Recombination Behaviour and Evolutionary History of the Extended Pseudoautosomal Region 1 (ePAR) on the Human Y chromosome

Thesis submitted for the degree of

Doctor of Philosophy

at the University of Leicester

Nitikorn Poriswanish

Department of Genetics and Genome Biology

College of Life Sciences, University of Leicester

September 2018

"If a man will begin with certainties, he shall end in doubts; but if he will be content to begin with doubts, he shall end in certainties."

Francis Bacon. The Advancement of Learning (1605)

"Better than a thousand utterances, comprising useless words, is one single beneficial word, by hearing which one is pacified."

Gautama Buddha. Dhammapada (translated by Narada Maha Thera. 1963)

Abstract

The human X and Y chromosomes are heterologous except at the tips of both arms - the pseudoautosomal regions (PARs). During male meiosis, crossover occurs only within these PARs, primarily in the Xp/Yp PAR1 which serves as an obligatory site of exchange. Although the boundary between PAR1 and the sex-specific region was thought to be conserved among primates, it has recently been found to differ among human males by a translocation of ~110 kb from the X to Y chromosome, generating an extended PAR1 (ePAR). This event occurred at least twice in human evolution. So far there has been limited indirect evidence to show that this extended homology domain actively supports recombination. This study asked if direct proof could be provided by investigating thousands of gametes from each of two ePAR-carrying men for two subregions, selected principally on published male X-chromosomal meiotic double-strand break (DSB) maps. Crossover behaviour between the X and the ePAR-borne segment on the Y is comparable to that of autosomal recombination hotspots for both the distal and proximal sites within the 110-kb extension. Other hallmarks of classical hotspots including evidence of transmission distortion and GC-biased gene conversion are also observed. This study also demonstrates correspondence between male DSB clusters and historical recombination activity in females within this region, as ascertained by linkage disequilibrium analysis. This suggests that this region is similarly primed for crossover in both males and females, although sex-specific differences may also exist. Additionally, extensive resequencing combined with the analysis of the Y phylogeny estimated a minimum crossover rate for ePAR that is at least six times higher than genome average. Finally, an examination of >200,000 Y chromosomes in UK Biobank revealed nine additional examples of ePAR, underlining the recurrent nature of the rearrangement, sponsored by non-allelic homologous recombination between specific dispersed repeats.

Work described in Chapters 3 - 6 of this thesis has been published as follows:

Poriswanish, N., Neumann, R., Wetton, J.H., Wagstaff, J., Larmuseau, M.H.D., Jobling, M.A. and May, C.A. 2018. **Recombination hotspots in an extended human pseudoautosomal domain predicted from double-strand break maps and characterized by sperm-based crossover analysis**. *PLoS Genet*, 14**,** e1007680.

(doi: 10.1371/journal.pgen.1007680)

Acknowledgement

Foremost, this study could not be accomplished without my supervisors, Dr. Celia A. May and Prof. Mark A. Jobling. My gratitude is first expressed to Celia who formed the concept and gave me hands-on training throughout the work. Her intense experience and expertise in the field did guide my blind intelligence to the right destination. And Mark, whom I wholeheartedly respect to his wisdom, had supervised by balancing every factors to reach the optimum outcome. His supervising technique, his keen ideas and his superb literacy skills directed my thesis towards the goal. May I pay tributes to both great teachers who are like wind beneath my wings not only by giving a marvellous guidance to my dullness, but by also supporting in other aspects during the time I was away from home.

Next, I would like to pay my sincerely enormous thanks to Rita Neumann who contributed a lot to my work via her magical laboratory knowledge and skills. Without her, the results might be too hard to analyse. To mention another person who are behind various parts of my work, my gratitude goes to Dr. Jon H. Wetton for his genius and skills as well as kindness and modesty.

This study would be impossible to progress without their bioinformatic helps from: Dr. John Wagstaff, Dr. James Eales, Dr. Xiaoguang (Allan) Xue, Dr. Chiara Batini, Dr. Pille Hallast and Tünde Huszar. John, Chiara, Pille and Tünde shared their knowledge as well as friendship through my time at Leicester. Specifically, I would pay my great thanks to James and Allan together with their group from Manchester led by Prof. Maciej Tomaszewski. Their huge efforts to deal with the big UK Biobank data significantly moved this part ahead.

Another great thanks to the sample contributors would go to Prof. Sir Nilesh Samani, Dr. Veryan Codd and her team at Glenfield Hospital who kindly supply their UK Biobank DNAs; and also to Dr. Maarten H.D. Larmuseau from Ghent who did not only provide the samples from the pioneer work, but also handed me an everlasting friendship. Moreover, some parts of this work were received helps from Jodie Lampert, Marwan El Khoury, Ithisham Ali, Poonam Thakkar and Toby Evans. They deserve mentioning with great thanks.

Some technical supports were from Gurdeep M. Lall, Diana Martin, Rachel Madison, Nicola Butler and Victoria E. Cotton. Though many of them had moved from the department, they are still in memory.

Also with many thanks to my PGR panels, Dr. Nicola J. Royle and Dr. Sandra Belesa who suggested good points along this four years and to the GGB staffs who, more or less, shared their knowledge via lectures, workshops or by personal contacts that are supplementary to my work.

My study cannot be achieved by only working, my life at Leicester was balanced by recreations, sports and warm friendship from people in G2, G3, G7, G18, G19 and Thai friends.

Finally, this study was sponsored by Faculty of Medicine Siriraj Hospital, Mahidol University, Thailand as well as strong supports from Aj. Somboon and Aj. Patcharee who pushed me to the present position. Last but not least, my heart goes to my family and ultimately, this work is dedicated to my parents who believe that education is for a benifit of humanity.

Table of contents

List of tables

List of figures

List of abbreviations

aCGH	Array comparative genomic hybridization
ASD	Average squared difference
ASO	Allele-specific oligonucleotide
ASP	Allele-specific primer
AS-PCR	Allele-specific PCR
ChIP-Seq	Chromatin immunoprecipitation with massively parallel DNA sequencing
CI	Confidence interval
cM	Centimorgan
CNV	Copy number variation
CO	Crossover
DSB	Double-strand break
$D, D', \lvert D' \rvert$	Linkage disequilibrium coefficients
d	Deviance
ePAR	Extended PAR1
ERV	Endogenous retroviral sequences
FISH	Fluorescent in situ hybridization
GSD	Glycogen storage disease
H3K4	Histone3-lysine4
H3K4me3	Histone3-lysine4 trimethylation
H3K9me3	Histone3-lysine9 thrimethylation
HERV	Human endogenous retroviral element
HWE	Hardy-Weinberg equilibrium
Indel	Insertion/deletion variant
kb	Kilobase
K_s	Estimated mean number of synonymous nucleotide substitutions per synonymous site
LD	Linkage disequilibrium
LE	Linkage equilibrium

LINE/L	Long interspersed nuclear element
LOD	Logarithm of odds
LR	Likelihood ratio
LTR6B	Long Terminal Repeat 6B
LTR6B-X	Long Terminal Repeat 6B on the distal X chromosome
LTR6B-YPAR1	Long Terminal Repeat 6B on the proximal part of PAR1 on the Y chromosome
MAF	Minor allele frequency
Mb	Megabase
MJN	Median-joining network
MLE (*L*)	Maximum likelihood estimate
MPS	Massively-parallel sequencing
MSY	Male-specific region of the Y chromosome
Mya	Million years ago
NAHR	Non-allelic homologous recombination
NCO	Non-crossover
PAB	Pseudoautosomal boundary
PABX or PABY	Pseudoautosomal boundary on the X or Y chromosome
PAR	Pseudoautosomal region
PRDM9	PR-domain 9 protein
Q	Phred score
QC	Quality control
RF	Recombination frequency
SCE	Sister chromatid exchange
SNP	Single nucleotide polymorphism
SNV	Single nucleotide variant
SRY	Sex-determining region of the Y chromosome
ssDNA	Single-stranded DNA
SSDS	Single-stranded DNA sequencing
STR	Short tandem repeat
SV	Structural variation
TD	Transmission distortion

TE	Transposable element
TFBS	Transcription factor binding site
T_m	Melting temperature
TMRCA	Time to the most recent common ancestor
Xa	Active X
XAR	X-added region
XCI	X-chromosome inactivation
XCR	X-conserved region
Xi	Inactivated X
XIST	X-inactive specific transcript
Xp or Yp	Short arm of the X or Y chromosome
Xpter/Ypter	Tip of short arms of the X and Y chromosomes
Xq or Yq	Long arm of the X or Y chromosome
Xqter/Yqter	Tip of long arms of the X and Y chromosomes
XTR	X-transposed region
YAR	Y-added region
Y-Hg	Y-haplogroup
YTR	Y-transposed region
θ	Recombination fraction

Chapter 1 Introduction

The sex of humans is determined by a pair of sex chromosomes. Men are the heterogametic sex, carrying two different sex chromosomes, X and Y – so each one is haploid - and the Y chromosome determines male sex in a dominant fashion. In contrast, while women are homogametic and carry diploid homologous X chromosomes. The X and Y chromosomes, though very different in size and structure (discussed further below), retain a segment of true sequence homology at the tips of their short arms that allows crossing-over to occur in male meiosis, permitting faithful segregation of the chromosomes into sperm [Rappold, 1993]. Because of the autosomal-like inheritance of markers in this segment, it is known as the pseudoautosomal region 1 (PAR1) [Burgoyne, 1982]. PAR1 has been long thought of as an evolutionary stable entity, but recently the discovery of a translocation X-chromosome material onto the Y [Mensah *et al.*, 2014] which elongates the distal short arm of the Y chromosome (Yp) adjacent to the boundary of PAR1 has challenged this idea.

This study is focused on this specific translocation on the human Y chromosome, which is named hereafter in this study as the extended PAR1 or "ePAR"; and is aimed to better understand ePAR in terms of its current recombination behaviour and also its recombinational and evolutionary histories.

1.1 Evolution of the human sex chromosomes

Both sex chromosomes originated from an ordinary pair of autosomes [Ohno, 1967], then X-Y differentiation began 240-320 million years ago (Mya) when proto-mammals diverged from the lineage leading to birds. The origin of the sex-determining region, Y (*SRY*) gene [Sinclair *et al.*, 1990] on one of the "proto" sex chromosomes made that chromosome become a proto-Y chromosome. Selection favoured those chromosomes with large segmental inversions that prevented the transfer of *SRY* to the X, since such inversions prevent homologous pairing and exchange via recombination. Later, the sex-chromosomes diverged from each other independently by rearrangement of their structures and gradually degeneration over time of the male-specific region of the Y chromosome (MSY) [Lahn and Page, 1999, Graves *et al.*, 2006] (Figure 1.1).

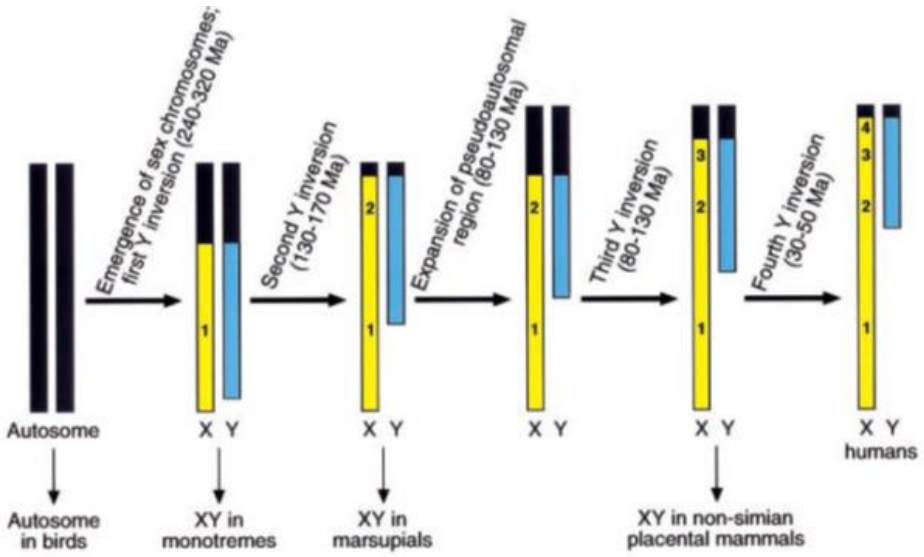


Figure 1.1 Proposed evolutionary events in the origins of the human sex chromosomes. (From [Lahn and Page, 1999])

The "finished" sequence of the human X chromosome (99.3% of the euchromatin) was described as a part of the Human Genome Project [Ross *et al.*, 2005]. It revealed five evolutionary strata on the X chromosome: each stratum is an approximate physical partition that is caused by evolutionary divergence resulting from successive suppression of X-Y recombination, such that each stratum has a characteristic level of X-Y sequence divergence. Four human X-chromosomal strata were first defined by comparing 19 X-Y homologous genes using clustering of the estimated mean number of synonymous nucleotide substitutions per synonymous site (K_s) to estimate evolutionary time [Lahn and Page, 1999]; however, after the finished version of the X chromosome sequence became available, more information could be retrieved and stratum 4 was split in to two, strata 4 and 5 [Ross *et al.*, 2005]. Comparative evolutionary analysis also showed that the X chromosome can be divided into two main evolutionary regions, the X-conserved region (XCR) and the X-added region (XAR). XCR, which occupies the largest portion of the X chromosome including the proximal third of the short arm and most of the long arm, comprises the two most ancient evolutionary strata (1 and 2). XAR comprises strata 3, 4 and 5 [Lahn and Page, 1999, Graves *et al.*, 2006] (Figure 1.2).

