ultimately, specificity and yield for all primer pairs were evaluated empirically using annealing temperature titrations. Examples of this work can be found in Chapters 3 and 4. ASP specificity was assessed using men who were homozygous for either SNP allele using annealing temperature titration PCRs and ASPs were selected for specificity and yield. Linkage phasing was done on individuals suitable for crossover analysis using radio-labelled ASO hybridisation of allele-specific PCR products from each man. As stated before, CO assay was done on both orientations where possible using two rounds of allele-specific PCR. The primary PCR was typically seeded with 4-5ng of sperm DNA and the secondary PCR was seeded with a 1 in 10 or 1 in 20 dilution of the primary PCR product. CO positive and negative reactions were visually identified by 0.8% gel electrophoresis. Then, the crossover breakpoints were mapped using hybridisation of radio-labelled ASOs to tertiary PCR products using internal universal primers. To improve the estimates of the true number of crossover events, CO Poisson Calculator by Prof AJ Jeffreys was used to Poisson adjust which account for the potential existence of more than

one recombinant molecule in a single reaction. Further details on the technique can be found in the studies conducted in Chapters 3 and 4.

2.2.3 Gel Electrophoresis

To make 0.8% Agarose gels, custom gel trays in 20x20cm, 20x35cm and 10x10cm formats were used with several 20-, 34- and 47-comb arrays placed in appropriately sized-gel tanks. The open ends of the gel tray were taped and comb arrays were inserted into the slots on the gel tray so that they were atleast 4mm from the inner base of the gel tray. In an appropriately sized duran bottle, weighed 1.6g of LE agarose was added to 200ml of 0.5x TBE w/o Et Br for a concentration of 0.8% agarose or when working with PRDM9 ZnF array amplicons, 3.2g of LE agarose was added to 200ml of 0.5x TBE w/o Et Br for a concentration of 1.6% agarose in long format 20x35cm gel trays. The mixture was heated first for 8-10min at medium power, taking out the bottle using thick rubber gloves, tightening the cap and shaking the mixture to help dissolve the agarose. The mixture was checked for streaks of undissolved agarose. The mixture was then heated at medium-high power for 5min, then taking the bottle out and mixing as before. If the agarose does not dissolve, the bottle was shaken more vigorously and heated at full power for 2min taking care not to let any liquid evaporate from the bottle during heating. When the agarose was fully dissolved, the solution was cool by standing on benchtop at room temperature or cooled under running water with vigorous shaking taking care not to cause the agarose to solidify prematurely. Then, 10 μ l of EtBr was added to make 0.5 μ g/ml EtBr in 200ml of 0.5xTBE. The mixture was shaken well to homogenise. The solution was poured into the gel tray after which 70% IMS gently sprayed over the solution to remove bubbles on the surface and around the comb arrays. The gel was left to set for atleast 30min in a cold room at 4-8°C. After the gel had solidified, the combs were removed and placed in the landing of the gel tank filled with 0.5x TBE w/ 0.5 μ g/ml EtBr buffer solution with the sample loading wells towards the negative electrode. The gel tank lid was closed and the gel

was allowed to equilibrate for at least 30 mins before loading samples. To make H₂O:Dye 2:1 Loading Dye, two volumes of dH₂O to one volume of 5x Loading Dye (30% Glycerol, 0.5x TBE) were mixed well by pipetting or flicking and spun briefly. To make λ:φ 1:1 DNA Ladder, 30μl aliquots of λ DNA– HindIII Digest at 200ng/μl and φX174 DNA- HaeIII Digest at 200ng/μl were mixed together with 60μl of water and 30μl of 5x loading dye. If the PCR products were not required for further downstream processes such as secondary PCRs and sequencing, then the loading dye was added directly to the PCR product, typically 2-2.5μl of H₂O:Dye 2:1 loading dye to 10μl of PCR product but 4-6μl was also used depending on the number of cycles used in the PCR. If the PCR products were required for further experiments, 2.5-4μl loading dye to dedicated 96-well plates and 1-2.5μl of PCR product was added depending on the number cycles used in the PCR and/or after adjusting for the expected yield of the product in the conditions used. Pipetting was used to mix. For the DNA ladder standard for comparison, 5μl i.e. 250ng (or a range such as 1, 2, 2.5, 4 and 5μl if yield of PCR product is to be estimated) of λ:φ 1:1 DNA Ladder was loaded in addition to 5-6.5μl of test sample PCR products, positive controls and negative controls. DC power supply at ~40-140V was used as volages below 40V were too weak and allowed DNA to diffuse into the gel instead of travelling towards the positive electrode. If the voltage is too high, the force with which DNA bands travel caused smears as the amplicons failed to move together. Smears can also be caused by amplicons collapsed during the PCR or the presence of a minisatellites within the target region resulting in a series incomplete amplification products. Therefore, great care had to be taken in producing PRDM9 ZnF array amplicons and also cutting out gel fragments only containing the target band whilst separating away the smears. After initiating the run, the gel was checked every 30min with a handheld UV wand to illuminate EtBr staining on the PCR products. Mostly, the run was stopped when the DNA ladder bands were fully resolved (roughly 2hr with 140V). In the case of multiple target bands such as two heterozygous PRDM9 ZnF arrays with only 84bp difference generated using PCR with universal primers, the run was allowed to

continue for longer at 90V up to 18hrs. The gel was then loaded on the Gene Genius Bio Imaging System (Syngene) and the GeneSnap program to open live images. EtBr has high UV absorbance at 300 and 360 nm, and an emission maximum at 590 nm. Limits of detection for EtBr-bound DNA is 0.5-5.0 ng/band. Aperture, motor-operated zoom, focus and exposure controls were used to obtain a focused and bright image. A still image was then captured and saved as in the proprietary .sgd file format. Further editing tools were used to obtain images with high contrast between the DNA bands and background. Print the image. The edited image was saved as a standard .jpeg file and a printout of the image was also made for annotation and comparisons. Image acquisition was repeated several times to obtain a collection of photos that could be helpful in data analysis. Images at different exposure levels helped to identify weak PCR samples which were not visible with lower exposures. Images were routinely printed with black and white colours inverted. The goal was to derive as much information as possible for interpretation. Yields of PCR products were estimated using the ladder marker information as a guide, where 5µl of λ:φ 1:1 DNA Ladder would be a total of 250ng. The ladder marker closest in size (or similarity in migration length) was compared by eye to the target PCR product. Hand-held UV wands and dark readers were used appropriately. Hand-held and benchtop dark readers were used if the PCR products were required for downstream processes such as SNP genotyping and Sanger sequencing.

2.2.4 SNP genotyping

Except for small-scale work done using Sanger sequencing, SNP genotyping in all the studies included in this work was predominantly done using γ -P³²-labelled ASO hybridisation as described by Kauppi, May and Jeffreys (2009). Briefly, PCR products of single target regions or overlapping multiple regions were generated and dotblotted onto nylon membranes, making replicates according to need. ASOs were hybridised TMAC-based solutions using the following conditions: Pre-hybridisation of dotblots at 54°C for a

minimum 10min, adding labelled probes, hybridisation at 52°C for a minimum 1-2hr and washing at 53-54°C for 10min. The dotblot membranes were then washed in SSC and signals captured by autoradiography. Once adequate films had been developed for genotype scoring, the radioactive probes were washed off with SDS, rinsed and stored with SSC. The scoring data was tabled and analysed in Apple Numbers software and/or Microsoft Excel.

Use of TMAC in this method eliminated the issue of the G-C content of oligonucleotide probe on binding with the PCR products. Thence, hybridisation depended on the length of the oligonucleotides designed. To make 1M Sodium Phosphate Buffer pH6.8 used in this method, a protocol from Molecular Cloning: A Laboratory Manual by T. Maniatis et al. (Malke, 1984) was used. Anhydrous Na_2HPO_4 and NaH_2PO_4 were dissolved in deionised water to make two separate 1M solutions and combined to a ratio of 1:1598, respectively, to reach at pH of 6.8. This was checked using universal indicator and adjustments were made using Phosphoric acid (H_3PO_4) if the buffer solution was too alkaline and Sodium Hydroxide (NaOH) if was too acidic.

2.2.5 Sanger Sequencing

The PCR amplicons for PRDM9 ZnF arrays and other amplicons prepared for sequencing (Appendix I) were separated by gel electrophoresis (0.8% (w/v) LE agarose), physically excising gel fragments containing PCR product bands with the help of a benchtop Dark Reader. Fragments were excised from the gel and purified using Zymoclean Gel DNA Recovery kit (Zymo Research) and Sanger sequenced. During universal and allele-specific PCRs, once the length heterozygous ZnF arrays or separated haplotypes had been run and excised from the gel, the PCR products were run on a new gel to further purify the target product. For the PRDM9 ZnF array, two primers for both ends of the amplicons were used to allow sufficient overlap for ZnF array reconstruction. From ABI file format files, the sequence read ends were trimmed using Factura™ Feature

Identification v1.2.0 (Applied Biosystems © 1993). Sequence reads were curated by eye for manual basecall corrections on AutoAssembler v1.4.0 (Perkin Elmer © 1995). For ZnF array structures, assembly was achieved using individual ZnFs and ZnF mini-motifs references (Appendix IV).

2.2.6 Ion Torrent sequencing

Equivalent amounts of gel purified PCR products containing PRDM9 A, C, L4 and L47 ZnF arrays were barcoded using Ion Xpress™ Library kit and barcodes according to for 200bp and 400bp sequencing format. Size selection was done on 1.6% LE agarose. Quantity and fragment sizes of gel purified PCR products were determined using Agilent 2100 Bioanalyzer with DNA 1000 chip. Sequencing on the Ion Personal Genome Machine® with Ion 316™ Chip Kit v2 (Thermo Fisher Scientific) was done using Ion PGM™ 200 Xpress™ Template Kit and Ion PGM™ 400 Xpress™ Template Kit chemistry according to manufacturer's instructions (Thermo Fisher Scientific). Ion Torrent Suite 5.0.2. was used in initial mapping to reference PRDM9 alleles to the GRCh37/hg19 (Feb 2009) reference genome which contained the PRDM9 B allele. BWA v0.7.10-r789 software was used to map the .bam files to the reference PRDM9 alleles confirmed via Sanger sequencing and also other known PRDM9 alleles to explore mapping detail. The data was examined on IGV v2.4.19.

2.2.7 Nanopore sequencing

Nanopore sequencing in 2015 with the with the MinIon Access Program MinION sequencer and R7 flow cell hardware was done with the Genomic004 kit (SQK-MAP004) library preparation chemistry according to manufacturer's instructions (Oxford Nanopore Technologies Limited). Basecalling was done on Metrichor and read mapping was done using an algorithm involving LAST sequence alignment software developed by Dr John H Wagstaff. Subsequently,

Burrows-Wheeler Aligner (BWA) (Li and Durbin, 2010) and minimap2 (Li, 2018) were also for mapping.

Nanopore sequencing in 2019 with the MinION Mk1B sequencer and R9.41.1 flow cell hardware was done with the Ligation Sequencing Kit (SQK-LSK109) library preparation chemistry according to manufacturer's instructions (Oxford Nanopore Technologies Limited), whilst omitting the optional 8-10kb fragmentation/size selection step. Basecalling was done on EPI2ME and minimap2 (Li, 2018) was used for mapping.

2.2.8 Illumina HiSeq data processing

The data processing pipelines and scripts for the Illumina dataset are detailed in Appendix V. Reads from bam files originally mapped to GRCh37/hg19 genome reference were extracted and remapped to 46 (45 on Genbank and 1 novel allele in a sample panel from the Illumina dataset) known reference PRDM9 alleles using BWA aligner with seed length increased from 19 to 100bp to effect virtually 100% stringency (zero mismatch). Read depth data was then obtained via VCF files for each sample-reference pair (sam/ref) and tabulated in Apple Numbers. This was the raw data used for analysis.

A 'IF minbymeans' filtering of read depth data based on DP_{min} / DP_{mean} values extracted from vcf files for each sample-reference measured lower limits of read depth across ZnF arrays and normalised the data across samples so that they could be compared. To make this filtering comparable and diagnostic for different alleles, the measure was fitted into thresholds levels 0.9, 0.8, 0.75, 0.7, 0.65, 0.6, 0.5 and 0.4. For example, if $min=112$ and $mean=140$, then $min/mean=0.8$. If $min/mean$ was equal to or more than 0.8, a score of '0' was returned for that sam/ref. For this sam/ref the score at 0.9 threshold level would be '1' as it was more intuitive to ignore 0s and inspect 1s. Since, there were no sample-references above a 0.9 threshold, this category was removed.

The remaining min/mean threshold levels were colour coded. Additionally sam/refs in the Sample column were coloured according to the highest DP_mean/DP_mean category for which they got a '0' score. Measures of central tendency was also obtained to get a general idea for read depth.

For DPmin where any sample-references returned a score of '0' were removed as they contained at least one base within the ZnF array for which no segment of any read had mapped. All of these instances showed that these 'gaps' were longer than a mere base (at least 6bp) as they corresponded to samples which do not contain a ZnF that is in the reference given. Hence, DPmin=0 sample-references were removed. Following this DPmin/mean-based identification of two PRDM9 alleles for each individual in the dataset was performed.

Variant site calling was also done for each sample-reference pair in the dataset and the results were tabulated with an additional step carried out to filter out variant sites such that only unique variant sites indicative of ZnF sequence identity remained. The results were evaluated to check whether these unique variant site could be used as signatures that confirmed the ZnF order in the ZnF arrays.

CHAPTER 3: CHILDHOOD ALL AND PRDM9 REGULATION OF TWO MHCII HOTSPOTS

3.1 Introduction

This study investigated the nature of PRDM9 regulation of two meiotic recombination hotspots in the MHCII region of chromosome 6. Batch sperm typing assays were used to determine the activating and non-activating PRDM9 alleles for the DNA3 hotspot and a closely located putative hotspot identified by Hinch et al. (2011) using a fine-scale SNP map derived from an African American population. The crossover activity data was used to examine a reported link of the DNA3 hotspot and a potential association for the so-called AA hotspot as having partial roles leading to the development of childhood B-ALL.

3.1.1 Childhood B-ALL susceptibility

Recent studies have been used as rationale for exploring a potential infectious, immune response aetiology for childhood B-ALL with the MHCII gene region as a potential locus associated with childhood ALL (Khor and Hibberd, 2012; Taylor et al., 2002; Taylor et al., 2009; Taylor et al., 2011; Urayama et al., 2013). For instance, the *HLA-DPB1* gene in the MHCII (Fig. 8) encodes a beta subunit of an alpha-beta heterodimer protein/peptide antigen receptor to CD4+ T helper cells and was found to contain alleles with weak associations for ALL susceptibility (Urayama et al., 2012). Following previous data linking B-ALL with an allele at the *HLA-DPB1*0201* locus, Taylor et al. (2002) typed *HLA-DPB1* alleles and found that children with B-ALL and T-ALL carried significantly more specific variants of the *HLA-DPB1* protein that changed the amino acids lining the antigen-binding site (OR 1.76 at 95% Confidence Interval (CI) of 1.20 to 2.56 for B-ALL, OR 1.93 at 95% CI of 1.01 to 3.80 for T-ALL when compared with healthy controls and OR 1.83 at 95% CI of

1.34 to 2.48 for B-ALL and OR 2.00 at 95% CI of 1.10 to 3.82 for T-ALL when compared with children with non-lymphoma solid tumours). Taylor et al. (2009) reiterated that the HLA-DBP1 holds primary association with B-ALL. Hence, this locus has been of interest in understanding the development of childhood B-ALL.

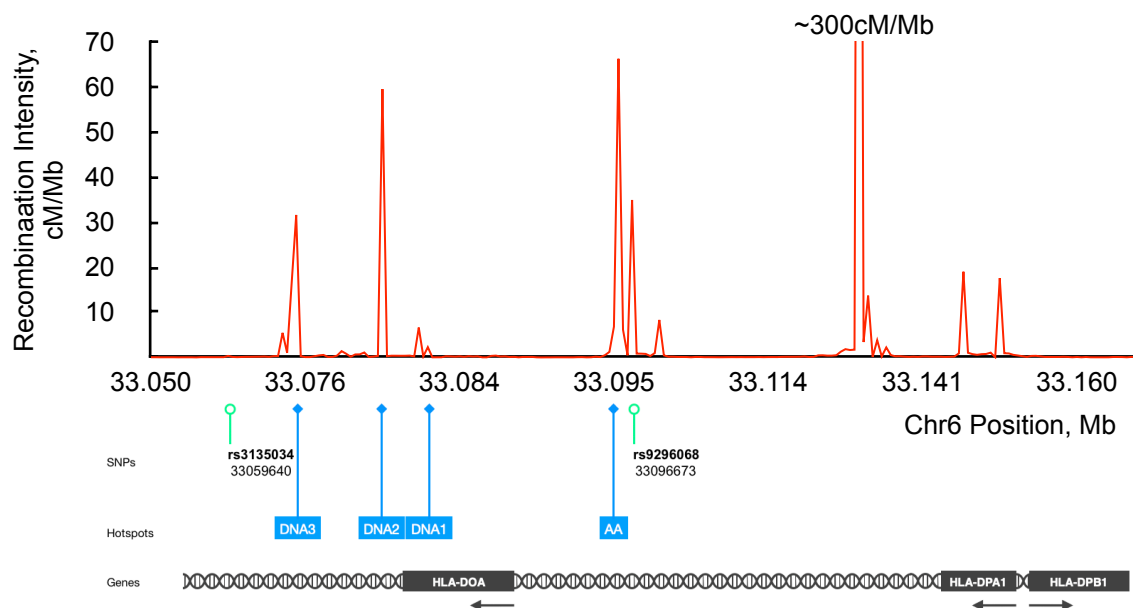


Fig. 8 Recombination profile of a 60kb section in the MHCII interval containing the DNA hotspot cluster DNA1, DNA2 and DNA3 hotspots and the putative African-enriched AA hotspot constructed using data from Hinch et al. (2011) on genome build is NCBI36/hg18 Mar2006. The location of two B-ALL-associated SNPs and study-relevant genes including the HLA-DOA gene are shown below the plot. SNP rs9296068 (Chr6:33096673, NCBI36/hg18) ~66kb telomeric of HLA-DPB1 (Chr6:33151738-33162954, NCBI36/hg18) and SNP rs3135034 (Chr6:33059640, NCBI36/hg18), ~92kb telomeric of HLA-DPB1 are marked with a green line with a looped end. Hotspots are marked in blue labels and lines with diamond ends. Approximate gene positions are marked in grey labels.

An extended MHC (xMHC) SNP-based analysis in a Californian population (Urayama et al., 2013) revealed SNP rs9296068, located ~66kb telomeric of HLA-DPB1, as being significantly associated with childhood ALL (OR=1.37, 95% CI=1.17-1.61, $P=1.2 \times 10^{-4}$ and OR 1.40 at 95% of CI 1.19-1.66, P corrected for multiple comparisons=0.036), especially in the ETV6-RUNX1+ and hyperdiploid cytogenetic subtypes. Additionally, a UK Genome Wide Association (GWAS) study revealed SNP rs3135034, located ~92kb telomeric of HLA-DPB1 within the MHCII as having a non-significant but notable association

with childhood ALL (Hosking et al., 2011). These two SNPs identified from the independent studies are in close proximity (~37kb apart) and they flank HLA-DOA locus. The HLA-DOA gene encodes an alpha subunit of an alpha-beta heterodimer protein/peptide antigen receptor in B-cells and Thompson et al. (2014) considered the possibility that previous associations with HLA-DPB1 may be the result of breakdown in LD between the HLA-DPB1 to HLA-DOA interval, and that in fact, HLA-DOA is the locus associated with childhood ALL as it is found in a region of patchwork breakdown in LD indicating recombination hotspots, as confirmed by Jeffreys, Kauppi and Neumann (2001), and flanked by these two SNPs.

3.1.2 DNA3 hotspot

A seminal study precisely co-locating breakdown in LD with meiotic sperm recombination hotspots in a 216-kb of the MHCII region in Chromosome 6 identified six hotspots including the DNA3 hotspot which exhibited a peak intensity of 140 cM/Mb in the North European male germline (Jeffreys, Kauppi and Neumann, 2001). The primary DNA sequence did not reveal the presence of an obvious 13bp consensus motif (Myers et al., 2008) associated with the PRDM9 A allele in the hotspot centre located at approximately Chr6:33073448 (NCBI36/hg18). No significant associations of these LD hotspots and aetiology of complex diseases were made until Thompson et al. (2014) reported elevated ancestral recombination rates in a British B-ALL cohort (228 times higher, $P=0.02$) compared to ethnically matched controls (117 x background) using the PHASE v2.1.1 software to estimate haplotype blocks with settings taking into account the presence of recombination hotspots (Stephens and Scheet, 2005; Stephens, Smith and Donnelly, 2001). This difference in recombination rates posed the question whether the DNA3 hotspot was being activated by a subset PRDM9 alleles and whether NAHR resulting in a genomic rearrangement might contribute to B-ALL disease aetiology.

B-ALL involves the B cell lymphoid lineage and the HLA-DPB1 and HLA-DOA loci have been associated with affected B-ALL children (Urayama et al., 2013). PRDM9 allele variation at recombination hotspots have been previously linked to instability causing genomic disorders (Berg et al., 2010). Genomic disorders are recurrent due to the underlying architecture of the DNA in the region. It is possible that there is an immune aetiology for B-ALL that can be explained by DNA3 hotspot activity and PRDM9 status-dependent abnormal genomic rearrangements affecting the HLA-DOA locus in the presence of minor allele of SNP rs9296068 (Urayama et al., 2013) and the relevant allele of SNP rs3135034 (Hosking et al., 2011). However, there has been no model for the type and location of the proposed rearrangement via NAHR misalignment. To investigate these links, it is important to study the effect of non-A or Ct variants of PRDM9 on the DNA3 hotspot. Furthermore, since the DNA3 hotspot was previously characterised using a panel of Europeans (Jeffreys, Kauppi and Neumann, 2001) carrying predominantly PRDM9 A alleles, the hotspot was presumed to be PRDM9 A-regulated, however, it is important to determine whether PRDM9 alleles more rarely found in Europeans have any influence over the regulation of the DNA3 hotspot.

3.1.3 An African-enriched hotspot

A concurrent investigation was initiated on a putative hotspot ~22kb centromeric of the DNA3 hotspot revealed by a highly detailed SNP-based genetic map of the MHCII region produced from an African American population with 80% African and 20% European ancestry (Hinch et al., 2011). This African American hotspot, hereon referred to as the AA hotspot, had a reported historical recombination intensity of 70cM/Mb. Comparisons with other recombination maps showed that this hotspot existed in the African YRI map but was not seen in the European DECODE and CEU maps, making it one of ~2450 African-only historical recombination peaks the Hinch et al. (2011) study. The AA hotspot also contains the 16bp motif (Kong et al., 2010) that PRDM9 Ct

alleles are predicted to bind. Additionally, the two B-ALL-associated SNPs (Hosking et al., 2011; Urayama et al., 2013) flank both the DNA3 and AA hotspots with the SNP rs9296068 (Urayama et al., 2013) located ~900bps from the approximate centre of the AA hotspot.

In order to inform the study of a possible link between hotspot activity within the MHCII region and B-ALL, it was decided that this AA hotspot needed to be more thoroughly characterised to develop a better profile of the recombination landscape in the MHCII region. This study therefore aimed to determine whether this historical AA hotspot was currently active and to confirm its PRDM9 regulation. Since hotspot activity would be measured using sperm crossover assays, the study would also distinguish paternal recombination rates from African American genetic map (Hinch et al., 2011) which is sex-averaged.

3.1.4 Study aims

The main aims of this study were to determine whether the putative AA hotspot is still active and confirm the nature of the PRDM9 regulation of the DNA3 hotspot. Batch sperm typing experiments using suitable individuals carrying A, Ct and other PRDM9 alleles were used to determine crossover activity over each of the DNA3 and AA hotspot intervals. The recombination frequencies were calculated according to Poisson distribution of rare recombinant molecules for each individual. It was anticipated that the results would provide a more detailed view of how PRDM9 status changes the recombination landscape in this MHCII region, which in turn would inform whether a crossover-related activity could be linked to B-ALL.

3.2 Results

3.2.1 SNP genotype survey of the DNA3 and AA hotspots

To identify heterozygous SNPs that could be exploited to selectively amplify recombinants over parental molecules, SNP genotyping was done for a University of Leicester panel of sperm donors consisting of 87 men, predominantly of sub-Saharan African origin (SSA panel), including 56 Zimbabweans, 17 Afro-Caribbeans, 10 Britons, 2 Indians, 1 African and 1 mixed British/Sri Lankan (Monckton et al., 1994; Kauppi, May and Jeffreys, 2009). The PRDM9 diversity of the SSA panel was represented by 26 alleles out of the 50 unique PRDM9 alleles published on GenBank genetic sequence database (Benson et al., 2008) and an additional 3 alleles characterised and reported by Berg et al. (2011) but not included in GenBank. In the SSA panel, 0.48 of the ZnF arrays were of the A type and 0.29 were classified as Ct variants (Berg et al., 2010).

To conserve the amount to be used from this finite source of sperm DNAs, a multiplex PCR strategy was developed to simultaneously amplify four overlapping regions within ~15kb intervals around both the DNA3 and AA hotspots (Fig. 9). Individual amplicons were then recovered using secondary nested and hemi-nested PCRs using the primary PCR product as a template. An additional amplicon covering the central AA hotspot region was generated using hemi-nested PCRs. The secondary PCR amplicons were denatured and dotblotted on nylon membranes (Fig. 10a). SNP genotyping was carried out by hybridisation of γ -³²P-labelled ASOs designed for each allele of targeted SNPs (Fig. 10b) and visualisation on X-ray films (Fig. 10c).

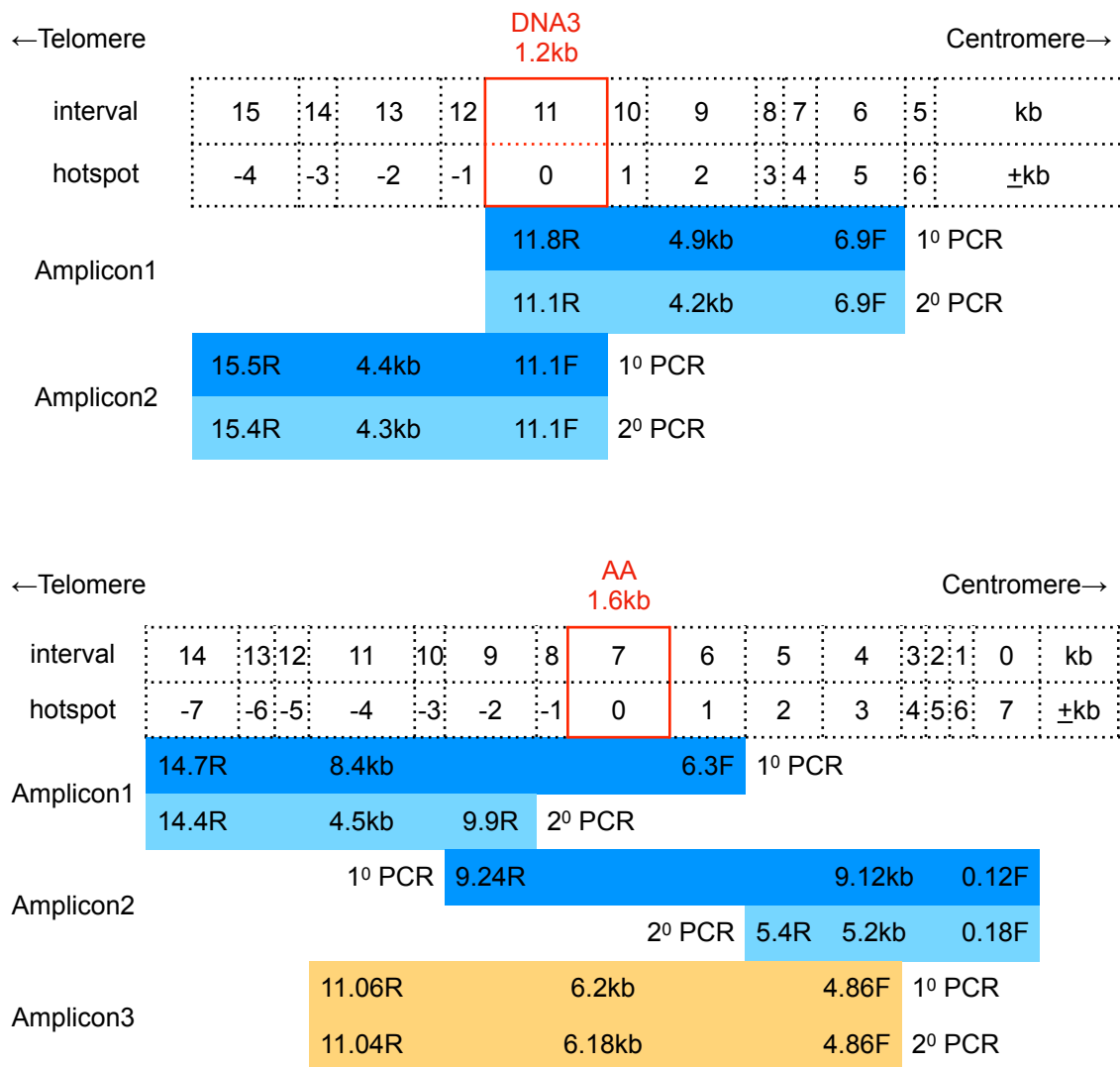


Fig. 9 Multiplex PCR strategy used to simultaneously amplify four overlapping regions (shown in aqua blue) in 15kb intervals around both hotspots. Nested and hemi-nested secondary PCRs were used to isolate two amplicons for DNA3 hotspot and two amplicons for AA hotspot (shown in sky blue). An additional amplicon covering the central AA hotspot region was generated using hemi-nested PCRs (shown in cantaloupe orange). The estimated hotspot centres are shown in red. Distances relative to the hotspot centre are given in kilobases (kb). The approximated hotspot intervals indicated in kb are derived from Jeffreys et al. (2001) for the DNA3 hotspot and Hinch et al. (2011) for the AA hotspot.

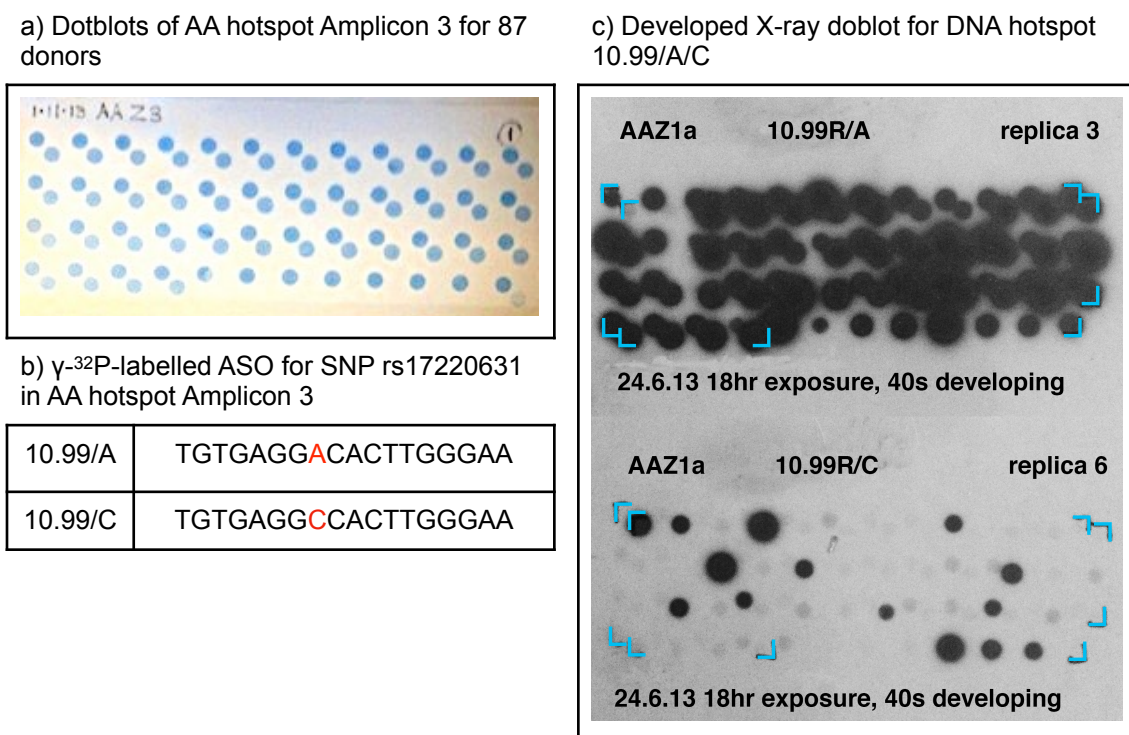


Fig. 10 SNP Genotyping in the DNA3 and AA hotspot intervals a) PCR products were denatured and dotblotted on nylon membranes. b) As shown by the example SNP rs17220631 A/C (local coordinate 10.99 A/C), ASOs were designed with the polymorphism typically occurring between the 8th and 12th position of a 18nt oligonucleotide. T4 polynucleotide kinase was used to catalyse the transfer and exchange of γ -³²P from the γ position of ATP to the 5'-hydroxyl terminus of the ASOs. In a 3M TMAC-based hybridisation solution, the PCR products on the dotblots were incubated with non-labelled competitor ASOs carrying alternative alleles for the SNPs and then hybridised with the γ -³²P labelled ASO probes to enforce specific binding of the correct version of the ASO to the PCR product. Hybridisation was repeated for the alternative competitor and labelled probes. c) The dotblot signals were visualising by exposing to a X-ray film sandwiched between dotblot and a phosphor intensifier screen stored at -80°C for 4-20hrs. The genotypes were scored for each donor where a signal for one or the other allele indicated that the individual was homozygous and signals for both alleles indicated that the individual was heterozygous for the SNP.

For the DNA3 hotspot, 4 SNPs previously used as selector sites for CO analysis were genotyped for the SSA panel of donors (Table 4; Jeffreys, Kauppi and Neumann, 2001) in addition to 36 internal SNPs (Appendix II). The Minor Allele Frequency (MAF) range of 40 SNPs typed were between 0.01-0.48 with a median of 0.21. For the AA hotspot, 14 SNPs found at the predicted central hotspot region and with a MAF of more than 0.05 for African populations as indicated by the major African HapMap data sets (YRI, ASW, LWK, MKK) were genotyped so that they could be used in mapping CO events and therefore enhance the resolution of the hotspot during characterisation (Appendix II). The MAF range of 26 SNPs typed were between 0.01-0.45 with a median of 0.10.

Additionally, 12 SNPs located approximately 4kb away from the hotspot centre in either direction (Table 4) were genotyped for use as potential selector sites.

Table 4 - List of SNPs used as selector sites for CO analysis at DNA3 and AA hotspots

Hotspot	#	SNP	Position	MAF, %
DNA3	1	rs34349704 +/-	33075971	0.33
	2	rs12190787 C/G	33075892	0.33
	3	rs172275 C/T	33069599	0.48
	4	rs206767 G/T	33070398	0.47
AA	1	rs6930608 C/T	33098789	0.1
	2	rs6930399 C/T	33098653	0.1
	3	rs6936620 C/T	33092429	0.24
	4	rs9276994 A/G	33092233	0.45
	5	rs17220631 A/C	33092209	0.09

Selector sites previously used for DNA3 hotspot with a European sperm donor panel (Jeffreys, Kauppi and Neumann, 2001) were genotyped for the current panel of donors. For the AA hotspot, SNPs 4kb away from the estimated hotspot centre were genotyped for the current panel of donors. MAF was determined for each SNP. SNPs ultimately used as selector sites are shown here. ASPs for the SNPs can be found in Appendix II. A detailed list for all SNPs genotyped for DNA3 and AA hotspot can be found in Appendix II. Chromosome 6 positions of SNPs are based on NCBI36/hg18 Mar2006 genome assembly.

3.2.2 Optimisation of ASPs

For the DNA3 hotspot, Allele-Specific Primers (ASPs) (Jeffreys, Kauppi and Neumann, 2001) were re-evaluated for the current study to determine the optimum primer annealing temperatures for each version. Individuals homozygous for either allele of a SNP were used to determine allele-specificity and yield at three or four annealing temperatures. For the AA hotspot, although some SNPs genotyped for use as potential selector sites in CO analysis had MAF range of 0.09-0.45, these SNPs were deemed suitable for a pilot study to determine whether the hotspot is still active, and ASPs were designed for the SNPs listed in Table 4 (Appendix I).

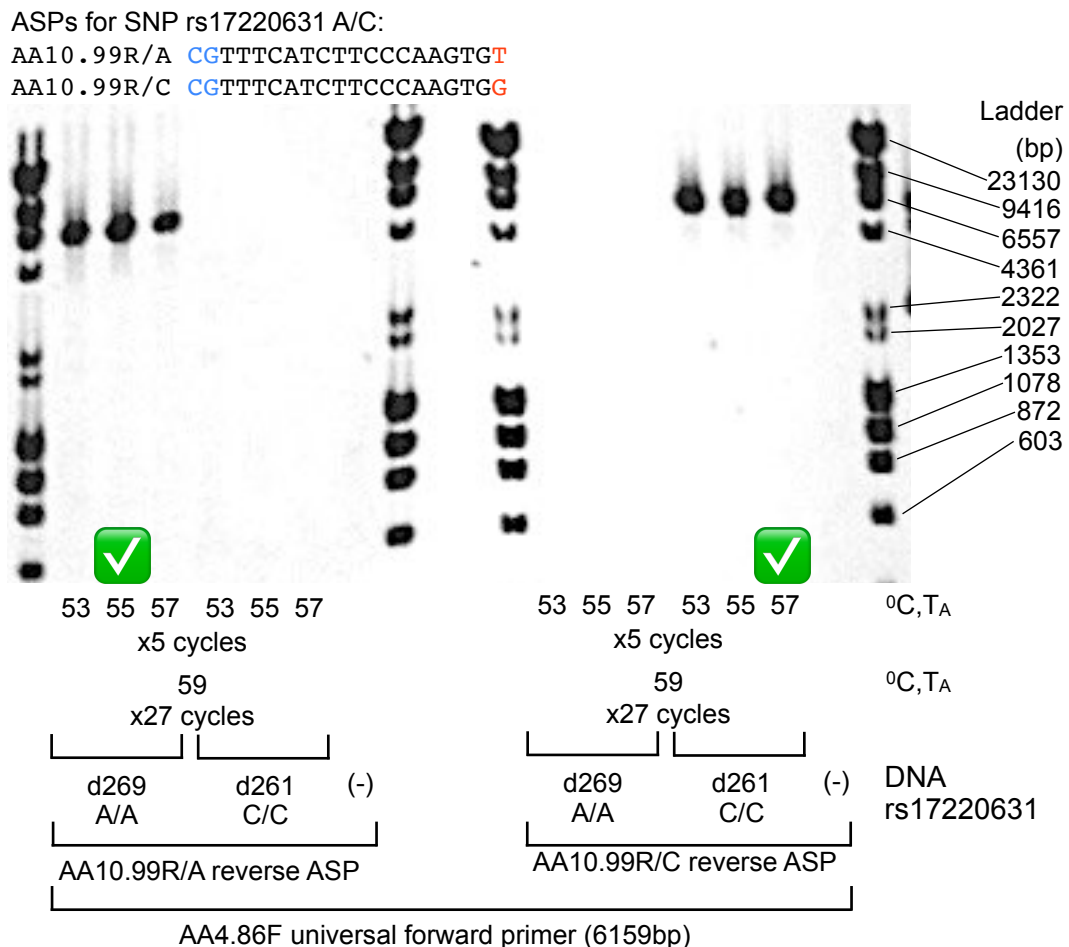


Fig. 11 Optimisation of ASPs. Individuals homozygous for either allele of a SNP were used to carry out Annealing Temperature (T_A in Degrees Celsius, °C) titrations covering the expected T_m range. SNP positions on ASPs are indicated in red. Genotypes of individuals are indicated below donor name. ASPs AA10.99R/A and 10.99R/C (for rs17220631) which have CG modifications at the 5' end (indicated in blue). A two-step thermal cycling PCR program was used to allow the complementary sequence of the ASP, which has lower T_m, to sufficiently anneal and amplify the target sequence. The generation of the amplicons with terminal sequences fully complementary to the modified ASPs then allowed the modified ASPs to fully anneal to the new amplicons at higher annealing temperatures and generate more amplicons. Optimum temperatures for the ASPs are indicated by green arrows.

Several ASPs had to be designed with 5' end modifications to the increase GC content. To allow these modified ASPs to anneal sufficiently to the target sequence, a two-step PCR cycling program was employed; 5 cycles were carried out a lower annealing temperature, to account for the lower melting temperature (T_m) of the complementary sequence of the ASP followed by further cycles at the higher annealing temperatures more suitable for the whole modified ASP to anneal to the target sequence (Fig. 11).

Low yield posed a problem for some of these modified ASPs and this was resolved by selecting the lowest annealing temperature that would give

specificity for a particular version of ASP, in the second step of thermal cycling. In some cases, ASPs which were completely complementary to the target sequence still showed low specificity (Fig. 12). This was especially pronounced in T allele versions of ASPs. To increase specificity, the optimisation was performed again with the first 5 cycles carried out at a higher annealing temperature followed by further cycles at a lower annealing temperature more suitable for sufficient annealing of the ASP. Since the DNA3 hotspot region has numerous AT-rich regions, lower extension temperatures (62°C as opposed to 65°C) were necessary for efficient amplification of the target sequence.

The same input of 4-5ng of DNA was used for all reactions and where there were quality differences between donors, the total number of cycles in the second step of thermal cycling was increased to increase the yield. Since the amplicon size was ~6kb long, this still led to some differences in the amount of product generated. However, the goal was to achieve specificity for the ASPs and it was accepted that visual identification of a recombinant product from selective amplification was sufficient even if the yield was low.