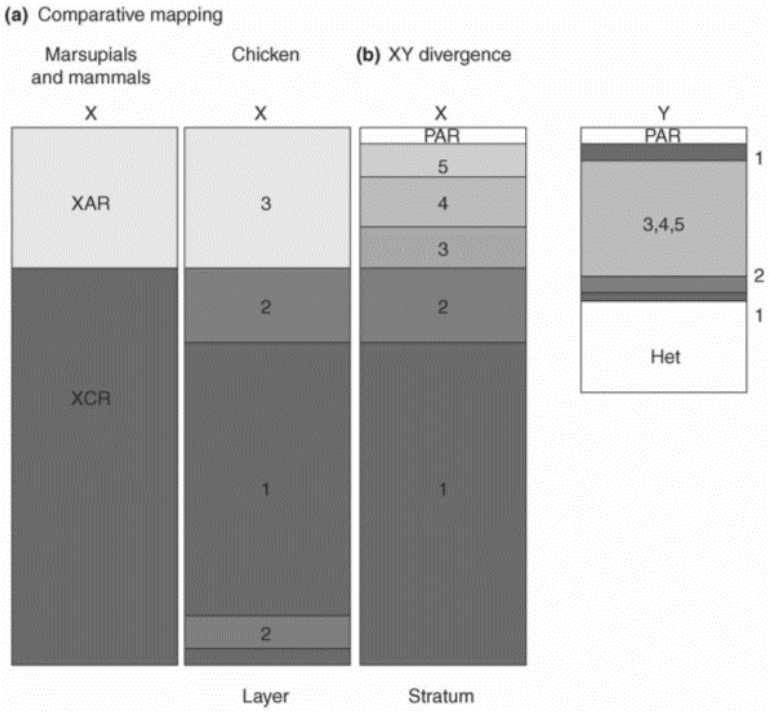


Figure 1.2 **The evolutionary strata in the human sex chromosomes**. The human XCR is orthologous to a bird autosome and is also conserved in mammals, while XAR is orthologous to a marsupial autosome which is thought to have been added to the proto-X chromosome. PAR, an addition from autosomes to both X and Y and is strictly homologous between the two chromosomes. X and Y also share clusters of genes in 5 strata. Strata 1 and 2 coincide in both chromosomes while strata 3-5 in the Y chromosome have been scrambled by multiple later rearrangements. (From [Graves *et al.*, 2006])

The molecular evolutionary history of the human X chromosome reveals that the XCR stratum 1 developed more than 240 Mya; some genes which are orthologous to those on chicken chromosome 4p have been conserved in all mammals until the present day [Ross *et al.*, 2005]. Stratum 2, also part of the XCR, corresponds to Xq28 and the pericentric region of Xp and emerged around 130-170 Mya. Stratum 3 of the XAR was finally added by autosomal translocation around 100-180 Mya and later rearranged and subdivided by inversion at 38-44 and 29-32 Mya into strata 4 and 5, coincidental with prosimian and simian divergence and New and Old World monkey divergence, respectively [Gläser *et al.*, 1997, Lahn and Page, 1999, Graves *et al.*, 2006]. The ~4-Mb X-transposed region (XTR) lies in the mid proximal one-third of the long-arm and is part of stratum 1; by duplication and transposition this region has transferred to the Y since the evolutionary separation between

human and chimpanzees approximately 4.7-6 Mya [Skaletsky *et al.*, 2003]. The corresponding region on the Y chromosome carries an inversion of 200-kb and has a 540-kb deletion to make it shorter than its original copy on X chromosome [Ross *et al.*, 2005] (Figure 1.3).

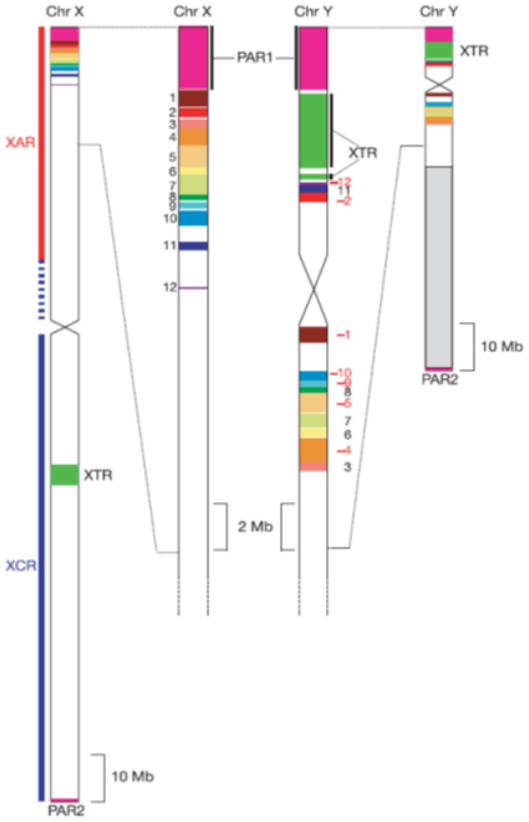


Figure 1.3 Diagram of the major segments of human X-Y homology. The heterochromatic part of the Y chromosome is presented in grey on the right-most representation of the full Y chromosome. In the expanded sections, coloured homology blocks show their corresponding regions on both chromosomes. Some of them are XAR and YAR (PAR1 and block 1-12) but some are the products of duplication from X to Y (XTR and PAR2) since divergence of humans and chimpanzees. Inverted blocks on the Y are represented by the negative numbers shown in red. (From [Ross *et al.*, 2005])

1.2 Dosage balancing of gene expression in the X-specific region between two sexes

Eventually, largely by means of sequence loss and divergence on the Y, both human sex chromosomes became significantly different in size and sequence content. The X chromosome is about 160 Mb long with around 1,098 known genes [Ross *et al.*, 2005], whereas the Y chromosome is much smaller at around 60 Mb in length and has approximately 78 known genes in the male-specific region [Skaletsky *et al.*, 2003]. Moreover, because of the diploid state of female sex chromosomes, one of the X chromosomes becomes inactivated in somatic tissue to compensate gene dosage between the two sexes. X-chromosome inactivation (XCI) is initiated by a non-translated spliced mRNA called X-

inactive specific transcript (XIST) whose gene, *XIST* [Brown *et al.*, 1991], is located at proximal Xq. XIST coating of the future inactive X in *cis* is possibly promoted by the relatively high density of long interspersed nuclear element 1 (LINE-1 or L1) throughout the entire length of X chromosome [Bailey *et al.*, 2000, Lyon, 1999]. The destined inactivated X (Xi) becomes modified into a heterochromatic state and stops transcription of about 85% of its genes [Balaton and Brown, 2016]. The genes on Xi which escape from XCI mostly lie on both tips of the X chromosome within the pseudoautosomal regions (PARs) [Balaton and Brown, 2016]. Recent studies have revealed more details of the mechanisms involved in human XCI including that of a long non-coding RNA called XACT which coats and prevents the active X (Xa) from becoming coated with XIST [Briggs and Reijo Pera, 2014]. Moreover, studies in mice indicate there are many more X-inactivation regulatory RNAs and proteins acting in *cis* (i.e. *Xite*, *Xpr*, *Ftx*, *Xce*), in *trans* (i.e. *Rnf12*, Oct4, Rex1, c-Myc, Klf4), or in as yet to be determined manner (i.e. *Jpx*) [Galupa and Heard, 2015].

1.3 Pseudoautosomal regions, segments supporting homologous-recombination in male sex chromosomes

PAR1 is located at the tip of short arms of X and Y chromosomes (Xpter/Ypter) while PAR2 is at the end of long arms (Xqter/Yqter). PAR2 is found only in humans; in other species the region is X-specific. Mapping of genes in comparison with other mammals including primates confirms its evolutionary history of recent translocation [Ciccodicola *et al.*, 2000]. While the X chromosomes will normally recombine with each other along their lengths, like autosomes, in female meiosis, the PARs maintain this function for the sex chromosomes in male meiosis. The very different pattern of crossover that is restricted to the two PARs in male meiosis, reflects an essential function for correct chromosomal segregation in Meiosis I to the resulting gametes.

PAR1 is the major pseudoautosomal region in humans: it is ~2.7 Mb long [Brown, 1988], contains at least 24 genes and is the site of an obligatory crossing-over event in male meiosis with a recombination rate of around 20-fold the average of that in autosomes [Rappold, 1993]. In contrast, the minor region, PAR2 (330 kb), contains only five known genes and recombination is not necessary or sufficient in this region for proper disjunction of X and Y [Bickmore and Cooke, 1987, Kvaløy *et al.*, 1994]. In the male germline PAR2 shows a ~6-fold higher recombination rate than the average for the autosomes. However, in the female germline, both PARs as well as the X-specific region have similar recombination rates to autosomes [Ross *et al.*, 2005, Morris and Mangs, 2007, Flaquer *et al.*, 2008, Otto *et al.*, 2011].

Fine-scale mapping of the recombination distribution within PAR1 using sperm typing and multi-SNP analysis in both pedigrees and populations unveiled the fact that this region contains some of the most active recombination hotspots in the human genome. Hotspots are 1- to 2-kb wide intervals showing highly elevated rates of crossing over compared with the surrounding DNA [Kauppi *et al.*, 2004]. A highly active hotspot with a peak activity of > 300 cM/Mb has been characterized in the PAR1 *SHOX* gene [May *et al.*, 2002] and high rates of recombination are also seen in the vicinity of the *XG* gene at the proximal end of PAR1 towards its boundary [Hinch *et al.*, 2014]. Patterns of breakdown of linkage disequilibrium (LD) in PAR1 differ among Europeans and Africans suggesting some differences in the location of hotspots in these populations [Hinch *et al.*, 2014]. Despite its secondary role in sex-chromosome segregation, PAR2 also contains some of the most active recombination hotspots such as that within the *SPRY3* gene which locates near its boundary (PAB2) [Sarbajna *et al.*, 2012].

1.4 Organization and function of the genes around the PAR1 boundary (PAB)

The junction of PAR1 and the sex-specific DNA is called the pseudoautosomal boundary (PAB) [Ellis *et al.*, 1989]. Since this boundary lies on both X and Y chromosomes, it can be called PABX or PABY to distinguish them. The PAB region includes the cluster of genes which represents one of the human minor blood groups, Xg and CD99 (the latter was previously called MIC2) [Buckle *et al.*, 1985, Cooke *et al.*, 1985]; proximally adjacent to the *XG* gene is the *GYG2* gene (Figure 1.4).

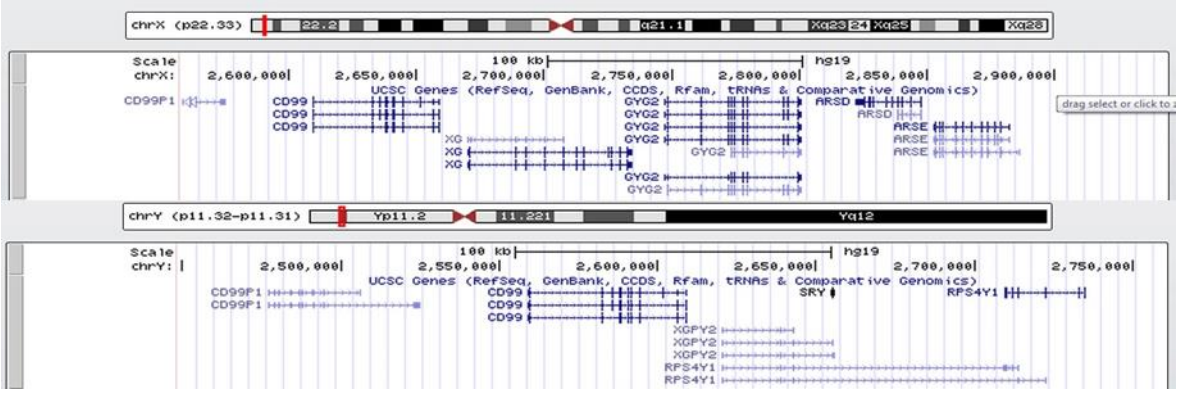


Figure 1.4 **Organisation of genes at the PAR1 boundary, PAB1**. PABX (red on the top) and PABY (red on the bottom) and the genes across these regions. (A screenshot from [UCSC Genome Browser]).

Xg is X-linked and the *XG* gene spans the proximal end of PAR1 across PABX into the X-specific DNA. It consists of 10 exons, the first three of which lie in PAR1, with the rest in the X-specific region [Ellis *et al.*, 1994, Weller *et al.*, 1995]. In males, there is no functional copy of this gene on the Y chromosome as only the first three exons within PAR1 are present. The full-length *XG* gene gives rise to Xg antigen on the surface of red blood cells. There are two alleles, XG*A which is dominant and XG*0 which is recessive or null allele. Females that are homozygous dominant (XG*A/A) or heterozygous (XG*A/0), have the Xg phenotype Xg(a+) i.e. Xg antigen is present on the membrane of their red cells, while homozygous recessive (XG*0/0) females have no Xg antigen and the phenotype Xg. The *XG* gene escapes from the process of XCI in females [Berletch *et al.*, 2011] thus, there should be no random effect. On the other hand, in males, only the *XG* gene on their X chromosome is capable of antigen production as a result of the hemizygous nature of this locus. Surprisingly, the antigenic Xg(a+) dosage shows no difference between male hemizygotes and female homozygotes, and interestingly, they also present stronger Xg(a+) antigenic expression than that of female heterozygotes [Daniel, 2013].

The high rates of recombination around PABX and PABY may affect the Xg phenotype. A number of cases have been reported that did not show clear Mendelian inheritance. One instance involves sons of Xg(a-) mothers (XG* 0/0 genotype) who were found to have Xg(a+) instead of the Xg(a-) that should be inherited from their mothers [Ferguson-Smith *et al.*, 1982]. Some authorities in transfusion medicine, such as Sanger and Race, hypothesised that this might be caused by a translocation of XG*A genotype of the father's X onto the Y and then transmission to the sons [Sanger and Race, 1975]. Therefore, this might cause an

expression of XG*A on Y over XG*0 on the mother's X. Another hypothesis is possibly from an unusual X-Y recombination [Daniel, 2013]. These incidences may therefore reflect examples of aberrant recombination occurring in the vicinity of the major pseudoautosomal boundary.

GYG2, one of the two paralogous glycogenin genes, is involved in the important initiation step of a glycogen biosynthesis by mediating a self-glucosylation [Zeqiraj and Sicheri, 2015, Zhai *et al.*, 2000]. Different from its paralogous *GYG1* gene on Chromosome 3, *GYG2* is predominantly expressed in liver, heart and pancreas rather than ubiquitously especially in skeletal muscle as *GYG1* [Mu *et al.*, 1997, Zhai *et al.*, 2000]. This gene is located in the vicinity of the X-specific region but is thought to have the X-Y homologous pseudogene for some of the exons on the Y chromosome that is the evidence of X-Y coevolution [Zhai *et al.*, 2000]. Similar to other genes outside PARs, it is also undergoes X-inactivation [Stabellini *et al.*, 2009]. Though lack of a gene involved in glycogen metabolism may be associated with some metabolic diseases, for instance, glycogen storage diseases (GSD) or diabetes mellitus [Zhai *et al.*, 2000], deletion of the *GYG2* gene is found to have no clear relation to diabetes [Irgens *et al.*, 2015]. However, in the patients with GSD type XV who are *GYG1*-deficient but do not lack *GYG2*, it has been shown that *GYG2* represents an alternative pathway to rescue glycogen synthesis in skeletal muscle instead of *GYG1*, though to a very mild degree; whereas *GYG2*-deficient individuals with presence of *GYG1* appear healthy [Krag *et al.*, 2017].

1.5 The extended pseudoautosomal region 1 (ePAR)

Recently, there was a discovery of PAR1 length polymorphism that effectively shifts the location of PABY proximally in some males. Direct evidence of the PAR1 plasticity came from a chance discovery during an aCGH (array comparative genomic hybridization) screen for pathogenic copy number variation (CNV) in ~4,300 patients with developmental disorders, which showed that 15 Belgian males carry the rearrangement which is a duplication spanning 98-136 kb on Xp22.33 (Figure 1.5), hereafter known as the extended pseudoautosomal region 1, or ePAR: this demonstrates that the PAR1 boundary is not static, but polymorphic in modern humans [Mensah *et al.*, 2014]. This event has been attributed to non-allelic homologous recombination (NAHR) sponsored by Long Terminal Repeats 6B sequences (LTR6Bs) that results in an insertional translocation of X-specific DNA, an event that has been shown to be recurrent. The segment of inserted X-specific sequence has the coordinates chrX:2,694,151-2,808,549 (GRCh37/hg19) [Mensah *et al.*, 2014].