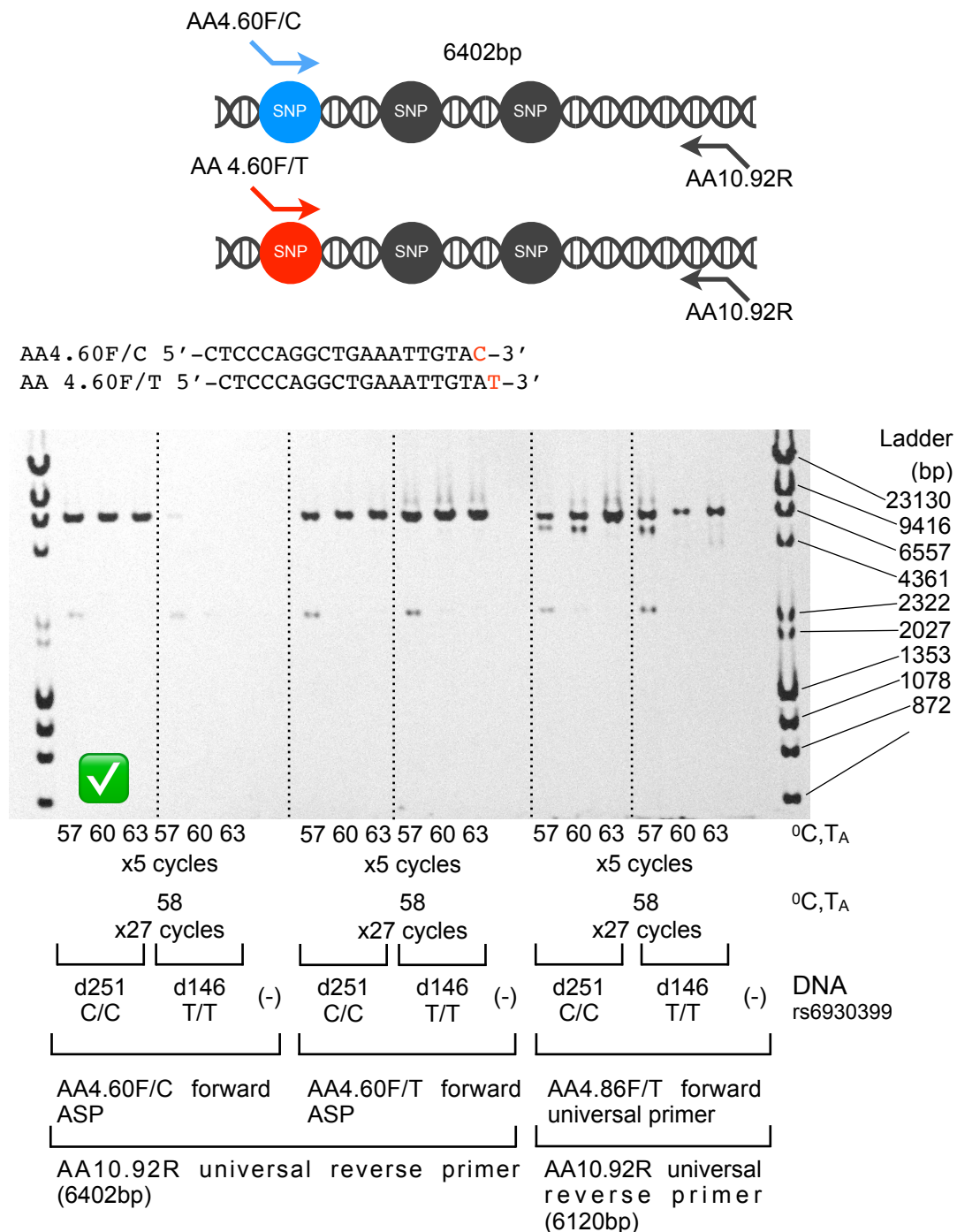


Fig. 12 Non-specific amplification of target sequence by both versions of ASPs for SNP rs6930399 C/T. Individuals homozygous for either allele of a SNP were used to carry out annealing temperature titrations covering the expected T_m range. SNP positions on ASPs are indicated in red. Genotypes of individuals are indicated below donor name. Annealing temperature titration of ASPs AA10.99R/A and 10.99R/C (for rs17220631) which have CG modifications at the 5' end (indicated in blue). A two-step thermal cycling PCR program was used to allow the complementary sequence of the ASP, which has lower T_m , to sufficiently anneal and amplify the target sequence. The generation of the shorter amplicons with terminal sequences fully complementary to the modified ASPs then allowed the modified ASPs to fully anneal to the new amplicons at higher annealing temperatures and generate more amplicons. Optimum temperatures for the ASPs are indicated by green arrows. For ASP AA 4.60F/T (rs6930399 C/T) non-specific amplification was observed for donor d251 in all tested annealing temperatures even though this individual carried the C allele for the SNP marker. Additionally, yield was comparably lower than that for d146 who carried the target T allele. Using the same ASP design, the most appropriate solution was to use a two-step thermal cycling with 5 cycles at higher annealing temperature for AA 4.60F/T followed by a lower annealing temperature more suitable for efficient annealing of modified reverse ASPs. Optimum annealing temperature for AA 4.60F/C is indicated by the green arrow. The AA4.86F/T and AA10.92R primers were used as a template control to ensure that this ~6kb amplicon was being amplified.

3.2.3 Selection of donors suitable for CO analysis

Owing to the inherent design of sperm CO analysis, men (individuals/donors) heterozygous for the SNPs used as selector sites had to be identified from SNP genotyping data (Jeffreys et al., 2004). For the DNA3 hotspot, there were eighteen men from the SSA panel of donors that were suitable for CO analysis. Due to quantity limitations, twelve of these candidate men (Table 5) were used in subsequent CO analysis.

Table 5 - PRDM9 status of individuals heterozygous for the SNPs used as selector sites in CO assays at DNA3 and AA hotspots

Hotspot	Men	Origin	PRDM9 Alleles		Forward Primary Selector Site	Reverse Primary Selector Site	Forward Secondary Selector Site	Reverse Secondary Selector Site
DNA3	d177 (3)	Afro-Caribbean	A	A	A6F +/- rs34349 704	BC4R C/T rs17227 5	A7F C/G rs1219078 7	B7R G/T rs206767
	d278* (6)	Zimbabwean	A	A				
	d288 (15)	Zimbabwean	A	B				
	d290 (14)	Zimbabwean	A	B				
	d244 (11)	Zimbabwean	L22	A				
	d249 (12)	Zimbabwean	C	A				
	d260 (1)	Zimbabwean	A	L11				
	d277 (2)	Zimbabwean	A	L5				
	d135 (13)	British	D	A				
	d185 (17)	Afro-Caribbean	C	L12				
	d236 (18)	Zimbabwean	L4	L22				
	d264 (16)	Zimbabwean	L6	L14				
	d6 (8)	European	A	A				

Hotspot	Men	Origin	PRDM9 Alleles		Forward Primary Selector Site	Reverse Primary Selector Site	Forward Secondary Selector Site	Reverse Secondary Selector Site
DNA3 from (Jeffreys, Kauppi and Neumann (2001))	d50 (7)	European	A	A				
	d22 (9)	European	A	A				
	d38 (10)	European	A	B				
	d45 (4)	European	A	A				
	d17 (5)	European	A	L20				
AA	d257 (20)	Zimbabwean	L14	L16	AA 4.47F C/T rs6930608	AA 10.99R A/C rs17220631	AA 4.60F C/T rs6930399	AA 10.96R A/G rs9276994
	d285 (21)	Zimbabwean	C	A				
	d262 (19)	Zimbabwean	L19	C		AA 10.96R A/G rs9276994		AA 10.77R C/T rs6936620
	d176 (22)	Afro-Caribbean	A	A				
	d278 (6)*	Zimbabwean	A	A				
	d264 (16)*	Zimbabwean	L6	L14	AA 4.31F C/T rs3097648	AA 10.96R A/G rs9276994		

Individuals previously tested (Jeffreys, Kauppi and Neumann, 2001) are also listed as CO activity data at DNA3 hotspot of these individuals were used in analysis of new data generated in the current study. PRDM9 A alleles are highlighted in blue and PRDM9 Ct alleles are highlighted in red. The 'd' numbers can be used to trace to the identity of the men and adjacent bracketed numbers assist to link each individual to figures, figure legends and elsewhere in the main text of this work. (*) Individuals d278(6) and d264(16) were used in CO assays for both hotspots.

For the AA hotspot, individuals/men d257 (20) and d285 (21) were heterozygous for SNPs rs6930608, rs6930399, rs9276994 and rs17220631, and individuals d262 (19), d176 (22) and d278 (6) were heterozygous for SNPs rs6930608, rs6930399, rs6936620 and rs9276994 (Table 5) and d264 (16) was heterozygous for SNPs rs3097648, rs6930399, rs9276994 and rs6936620. ASPs were designed and optimised for these selector sites as previously described.

The PRDM9 status of the individuals amenable to CO analysis is a prerequisite for determining PRDM9 regulation of the DNA3 and AA hotspots. For the DNA3 hotspot, the selection of individuals heterozygous for the selector sites included two A/A, seven A/N and three N/N individuals, with N denoting non-A alleles. For the AA hotspot, the pilot selection included two Ct/Ct, one Ct/N and two N/N individuals, with N denoting non-Ct alleles. Hence, successful CO analysis would be able to determine whether the DNA3 hotspot was PRDM9 A-regulated and the AA hotspot was PRDM9 Ct-regulated as predicted.

3.2.4 Linkage phasing and sperm CO assay design

	10.96R	10.77R		4.60F	4.47F
	A	C	hotspot	C	C
	G	T	centre	T	T
				4.60F	4.47F
1 ^o PCR	10.88R	C	6.41kb	C	C
	10.88R	T		T	T
2 ^o PCR	10.88R	C	6.36kb	C	4.52F
	10.88R	T		T	4.52F
		ASO		ASO	
	10.96R	10.77R		4.60F	
1 ^o PCR	A	C	6.44kb	C	4.52F
	G	T		T	4.52F
2 ^o PCR	10.88R	C	6.36kb	C	4.52F
	10.88R	T		T	4.52F
		ASO		ASO	

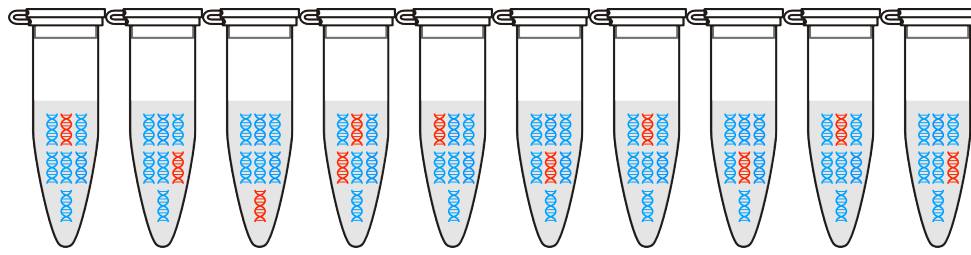
Fig. 13 Linkage phasing of individuals heterozygous for forward primary selector site rs6930608 (AA 4.47F C/T), forward secondary selector site rs6930399 (AA 4.60F C/T), reverse primary selector site rs9276994 (AA 10.96R A/G) and reverse secondary selector site rs6936620 (AA 10.77R C/T) for use in CO analysis at the AA hotspot. AA 4.52F universal forward primer and AA 10.88R universal reverse primer were used with ASPs in the primary PCR to selectively amplify haplotypes. Secondary PCRs were used to generate sufficient PCR product for dotblotting. Internal selector sites, highlighted in red, were genotyped using γ -³²P-labelled ASOs for both alleles for these sites.

For each individual selected for CO analysis, the alleles of the selector site SNPs were phased to determine alleles located on the same haplotype. Further conservation of sperm DNA was achieved by using ASPs to selectively amplify haplotypes and then genotyping the other selector sites using hybridisation with

γ -³²P-labelled ASOs (Fig. 13). The alternative method would have been to use all four selector site ASPs in different combinations to determine linkage phasing but this would have required more sperm DNA.

Multiple pools of sperm DNA, some of which are expected to contain rare recombinant molecules, were used in CO PCR assays typically with 250, 500, 1000, 2000 molecules per pool and 10 reactions per pool size (Fig. 14a; Kauppi, May and Jeffreys, 2009). Using these pool sizes and number of reactions, the assays could screen 37,250 amplifiable molecules per individual. From the linkage phasing, it was possible to design a CO PCR assay to selectively amplify recombinant molecules using two rounds of repulsion phase allele-specific PCR using combinations of primary and secondary ASPs which would prevent amplification of false positives resulting from 'bleed-through' of parental haplotypes (Fig. 14b). Reciprocal CO assays were performed on individuals d6 (8) and d50 (7) (Table 5) to determine which orientations/ASP combinations provided the most specific reactions.

a Pools of sperm DNA containing rare recombinants (illustration)



b Sperm CO assay (schematic)

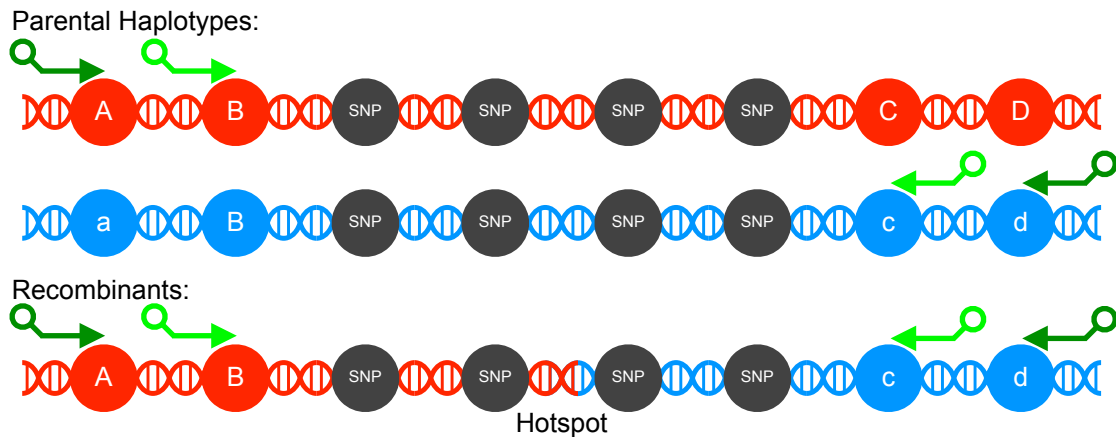
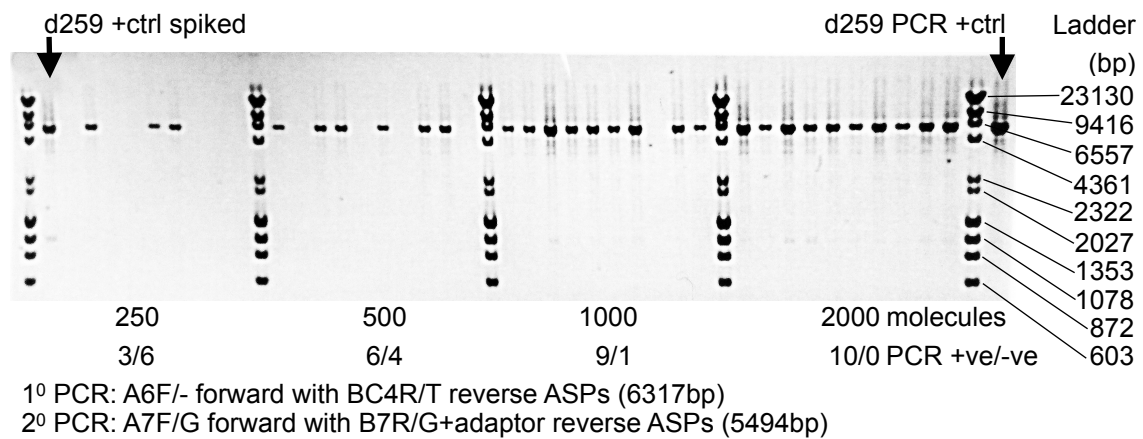


Fig. 14 CO assay design. Figures not to scale. a) Multiple pools of sperm DNA, typical pool sizes of 250, 500, 1000, 2000 molecules (10 reactions per pool size). Rare recombinant molecules in red. b) CO PCR assay design to selectively amplified recombinant molecules using two rounds of repulsion phase allele-specific PCR. Primary ASPs are shown in dark green and secondary ASPs are shown in light green. The repulsion phase ASPs allow only recombinants (containing A, B, c and d alleles) to be amplified while the parental haplotypes (A, B, C, D and a, b, c, d) remain unamplified.

3.2.5 DNA3 hotspot is activated by PRDM9 A

Sperm CO analysis was carried out over a 5.46kb interval spanning the DNA3 hotspot centre for 12 individuals from the SSA panel as detailed in Table 5. Fig. 15 shows the data for d177 (3), a PRDM9 A homozygous donor.



RF=0.20% at DNA3 hotspot
Individual d177 (3), PRDM9 A/A

Fig. 15 DNA3 hotspot CO PCR assay products for individual d177 (3) run on a 0.8% agarose gel with 0.5µg/ml Ethidium Bromide and visualized by UV transillumination. ASP combinations used are indicated below the gel photograph. Individual d259 who carried a parental haplotype identical to the recombinants targeted by the ASPs was used as a positive control. PCR-positive and PCR-negative reactions were counted and analysed using the Poisson correction program. Individual d177 (3) carried two PRDM9 A alleles and showed a recombination frequency of 0.20%. See Appendix II for PCR design.

From the number of recombinants recovered per pool size, the CO frequency (recombination frequency, RF %) was estimated according to the Poisson distribution for each of the twelve individuals. Additional data from 6 men of North European descent previously assayed at the DNA3 hotspot (Jeffreys, Kauppi and Neumann, 2001) were collated with the new data. CO analysis showed that the DNA3 hotspot, which lacks a central 13bp motif, was nevertheless activated in the fifteen men carrying at least one PRDM9 A allele with a mean CO frequency of 0.13% +0.056 s.d. More importantly, recombination was significantly suppressed in the three non-A (including Ct allele) carriers (Mann-Whitney U , $U_s=45$, $P<0.005$) with a mean CO frequency of 0.01%.

There was considerable variation in the activity of PRDM9 A carriers, with 5-23-fold higher RFs compared to non-A carriers. To ensure that there were no systemic differences between the new and pre-existing datasets, RF from the new data set that includes individuals 1-5, 10 and 13-15 was compared with the data set that includes individuals 6-9, 11 and 12 (Jeffreys, Kauppi and Neumann,

2001). Although there was less variation among individuals in the dataset from Jeffreys, Kauppi and Neumann (2001), there was no obvious clustering away from the range of frequencies seen in the new data set (Fig. 16). Mean CO frequencies of nine A/N individuals in the new data set and six A/N individuals from the Jeffreys, Kauppi and Neumann (2001) dataset were 0.13 ± 0.07 s.d. and 0.13 ± 0.03 s.d., respectively. To ascertain that the two data sets were indeed comparable, a Mann-Whitney U test was performed and the variation between the datasets proved to be insignificant (Mann-Whitney U, $U_s=32$, $P>0.10$) (Appendix III). To assess the variation in frequencies within the combined A/A and A/N group, a contingency table was drawn up for the number of COs against the number of molecules screened for each individual (Appendix III). A chi-squared test proved that there is significant heterogeneity between the men carrying at least one PRDM9 A allele (15 d.f., $X^2=331.81$, $P<0.001$). The pre-existing dataset involved screening of up to 164180 molecules whereas the new data was generated from the screening of 37,250 molecules per individual. The variation within the combined A/A and A/N group may simply be due to the lesser number of molecules screened per individual in the new study.

Subsequently, ASP combinations used in the CO assays were compared between the fifteen A/N individuals tested to determine whether the variation was due to a technical effect of the CO assay. Individuals 2, 3, 5 and 10 (Fig. 16) had the same linkage phasing and the same ASP combination of A6F/+ with BC4R/T in 1^oPCR and A7F/C with B7R/G in 2^oPCR were used in their CO assays. It could be speculated that the frequencies of these individuals form a cluster although individual 10, with a mid-range frequency, is an outlier. Reciprocal CO assays carried out for individual 8 during the current study to determine best orientations for CO analysis showed that while reciprocal assays gave similar recombination frequencies for either orientation, this ASP combination was more specific and easier to score CO-positive PCRs from the gel. Similarly, the remaining five individuals (1, 4, 13, 14 and 15) had the same phasing and an

ASP combination of A6F/- with BC4R/T in 1^oPCR and A7F/G with B7R/G in 2^oPCR were used to recover recombinant molecules. In this case, reciprocal crossover assays for individual 9 who had the same phasing showed that this ASP combination was more specific and easier to score. ICO frequencies in individuals 1, 4, 13, 14 and 15 cover the range of frequencies seen in the A/N group. In summary, ASP combinations could not have contributed to the observed variation.

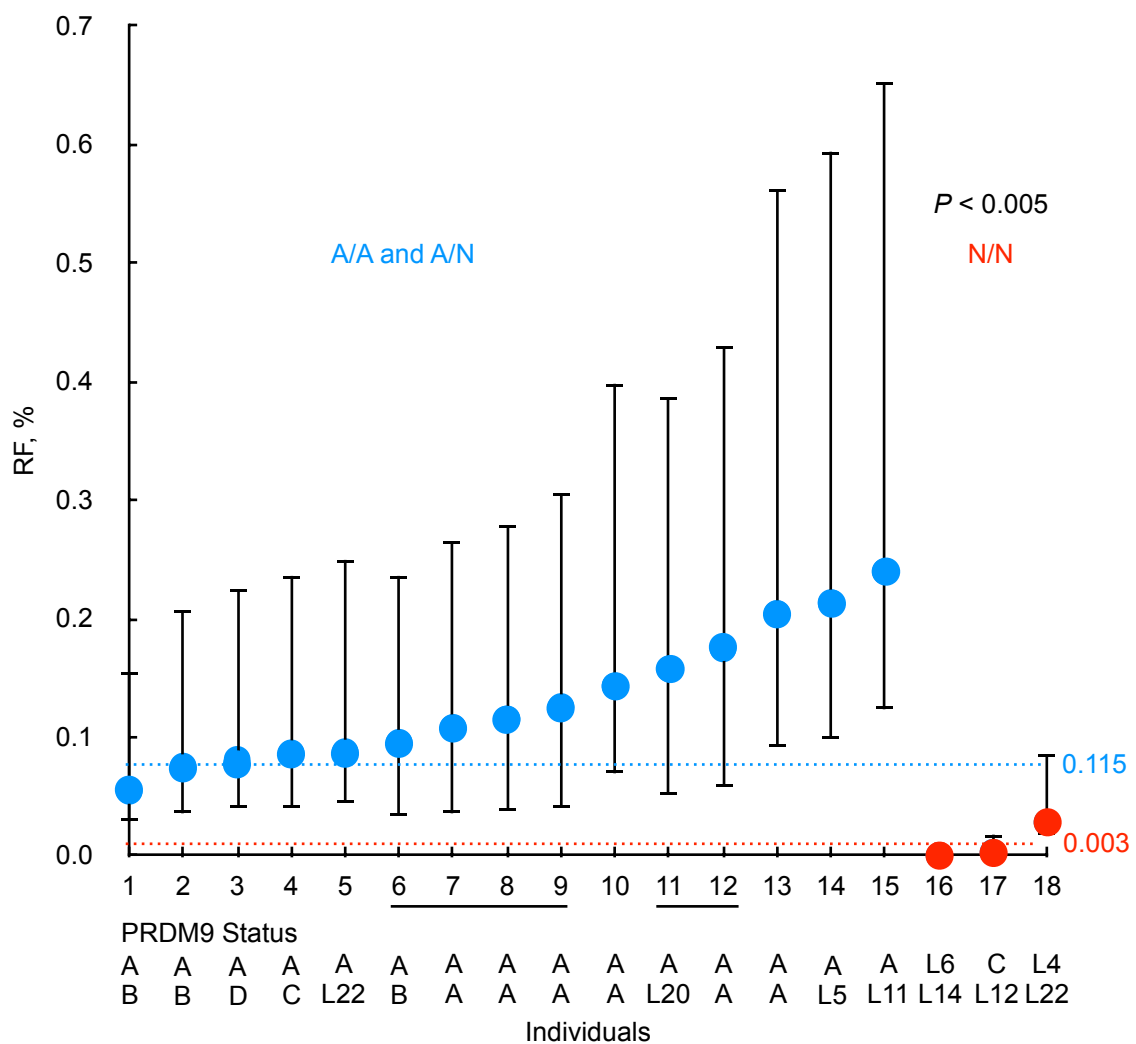
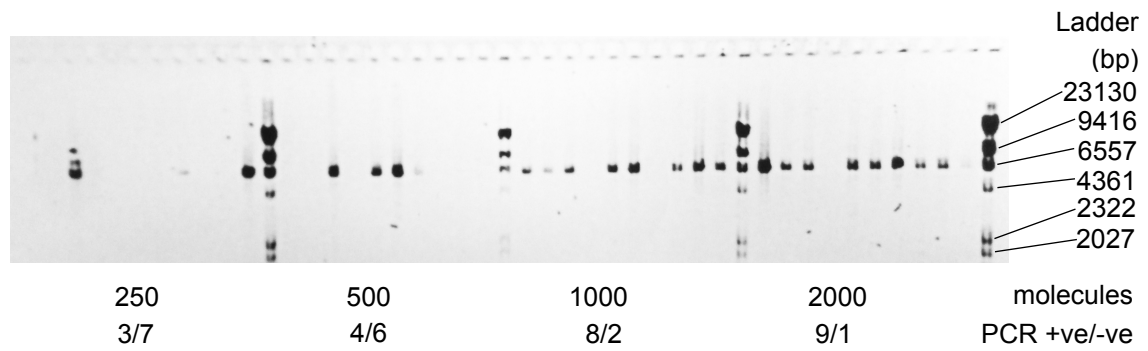


Fig. 16 Variation of sperm CO activity between men at the DNA3 hotspot. Men carrying at least one PRDM9 A (A/A and A/N) are shown in blue and those with two non-A alleles (N/N) are shown in red. Recombination frequency data from Individuals 6-9, 11 and 12 (underlined) were obtained from Jeffreys, Kauppi and Neumann (2001). Estimated recombination frequencies and 95% confidence intervals, after Poisson correction, are shown. Median recombination frequencies for the two groups groups are indicated by the dotted lines. Mann-Whitney U test results for the significance of differences between the two groups are displayed a the top right.

Of the men tested, two men with PRDM9 A/A genotype (individuals 10 and 13, Fig. 16) and two men with PRDM9 A/B genotype (individuals 1 and 2, Fig. 9) had lower recombination frequencies, 0.14 and 0.2%, and 0.05 and 0.08% respectively, compared to two men carrying A/L11 (0.24%) and A/L5 (0.21%) genotypes. Together with the significance of the variation seen in the A/N group, it was possible that the variation in the A/N group could be due to cis-effects. Internal SNPs (between the selector sites and encompassing the hotspot centre) were examined to look for any cis-acting elements that may affect binding and hotspot activation by PRDM9 A alleles. Individuals were placed into three groups; group 1 consisted of two individuals carrying A/L5 and A/L11, group 2 of two individuals carrying A/A alleles and two individuals carrying A/B alleles, and group 3 consisted of the three remaining A/N individuals. No clear differences were observed in these three groups.

3.2.6 AA hotspot is still active but weak

Sperm CO analysis was carried out over a 6.58kb interval for individuals d257 (20) (Fig. 17) and d285 (21), over a 6.56kb interval for individuals d262 (19), d176 (22) and d278 (6) and over a 6.72kb interval for individual d264 (18to16) spanning the predicted AA hotspot according to the ASPs used (Appendix II). The data from the six men (Fig. 18) showed that the AA hotspot is still active but weak with the highest CO frequency of 0.13% observed in individual d257 (20) (Fig. 17). There is obvious clustering of RF of Ct/Ct individuals compared to Ct/A and A/A men and the difference between these two groups is significant (Mann–Whitney *U* test, $P < 0.005$) (Appendix III).



RF=0.13% at AA hotspot

Individual d257 (20), PRDM9 L14/L16

1^o PCR: 4.47F/C forward with 10.99R/A reverse ASPs (6616bp)

2^o PCR: 4.60F/C forward with 10.96R/G reverse ASPs (6420bp)

Fig. 17 AA hotspot CO PCR assay products for individual d257 (20) run on a 0.8% agarose gel with 0.5µg/ml Ethidium Bromide and visualized by UV transillumination. ASP combinations used are indicated below the gel photograph. See Appendix I for details on ASP combinations and PCR conditions used. PCR-positive and PCR-negative reactions were counted and analysed using the Poisson correction program. Individual d257 (20) carried PRDM9 L14 and L16 alleles and showed a recombination frequency of 0.13%.

In terms of PRDM9 regulation, the results suggest that the hotspot is PRDM9 Ct-regulated, even though the predicted 16bp motif could not be found at its centre which was estimated to be at Chr6:33095891-33097368 (NCBI36/hg18) (Hinch et al., 2011). Individuals d176 (22) and d278 (6) who both carry non-Ct alleles, specifically PRDM9 A homozygotes, had 10-to-43-fold lower recombination frequencies compared to individuals d262 (19) and d257 (20) both of which carry two Ct alleles. However, individual d262 (19) had 4-fold lower frequency than individual d257 (20). Individual d285 (21) who carries one Ct allele had 22-fold lower recombination frequency compared to individual d257 and 5-fold lower frequency compared to individual d262 (19). CO assays in more individuals carrying Ct and non-Ct alleles will be required for a more complete analysis of PRDM9 regulation at the AA hotspot.

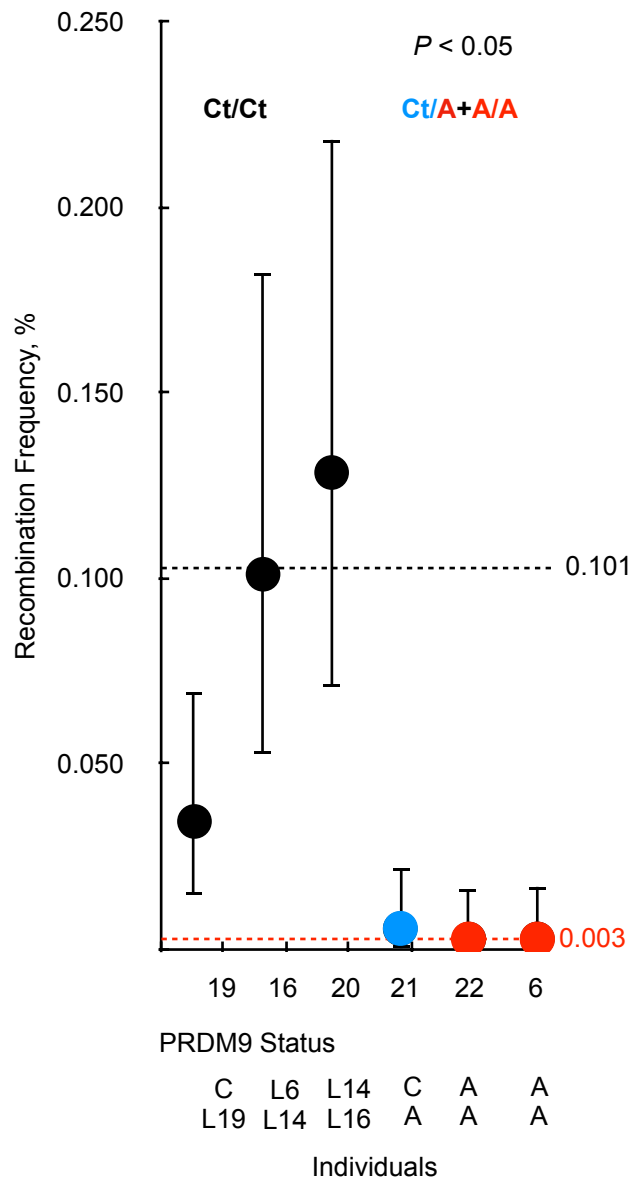


Fig. 18 Variation of sperm CO activity between men at AA hotspot. Men carrying two PRDM9 Ct alleles (Ct/Ct, shown in black) at least one PRDM9 Ct alleles (Ct/A, shown in blue) and two non-Ct alleles (A/A, shown in red) are grouped and displayed in ascending order of activity. Estimated recombination frequency and 95% confidence intervals, after Poisson correction, are shown. Median recombination frequencies for Ct/Ct and A/A groups are indicated by the dotted lines. Mann-Whitney *U* test results for the significance of differences between the Ct/Ct and Ct/A+A/A (Ct/A and A/A) groups is displayed.

When Ct/Ct and Ct/A individuals are grouped against A/A individuals, there is no significant difference in activity (Mann-Whitney *U* test, $P > 0.10$). The number of individuals tested may not be sufficient to provide statistical power. It has also been demonstrated that Ct/Ct and Ct/A groups have considerable overlap in range of activity as seen with the 5A, 12D and other Ct-regulated

hotspots (Berg et al., 2011). It is even possible that the C allele may be non-activating for this hotspot as observed for hotspot 12B which appears only to be activated by L6 and L8.

3.3 Discussion

3.3.1 PRDM9 A binding at DNA3 hotspot

Determination of PRDM9 A regulation at DNA3 hotspot in the absence of a 13bp motif (Myers et al., 2008) in the hotspot centre, brings into question whether the binding of PRDM9 A to this motif is the determining factor for subsequent chromatin remodelling and recombination at the DNA3 hotspot. Five other hotspots, also not containing a central motif, have been shown to be active only in PRDM9 A allele carrying individuals (Berg et al., 2010) and this points to a more complex PRDM9-DNA interaction than our current theoretical predictions. Patel (2016) noted the impact of the ZnF mini-motifs towards the 3' end of the array along with the non-interaction of the terminal J-ZnF. The FGHF mini-motif in particular bound exclusively to the major groove.

In the DNA3 hotspot interval analysed in this study, three instances of the 13bp motif were found with the nearest occurring 1.6 kb centromeric of the hotspot centre. The DNA3 hotspot is 1.2kb wide and hotspot initiation sites have been shown to be highly localised zones at the centres of the hotspots (Jeffreys, Kauppi and Neumann, 2001; May et al., 2002; Jeffreys and May, 2004; Jeffreys et al., 2004). Since PRDM9 alleles are predicted to bind to specific DNA sequences and its H3K4me3 methyltransferase capability, which presumably opens up the chromatin to allow Spo11 to induce DSBs, is highly local (Baudat et al., 2010; Hayashi, Yoshida and Matsui, 2005; (Brick et al., 2012), the more logical model is that PRDM9 binding, H3K4me3 and DSBs sites are co-localised. So PRDM9 binding several hundred bases away seems unlikely to be associated with the DNA3 or any other studied hotspot.

PRDM9 may be binding to a sequence at the centre of the DNA3 hotspot that does not fit the 13bp consensus motif. Our understanding of ZnF-DNA binding has been largely based on single ZnFs and not on long arrays with variable repeat types. Recently, newer evidence supporting complex PRDM9-motif interactions was reported (Patel et al., 2016; Patel et al., 2017). Different PRDM9 variants have been predicted to use distinct sets of ZnFs to bind to the same 13bp motif. Even subtle changes in the ZnF array may contribute to the complex interactions between DNA. Allele L13, found to be non-activating with reference to the 13bp motif, is identical to the PRDM9 A except for a Ser→Arg substitution at the non-contact position -2 in the 11th ZnF (Berg et al., 2010). This substitution is not predicted to alter DNA binding characteristics yet allele L13 is unable to activate the hotspots enriched with the 13bp motif. Also, Allele L20, which is identical to allele A except for a Asn→His substitution in the tenth ZnF and which falls on helical position 3 of the four DNA-contact residues of this ZnF, is also unable to activate the 13bp motif-containing hotspots.

Alternatively, subsets of PRDM9 variants may be fine-tuned to a particular version of the motif due to interactions with specific bases in the degenerate positions of a consensus motif. For example, in Ct-regulated hotspots where, even though all the Ct alleles are predicted to bind to the 16bp sequence motif, they were not all able to activate these hotspots due to insertions/deletions of ZnFs in regions distal to the motif-defining region and in the C-terminal of the ZnF array (Berg et al., 2011). Additional proteins may also be required to promote binding and initiate chromatin remodelling.

3.3.2 Variation in activity of A/N individuals at DNA3 hotspot

The variation in CO activity among A/N men at the DNA3 hotspot needs further investigation. The two men carrying PRDM9 A/L11 and A/L5 alleles showed the highest estimates of recombination activity but whether L5 or L11

are capable of activating the DNA3 hotspot on their own cannot not be tested as there were no L5/N and L11/N individuals in the current panel of donors. Berg et al. (2010) also see lots of variation in amongst the A/A men (or indeed A/N men) (eg. F and S hotspots) so this is not a novel finding but one that might simply reflect other factors – indeed other genetic factors have been implicated to play a role in CO rate variation, namely, RNF212, KIAA1462, PDZK1, UGCG, NUB1 (Chowdhury et al., 2009; Reynolds et al., 2013). It is also conceivable that epigenetic and /or environmental differences at play.

Even though no distinguishing features could be gleaned from the SNP genotyping information, the variation may still be attributed to cis-effects which alter PRDM9 A binding to the target sequence motif. In this instance, a SNP variation in certain individuals may enhance or suppress activity compared to A/A men and this could be an explanation for the higher recombination frequencies seen in the A/L5 and A/L11 men. Sequencing of the central hotspot region of the DNA3 hotspot in these individuals and the A/A men may highlight additional differences in sequences between men that may be the cause of any cis-acting effects that enhance hotspot activation.

3.3.3 Conclusions

The DNA3 and AA hotspots were targeted for investigation as elevated recombination rates had been reported in a B-ALL cohort for this region in the MHCII (Thompson et al., 2014) and rare PRDM9 alleles have been linked with B-ALL (Hussin et al., 2013). The DNA3 hotspot was confirmed to be PRDM9 A-regulated and suppressed in individuals carrying Ct alleles. Since Ct alleles are equivalent to the rare PRDM9 alleles previously reported as having an association with B-ALL (Hussin et al., 2013) as they are both noted for containing K-type ZnF repeats in the PRDM9 ZnF arrays, it is unlikely that a NAHR mechanism via DNA3 hotspot is involved.

The AA hotspot appears to be active in the male germline, albeit weakly, in men carrying two copies of Ct alleles, though ideally a more extensive panel of men should be screened for a definitive conclusion to establish whether the C allele effectively interacts with this hotspot. If NAHR within the MHCII underlies the disease aetiology of B-ALL, it therefore seems more likely that it is associated with the AA hotspot rather than the DNA3 hotspot. However, the nature of any such ectopic recombination exchange points remain elusive. However, the AA hotspot does lie within a region containing copy number variations. The fine-scale characterisation of the AA hotspot morphology is described in Chapter 4.

CHAPTER 4: CHARACTERISATION OF THE AA HOTSPOT

4.1 Introduction

This chapter describes the detailed characterisation study, using sperm typing methodology, of the African-American (AA) recombination hotspot in the MHCII region first identified by Hinch et al. (2011) using SNP genotyping, recombination maps and bioinformatic methods.

4.1.1 African-enriched hotspots

A genetic map produced from individuals with 80% African and 20% European ancestry has uncovered ~2450 African-only historical recombination peaks including one historical peak in the MHCII region (Hinch et al., 2011). This AA hotspot is ~22kb centromeric of the DNA1-3 hotspot cluster. It has a recombination intensity of 70cM/Mb, is also found in the African YRI map but not observed in the European DECODE and CEU maps. It was predicted that the hotspot would be Ct regulated and it has been confirmed to be activated by a subset of Ct alleles in Chapter 3. Due to the intense historical peak, it can be also speculated that currently unknown PRDM9 alleles might activate this hotspot. Equally, it is possible that the contribution may be from the female germline for which there is no direct experimental method available apart from expensive single molecule sequencing methods (Ottolini et al., 2015). To develop a better profile of the recombination landscape in MHCII, it was decided to characterize sperm recombinants for this hotspot. This would also permit paternal recombination rates to be compared with the sex-averaged African American genetic map produced by Hinch et al. (2011) and infer whether this is the result of high recombination activity in the female germline.

4.1.2 Crossover:Non-crossover variation and GC-biased gene conversions

COs and NCOs co-localise in mice and human hotspots (Jeffreys and May, 2004; Cole, Keeney and Jasin, 2010), suggesting that both events are initiated by the same programmed DSBs made by Spo11 during leptotene (Keeney, Giroux and Kleckner, 1997). Yet, evidence suggests that COs and NCOs may result from different processes (Guillon et al., 2005; Baudat and de Massy, 2007) with two pathways diverging early on after initiation of recombination. This is similar to the budding yeast model where COs are primarily the result of the resolution of the dHJ and NCOs are formed through single-strand end invasion of intermediates by the SDSA pathway (Paques and Haber, 1999; Hunter and Kleckner, 2001; Allers and Lichten, 2001).

As to be expected, both CO and NCO frequencies are heavily influenced by PRDM9 status. For example, the PAR2 hotspot SPRY3 is PRDM9-A regulated with the highest frequencies of COs and NCOs seen in men homozygous for the A allele (Sarbjana et al., 2012). PRDM9 influences the frequencies of both NCOs (0 to 0.35%) and COs (0.017 to 0.845%), supporting the theory that activation by PRDM9 is an upstream specifier. There is no evidence of PRDM9 influencing NCO/CO decision.

However, there is evidence for the influence of cis-acting elements causing variation in CO and NCO frequencies. The central SNP rs700442 in the SPRY3 hotspot has a 7/8 match with the 13bp motif with the SNP being in the seventh position. The C allele disrupts the activation of the hotspots by PRDM9 A alleles (Sarbjana et al., 2012). Similar cis-effects were seen in the DNA2 (Jeffreys and Neumann, 2002) and NID1 hotspots (Jeffreys and Neumann, 2005). Sperm typing for these three hotspots showed that the distribution of CO junctions appear to lack CO symmetry when reciprocal events are examined in men with certain SNP heterozygosities, with displacements of up to 440bp between orientations.

Symmetry is observed when reciprocal CO products have equal opportunity to result from DSB repair. For reciprocal COs to occur in a Mendelian sense, the symmetrical distribution of CO junctions require a) mean position of initiating DSBs to be the same on both homologues, b) the frequency of initiating DSBs to be the same on both homologues and c) subsequent biochemical steps leading to the CO event must give rise to a CO junction with equal probability to both sides flanking the initiating DSBs. At the SPRY3 hotspot, men heterozygous for the SNP rs700442 C/T allele experience more transmission distortion (TD) with the over-transmission of the C allele into the recombinant sperm (Sarbjana et al., 2012). This can be interpreted as T allele-containing haplotypes experiencing more DSBs and therefore being repaired using information from the homologue in a CO with gene conversion pathway.