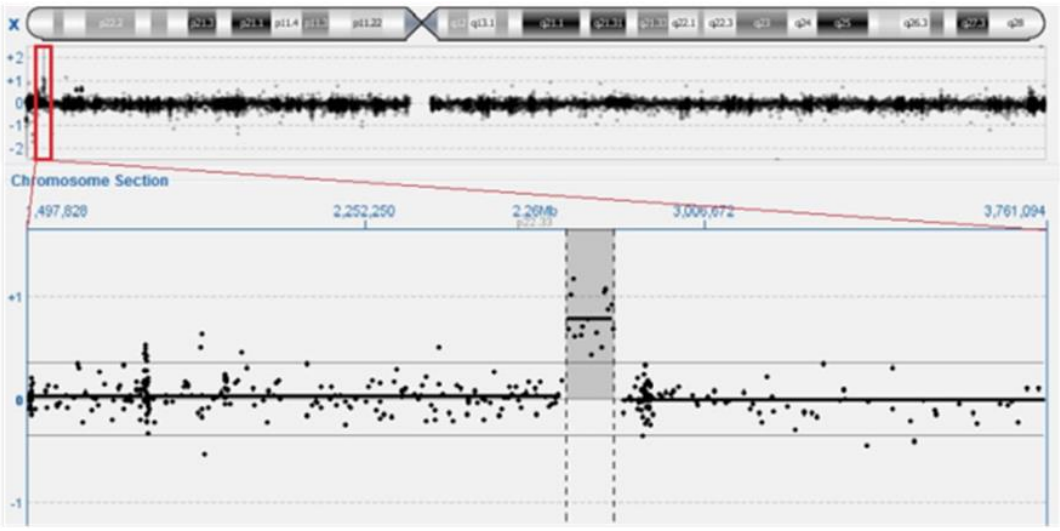


Figure 1.5 **Duplication of X-specific SNPs in Xp22.33 in the Mensah et al. (2014) study**. The area of duplication (red rectangle) covering PABX showed an increase in SNP signals in the X-specific region adjacent to PABX (grey block in zoom-in view) by aCGH using a 180K Custom Microarray. This indicates a duplication of the region. (Adapted from [Mensah *et al.*, 2014])

By using a PCR assay in six families, ePAR was shown to be paternally inherited (Figure 1.6).

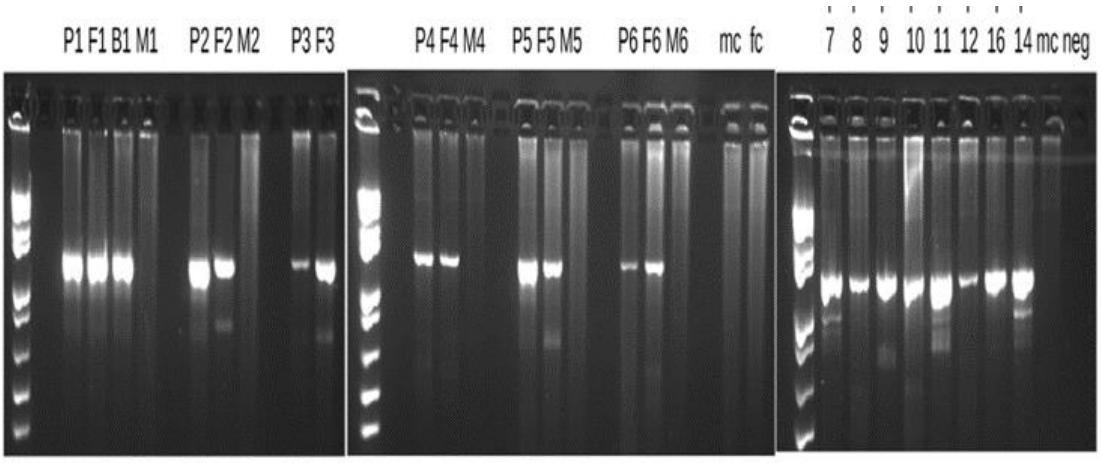


Figure 1.6 **Demonstration of paternal inheritance of the duplication**. PCR bands across the junction showed positive in the male individuals (P) and fathers (F), but not in mothers (M), controls (mc, fc) or negative controls [Nelson *et al.*]. It could therefore be inferred that the extended PAR1 was on the Y chromosome. (From [Mensah *et al.*, 2014])

The ePAR was inferred to be a product of NAHR between LTR6B adjacent to the proximal of the most distal 110-kb X-specific piece (LTR6B-X, chrX: 2,808,549-2,809,097; hg19) and LTR6B adjacent to the distal of the most proximal 5-kb PAR1 on Y chromosome (LTR6B-YPAR1, chrY:2,644,151-2,644,702; hg19). The LTR6Bs are not perfectly identical due to variation in their sizes (549 bp for LTR6B-X *v.* 552 bp for LTR6B-YPAR1) and sequences (SNPs and indels); however, they still share long identical sequences in some parts. In strong support of a NAHR mechanism, the researchers also detected a reciprocal deletion event in the two female daughters and their carrier father (Figure 1.7). The structure of ePAR is demonstrated in Figure 1.7 and details are presented in Appendix.

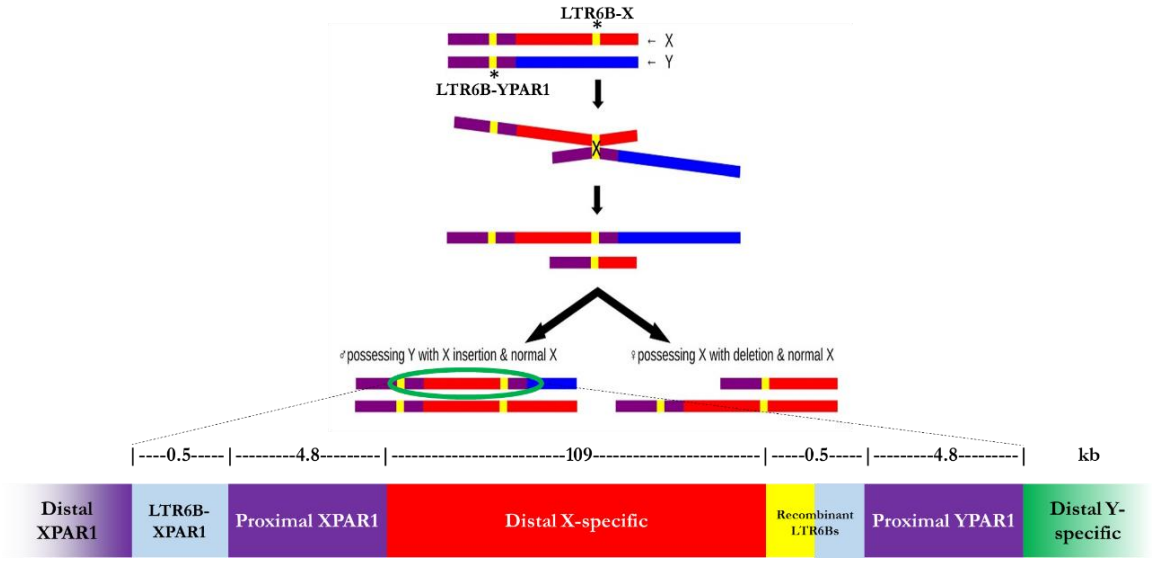


Figure 1.7 **NAHR mechanism of formation of the ePAR and schematic diagram of the 120-kb ePAR structure.** Each composition is named according to the original sequence positions and the diagram is not to scale. (Adapted from [Mensah *et al.*, 2014])

The ePAR is believed to occur recurrently because it was found amongst men falling into two Y-haplogroups (Y-Hg) I2a-P37.2* and R1b-P312* which are phylogenetically distant from one another (Figure 1.8). Details of Y-haplogroups are presented thereafter in the introduction section 6.1.3 in Chapter 6.

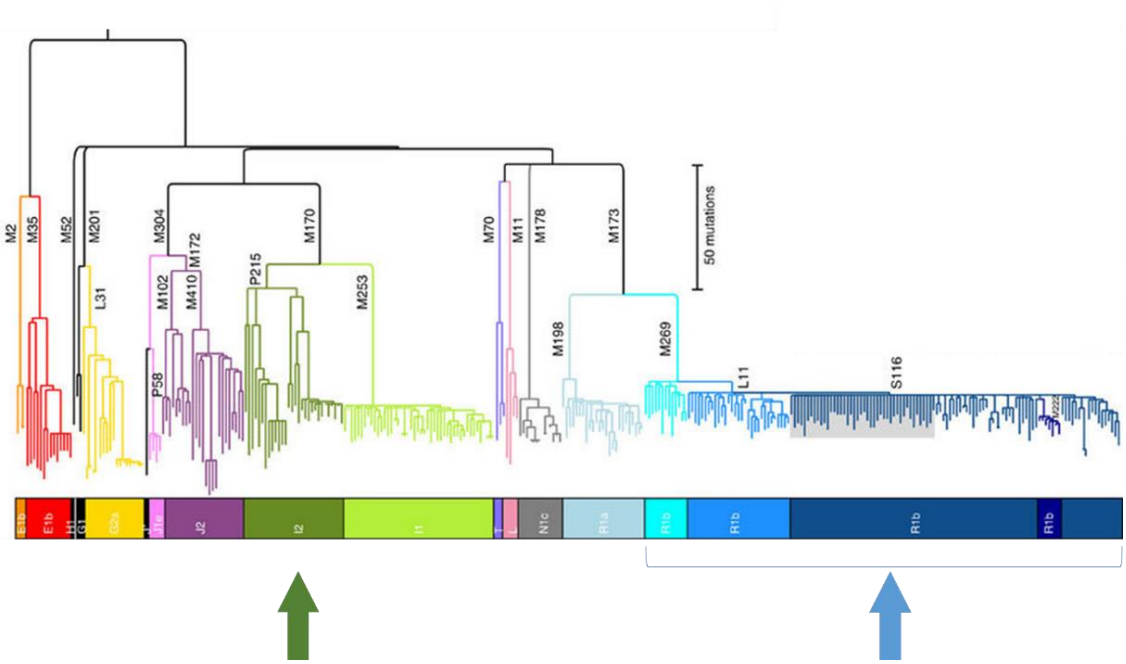


Figure 1.8 **Phylogenetic tree depicting men in Y-Hg I2 and R1b.** The tree shows how distant both Y-Hgs are. Hg I2 is indicated by green arrow while the Hg R1b is indicated by blue arrow. (Adapted from [Batini *et al.*, 2015])

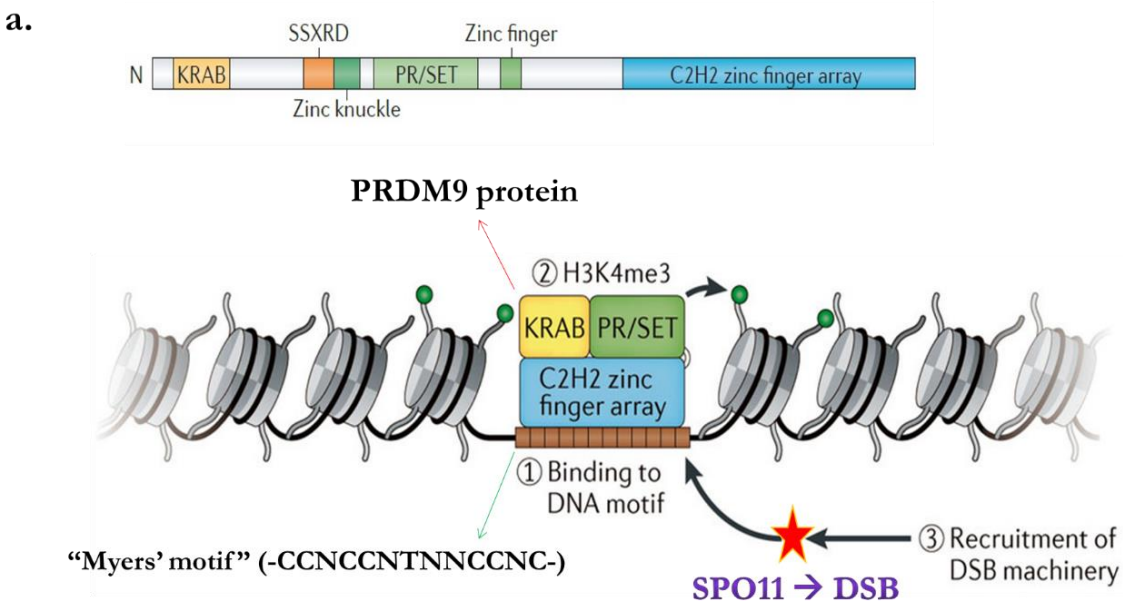
1.6 Meiotic recombination and ePAR

Meiotic recombination has been intensively studied for decades to help elucidate various biological disciplines, including development, cancer, and genetic diseases, and also evolution. Recombination is not only essential to ensure correct chromosomal segregation during gametic cell division, but is also one of the major processes in generating genetic diversity. Moreover, it might be a mechanism to counteract the processes resulting in loss or degradation of genetic material, and prevent organisms from extinction [Aitken and Marshall Graves, 2002, Mensah *et al.*, 2014].

Mensah *et al.* presented indirect evidence that the X-translocated part within ePAR actually functions pseudoautosomally by observing that at least two haplotypes exist among the Y-Hg I2 ePAR men. These findings were interpreted as a consequence of recombination between X and ePAR rather than mutation accumulation because they differed by twelve SNP variants all of which are also observed on the X chromosome [Mensah *et al.*, 2014].

Analysis of recombination clusters in 1- to 2-kb wide hotspots and the finest-scale mapping of hotspots is usually performed by sperm typing [Kauppi *et al.*, 2009]. Even though the initiation of recombination in humans is poorly understood, the most up-to-date knowledge

suggests that it is determined by the histone3-lysine4 (H3K4) *N*-methyltransferase PR-domain 9 protein (PRDM9) [Baudat *et al.*, 2010, Parvanov *et al.*, 2010, Berg *et al.*, 2010] (Figure 1.9). *PRDM9* is a highly polymorphic gene located on the short arm of chromosome 5 in humans. There are many alleles of this gene, defined by the structure of a zinc finger (ZnF) domain, but the most common is the so-called A allele (*PRDM9*A) [Berg *et al.*, 2010]. PRDM9A protein is predicted to bind via its polymorphic tandem array of zinc fingers to a specific DNA sequence called a Myers' motif (-CCNCCNTNNCCNC-) that is present at the centre of ~40% of hotspots identified via patterns of LD [Myers *et al.*, 2008, Myers *et al.*, 2010]. This then initiates the recombination cascade by recruiting several proteins including the topoisomerase-like SPO11 protein [Pratto *et al.*, 2014] (Figure 1.9 a). Commitment to recombination occurs when SPO11 introduces a DNA double-strand break (DSB) and this can be repaired by either a crossover or a non-crossover pathway [Baudat *et al.*, 2013]. Only ~10% of DSBs will be repaired by homologous recombination resulting in a crossover and NAHR (or a misplaced crossover), while the rest may be processed via various mechanisms, as non-crossovers [Pratto *et al.*, 2014] (Figure 1.9 b).



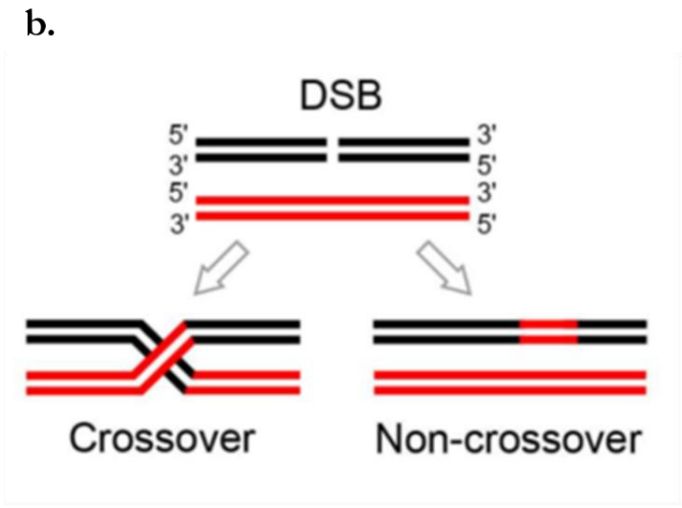


Figure 1.9 **PRDM9 initiating DSB and outcome of DSB repair. a)** shows the structure of PRDM9 protein with Zn finger array (on the top) which binds to Myers' motif on DNA and recruits several proteins, including SPO11 which is one of the major proteins involved in the DSB mechanism. **b)** shows two major repairs of DSB which is resolved by either crossover, found as a minority of events, or non-crossover, as a majority of events. (Figure 1.9 a is adapted from [Baudat *et al.*, 2013] and b is adapted from [Pratto *et al.*, 2014])

1.7 *De novo* DSB mapping enhancing an approach to study recombination hotspots

Pratto *et al.* (2014) used an antibody against DMC1, one of the major meiotic recombination proteins which binds to single-stranded DNA after a DSB event (Figure 1.10), to perform single-stranded DNA sequencing (SSDS) by chromatin immunoprecipitation (ChIP) with massively parallel DNA sequencing (ChIP-Seq). They used human testis tissue to identify DSBs throughout the whole genome. The results showed that DSBs are strongly correlated with recombination hotspots and the range spans 1-2 kb, while the *PRDM9*A allele is stronger in hotspot strength than the other *PRDM9* zinc finger alleles [Pratto *et al.*, 2014]. These data sets are useful to build a preliminary crossover breakpoint map that is much more informative than a previously-performed LD breakdown map, in that it is reporting directly on contemporary distribution of recombination, whereas LD is a population-based measure which is affected by factors other than historical recombination [Kauppi *et al.*, 2009].

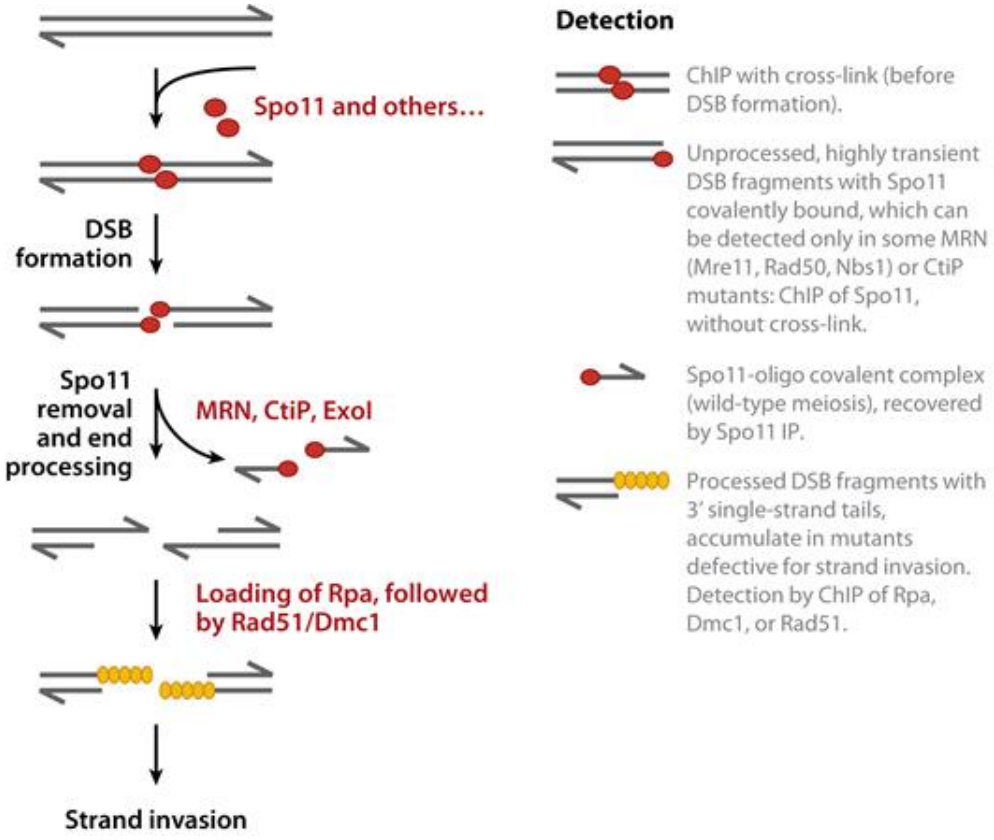


Figure 1.10 **Induction of DSBs and the subsequent events leading to single-stranded DNA invasion.** The left diagram indicates the roles of SPO11 and other proteins which initiate DSB and then several proteins including meiosis-specific Rad51/DMC1 bind to ssDNA ends to start the process of DSB repair. In the right-hand diagram, ChIP-Seq, a method for detecting sequences where DSBs occur, is outlined. The general principle is to use a specific Antibody to extract the specific DSB-related proteins, including Dmc1, and also the DNA sequences which are bound by the protein, which can then be sequenced and mapped back to the reference genome [de Massy, 2013].

1.8 Aims of this study

This study addresses three main research questions. The first aim is to study directly the recombination behaviour of the ePAR. There is indirect evidence that recombination occurs in the region [Mensah *et al.*, 2014], but no direct evidence, and no information on the nature of the crossover process. This aim was approached by using PCR-based analysis of sperm DNA to analyse recombination events in males who carry the ePAR.

The second aim is to ask whether full ePAR-haplotypes reflect a history of recombination activity, including hotspots. The aim was approached by using massively parallel sequencing analysis (Ion Torrent™) and phasing X-haplotypes on the ePAR in each family; then making a comparison to the known-phased X-haplotypes from publicly available data (1000 Genomes Project) together with estimating the pedigree recombination rate through a phylogenetic approach.

The third aim is to ask if other ePARs than those already discovered, independently generated by NAHR, can be detected and characterised. This included searching in the very large population of UK Biobank, and the characterisation of novel ePARs by Y haplotyping and junction sequence analysis.

Chapter 2 Materials and Methods

2.1 Materials

2.1.1 Chemical reagents and enzymes

- Water: Molecular Biology Reagent (dH_2O) (Sigma-Aldrich)
- 5 U/μl *Taq* DNA polymerase (Bioline)
- 2.5 U/μl *Pfu* DNA polymerase (Thermo Fisher Scientific)
- 1M Tris base
- 11.1× PCR buffer, recipe shown as follows [Kauppi *et al.*, 2009]:
  a. 2M Tris-HCl pH 8.8 (ratio of input:total volume =167:676)
  b. 1M $(NH4)_2SO_4$ (ratio of input:total volume = 83:676)
  c. 1M $MgCl_2$ (ratio of input:total volume = 33.5:676)
  d. 100% 2-Mercaptoethanol (ratio of input:total volume = 3.6:676)
  e. 10 mM EDTA pH 8.0 (ratio of input:total volume = 3.4:676)
  f. 100 mM dATP (ratio of input:total volume = 75:676)
  g. 100 mM dCTP (ratio of input:total volume = 75:676)
  h. 100 mM dGTP (ratio of input:total volume = 75:676)
  i. 100 mM dTTP (ratio of input:total volume = 75:676)
  j. 10 mg/ml Bovine Serum Albumin (BSA) (Ambion) (ratio of input:total volume = 85:676)
- 1M Tris-HCl pH 7.5
- 10% (w/v) SDS
- 1× Tris-Borate EDTA (TBE) with 0.5 μg/ml Ethidium Bromide (EtBr)
- 100% Glycerol
- SeaKem® LE Agarose (Lonza)
- Loading dye: Bromophenol Blue (Bioline)
- ΦX174 DNA/BsuRI (HaeIII digest) marker (72-1,353 bp) (Thermo Fisher Scientific)
- λ DNA (HindIII digest) marker (125-23,130 bp) (Thermo Fisher Scientific)
- BigDye® Terminator v3.1 Ready Reaction Mix (Life Technologies)
- BigDye® Terminator v3.1 5× Sequencing Buffer (Life Technologies)
- PowerPlex® Y23 5× Master Mix (Promega)

- PowerPlex® Y23 10× Primer Pair Mix (Promega)
- PowerPlex® Y23 Allelic Ladder Mix (Promega)
- CC5 Internal Lane Standard 500 Y23 (Promega)
- Hi-Di™ formamide (Applied Biosystems)
- 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)
- Exonuclease I (ExoI) (New England Biolabs)
- Shrimp Alkaline Phosphatase (SAP) (New England Biolabs)
- SNaPshot® single-base extension assay (Thermo Fisher Scientific)
- GeneScan™ 120 LIZ® Size Standard (Thermo Fisher Scientific)
- 20× NaCl-Sodium citrate buffer (SSC): 3 M NaCl, 0.3 M sodium citrate-dihydrate [Kauppi *et al.*, 2009]
- 10 M NaOH
- 5 M NaCl
- 0.5 M EDTA
- 0.5M spermidine trichloride
- 0.5M disodium EDTA
- 0.5M dithiothreitol
- 100mM ATP
- 0.1M Na_3PO_4 pH 6.8
- 10 mg/mL yeast RNA
- T4 polynucleotide kinase (New England Biolabs)
- T4 kinase labelling mix according to [Kauppi *et al.*, 2009]
- T4 kinase stop mix according to [Kauppi *et al.*, 2009]
- 10 mCi/mL γ-^{32}P ATP (Perkin-Elmer)
- 50× Denhardt's solution: 5g Ficoll® PM 400 (Sigma-Aldrich), 5g polyvinylpyrrolidone, 5g BSA, 500 mL dH_2O [Kauppi *et al.*, 2009]
- 3 mg/mL high-molecular-weight salmon-sperm DNA in dH_2O
- 5M Tetramethylammonium chloride (TMAC) (Sigma-Aldrich)
- Denaturing mix according to [Kauppi *et al.*, 2009]

- TMAC hybridisation solution according to [Kauppi *et al.*, 2009]
- TMAC wash solution according to [Kauppi *et al.*, 2009]
- Autoradiograph-film developing solutions (developer, stop and fixer)

2.1.2 Oligonucleotides

Oligonucleotides used in this study were primers and hybridising probes. Primers were mostly synthesised by Sigma-Aldrich, except some already-available ones which had been synthesised by the University of Leicester Facility. All of the oligoprobes were synthesised by Sigma-Aldrich. Purification of the oligonucleotides was by the desalting method except for the long-polyadenylated primers used for the SNaPshot® Multiplex System which required HPLC purification. Names and details are presented in Appendix.

2.1.3 Kits

- Zymoclean™ Gel DNA Recovery Kit (Zymo Research)
- QIAamp® DNA Mini Kit (Qiagen)
- Performa® Gel Filtration Cartridge (EdgeBio)
- E.Z.N.A.® Cycle-Pure Kit (Omega Bio-Tek)
- Agilent DNA 1000 chip (Agilent)
- Ion 316™ Chip Kit v2 BC (Thermo Fisher Scientific)

2.1.4 Plasticwares, film, paper and membranes

- DNA LoBind tubes (Eppendorf)
- Screw-top 1.5-mL microcentrifuge tubes
- PCR tubes
- 8-tube strips
- 96-well plates and septa
- 0.1–10, 1–200 and 100–1,000 μl pipette tips
- Whatman™ 3MM blotting paper
- Nylon membrane filter (Bio-Rad)
- X-ray films (Fuji)

2.1.5 Instruments and apparatus

- Level-2 laminar-flow hood
- Heat block and water bath
- Centrifuge
- Veriti® thermal cycler (Life Technology)
- DNA Engine Tetrad® 2 (Bio-Rad)
- Agarose gel trey, comb, tank, and power generator
- GeneGenius Gel Imaging System (Syngene)
- Applied Biosystems® 3130xl Genetic Analyzer
- 2100 Bioanalyzer (Agilent)
- Ion Torrent™ PGM™ 316 Sequencer
- Dot blot mannifold
- Hybridisation tubes
- Pumping reservoir-flask
- Suction pump
- Geiger counter

2.1.6 Databases

- Homo sapiens reference sequence in GRCh37/hg19 assembly was downloaded from UCSC Genome Browser including SNPs identification (https://genome.ucsc.edu).
- *PRDM9A* motif coordinates along chrX:2,694,150-2,809,097 were provided by Dr John Wagstaff (bioinformatician, University of Leicester).
- SNPs were checked in NCBI SNP database (https://www.ncbi.nlm.nih.gov/projects/SNP/).
- LTR6B sequences were checked in Human LTR-retrotransposon Genome Browser (http://herv.pparser.net/GenomeBrowser.php?x=171266043&s=3&GeneTrGen=1&chr=chr1&rep=491).
- Transcription factor LTR6B database was obtain via the dbHERVs-Res database (http://herv-tfbs.com).