The SPRY3 hotspot highlights one of the aspects that is still unclear about recombination hotspots. There is high variability in the NCO:CO frequency ratio (1:10 to 3.8:1), regardless of PRDM9 status, in the men tested. Even though the frequencies of COs across all men heterozygous for the SNP rs700442 had a consistent 2:1 transmission bias in favour of the C allele, the frequencies of transmitted alleles substantially varied in NCOs with a range from 1:1 to >85:1 in favour of the C allele.

4.1.3 Study aims

The main aim of the study was to characterise the AA hotspot in terms of morphology and further evaluate its PRDM9 regulation. If the AA hotspot was intense enough with any tested PRDM9 allele, then it was planned to evaluate Non-crossover:Crossover (NCO:CO) variation and GC-biased gene conversion and define how these processes affect the outcomes of recombination.

4.2 Results

4.2.1 Comparison of ZnF array sub-motifs in activating Ct alleles

As detailed in Chapter 3, sperm CO assays covering a ~6kb region around the hotspot interval were carried out for 6 individuals. There was obvious clustering of RF for men carrying two Ct alleles compared to the group of men who were heterozygous (Ct/A) and men who were homozygous for the PRDM9 A allele. The difference between these two groups is significant. When Ct/Ct and Ct/A individuals are grouped against A/A individuals, there was no significant difference in activity. The number of individuals tested is modest and each Ct allele could not be tested in isolation. Hence, there was no sufficient statistical power to the results. It has also been demonstrated that Ct/Ct and Ct/A groups have considerable overlap in range of activity (Berg et al., 2011). It is quite possible that the C allele may be non-activating for this hotspot.

ZnF arrays have most variation towards the 3' end and are predicted to be more actively involved in binding (Berg et al., 2011). ZnF order in Ct alleles were examined to define any subgroups that may possess similar mini-motifs. As seen in Table 6, L6 and L14 share the same FKHLOIJ motif. Individual 18 carrying L6/L14 alleles showed a RF of 0.10% (95% CI 0.053-0.183) (Fig.17), a similar magnitude to that of individual 20 who had an RF of 0.13% (95% CI 0.071-0.218). However, both C and L16 share the same motifs and yet the C allele appears to be a poor activator whereas L16 in a L14/L16 genotype has shown highest activity suggesting that it is also activating. Hence, these motifs cannot be used to predict activating alleles.

Table 6 Ct alleles of SSA individuals amenable to sperm typing assays

Ct allele	ZnF order	Availability in SSA panel
C	ABCDDCC FKHLH IJ	tested C/A, C/L19
L4	ABCDDCC CDDC FKHLH IJ	in donor panel
L6	ABCDDCC C FKHLO IJ	tested L6/L14
L8	ABCD DECC FKHLH IJ	no donor
L14	ABCDDCC FKHLO IJ	tested L14/L16 and L6/L14
L16	ABCD RC FKHLH IJ	tested L14/L16
L19	ABCDDCC FKHLH QIJ	tested C/L19
A (non-Ct reference)	ABCDDEC FGHF IJ	tested A/A

4.2.2 Extended LD survey of the AA hotspot region

The recombination map by Hinch et al. (2011) had a 10kb smoothing process and showed one strong historical recombination peak. A reconstructed AA map (Fig. 8) using data from the Hinch et al. (2011) showed high two recombination peaks with recombination intensities of 30 and 66cM/Mb over a 401bp span. Smoothing may have shifted the AA peak position and made it difficult to pinpoint the hotspot location for experimental purposes. Hence, sperm typing assays were designed using the reconstructed AA map as a reference to locate the peak at a finer scale.

In fact, it remained possible that the low RFs determined in Chapter 3, were not a reflection of lower than historical levels of recombination in this region, rather that they might be a technical artifact caused by the assays not fully encompassing the hotspot. Hence, LD maps were used to check whether crossover assays were targeted to the correct region.

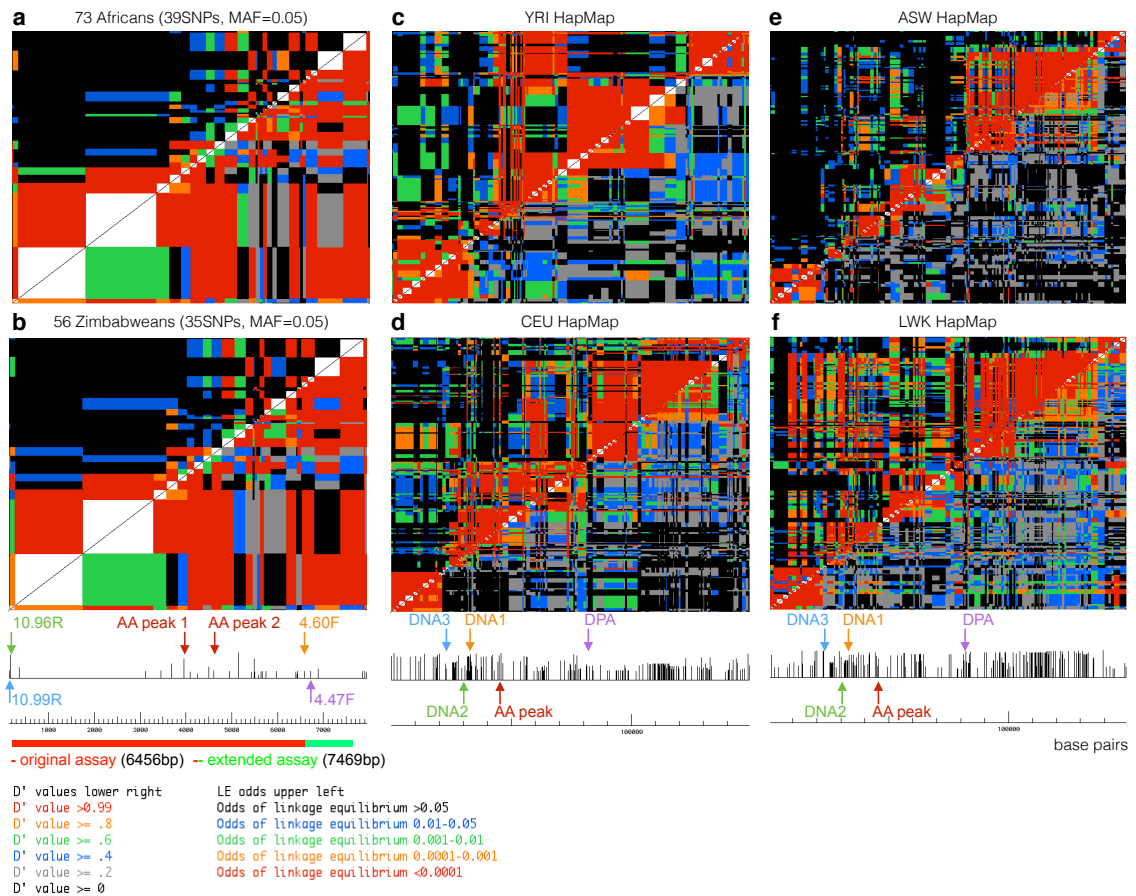


Fig. 19 Patterns of Linkage Disequilibrium in the AA hotspot interval. $|D'|$ is a measure of LD between each pair of bi-allelic SNPs. $|D'|$ plots were constructed using the 'Diploid LD' software developed in TrueBASIC by Alec Jeffreys in Ref. 11. SNPs where genotyping suggested that they were not in Hardy-Weinberg Equilibrium were excluded. **a** $|D'|$ plot of a ~8kb region encompassing the AA crossover assay region using 39SNPs in 73 Africans **b** $|D'|$ plot of a ~8kb region encompassing the AA crossover assay region using 35 SNPs in 56 Zimbabweans. **c** $|D'|$ plot of a 15kb region of the MHCII based on HapMap-YRI genotype information **d** $|D'|$ plot of a 15kb region of the MHCII based on HapMap-CEU genotype information **e** $|D'|$ plot of a 15kb region of the MHCII based on HapMap-ASW genotype information **f** $|D'|$ plot of a 15kb region of the MHCII based on HapMap-LWK genotype information. In all plots, chromosome positions are with reference to NCBI36/hg18 genome assembly. SNPs where genotyping suggested that they were not in Hardy-Weinberg Equilibrium were excluded. The legend shows the D' categories for the colouring the squares of D' and linkage equilibrium (LE) values. Breakdown in LD is indicated by black/grey and blue squares) and blocks where SNPs are in strong association blocks are indicated in red and orange. The CO selector sites and hotspot peaks are 10.99R (rs17220631 A/C, chr6:33092233), 10.96R (rs9276994 A/G, chr6:33092233), 4.60F (rs6930399 C/T, chr6:33098653), 4.47F (rs6930608 C/T, chr6:33098789) whilst the AA peak 1 is (7.51/rs423639, 33095752) and AA peak 2 (7.10, rs17214311, 33096153). The HapMap-based plots situate the AA peak at a small region of LD breakdown within a large LD block flanked by the DNA hotspot cluster towards the distal end and the DPA hotspot towards the proximal end. The LD block is most pronounced in the CEU plot. The LD breakdown is most apparent in the African YRI, ASW and LWK and weakest in the CEU. In the SSA panel plot a and b, LD breakdown is seen at the proximal end of the assay region.

As disequilibrium, D , is measure used for the non-random association of alleles of SNPs and other loci, D would be useful to define haplotype blocks formed by the SNP alleles in this region. Specifically, D' divided by D_{\max} resulting in 'D Prime' or $|D'|$ was used to negate the influence of deficiencies in allele frequency data for underrepresented SNPs in datasets, enabling the measurement of pairwise association of bi-allelic SNPs within and between populations (May, Slingsby and Jeffreys, 2008). $|D'|=1$ means recombination between the loci is completely non-random with only two of the four possible haplotypes being observed as compared to the expected frequency of all four haplotypes if there was complete free association or $|D'|=0$, indicating random recombination in the region between each pairwise SNP. SNP genotyping information was used to construct $|D'|$ plots (Jeffreys, Kauppi and Neumann, 2001) for 73 Africans and 56 Zimbabweans in the SSA panel along with HapMap population reference data sets (Fig. 19). Evidence of SNPs in free association is observed towards the proximal end of the assay region in $|D'|$ plots.

To further visualise the extent of this LD breakdown, these 6 datasets plus the MKK HapMap population, were used to construct LD Unit (LDU) maps (Fig. 20). LDU maps originate from Malecot model of decay in association where the distance is determined by recombination rates (Malécot, G.,Blaringhem, L., 1948). The LDU map visualises as a series of additive steps spaced by plateaus that represent the distance of haplotype blocks. In addition to showing physical distance, LDU maps are comparable to genetic maps as the length of the LD units equate to the genetic distance in terms of recombination. An LDU step of ~ 1.5 units is observed for the two SSA datasets. Modest LD breakdown is also seen in the YRI and MKK maps.

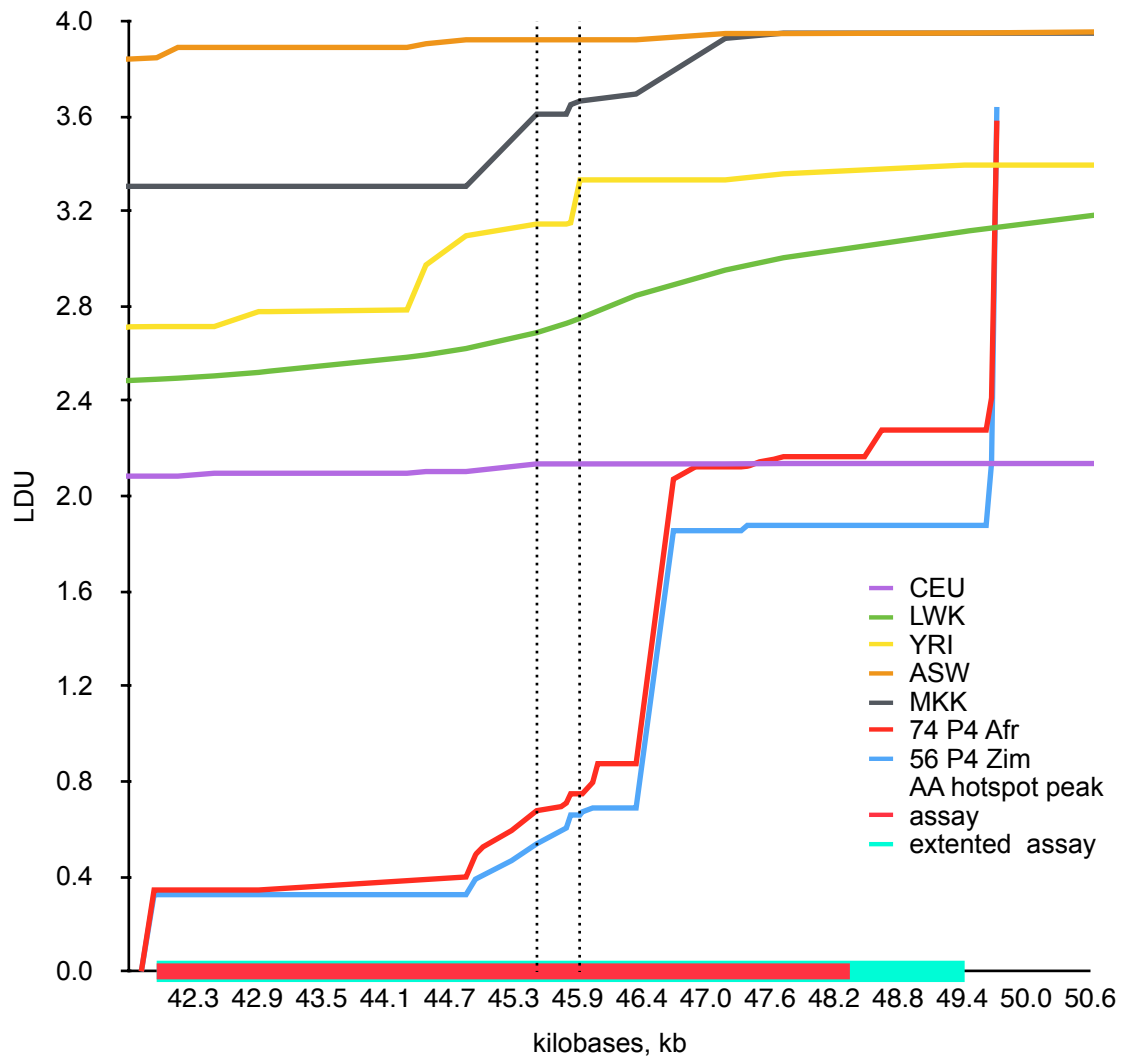
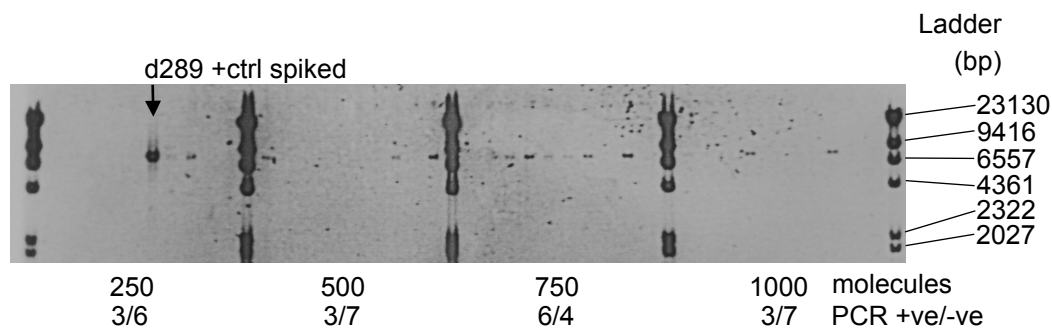


Fig. 20 Linkage Disequilibrium Unit (LDU) Map constructed from LDUs over physical distance using data sets from 74 Africans and the 56 Zimbabwean subset from the SSA panel and from CEU, YRI, ASW and LWK HapMap data sets. The AA hotspot peaks, the original assay region for individual carrying L14/L16 alleles and the proposed extended assay region are also indicated. The CEU and ASW maps are nearly flat over the region suggesting a large conserved LD block extending beyond this region whilst the YRI and MKK maps showed more modest increases with small steep steps over the predicted AA hotspot peak. The LWK map showed a more gradual slope consistent with smaller LD blocks seen in its $|D'|$ plot.

Based on these findings, there remained the possibility that the assay might be underestimating the number of recombinants recovered due to the proximal selector sites being too close to the region of LD breakdown. Therefore, the assay was extended by ~1kb towards the proximal end (Fig. 21). However, for individual d257 (20), a RF of 0.08% (95% CI 0.04-0.14) was obtained over this extended interval, demonstrating that the original assay was in fact recovering all crossovers initiating from the AA hotspot.



1⁰ PCR: 3.34F/G forward with 10.99R/A reverse ASPs (7721bp)
 2⁰ PCR: 3.53F/A forward with 10.96R reverse ASPs (7469bp)

RF=0.08% at AA hotspot

Individual d257, PRDM9 L14/L116

Fig. 21 Crossover PCR assay products run on a 0.8% agarose gel with 0.5µg/ml Ethidium Bromide and visualized by ultraviolet (UV) transillumination. Allele-specific primer (ASP) combinations and amplicon lengths are indicated below the gel photograph. The CO selector sites are 10.99R (rs17220631 A/C, chr6:33092233), 10.96R (rs9276994 A/G, chr6:33092233), 3.53F (rs13217173 A/C, chr6:33099716) 3.34F (rs72860818 A/G, chr6:33099915). Individual d289 who carried a parental haplotype identical to the recombinants at the reverse selector sites but heterozygous at the forward selector sites was used as the best available control targeted by the ASPs was used as a positive control. PCR-positive and PCR-negative reactions were counted and analysed using the Poisson correction program written by Alec Jeffreys. Individual d257 carried two Ct alleles (L14/L16) and showed a recombination frequency of 0.08%.

4.2.3 Hotspot morphology

The redesigned crossover interval contained a total of 97 SNPs, of which 26 were within ~1kb of the predicted hotspot centre. Man 20, carrying L14/L16 PRDM9 alleles exhibited a recombination frequency (RF) of 0.13% (95% CI 0.071-0.218) which enabled the recovery of sufficient numbers of recombinants to explore AA hotspot morphology and investigate potential cis-effects.

Haplotype sequencing over the assay interval for this sperm donor enabled phasing of 28 SNPs with a range of 9-1097bp (mean, 120bp; median, 229bp) between adjacent markers. The central ~1kb of the predicted hotspot contained 9 SNPs from rs423639 to rs9296068. Crossover assays were carried out in alternate orientations and SNP genotyping of internal markers allowed CO resolution breakpoints to be mapped (Fig. 22). This showed that the hotspot

centre was slightly shifted centromeric to the predicted hotspot region. All crossovers were mapped between 10 SNPs from rs6457702 to rs72860807 and range of 16-346bp between adjacent markers (mean, 129bp; median, 107bp). Orientation A assays yielded 114 COs out of 114,000 molecules screened, at a frequency of 0.101% (95% CI 0.082-0.12), and orientation B assays yielded 123 COs out of 136,240 molecules screened, at a frequency of 0.09% (95% CI 0.075-0.108), (Fisher's Exact test, $P>0.10$). In both orientations, the peak number of COs mapped to the same 346bp interval between rs9296068 and rs72858101. Cumulative frequency of COs across the assay region for each orientation showed a similar profile. Hotspot widths for orientation A and B were estimated at 1447bp and 1408bp, respectively. The shift between hotspot centres was only 70bp indicating that an accurate profile can be obtained from both assay orientations. Hence, the hotspot centre occurs at Chr6:~33096471 and ~580bp proximal to that predicted by the map.

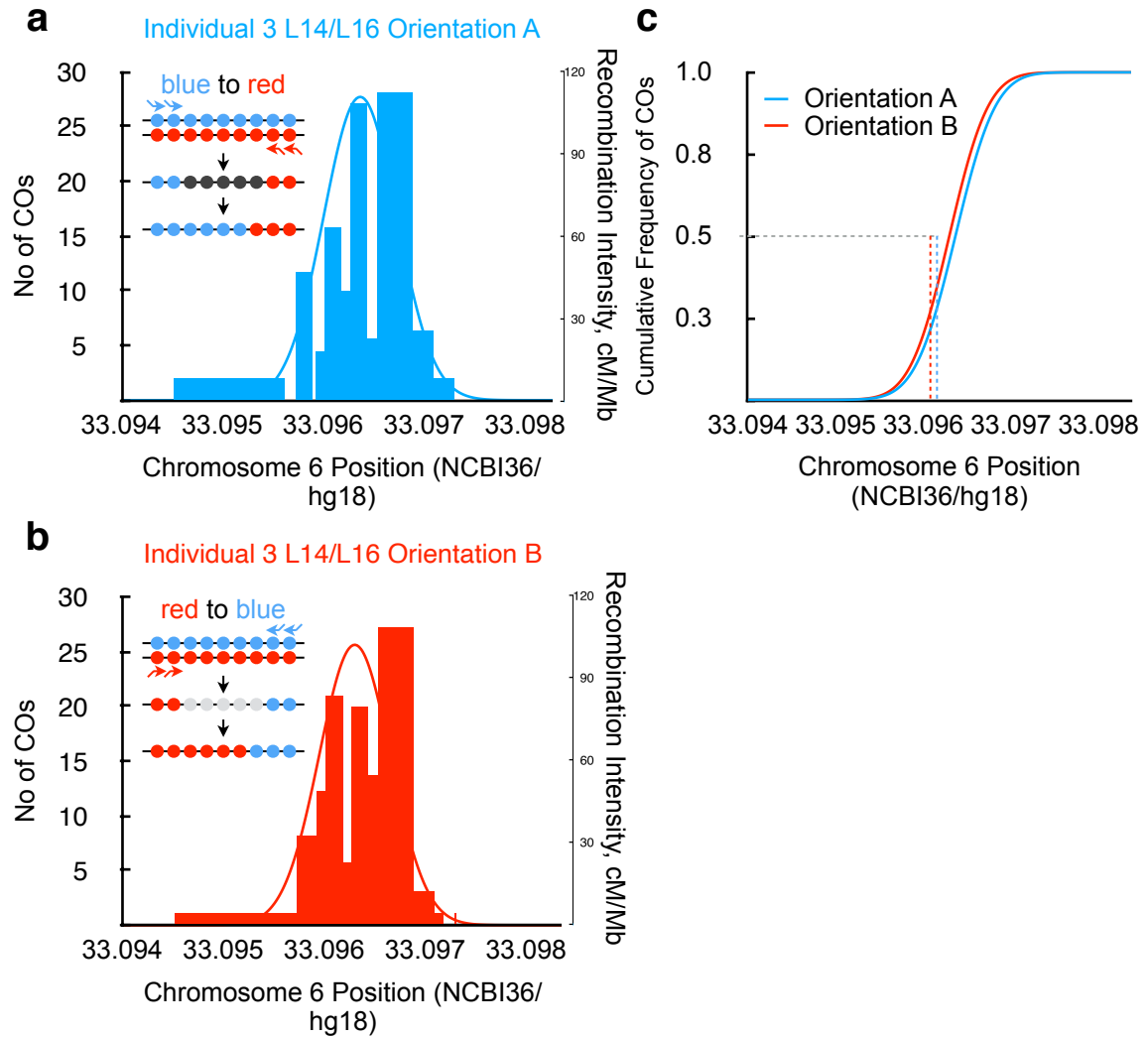


Fig. 22 AA hotspot morphology. Chromosome positions are with reference to NCBI36/hg18 genome assembly **a** Distribution of COs between informative markers and the recombination intensity across the span of the hotspot in Orientation A using CO selector sites 10.99R/A (rs17220631 A/C, chr6:33092233), 10.96R/G (rs9276994 A/G, chr6:33092233), 4.60F/C (rs6930399 C/T, chr6:33098653), 4.47F/C (rs6930608 C/T, chr6:33098789). **b** Distribution of crossovers between informative markers and the recombination intensity across the span of the hotspot in Orientation B using CO selector sites 10.99R/C (rs17220631 A/C, chr6:33092233), 10.96R/A (rs9276994 A/G, chr6:33092233), 4.60F/T (rs6930399 C/T, chr6:33098653), 4.47F/T (rs6930608 C/T, chr6:33098789). **c** cumulative frequency of crossover across the assay region with both orientations. The midpoint on the y-scale is used to determine the hotspot centre.

4.3 Discussion

The work has shown that the AA hotspot is used in contemporary populations and is activated by a subset of Ct alleles, namely L6, L14, L16 and L19 but not by allele C. The historical peak suggested a very active hotspot (Hinch et al., 2011) but this is not reflected in this study. If it is not due to a

strong signal left by female germline recombination, it may be that the hotspot was used more extensively by Ct alleles that were more prevalent in the past and/or Ct alleles not tested in this work.

Differential Ct activation of hotspots has been noted previously; all Ct alleles tested activate the 12D hotspot whereas only L6 and L8, which possess dissimilar 3' end motifs, activate the 12B hotspot (Berg et al., 2011). Hence, motif similarity is not a reliable indicator of activating alleles for a specific hotspot. Upstream ZnFs may play a role in these subtle interactions. It would be helpful to examine the Ct alleles tested at the AA hotspot in their isolated states. However, there are currently no suitable sperm donors available amongst the Leicester collection to enable these comparisons.

The nearest 16bp consensus motif is ~1200bp upstream of the hotspot centre and so cannot be involved in activation. However, it is possible that Ct alleles may nonetheless be interacting with higher affinity to a sequence similar to the 16bp consensus motif. Patel et al. (2016) similarly noted that the THE1B hotspot sequence first described by Myers et al., in 2008, was bound by PRDM9 A with higher affinity than the 13bp consensus motif. Assaying both orientations for individual 3 showed no initiation differences indicating that no cis-effects are operating at this hotspot. Since very small pool sizes are necessary for obtaining NCOs, the low activity of this hotspot makes this work impractical. However, there are other candidate LD hotspots (Hinch et al., 2011) and meiotic DSB hotspots (Pratto et al., 2014) that might be more amenable to such analysis and therefore further our understanding of CO:NCO ratios.

Uncharacterised alleles in the more diverse African populations may be activating the AA hotspot. New strategies could be explored to use available Next-Generation Sequencing (NGS) data and samples from African and other populations to discover novel alleles. Such approaches are the focus of Chapter 6.

CHAPTER 5: CHILDHOOD ALL AND PRDM9 ALLELES RARE TO EUROPEANS

5.1 Introduction

This chapter describes a study primarily carried out to establish whether K-type ZnF coding repeat containing PRDM9 alleles are associated with the development of childhood ALL in a British cohort to corroborate the same association demonstrated in a French-Canadian ALL cohort. Additionally, four coding variants in the FIGNL1 gene region and two other SNPs, all identified by Genome-Wide Association Studies (GWAS) for ALL susceptibility were examined in this British cohort to verify any connections to childhood ALL. This study was a collaboration with Dr Pamela Thompson at the University of Manchester.

5.1.1 K-ZnF containing PRDM9 alleles

A subset of PRDM9 alleles contain K-type ZnF coding repeats (K-ZnF) in their arrays (Berg et al., 2010; Berg et al., 2011). These K-ZnF repeats have been found in the D and L20 alleles in addition to all Ct alleles. Ct alleles have been defined as having a 5/8 match to the non-degenerate positions of the 13bp CCNCCNTNNCCNC motif (Berg et al., 2010) demonstrated to be bound by the common PRDM9 A allele (Baudat et al., 2010). In comparison, D and L20 alleles have a 7/8 match to this motif (Berg et al., 2011). The frequency of Ct alleles in African populations is higher (34.5%) than in European populations (1.3%) (Berg et al., 2010). However, whilst D and L20 alleles have been found in British/North European populations with frequencies of 4% and 1% respectively, these alleles have yet to be reported in individuals of African origin. The order of ZnFs coding repeats in these alleles also do not share a distinct similarity (Berg et al., 2011). Hence, the only common feature of D, L20 and Ct alleles is that they all carry K-ZnFs.

In 2013, Hussin and colleagues reported an overrepresentation of K-ZnF containing PRDM9 alleles in the parental population of French-Canadian nuclear family units with B-ALL affected children (Hussin et al. 2013). This cohort consisted of 22 parental trios and one family of both parents and two affected male siblings. When ThermoFisher Scientific SOLiD 4.0 System exome sequencing reads were aligned to the PRDM9 gene 12 of 46 parents were found to carry PRDM9 alleles containing K-ZnFs, which are ordinarily rare in European populations compared to the FCEXOME cohort, which consists of 68 healthy parents from three disease cohorts (two-tailed Fisher's Exact test, $P=0.0181$). Additionally, Sanger re-sequencing of 13 parent duos from this cohort plus 76 parents of an ethnically matched French-Canadian family cohort further showed an excess of K-ZnF (C, D, L20) alleles (two-tailed Fisher's Exact test, $P=6.25 \times 10^{-3}$) with 76.9% of families having at least one parent carrying a K-ZnF allele.

The association was found in the B-ALL children as well (two-tailed Fisher's Exact test, $P=0.0123$) and was reproduced in an independent American cohort of 50 affected children from Tennessee, USA (one-tailed Fisher's exact test, $P=0.0353$, compared to 1000 Genomes Project CEU controls). In this American cohort, there were no children with African ancestry as ascertained by principle component analysis of genome wide data and hence the probability of Ct alleles contributing to this overrepresentation seems unlikely due to the lower frequency of Ct alleles in populations of European ancestry. Since read mapping data was not available, it was not possible to investigate whether K-ZnF containing sequence reads actually belonged to another part of the target region. Moreover, the relatively high frequency of these alleles in African populations (Berg et al., 2011; Hinch et al., 2011) combined with a lower incidence of childhood leukaemia among African Americans (Gurney et al., 1995) suggest that PRDM9 variation alone cannot explain disease aetiology. The second trigger for B-ALL has therefore been speculated to be some kind of environmental exposure.

As stated previously, the frequency of Ct alleles is predicted to be low in a British ALL cohort since these are rare in European populations (Berg et al., 2010). However, the D and L20 alleles, which have so far only been observed in Europeans, also contain K-ZnFs so in this study I sought to establish their frequency in the British ALL cohort and determine whether these alleles are similarly overrepresented, thereby providing support for the findings from the French-Canadian and USA cohorts (Hussin et al., 2013).

5.1.2 FIGNL1 variants and other GWAS SNPs

GWAS has been extensively used to identify allelic variants that increase susceptibility to complex genetic disorders. The Ikaros Family Zinc Finger 1 (*IKZF1*) gene, which codes for a protein that regulates lymphocyte differentiation, has previously been identified for a role in childhood ALL disease aetiology (Papaemmanuil et al., 2009; Trevino et al., 2009). It is one of several genes identified by GWAS as being in LD with SNPs associated with the development of childhood ALL.

A study was carried out on seven childhood ALL patients derived from a British ALL cohort consisting of local (Manchester) and national British children diagnosed with B-ALL between 1991 and 2001. Using targeted next generation sequencing (NGS) with Sure Select Target Enrichment on a SOLiD 5500xl platform, Dr Pamela Thompson at the University of Manchester reported that four patients carried four different germline coding variants in exon 4 Fidgetin-like 1 (*FIGNL1*) gene region, which is in close proximity to the *IKZF1* gene (May et al, unpublished). The *FIGNL1* gene encodes a member of the AAA-ATPase family of proteins that dephosphorylates ATP into ADP and a phosphate ion (PO_3^{-4}). The *FIGNL1* protein is recruited to sites of DNA damage and participates in double-strand break repair via homologous recombination. *FIGNL1* also binds RAD51, an important protein involved in stabilisation and DSB repair in homologous recombination. (Yuan and Chen, 2013). It is thought to

play a role in the proliferation and differentiation of osteoblasts. FIGNL1 is also linked to the CDKN2A and HSPA1L genes and seems to be an interesting candidate for ALL susceptibility. Among the 4 SNPs identified in this study, SNP rs10235371 G/A (FIGNL1_4) is a non-synonymous and putatively deleterious mutation. SNP rs62445870 G/A (FIGNL1_1) is also a non-missense type mutation, which may be expected to impact on the protein structure or function though different software packages give conflicting predictions. In contrast, SNP rs112666980 (FIGNL1_3), also leading to a non-synonymous mutation is not predicted to affect the protein and SNP rs61735234 C/A (FIGNL1_2) is a synonymous mutation. Two other SNPs were also identified via haplotype analysis by Dr. Pamela Thompson. SNP rs146179135 C/G is in the CDKN2A gene region, which is associated with susceptibility to childhood ALL (Sherborne et al., 2010). SNP rs148661414 is a rare SNP (0.007 for Europeans in 1000Genomes) (Sherry et al., 2001) purported to be in a PRDM9-recognised motif.

Collectively, these SNPs represent candidates for association with childhood ALL that could be verified using the British ALL cohort made available for study by our University of Manchester collaborator. The cohort consists of 373 B-ALL patients, 260 B-ALL patients' fathers, 303 B-ALL patients' mothers, 43 B-ALL patients' siblings, 57 T-ALL patients, 49 T-ALL patients' fathers, 48 T-ALL patients' mothers and 25 T-ALL patients' siblings, along with 565 healthy controls. SNP genotyping methods and comparisons with healthy controls and between cytogenetic subtypes of ALL could therefore be used to investigate these links.

5.1.3 Study aims

The main aim of this study was to identify whether any of the K-ZnF containing PRDM9 alleles (D, L20 and Ct alleles) and /or the four FIGNL1 and additional GWAS SNPs have significant associations with childhood ALL. The

study leveraged predictive SNPs for D, L20 and Ct alleles based on a PRDM9-SNP haplotype network to screen the British ALL cohort for potential carriers of K-ZnF alleles. Of the potential carriers identified, a representative sample was then further characterised using allele-specific PCR and subsequent Sanger sequencing of the PRDM9 ZnF arrays. In order to verify the results from the British ALL cohort, bioinformatic methods were also used to phase and impute FIGNL1 SNPs genotypes in an independent German ALL cohort dataset consisting of primary genotype data kindly provided by Prof Martin Stanulla at Hannover Medical School (Ellinghaus et al., 2012).

5.2 Results

5.2.1 Further validation of a PRDM9-SNP Haplotype Network

To screen the British ALL cohort for individuals carrying potential K-ZnF-containing PRDM9 alleles, an existing PRDM9 allele-associated SNP haplotype network (CA May and JH Wetton, unpublished) was used to select predictive SNPs for D, L20 and Ct alleles (Fig. 23). The network had been based on PRDM9 allele characterisation and SNP genotyping in the PRDM9 gene region in 183 men (R Neumann, unpublished). SNP haplotypes were inferred using two software programs in tandem. Since both programs overestimate the number of possible haplotypes, the resulting data were curated to minimise the number of haplotypes and the most parsimonious connections were used to draw the network (Bandelt et al., 1995) (Fig. 23).

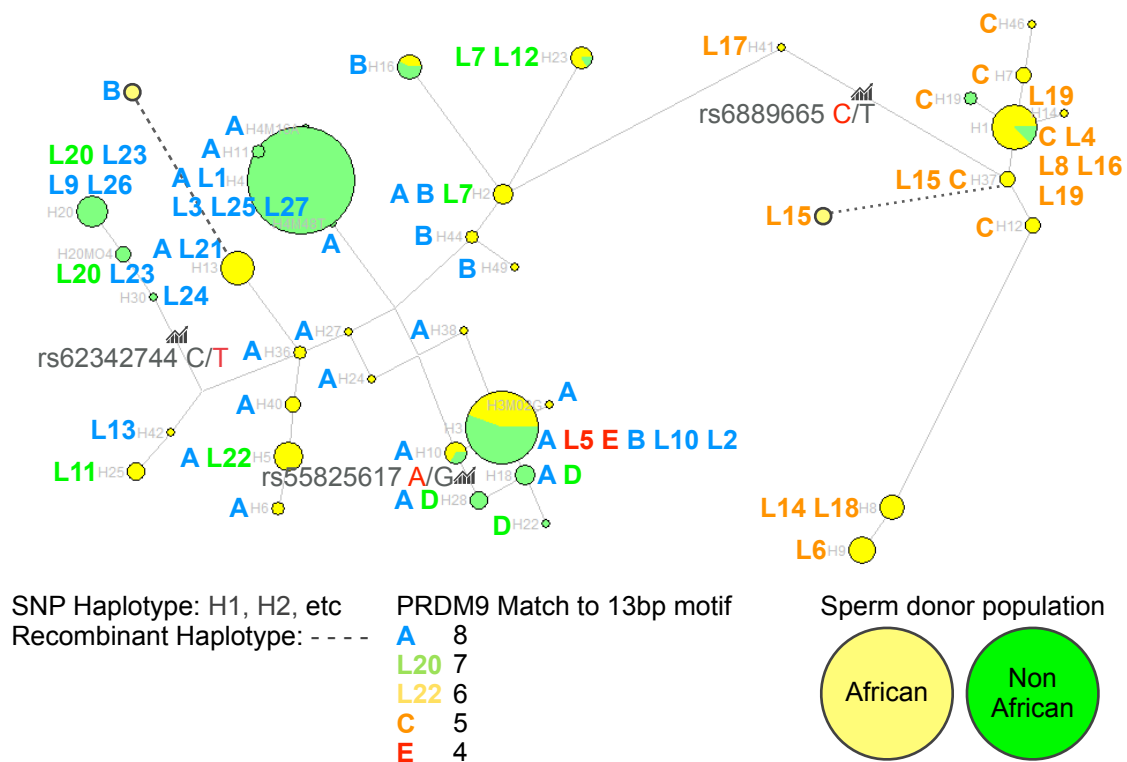


Fig. 23 PRDM9-associated SNP haplotype network. The network is based on genotyping of 48 SNPs in a ~35.5kb interval spanning the PRDM9 gene region of Chromosome 5 and the PRDM9 allele characterisation of 183 men (102 British, 56 Zimbabwean, 17 Afro-Caribbean, 2 Indian, 2 French, 1 African, 1 Greek Cypriot, 1 mixed British/Indian and 1 mixed British/Sri Lankan). SNP haplotypes were inferred using both the 'Haplotype Extractor Program' (AJJ Jeffreys, unpublished) and PHASE (Stephens, Smith and Donnelly, 2001; Stephens and Scheet, 2005). The resulting haplotypes were curated by eye to achieve the most parsimony. The network was then drawn using Network by Fluxus (Jon H Wetton). PRDM9 alleles and relevant predictive SNPs were then annotated manually. In the network, rs6889665 C/T (Chr5:23568400, NCBI36/hg18), the predictive SNP for PRDM9 Ct allele clusters out all Ct alleles. The minor T allele of SNP rs62342744 C/T (Chr5:23558129, NCBI36/hg18) clusters the L20 alleles and the minor A allele of SNP rs55825617 G/A (Chr5:23559650, NCBI36/hg18) clusters all D alleles.

The network concurs with Hinch et al. (2011) in that SNP rs6889665 C/T is a definitive predictor of Ct alleles as all haplotypes for Ct alleles cluster exclusively with this SNP in the network. In the Hinch and colleagues study, SNP rs6889665 C/T (5861bp upstream relative to the 5' end of the PRDM9 ZnF array) had been shown to be strongly associated with COs in African-enriched hotspots in an African-American cohort (GWAS/OR, $P=1.5 \times 10^{-246}$). This SNP also had the strongest link to the usage of LD hotspots (GWAS/OR, $P=1.8 \times 10^{-52}$) and coupled with its association with the PRDM9 gene and allele status, points to its importance as an indicator for strong historical peaks. By analysing 1000 Genomes Project sequencing data from 139 individuals, the ancestral T allele

was shown to be strongly associated with PRDM9 alleles A and B which have exact 8/8 match to the 13bp motif whilst the derived C allele, which has a frequency of 0.29 and 0.02 in YRI and CEU maps, respectively, was linked to alleles which only have a 5/8 match to the 13bp motif and predicted to bind to 16bp motif that Ct alleles are predicted to bind. Hence, the ZnF array length of 345 individuals including 166 African Americans were measured and it was found that the ancestral T allele was linked to PRDM9 alleles with <14ZnFs whilst the derived C allele was linked to PRDM9 alleles with >14ZnFs (Hinch et al., 2011). This correlation was almost perfect for Ct alleles apart from L15, L16 and L18 that have only 13 ZnFs in their arrays.

In the tested North Europeans (102 British, 2 French, 1 Greek Cypriot), the frequency of L20 and D was 0.04 and 0.01, respectively (Berg et al. 2010). No men from the African populations carried an L20 allele. The mixed British/Sri Lankan individual carried one D allele. The network indicated, there were no unique SNP haplotypes separating either L20 or D alleles. The L20 allele was associated with two haplotypes, H20 where it clustered with L9, L24 and L26, and H20MO4 where it clustered with L23. The minor T allele of SNP rs62342744 C/T was associated with both of these two haplotypes and so was selected as a predictive SNP for the L20 allele. The D allele clustered with the common A allele in the H18 and H28 haplotypes whilst it was found by itself in the H22 haplotype but this contribution originated from the mixed British/Sri Lankan individual. The minor A allele of SNP rs55825617 G/A was associated with these three haplotypes and was selected as a predictive SNP for the D allele. To further validate this PRDM9-SNP haplotype network as reliable predictors of D and L20, phasing experiments were performed to demonstrate the physical link between these PRDM9 ZnF arrays and their respective SNP alleles in sperm donors. These proof-of-principle experiments used allele-specific PCR to demonstrate that the minor alleles of SNPs rs62342744 and rs55825617, T and A, are physically linked to L20 and D alleles, respectively.

For the L20 allele, ASPs for SNP rs62342744 C/T were used to selectively amplify a 5.6kb region encompassing the PRDM9 ZnF array in 12 sperm donors (Fig. 24). Individuals who were heterozygous for both alleles generated the correct size of amplicons whilst those who were homozygous for the C allele did not generate any amplicons for the T allele version of the ASP.