- DSB database was downloaded from Pratto et al.'s supplemental data [Pratto *et al.*, 2014].
- Individuals carrying genome duplication at the X-translocated region of ePAR were searched through DGV database (http://dgv.tcag.ca/dgv/app/home)
- Recombination hotspot data were downloaded from International HapMap Project phase II (http://hapmap.ncbi.nlm.nih.gov/downloads/recombination/2011-01_phaseII_B37/).
- vcf files from CEU and GBR populations were taken from the 1000 Genomes Project (ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/ALL.chrX.phpha3_shapeit2_mvncall_integrated_v1b.20130502.genotypes.vcf.gz)

2.1.7 Softwares

- Primer design was guided for preliminary Tm°, possibility of secondary structure formation and possibility of primer dimer formation by: http://www.oligoevaluator.com/Login.jsp
- Interaction between primers was guided by: https://www.lifetechnologies.com/uk/en/home/brands/thermo-scientific/molecular-biology/molecular-biology-learning-center/molecular-biology-resource-library/thermo-scientific-web-tools/multiple-primer-analyzer.html
- DNA sequence reverse-complement was converted by: http://www.bioinformatics.org/sms/rev_comp.html
- Primer sequence specificity was checked via UCSC BLAT Search Genome: http://genome.ucsc.edu/cgi-bin/hgBlat?command=start
- Sequence format converter helps converting a sequence in to a file via : http://genome.nci.nih.gov/tools/reformat.html
- Alignment of sequence was done via BioEdit v.7.2.5, downloaded via: http://www.mbio.ncsu.edu/bioedit/bioedit.html

- Y-STR profiles were analyzed by GeneMapper® ID v.4.0 (Life Technologies)
- Y-STR profiles were used to predict Y-Hg via the online software NEVGEN (http://www.nevgen.org/), batch capability function in the desktop version of NevGen Genealogy Tools Software v1.1, and also via the personal program in Microsoft® Excel by Dr. Jon H. Wetton (Department of Genetics & Genome Biology, University of Leicester).
- Statistical analysis and line plotting was done using Microsoft® Excel 2010, SPSS v. 10 (IBM), Prism 7 (GraphPad) and online: https://www.causascientia.org/math_stat/ProportionCI.html (95% CI by Bayesian approach) and https://www.graphpad.com/quickcalcs/binomial1.cfm (binomial test)
- Ion Torrent sequence data were analyzed using Torrent Suite™ Software 5.0.2
- Bam files were visualised via IGV 2.3 (https://software.broadinstitute.org/software/igv/)
- vcf files were generated from bam files by SAMtools Software 1.3.2 (http://samtools.sourceforge.net/)
- vcf files were extracted using the Data Slicer Tool from Ensembl (http://grch37.ensembl.org/Homo_sapiens/Tools/DataSlicer?db=core)
- Haplotype phasing was performed using PHASE Software 2.1.1 (http://stephenslab.uchicago.edu/phase/download.html)
- Median Joining Networks were generated and presented using Network 5 and Network Publisher Software (http://www.fluxus-engineering.com/)
- Venn diagrams were created using the online software from University of Ghent (http://bioinformatics.psb.ugent.be/webtools/Venn/)
- LD and LOD scores were analyzed suing the software written and run in True Basic® v.4.1 language by Prof. Sir Alec Jeffreys
- Sperm recombination data were analyzed using the software written and run in True Basic® v.4.1 language by Prof. Sir Alec Jeffreys

2.1.8 Samples and ethical approval

- North European semen samples were collected with informed consent, and ethical approval for their use in recombination studies has been granted to Dr. Celia A. May by NRES-East Midlands (REC ref. 6659). Sperm DNA was prepared as described in [Kauppi *et al.*, 2009].

- Additional DNA samples were also collected with informed consent following University of Leicester ethical review (refs.: maj4-46d9 and maj4-cb66).

- Blood DNA samples originally analysed in [Mensah *et al.*, 2014] were part of an institutional genome-wide CNV study that was approved by KU Leuven review board (protocol number S55513).

- Lymphoblastoid cell-line DNA from CEPH family 1334 is available from the Coriell Institute (https://www.coriell.org/).

- Monochromosomal hybrid cell-line DNA controls were obtained from the Coriell Institute (https://www.coriell.org/0/Sections/Collections/NIGMS/Map02.aspx?PgId=496)

- Data from UK Biobank were made available as a collaboration with Prof. Maciej Tomaszewski of the University of Manchester, as part of UK Biobank Main Application 15915 'Paternal lineages of the Y chromosome and men's health and disease'. Access was agreed under a Material Transfer Agreement to UK Biobank to DNA samples held under UK Biobank Main Application 6077, led by Prof. Sir Nilesh Samani and Dr. Veryan Codd at the Department of Cardiovascular Sciences, Glenfield Hospital, University of Leicester.

2.2 Methods

2.2.1 PCR

Extracted DNA was diluted with dH_2O or 5mM Tris-HCl pH 7.5 to 20 ng/μl concentration. Preparation of reactions was done in a level-2 laminar-flow hood. PCR from genomic DNA was prepared for a total reaction volume of 10 μl as follows:

1) 11× PCR buffer → final conc. 1×
2) 1M Tris base → final conc. 12.5 mM
3) 10μM F-primer → final conc. 0.3 μM
4) 10μM R-primer → final conc. 0.3 μM
5) 20:1 *Taq*:*Pfu* → final conc. 0.03 U/μl *Taq* : 0.0015U/μl *Pfu*
6) DNA → final conc. 5-10 ng/μl
7) dH_2O → to make the total volume of 10 μl

2.2.2 Agarose gel electrophoresis

0.8% (w/v) (for PCR products > 0.8 kb in size) or 1.6 – 1.8% (w/v) (for PCR products < 0.8 kb in size) agarose gel were prepared in 1× TBE with 0.5 μg/ml EtBr. Set gels were immersed in a TBE-filled gel running tank. ΦX174 and/or λ DNA markers were diluted in loading dye to give an amount of 50 ng of the 1078-bp band of ΦX174 and/or 9416 band of λ DNA when loading 6 μl of marker mixture. In cases requiring a ΦX174 and λ DNA-marker mix, both were mixed together in loading dye (each marker 1:1:2 loading dye:8 dH_2O) to give the same overall concentration of each marker. To check amplicons, finished gels were visualised under UV light (λ ~302 nm) and captured using the GeneGenius Gel Imaging System.

2.2.3 Purifying PCR products for sequencing

Each amplicon was excised from the gel using a blue-light transilluminator (Dark Reader™), and a Zymoclean™ Gel DNA Recovery Kit was used for purification using the manufacturer's protocol. To purify a batch of PCR amplicons, the E.Z.N.A.® Cycle-Pure Kit was used as per the manufacturer's protocol. Any product which failed to purify by this column kit was purified by gel excision. Purified DNA was finally diluted with dH_2O or 5mM Tris-HCl to 10 ng/μl concentration.

2.2.4 Sequencing reactions

Purified PCR product was prepared for a sequencing reaction (total volume 20 μl/reaction) using the manufacturer's protocol and an input 20-30 ng/kb of purified DNA. The sequencing protocol was run using manufacturer's conditions. After the sequencing reaction, excess dye removal was performed by this protocol:

1) Add 2 μl of 2.2% (w/v) SDS and mix
2) Incubate in PCR machine: 98°C-5 min, 25°C-10 min
3) Use Performa® Gel Filtration Cartridge under the manufacturer's protocol

Finally, the ready reaction as then run on an Applied Biosystems 3700 Genetic Analyzer by the Protein and Nucleic Acid Laboratory of the University of Leicester (PNACL), and data returned via email.

2.2.5 Y-STR typing

PCRs were prepared in a total volume of 25 μl/reaction using the PowerPlex® Y23 PCR reaction kit using the manufacturer's protocol. Genomic DNA was diluted to 1 ng/μl and 0.5 ng taken per reaction. When not using PCR product immediately, it was stored at -20°C. One microlitre of ready reaction was prepared to run on an Applied Biosystems 3130xl Genetic Analyzer capillary electrophoresis apparatus by adding 1 μl of CC5 Internal Lane Standard 500 Y23, 10 μl of Hi-Di™ formamide and 1 μl of PowerPlex® Y23 Allelic Ladder Mix in each well according to the manufacturer's protocol. The ready-to-run product in 96-well plates was then denatured in a PCR machine at 95°C for 3 minutes, then immediately chilled on crushed ice or a freezer plate block or in an ice-water bath for 3 minutes. This process was done just prior to loading the instrument. The product was run through a 36-cm electrophoresis capillary filled with POP-4® polymer in the ABI 3130xl Genetic Analyzer, using an injection voltage of 1-3 kV; fluorescent emission was detected via a CCD camera. Following the run, the run data were analyzed by GeneMapper® ID v.4.0 Software.

2.2.6 SNP sub-typing in Y-chromosome haplogroup I2a

A multiplex PCR encompassing nine haplogroup-identifying SNPs within I2a was developed with annealing and extension temperatures as presented in the Appendix. The resulting products were subsequently treated by ExoI together with SAP to remove excess primers and inactivate nucleotide triphosphates, then used in a SNaPshot® (Applied Biosystems) single-base extension assay using the primer mix detailed in Appendix, according to the

manufacturer's instructions. The SNaPshot® product was finally treated with SAP as per the SNaPshot® protocol and the final product prepared to run on Applied Biosystems 3130xl Genetic Analyzer capillary electrophoresis system by adding 1 μl of GeneScan™ 120 LIZ® Size Standard and 10 μl of Hi-Di™ formamide into each well according to the manufacturer's protocol. Prior to loading, the ready-prepared assay in the 96-well plate was denatured in a PCR machine as for Y-STR typing. The data were analyzed by GeneMapper® ID v.4.0 Software. The phylogenetic relationships of the haplogroups detected by the SNaPshot® assay are as shown in Figure 3.4 (Chapter 3) and Figure 6.2 (Chapter 6).

2.2.7 Ion Torrent™ sequence analysis of the ePAR

Overlapping long-PCR amplicons were designed to cover the ePAR (details of the primer pairs are given in the Appendix). The amplicons were pooled equimolar for each individual in two sets, cleaned with Agencourt® AMPure XP beads and used to make individual-specific libraries using the Ion Xpress™ Library kit and barcodes according to the manufacturer's instructions for 400-bp sequencing. Libraries were size-selected on 1.8% (w/v) LE agarose, gel-purified using a Zymoclean™ Gel DNA Recovery Kit, quantified using an Agilent 2100 Bioanalyzer with Agilent DNA 1000 chip and pooled equimolar. Sequencing templates were prepared using the Ion PGM™ Hi-Q View OT2 Kit and sequencing was performed according to manufacturer's instructions in two runs on an Ion Torrent™ PGM™ 316 Sequencer using the Ion PGM™ HI-Q Sequencing Kit and Ion 316™ Chip Kit v2 BC (Thermo Fisher Scientific). Reads were mapped to the human reference sequence (hg19) using the Torrent Suite™ Software 5.0.2.

2.2.8 Phasing of the ePAR

Variant calls were generated by SAMtools 1.3.2 using the bam files and selecting only reads with a minimum mapping quality of 50 and a minimum base quality of 20. The variant calls from the two runs were merged for each individual. Inclusion of female samples and appropriate monochromosomal hybrid cell-line DNA controls at the template preparation stage indicated that despite careful design of primer pairs, it was impossible to prevent amplification of portions of Yq11.2; genotype calls for these regions were therefore excluded from further analysis along with indels and markers mapping to tandemly repetitive sequences. Haplotypes were derived using the program PHASE 2.1.1 Software [Stephens and Donnelly, 2003, Stephens *et al.*, 2001]. Phased X haplotypes over the interval involved

in the ePAR translocation were obtained from the CEU and GBR males from the 1000 Genomes Project [1000 GP III FTP] for comparison.

2.2.9 TMRCA of haplogroup I2a ePARs

A median-joining Y-STR network of the haplogroup I2a ePARs was constructed using the Network 5 and Network Publisher softwares; the bilocal DYS385a and b was included because these Y chromosomes are closely related and the 'phasing' issue can be ignored. The TMRCA was estimated from the 23 Y-STR data using the ASD method [Goldstein *et al.*, 1995a, Goldstein *et al.*, 1995b], assuming a generation time of 31 yrs [Fenner, 2005].

2.2.10 Detection of sperm *de novo* recombinants

Assays capable of detecting *de novo* reciprocal crossovers spanning the most distal DSB cluster, called in this study crossover assays, were designed for each semen donor following the guidelines in [Kauppi *et al.*, 2009]. Similarly, assays able to simultaneously detect reciprocal *de novo* crossovers as well as non-crossover gene conversion events, called a half-crossover assay in this study, were designed for the proximal target region [Kauppi *et al.*, 2009]. Details of the allele-specific primers (ASPs) directed against SNP variants used for recombinant selection are given in the Appendix. Phasing of these markers was established empirically using ASP-derived amplicons as templates for allele specific oligonucleotide (ASO) typing [Kauppi *et al.*, 2009]. *De novo* recombinants were also characterised by the same method. Details of ASOs are given in Appendix. ASO hybridisations were done as described in [Kauppi *et al.*, 2009], however, for a half-crossover assay, it is necessary to set up control series by mixing the secondary PCR products of the opposite haplotypes prior to dot-blotting onto nylon membrane. To do this, in one haplotype given as Haplotype B, four adjacent wells, namely here Wells 1 – 4, are set to be a control series and will be subtracted from the analysis. Then secondary PCR products from two wells of Haplotype A are subjected to be taken to the control wells by taking 1.5 μl to the outmost well, i.e. Well 4, in the control series to make ~9% (given the prior total volume of 15 μl) and mixed thoroughly. Then taking 1.5 μl of the mixed 9% of Haplotype A in B to the well next to the adjacent well, i.e. Well 2, and mixed; now the concentration of Haplotype A in B of Well 3 is ~0.9%. The further two control wells were similarly made by taking 0.5 μl from the new well in Haplotype A to Well 3 in Haplotype B then mixed to make ~3.2% of Haplotype A in B and transfer 1.5 μl of the mix to Well 1 to make ~0.3%. Similarly, the same operations were done *vice versa* from Haplotype B to A. This complicated setting up of control series is depicted in

Figure 2.1. Moreover, a half-crossover assay must be hybridised exclusively with ASOs of the opposite haplotype.

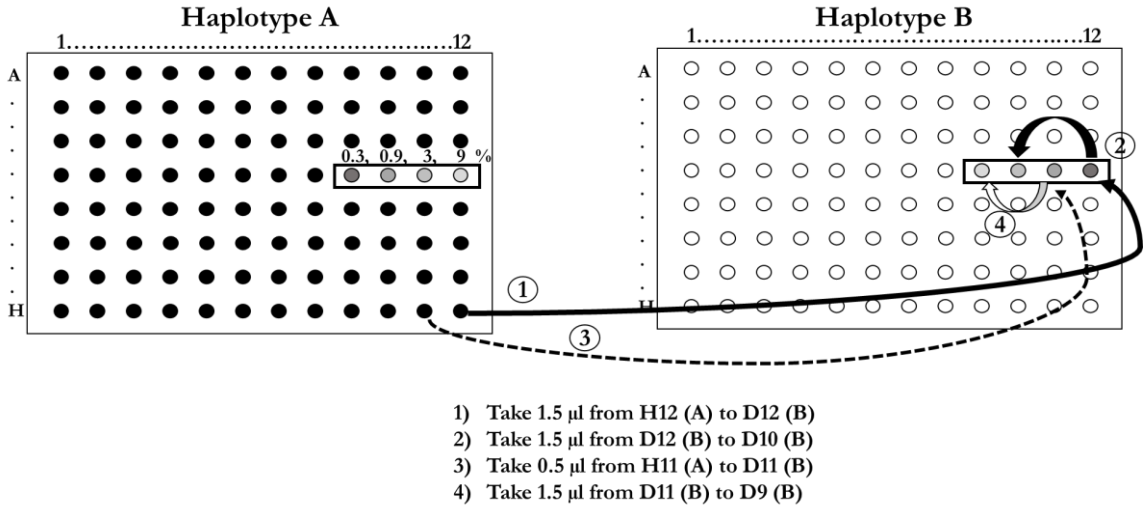


Figure 2.1 Set control series of both haplotypes in a half-crossover assay. Diagram demonstrates a preparation of a control series in the Haplotype B plate by taking the secondary PCR product from the opposite Haplotype A plate. Arrows with numerals indicate the procedure step-by-step. Control series in both haplotypes are shown inside rectangles. For Haplotype A, the process follows the same procedure *vice versa* and the final concentrations of Haplotype B in each control well of Haplotype A are indicated above the series.

Chapter 3 Confirmation of ePAR status of semen donors and diagnostic PCR assay development

3.1 Introduction

The formation of ePAR is an example of NAHR mediated by endogenous retroviral sequences (ERVs) – a process that plays an important role in shaping the evolution of genomes [Cordaux and Batzer, 2009, Hughes and Coffin, 2005, Nelson *et al.*, 2003, Tristem, 2000, Trombetta *et al.*, 2017, Vitte and Panaud, 2003, Yi *et al.*, 2004]. As well as ePAR, other ERV-mediated processes affecting the Y chromosome are known: recurrent ~780-kb deletions of the *AZFa* region, which cause male infertility [Blanco *et al.*, 2000, Kamp *et al.*, 2000, Sun *et al.*, 2000], and reciprocal asymptomatic duplications [Bosch and Jobling, 2003]. Formation of recombination intermediates in this region is also associated with gene conversion [Blanco *et al.*, 2000, Bosch *et al.*, 2004]. Indeed, NAHR activity among LTRs in the human genome is most often observed as gene conversions on the Y chromosome [Trombetta *et al.*, 2016], although human endogenous retroviral elements (HERVs) could potentially generate unequal crossovers in primates on other chromosomes [Hughes and Coffin, 2005]. LTRs prone to NAHR can cause sequence exchange either intra- or inter-chromosomally [Trombetta *et al.*, 2016].

As an ERV-mediated event, the formation of ePAR might have recurred sporadically through human history [Mensah *et al.*, 2014]. Given the publication of Mensah *et al.*'s work describing the ePAR in Belgians in 2014, two putative ePAR-carrying English semen donors, one who had previously been discovered to carry a duplication of several kb of X-chromosome sequence extending into PAR1 by Dr. Celia May at the University of Leicester, and another who was identified by virtue of his predicted Y-Hg being similar to some of the ePAR carriers in Mensah *et al.* (2014), should be worth investigating to determine the precise nature of their chromosome rearrangement. If this proves to be the ePAR, then detailed work on the recombination behaviour in this region will be carried out

3.1.1 LTR6Bs and NAHR

LTR6Bs are retrotransposons, an example of transposable elements (TEs), which are originated from HERVs. This specific LTR type evolved from HERV-S71, which is a part

of the superfamily ERV1 [Kojima, 2018, Tristem, 2000]. Generally, the basic HERV structure consists of three proviral genes, namely *GAG* (group-specific antigen), *POL* (polymerase) and *ENV* (envelope) flanked by 5´- and 3´-LTRs [Nelson *et al.*, 2003] and spans approximately 6.2-8.7 kb in total [Tristem, 2000]. There are other forms of sequence organization, which may contain other elements and genes such as open reading frame (ORF) and protease genes (*PRO*) which can extend the whole length to ~7-11 kb [Jern and Coffin, 2008] (Figure 3.1a.). However, either homologous recombination or NAHR between LTRs might obliterate any of the proviral genes, leaving only a single LTR which is called a solitary or 'solo' LTR [Hughes and Coffin, 2004, Jern and Coffin, 2008] (Figure 3.1b.).

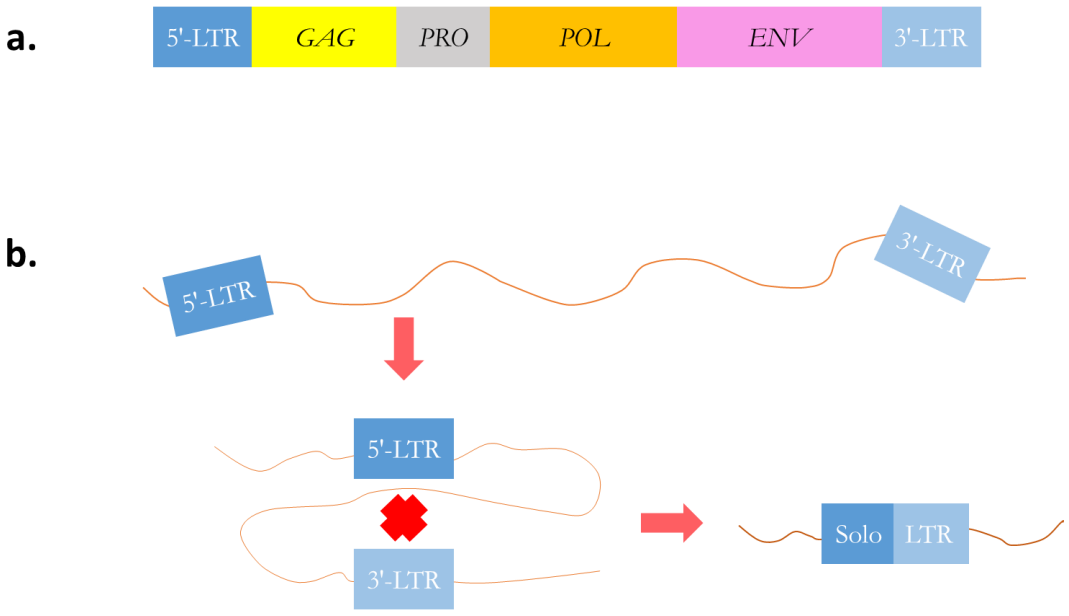


Figure 3.1 Organization of HERV genome and generation of solo LTR. a. Structure showing the order of proviral genes and LTRs flanking at both ends (adapted from [Jern and Coffin, 2008]). **b.** Schematic diagram showing one of the NAHRs between 5´- and 3´-LTRs in *cis* generating a solo LTR.

Solo LTRs have been estimated to be 10 to 100 times more numerous than full-length HERV sequences, and sometimes act as gene regulators [Jern and Coffin, 2008]. As well as this they are involved in evolutionary dynamics mainly via recombination processes, contributing to the plasticity of the genome by shuffling of genomic and gene regulatory contents [Jern and Coffin, 2008]. In ePAR, although the LTR6Bs which mediate ectopic crossover flank a ~110-kb stretch of genomic DNA from the distal part of the X-specific

region across PAB1 to the proximal part of PAR1, the structure between the two LTRs does not consist of proviral elements. Therefore LTR6B-X and LTR6B-PAR1 are solo LTRs. Both LTR6Bs share very similar sequences which are consistent with the 558-bp prototype of HERV-S71 according to Repbase, the public archive of repetitive DNA sequences [Kojima, 2018, Repbase]. In the database of HERV/LTR regulatory elements (dbHERV-REs) [dbHERV-REs, Ito *et al.*, 2017], there are at least 449 LTR6Bs, 79-590 bp in length, which were identified as transcription factor binding sites (TFBSs) distributed across the whole human genome and on every chromosome except Chromosome 18. In Chapter 7, LTR6Bs which have the greatest sequence and size similarity to the LTR6Bs that mediate ePAR formation will be systematically analysed.

3.1.2 Extension of PAB1 and duplication of distal X-specific region leading to discovery of putative ePAR-carrying semen donors

Prior to 2014, one semen donor from the University of Leicester's collection was identified by Dr. Celia A. May and team to carry duplicated X-specific markers, via allele-specific PCR and SNP typing. It was found that there was heterozygosity in X-specific SNPs in a region spanning 5 kb from the most proximal part of PAR1 across PABX, together with a predicted diploid copy number determined by a qPCR assay some 12-kb proximal to PABX (Figure 3.2 – the semen donor was called Man 20). These experiments suggested that there may be a duplication of at least 17 kb of sequence, stretching from the most proximal PAR1 across PABX through X-specific portion, somewhere in the genome. After the article by Mensah *et al.* was published [Mensah *et al.*, 2014], the specimen seemed worthy of thorough study, not only to ask if it does carry the ePAR, but if so, also to study its recombination behaviour. In addition, to search for more putative ePAR semen donors based on Y-Hg I-P37.2* as the most prominent ePAR-carrying Y-Hg known so far [Mensah *et al.*, 2014], the profiles of Y-chromosome short-tandem repeats (Y-STRs) from the semen donor collection genotyped at the University of Leicester were used to predict their Y-Hgs using Dr. Jon H. Wetton's (University of Leicester) nearest-neighbour-based prediction program (Jon Wetton, unpublished). Y-Hgs and the methodology of prediction from Y-STR haplotypes will be described fully in Chapter 7. After obtaining Y-Hg predictions, it was found that another semen donor, namely Man 53, also fell in the I-P37.2* Y-Hg. As this Y-Hg was the major ePAR-carrying lineage described in Mensah *et al.* (2014), Man 53 was expected to potentially carry the ePAR. Confirmation of ePAR status in both semen donors is described in the Results part of this Chapter.

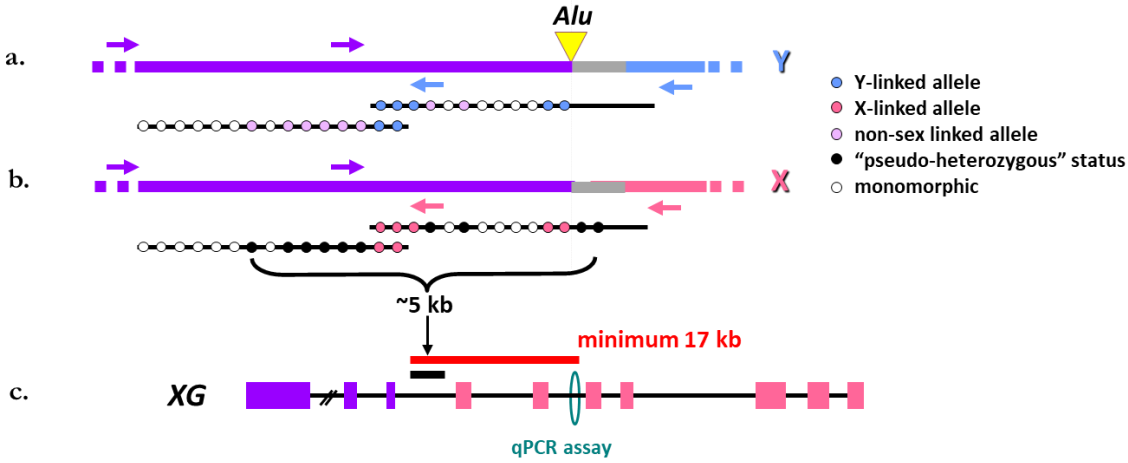


Figure 3.2 Schematic diagram of SNP genotyping and qPCR-assay detecting duplication of X-chromosomal material in the Leicester semen donor, Man 20. In **a** and **b**, Man 20 was detected to have a duplication around PABX by SNP typing (coloured dots on the black lines refer to each SNP). Allele-specific primers are indicated by coloured arrows; the forward primers in the proximal PAR1 are shown in purple while the reverse ones are in blue for Y-specific and pink for X-specific regions. The boundary, PAB1, on the Y chromosome is currently demarcated by an inserted *Alu* element (yellow triangle); immediately distal material was previously pseudoautosomal, and shows 77% similarity between the X and Y (grey block). Diagram **a** shows the 5-kb PCR amplicon using primers in the most proximal PAR1 through the Y-specific region, where Y-linked SNPs identified by allele-specific typing were haploid as expected; In diagram **b,** on the other hand, X-linked SNPs, contained within a PCR amplicon amplified by a similar methodology but across PABX into the X-specific region, gave pseudo-heterozygous signals indicative of this region being duplicated in the genome. Diagram **c** shows the position of the duplicated regions in relation to the exons of the *Xg* gene extending from the most proximal PAR1 across PABX to the most distal X-specific part. qPCR was performed using an X-specific amplicon 12 kb proximal to PABX (chrX:2711371-2711465, hg 19) and a published assay for the estrogen receptor located on chromosome 6 (chr6:152265487-152265570, hg19) [Mhlanga & Malmberg, 2001] using qPCR MasterMix Plus for SYBR® Green I without UNG (Eurogentec) as per the manufacturer's instructions. The resulting normalised data from females could be distinguished from those derived from males ($p<0.05$, Student's t test) yet Man 20 clustered with the former and not the latter.

By combining both results, it could be inferred that Man 20 must carry, somewhere in his genome, at least a 17-kb duplication from the most proximal PAR1 across PABX through the X-specific region. However, the precise size and location of the duplication was not established at that time (Adapted by permission of Dr. Celia A. May, unpublished)

3.2 Results

3.2.1 Sequence similarity of ePAR-mediating LTR6Bs

First, the human reference sequences (hg19) for both LTR6Bs (Figure 3.3 a.) forming the ePAR were taken from the UCSC Genome Browser [UCSC Genome Browser], then aligned. LTR6B-X and LTR6B-PAR1 shared ~97% similarity, with 15 distinguishing SNVs and two indels based on the reference sequence as shown in Figure 3.3 b; the distinguishing SNVs are listed in the Appendix (some SNVs are not in the dbSNP database).

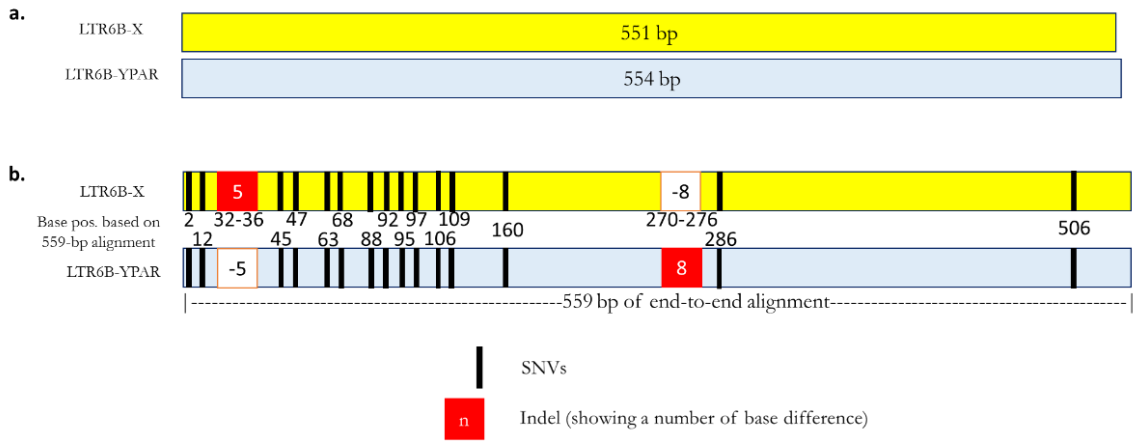


Figure 3.3 Alignment of LTR6B-X and LTR6B-YPAR showing distinguishing variants. Diagram **a.** shows simple models of both LTR6Bs (here, using YPAR to represent the real ePAR-mediating LTR6B which is on Y chromosome) which were 3 bp different in length. Diagram **b.** shows their end-to-end alignment of human reference sequences showing distinct SNVs (black lines) and indels (red and white rectangles representing gain (+) and loss (-) of bases indicated in number).