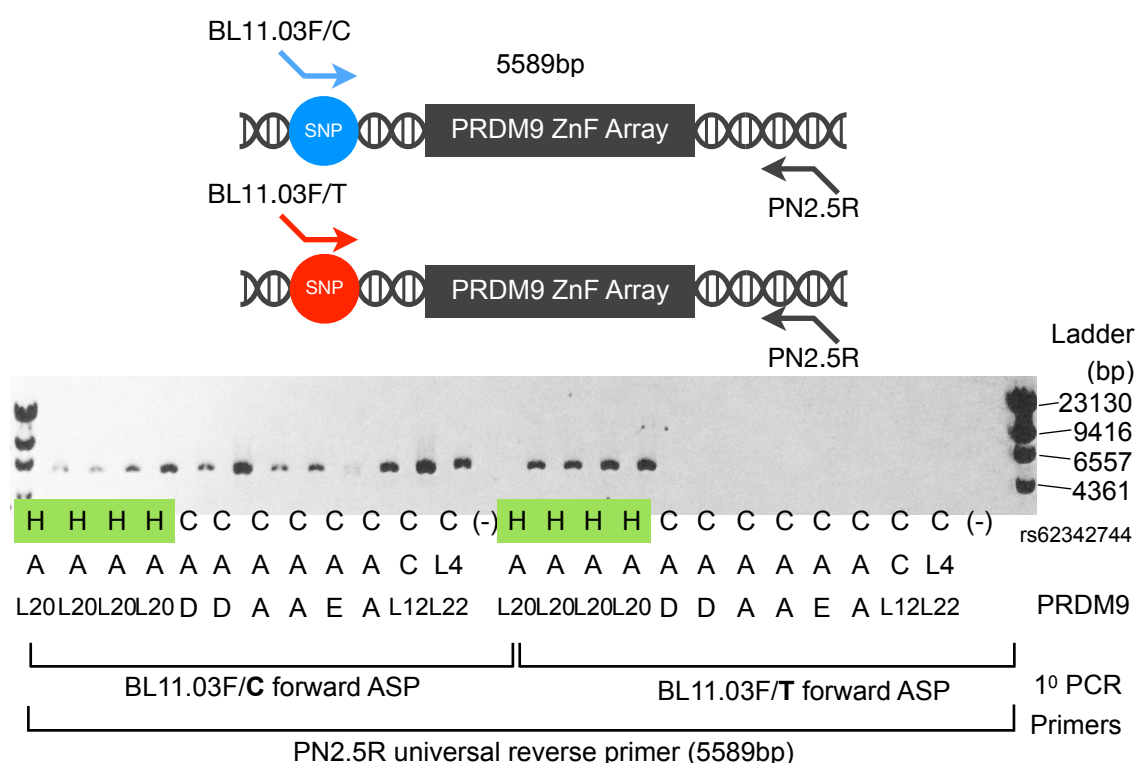


Fig. 24 Phasing rs62342744 C/T with PRDM9 L20 alleles. A PCR using ASPs BL11.03F/C and BL11.03F/T for the SNP site and a universal reverse primer PN2.5R covering the PRDM9 ZnF array was tested in 12 individuals including those carrying L20 alleles. The assay shows good specificity and adequate yield for Sanger sequencing.

The PCR positive amplicons were sequenced using the Sanger method over a ~1900bp region targeting the PRDM9 ZnF array. The sequences were aligned with individual ZnF and mini-motifs reference sequences (Appendix IV) to construct a consensus reference array that could be compared with all known PRDM9 ZnF arrays (Benson et al., 2008; Berg et al., 2011). This characterisation of the PRDM9 alleles confirmed that the minor T allele of SNP rs62342744 C/T formed a haplotype with PRDM9 L20.

The predictive SNP rs55825617 G/A for PRDM9 D was found at the end of a L2c-type Long Interspersed Nuclear Elements (LINE) repeat element (Chr5: 23523759-23,523,905) that also had a 100% match in chromosome 16 and with a run of five A-nucleotide bases just preceding the SNP where the PCR primers could be designed. Annealing temperature titrations with designed versions of ASPs produced mixed results, mainly a lack of allele-specificity. Hence, a universal PCR covering rs55825617 G/A and the PRDM9 ZnF array was used followed by a hemi-nested PCR with the ASPs (Fig. 25). The PCR amplicons were separated by gel electrophoresis since the D allele, which has 14 ZnFs in the array, can be distinguished from an A PRDM9 allele, which has 13 ZnFs. Fragments were excised from the gel were purified and Sanger sequenced (Chapter 2). The resulting consensus ZnF array confirmed that the minor A allele of rs55825617 was indeed linked to PRDM9 D.

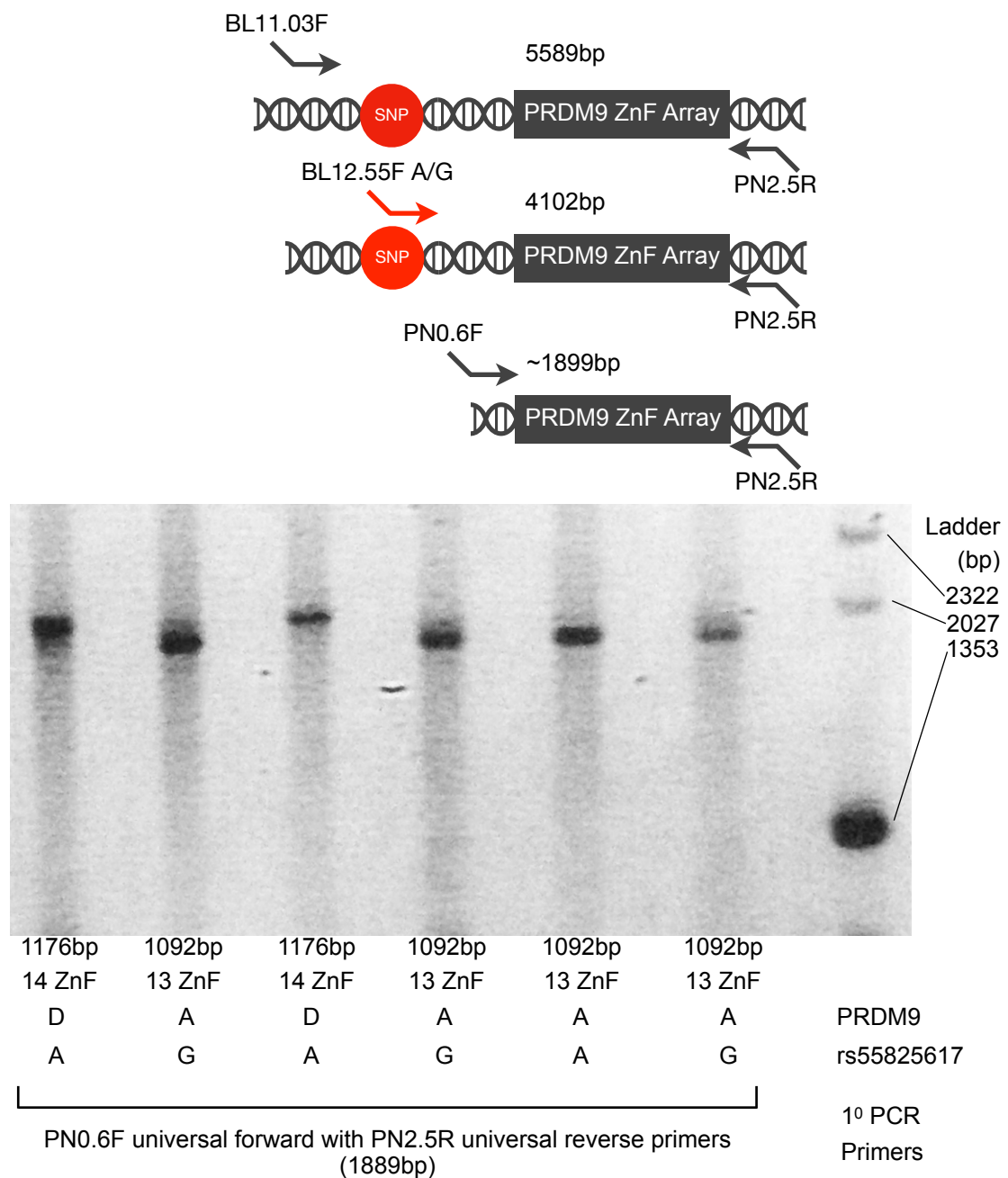


Fig. 25 Phasing rs55825617 A/G and PRDM9 D alleles. The primary PCR product amplified using universal primers was seeded into a hemi-nested secondary PCR with ASPs BL12.55F/A and BL12.55F/G and the previous reverse primer PN2.5R. The secondary PCR product was seeded into a hemi-nested PCR with PN0.6F and PN2.5R primers that are normally used for creating PRDM9 sequencing templates. Due to the size difference that could be observed via gel electrophoresis at ~2kb amplicon sizes, it was possible to differentiate the D (14 ZnF repeats) and A (13 ZnFs repeats) alleles of the tested individuals. Negative control is homozygous for PRDM9 A and homozygous G for rs55825617. However, a product was seen for the reaction amplified via BL12.55F/A which was interpreted as the ASP being non-specific. Further optimisation is required.

5.2.2 SNP genotyping on the British ALL cohort

To conserve the DNA available from the British ALL cohort for the proposed investigations and to most easily comply with the associated ethical approval for their use, multiplex PCRs were carried out by Dr Pamela Thompson at the University of Manchester and sent to Leicester for the subsequent SNP genotyping. Thus, the screening for Ct alleles via SNP rs6889665 C/T amongst the British ALL cohort and controls was done from a multiplex PCR that also contained amplicons for the four FIGNL1, one CDKN2A and one MHCII AluSg SINE SNPs.

Initially this multiplex PCR included the first 3 amplicons (Table 7), consisting of a 160bp region containing SNP rs6889665 C/T in addition to two more amplicons, a 554bp region covering the four FIGNL1 SNPs and a 1.7kb region containing SNP rs148661414. Hence SNP genotyping was successfully done for SNP rs148661414 A/G for 93 cord blood control samples, 91 B-ALL patients, 32 fathers, 48 mothers and 5 siblings of patients. The SNP was found to be completely monomorphic for the G allele in both case and controls except for one control individual who was heterozygous and as a consequence no further genotyping was carried out for this SNP. Subsequent multiplex PCRs omitted the relevant amplicon (number 3 in Table 7). Due to difficulties in stabilising the ASOs for the SNP rs146179135 C/G, reliable scoring could not be achieved for this marker so genotyping of this marker was also abandoned. Finally, genotyping of rs112666980 G/T, one of the FIGNL1 SNPs, showed this marker to be completely monomorphic with respect to the G allele amongst both cases and controls.

Table 7 SNPs identified by haplotype analysis for B-ALL susceptibility

SNP ID	Chr6 Position (GRCh37/hg19 Feb 2009)	Amplicon	Location
rs6889665 C/T	chr5:23532643	1: chr5:23532552-23532711 (160bp)	downstream of PRDM9
rs62445870 A/G	chr7:50514904		
rs61735234 A/C	chr7:50514833		
rs112666980 G/ T	chr7:50514651	2: chr7:50514546-50515099 (554bp)	FIGNL1 exon
rs10235371 A/G	chr7:50514577		
rs148661414 A/ G	chr6:32967118	3: chr6:32966002-32967731 (1730bp)	MHCII
rs146179135 C/ G	chr9:21960985	4: chr9:21970829-21971286 (458bp)	CDKN2A

To determine the SNP genotypes of the predictive SNPs for the K-ZnF containing PRDM9 alleles D and L20 in the British ALL cohort, a separate screening method for SNPs rs62342744 C/T and rs55825617 G/A was developed using a duplex PCR and ASO genotyping (Fig. 26).

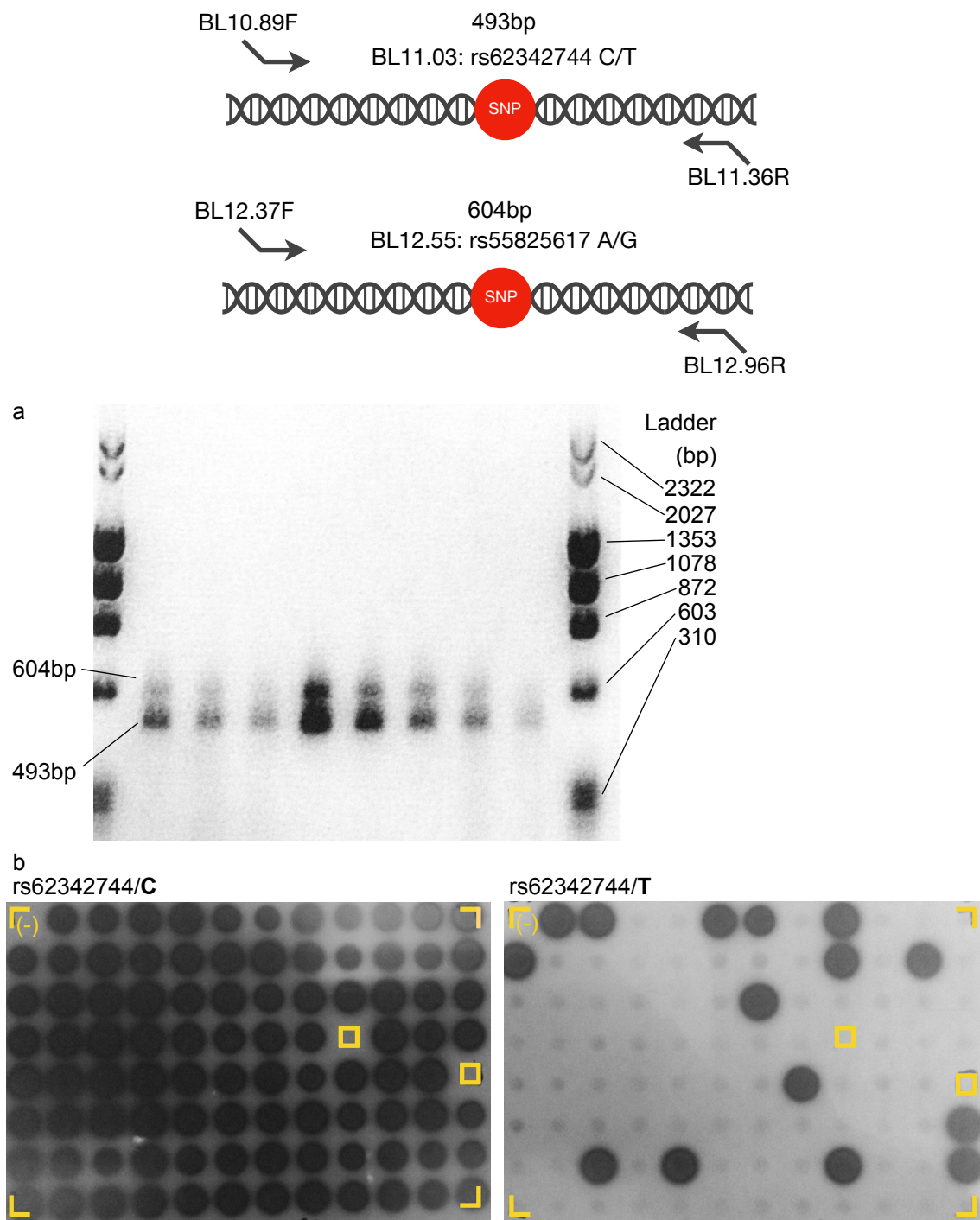


Fig. 26 a) Duplex PCR for generating 2 amplicons: BL10.89F universal forward with BL11.36R universal reverse primers (493bp) containing rs62342744 C/T and BL12.37F universal forward with BL12.96R universal reverse primers (604bp) containing rs55825617 A/G using PCR Biosystems *Taq* at 0.03U 0.003U/ μ l with TA 55°C in 36 cycles was developed and carried out on 95 B-ALL Cord Blood control samples. PCR products from 8 samples are shown here. b) SNP Genotyping of rs62342744. In a 3M TMAC-based hybridisation solution, the PCR products on the dotblots were incubated with non-labelled competitor ASOs carrying alternative alleles for the SNPs and then hybridised with the γ - 32 P-labelled ASO probes to enforce specific binding of the correct version of the ASO to the PCR product. Hybridisation was repeated for the alternative competitor and labelled probes. The dotblot signals were visualising by exposing to a X-ray film sandwiched between dotblot and a phosphor intensifier screen stored at -80°C for 20hrs. The genotypes were scored for each donor where a signal for one or the other allele indicated that the individual was homozygous for an allele and signals for both alleles indicated that the individual was heterozygous for the SNP.

5.2.3 Testing for association between K-ZnF and B-ALL

OR for case-control samples in cancer studies is a ratio representation of two propositional odds, for example, the odds of a specific SNP allele occurring in an exposed or patient group and the odds of the same allele occurring in a healthy control group. A high OR predicts a higher likelihood that the allele would be found in case individuals. This in turn implicates the SNP as having a notable association with the disease in question. The SNP genotyping demonstrated a significant association between the minor A allele of rs55825617, the marker for PRDM9 D, and ALL patients (OR 0.436 at 95% CI 0.2192 to 0.8659, $P=0.0178$) and also the cohort as a whole (OR 0.575 at 95% CI 0.3553-0.9307, $P=0.0243$) (Table 8).

The SNP genotyping results from all 3 SNPs were also used to categorise the genotypes into K-ZnF containing (by minor alleles of each predictive SNP) and non-K-ZnF containing genotypes. Here there was no significant relationship between K-ZnF containing alleles and the ALL cohort as a whole. However, an association found between K-ZnF containing alleles and siblings of B-ALL patients was noted (OR 0.177 at 95% CI 0.0433 to 0.7263, $P=0.0162$). The sibling individuals (N=112) is sufficient in the cohort but this needs to be further evaluated in the context of the genotyping information of the specific family units to which these siblings belong and potential similarities with the French-Canadian ALL quartet family studied by Hussin et al. (2013).

Table 8 Odds Ratios for increased susceptibility to B-ALL based on SNP haplotype

Ct marker rs6889665 T/C	N	T/T	T/C + C/C	MAF	Odds Ratio	95% CI	z-statistic	P value
British Control group	470	435	35	0.038	Ref	Ref	Ref	Ref
Patient	366	340	26	0.036	0.950	0.5611-1.6097	0.189	0.8500
Father	257	242	15	0.029	0.770	0.4124-1.4392	0.818	0.4132
Mother	301	284	17	0.030	0.744	0.4089-1.3534	0.969	0.3327
Sibling	43	42	1	0.012	0.296	0.0395-2.2149	1.186	0.2357
Total B-ALL	967	908	59	0.031	0.808	0.5235-1.2458	0.966	0.3339

L20 marker rs62342744 C/T	N	T/T	T/C + C/C	MAF	Odds Ratio	95% CI	z-statistic	P value
British Control group	508	445	63	0.063	Ref	Ref	Ref	Ref
Patient	321	283	38	0.062	0.948	0.6174-1.4570	0.242	0.8091
Father	205	182	23	0.059	0.893	0.5373-1.4831	0.438	0.6610
Mother	240	201	39	0.083	1.371	0.8891-2.1125	1.428	0.1534
Sibling	34	33	1	0.015	0.214	0.0288-1.5925	1.506	0.1322
Total B-ALL	800	699	101	0.066	1.021	0.7291-1.4287	0.119	0.9053
D marker rs55825617 G/A	N	G/ G	G/A + A/A	MAF	Odds Ratio	95% CI	z-statistic	P value
British Control group	485	447	38	0.039	Ref	Ref	Ref	Ref
Patient	308	297	11	0.018	0.436	0.2192-0.8659	2.371	0.0178
Father	163	154	9	0.028	0.687	0.3250-1.4544	0.980	0.3270
Mother	202	189	13	0.032	0.809	0.4214-1.5535	0.636	0.5245
Sibling	35	35	0					
Total B-ALL	708	675	33	0.023	0.575	0.3553 to 0.9307	2.252	0.0243
Combined Ct, L20 and D markers	N			MAF	Odds Ratio	95% CI	z-statistic	P value
British Control group	1463	1327	136		Ref	Ref	Ref	Ref
Patient	995	920	75		0.795	0.5927-1.0675	1.525	0.1273
Father	625	578	47		0.793	0.5615-1.1211	1.312	0.1896
Mother	743	674	69		0.999	0.7369-1.3540	0.007	0.9943
Sibling	112	110	2		0.177	0.0433-0.7263	2.405	0.0162
Total B-ALL	2475	2282	193		0.825	0.6559-1.0382	1.640	0.1011

The British ALL cohort was also categorised into different cytogenetic subtypes, namely ETV6-RUNX1+, hyperdiploid, all other abnormal, normal, no result and untested categories. For SNP rs62342744 C/T, the marker for PRDM9 L20, there was an apparent borderline difference between the mothers of patients of normal subtype patients compared to healthy controls (OR 2.649, $P=0.0501$) but given that the 95% CI was 0.9994-7.0201 i.e. passed through 1, this result was not significant (Table 9). Furthermore, when the previously described phasing and Sanger sequencing method was applied to a shortlist of individuals from the cohort (Table 10), the results showed that these individuals carry not only L20 but all the PRDM9 alleles in the H20 haplotype of the PRDM9-SNP haplotype network i.e. L9, L24 and L26. Of these alleles, only L20 carries a K-ZnF and of the individuals tested, only one mother (MOTHER-51.2)

whose B-ALL patient child gave a ALL normal result for cytogenetic subtype, carried an L20 allele. Conversely one mother (MOTHER-10.2) whose B-ALL patient child gave a no result for cytogenetic subtype, carried an L20 allele. The remaining tested mother carrying a L20 allele had a T-ALL patient child and cytogenetic subtypes were not tested for any T-ALL families.

Table 9 Odds Ratios for increased susceptibility to normal B-ALL subtype based on SNP haplotype for L20 markers and no result or untested for B-ALL subtype based on SNP haplotype for D marker

L20 marker rs62342744 C/T	N	C/ C	C/T + T/T	MAF	Odds Ratio	95% CI	z-statistic	P value
British Control group	508	445	63	0.063	Ref	Ref	Ref	Ref
Patient	24	23	1	0.021	0.307	0.0408-2.3138	1.146	0.2519
Father	12	10	2	0.083	1.413	0.3026-6.5960	0.439	0.6603
Mother	22	16	6	0.136	2.649	0.9994-7.0201	1.959	0.0501
Sibling	2	2	0	0.000	0.000			
Total BALL	60	51	9	0.075	1.246	0.5852-2.6551	0.571	0.5679

Table 10 PRDM9 Alleles phased by the minor T allele of SNP rs62342744

PRDM9 Allele	Allele Count	% of Sequenced Individuals
L9	3	9%
L20	8	25%
L24	16	50%
L37/L26	5	16%

It is important to note that L26 (Berg et al., 2011) which was not submitted to GenBank genetic sequence database and L37 (Hussin et al., 2013) are identical in sequence (Benson et al., 2008) with a PRDM9 ZnF array of ABCDDECFGHFQJ.

For SNP rs55825617 G/A, the marker for PRDM9 D, there was a significant difference between the total subgroup of no result and untested families compared to healthy controls (OR 0.157 at 95% CI 0.0374 to 0.6580, $P=0.0113$) (Table 11). After applying for Bonferroni correction ($\alpha/m=0.05/2=0.025$) for subtype and family, the significance held for this association between rs55825617 G/A and mothers of patients whose cytogenetic subtypes gave no result or were untested.

As for SNP rs62342744, it was hoped that phasing followed by ZnF array sequencing would be used to demonstrate the physical link between PRDM9 D alleles and the minor A allele of SNP rs55825617 within this cohort. However, during the 3-year time period from receiving the DNA samples and these particular experiments, the quality of DNA had deteriorated such that PCR amplicons of ~5kb in size that included both the SNP and the PRDM9 ZnF array could not be produced. To compensate for this, DNA samples were obtained from Prof Mark Jobling (University of Leicester) that included 3 individuals who were homozygous A for SNP rs55825617; the PRDM9 alleles of these individuals were characterised via Sanger sequencing. One Turkish individuals carried a novel PRDM9 allele thus termed L49 with the ZnF array ABCDNECFGHFIJ. Another Turkish individual carried the A allele. The remaining Norwegian individual also carried an A allele. Since PRDM9 A alleles are clustered with D alleles in the PRDM9-SNP haplotype network, the appearance of the A allele in these two individuals agrees with the network. The existence of a novel allele does raise questions about the applicability of the network for populations with more PRDM9 diversity.

Table 11 Odds Ratios for increased susceptibility to normal B-ALL subtype based on SNP haplotype for L20 markers and no result or untested for B-ALL subtype based on SNP haplotype for D marker

D marker rs55825617 G/A	N	G/ G	G/A + A/A	MAF	Odds Ratio	95% CI	<i>z-statistic</i>	<i>P value</i>
British Control group	485	447	38	0.039	Ref	Ref	Ref	Ref
Patient	61	61	0	0.000	0.000			
Father	29	29	0	0.000	0.000			
Mother	45	43	2	0.022	0.547	0.1276-2.3464	0.812	0.4169
Sibling	17	17	0	0.000	0.000			
Total B-ALL	152	150	2	0.007	0.157	0.0374-0.6580	2.532	0.0113

5.2.4 FIGNL1, CDKN2A and MHCII SNPs

The four FIGNL1 SNPs were genotyped in the British B-ALL cohort and controls (Table 12). Since, SNP rs112666980 G/T was monomorphic for ALL patients and family members, no further analysis was done. No significant relationship with the B-ALL families as a whole or family groups were found with the remaining SNPs rs62445870 G/A, rs61735234 C/A and rs10235371 G/A.

Table 12 Odds Ratios for increased susceptibility to B-ALL based on SNP haplotype

FIGLN_1 rs62445870 G/A	N	G/G	G/A + A/A	MAF	Odds Ratio	95% CI	<i>z-statistic</i>	<i>P value</i>
British Control group	468	451	17	0.018	Ref	Ref	Ref	Ref
Patients	367	346	21	0.030	1.610	0.8367 to 3.0986	1.426	0.1538
Father	255	244	11	0.022	1.196	0.5514 to 2.5941	0.453	0.6505
Mother	302	289	13	0.023	1.193	0.5710 to 2.4939	0.470	0.6383
Sibling	42	41	1	0.012	0.647	0.0840 to 4.9863	0.418	0.6761
Total BALL	966	920	46	0.025	1.326	0.7519 to 2.3400	0.976	0.3293
FIGLN_2 rs61735234 C/A	N	C/C	C/A + A/A	MAF	Odds Ratio	95% CI	<i>z-statistic</i>	<i>P value</i>
British Control group	467	465	2	0.002	Ref	Ref	Ref	Ref
Patients	365	362	3	0.004	1.927	0.3203 to 11.5924	0.716	0.4738
Father	256	254	2	0.004	1.831	0.2563 to 13.0744	0.603	0.5466
Mother	302	300	2	0.003	1.550	0.2172 to 11.0632	0.437	0.6621
Sibling	42	41	1	0.012	5.671	0.5034 to 63.8772	1.405	0.1602
Total BALL	965	957	8	0.004	1.944	0.4111 to 9.1889	0.838	0.4018
FIGLN_4 rs10235371 G/A	N	G/G	G/A + A/A	MAF	Odds Ratio	95% CI	<i>z-statistic</i>	<i>P value</i>
British Control group	466	372	94	0.107	Ref	Ref	Ref	Ref
Patients	366	277	89	0.130	1.272	0.9151 to 1.7668	1.431	0.1524
Father	256	209	47	0.092	0.890	0.6032 to 1.3131	0.587	0.5569
Mother	302	247	55	0.094	0.881	0.6089 to 1.2752	0.671	0.5025
Sibling	42	38	4	0.048	0.417	0.1451 to 1.1962	1.627	0.1037
Total BALL	966	771	195	0.105	1.001	0.7599 to 1.3183	0.006	0.9948

As before, the B-ALL families of the cohort were then categorised by cytogenetic subtype. The results showed that for SNP rs62445870 G/A, patients with the ETV6-RUNX1 + subtype had a significant relationship (OR 4.319 at 95%

CI 1.6969 to 10.9914, $P=0.0021$) and the B-ALL group as whole also had a significant relationship (OR 2.948 at 95% CI 1.3672 to 6.3552, $P=0.0058$) compared to healthy controls (Table 13). The results also showed that for SNP rs10235371 G/A, patients with the ETV6-RUNX1+subtype had a significant relationship compared to healthy controls (OR 2.102 at 95% CI 1.1194 to 3.9485, $P=0.0208$). After application of Bonferroni correction for grouping by family and subtype ($\alpha/m=0.05/2=0.025$), the odds were still significant in all three cases.

Table 13 Odds Ratios for increased susceptibility to ETV6-RUNX1+ B-ALL subtype based on SNP haplotype

FIGLN_1 rs62445870 G/A	N	G/ G	G/A + A/A	MAF	Odds Ratio	95% CI	z-statistic	P value
British Control group	468	451	17	0.018	Ref	Ref	Ref	Ref
Patients	50	43	7	0.070	4.319	1.6969 to 10.9914	3.070	0.0021
Father	30	28	2	0.033	1.895	0.4169 to 8.6132	0.827	0.4080
Mother	33	30	3	0.045	2.653	0.7363 to 9.5591	1.492	0.1357
Sibling	7	7	0	0.000	0.000			
Total BALL	120	108	12	0.050	2.948	1.3672 to 6.3552	2.758	0.0058
FIGLN_4 rs10235371 G/A	N	G/ G	G/A + A/A	MAF	Odds Ratio	95% CI	z-statistic	P value
British Control group	466	372	94	0.107	Ref	Ref	Ref	Ref
Patients	49	32	17	0.184	2.102	1.1194 to 3.9485	2.311	0.0208
Father	30	23	7	0.117	1.204	0.5017 to 2.8916	0.416	0.6772
Mother	33	27	6	0.091	0.879	0.3529 to 2.1916	0.276	0.7827
Sibling	7	7	0					
Total BALL	119	89	30	0.130	1.334	0.8324 to 2.1378	1.198	0.2311

In the other abnormal subtype, several moderate associations were observed (Table 14) but after application of Bonferroni correction for grouping by family and subtype ($\alpha/m=0.05/2=0.025$), a significant association was only observed with the siblings in B-ALL families who carried the minor A allele of SNP rs10235371 G/A. However, due to the low power of 7 individuals, further evaluation is required to explain the association and the implications it has for development of childhood ALL.

Table 14 Odds Ratios for increased susceptibility to other abnormal B-ALL subtypes based on SNP haplotype

FIGLN_2 rs61735234 C/A	N	C/ C	C/A + A/A	MAF	Odds Ratio	95% CI	<i>z</i> -statistic	<i>P</i> value
British Control group	467	465	2	0.002	Ref	Ref	Ref	Ref
Patients	76	74	2	0.013	6.284	0.8717 to 45.2980	1.824	0.0682
Father	53	53	0	0.000	0.000			
Mother	67	66	1	0.007	3.523	0.3150 to 39.3911	1.022	0.3066
Sibling	8	7	1	0.063	33.214	2.6889 to 410.2776	2.731	0.0063
Total BALL	204	200	4	0.010	4.650	0.8448 to 25.5938	1.766	0.0774

FIGLN_4 rs10235371 G/A	N	G/ G	G/A + A/A	MAF	Odds Ratio	95% CI	<i>z</i> -statistic	<i>P</i> value
British Control group	466	372	94	0.107	Ref	Ref	Ref	Ref
Patients	76	54	22	0.151	1.612	0.9349 to 2.7804	1.718	0.0858
Father	53	44	9	0.085	0.809	0.3816 to 1.7169	0.551	0.5816
Mother	67	58	9	0.075	0.614	0.2936 to 1.2842	1.295	0.1952
Sibling	8	4	4	0.250	3.957	0.9717 to 16.1170	1.920	0.0549
Total BALL	204	160	44	0.113	1.088	0.7272 to 1.6286	0.411	0.6808

SNP rs148661414_A/G was monomorphic for G allele for all tested (176 B-ALL patients, 130 T-ALL patients, 92 control individuals) and SNP rs146179135C/G was monomorphic for C allele for all tested (151 B-ALL patients, 135 T-ALL patients, 154 control individuals). Hence, no further analysis could be carried out with British ALL cohort individuals with respect to these markers.

5.2.5 Imputation of FIGNL1 SNP genotypes in a German cohort

To test the relationship between childhood ALL and the 4 FIGNL1 coding variants in an independent ALL cohort, a SNP genotype dataset of a German ALL cohort was provided by Prof Martin Stanulla (Hannover Medical School), who also provided associated phenotypic information, namely cytogenetic subtype (Ellinghaus et al., 2012).

Suitable haplotype reference panels and software tools were sought to analyse the German ALL cohort SNP dataset. The High Performance Computing cluster at the University of Leicester was used to conduct all data processing. SHAPEIT v2 is a software program used for estimating haplotypes from sequencing or genotyping data, developed by Dr Olivier Delaneau, CNAM and University of Oxford (Delaneau, Marchini and Zagury, 2011). The program was used to phase genotype data of the four FIGNL1 SNPs previously obtained from the British ALL cohort by experimental methods. The phased British dataset would be as a reference haplotype panel to complement a larger phased reference panel for chromosome 7 from the publicly available 1000 Genomes data (1000 Genomes Project Consortium et al., 2015). Input files in the Oxford gen/sample format were created from the Excel spreadsheet provided by Prof Martin Stanulla. All duplicates, individuals with missing data on any one of the four SNPs and genotypes with ambiguous scores were removed in the process. Where necessary, SNP reference and alternate alleles were flipped to the positive strand. A strand file containing information on the positive strand status of the SNPs was made manually. SHAPEIT has a 'check' mode that identifies individuals or SNP loci that have more than 5% missing data and SNPs where the genotyping shows that the SNP is monomorphic for all individuals in the sample. The SNP rs112666980 G/T (FIGNL1_3) was removed prior to the SHAPEIT check as it was monomorphic and therefore would not be informative in either phasing or imputing genotypes. SHAPEIT then phased the SNP genotype data against a reference haplotype panel consisting of 1000 Genomes Pilot + HapMap 3 (Jun 2010 / Feb 2009) Chromosome 7 (haps/legend file format) and a Chromosome 7 recombination map.

5.2.5.1 Imputation model

The 1000 Genomes Pilot + HapMap 3 (Jun 2010 / Feb 2009) CEU data was chosen as a reference panel for three reasons. It is the only panel available that contained haplotype data on one of the FIGNL1 SNPs (FIGNL1_4). Neither

1000 Genomes Phase I or Phase 3 datasets contained haplotype information that included any of the four FIGNL1 SNPs (i.e. these two datasets did not contain these four SNPs at all). Additionally, the German data set was based on the NCBI36/hg18 genome build. For the SNP genotypes of the British ALL cohort, the CEU (Utah Residents (CEPH) with Northern and Western European Ancestry) population group was deemed the most suitable reference population compared with the only other alternative, YRI (Yoruba in Ibadan, Nigeria) that is available for this dataset. Hence, the CEU haplotype data was used as a reference panel for phasing. The 1000 Genomes Pilot + HapMap 3 was the last panel to use the NCBI36/hg18 genome build. If later haplotype datasets had contained haplotype data on the FIGNL SNPs, then they would have provided more accurate imputation due to the richer data in terms of SNP density and number of individuals. 1000 Genomes Pilot + HapMap 3 contains haplotype data on 75,242 SNPs from 120 CEU individuals. Combining phased reference panels such as 1000Genomes (Phase I, Phase 3) and Hapmap 3 with a population specific unphased haplotypes or unphased genotypes of a target region can improve how representative imputed genotypes in a different population (exhibiting the same phenotype i.e. childhood B-ALL types) can be in terms of accuracy.

Impute2 is a genotype imputation software with haplotype estimation functionality. It was developed at the University of Oxford (Howie, Donnelly and Marchini, 2009) and uses reference haplotype data to impute missing genotypes in a study dataset. In Scenario B, Impute2 merges the two reference panels by identification of four (04) types of haplotype information. Type 0 SNPs have haplotype data in Panel 0 only and can be used for imputation. Type 1 SNPs have haplotype data in Panel 1 but may have data in Panel 0. If not, a hole-filling function of Impute simulates Panel 0 alleles. Type 1 SNPs can be used for imputation. Type 2 SNPs have haplotype data in Panel 1 and 2, but may have data in Panel 0. If not, a hole-filling function of Impute simulates Panel 0 alleles. Type 2 SNPs do not occur in the study dataset here and so cannot be

used for imputation in this study. Type 3 SNPs have haplotype data in Panel 2 only. These SNPs are not suitable for imputation. In the merging, Type 0 SNP alleles are simulated in Panel 1 and Panel 1 SNPs are simulated in Panel 0. Impute2 can produce the merged reference haplotype panel for inspection using the `-merge_ref_panels_output_ref` or `-merge_ref_panels_output_gen` options.

To improve imputation quality at the edges of the specified analysis interval, Impute2 uses a default upstream and downstream buffer interval of 250kb beyond the specified analysis interval. The analysis region was set as Chr7:50,300,000-50,600,000 which is a range of 300,000bps. Impute2 has a basic limit of 7Mb for imputation analysis range although this can be overcome using the `-allow_large_regions` option. For imputing the FIGNL1 SNP genotypes in the German ALL dataset, the phased British ALL panel and 1000 Genomes Pilot + HapMap 3 (Jun 2010 / Feb 2009) applying Scenario B of Impute2 (Fig. 27) where the larger HapMap reference panel containing one of the FIGNL1 SNPs was leveraged to merge with the phased British reference panel and then impute the genotypes of the FIGNL1 SNPs in the German dataset.

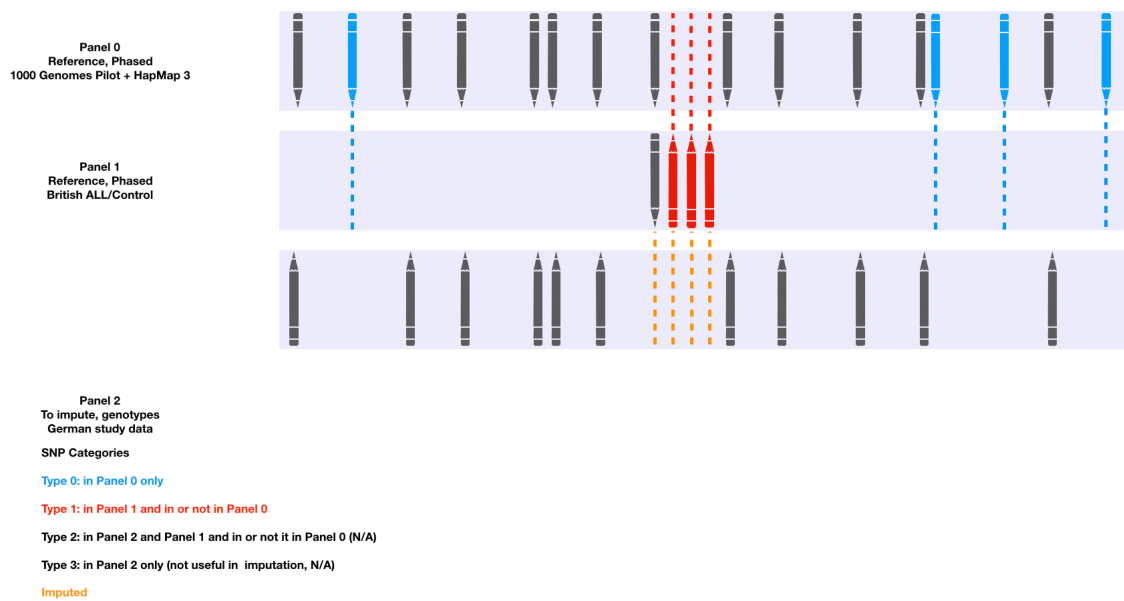


Fig. 27 The specific imputation model used on based on https://mathgen.stats.ox.ac.uk/impute/merging_reference_panels.html. This is known as Scenario B for Impute2 since there are two reference panels that need to be merged. Panel 0 depicts the 1000 Genomes Pilot + HapMap 3 reference panel containing haplotype data on 75,242 SNPs (including rs10235371 SNP) from 120 CEU individuals. Panel 1 is the British ALL cohort of 559 control individuals or 942 case and 559 control individuals combined phased by SHAPEIT. It contains all for FIGNL1 SNPs. Panel 2 is the study genotype dataset containing 419 ALL individuals and 474 control individuals from the German ALL cohort.

The dataset consisted of 18,743 bi-allelic SNPs in Chromosome 7 (Chr7:145340-158798338) for 419 ETV6–RUNX1-positive ALL patients and 474 healthy control samples (Ellinghaus et al., 2012) in the VCF file format (v4.2) generated using PLINKv1.90. There is no family information in this dataset (duo, trio, sibling). This data set did not include SNP genotyping information on the four FIGNL1 SNPs. The FIGNL1 SNPs occur in the range Chr7:50482071-50482398. The SNP density in the range Chr7:50305027-50593392 was 54 SNPs in 288 kb = 0.1875 SNP/kb. The dataset was converted into Oxford gen/sample file format using PLINKv2 for Impute2.

5.2.5.2 Imputation Results

Table 15 Odds Ratios for increased susceptibility to ETV6–RUNX1-positive ALL based on SNP haplotype

FIGLN_1 rs62445870 A/G	N	G/ G	G/A + A/A	MAF	Odds Ratio	95% CI	z-statistic	P value
Control	579	561	18		Ref			
Patients	331	311	20		2.004	1.0446-3.8458	2.091	0.0365
Father	264	253	11		1.355	0.6308-2.9110	0.779	0.4361
Mother	297	284	13		1.427	0.6892-2.9531	0.957	0.3384
Sibling	40	39	1		0.799	0.1039-6.1441	0.215	0.8294
Total B-ALL	932	887	45		1.581	0.9061 to 2.759 1	1.613	0.1068
German Ctrl (British ALL+Ctrl)	474	465	9		Ref			
German ALL (British ALL+Ctrl)	419	412	7		0.8778	0.3240-2.3781	0.256	0.7978
FIGLN_4 rs10235371 G/A	N	G/ G	G/A + A/A	MAF	Odds Ratio	95% CI	z-statistic	P value
German Ctrl (British Ctrl)	474	376	98		Ref	Ref	Ref	Ref
German ALL (British Ctrl)	419	303	116		1.469	1.092 to 2.0270	2.519	0.0118
German Ctrl (British ALL+Ctrl)	474	376	98		Ref	Ref	Ref	Ref
German ALL (British ALL+Ctrl)	419	305	114		1.434	1.0525 to 1.9540	2.284	0.0224

Results of SNPs genotyped experimentally in the British ALL cohort and SNPs imputed in the German ALL cohort were tabulated and Odds Ratios calculated to determine the strength of association between the SNPs and ALL case individuals (Table 15). In the British ALL cohort, rs62445870 A/G SNP (plus strand C/T) showed a significant relationship in B-ALL ETV6–RUNX1-positive patients and the whole family group compared to controls. For the German cohort, this relationship between rs62445870 and ETV6–RUNX1-positive patients was upheld compared to controls (OR 2.004, at 95% CI 1.0446-3.8458, $P=0.0365$). SNP rs10235371 A/G SNP (plus strand C/T) also showed a significant relationship in ALL cases (OR 1.4688 at 95% CI 1.092-2.0270, $P=0.0118$) compared to controls. When the ALL cases of the British cohort were included in Panel 1 British haplotype reference panel to enrich the data in this panel,

there was still a significant association (OR 1.4341 at 95% CI 1.0525-1.9540, $P=0.0224$)

5.3 Discussion

5.3.1 Association of D, L20 and Ct alleles with childhood ALL

SNP genotyping results demonstrate a significant association between the minor A allele of rs55825617, the marker for PRDM9 D, and the British ALL patients and the British ALL cohort (patient and family members) collectively. This suggests that PRDM9 D may have contributed to the reported excess of K-ZnF containing PRDM9 alleles in the French-Canadian and Tennessee, USA cohorts (Hussin et al., 2013). Unfortunately, it is not possible to confirm this due the sequencing methods used in that previous work where single-end short read exome sequencing data aligned to known ZnF types would not allow the assembly of entire ZnF array structure. However, since the relationship was found in patients carrying PRDM9 D alleles, the mechanism and location for causing an abnormal genomic rearrangement via NAHR cannot be explained (Berg et al., 2010) as it could not have occurred in the germline as the results do not indicate an excess of D alleles in the parental population. Hussin et al. (2013) noted an excess of K-ZnF alleles in the parental population as well and proposed that, given the current understanding of leukaemogenesis, genomic rearrangements via PRDM9-initiated NAHR might be occurring in the germline (Borel et al., 2012). The results shown here with an excess of PRDM9 D alleles in patients could mean that the rearrangement occurs in pre-leukaemic progenitor cells at the prenatal stage too.