3.2.2 Initial confirmation of ePAR in semen donors by long-range PCR and study of junction diversity by nested PCR

Two Leicester sperm-donors, Man 20 and Man 53, were suspected of carrying the ePAR described in Mensah *et al.* (2014). However, only one (Man 20) had already been confirmed by virtue of SNP haplotyping and qPCR (Figure 3.2) to have pseudo-heterozygosity of the most distal part of the X-specific region by obtaining a continuous haplotype near PABX. This showed that he must carry a duplication of a portion of the X chromosome somewhere in his genome. Nevertheless, the actual size of this duplicated piece and pattern of generation had not yet been defined - it was only known to be at least 17 kb in length (Figure 3.2). Because the majority of ePAR cases reported in Mensah *et al.* (2014) clustered in Y-Hg I-P37.2*, Y-STR profiles and predicted haplogroups of all the men in the Leicester semen donor collection (n=174) were surveyed for potential ePAR-carrier status. Y-STR profiling and Y-Hg prediction was performed by the in-house program of Dr. Jon H. Wetton, which is able to predict to a finer sub-branch level than that used in Mensah *et al.* (2014) [Athey, 2006]. Following this survey, Man 53 was the only additional case, together with Man 20 (Table 3.1), to be predicted to carry a Y chromosome belonging to Y-Hg I-L233 – a branch of I-P37.2* (Figure 3.4), and therefore to potentially carry the ePAR according to Mensah *et al.* (2014).

Table 3.1 PowerPlex® Y23 STR profiles of semen donors and Y-Hg prediction by two approaches.

Locus / Sample	DYS19	DYS389I	DYS389II	DYS390	DYS391	DYS392	DYS393	DYS385a	DYS385b	DYS437	DYS438	DYS439	DYS448	DYS456	DYS458	DYS635	YGATAH4	DYS481	DYS533	DYS549	DYS570	DYS576	DYS643	Prediction method*	
																								A	J
d20	15	14	31	22	10	11	13	12	15	14	10	11	18	14	18	21	11	28	13	12	17	17	12	I2a	I-L233
d53	16	14	30	23	10	11	13	12	15	14	10	11	18	14	17	21	11	27	13	12	19	17	12	I2a	I-L233

*Prediction method A = Athey's method [Athey, 2006, Athey Haplogroup Predictor] and J = Jon Wetton's method (unpublished)

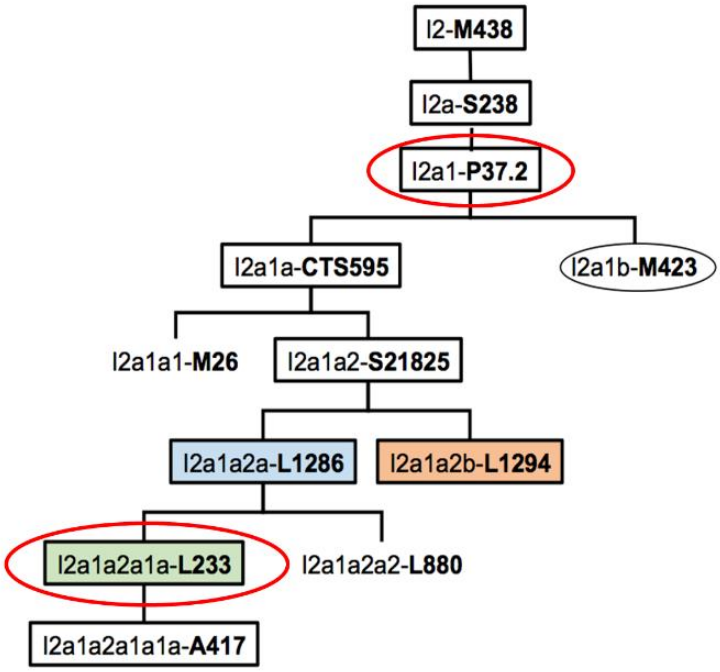


Figure 3.4 Phylogenetic tree of the Y-Hg I2a-P37.2*. Names of Y-Hg and sub-Hg are usually written as a ≤3-character alphanumeric designation followed by '-' and the name of the SNP defining that specific subgroup, e.g. I-L233 or I2a-L233. This diagram shows full names of sub-Hgs according to ISOGG nomenclature [ISOGG] in Hg I2a1, a sub-lineage of I2a. Red ellipses highlight Y-Hg I2a-L233 which is derived from the lineage I2a-P37.2*.

To confirm both semen donors carried the ePAR, primers were selected from Dr. Celia A. May's existing collection (see Appendix). These primers should amplify across the translocation boundary running from X-specific to Y-specific portions. In addition, if positive, the large size of the product (12 kb) may be useful to demonstrate the overall DNA condition of the samples for future recombination work. The primers cover the amplicon over the translocation junction as shown in Figure 3.5. After optimising T_m of the primers, both semen donors showed a 12-kb product in PCR, indicated that they indeed likely carried the ePAR (Figure 3.5).

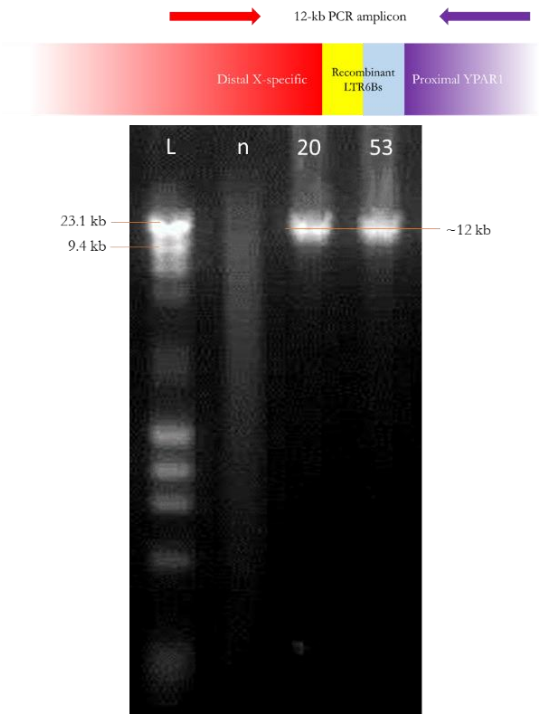


Figure 3.5. Long-range PCR confirming ePAR in two semen donors. The diagram above the gel picture shows the primers (red and purple arrows) amplifying from X- to Y-specific portions across the recombinant junction. In the gel picture, Men 20 and 53 produced the predicted 12-kb amplicons indicative of ePAR. By contrast, a donor not predicted to carry this rearrangement did not amplify (negative lane: n). L = λ (HindIII digest) and ΦX174 (HaeIII digest) DNA markers.

After successful amplification of the 12-kb amplicon, the PCR products were subsequently reamplified to focus only on the translocation junction with another pair of primers (see Appendix) to give a 3.8-kb amplicon and to save as much genomic DNA as possible. The nested-PCR amplicons from the previous step were Sanger-sequenced using a number of primers. Even though this sequence is not very long, there are abundant homopolymeric A- or T-tracts scattered along its length, obstructing the sequencing process by generating stuttering. Therefore, it needed more sequencing primers than usual to achieve the whole sequence.

The 3.8-kb sequence in both samples spanned from the proximal end of the X-specific part through the LTR6B segment and through to the distal end of PAR1 on Y in the extended region, as in Figure 3.5. Not surprisingly, the sequences in the X-specific segment and in PAR1 on the Y were found to be similar to those in the reference sequence. However, the issue to be considered was in the LTR6B segment, which was not expected to be exactly the entire sequence from either LTR6B-X or the copy on YPAR1 (LTR6B-YPAR1), but rather the hybrid junction-product of translocation by means of NAHR between both LTR6Bs. The analysed sequences in both semen donors were completely identical, and were consistent with "Junction 1" in Mensah *et al.* (2014); details are as shown in Figure 3.6.

Figure 3.6 Alignment of Sanger-sequencing data at the LTR6B-recombinant junction. Diagram **a** shows an alignment of Junction 1 (or Jn 1 in **b**) against the reference sequences of LTR6B-X and LTR6B–YPAR1. Green highlights show distinguishing bases between both LTR6Bs. The total of 559 bp of sequence is based on end-to-end alignment whereby the first position is taken as the coordinate chrX (hg19):2,808,547. The 559-bp sequence of the junction LTR6B can be divided into three parts regarding similarity to each originating LTR6B by alignment. The 5' 160-bp portion highlighted in yellow is identical to LTR6B-X (chrX:2,808,547-2,808,706). The mid-portion highlighted in red is identical to

the common sequence in the mid-part of both LTR6Bs (X:2,808,707-2,808,815 equivalent to Y:2,644,304-2,644,412 ; 109 bp). The 3'-most 290 bp segment highlighted in blue is highly similar to LTR6B-YPAR1 (chrY(hg19):2,644,413-2,644,702). Position 451 marked with the red rectangle shows a SNV which is not found in either LTR6B-X or LTR6B-YPAR1 while Position 506, a SNP marked with a purple rectangle, shows the same allele as LTR6B-X. Diagram **b**, corresponding to Junction-1 sequence in **a,** shows a simplified structure of the recombinant junction. The coloured circles indicate SNVs where alleles are different from the human reference sequence as described in **a**.

3.2.3 Development of *de novo* ePAR 'Junction PCR' assay

In this study, it was hoped to identify more ePAR-carrying samples to study the relation between phylogenetic diversity and the rearrangement event(s), and to explore the diversity of the junctional sequence. Therefore an efficient and easy-to-use assay needed to be designed and established, because the initial PCR-based method which was used to confirm the ePAR in both semen donors produced too long an amplicon, and indeed may not work in the case of degradation of DNA in any sample.

In designing the assay, two principles were taken into consideration. First, it should amplify a small targeted amplicon to avoid false negative results in the presence of degraded DNA. Second, it should co-amplify a segment which is longer in size, to act as an internal positive control for successful PCR. With these concepts, the primers were designed for two amplicons in a duplex PCR assay as described in Appendix; the "test" primers cover an 847-bp amplicon running from X-specific to Y-PAR1, and the "control" primers cover a 1551-bp amplicon on the Y-specific portion. Although the hybrid junction LTR6B spans only 551-559 bp, the presence of a long repetitive sequence and abundant SNPs required the design of a longer amplicon in order to provide stable primer-binding sites. After optimizing the annealing temperature, the result of the PCR-based assay is shown as in Figure 3.7. The finalized PCR conditions are presented in the Appendix.

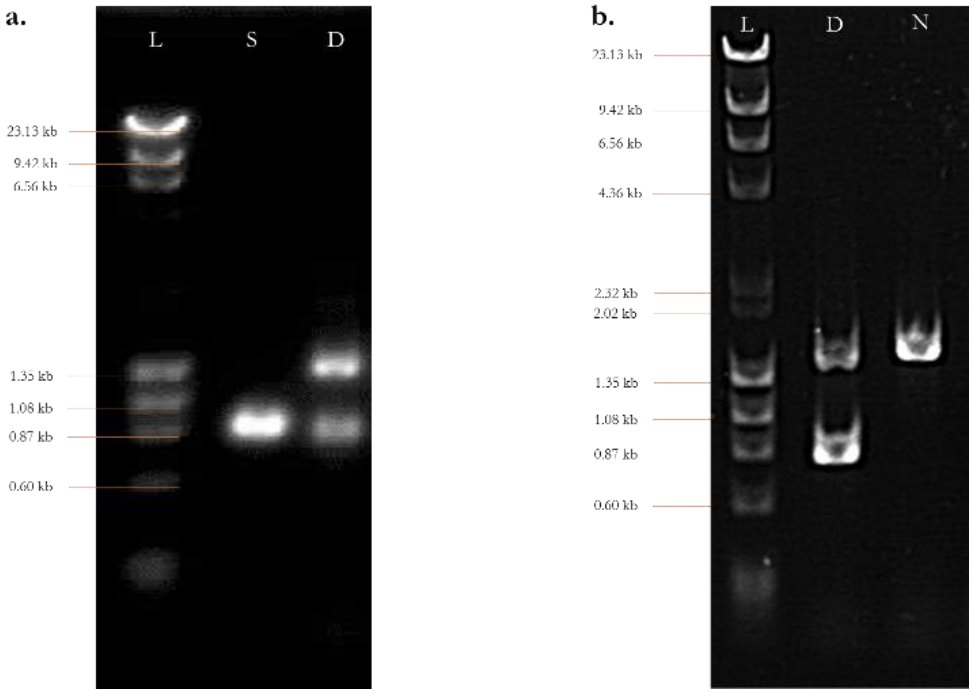


Figure 3.7 A duplex-PCR assay for the extended PAR1 detection. L = λ (HindIII digest) and ΦX174 (HaeIII digest) DNA markers, S = singleplex PCR for testing primers only, D = positive sample in duplex PCR, and N = non-ePAR male sample amplified with the duplex PCR assay. Picture **a**. shows the PCR results from the same sample selected from the ePAR positive cases of Mensah *et al.* (2014) tested with the singleplex PCR of the test primers (S) and the duplex PCR assay (D). Picture **b.** shows the results of duplex PCR assay of the same sample in **a** (D) and non-ePAR (N). The 847-bp test amplicon will show only in an ePAR case both in singleplex (S) and duplex PCR (D), while the 1551-bp control must appear in all male DNAs, unless severe degradation occurs.

To validate its accuracy and robustness, this "Junction PCR" assay was performed for 24 Belgian samples published in Mensah *et al.* (2014) and supplied under a collaboration from Dr. Maarten H. D. Larmuseau, one of the authors of the 2014 paper, and with permission from the principal investigator. The result confirms that the assay performs as expected.

3.3 Discussion

To confirm that two semen donors in our collection carried ePAR, ~12-kb long-range PCR was performed as described and followed by nested PCR (~3.8 kb) to narrow the range to cover just the recombinant junction and some flanking sequence. To do this is time-consuming (>8 hrs. in the long-range PCR) and it is likely to fail if DNA quality is poor or degraded. A more useful PCR assay was developed to overcome these issues by shortening the amplicon length and including a positive control amplicon, and thus a duplex 'Junction' PCR assay was available. Tests with positive control samples from Mensah *et al.* (2014) showed good results. This assay not only makes the test more convenient and quicker, but is also useful as a prelude to Sanger sequencing for the study of junction diversity (see Chapter 7). Here, the sequencing results were compared between the junctions using primers from nested PCR and those from the Junction assay, and showed the same sequence.

The junctional sequences of both semen donors, Man 20 and 53, were consistent with the so-called 'Junction 1' in Mensah *et al.* (2014) which was assumed here to have a recombinant breakpoint within a 109-bp segment in the mid-part of both LTR6Bs (X:2,808,707-2,808,815 equivalent to Y:2,644,304-2,644,412) as these portions are identical. Remarkably, the proximal or 3´-most part of the junction is highly similar to LTR6B-YPAR1 as its 5´ end begins with the 8-bp indel as a signature of that LTR6B despite of two different SNVs at the positions 451 and 506 as shown in Figure 3.6 b. It can be possibly explained, in a more parsimonious way, that position 451, rs2534626 (build 150 dbSNP), and position 506, rs311162, of the ePAR samples, both Belgian and English, originally had different alleles to the human reference sequence, rather than that position 451 has a different original allele while position 506's identity to LTR6B-X is caused by a gene conversion as a subsequent and separate event after ePAR generation.

One DSB hotspot was detected in PRDM9-*AA* males covering LTR6B-X (X:2,806,301-2,809,274) but none was found on LTR6B-PAR1 [Pratto *et al.*, 2014]. Despite no detected DSB hotspot in one recombinant site, NAHR did occur resulting in ePAR. In addition, this recombinant junction was found to be prominent in one particular Y-Hg, I-P37.2* [Mensah *et al.*, 2014], which was consistent with both semen donors. It is likely that they are related via a common patrilineal ancestor, and carry ePAR identical-by-descent rather than identical-by-state.