5.3.2 Association of FIGNL1 SNPs with childhood ALL

The role of a FIGNL1-containing protein complex in homologous recombination and the role of PRDM9 allelic variants in selecting different sites

for recombination initiation provided the main context for an investigation into the coding variants of this gene found through GWAS as being associated with childhood ALL. SNP rs62445870 G/A is non-synonymous causing a Gly>Arg substitution and although a SIFT check by Dr Pamela Thompson showed that the change might be tolerated in terms of protein being able to function, PSIPRED modelling of the 2-dimensional structure showed that the change may cause a disruption of the boundary of an alpha helical domain and Polyphen also indicated that it is a damaging change (May et al, unpublished). SNP rs10235371 G/A is also a non-synonymous and deleterious mutation that was shown to have a significant association to the ETV6-RUNX1+ leukaemia. This SNP represents a Val>Met substitution. PSIPRED modelling of the 2-dimensional structure by Dr Pamela Thompson suggests that the Met coding variant disrupts a short beta-sheet between the N-terminal domain required for homologous recombination foci formation and the recruitment of RAD51. Hence, these SNPs may contribute to the poorly understood underlying genetic changes that contribute to the development of the ETV6-RUNX1+ subtype of childhood ALL.

The association between these SNPs and ETV6-RUNX1+ subtype was also reproduced in the German cohort. The imputation results were obtained using an analysis range of 300,000 bases which contained 199 SNPs with 126 SNPs upstream and 196 SNPs downstream. In the future, the imputation could be repeated over a larger range which would increase the imputation accuracy but the aim would be to check whether this has any significant effects on the imputed genotypes in the German dataset that may change the relevance of the SNPs for ALL association. Although the 1000 Genomes, Hapmap and UK10K Rare Genetic Variants in Health and Disease Projects did not contain haplotype data on the FIGNL1 coding variants, there may be other publicly available datasets that do. These resources can be sought out to conduct further imputation on the German data set including FIGNL1-3, which was monomorphic in the British ALL and so could not be used in the imputation process.

CHAPTER 6: PRDM9 ALLELE DISCOVERY USING SECOND AND THIRD GENERATION SEQUENCING

6.1 Introduction

This chapter describes studies performed to evaluate the capability of second and third generation sequencing to confidently characterising known and novel PRDM9 ZnF alleles with the overarching aim to characterise novel PRDM9 alleles from more diverse populations in the process. To this end, an Illumina HiSeq2000 100bp paired-end dataset consisting of ~456 individuals from over 30 countries/sub-populations (Batini et al., 2015) was reanalysed using bioinformatic methods. In a comparative study of the Illumina data and Sanger, Ion Torrent and MinION nanopore sequencing platforms, independent characterisation of specific PRDM9 alleles was attempted.

6.1.1 Novel PRDM9 alleles

Previous studies and this work has demonstrated that PRDM9 alleles activate different sets of recombination hotspots, this in turn altering recombination landscapes according to the PRDM9 alleles existing in the studied populations. In the case of PRDM9 A and Ct alleles, hotspot activation appears to be mutually exclusive for a particular hotspot as demonstrated with the DNA3 and AA hotspots. Whilst PRDM9 alleles of major populations such as Europeans and to an extent Africans have been sufficiently characterised (Genbank, 2008; Oliver et al., 2009; Berg et al. 2010; Baudat et al. 2010; Berg et al., 2011; Hussin et al. 2013; Oliver-Bonet, 2013), they are not representative of all world populations. Added to this, of ~458 sperm mutants found by (Jeffreys et al (2013), 95% have not been found in the wild i.e. in individuals. This suggested that there is an abundance of unsampled PRDM9 alleles in various populations. Even though PRDM9 A and Ct alleles account for 85% of hotspot activation, there remains potentially more hotspots activated by as yet uncharacterised

alleles of PRDM9. Additionally, Ct alleles are further diversified whereby they exhibit varying levels of ability to activate the same hotspots as seen in Berg et al (2011) and in Chapter 3 of this work. Also, whilst the Hinch et al (2011) data used the Ct allele associated SNP rs6889665 C allele as a marker for Ct alleles and presented PRDM9 A and Ct alleles classified via ZnF array length being ≤ 14 and ≥ 14 ZnFs respectively, this does not preclude the possibility that non-A and non-Ct alleles with ≤ 14 and ≥ 14 ZnFs are activating all African-enriched hotspots. Hence, to gain a fuller understanding of differentiated recombination landscapes in more diverse populations and sub-populations, there is a need to further explore PRDM9 diversity. It may also uncover links to genomic instability as observed with CMT1A/HNPP-associated rearrangements and the PRDM9 A-promoted hotspot (Berg et al., 2010).

6.1.2 Sanger versus NGS platforms

For over a decade, Sanger sequencing has been the standard method used to characterise PRDM9 alleles via their ZnF array configurations (Berg et al., 2010). However, the impracticality of sequencing ZnF arrays for large population samples using this method and the current availability of Massively Parallel Sequencing (MPS) methods and large datasets generated using these methods make it necessary to evaluate their capacity for accurate ZnF array assembly as a way to overcome this limitation of throughput in Sanger sequencing. The obvious stumbling block for most of these methods is the short read lengths (100-400bp) inherent to these platforms. However, greater coverage and read depth, and innovations such as paired-end reads have shown the potential to overcome the potential mismapping expected over tandem 84bp repetitive sequences over a ~ 1000 -2300bp stretch of ZnF arrays. Additionally, long read formats such as MinION nanopore sequencing aim to neutralise this issue completely and provide more sequence context than possible before, especially in terms of haplotyping PRDM9 alleles and genetically significant markers across chromosomes.

6.1.3 Study Aims

To aid in potential PRDM9 allele discovery, ZnF arrays from a set of individuals were characterised using the Sanger sequencing. Using these validated alleles as a guide, an Illumina HiSeq 2000 dataset containing ~456 individuals were used to remap to ~50 of the known reference PRDM9 ZnF arrays on GenBank. The remapping was visually examined, and read depth and variant site information were extracted (for each individual/PRDM9 allele or sample-reference combination in the dataset) for analysis. De novo assembly was also used as additional method of identification. Further to this, the set of alleles that were Sanger characterised from the Illumina sample set were sequenced on second generation Ion Torrent and third generation MinION nanopore sequencing platforms.

Since read length was predicted to be the major limiting factor in NGS or MPS platforms, this was the main assessment criteria and examined via read depth analysis both visually and mathematically. If read lengths were insufficient, then it was hoped that overlapping reads, platform specific innovations and use of different mapping tools would demonstrate the strengths and limitations of each platform. The mapping of these data to their respective alleles using various assembly methods formed the basis for a platform comparison of the sequencing methods examining the major factors such as read length which enable confident characterisation.

6.2 Results

6.2.1 Validation by Sanger sequencing

Characterisation of PRDM9 alleles was done using Sanger sequencing of the ZnF arrays as previously described (Berg et al., 2010). The method was optimised in a trial of four individuals from the sperm DNA-based SSA panel

(Appendix VII: Table S1). A ~2kb amplicon spanning the ZnF array was generated. The samples were run on a 0.5xTBE gel and the correctly-sized bands were excised to isolate them from any collapsed PCR products, unincorporated primers, dNTPs and primer dimers (Appendix VII: Fig. S1). The gel extracts were purified using a Zymoclean kit and used to generate sequencing templates using PN0.6F forward and PN0.2.5R reverse primers (Appendix I) in two separate reactions delivering 700-950bp long reads that created sufficient overlap for reconstructing the full ZnF array post-sequencing. The forward and reverse Sanger sequencing reads for each of the four individuals were trimmed to define the boundaries of interpretable reads and minor editing was carried out. Sequence alignment was done on a mixture of individual ZnF and ZnF mini-motif reference sequences to facilitate the full assembly of the ZnF arrays. A total of 8 alleles including 4 A alleles, 1 C allele, 1 L4 allele, 1 L8 allele and 1 L22 alleles which ranged in size from 13 to 18 ZnF repeats were successfully sequenced and their array structures assembled.

In order to evaluate the capability of the Illumina HiSeq2000 NGS data to accurately determine the ZnF structures of PRDM9 alleles, DNA samples from 5 individuals used in the Batini et al. (2015) study were kindly made available by Dr Pille Hallast from Professor Mark Jobling's group at the University of Leicester for traditional Sanger sequencing. ZnF arrays were characterised by the optimised Sanger method. From the 10 alleles sequenced, 5 A alleles, 2 C alleles and 1 L4 allele and 1 L7 allele were obtained along with a novel 15-ZnF allele in an individual from the BiAka Pygmy population, a nomadic sub-Saharan African people (Table 16).

Table 16 PRDM9 alleles carried by a selection of individuals from rarely sampled populations

Origin	allele 1	allele 2	ZnF array 1	ZnF array 2	allele 1 Repeat No	allele 2 Repeat No	allele 1 Myer's match	allele 2 Myer's match
Australian Aborigine	A	L7	ABCDDECFG HFIJ	ABCDDCCF GHFJ	13	12	8	7
Australian Aborigine	A	A	ABCDDECFG HFIJ	ABCDDECF GHFIJ	13	13	8	8
BiAka Pygmy	L47 (novel)	C	ABCDDCCFK HLHOIJ	ABCDDCCF KHLHIJ	15	14	5	5
Kung-speaking San	C	A	ABCDDCCFK HLHIJ	ABCDDECF GHFIJ	14	13	5	8
Palestine	L4	A	ABCDDCCCD DCFKHLHIJ	ABCDDECF GHFIJ	18	13	5	8

This novel PRDM9 allele, named L47 hereon, was composed of 15 repeats: ABCDDCCFKHLHOIJ (Fig. 28). Except for the addition of an O-ZnF, this newly found L47 allele is identical to the C allele (ABCDDCCFKHLHIJ). Interestingly, a sperm mutant with the same ZnF array configuration has also been described (Jeffreys et al, 2013).

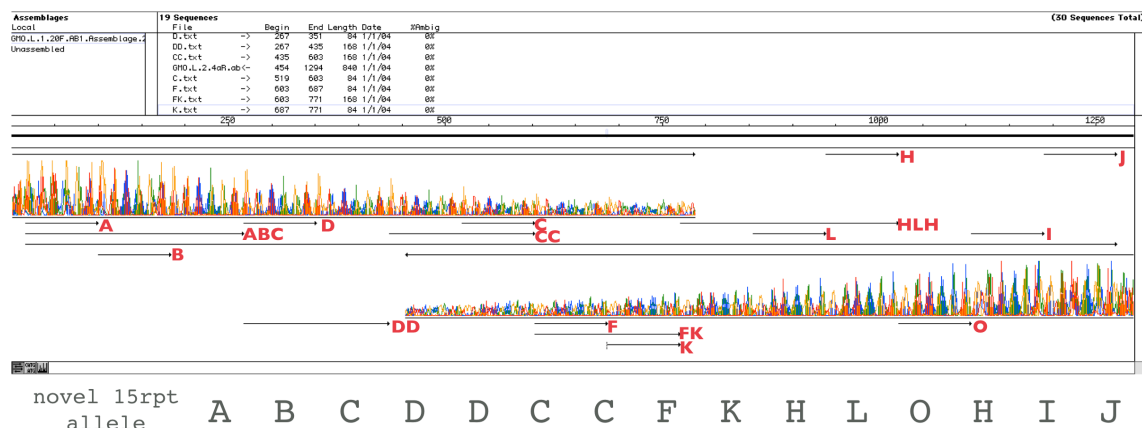


Fig. 28 Proposed assembly of a novel 15-repeat ZnF array of a Kung-speaking San 'Bushmen' individual using ABI AutoAssembler v1.4.0. A mixture of individual ZnF and ZnF mini-motif reference sequences were used to locate the known ZnFs and create a full assembly of the ZnF array from the two Sanger reads for each individual.

When motif binding was predicted according to Berg et al. (2010), this allele had a 5/8 match for the 13bp motif but also only a 6/7 match for the 16bp motif. Both H- and O-ZnFs, which are adjacent to each other, are predicted to bind to the same bases (Table 17). This allele needs to be tested in known hotspots to determine its regulatory influence. Although not covered in this work, investigating binding affinity to hotspot motifs may also advance our understanding on ZnF array sequence recognition. Its discovery in an individual alludes to the possibility that the alleles matching PRDM9 gene sperm mutants (Jeffreys et al., 2013) may be found in contemporary populations.

Table 17 Predicted binding of novel 15-ZnF PRDM9 allele to 16bp motif in comparison to the C allele

C	A	B	C	D	D	C	C	F	K	H	L	H	I	J	Match
Predicted binding	.G.	.Ca	...	c..	c..	CC	gc.	gt.	..C	gt.	.CC	g..		7/7
16bp motif								CC	nCn	tn	nnC	ntn	nC		
L47	A	B	C	D	D	C	C	F	K	H	L	H	O	I	J
Predicted binding	.G.	.Ca	...	c..	c..	CC	gc.	gt.	..C	gt.	gt.	.CC	g..	6/7
16bp motif								CC	nCn	tn	nnC	ntn	nC		

PRDM9 ZnF arrays in these five individuals were determined by Sanger method in order to serve as means of validating the interpretation of data generated by both second and third generation sequencing platforms. A comparison was made between data obtained via Illumina sequencing, Ion Torrent sequencing and MinION nanopore sequencing, as discussed below.

6.2.2 Illumina dataset remapping

This investigation was a collaboration with Dr Pille Hallast from Professor Mark Jobling's group at the University of Leicester mainly exploring the potential use of an existing Illumina HiSeq2000 NGS dataset to uncover novel PRDM9 ZnF arrays in populations of interest including Australian aborigines and

rarely sampled African ethnic groups such as Pygmies as listed in Chapter 2. Illumina sequencing offers high fidelity short reads and high read depth which has the potential to mitigate the difficulties associated with building a consensus assembly of ZnF arrays containing 84bp long ZnFs with many instances of identical ZnFs occurring in tandem. This dataset offered 100bp paired end reads where the two ends of the same DNA fragment were sequenced up to ~100bp providing an artificially longer read that facilitates mapping uniquely to the correct positions along the reference especially in the case of repetitive regions such as ZnF arrays.

IGV was used to locate the PRDM9 ZnF array region as one of the functional aspects of the IGV is to assess read length and coverage along the ZnF array region. Due to the short reads of Illumina data, even shorter length of individual ZnFs (84bp) and a lack of sufficient overlap of reads to enable the ZnFs to be aligned in the right order, *de novo* assembly was inspected as an alternate approach to mapping against the hg19 reference genome.

Since the PRDM9 ZnF arrays of five individuals included in the dataset were determined by Sanger sequencing as a means of validation so that the Illumina data for these individuals could be mapped to the correct references as a additional validation of the mapping pipeline developed by Dr Pille Hallast. The aim was to first check whether the Illumina data mapping agreed with the characterisation by Sanger sequencing. In IGV, individual ZnFs subsequently were located using the motif finder tool. The overlap of reads, uniformity of read depth and overall coverage were examined visually.

During the initial assembly of the dataset, the Stampy mapping pipeline was found to force-map the ZnFs in the reads against the hg19 reference genome which contains PRDM9 B allele. In contrast, Sanger sequencing showed that these individuals carried A, L4, L7, C and the novel L47 which had different orders and types of ZnFs.

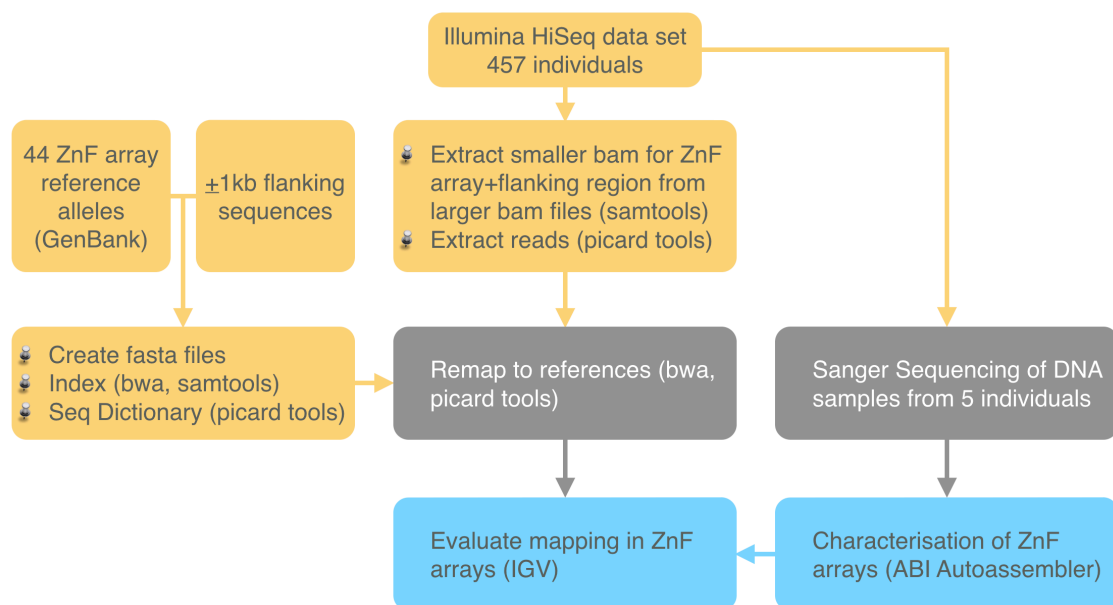


Fig. 29 Remapping pipeline for the Illumina HiSeq2000 NGS dataset. The ZnF array region plus ~1kb flanking (chr5:23,525,740-23,528,905, hg19 human genome assembly) were extracted from the whole genome into smaller BAM files. The reads in the BAM files were then extracted and remapped to 44 alleles (including ~1kb flanking region on either side) from GenBank. The remapped PRDM9 alleles were visualised on IGV. Sanger sequencing of the ZnF arrays of five individuals were used to evaluate the performance of the remapping.

Dr Pille Hallast developed a data analysis pipeline (Fig. 29) to process the Illumina HiSeq2000 dataset for 456 individuals which had already been mapped to the hg19 human genome assembly. This assembly contained the PRDM9 B allele and there was concern that much information in terms of reads may have been lost in this process. To explore this, remapping was done by Dr Pille Hallast initially to incorrect reference alleles to check whether the read quality and mapping can overcome this deficiency and it was reported that the dataset may be able to identify discrepancies via mismatches. This was replicated in this study as demonstrated below. A new pipeline was written to re-extract the reads from the larger BAM files for the region chr5:23525740-23528905 containing the ZnF array and ~1kb of flanking DNA on either side. The extracted reads were then remapped using BWA sequence alignment software to 50 of the known PRDM9 alleles on Genbank (Benson, 2008). The results of the remapping were

assessed visually on IGV (v2.3). The Sanger characterisation of the ZnF arrays of the five individuals were used as means for evaluating the remapping.

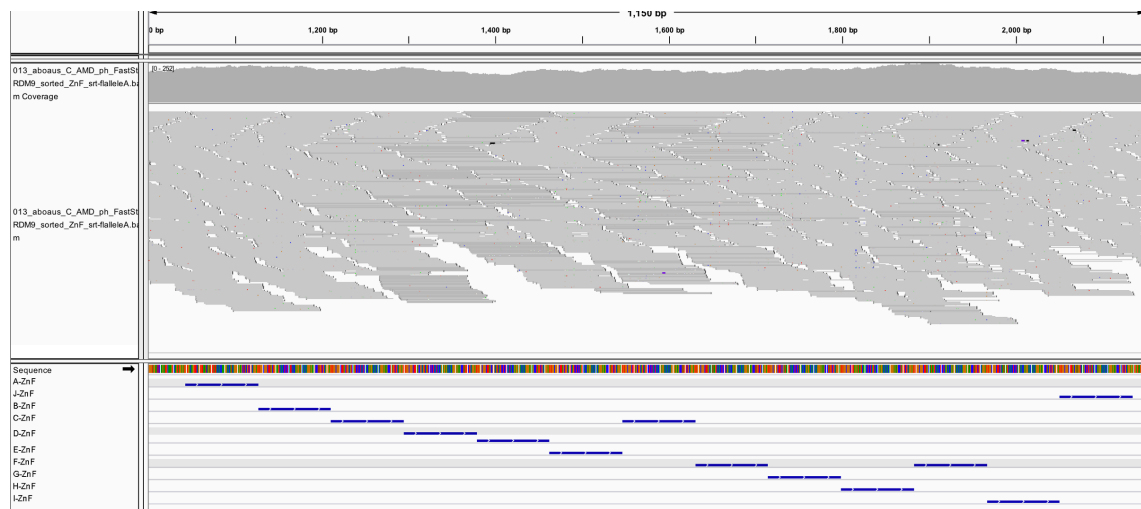


Fig. 30 IGV2.3 visualisation of the remapping of PRDM9 A ZnF array Illumina 100bp paired end sequence reads with the individual ZnFs of the array indicated below, for the Australian Aborigine individual confirmed via Sanger as carrier of A/A alleles.

For the Australian Aborigine individual homozygous for the PRDM9 A allele as validated by Sanger, the remapping shows uniform read depth and tight overlap and continuity of reads without gaps in expanded or collapsed read view (Fig. 30). There is also no signature of mismatches even at MAF of 0.2. To test the mapping accuracy of the pipeline and the quality of the reads themselves, the reads for this individual were mapped to the genome reference B allele (Fig. 31). The difference between the PRDM9 A and B allele ZnF arrays is a single E- to C-ZnF change on the 6th ZnF of the array. Also, the difference between the E- and C-ZnFs is a single G to C base substitution on the 39th position of the ZnF sequence. It is evident from this remapping that the problems are due to a percentage of E-ZnF reads being swapped or mismatched to the 7th ZnF region of the array as well as a larger percentage of E-ZnF containing reads mapping to the 6th position.

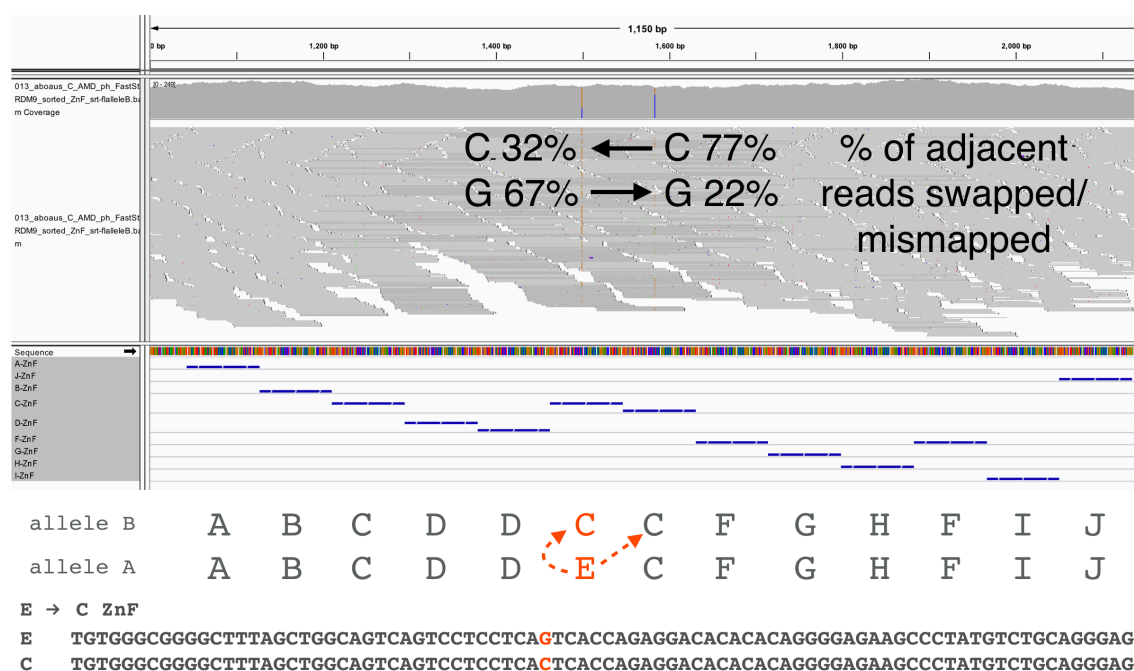


Fig. 31 IGV2.3 visualisation of the remapping Illumina 100bp paired end sequence reads of the Australian Aborigine PRDM9 A/A carried to the PRDM9 B ZnF array with the individual ZnFs of the array indicated below.

The remapping was further tested for this A/A carrier with the PRDM9 C allele as in addition to the change from E- to C-ZnF, the C allele ZnF array has different order of shared ZnFs towards the 3' end and a K-ZnF not found in the A allele. As seen with the mapping to the B allele, the E- to C-ZnF signature is seen when mapped to the C allele ZnF array (Fig. 32). More importantly here, there are two regions where there is a dramatic drop in read depth on either of the position where the H-ZnF on the 10th position is situated. It can be surmised that since the reads actually contain overlapping sequences with continuity from F- to G-ZnF but given the reference of F- to K- ZnF, the mapping drops these reads. The reads containing the first H-ZnF and enough of the adjacent G- and F-ZnF sequence that do not mismatch with the reference are able to map to the H-ZnF area. The second major drop in read depth occur for where the reference indicates a L-ZnF which has a stretch of 6bp sequence to the F-ZnF. This causes the mapping to drop the reads containing sequences in the HFI ZnF mini-motif. The observed 'stacked' reads in the position of the first F-ZnF of both PRDM9 alleles are F-ZnF containing sequences from the I-ZnF which could not

align to the FI ZnF mini-motif towards the 3' end. Similarly, the stacked reads over the penultimate I-ZnF are mostly reads that did have overlap with the F-ZnF of A allele ZnF array and shows the mismatches due to force-mapping to the HI mini-motif of the C allele ZnF array.

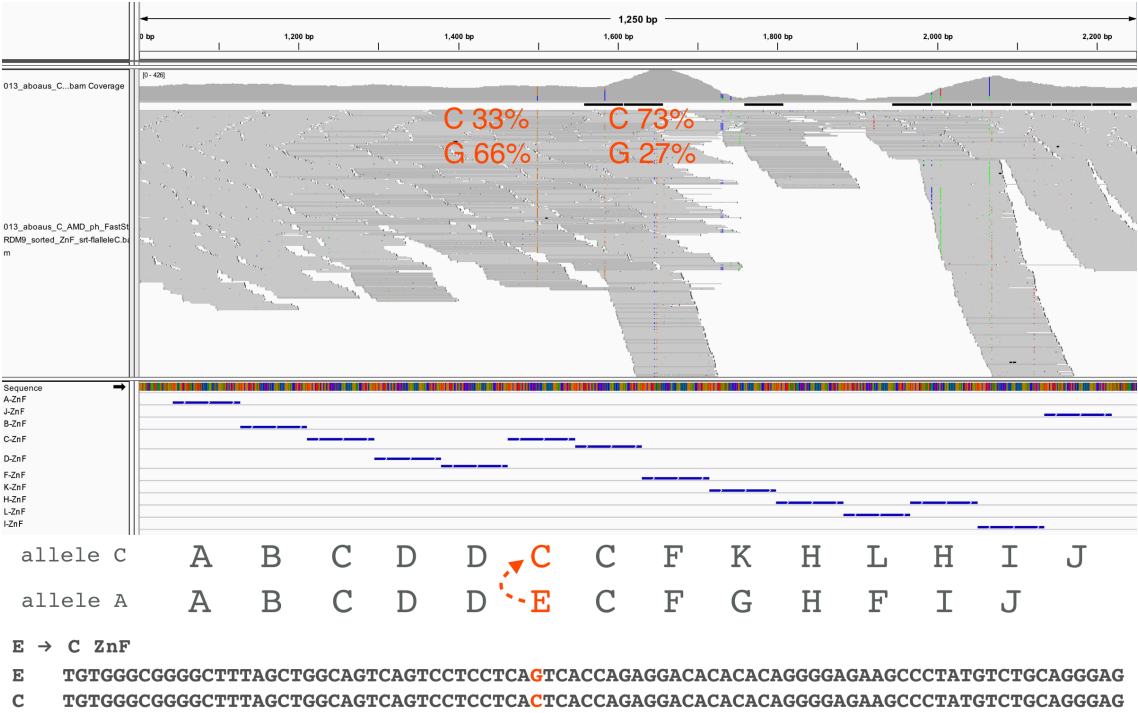


Fig. 32 IGV2.3 visualisation of the remapping Illumina 100bp paired end sequence reads of the Australian Aborigine PRDM9 A/A carried to the PRDM9 C ZnF array with the individual ZnFs of the array indicated below.

Remapping in the F-ZnF containing reads showed traces reads containing G-ZnFs (Fig. 33). The sequence of CNNG on same reads can be seen. The only other ZnF with same mini-motif is T-ZnF which is not in either A or C allele arrays. This mapping of G-ZnF may be caused due to some C-ZnFs aligning to where E-ZnF is mapped in preceding position in the array, causing a shift in subsequent ZnFs to the left by overlapping reads.

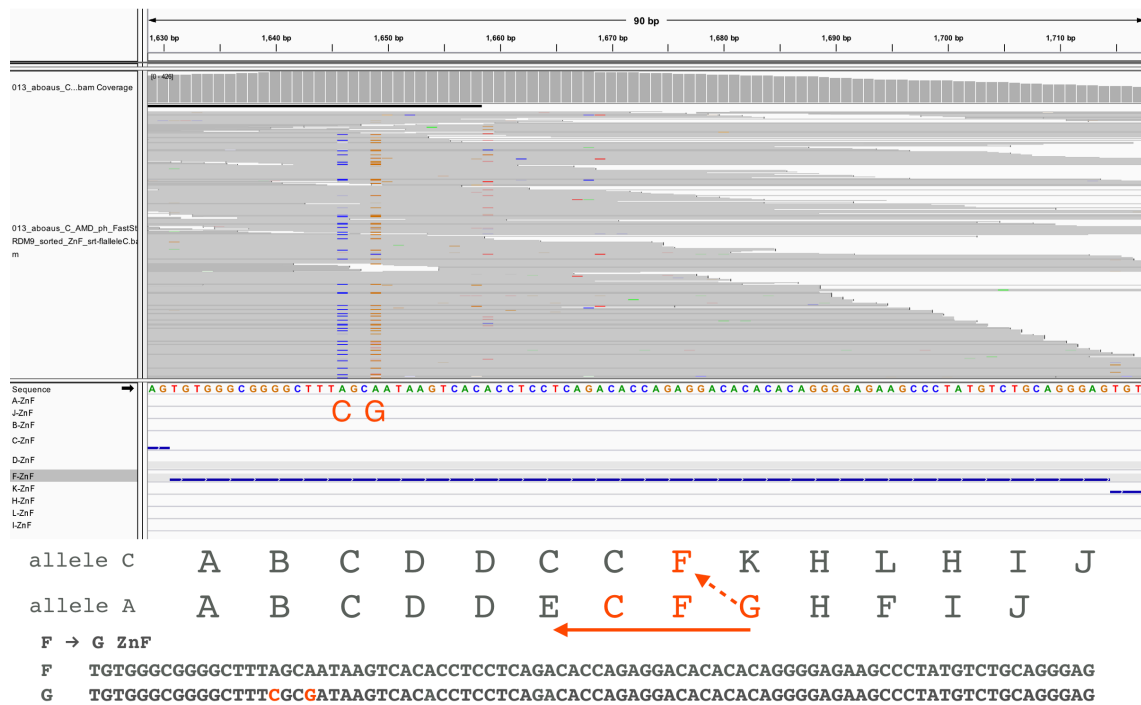


Fig. 33 IGV2.3 visualisation of the remapping Illumina 100bp paired end sequence reads of the Australian Aborigine PRDM9 A/A carried to the PRDM9 C ZnF array with the individual ZnFs of the array indicated below.

Remapping in the individual carrying the C allele and the novel 15-repeat L47 showed uniform read depth across the ZnF array when mapped to the C allele (Fig. 34). Yet since the Sanger validation confirmed this individual as being heterozygous for PRDM9 C/L47, the mapping was slightly misleading. At allele frequency threshold of 0.2, a significant mismatch signature was observed (Fig. 35). The ZnF arrays of the C and L47 alleles are ABCDDCCFKHLHIJ and ACDDDDCCFKHLOHIJ, respectively, with a single addition of a O-ZnF in L47. The difference between the O- and H-ZnFs is a single G to C base change on the 63rd position of the ZnFs. Further examination of this mismatch explained the O-ZnF containing reads mapping to the same position that the H-ZnF reads had mapped to.

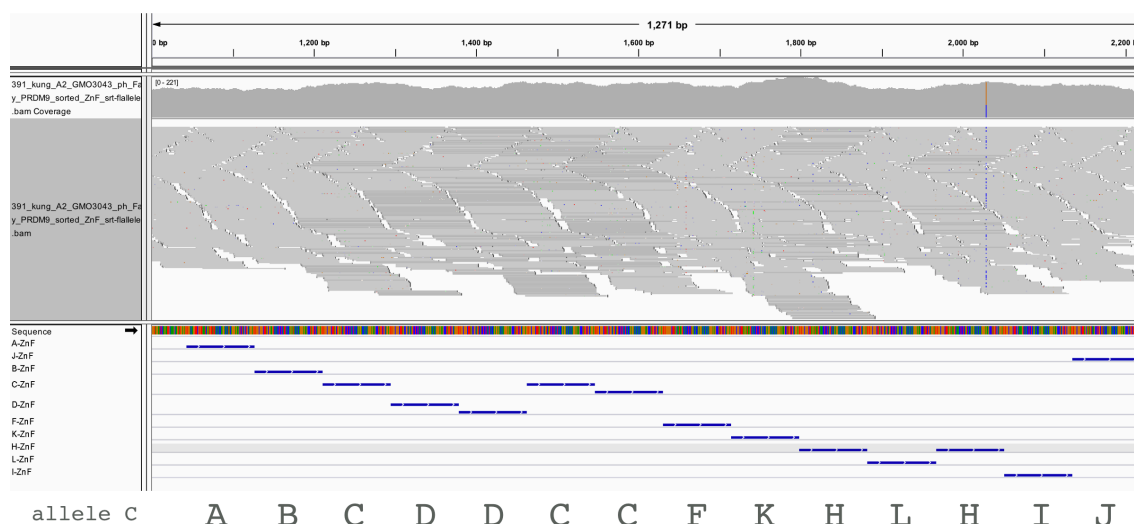


Fig. 34 IGV2.3 visualisation of the remapping Illumina 100bp paired end sequence reads of the Kung-speaking San 'Bushmen' individual carried to the PRDM9 C ZnF array with the individual ZnFs of the array indicated below.

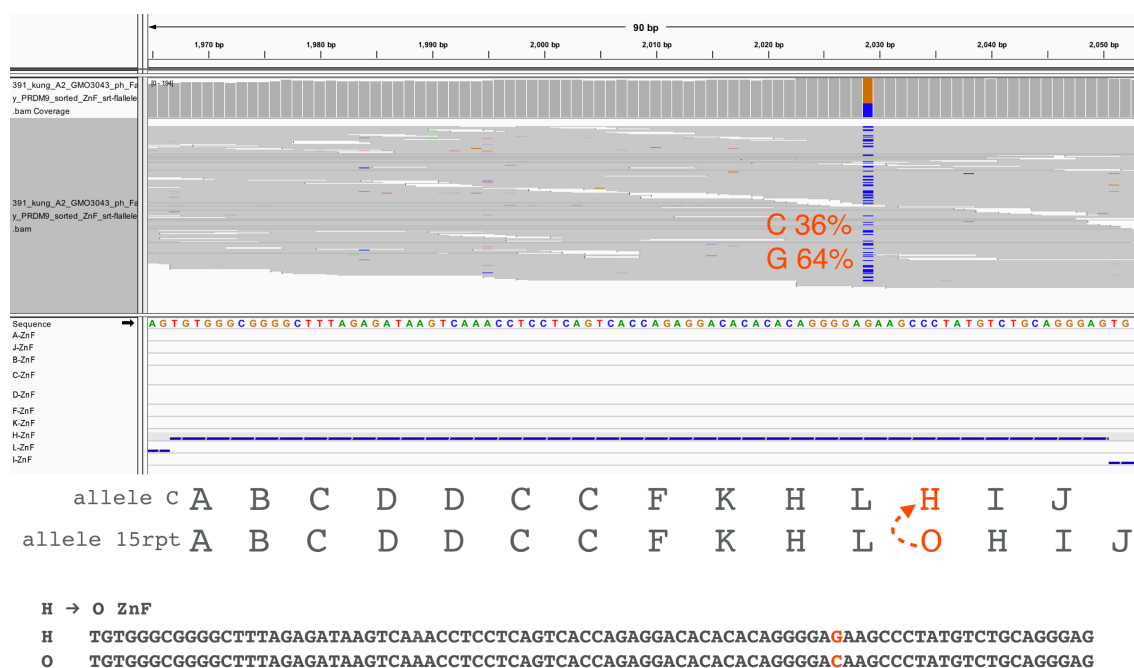


Fig. 35 IGV2.3 visualisation of the remapping Illumina 100bp paired end sequence reads of the Kung-speaking San 'Bushmen' individual carried to the PRDM9 C ZnF array with the individual ZnFs of the array indicated below. The mismatch observed is zoomed in to show the mark of O-ZnF containing reads with base C that maps to the same position as the H-ZnF containing reads with a base G in the same position.

Remapping to the L47 allele also produced similar results and also served to highlight how this pipeline can distinguish between alleles in a heterozygous

individual (Fig. 36). When mapped to L47, essentially the reverse situation arises where O- to H-ZnF position mapping occurs. Yet here, a loss of read depth can also be seen indicating the heterozygosity in repeat length since the reference L47 is one repeat longer than the 14-repeat C allele.

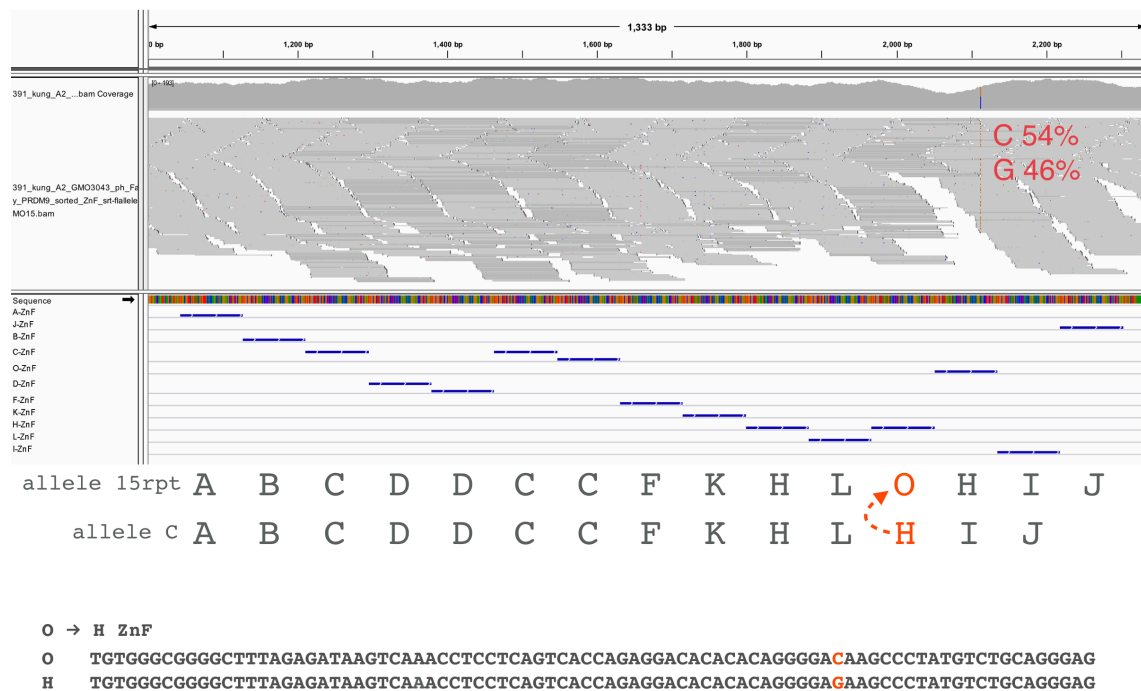


Fig. 36 IGV2.3 visualisation of the remapping Illumina 100bp paired end sequence reads of the Kung-speaking San 'Bushman' individual carried to the PRDM9 L47 ZnF array with the individual ZnFs of the array indicated below. The mismatch observed is shows the mark of H-ZnF containing reads with base G that maps to the same position as the O-ZnF containing reads with a base C in the same position.

Remapping of the Palestinian PRDM9 A/L4 carrier also gave insight into the quality of the dataset. When mapped to the A allele, even though there is no complete loss of read depth, there is variation in read depth with stacked reads presumably from the L4 ZnF array containing reads force-mapped to the A allele reference given (Fig. 37). The assumption can be made for ZnFs in similar positions from the alternate L4 allele trying to align. For example, it is likely that the reads containing C-ZnFs are mapping to the same positions as where the single E-ZnF reads are mapping (Fig. 38). There is only a single C to G base change on the 38th position from C- to E-ZnF. In contrast, the difference

between C- and D-ZnFs is a single T to C base substitution on the 19th position of the ZnFs but this mark cannot be seen which accounts for the loss of read depth in the second D-ZnF, the fifth ZnF in the reference A array. From the further mismatch, there is evidence of F-ZnF containing reads mapping to where the E-ZnF reads are mapped.

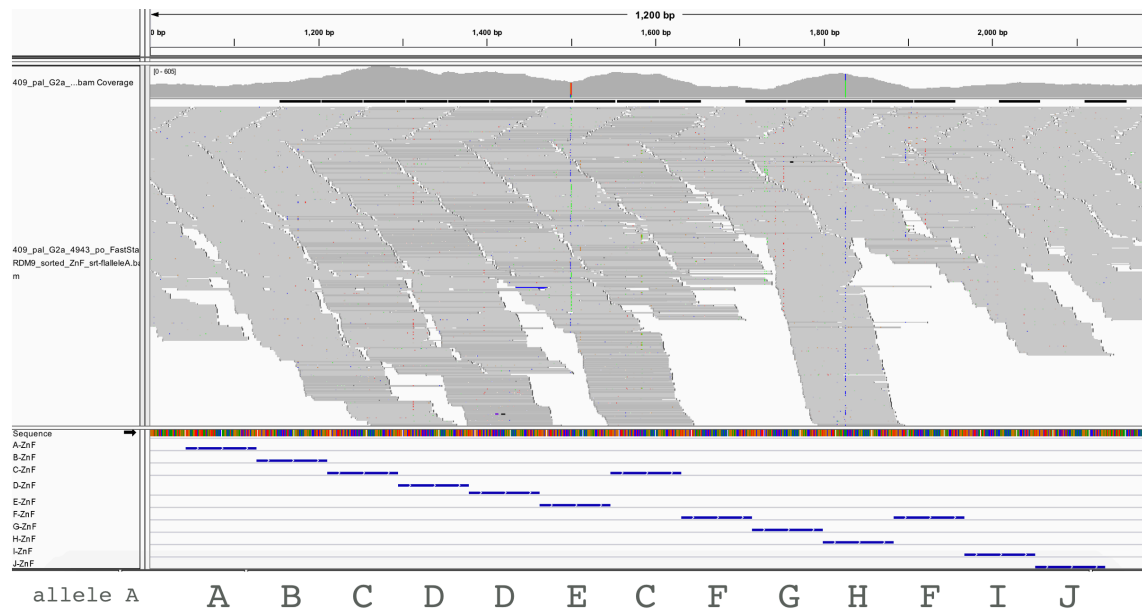


Fig. 37 IGV2.3 visualisation of the remapping Illumina 100bp paired end sequence reads of the Palestinian individual carried to the PRDM9 A ZnF array with the individual ZnFs of the array indicated below.

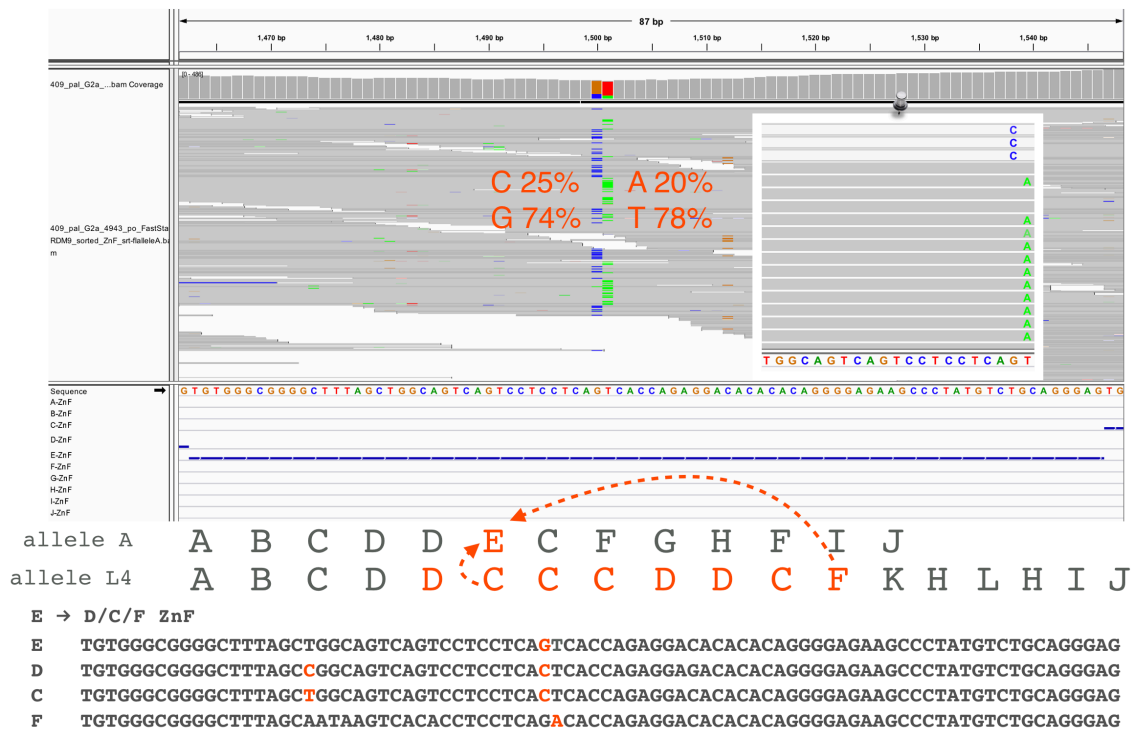


Fig. 38 IGV2.3 visualisation of the remapping Illumina 100bp paired end sequence reads of the Palestinian individual carried to the PRDM9 A ZnF array with the individual ZnFs of the array indicated below. The mismatch caused due to both C- and F-ZnFs mapping to

When the reads for this Palestinian individual were mapped to the L4 allele, there was uniform read depth across the array but a notable presence of stacked reads (Fig. 39). Here too, it was presumed that these stacked reads originated from the A allele. This was corroborated to a certain extent when the sequence of the possible flanking ZnFs were considered from the A allele (Fig. 40). For example, the CNNG mismatches appear on the same reads mapped to the area where the F-ZnF containing reads mapped. This indicates that FG mini-motifs containing reads from the A allele were mapped here. Stacked reads seem to indicate the signature for misaligned ZnFs.

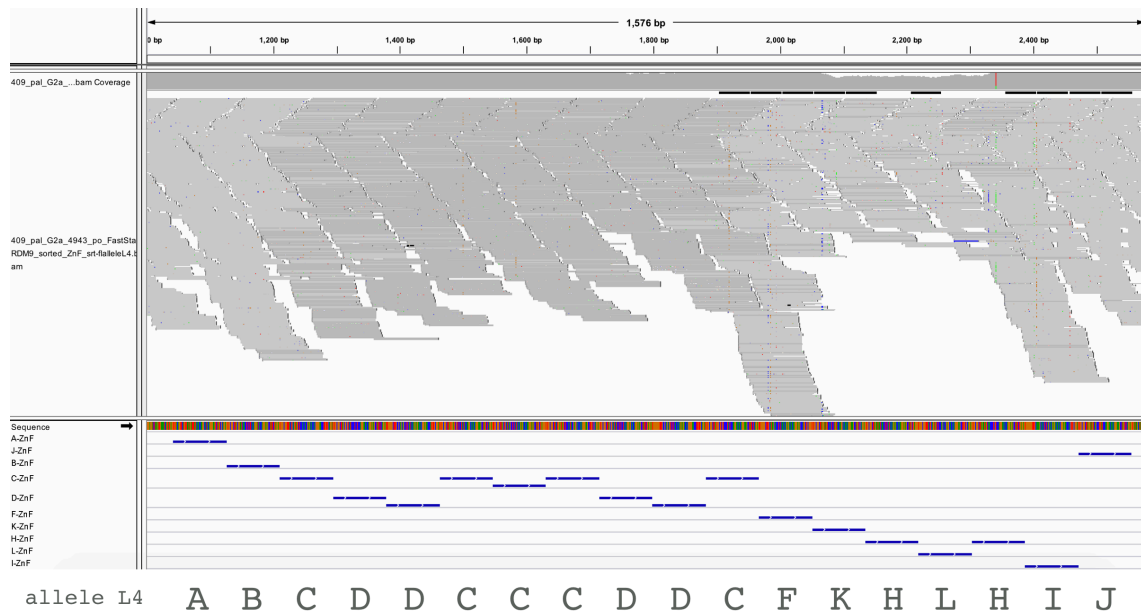
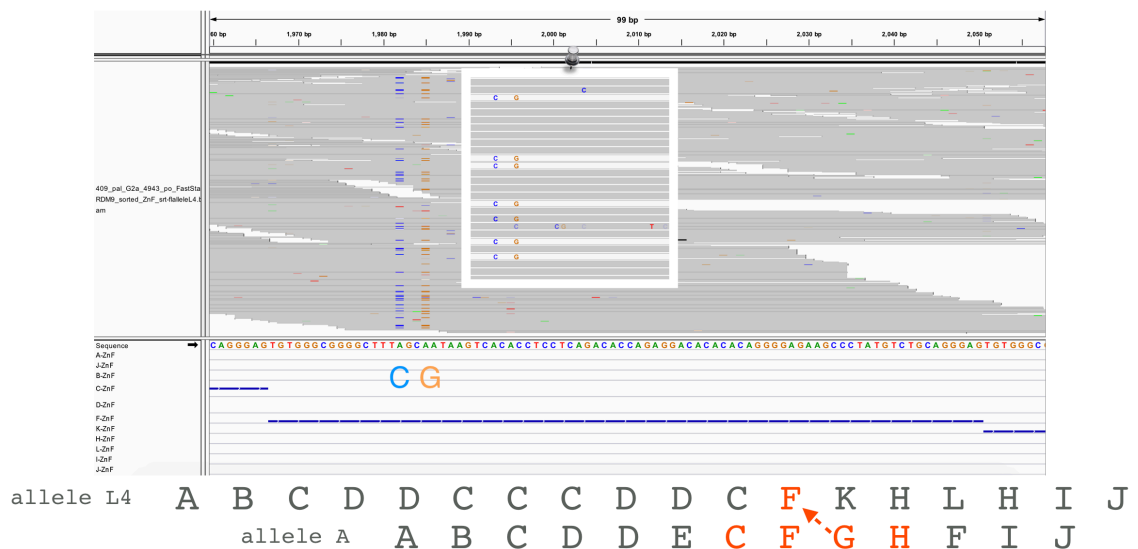


Fig. 39 IGV2.3 visualisation of the remapping Illumina 100bp paired end sequence reads of the Palestinian individual carried to the PRDM9 L4 ZnF array with the individual ZnFs of the array indicated below.



F to C/G/H ZnF

F TGTGGGCGGGGCTTTAGCAATAAGTCACACCTCCTCAGACACCAGAGGACACACACAGGGGAGAAGCCCTATGTCTGCAGGGAG
C TGTGGGCGGGGCTTTAGCTGGCAGTCAGTCCTCCTCAGTACCAGAGGACACACACAGGGGAGAAGCCCTATGTCTGCAGGGAG
G TGTGGGCGGGGCTTTCCGATAAGTCACACCTCCTCAGACACCAGAGGACACACACAGGGGAGAAGCCCTATGTCTGCAGGGAG
H TGTGGGCGGGGCTTTAGAGATAAGTCAAACCTCCTCAGTCACCAGAGGACACACACAGGGGAGAAGCCCTATGTCTGCAGGGAG

Fig. 40 IGV2.3 visualisation of the remapping Illumina 100bp paired end sequence reads of the Palestinian individual carried to the PRDM9 L4 ZnF array with the individual ZnFs of the array indicated below. Additional ZnF shows the possible ZnFs from the alternate A allele underlying the mismatch signature at MAF 0.2.

6.2.3 Ion Torrent sequencing and mapping

A selection of the PRDM9 ZnF array amplicons (PRDM9 A, C, L4 and L47) Sanger sequenced from the five individuals discussed in section 6.2.1, were sequenced using Ion PGM™ 200 Xpress™ Template and Ion PGM™ 400 Xpress™ Template kits with the Ion Personal Genome Machine® (PGM™) System, which are marketed to generate ~ 200 and ~400 bp sequencing libraries respectively. After mapping the sequence reads generated to their respective reference ZnF arrays containing ~1kb of flanking DNA sequence using BWA v0.7.10-r789, the data were viewed on IGV v2.4.19. The consensus sequence was used as a query for a BLAST search against GenBank (Benson, 2008).

Most of the reads fell below a mapping quality of 60 which BWA software caps as being of highest quality for short read data. To ensure a full view of the capability of mapping and overlapping reads, the amount of reads in the coverage were considered between 20 and 60 mapping quality to determine whether useful reads were being discarded due to being below mapping quality 60. However, this resulted in complete loss of coverage in several areas of the ZnF array. To identify the problem, sequence reads with mapping quality of 0 were also added (Fig. 41). These gave a full coverage over the length of the ZnF array although the coverage was markedly lower compared to the 5' flanking region of the ZnF array begins, especially in the case of the PRDM9 A ZnF array amplicon mapping to the reference A allele (ABCDDECFGHFIJ).

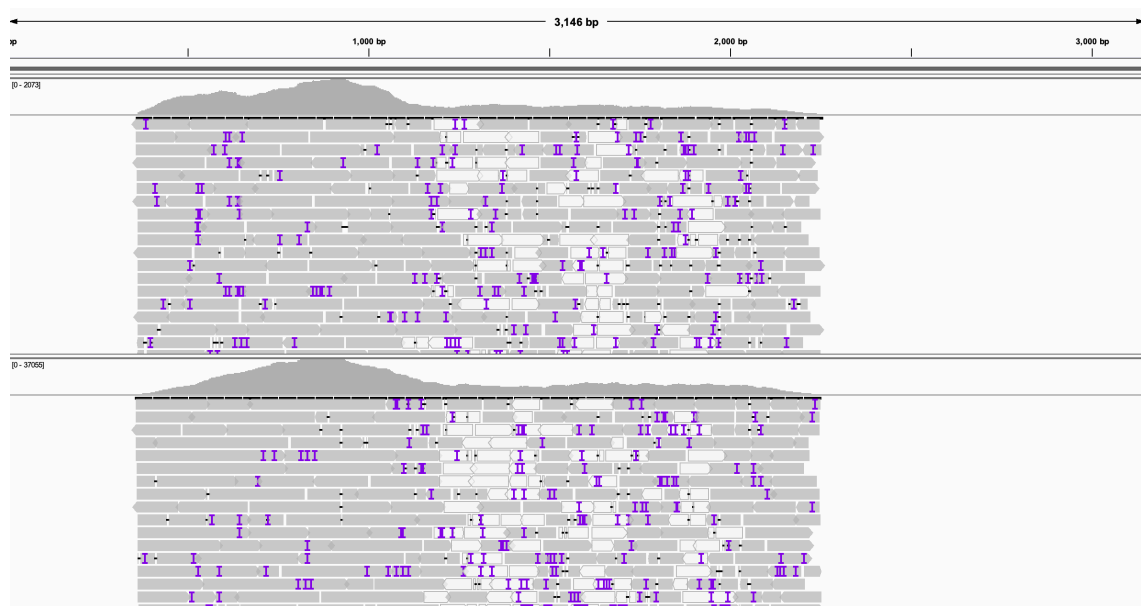


Fig. 41 IGV visualisation of the mapping of PRDM9 A ZnF array sequence reads from 200bp (top) and 400bp (bottom) to reference ZnF array (ABCDDECFGHFIJ) with ~1kb of flanking DNA. Purple vertical markings denote single base insertions; black dashes between the sequence reads (in grey for mapping quality >0) denote deletions/gaps in the sequence reads; white sequence reads have mapping quality of 0.

The loss of coverage was more pronounced for L4 (ABCDDCCCDDCFKHLHIJ) when mapping quality of reads was set at 20, with the ZnF arrangement of tandem C- and D-ZnFs at first being identified as the being the main site of complete breakdown in coverage (Fig. 42). Reads with mapping quality 0 were added back in and they were concentrated but not limited to this area as with the mapping for PRMD9 A. It can be surmised from the sequence similarity of these two ZnF types, which have a single base pair difference (C/T), that BWA has particular difficulty in reconciling this small difference in the sequence reads.

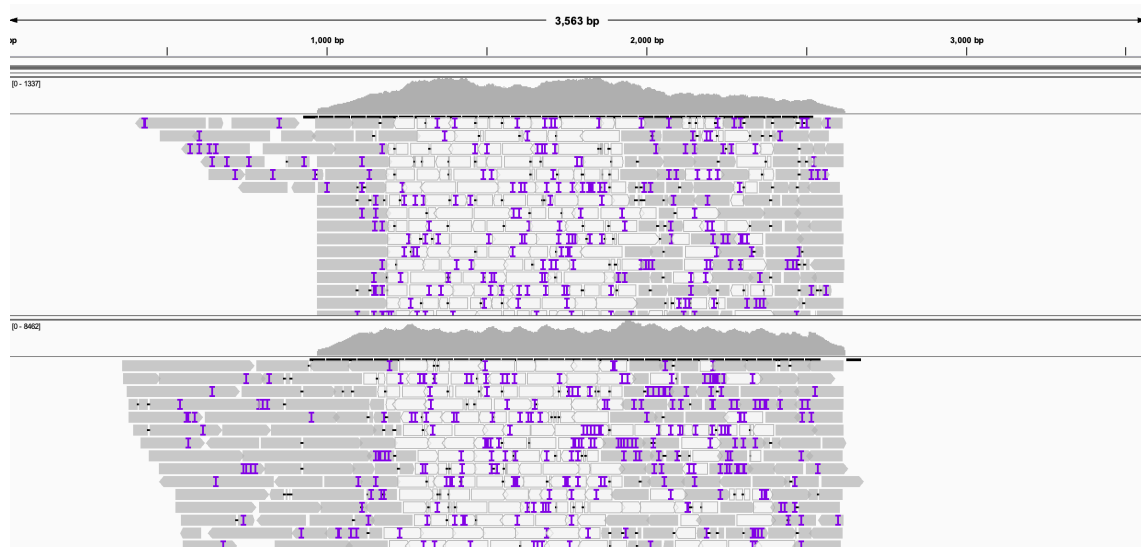


Fig. 42 IGV visualisation of the mapping of PRDM9 L4 ZnF array sequence reads from 200bp (top) and 400bp (bottom) to reference ZnF array (ABCDDCCDDCFKHLHIJ) with ~1kb of flanking DNA. Purple vertical markings denote single base insertions; black dashes between the sequence reads (in grey for mapping quality >0) denote deletions/gaps in the sequence reads; white sequence reads have mapping quality of 0.

When the individual ZnFs were located along the length of the array, it was seen that in general that both 200bp and 400bp reads delivered poor mapping within the ZnF array (Fig. 43, 200bp one shown for simplicity). An individual ZnF being 84bp and 400bp reads would essentially cover four ZnFs, it was expected that this NGS platform and kit would provide good mapping quality and coverage over the DDCCCDD motif of the ZnF array of L4. However, there was no noticeable improvement in the mapping quality over this motif with the 400bp kit.

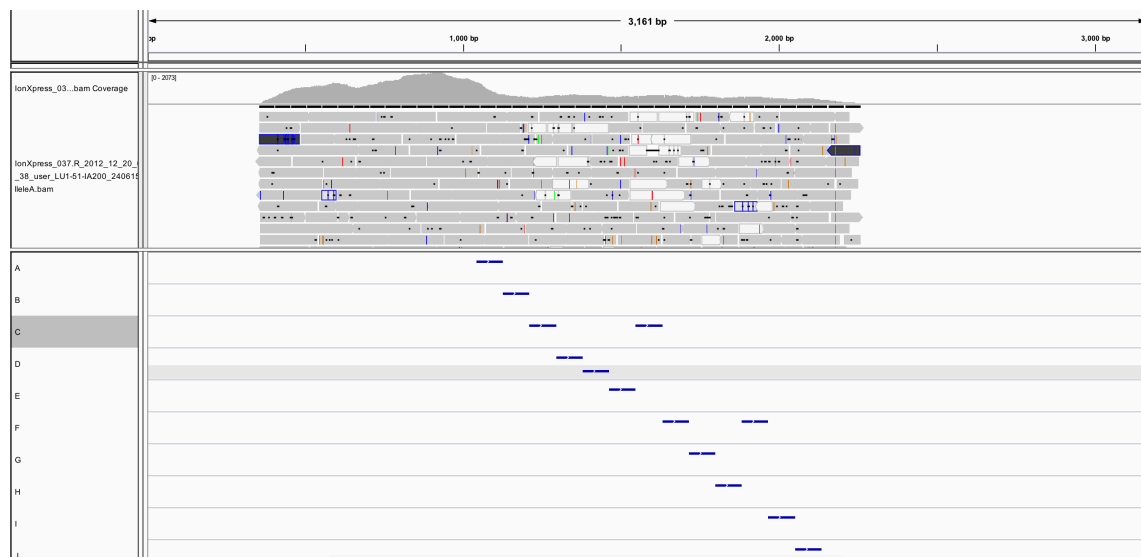


Fig. 43 IGV2.3 visualisation of the mapping of PRDM9 A ZnF array sequence reads from 200bp with the individual ZnFs of the reference ZnF array indicated below. Black dots denote single base insertions; black dashes between the sequence reads (in grey for mapping quality >0) denote deletions/gaps in the sequence reads; white sequence reads have mapping quality of 0.

Re-examination of the sequence reads showed that for the PRDM9 A allele sequencing described up to this point, the mean read lengths from the 200bp and 400bp kits were 122bp and 136bp, respectively. This probably accounted for this lack of improvement in mapping quality with added read length. The other PRDM9 A allele ZnF array used in 200bp and 400bp sequencing also had mean read lengths of 122bp and 135bp, respectively. The visual examination of the mapping showed a similar profile. It must be noted that the mean read lengths do not represent the entire picture of the quality of the reads obtained from both 200bp and 400bp version kits as maximum read lengths identified from IGV include read lengths over 250bp for 200bp kit and read lengths over 400bp for the 400bp kit. However, the longer reads were concentrated towards the 5' prime end of the reference sequence given. It is accepted that the issue may lie with inaccurate size-selection of fragments for sequencing. In fact, the fragment sizes from Bioanalyzer estimation were ~80bp for all tested samples (Appendix VII: Fig. S2).

For the mapping of the 400bp read length version of L4 allele to its reference, a drop in mapping quality of reads were observed around the DDCCCDD motif of the L4 allele (Fig. 44). In contrast to the results of the 400bp kit for the PRDM9 A allele, coverage remained fairly uniform across ZnF array. This may partly be due to the 400bp kit reads obtained have an four-fold higher number of reads compared to the 200bp kit reads for all tested alleles.

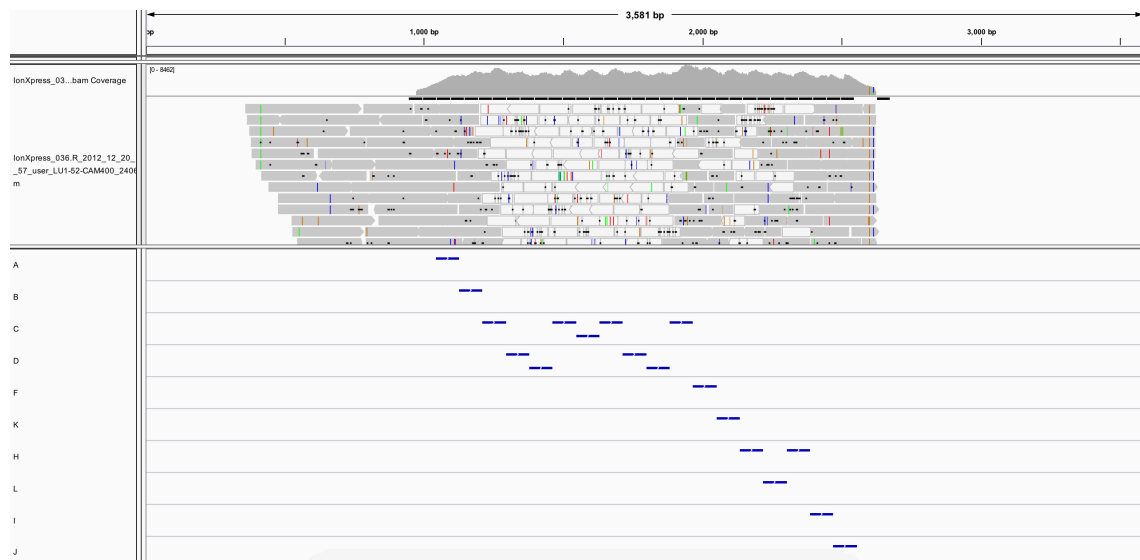


Fig. 44 IGV2.3 visualisation of the mapping of PRDM9 L4 ZnF array sequence reads from 400bp with the individual ZnFs of the array indicated below. Black dots denote single base insertions; black dashes between the sequence reads (in grey for mapping quality >0) denote deletions/gaps in the sequence reads; white sequence reads have mapping quality of 0.

6.2.4 Nanopore sequencing and mapping

The first series of MinION nanopore sequencing was done in 2015 during the time of the MinION Access Program (MAP) and used the same PRDM9 A and L4 allele amplicons that were selected from the DNA samples of individuals belonging to the Illumina HiSeq2000 dataset. The MAP004/Genomic004 kit was used to prepare the library with an input of 445ng of L4 and A allele amplicons, roughly half the recommended 1µg input amount for prepared library. The library was loaded onto a R7 flow cell and run for 24hrs with 197 active pores obtaining 6506 reads. Raw squiggle data were uploaded to Metrichor for 2D

basecalling online. Output fast5 files were downloaded with 761 pass reads and 5745 fail reads and used in a custom mapping pipeline, that integrated the LAST sequence alignment software, developed by Dr John Wagstaff at the University of Leicester. This version of LAST was configured to be less stringent towards the numerous deletions and gaps in signals observed in nanopore sequence data at the time. The reads were mapped to two types of reference sequences. The first reference type contained either L4 or A allele with flanking regions plus the Control Sequence (CS), a 3.6kb positive control added to the library preparation. The second reference type was a concatenated reference sequence containing both alleles and CS. IGV v2.3.46 was used to visualise the mapping and filter according to allele frequency and mapping quality.

The second series of sequencing was carried out in early 2019 after the MinION device had been commercialised for a few years and validated Ligation Sequencing reagent (SQK-LSK109) kit had been in use by research community. This second experiment used the same alleles but from Zimbabwean sperm donor from the SSA panel. For the L4 allele amplicon of 2319bp, 30ng or 19.93fmols were generated by PCR using PN0.5F and PN2.5R primers (Appendix 2), separated by gel electrophoresis, excised to remove collapsed product and primer dimers, and used as input for the library preparation. For the A allele of 1899bp, 95ng or 77.08fmols were used. Hence, there was a total of ~97fmols as input for library preparation. No barcoding was done as the pipeline designed was for the reads to be mapped to a concatenated reference sequence containing both L4 and A allele sequences with flanking regions and poly-N bridges. For the Ligation Sequencing kit, the manufacturer QC test guide estimated a 25% loss of DNA during end repair, dA tailing and cleanup with magnetic beads and a 33% loss of DNA during migration of adaptors. At the end of these two steps, ~24fmols of the L4 and A allele mixture were estimated to be in the prepared library. The Ligation Sequencing kit guide and other guidelines in the ONT Community recommend 100-200fmols input DNA to begin library preparation but they specifically refer to DNA fragments or PCR amplicons of

8-10kb in size. Since the L4 and A allele amplicons were ~2kb long, this total amount of prepared library in femtomols was within the recommended range of 5-50fmols input of prepared library per run for the R9.4.1 flow cell used. Secondary PCR amplicons using PN0.5F and PN2.4aR (Appendix I) were generated from the primary PCRs of the L4 (1655bp, 23.1ng, 21.51fmol) and A allele (1235bp 46.2ng, 57.64fmol) in case the experiments yielded problematic results for the L4 allele. Ideally, equal amounts of L4 and A allele prepared library would have been added to the flow cell.

The prepared library was run on the MinION sequencer with 1639 active pores within the flow cell, base calling was performed in real-time using the EPI2ME online basecaller. After 17min, 31.61k reads were obtained and when the run was stopped at 3hrs, this reached 533.26k. Basecalling was completed in ~18hrs, yielding 428k pass reads and 116k fail reads. Post-run QC check showed that the flow cell contained 1576 active pores. Both fast5 and fastq files were downloaded, each containing 4000 individual reads. The fastq files were merged using command line tools and mapped to the same reference sequences used for the 2015 sequence data using minimap2 v2.15 sequence alignment software. Samtools v1.9 was used to convert the sam files into bam format, sort the bam files and create the bai files. The results were visualised on IGV v2.4.19 with filtering according to allele frequency and mapping quality to assess and compare with the earlier dataset.

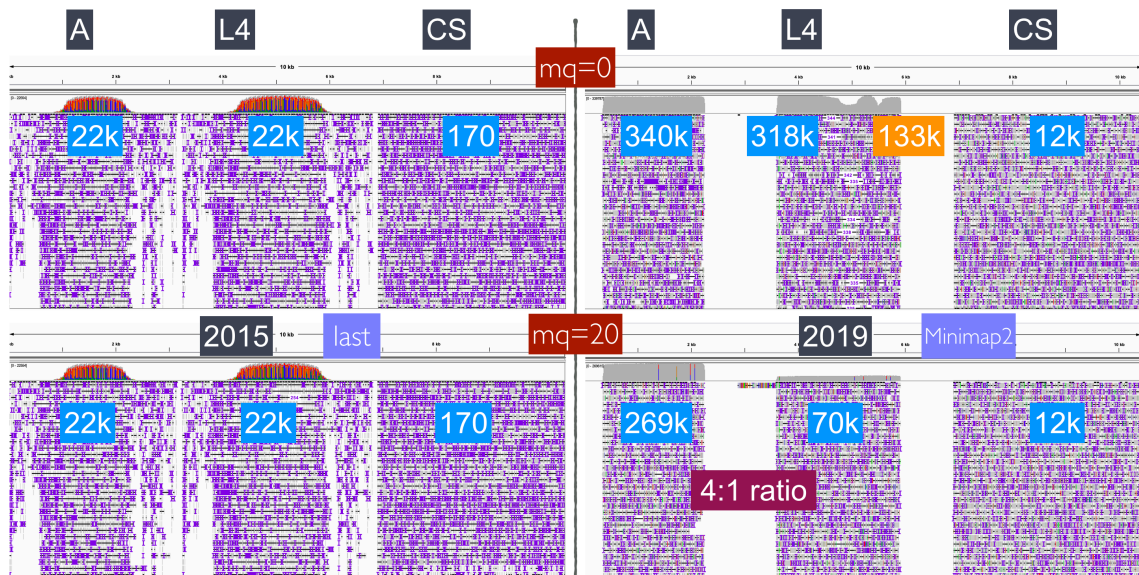


Fig. 45 IGV2.4.19 visualisation of the mapping of PRDM9 L4 and A ZnF array MinION nanopore sequence reads from MAP004 kit/R7 flow cell (2015) and Ligation Sequencing kit/R9.4.1 flow cell mapped to concatenated reference sequences containing L4 and A alleles and CS sequence. Purple vertical marking denote single base insertions; black dashes between the sequence reads (in grey for mapping quality >0) denote deletions/gaps in the sequence reads; white sequence reads have mapping quality of 0; blue labels indicate maximum read depth; dark grey labels indicate region of the reference alleles and CS; light purple labels indicate alignment software used; red labels indicate mapping quality threshold applied; magenta label indicate relative ratio of read depth between A and L4 alleles.

Although the 2015 dataset yield 761 pass reads, there was an observed 28-fold increase in the number of mapped reads (Fig. 45). This is due to LAST alignment software placing multiple instances of the same read in different positions along the concatenated reference sequence, whereas minimapp2 only mapped each read once. The results from both datasets were compared on IGV v.2.4.19 with mapping quality thresholds of 0 and 20. The change in mapping quality threshold significantly affected the 2019 data but not the 2015. This is also due to the process that LAST software applied to low quality reads. For the 2019 data, it was observed that increasing the mapping quality threshold caused the read depth for the L4 allele to decrease and the read depth across the ZnF array to become more uniform. The fall in read depth was less dramatic for the A allele and a potential reason is that the reads that were removed from the L4 allele were proportionally of poorer sequence read quality than the sequence

reads already aligned to the A allele. Another possibility is that many of the reads that previously were aligned to the L4 allele ZnF array moved to the ZnF array of the A allele. The maximum read depth change from 340k to 269k alone does not negate this possibility as this metric does not indicate the over mean read depth observed for the L4 allele and A allele. At the application of allele frequency threshold of 0.7 to both datasets, there was a significant and characteristic pattern of mismatches or mixed positions observed in the 2015 dataset. The relative amount of deletions or gaps in the reads with no real base data common in nanopore sequencing was much greater in the 2015 dataset.

To check whether the 2015 dataset could be improved via alternate sequence mapping software, BWA sequence alignment software was used (Fig. 46). This application cleared up the overall quantity of mismatches but still left a mark on the significant and repeating mismatches observed when LAST was used. The placement of single instances of sequence reads with BWA also uncovered the loss of reads in the same parts of the ZnF arrays observed with Ion Torrent mapping with BWA, particularly with identical ZnFs occurring in tandem along ZnF arrays of both L4 and A alleles.

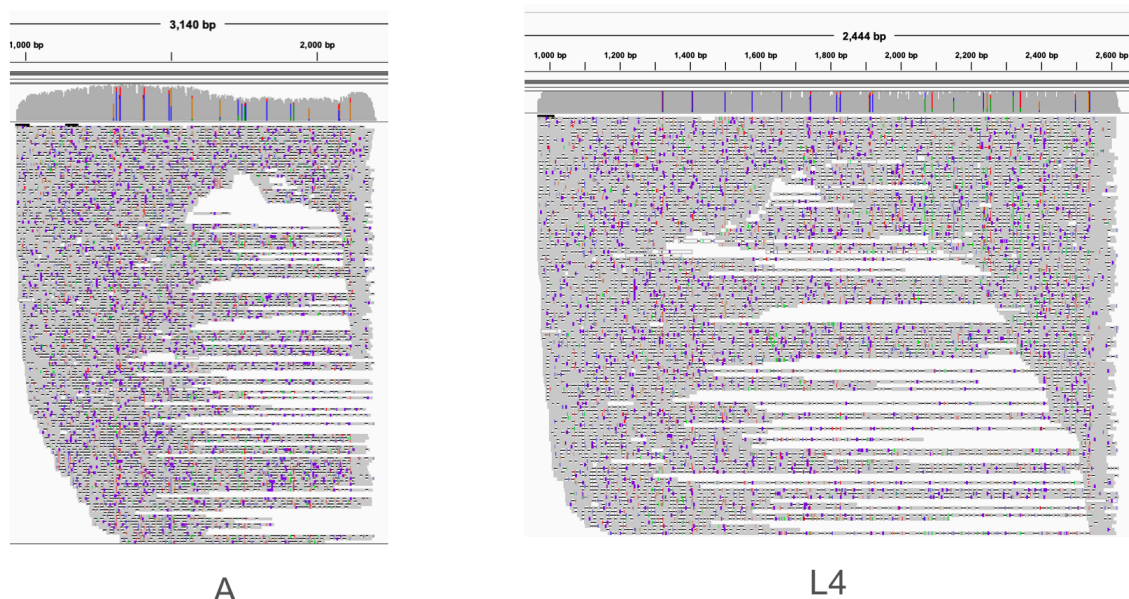


Fig. 46 IGV2.4.19 visualisation of the mapping of PRDM9 L4 and A ZnF array MinION nanopore sequence reads from MAP004 kit/R7 flow cell (2015) mapped to separate reference sequences containing L4 and A alleles using BWA sequence alignment software. Purple vertical marking denote single base insertions; black dashes between the sequence reads (in grey for mapping quality >0) denote deletions/gaps in the sequence reads; white sequence reads have mapping quality of 0; coloured vertical lines on the reads and coverage track indicate mismatches over 0.7.

As minimap2 became available, it was decided that mapping of both datasets should be done with the same sequence alignment pipeline. To this end, after converting the fast5 files to fastq files using poretools, the mapping pipeline for the 2019 dataset was adopted for the 2015 dataset and both datasets were compared again (Fig. 47). The new cross-comparison showed a more similar profile for both datasets with no significant mismatches (over 0.7) and a general uniformity of read depth across the ZnF arrays at mapping quality of 0. However, when the mapping quality was increased to 20, there was a relatively higher number of significant mismatches in the 2015 data. As previously shown (Fig. 45), the 2019 data also retained mismatches over 0.7 even when considering that these L4 and A alleles were purified and separate until the library preparation. A possible explanation is that a number of reads are aligning to the incorrect ZnF array.

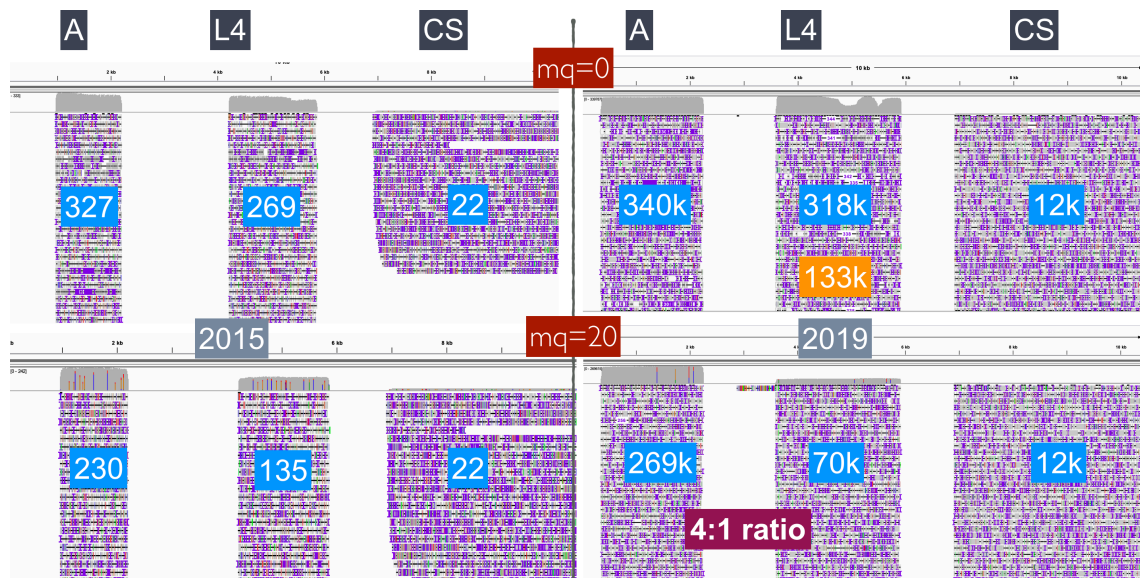


Fig. 47 IGV2.4.19 visualisation of the mapping of PRDM9 L4 and A ZnF array MinION nanopore sequence reads from MAP004 kit/R7 flow cell (2015) and Ligation Sequencing kit/R9.4.1 flow cell mapped to concatenated reference sequences containing L4 and A alleles and CS sequence using minimap2 sequence alignment software. Purple vertical marking denote single base insertions; black dashes between the sequence reads (in grey for mapping quality >0) denote deletions/gaps in the sequence reads; white sequence reads have mapping quality of 0; blue labels indicate maximum read depth; dark grey labels indicate region of the reference alleles and CS; light purple labels indicate alignment software used; red labels indicate mapping quality threshold applied; magenta label indicate relative ratio of read depth between A and L4 alleles.

6.2.5 Illumina dataset read depth and variant site analysis

Preliminary findings from the visualisation of the remapping of the Illumina dataset showed that when the alleles carried by an individual is length homozygous, it is possible to distinguish the diploid genotype of the individual from assessing the read depth and mismatches observed. It is also partially possible to identify both alleles even if there is length heterozygosity depending on the type of known PRDM9 alleles. However, Sanger sequencing would still be required for confirmation. Also in the Illumina dataset, distance between cognate ends varied with typical insert sizes ranging ~200-500bp. It was evident at this point that 100bp reads, even if paired end facilitated better mapping, was not sufficient to characterise novel alleles by itself. Hence, attention turned to

the whole Illumina dataset to identify known alleles carried by all the individuals.

An alternative bioinformatic approach involving a minimal number of steps and one which did not rely on Sanger sequencing confirmation and could therefore be applied to the whole dataset was then developed in association with Dr Pille Hallast. A detailed breakdown of the steps involved in this approach is given in Appendix V. Scripts were written to extract VCF data, namely read depth and variant sites, for all the individuals. As a test case, data from a Dutch (Netherlands) individual was examined since previous visualization in IGV had showed that this individual was most likely heterozygous for PRDM9 A and L24. Read depth was evaluated when the sequences this individual were mapped against each of 50 different known alleles of PRDM9 (Fig. 48). The mean read depth was ~ 174 and since there are mismatches between the two alleles at same variant site, there was no drop in coverage. Various allele combinations that differed greatly in size were eliminated. Sample profiles with raw read depth demonstrated more than 2 candidate alleles, namely A (ABCDDECFGHFHFIJ), L24 (ABCDDECFTPFQJ) , L37 (ABCDDECFGHFQJ), L42 (ABCDDECFGsFIJ) and L9 (ABCDDECFGPFQJ) (Fig. 49).

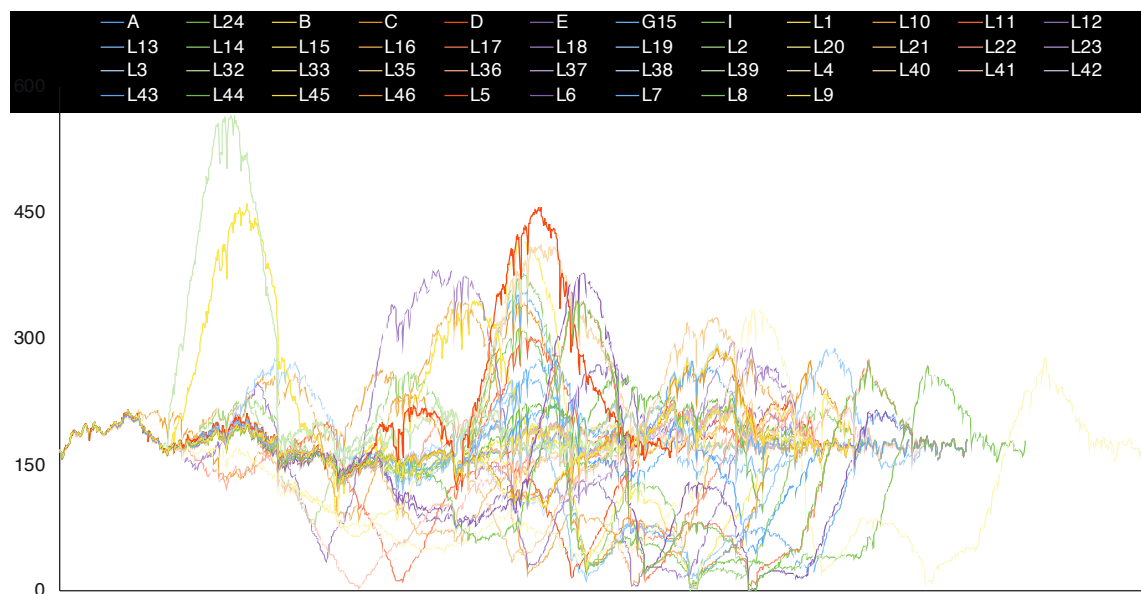


Fig. 48 Read depth data over the ZnF array region for a Dutch (Netherlands) individual thought to carry PRDM9 A/L24 alleles according to IGV visualisation after remapping. The legend and colours indicate the read depth when mapped 45 PRDM9 alleles on GenBank (Benson, 2008).

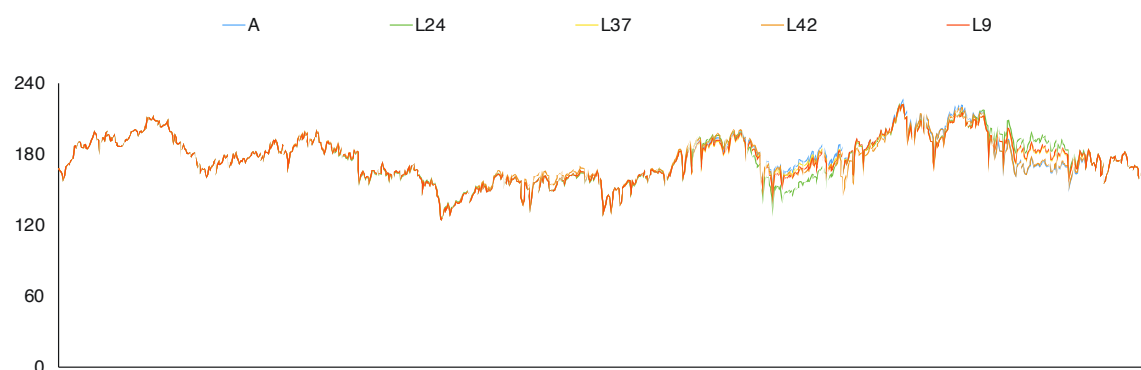


Fig. 49 Read depth data over the ZnF array region for a Dutch (Netherlands) individual thought to carry PRDM9 A/L24 alleles according to IGV visualisation after remapping. The legend and colours indicate the read depth when mapped 45 PRDM9 alleles on GenBank (Benson, 2008).

When this methodology was applied to the Australian Aborigine individual confirmed to be homozygous for the A allele by Sanger sequencing and indicative of the same by IGV visualisation of the remapping, the dip in read depth was observed at the point where the alignment software tries to force-

map C-ZnF containing reads (according to B allele reference) to where there is actually an E-ZnF of the A allele (Fig. 50).

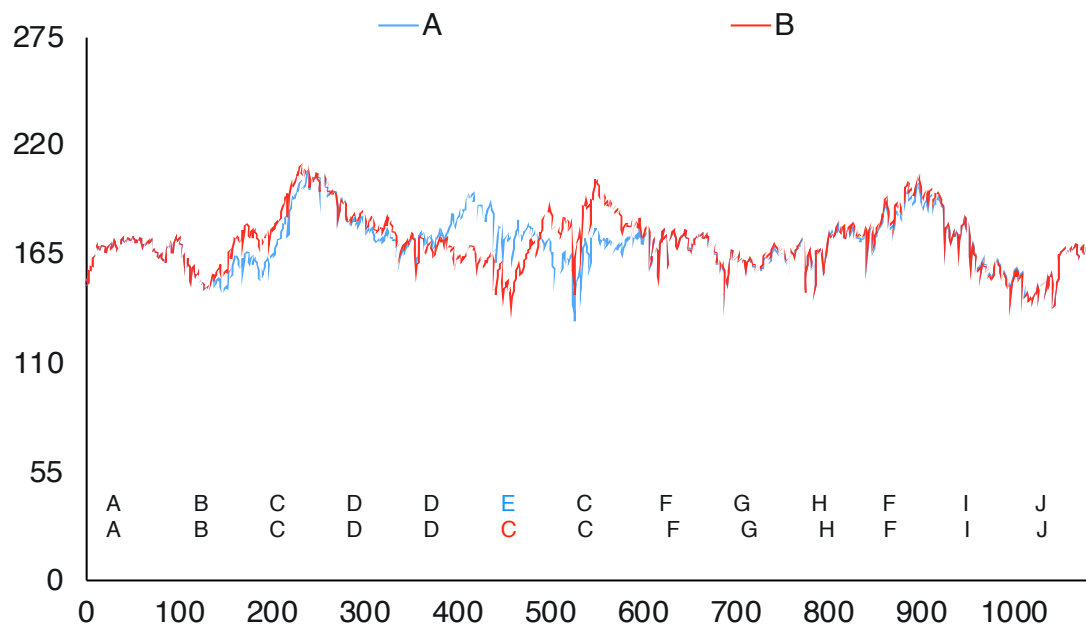


Fig. 50 Read depth data over the ZnF array region for a Dutch (Netherlands) individual thought to carry PRDM9 A/L24 alleles according to IGV visualisation after remapping. The legend and colours indicate the read depth when mapped 45 PRDM9 alleles on GenBank (Benson, 2008).

The variant site analysis method was devised to gain further insight into these minute changes, locating the exact opposite alternate alleles which had the variants at the same sites but no other sites (Table 18). All variant sites data for each sample-references were extracted, reordered according to base position and conditional highlighting was used for all bases (A, G, C, T) in Ref and Alt alleles.

Table 18 Variant site identification of candidate alleles inferred from remapping of

Base Position	Allele A Ref/Alt	L9 Ref/Alt	L37 Ref/Alt	L24 Ref/Alt	Further candidates
1500	C/T	C/T	C/T	T/C	
1798	A/G	G/A	A/G	G/A	
2038	C/T	T/C	T/C	T/C	

L39	187	T	C	78
L1	206	C	G	81
L39	218	C	G	222
L32	323	T	A	5.73279
L16	374	C	G	89
L18	374	C	G	170
L1	411	C	T	222
L39	411	C	T	222
L1	442	A	G	222
L17	442	A	G	18.4575
L39	442	A	G	222
L17	448	C	A	73
B	458	C	G	122
C	458	C	G	190
G15	458	C	G	178
I	458	C	G	178
L12	458	C	G	74
L13	458	C	G	189
L15	458	C	G	189
L19	458	C	G	189
L2	458	C	G	90
L6	458	C	G	73
L7	458	C	G	126
G15	542	C	G	4.08622
L13	542	C	G	7.82705
L17	542	C	G	132
L40	542	C	G	22.2085
L5	542	C	G	52
L16	604	A	C	114

Fig. 51 Exemplary positions of unique reference and alternate allele bases for all 45 references mapped to the Dutch individual.

Monomorphic sites for all references were removed leaving only the unique variant sites (Fig. 51) as removing all differences in a potentially heterozygous individual would result in loss of information which would make confirmation of the correct two alleles impossible. Producing a matrix with base

position from 1 to n vertically and the PRDM9 allele references moving across, with the intersection showing the variant (ref/alt) base configuration proved time consuming (Fig. 52).

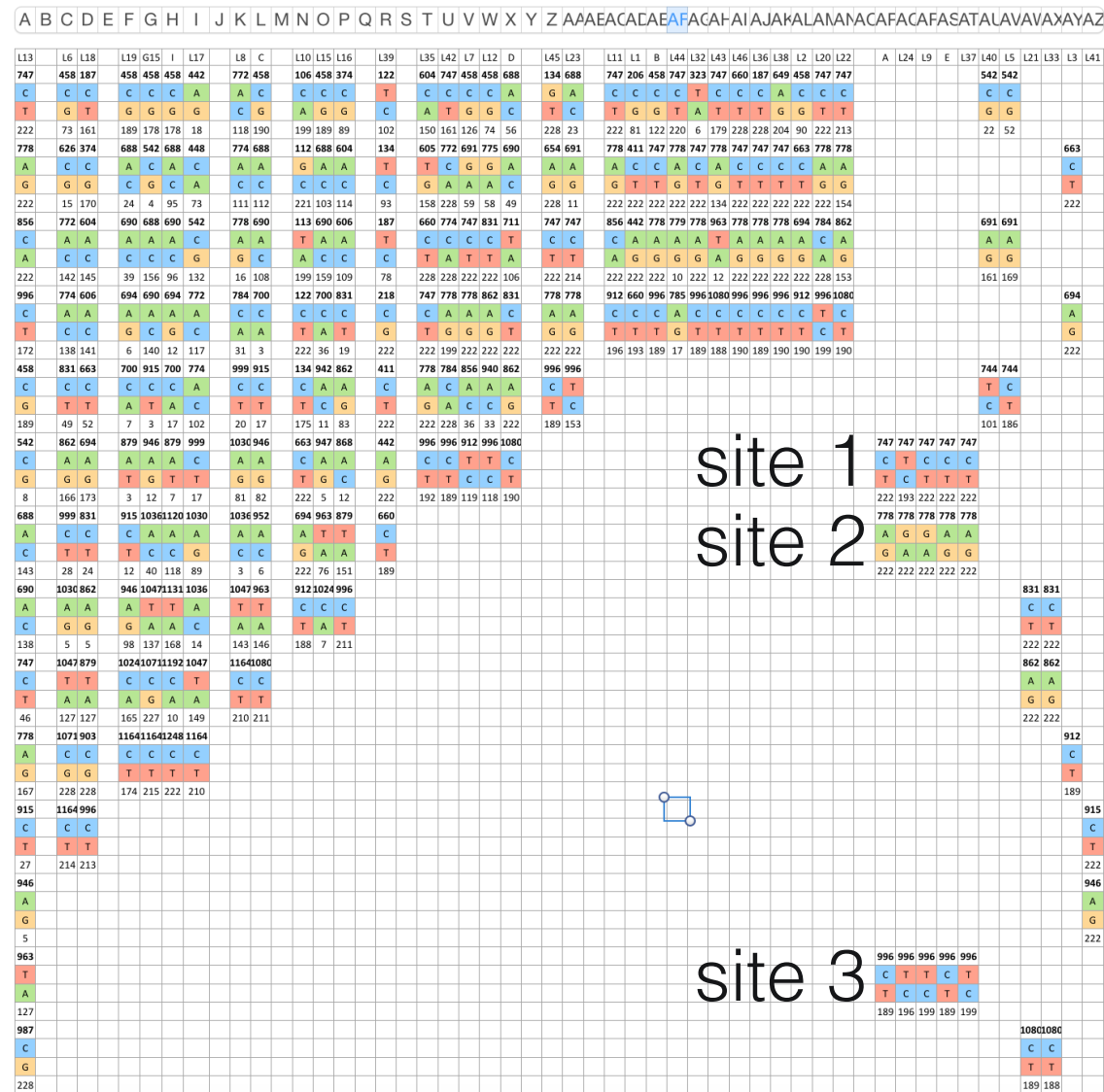


Fig. 52 Matrix showing the base positions and variant configuration in reference and alternate allele for all 45 references mapped to the Dutch individual.

To make this process more efficient and automated, Dr Pille Hallast worked to develop a script to extract variant site information from the remapped BAM files. The output of this script was checked but due to a lack of time this could not be explored any further. Although IGV visualisation was more straightforward to discern whether alleles could be identified or not, the aim of

the proposed variant calling method was for it be independent. However, all three methods may need to be used to reinforce conclusions of allele identification.

6.2.6 De novo assembly of Illumina dataset

The bam files used for remapping had already been mapped to the human genome reference assembly which contains the PRDM9 B allele. As such, it was reasonable to assume that many useful reads would have been filtered out of these bam files in the initial mapping. Still, since the remapping of these reads to known ZnF arrays also forces reads to be filtered out again, the remapping would not ultimately assist in characterising novel alleles. Instead, through signatures complete losses in read depth and stacked reads with obvious mismapping, the remapping only identified alleles which may be either novel or more likely unresolvable due to the existence of length heterozygous alleles in the sample. To investigate whether the former case could be addressed in the dataset, de novo assembly was attempted for Illumina read data from individuals verified by Sanger method to be carriers of A and L4 alleles. Generating these assemblies required information about the FASTQ files for these alleles, which was obtained by running FastQC Report on Galaxy (Afgan et al., 2018). For the PRDM9 A/L4 individual, total number of reads were 3840 in each of two files and the coverage expectation was just under x200 for the ZnF array and flanking DNA. The read length was 100bp which would inform the k-mer size value for de novo assembly. The quality encoding type was Illumina 1.9 type and drops in base quality were only observed over the last ~3 bases of the reads with far outliers close to a Phred score of 20 so at first it seemed that opportunistic trimming could be done. However, the number of recurring k-mers was less than 8 counts and there were no signs of adaptor or barcode sequences, so no trimming was ultimately done. Poly-Ns were also not observed. With a total number of 2578 reads, similar attributes were found for the A/A allele carrier sequence data.

The hash size (k), expected coverage (e), and coverage cutoff (c) needed to be set to make a viable de novo assembly. A VelvetOptimiser script, which wraps Velvet assembler and conducts sequence read survey and assembly of contigs in a pipeline, was initially used to find the best values for these variables. Different combinations of these variables such as shuffling through different k values (based on around half the read length of 100bp for these reads), N50, the number of base pairs in the longest contigs, and the size of the largest contains were changed to find the best assembly with least number of contigs (nodes in file) whilst not losing useful read data and an N50 close to the half of the ~2000bp region of the PRDM9 ZnF array. Since VelvetOptimiser also carried out the 'velvetg' assembly function, the results were compared simultaneously. VelvetOptimiser was unable to confirm the best variables to use to get the essential k-mer length for these data. It was also unable to confirm best combination of these variables to check (Appendix V).

Velvet sequence alignment software was then run manually by only setting the k-mer sizes to around half the 100bp read length which was shown to be consistent for all reads in the data. A range of k-mer sizes beginning with 45 and upwards was checked using 'velveth' function and contigs were assembled using 'velvetg' functions. The resulting contigs were aligned to the relevant A and L4 allele reference sequences with 1000bp flanking DNA on both ends. For the A/A carrier, at k-mer set at 50, yielding the least number of contigs and reaching a target N50 of ~1000, a 84% alignment to the A allele was achieved using Serial Cloner v2.6 (Appendix V). In contrast, the best results for the A/L4 carrier contigs aligned to the L4 allele achieved only ~30% alignment. Removing the flanking region from the reference in fact lowered the extent of alignment observed. This is likely due to the flanking region of non-repetitive sequence serving to facilitate better alignment.

Since VelvetOptimiser did not provide the required information and despite the mixed results from using Velvet by itself, de novo assemblies were

obtained with Velvet for the five individuals validated by Sanger method. Since the contigs themselves are meant to overlap, for each individual they were aligned to themselves and then with reference alleles using MacVector v17 (Fig. 53). At the ideal k-mer sizes, longer contigs were obtained but only mini motifs of up to 3 ZnFs (eg. ABC) were observed (Fig. 53a). With lower k-mer sizes, more contigs overlapped but with a higher level of mismatches (Fig. 53b). MacVector was used to generate a consensus assembly for each after alignment with the relevant reference alleles. Due to the overall deficiency of the consensus and high number of ambiguous sites, a complete ZnF array was not possible for any of the tested alleles.

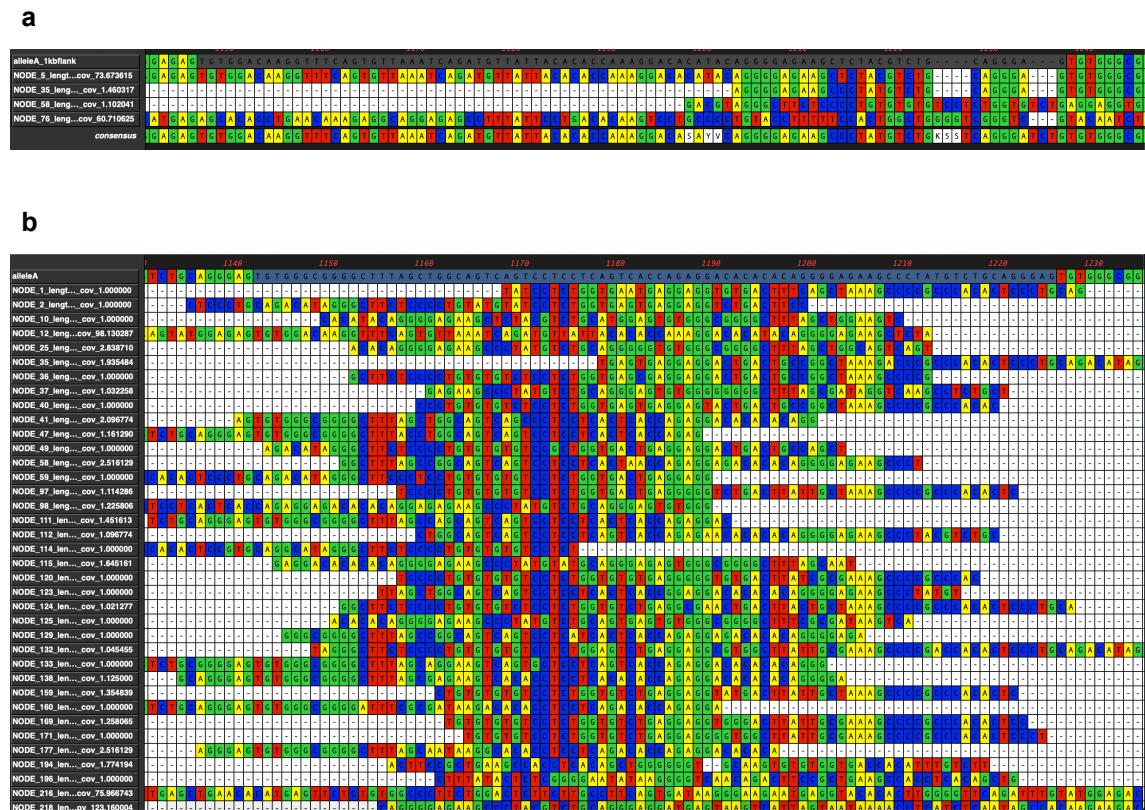


Fig. 53 Contigs derived from de novo assembly of Illumina data for individual homozygous for PRDM9 A aligned to A allele ZnF array. a. K-mer size set to 50 b. K-mer size set to 31. The instance highlights the single E-ZnF of this array. Most of the contigs obtained were concentrated on the E-ZnF.

6.2.8 Overall platform comparison

For comparison, consensus sequences for PRDM9 alleles A, C, L4, L47 were generated via IGV from the Sanger, Ion Torrent (both versions) and MinION nanopore. It should be noted that generating consensus sequences via IGV is not entirely suitable if there are gaps in the sequence but the purpose was to do a preliminary comparison. The de novo assembly consensus sequences were not suitable for comparison. Additionally, an individual predicted to carry L3/L32 and an individual predicted to carry L2/A according to read depth analysis of the Illumina dataset were selected at random for comparison against the known structure of these alleles. Two examples (PRDM9 L4/A and L47/C carriers) of sequencing comparison for three platforms is in Appendix VI.

When comparing Sanger, Ion Torrent (both 200bp and 400bp versions) and the Illumina remapping for the L4 and A allele and flanking DNA, 25 identified sites were predominantly due to the manual editing of the Sanger sequence bases. The remaining 789 bases were genuine mismatches where the assemblies of the L4 and A alleles differed after the 7th ZnF of the A allele from which point the order of ZnF between these alleles started to differ greatly. However, there were only 3 discordancies noted for the L4 allele sequences across the three platforms and they all occurred outside of the ZnF array. A heterozygous site was found on the 5' flanking region at chr5:23526045 (hg38) with A/T (W) polymorphism which was identified as SNP rs2934795. The nature of the other 2 sites could not be determined.

For the sequencing of the PRDM9 C/L47 carrier, only one discordancy was observed across platforms for mapping to L47, specifically C/G heterozygosity observed for base C on the O-ZNF and base G on the H-ZNF to G but not from I-ZNF (6 differences) which is consistent with the IGV visualisation method. No other discordancies were identified. In contrast, the mapping of the

C allele for this individual showed 164 discordancies. Whilst, the Sanger and Illumina consensus matched, the two Ion Torrent versions matched but differed to the former two sequences. The nature of these differences could not be determined.

6.3 Discussion

The main aim of this study was to determine whether it was feasible to identify novel PRDM9 alleles from the Illumina HiSeq2000 dataset because it contained individuals representing a more diverse population set than studied before. The underlying reason for doing a feasibility study was to determine the critical factors such as read length, read depth and base accuracy (variant sites) to enable confident characterisation of the ZnF arrays that distinguish PRDM9 alleles.

Several lessons were learnt from this exercise. It is very possible that the initial data processing on the raw sequencing data on the human genome reference containing the B allele may have caused the loss of valuable read data that may have otherwise strengthened the mapping and confident characterisation of PRDM9 alleles in this study, especially in the case of de novo assembly with Velvet. Further work with the original FASTQ files after base calling may help to confirm this. There was also the added difficulty of having to the Illumina HiSeq2000 dataset being diploid and therefore no possibility applying a mapping pipeline for each two alleles for each individual in isolation. For the time being, even 100bp paired end reads have been shown to map well to very similar ZnF arrays making it difficult to assign only two PRDM9 alleles confidently to any one individual. Instead, a candidate of list alleles has been generated. This however allows a list of individuals to be identified whose PRDM9 alleles are either length heterozygous and/or novel.

The 200bp and 400bp versions of Ion Torrent sequencing, and the MinION-based nanopore sequencing with its variable lengths of long reads applied to ~2000bp amplicons were essentially to check the extent to which read length affects reliable definition of ZnF array structure. Comparing between 200bp and 400bp Ion Torrent sequencing, it must be noted that the trial was done only once. There was an indication of weak coverage in the 200bp kit chip map, probably due to size selection issues. This was reflected in the maximum read depth varying between 6-17 times from the median read depth in the 400bp version compared to the 200bp kit. Yet, the actual read depth across the ZnF array and flanking region were still comparable with both versions. The 5' flanking non-repetitive region and the first few ZnFs of the array had higher read depth owing to the better mapping and the advantage of the longer 400bp kit was observed. The read depth then trailed off as lower quality reads were found mostly inhabiting the middle of the ZnF arrays. These lower quality reads populated only when the mapping quality threshold was incrementally decreased. At higher threshold, there were visible loss of coverage in the middle of the ZnF array. The mapping quality of reads only mildly improved towards the 3' end of the arrays. The possible explanation is that the 400bp kit improves the BWA mapping but still does not have enough read length to manage the tandem repeats of ZnFs such as those found in the L4 allele. This was also seen to a similar extent with the A allele mapping. It was interesting to note the read depth for L4 allele was more uniform across the ZnF array compared to the A allele.

The expected advantage of MinION nanopore sequencing in its ability to generate ultra-long reads of up to 2.2Mb was not fully explored as the amplicons used for sequencing were ~2kb in length. Theoretically using longer amplicons or genomic DNA to sequence the ZnF array region would yield better results. It may also compensate for the sheer number of gaps/deletions in these reads. A de novo assembly for the remapping the nanopore data was still a work in progress at the time of writing and given the inability of the Illumina data de

novo assembly to characterise the known A allele on its own shows that even good read quality alone cannot enable confident characterisation.

Novel PRDM9 alleles found in this study can be used to target recombination hotspots of interest in terms of how efficiently they use particular hotspots and how that changes the activity of nearby hotspots. Binding affinity assays similar to Patel et al (2016) and Patel et al (2017) may further illuminate the spatial complexities of DNA binding for long arrays of ZnFs. This includes understanding degenerate and non-degenerate positions, redundancy, ZnFs that are not required for binding, etc. For now, it is clear that Sanger sequencing remains the method of choice for accurate characterisation of ZnF array configuration.

CHAPTER 7: DISCUSSION

PRDM9-regulated hotspot activation has been well studied in the last decade. PRDM9's role in events leading to genetic disorders in individuals has also been directly demonstrated using sperm DNA approaches (Berg et al., 2010) and implicated in others (Pratto et al. 2014). Its ability to change recombination landscapes (Hinch et al., 2011) and its potential role in childhood leukaemia (Hussin et al., 2013; Thompson et al., 2014) has also created interest in the extent of PRDM9 diversity in different population groups.

Until very recently, PRDM9 was only ever considered to be a meiotically-expressed gene. However, recent findings have shown aberrant PRDM9 expression in tumour cells of 39 different cancer types (Houle et al., 2018), though not in cell types associated with childhood leukaemia, which has been a focus of this work. In the 39 tumours samples studied by Houle et al., PRDM9 A-associated sequence motifs appeared to be in the vicinity of structural variant breakpoints suggesting a direct link. The probability of recombination in these locations was verified in part by comparing with the meiotic DSB maps from the Pratto et al. (2014) study. This agrees with the Hussin et al. (2013) study that proposes an inheritance model whereby PRDM9-activated recombination sites in parents increase the likelihood for genomic instability in the offspring. The observation of elevated recombination in the MHCII region associated with the British ALL cohort by our collaborator Dr Pamela Thompson prompted further investigation of hotspot control in this region in this thesis. Of the two meiotic hotspots studied, one was found to be PRDM9-A regulated and the other Ct-regulated. Given the findings of Hussin et al. (2013), namely that K-type ZnF alleles are overrepresented in the parents of children with ALL in French-Canadian families, the latter Ct-regulated AA hotspot was deemed the most relevant in in this context. However, comparing expression from post-diagnosis and remission samples amongst UK samples, our collaborator was unable to observe detectable levels of PRDM9 proteins in the case samples and hence

concluded that PRDM9 had no role in the disease aetiology of childhood ALL, at least in this cohort (Thompson, 2015). Furthermore, work presented in this thesis did not replicate the relation between childhood ALL and K-ZnF containing Ct, D and L20 alleles, noted by Hussin et al. (2013) for comparable rare alleles containing K-ZnFs. However, SNP genotyping of the UK cohort and controls did strengthen FIGNL1 as being a disease susceptibility locus, in line with previous work.

In this work DNA3 hotspot was confirmed to be PRDM9 A-regulated and AA hotspot was demonstrated to be activated by a subset of Ct alleles. It should be noted that for both DNA3 and AA hotspots, cis-effects such as SNPs in the central hotspot region may affect PRDM9 binding as seen in previous works (eg. Jeffreys and Neumann, 2002; Jeffreys and Neumann, 2005). Hence, hotspot activity would not be solely dependent of PRDM9 variants. Individual-specific sequence variation can change the recombination landscape and hotspots may decline as variation accumulates such that sequence motifs associated with A and Ct alleles disappear.

7.1 Future work

The DNA3 hotspot regulation was revisited due to the elevated ancestral recombination rates at this hotspot (Thompson et al., 2014) directly to address the propositions of Hussin et al. (2013) and Thompson et al. (2014). The AA hotspot was however found to be the more likely candidate for NAHR mechanism since it was found to be activated by some K-finger containing ZnF arrays. Further exploration of DSB maps and structural variations associated with the AA hotspot and potentially detection of de novo NAHR events using batch sperm methods as employed by Lam et al. (2006) in relation to globin gene rearrangements leading to thalassaemia could shed further light on this.

As well as an opportunity to characterise novel PRDM9 alleles in rarely sampled populations, the comparison of second and third generation sequencing platforms provided insight into the prospects of using other publicly available existing datasets such as the 1000 Genomes data (1000 Genomes Project Consortium, 2015) could be used to characterise novel alleles. At the time of writing none of these platforms performed sufficiently well to challenge the 'gold standard' of Sanger sequencing. Oxford Nanopore Technologies' MinION platform held the most promise but software improvements are required for this to become a real contender. At present, combining the MinION data with the corresponding Illumina data is likely to improve the de novo assembly, though this is not a practical solution for exploring existing datasets in the public domain that will mostly have been generated using a single technology.

This work did however show that the PRDM9-SNP haplotype network generated from the North European and Sub-Saharan sperm donor panels held at the University of Leicester was reliable for screening for K-ZnF containing alleles. Although there was no connection with B-ALL families, the prevalence of these alleles in different populations is of interest as they might also reveal other Ct alleles with 5/8 match to the Myer's motif. Hence, a K-ZnF 'search' on the Illumina dataset might prove fruitful in the quest for identifying as yet uncharacterised PRDM9 alleles.

Appendices

Appendix I: PCR design

This appendix includes PCR designs for the recombination hotspot crossover assays done in the DNA3 and newly described AA hotspots in the MHCII, duplex PCR for screening for K-ZnF containing PRDM9 alleles in a British ALL cohort and Sanger sequencing as carried out in this thesis. PCR primer characteristics, PCR cycling conditions, PCR strategy for target regions and sequence context of target regions and AA hotspot morphology analysis are included. This appendix includes:

- A. DNA3 and AA hotspots
- B. ALL K-ZnF Screening duplex PCR
- C. PCR product for Sanger Sequencing PRDM9 ZnF array template
- D. AA Hotspot ASP Design
- E. AA Hotspot Universal Primer Design
- F. DNA3 Hotspot Universal Primer Design
- G. AA Hotspot Linkage Phasing

Ali, Iththisham (2020): Appendix I: PCR Design. University of Leicester. Dataset.
<https://doi.org/10.25392/leicester.data.12318695>

Appendix II: SNP genotyping

This appendix contains SNP genotyping data from complete SNP genotyping results from studies on DNA3 and AA hotspots, and the childhood ALL study. For recombination hotspot analysis, these data formed the basis of the crossover analysis design. SNP genotyping data from the British ALL cohort was classified by major subtypes and compared with health control groups. This appendix includes:

- A. DNA3 and AA hotspots

B. ALL study

Ali, Ihthisham (2020): Appendix II: SNP Genotyping. University of Leicester. Dataset. <https://doi.org/10.25392/leicester.data.12235397>

Appendix III: Crossover activity

Appendix III contains hotspot activity analysis data on DNA3 and AA hotspots in a MHCII sub-region. These data is indicative of the local recombination profile in the MHCII and provides confirmatory evidence for the differential activation of PRDM9 A and a subset of Ct alleles for these two hotspots. Additionally, these data provide fine-scale information on the morphology of the AA hotspot. This appendix includes:

- A. DNA3 hotspot CO activity data
- B. DNA3 hotspot Chi-Squared Test
- C. AA hotspot CO activity data
- D. AA hotspot d257 (20) breakpoint mapping
- E. AA hotspot d257 (20) hotspot morphology graph data

Ali, Ihthisham (2020): Appendix III: Crossover activity. University of Leicester. Dataset. <https://doi.org/10.25392/leicester.data.12364031>

Appendix IV: PRDM9 ZnF arrays

Appendix IV contains information on individual ZnFs, ZnF mini-motifs, ZnF arrays used for PRDM9 ZnF array characterisation and also the descriptions of novel PRDM9 alleles characterised in this thesis. This appendix includes:

- A. Individual ZnFs
- B. ZnF minimotifs
- C. Known and novel ZnF arrays
- D. GenBank ZnF arrays defined by individual ZnFs

Ali, Ihthisham (2020): Appendix IV: PRDM9 ZnF arrays. University of Leicester. Dataset. <https://doi.org/10.25392/leicester.data.12364007>

Appendix V: Bioinformatic pipelines/scripts

Appendix V contains various data pipelines and scripts used for the remapping of Illumina HiSeq2000 dataset to known PRDM9 ZnF arrays, read depth and variant calling vcf file generation, haplotype estimation and imputation of FIGNL1 coding variants in relation to the British ALL cohort, de novo assembly of read data and mapping of MinION read data. This appendix includes:

- A. ALL study phasing and imputation
- B. Illumina HiSeq 2000 dataset - Read depth (DP) and variant calling pipeline
- C. Illumina HiSeq 2000 dataset - data treatment
- D. VelvetOptimiser best k-mer determination log (exemplary)
- E. Alignment of contigs generated by Velvet de novo assembly for the PRDM9 A/A carrier and aligned against the PRDM9 A ZnF array
- F. MinION nanopore reads - minimap2 pipeline

Ali, Ihthisham (2020): Appendix V - Bioinformatic pipelines/scripts. University of Leicester. Dataset. <https://doi.org/10.25392/leicester.data.12363785>

Appendix VI: Sequencing platform comparison

Appendix VI contains the consensus sequences from reads from different sequencing platforms (Sanger, Ion Torrent 200bp and 400bp formats and Illumina HiSeq2000 mapped to reference ZnF arrays confirmed via Sanger. This appendix includes:

Appendix VI: A. L4 and A allele fragments mapped across platforms

Appendix VI: B. L47 mapping across platforms for the PRDM9 L47/C carrier

Appendix VI: C. C allele mapping across platforms for the PRDM9 L47/C carrier

Ali, Iththisham (2020): Appendix VI: Sequencing platform comparison. University of Leicester. Dataset. <https://doi.org/10.25392/leicester.data.12408119>

Appendix VII: Supplementary figures and table

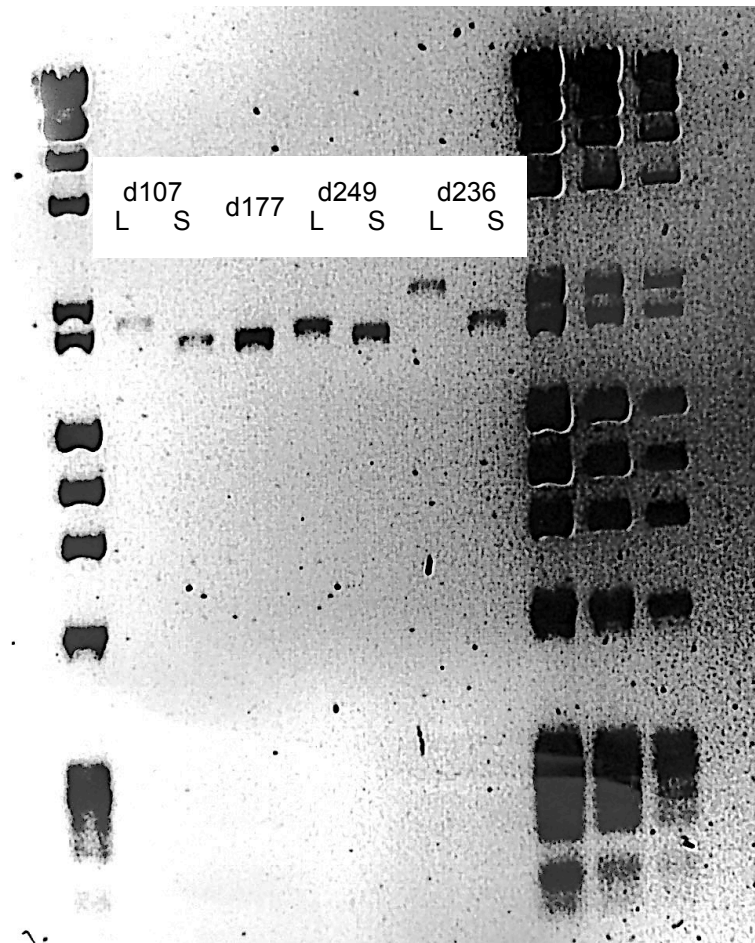


Fig. S1 PCR of amplicon for generating Sanger sequencing templates for PRDM9 ZNF Array in four trial individuals. A forward primer PN0.6F-TGAGGTTACCTAGTCTGGCA (forward primer) and PN0.2.5R-GAAGTCTGCTGACCCCTTAT (reverse primer) were used to create a ~2kb amplicon (for PRDM9 A) which was then used for creating sequencing templates. PRDM9 allele names, ZnF repeat lengths and Myer's match are given in Table S1.

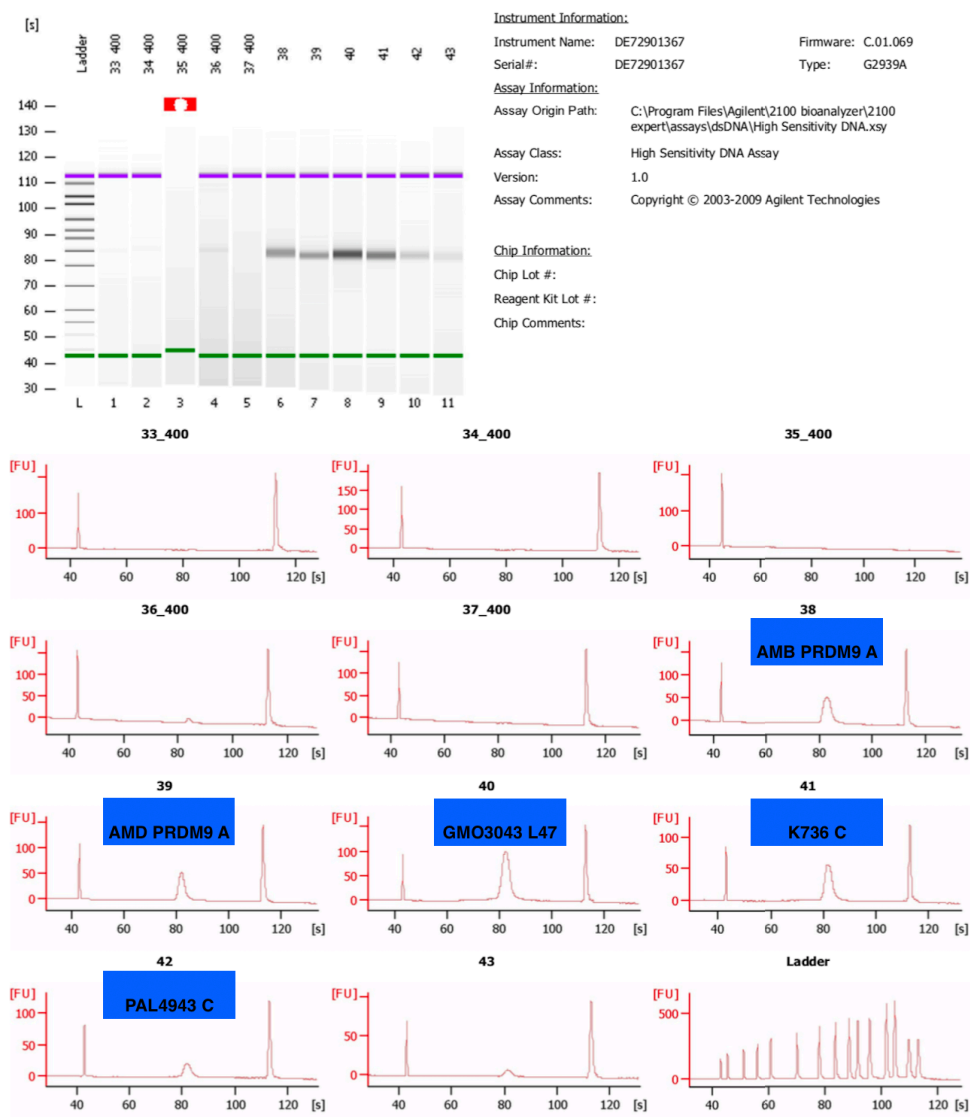


Fig. S2 Size estimation of fragments used for Ion Torrent sequencing

Table S1 PRDM9 alleles of 4 trial individuals characterised during Sanger optimisation

Panel no.	Origin	allele 1 rpts	allele 2 rpts	allele types		Myer's match	
d107	British	15	13	L8	A	5	8
d177 (3)	Afro-Caribbean	13	13	A	A	8	8
d249	Zimbabwean	14	13	C	A	5	8
d236	Zimbabwean	18	14	L4	L22	5	6

Bibliography

- 👤 1000 Genomes Project Consortium, Auton, A., Brooks, L.D., Durbin, R.M., Garrison, E.P., Kang, H.M., Korbel, J.O., Marchini, J.L., McCarthy, S., McVean, G.A. and Abecasis, G.R. (2015) 'A global reference for human genetic variation', *Nature*, 526(7571), pp. 68-74.
- 👤 Afgan, E., Baker, D., Batut, B., van den Beek, M., Bouvier, D., Cech, M., Chilton, J., Clements, D., Coraor, N., Gruning, B.A., Guerler, A., Hillman-Jackson, J., Hiltemann, S., Jalili, V., Rasche, H., Soranzo, N., Goecks, J., Taylor, J., Nekrutenko, A. and Blankenberg, D. (2018) 'The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2018 update', *Nucleic acids research*, 46(W1), pp. W537-W544.
- 👤 Allers, T. and Lichten, M. (2001) 'Intermediates of yeast meiotic recombination contain heteroduplex DNA', *Molecular cell*, 8(1), pp. 225-231.
- 👤 Antonarakis, S.E., Petersen, M.B., McInnis, M.G., Adelsberger, P.A., Schinzel, A.A., Binkert, F., Pangalos, C., Raoul, O., Slaugenhaupt, S.A. and Hafez, M. (1992) 'The meiotic stage of nondisjunction in trisomy 21: determination by using DNA polymorphisms', *American Journal of Human Genetics*, 50(3), pp. 544-550.
- 👤 Armstrong, S.J., Kirkham, A.J. and Hulten, M.A. (1994) 'XY chromosome behaviour in the germ-line of the human male: a FISH analysis of spatial orientation, chromatin condensation and pairing', *Chromosome research : an international journal on the molecular, supramolecular and evolutionary aspects of chromosome biology*, 2(6), pp. 445-452.
- 👤 Arnheim, N., Calabrese, P. and Tiemann-Boege, I. (2007) 'Mammalian meiotic recombination hot spots', *Annual Review of Genetics*, 41, pp. 369-399.
- 👤 Bandelt, H.J., Forster, P., Sykes, B.C. and Richards, M.B. (1995) 'Mitochondrial portraits of human populations using median networks', *Genetics*, 141(2), pp. 743-753.

- 👤 Batini, C., Hallast, P., Zadik, D., Delser, P.M., Benazzo, A., Ghirotto, S., Arroyo-Pardo, E., Cavalleri, G.L., de Knijff, P., Dupuy, B.M., Eriksen, H.A., King, T.E., de Munain, A.L., Lopez-Parra, A.M., Loutradis, A., Milasin, J., Novelletto, A., Pamjav, H., Sajantila, A., Tolun, A., Winney, B. and Jobling, M.A. (2015) 'Large-scale recent expansion of European patrilineages shown by population resequencing', *Nature communications*, 6, pp. 7152.
- 👤 Baudat, F., Buard, J., Grey, C., Fledel-Alon, A., Ober, C., Przeworski, M., Coop, G. and de Massy, B. (2010) 'PRDM9 is a major determinant of meiotic recombination hotspots in humans and mice', *Science (New York, N.Y.)*, 327(5967), pp. 836-840.
- 👤 Baudat, F. and de Massy, B. (2007) 'Regulating double-stranded DNA break repair towards crossover or non-crossover during mammalian meiosis', *Chromosome research : an international journal on the molecular, supramolecular and evolutionary aspects of chromosome biology*, 15(5), pp. 565-577.
- 👤 Baudat, F., Imai, Y. and de Massy, B. (2013) 'Meiotic recombination in mammals: localization and regulation', *Nature reviews.Genetics*, 14(11), pp. 794-806.
- 👤 Benson, D.A., Karsch-Mizrachi, I., Lipman, D.J., Ostell, J. and Wheeler, D.L. (2008) 'GenBank', *Nucleic Acids Research*, 36(suppl_1), pp. D25-D30.
- 👤 Berg, I.L., Neumann, R., Lam, K.W., Sarbajna, S., Odenthal-Hesse, L., May, C.A. and Jeffreys, A.J. (2010) 'PRDM9 variation strongly influences recombination hot-spot activity and meiotic instability in humans', *Nature genetics*, 42(10), pp. 859-863.
- 👤 Berg, I.L., Neumann, R., Sarbajna, S., Odenthal-Hesse, L., Butler, N.J. and Jeffreys, A.J. (2011) 'Variants of the protein PRDM9 differentially regulate a set of human meiotic recombination hotspots highly active in African populations', *Proceedings of the National Academy of Sciences of the United States of America*, 108(30), pp. 12378-12383.
- 👤 Berg, I., Neumann, R., Cederberg, H., Rannug, U. and Jeffreys, A.J. (2003) 'Two Modes of Germline Instability at Human Minisatellite MS1 (Locus

- D1S7): Complex Rearrangements and Paradoxical Hyperdeletion', *The American Journal of Human Genetics*, 72(6), pp. 1436-1447.
- 👤 Billings, T., Parvanov, E.D., Baker, C.L., Walker, M., Paigen, K. and Petkov, P.M. (eds.) (2013) *DNA binding specificities of the long zinc-finger recombination protein PRDM9*. England: BioMed Central Ltd.
 - 👤 Borel, C., Cheung, F., Stewart, H., Koolen, D.A., Phillips, C., Thomas, N.S., Jacobs, P.A., Eliez, S. and Sharp, A.J. (2012) 'Evaluation of PRDM9 variation as a risk factor for recurrent genomic disorders and chromosomal non-disjunction', *Human genetics*, 131(9), pp. 1519-1524.
 - 👤 Brick, K., Smagulova, F., Khil, P., Camerini-Otero, R.D. and Petukhova, G.V. (2012) 'Genetic recombination is directed away from functional genomic elements in mice', *Nature*, 485(7400), pp. 642-645.
 - 👤 Buard, J., Bourdet, A., Yardley, J., Dubrova, Y. and Jeffreys, A.J. (1998) 'Influences of array size and homogeneity on minisatellite mutation', *EMBO Journal*, 17(12), pp. 3495-3502.
 - 👤 Campos, E.I. and Reinberg, D. (2009) 'Histones: annotating chromatin', *Annual Review of Genetics*, 43, pp. 559-599.
 - 👤 Chandley, A.C., Goetz, P., Hargreave, T.B., Joseph, A.M. and Speed, R.M. (1984) 'On the nature and extent of XY pairing at meiotic prophase in man', *Cytogenetics and cell genetics*, 38(4), pp. 241-247.
 - 👤 Chandley, A.C., Hargreave, T.B., McBeath, S., Mitchell, A.R. and Speed, R.M. (1987) 'Ring XY bivalent: a new phenomenon at metaphase I of meiosis in man', *Journal of medical genetics*, 24(2), pp. 101-106.
 - 👤 Chowdhury, R., Bois, P.R., Feingold, E., Sherman, S.L. and Cheung, V.G. (2009) 'Genetic analysis of variation in human meiotic recombination', *PLoS genetics*, 5(9), pp. e1000648.
 - 👤 Cole, F., Keeney, S. and Jasin, M. (2010) 'Comprehensive, fine-scale dissection of homologous recombination outcomes at a hot spot in mouse meiosis', *Molecular cell*, 39(5), pp. 700-710.
 - 👤 de Lannoy, C., de Ridder, D. and Risse, J. (2017) 'The long reads ahead: de novo genome assembly using the MinION', *F1000Research*, 6, pp. 1083.

- 👤 de Massy, B. (2013) 'Initiation of meiotic recombination: how and where? Conservation and specificities among eukaryotes', *Annual Review of Genetics*, 47(1), pp. 563.
- 👤 Delaneau, O., Marchini, J. and Zagury, J.F. (2011) 'A linear complexity phasing method for thousands of genomes', *Nature methods*, 9(2), pp. 179-181.
- 👤 Donis-Keller, H., Green, P., Helms, C., Cartinhour, S., Weiffenbach, B., Stephens, K., Keith, T.P., Bowden, D.W., Smith, D.R. and Lander, E.S. (1987) 'A genetic linkage map of the human genome', *Cell*, 51(2), pp. 319-337.
- 👤 Dunson, D.B., Weinberg, C.R., Baird, D.D., Kesner, J.S. and Wilcox, A.J. (2001) 'Assessing human fertility using several markers of ovulation', *Statistics in medicine*, 20(6), pp. 965-978.
- 👤 Egel, R. and Lankenau, D. (eds.) (2008) *Recombination and Meiosis*. Berlin, Heidelberg: Springer Berlin Heidelberg.
- 👤 Ellinghaus, E., Stanulla, M., Richter, G., Ellinghaus, D., te Kronnie, G., Cario, G., Cazzaniga, G., Horstmann, M., Panzer Grumayer, R., Cave, H., Trka, J., Cinek, O., Teigler-Schlegel, A., ElSharawy, A., Hasler, R., Nebel, A., Meissner, B., Bartram, T., Lescai, F., Franceschi, C., Giordan, M., Nurnberg, P., Heinzow, B., Zimmermann, M., Schreiber, S., Schrappe, M. and Franke, A. (2012) 'Identification of germline susceptibility loci in ETV6-RUNX1-rearranged childhood acute lymphoblastic leukemia', *Leukemia*, 26(5), pp. 902-909.
- 👤 Emerson, R.O. and Thomas, J.H. (2009) 'Adaptive evolution in zinc finger transcription factors', *PLoS genetics*, 5(1), pp. e1000325.
- 👤 Freije, D., Helms, C., Watson, M.S. and Donis-Keller, H. (1992) 'Identification of a second pseudoautosomal region near the Xq and Yq telomeres', *Science (New York, N.Y.)*, 258(5089), pp. 1784-1787.
- 👤 Fumasoni, I., Meani, N., Rambaldi, D., Scafetta, G., Alcalay, M. and Ciccarelli, F.D. (2007) 'Family expansion and gene rearrangements contributed to the functional specialization of PRDM genes in vertebrates', *BMC evolutionary biology*, 7, pp. 187-187.

- 👤 Gale, K.B., Ford, A.M., Repp, R., Borkhardt, A., Keller, C., Eden, O.B. and Greaves, M.F. (1997) 'Backtracking leukemia to birth: identification of clonotypic gene fusion sequences in neonatal blood spots', *Proceedings of the National Academy of Sciences of the United States of America*, 94(25), pp. 13950-13954.
- 👤 Greaves, M. (1999) 'Molecular genetics, natural history and the demise of childhood leukaemia', *European journal of cancer*, 35(2), pp. 173-185.
- 👤 Guillon, H., Baudat, F., Grey, C., Liskay, R.M. and de Massy, B. (2005) 'Crossover and noncrossover pathways in mouse meiosis', *Molecular cell*, 20(4), pp. 563-573.
- 👤 Gurney, J.G., Severson, R.K., Davis, S. and Robison, L.L. (1995) 'Incidence of cancer in children in the United States. Sex-, race-, and 1-year age-specific rates by histologic type', *Cancer*, 75(8), pp. 2186-2195.
- 👤 Hallast, P., Batini, C., Zadik, D., Maisano Delser, P., Wetton, J.H., Arroyo-Pardo, E., Cavalleri, G.L., de Knijff, P., Destro Bisol, G., Dupuy, B.M., Eriksen, H.A., Jorde, L.B., King, T.E., Larmuseau, M.H., Lopez de Munain, A., Lopez-Parra, A.M., Loutradis, A., Milasin, J., Novelletto, A., Pamjav, H., Sajantila, A., Schempp, W., Sears, M., Tolun, A., Tyler-Smith, C., Van Geystelen, A., Watkins, S., Winney, B. and Jobling, M.A. (2015) 'The Y-Chromosome Tree Bursts into Leaf: 13,000 High-Confidence SNPs Covering the Majority of Known Clades', *Molecular biology and evolution*, 32(3), pp. 661-673.
- 👤 Hanawalt, P.C. (2004) 'Density matters: The semiconservative replication of DNA', *Proceedings of the National Academy of Sciences of the United States of America*, 101(52), pp. 17889.
- 👤 Hassold, T.J., Sherman, S.L., Pettay, D., Page, D.C. and Jacobs, P.A. (1991) 'XY chromosome nondisjunction in man is associated with diminished recombination in the pseudoautosomal region', *American Journal of Human Genetics*, 49(2), pp. 253-260.
- 👤 Hayashi, K., Yoshida, K. and Matsui, Y. (2005) 'A histone H3 methyltransferase controls epigenetic events required for meiotic prophase', *Nature*, 438(7066), pp. 374-378.

- 👤 Hinch, A.G., Tandon, A., Patterson, N., Song, Y., Rohland, N., Palmer, C.D., Chen, G.K., Wang, K., Buxbaum, S.G., Akylbekova, E.L., Aldrich, M.C., Ambrosone, C.B., Amos, C., Bandera, E.V., Berndt, S.I., Bernstein, L., Blot, W.J., Bock, C.H., Boerwinkle, E., Cai, Q., Caporaso, N., Casey, G., Cupples, L.A., Deming, S.L., Diver, W.R., Divers, J., Fornage, M., Gillanders, E.M., Glessner, J., Harris, C.C., Hu, J.J., Ingles, S.A., Isaacs, W., John, E.M., Kao, W.H., Keating, B., Kittles, R.A., Kolonel, L.N., Larkin, E., Le Marchand, L., McNeill, L.H., Millikan, R.C., Murphy, A., Musani, S., Neslund-Dudas, C., Nyante, S., Papanicolaou, G.J., Press, M.F., Psaty, B.M., Reiner, A.P., Rich, S.S., Rodriguez-Gil, J.L., Rotter, J.I., Rybicki, B.A., Schwartz, A.G., Signorello, L.B., Spitz, M., Strom, S.S., Thun, M.J., Tucker, M.A., Wang, Z., Wiencke, J.K., Witte, J.S., Wrensch, M., Wu, X., Yamamura, Y., Zanetti, K.A., Zheng, W., Ziegler, R.G., Zhu, X., Redline, S., Hirschhorn, J.N., Henderson, B.E., Taylor, H.A., Price, A.L., Hakonarson, H., Chanock, S.J., Haiman, C.A., Wilson, J.G., Reich, D. and Myers, S.R. (2011) 'The landscape of recombination in African Americans', *Nature*, 476(7359), pp. 170-175.
- 👤 Hinds, D.A., Stuve, L.L., Nilsen, G.B., Halperin, E., Eskin, E., Ballinger, D.G., Frazer, K.A. and Cox, D.R. (2005) 'Whole-genome patterns of common DNA variation in three human populations', *Science (New York, N.Y.)*, 307(5712), pp. 1072-1079.
- 👤 Hohenauer, T. and Moore, A.W. (2012) 'The Prdm family: expanding roles in stem cells and development', *Development (Cambridge, England)*, 139(13), pp. 2267-2282.
- 👤 Holloway, K., Lawson, V.E. and Jeffreys, A.J. (2006) 'Allelic recombination and de novo deletions in sperm in the human beta-globin gene region', *Human molecular genetics*, 15(7), pp. 1099-1111.
- 👤 Hosking, F.J., Leslie, S., Dilthey, A., Moutsianas, L., Wang, Y., Dobbins, S.E., Papaemmanuil, E., Sheridan, E., Kinsey, S.E., Lightfoot, T., Roman, E., Irving, J.A., Allan, J.M., Taylor, M., Greaves, M., McVean, G. and Houlston, R.S. (2011) 'MHC variation and risk of childhood B-cell precursor acute lymphoblastic leukemia', *Blood*, 117(5), pp. 1633-1640.

- 👤 Houle, A.A., Gibling, H., Lamaze, F.C., Edgington, H.A., Soave, D., Fave, M.J., Agbessi, M., Bruat, V., Stein, L.D. and Awadalla, P. (2018) 'Aberrant PRDM9 expression impacts the pan-cancer genomic landscape', *Genome research*, 28(11), pp. 1611-1620.
- 👤 Howie, B.N., Donnelly, P. and Marchini, J. (2009) 'A flexible and accurate genotype imputation method for the next generation of genome-wide association studies', *PLoS genetics*, 5(6), pp. e1000529.
- 👤 Hunter, N. and Kleckner, N. (2001) 'The single-end invasion: an asymmetric intermediate at the double-strand break to double-holliday junction transition of meiotic recombination', *Cell*, 106(1), pp. 59-70.
- 👤 Hussin, J., Sinnott, D., Casals, F., Idaghdour, Y., Bruat, V., Saillour, V., Healy, J., Grenier, J.C., de Malliard, T., Busche, S., Spinella, J.F., Lariviere, M., Gibson, G., Andersson, A., Holmfeldt, L., Ma, J., Wei, L., Zhang, J., Andelfinger, G., Downing, J.R., Mullighan, C.G. and Awadalla, P. (2013) 'Rare allelic forms of PRDM9 associated with childhood leukemogenesis', *Genome research*, 23(3), pp. 419-430.
- 👤 Inaba, H., Greaves, M. and Mullighan, C.G. (2013) 'Acute lymphoblastic leukaemia', *The Lancet*, 381(9881), pp. 1943-1955.
- 👤 Inoue, K. and Lupski, J.R. (2002) 'Molecular mechanisms for genomic disorders', *Annual review of genomics and human genetics*, 3, pp. 199.
- 👤 International HapMap Consortium (2005) 'A haplotype map of the human genome', *Nature*, 437(7063), pp. 1299-1320.
- 👤 Jain, M., Fiddes, I.T., Miga, K.H., Olsen, H.E., Paten, B. and Akeson, M. (2015) 'Improved data analysis for the MinION nanopore sequencer', *Nature methods*, 12(4), pp. 351-356.
- 👤 Jeffreys, A.J., Holloway, J.K., Kauppi, L., May, C.A., Neumann, R., Slingsby, M.T. and Webb, A.J. (2004) 'Meiotic recombination hot spots and human DNA diversity', *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, 359(1441), pp. 141-152.

- 👤 Jeffreys, A.J. and May, C.A. (2004) 'Intense and highly localized gene conversion activity in human meiotic crossover hot spots', *Nature genetics*, 36(2), pp. 151-156.
- 👤 Jeffreys, A.J. and Neumann, R. (2005) 'Factors influencing recombination frequency and distribution in a human meiotic crossover hotspot', *Human molecular genetics*, 14(15), pp. 2277-2287.
- 👤 Jeffreys, A.J. and Neumann, R. (2002) 'Reciprocal crossover asymmetry and meiotic drive in a human recombination hot spot', *Nature genetics*, 31(3), pp. 267-271.
- 👤 Jeffreys, A.J., Cotton, V.E., Neumann, R. and Lam, K.W. (2013) 'Recombination regulator PRDM9 influences the instability of its own coding sequence in humans', *Proceedings of the National Academy of Sciences of the United States of America*, 110(2), pp. 600-605.
- 👤 Jeffreys, A.J., Tamaki, K., MacLeod, A., Monckton, D.G., Neil, D.L. and Armour, J.A. (1994) 'Complex gene conversion events in germline mutation at human minisatellites.', *Nature genetics*, 6(2), pp. 136-145.
- 👤 Jeffreys, A.J., Kauppi, L. and Neumann, R. (2001) 'Intensely punctate meiotic recombination in the class II region of the major histocompatibility complex', *Nature genetics*, 29(2), pp. 217-222.
- 👤 Johnson, R.E., Washington, M.T., Prakash, S. and Prakash, L. (2000) 'Fidelity of human DNA polymerase ϵ ', *The Journal of biological chemistry*, 275(11), pp. 7447-7450.
- 👤 Kauppi, L., May, C. and Jeffreys, A. (2009) 'Analysis of Meiotic Recombination Products from Human Sperm', in Keeney, S. (ed.) *Humana Press*, pp. 323-355.
- 👤 Kauppi, L., May, C.A. and Jeffreys, A.J. (2009) 'Analysis of meiotic recombination products from human sperm', *Methods in molecular biology* (Clifton, N.J.), 557, pp. 323-355.
- 👤 Keeney, S., Giroux, C.N. and Kleckner, N. (1997) 'Meiosis-specific DNA double-strand breaks are catalyzed by Spo11, a member of a widely conserved protein family', *Cell*, 88(3), pp. 375-384.

- 👤 Khor, C.C. and Hibberd, M.L. (2012) 'Host-pathogen interactions revealed by human genome-wide surveys', *Trends in genetics : TIG*, 28(5), pp. 233-243.
- 👤 Kleckner, N. (1996) 'Meiosis: how could it work?', *Proceedings of the National Academy of Sciences of the United States of America*, 93(16), pp. 8167-8174.
- 👤 Klug, A. (2010) 'The discovery of zinc fingers and their development for practical applications in gene regulation and genome manipulation', *Quarterly reviews of biophysics*, 43(1), pp. 1-21.
- 👤 Kong, A., Gudbjartsson, D.F., Sainz, J., Jonsdottir, G.M., Gudjonsson, S.A., Richardsson, B., Sigurdardottir, S., Barnard, J., Hallbeck, B., Masson, G., Shlien, A., Palsson, S.T., Frigge, M.L., Thorgeirsson, T.E., Gulcher, J.R. and Stefansson, K. (2002) 'A high-resolution recombination map of the human genome', *Nature genetics*, 31(3), pp. 241-247.
- 👤 Kong, A., Thorleifsson, G., Frigge, M.L., Masson, G., Gudbjartsson, D.F., Villemoes, R., Magnusdottir, E., Olafsdottir, S.B., Thorsteinsdottir, U. and Stefansson, K. (2014) 'Common and low-frequency variants associated with genome-wide recombination rate', *Nature genetics*, 46(1), pp. 11-16.
- 👤 Kong, A., Thorleifsson, G., Gudbjartsson, D.F., Masson, G., Sigurdsson, A., Jonasdottir, A., Walters, G.B., Jonasdottir, A., Gylfason, A., Kristinsson, K.T., Gudjonsson, S.A., Frigge, M.L., Helgason, A., Thorsteinsdottir, U. and Stefansson, K. (2010) 'Fine-scale recombination rate differences between sexes, populations and individuals', *Nature*, 467(7319), pp. 1099-1103.
- 👤 Lam, K.W. and Jeffreys, A.J. (2006) 'Processes of copy-number change in human DNA: the dynamics of α -globin gene deletion', *Proceedings of the National Academy of Sciences of the United States of America*, 103(24), pp. 8921-8927.
- 👤 Li, H. (2018) 'Minimap2: pairwise alignment for nucleotide sequences', *Bioinformatics (Oxford, England)*, 34(18), pp. 3094-3100.
- 👤 Li, H. and Durbin, R. (2010) 'Fast and accurate long-read alignment with Burrows-Wheeler transform', *Bioinformatics (Oxford, England)*, 26(5), pp. 589-595.

- 👤 Lichten, M. and Goldman, A.S. (1995) 'Meiotic recombination hotspots', *Annual Review of Genetics*, 29, pp. 423-444.
- 👤 Linet, M.S. and Devesa, S.S. (1991) 'Descriptive epidemiology of childhood leukaemia', *British journal of cancer*, 63(3), pp. 424-429.
- 👤 Liu, B., Yip, R.K. and Zhou, Z. (2012) 'Chromatin remodeling, DNA damage repair and aging', *Current Genomics*, 13(7), pp. 533-547.
- 👤 Liu, D.Y. and Baker, H.W. (1992) 'Tests of human sperm function and fertilization in vitro', *Fertility and sterility*, 58(3), pp. 465-483.
- 👤 Loutradis, D., Theofanakis, C., Anagnostou, E., Mavrogianni, D. and Partsinevelos, G.A. (2012) 'Genetic profile of SNP(s) and ovulation induction', *Current Pharmaceutical Biotechnology*, 13(3), pp. 417-425.
- 👤 Lupski, J.R. (2004) 'Hotspots of homologous recombination in the human genome: not all homologous sequences are equal', *Genome biology*, 5(10), pp. 242-242. Epub 2004 Sep 28.
- 👤 Malke, H. (1984) 'T. Maniatis, E. F. Fritsch and J. Sambrook, *Molecular Cloning: A Laboratory Manual*, X + 545 S., 61 Abb., 28 Tab. Cold Spring Harbor, N. Y. 1982. Cold Spring Harbor Laboratory', *Zeitschrift für allgemeine Mikrobiologie*, 24(1), pp. 32.
- 👤 Malouf, C. and Ottersbach, K. (2018) 'Molecular processes involved in B cell acute lymphoblastic leukaemia', *Cellular and molecular life sciences: CMLS*, 75(3), pp. 417-446.
- 👤 Malécot, G., Blaringhem, L., (1948) *Les mathématiques de l'hérédité*. Paris: Masson & Cie.
- 👤 Mardis, E.R. (2008) 'Next-generation DNA sequencing methods', *Annual review of genomics and human genetics*, 9, pp. 387-402.
- 👤 May, C.A., Shone, A.C., Kalaydjieva, L., Sajantila, A. and Jeffreys, A.J. (2002) 'Crossover clustering and rapid decay of linkage disequilibrium in the Xp/Yp pseudoautosomal gene SHOX.', *Nature genetics*, 31(3), pp. 272-275.
- 👤 May, C.A., Slingsby, M.T. and Jeffreys, A.J. (2008) *Human Recombination Hotspots: Before and After the HapMap Project*. Berlin, Heidelberg: Berlin, Heidelberg: Springer Berlin Heidelberg.

- 👤 May, C., O'Sullivan, J., Ali, I., Dickerson, J., Birch, J., Taylor, M. and Thompson, P. (eds.) ETV6-RUNX1+ Childhood ALL is associated with coding variants in FIGNL1 at the Susceptibility Locus on Chromosome 7p12.2. Unpublished.
- 👤 McGinty, R.J., Rubinstein, R.G., Neil, A.J., Dominska, M., Kiktev, D., Petes, T.D. and Mirkin, S.M. (2017) 'Nanopore sequencing of complex genomic rearrangements in yeast reveals mechanisms of repeat-mediated double-strand break repair', *Genome research*, 27(12), pp. 2072-2082.
- 👤 McVean, G.A., Myers, S.R., Hunt, S., Deloukas, P., Bentley, D.R. and Donnelly, P. (2004) 'The fine-scale structure of recombination rate variation in the human genome', *Science (New York, N.Y.)*, 304(5670), pp. 581-584.
- 👤 Mikheyev, A.S. and Tin, M.M.Y. (2014) 'A first look at the Oxford Nanopore MinION sequencer', *Molecular Ecology Resources*, 14(6), pp. 1097-1102.
- 👤 Monckton, D.G., Neumann, R., Guram, T., Fretwell, N., Tamaki, K., MacLeod, A. and Jeffreys, A.J. (1994) 'Minisatellite mutation rate variation associated with a flanking DNA sequence polymorphism', *Nature genetics*, 8(2), pp. 162-170.
- 👤 Myers, S., Bottolo, L., Freeman, C., McVean, G. and Donnelly, P. (2005) 'A fine-scale map of recombination rates and hotspots across the human genome', *Science (New York, N.Y.)*, 310(5746), pp. 321-324.
- 👤 Myers, S., Freeman, C., Auton, A., Donnelly, P. and McVean, G. (2008) 'A common sequence motif associated with recombination hot spots and genome instability in humans', *Nature genetics*, 40(9), pp. 1124-1129.
- 👤 Nicholas, J.L., Raju, V.M., Timothy, J.D., Constantinidou, C., Saheer, E.G., Wain, J. and Mark, J.P. (2012) 'Performance comparison of benchtop high-throughput sequencing platforms', *Nature biotechnology*, 30(5), pp. 434.
- 👤 Odenthal-Hesse, L., Berg, I.L., Veselis, A., Jeffreys, A.J. and May, C.A. (2014) 'Transmission distortion affecting human noncrossover but not crossover recombination: a hidden source of meiotic drive', *PLoS genetics*, 10(2), pp. e1004106.

- 👤 Ohno, S. (1967) Sex chromosomes and sex-linked genes. Berlin: Springer-Verlag.
- 👤 Ottolini, C.S., Newnham, L., Capalbo, A., Natesan, S.A., Joshi, H.A., Cimadomo, D., Griffin, D.K., Sage, K., Summers, M.C., Thornhill, A.R., Housworth, E., Herbert, A.D., Rienzi, L., Ubaldi, F.M., Handyside, A.H. and Hoffmann, E.R. (2015) 'Genome-wide maps of recombination and chromosome segregation in human oocytes and embryos show selection for maternal recombination rates', *Nature genetics*, 47(7), pp. 727-735.
- 👤 Paigen, K. and Petkov, P. (2010) 'Mammalian recombination hot spots: properties, control and evolution', *Nature reviews.Genetics*, 11(3), pp. 221-233.
- 👤 Papaemmanuil, E., Hosking, F.J., Vijayakrishnan, J., Price, A., Olver, B., Sheridan, E., Kinsey, S.E., Lightfoot, T., Roman, E., Irving, J.A., Allan, J.M., Tomlinson, I.P., Taylor, M., Greaves, M. and Houlston, R.S. (2009) 'Loci on 7p12.2, 10q21.2 and 14q11.2 are associated with risk of childhood acute lymphoblastic leukemia', *Nature genetics*, 41(9), pp. 1006-1010.
- 👤 Paques, F. and Haber, J.E. (1999) 'Multiple pathways of recombination induced by double-strand breaks in *Saccharomyces cerevisiae*', *Microbiology and molecular biology reviews : MMBR*, 63(2), pp. 349-404.
- 👤 Parkin, D.M., Stiller, C.A., Draper, G.J. and Bieber, C.A. (1988) 'The international incidence of childhood cancer', *International journal of cancer*, 42(4), pp. 511-520.
- 👤 Patel, A., Horton, J.R., Wilson, G.G., Zhang, X. and Cheng, X. (2016) 'Structural basis for human PRDM9 action at recombination hot spots', *Genes & development*, 30(3), pp. 257-265.
- 👤 Patel, A., Zhang, X., Blumenthal, R.M. and Cheng, X. (2017) 'Structural basis of human PR/SET domain 9 (PRDM9) allele C-specific recognition of its cognate DNA sequence', *The Journal of biological chemistry*, 292(39), pp. 15994-16002.

- 👤 Payne, A., Holmes, N., Rakyan, V. and Loose, M. (2019) 'BulkVis: a graphical viewer for Oxford nanopore bulk FAST5 files', *Bioinformatics* (Oxford, England), 35(13), pp. 2193-2198.
- 👤 Petes, T.D. (2001) 'Meiotic recombination hot spots and cold spots', *Nature reviews.Genetics*, 2(5), pp. 360-369.
- 👤 Piazza, A. and Heyer, W.D. (2019) 'Homologous Recombination and the Formation of Complex Genomic Rearrangements', *Trends in cell biology*, 29(2), pp. 135-149.
- 👤 Pratto, F., Brick, K., Khil, P., Smagulova, F., Petukhova, G.V. and Camerini-Otero, R.D. (2014) 'DNA recombination. Recombination initiation maps of individual human genomes', *Science* (New York, N.Y.), 346(6211), pp. 1256442.
- 👤 Price, B.D. and D'Andrea, A.D. (2013) 'Chromatin remodeling at DNA double-strand breaks', *Cell*, 152(6), pp. 1344-1354.
- 👤 Pui, C.H., Behm, F.G. and Crist, W.M. (1993) 'Clinical and biologic relevance of immunologic marker studies in childhood acute lymphoblastic leukemia', *Blood*, 82(2), pp. 343.
- 👤 Quail Michael, A., Miriam, S., Paul, C., Otto Thomas, D., Harris Simon, R., Connor Thomas, R., Anna, B., Swerdlow Harold, P. and Yong, G. (2012) 'A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers', *BMC Genomics*, 13(1), pp. 341.
- 👤 Reynolds, A., Qiao, H., Yang, Y., Chen, J.K., Jackson, N., Biswas, K., Holloway, J.K., Baudat, F., de Massy, B., Wang, J., Hoog, C., Cohen, P.E. and Hunter, N. (2013) 'RNF212 is a dosage-sensitive regulator of crossing-over during mammalian meiosis', *Nature genetics*, 45(3), pp. 269-278.
- 👤 Rouyer, F., Simmler, M.C., Johnsson, C., Vergnaud, G., Cooke, H.J. and Weissenbach, J. (2029) 'A gradient of sex linkage in the pseudoautosomal region of the human sex chromosomes', *Nature*, 319(6051), pp. 291-295.
- 👤 Sarbajna, S., Denniff, M., Jeffreys, A.J., Neumann, R., Soler Artigas, M., Veselis, A. and May, C.A. (2012) 'A major recombination hotspot in the XqYq

- pseudoautosomal region gives new insight into processing of human gene conversion events', *Human molecular genetics*, 21(9), pp. 2029-2038.
- 👤 Sarbajna, S., Denniff, M., Jeffreys, A.J., Neumann, R., Soler Artigas, Mar 'ia, Veselis, A. and May, C.A. (2012) 'A major recombination hotspot in the XqYq pseudoautosomal region gives new insight into processing of human gene conversion events.', *Human molecular genetics*, 21(9), pp. 2029-2038.
 - 👤 Shendure, J. and Lieberman Aiden, E. (2012) 'The expanding scope of DNA sequencing', *Nature biotechnology*, 30(11), pp. 1084-1094.
 - 👤 Sherborne, A.L., Hosking, F.J., Prasad, R.B., Kumar, R., Koehler, R., Vijayakrishnan, J., Papaemmanuil, E., Bartram, C.R., Stanulla, M., Schrappe, M., Gast, A., Dobbins, S.E., Ma, Y., Sheridan, E., Taylor, M., Kinsey, S.E., Lightfoot, T., Roman, E., Irving, J.A., Allan, J.M., Moorman, A.V., Harrison, C.J., Tomlinson, I.P., Richards, S., Zimmermann, M., Szalai, C., Semsei, A.F., Erdelyi, D.J., Krajcinovic, M., Sinnett, D., Healy, J., Gonzalez Neira, A., Kawamata, N., Ogawa, S., Koeffler, H.P., Hemminki, K., Greaves, M. and Houlston, R.S. (2010) 'Variation in CDKN2A at 9p21.3 influences childhood acute lymphoblastic leukemia risk', *Nature genetics*, 42(6), pp. 492-494.
 - 👤 Sherry, S.T., Ward, M.H., Kholodov, M., Baker, J., Phan, L., Smigielski, E.M. and Sirotkin, K. (2001) 'dbSNP: the NCBI database of genetic variation', *Nucleic acids research*, 29(1), pp. 308-311.
 - 👤 Shi, Q., Spriggs, E., Field, L.L., Ko, E., Barclay, L. and Martin, R.H. (2001) 'Single sperm typing demonstrates that reduced recombination is associated with the production of aneuploid 24,XY human sperm', *American Journal of Medical Genetics*, 99(1), pp. 34-38.
 - 👤 Shin, S., Kim, Y., Chul Oh, S., Yu, N., Lee, S., Rak Choi, J. and Lee, K. (2017) 'Validation and optimization of the Ion Torrent S5 XL sequencer and OncoPrint workflow for BRCA1 and BRCA2 genetic testing', *Oncotarget*, 8(21), pp. 34858-34866.
 - 👤 Siegel, R.L., Miller, K.D. and Jemal, A. (2016) 'Cancer statistics, 2016', *CA: a cancer journal for clinicians*, 66(1), pp. 7-30.

- 👤 Siegel, R., Naishadham, D. and Jemal, A. (2012) 'Cancer statistics, 2012', CA: a cancer journal for clinicians, 62(1), pp. 10-29.
- 👤 Simmler, M.C., Rouyer, F., Vergnaud, G., Nystrom-Lahti, M., Ngo, K.Y., de la Chapelle, A. and Weissenbach, J. (1985) 'Pseudoautosomal DNA sequences in the pairing region of the human sex chromosomes', Nature, 317(6039), pp. 692-697.
- 👤 Speed, R.M. and Chandley, A.C. (1990) 'Prophase of meiosis in human spermatocytes analysed by EM microspreading in infertile men and their controls and comparisons with human oocytes', Human genetics, 84(6), pp. 547-554.
- 👤 Speicher, M.R., Gwyn Ballard, S. and Ward, D.C. (1996) 'Karyotyping human chromosomes by combinatorial multi-fluor FISH', Nature genetics, 12(4), pp. 368-375.
- 👤 Steinmetz, M., Uematsu, Y. and Lindahl, K.F. (eds.) (1987) Hotspots of homologous recombination in mammalian genomes. Elsevier Trends Journals.
- 👤 Stephens, M. and Scheet, P. (2005) 'Accounting for decay of linkage disequilibrium in haplotype inference and missing-data imputation', American Journal of Human Genetics, 76(3), pp. 449-462.
- 👤 Stephens, M., Smith, N.J. and Donnelly, P. (2001) 'A new statistical method for haplotype reconstruction from population data', American Journal of Human Genetics, 68(4), pp. 978-989.
- 👤 Stiller, C.A., Allen, M.B. and Eatock, E.M. (1995) 'Childhood cancer in Britain: The National Registry of Childhood Tumours and incidence rates 1978–1987', European journal of cancer, 31(12), pp. 2028-2034.
- 👤 Striedner, Y., Schwarz, T., Welte, T., Futschik, A., Rant, U. and Tiemann-Boege, I. (2017) 'The long zinc finger domain of PRDM9 forms a highly stable and long-lived complex with its DNA recognition sequence', Chromosome research : an international journal on the molecular, supramolecular and evolutionary aspects of chromosome biology, 25(2), pp. 155-172.

- 👤 Tamaki, K., May, C.A., Dubrova, Y.E. and Jeffreys, A.J. (1999) 'Extremely complex repeat shuffling during germline mutation at human minisatellite B6.7', *Human molecular genetics*, 8(5), pp. 879-888.
- 👤 Tan, M.H., Austin, C.M., Hammer, M.P., Lee, Y.P., Croft, L.J. and Gan, H.M. (2018) 'Finding Nemo: hybrid assembly with Oxford Nanopore and Illumina reads greatly improves the clownfish (*Amphiprion ocellaris*) genome assembly', *GigaScience*, 7(3), pp. 1-6.
- 👤 Taylor, G.M., Dearden, S., Ravetto, P., Ayres, M., Watson, P., Hussain, A., Greaves, M., Alexander, F., Eden, O.B. and UKCCS Investigators United Kingdom Childhood Cancer Study (2002) 'Genetic susceptibility to childhood common acute lymphoblastic leukaemia is associated with polymorphic peptide-binding pocket profiles in HLA-DPB1*0201', *Human molecular genetics*, 11(14), pp. 1585-1597.
- 👤 Taylor, G.M., Hussain, A., Verhage, V., Thompson, P.D., Fergusson, W.D., Watkins, G., Lightfoot, T., Harrison, C.J., Birch, J.M. and UKCCS Investigators (2009) 'Strong association of the HLA-DP6 supertype with childhood leukaemia is due to a single allele, DPB1*0601', *Leukemia*, 23(5), pp. 863-869.
- 👤 Taylor, G.M., Robinson, M.D., Binchy, A., Birch, J.M., Stevens, R.F., Jones, P.M., Carr, T., Dearden, S. and Gokhale, D.A. (1995) 'Preliminary evidence of an association between HLA-DPB1*0201 and childhood common acute lymphoblastic leukaemia supports an infectious aetiology', *Leukemia*, 9(3), pp. 440-443.
- 👤 Taylor, M., Bergemann, T.L., Hussain, A., Thompson, P.D. and Spector, L. (2011) 'Transmission of HLA-DP variants from parents to children with B-cell precursor acute lymphoblastic leukemia: log-linear analysis using the case-parent design', *Human immunology*, 72(10), pp. 897-903.
- 👤 Taylor, M., Hussain, A., Urayama, K., Chokkalingam, A., Thompson, P., Trachtenberg, E. and Buffler, P. (2009) 'The human major histocompatibility complex and childhood leukemia: an etiological hypothesis based on molecular mimicry', *Blood cells, molecules & diseases*, 42(2), pp. 129-135.

- 👤 Thompson, P. (2015) Pilot project to study the expression of the meiotic regulator, PRDM9, by childhood leukaemias: Final Grant Report to Children with Cancer UK.
- 👤 Thompson, P., Urayama, K., Zheng, J., Yang, P., Ford, M., Buffler, P., Chokkalingam, A., Lightfoot, T. and Taylor, M. (2014) 'Differences in meiotic recombination rates in childhood acute lymphoblastic leukemia at an MHC class II hotspot close to disease associated haplotypes', PLoS one, 9(6), pp. e100480.
- 👤 Tiemann-Boege, I., Calabrese, P., Cochran, D.M., Sokol, R. and Arnheim, N. (2006) 'High-resolution recombination patterns in a region of human chromosome 21 measured by sperm typing', PLoS genetics, 2(5), pp. e70.
- 👤 Trevino, L.R., Yang, W., French, D., Hunger, S.P., Carroll, W.L., Devidas, M., Willman, C., Neale, G., Downing, J., Raimondi, S.C., Pui, C.H., Evans, W.E. and Relling, M.V. (2009) 'Germline genomic variants associated with childhood acute lymphoblastic leukemia', Nature genetics, 41(9), pp. 1001-1005.
- 👤 Tsai, M. (2014) 'How DNA polymerases catalyze DNA replication, repair, and mutation', Biochemistry, 53(17), pp. 2749.
- 👤 Turner, D.J., Miretti, M., Rajan, D., Fiegler, H., Carter, N.P., Blayney, M.L., Beck, S. and Hurles, M.E. (2008) 'Germline rates of de novo meiotic deletions and duplications causing several genomic disorders', Nature genetics, 40(1), pp. 90-95.
- 👤 'The United Kingdom Childhood Cancer Study: objectives, materials and methods. UK Childhood Cancer Study Investigators', (2000) British journal of cancer, 82(5), pp. 1073-1102.
- 👤 Urayama, K.Y., Chokkalingam, A.P., Metayer, C., Hansen, H., May, S., Ramsay, P., Wiemels, J.L., Wiencke, J.K., Trachtenberg, E., Thompson, P., Ishida, Y., Brennan, P., Jolly, K.W., Termuhlen, A.M., Taylor, M., Barcellos, L.F. and Buffler, P.A. (2013) 'SNP association mapping across the extended major histocompatibility complex and risk of B-cell precursor acute lymphoblastic leukemia in children', PLoS one, 8(8), pp. e72557.

- 👤 Urayama, K.Y., Chokkalingam, A.P., Metayer, C., Ma, X., Selvin, S., Barcellos, L.F., Wiemels, J.L., Wiencke, J.K., Taylor, M., Brennan, P., Dahl, G.V., Moonsamy, P., Erlich, H.A., Trachtenberg, E. and Buffler, P.A. (2012) 'HLA-DP genetic variation, proxies for early life immune modulation and childhood acute lymphoblastic leukemia risk', *Blood*, 120(15), pp. 3039-3047.
- 👤 Voelkerding, K.V., Dames, S.A. and Durtschi, J.D. (2009) 'Next-generation sequencing: from basic research to diagnostics', *Clinical chemistry*, 55(4), pp. 641-658.
- 👤 Wolf, G., Greenberg, D. and Macfarlan, T.S. (2015) 'Spotting the enemy within: Targeted silencing of foreign DNA in mammalian genomes by the Kruppel-associated box zinc finger protein family', *Mobile DNA*, 6, pp. 17-8. eCollection 2015.
- 👤 Wolfe, K.H. (1991) 'Mammalian DNA replication: Mutation biases and the mutation rate', *Journal of theoretical biology*, 149(4), pp. 441-451.
- 👤 Wolfe, S.A., Nekludova, L. and Pabo, C.O. (2000) 'DNA recognition by Cys2His2 zinc finger proteins', *Annual Review of Biophysics and Biomolecular Structure*, 29, pp. 183-212.
- 👤 Xie, L., Onysko, J. and Morrison, H. (2018) 'Childhood cancer incidence in Canada: demographic and geographic variation of temporal trends (1992–2010)', *Health Promotion and Chronic Disease Prevention in Canada*, 38(3), pp. 79-115.
- 👤 Yuan, J. and Chen, J. (2013) 'FIGNL1-containing protein complex is required for efficient homologous recombination repair', *Proceedings of the National Academy of Sciences of the United States of America*, 110(26), pp. 10640-10645.
- 👤 Zhang, F., Gu, W., Hurles, M.E. and Lupski, J.R. (2009) 'Copy number variation in human health, disease, and evolution', *Annual review of genomics and human genetics*, 10, pp. 451-481.
- 👤 Zhou, Z., Ni, C., Wu, L., Chen, B., Xu, Y., Zhang, Z., Mu, J., Li, B., Yan, Z., Fu, J., Wang, W., Zhao, L., Dong, J., Sun, X., Kuang, Y., Sang, Q. and Wang, L.

(2019) 'Novel mutations in ZP1, ZP2, and ZP3 cause female infertility due to abnormal zona pellucida formation', Human genetics, 138(4), pp. 327-337.

👤 Zickler, D. and Kleckner, N. (1998) 'The leptotene-zygotene transition of meiosis', Annual Review of Genetics, 32, pp. 619-697.