3.4 Conclusion

The ePAR mediated by NAHR between two highly similar LTR6Bs, one located on YPAR1 and the other in the distal X-specific region, has been found in some particular Y-Hgs, I2a-P37.2* and R1b-P312*. The more predominant Hg is I2a-P37.2*, which seems to carry the ePAR event as result of a single event in its evolutionary history. Two unrelated English semen donors, though living far from those in Belgium and being apparently unrelated to any of them, carry Y chromosomes from this same ePAR dominant Y-Hg and were here confirmed to also carry the rearrangement. Moreover, they also showed the same recombinant junction type as those Belgian ePAR individuals. To facilitate confirmation of ePAR status and to search for additional potential ePAR cases, a robust and convenient PCR assay has been developed and validated with control ePAR samples.

Chapter 4 Identification of Potential Recombination Hotspots within ePAR

4.1 Introduction

Human recombination hotspots have long been identified via LD analysis in populations [Kauppi *et al.*, 2004, The International HapMap Consortium, 2007, Myers *et al.*, 2005, 1000 Genomes Project Consortium, 2010], where breakdown between two stretches of linkage of markers suggests a concentration of historic recombination events over time. In the present decade, emerging molecular genetic technologies, e.g. genome-wide SNP typing and high throughput sequencing have rapidly broadened our understanding of recombination including the mechanisms underlying DSB formation [Altemose *et al.*, 2017, Davies *et al.*, 2016, Pratto *et al.*, 2014, Brunschwig *et al.*, 2012]. DSB mapping across the genome has been attempted via the detection of meiosis-specific protein-DNA complexes followed by massive-parallel sequencing either in humans [Pratto *et al.*, 2014] or mice [Lange *et al.*, 2016]. DSBs are crucial initiating factors in meiotic recombination though there has never been used of the first human DSB maps [Pratto *et al.*, 2014] alongside LD maps to approach to identifying recombination hotspots since its launch. However, the DSB hotspot mapping, though showing much stronger strength (>3×) than the highest crossover rate in autosomes from the sperm study, has never validated in sex chromosomes by such method therefore this study may present a novel approach in identifying and interesting results of human recombination hotspots in the sex chromosomes. This chapter interrogates how well DSBs coincide with regions of LD breakdown in ePAR, and asks if this approach is able to supplement the conventional approach based on LD analysis. The aim is to identify target regions for direct analysis of recombination in sperm DNA.

4.1.1 LD breakdown and sperm crossover study

Sperm crossover studies have long been established as a powerful high-resolution experimental approach to characterizing any crossover hotspot from a pool of germ cells. Since 1998 when the first report of characterization of the hotspot in the human minisatellite MS32 was published [Jeffreys *et al.*, 1998], several publications on recombination hotspots demonstrated by the assay have been produced [Berg *et al.*, 2010, Holloway *et al.*, 2006, Kauppi *et al.*, 2005, May *et al.*, 2002, Berg *et al.*, 2011, Jeffreys *et al.*, 2013, Jeffreys *et al.*, 2004, Jeffreys *et al.*, 2001, Jeffreys and May, 2004, Jeffreys and Neumann, 2002, Jeffreys and

Neumann, 2005, Jeffreys and Neumann, 2009, Odenthal-Hesse *et al.*, 2014, Sarbajna *et al.*, 2012, Webb *et al.*, 2008]. Most of these studies identified the areas in which hotspots might lie using LD analysis. Regions of LD breakdown are targeted as candidate hotpots, and subsequently characterized by a sperm crossover assay to detect rare recombinant molecules in pools of sperm DNA molecules. However, regions of LD breakdown do not always correspond to the currently active recombination hotspots [Kauppi *et al.*, 2005]. This likely due to several reasons, for instance, such LD breakdown may be the consequence of rarer PRDM9 alleles that have not been tested in sperm assays [Altemose *et al.*, 2017, Berg *et al.*, 2010, Pratto *et al.*, 2014] or may correspond to historic hotspots that through meiotic drive have become inactive [Jeffreys and Neumann, 2009]. So an LD analysis only indicates the location of potential crossover activity, and this requires a sperm crossover study to explore its current activity.

To identify the area of a sperm crossover, LD analysis will be performed by selecting a number of genetic markers, mostly SNPs or indels, covering the region of interest. Useful markers should have above-threshold heterozygosity, i.e. generally MAF ≥ 0.05, which are so-called 'informative markers' or 'tagged SNPs'.

4.1.2 Statistics to assess LD and its significance: $|D'|$ and the LOD score

A useful statistic used to infer linkage status between two pair of markers is $|D'|$. $|D'|$ was first introduced by Lewontin in 1964 [Lewontin, 1964] as a normalization value of D which is a linkage disequilibrium determinant. D, in the simplest way, is a difference of frequencies between observed and expected haplotypes of two markers and is calculated from allele frequencies as shown below. Among other statistical measures of LD such as r^2, $|D'|$ is one of the most popular. Mathematical calculation of $|D'|$ can be seen elsewhere, e.g. [Hedrick, 1987, Kauppi *et al.*, 2004]; however, in this chapter it is presented as a brief principle to engage the reader's understanding. Locus A and B are located on the same chromosome arm and each contains i and j alleles for locus A and B, respectively. Let p and q = frequency of any allele at A and B, respectively, so p_i and q_j = frequency of ith and jth alleles at A and B. If these two loci freely recombine, or in other words are in linkage equilibrium, then the expected probability to find haplotype A_iB_j equals p_iq_j.

Let D = coefficient of LD or disequilibrium, with a value either ≥ or < 0.

As $p_1 + p_2 = 1$ and so do q_1 and q_2, then we have:

$$x_{11} = p_1q_1 + D$$

$$x_{12} = p_1q_2 - D$$

$$x_{22} = p_2q_2 + D$$

$$x_{21} = p_2q_1 - D$$

where x_{11} = observed frequencies of haplotype A_1B_1 and p_1q_1 = expected frequencies of haplotype A_1B_1, and so on. D in any of these notations is the same and based on the 1st notation can be rearranged into

$$D = x_{11} - p_1q_1 \tag{4.1}$$

The alternative approach is to calculate D from all possible haplotypes as:

$$D = (x_{11} \cdot x_{22}) - (x_{12} \cdot x_{21}) \tag{4.2}$$

Normally, D has a maximum value as 0.25 because the maximum allele frequency equals 0.5 and D can be either + or – depending on which value is higher between expected and observed haplotype frequencies. Therefore, in summary, we can write that $-0.25 \leq D \leq 0.25$ and may be affected by a large difference between allele frequencies. To reduce this effect, it is suggested to be normalised by dividing by the maximum possible value of D [Lewontin, 1964, Slatkin, 2008] to obtain D' as:

$$D' = \frac{D}{D_{max}} \tag{4.3}$$

where $D_{max} = \min[p_1q_1, (1-p_1)(1-q_1)]$ when $D < 0$ or

$$D_{max} = \min[p_1(1-q_1), (1-p_1)q_1] \quad \text{when } D \geq 0$$

D' ranges from -1 to 1, thus $0 \leq |D'| \leq 1$. When $|D'| = 0$, it means there is no linkage at all while $|D'| = 1$ refers to an absolute LD.

Although $|D'|$ indicates the linkage status between two markers, another statistic may help to evaluate the relationship. The likelihood ratio (LR) of the recombination fraction, written $LR(\theta)$, is a comparison of a probability of the recombination fraction (θ), i.e. the frequency of recombination, between the pair of markers of interest over that of linkage equilibrium (LE). Significant LD is usually justified by high $|D'|$ and high LR [Jeffreys *et al.*, 2001].

θ can be obtained via a breeding experiment in organisms or observed through a pedigree. Figure 4.1 demonstrates how we can obtain θ from a crossing experiment.

Let locus A and B contain allele [A, a] and [B, b], respectively.

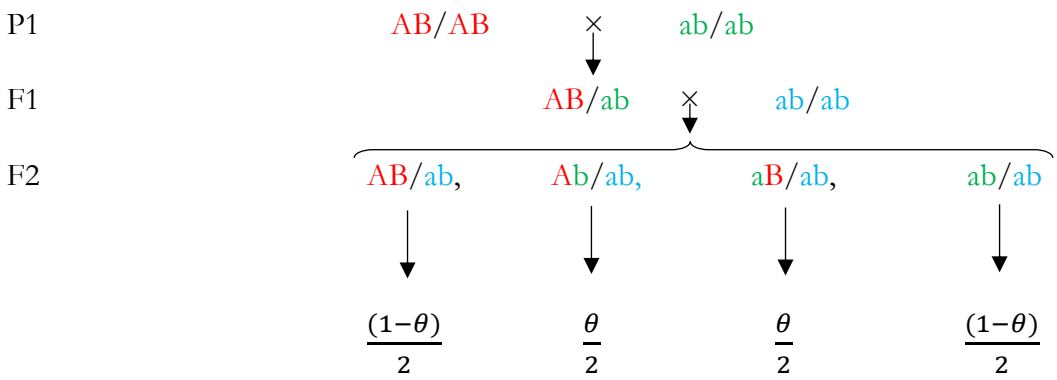


Figure 4.1 Crossing experiment in an organism to detect θ. P1 is a parental generation with homozygous AA and BB of one parent crossing with homozygous aa and bb of the other. In the F1 generation, a back-cross enables us to see four possible genotypes of gametes in F2. The term $(1-\theta)$ represents a non-recombination frequency (adapted from [Van Ooijen and Jansen, 2013]).

From Figure 4.1, the recombinant gametes, i.e. Ab and aB, have an equal probability so the recombination frequency for each haplotype is $\frac{\theta}{2}$ and in the same way the non-recombinant frequency equals $\frac{(1-\theta)}{2}$. If each event occurs equally probably, then for LE the maximum likelihood of $\theta = \frac{1}{2}$. In contrast, $\theta = 0$ for absolute LD. In other words, θ can range from 0 to 0.5 and this principle can be applied in general.

The maximum likelihood estimate (MLE) of θ, or $L(\theta)$, which follows a multinomial probability, can be shown as:

$$L(\theta) = \theta^r(1-\theta)^{n-r} \tag{4.4}$$

where n = total number of individuals and r = total number of observed recombinant individuals.

$LR(\theta)$ is a ratio between the real $L(\theta)$ against $L(\theta)$ under LE ($\theta = \frac{1}{2}$). It presents the magnitude between the probability of LD as a tested hypothesis against the probability of LE as a null hypothesis. The equation is shown as:

$$LR(\theta) = \frac{L(\theta)}{L(\frac{1}{2})}$$

$$= \frac{\theta^r(1-\theta)^{n-r}}{(\frac{1}{2})^r + (1-\frac{1}{2})^{n-r}}$$

hence $$LR(\theta) = 2^n\theta^r(1-\theta)^{n-r} \quad (4.5)$$

In 1955 Morton introduced a statistic called logarithm of odds, or LOD score [Morton, 1955]. LOD is defined as the base-10 logarithm of LR or a log likelihood-ratio statistic of the maximum likelihood of θ, so:

$$LOD(\theta) = log_{10}(LR) \quad (4.6)$$

However, $LOD(\theta)$ is preferred to be written in terms of a natural logarithm, sometimes called deviance (d), that is differentiated more conveniently. The d is also approximately equal to a chi-square distribution (χ^2) with 1 degree of freedom [Van Ooijen and Jansen, 2013, Ziegler and König, 2010]. The equation can be shown as:

$$d = 2lnLR(\theta) \approx \chi_1^2 \quad (4.7)$$

From (4.6) and the rule of logarithms that $log_b a = \frac{log_c a}{log_c b}$, $LOD(\theta)$ can be rewritten as:

$$LOD(\theta) = \frac{ln\, LR(\theta)}{ln\, 10}$$

And multiplied by $\frac{2}{2}$, then $$LOD(\theta) = \frac{2\, ln\, LR\, (\theta)}{2\, ln\, 10} = \frac{d}{2\, ln\, 10} \quad (4.8)$$

LOD is used interchangeably with LR. A rule of thumb for a significant cut-off value is LOD $\geq$ 3 or $LR \geq 1000$ which suggests that $L(\theta)$ is equal to or greater than 1000 times the probability of LE. LOD ensures LD when $|D'|$ tends towards 1 and $LOD \geq 3$.

4.1.3 Exploiting DSB data as a novel approach for targeting sperm crossover analyses

As noted above, LD breakdown has been extensively used as a means to identify target regions for sperm crossover assays. However, as introduced above, not all regions of LD breakdown correspond to active sperm recombination hotspots and if this is the case, it may lead to failure in performing a sperm crossover assay. Since meiotic recombination depends on SPO11-induced DSBs, an alternative approach to using LD would be to target regions known to support a high frequency of DSBs. Using the male meiotic DSB maps of Pratto *et al.* [Pratto *et al.*, 2014] to identify candidate hotspots may be a better strategy than conventional use of LD alone.

Since meiotic DSBs are dominantly determined by PRDM9 allele-specific motifs, this study also considered the positions of PRDM9*AA*-motifs within ePAR as both semen donors recruited in this study possess AA genotype. Finally, because designing a sperm crossover assay requires a number of informative, i.e. heterozygous, SNPs spread over the region, the density of these was also taken into account.

4.2 Results

4.2.1 Mapping of PRDM9*AA*-specific motifs and SNP distribution in the ePAR and selection of candidate SNPs for LD analysis

Since both ePAR-carrying semen donors in this study possess a PRDM9*AA* genotype, locations of Myers' motifs (CCnCCnTnnCCnC) were identified in the X-specific part of ePAR (coordinates X:2,699,521-2,808,548 [hg19]), by computational methods which were performed by Dr. John Wagstaff, a bioinformatician at the University of Leicester. The criteria for motifs of interest were the perfectly matched (8/8), or eight possible singly-mismatched (7/8) sequences at the eight specified bases of the otherwise degenerate 13 bp long sequence. The number of locations detected are shown in Table 4.1; see also in details in Appendix.

Table 4.1 Number of Myers's motifs across region X: 2,699,521-2,808,548 (hg19) of ePAR

Category	Number of locations
8/8 (-CCnCCnTnnCCnC-)	15
7/8 (-CnnCCnTnnCCnC-)	33
7/8 (-nCnCCnTnnCCnC-)	19
7/8 (-CCnCCnTnnCCnn-)	28
7/8 (-CCnCnnTnnCCnC-)	11
7/8 (-CCnCCnTnnnCnC-)	22
7/8 (-CCnnCnTnnCCnC-)	11
7/8 (-CCnCCnnnnCCnC-)	47
7/8 (-CCnCCnTnnCnnC-)	11
Total	197

To assess how SNPs clustered across this X region, all common SNPs in dbSNP build 150 were taken from UCSC Genome Browser [UCSC Genome Browser]. The number and distribution of PRDM9*AA* motifs and SNPs are discussed further in Section 4.2.3.

4.2.2 LD analysis of the distal X-specific region in European females

To make an LD analysis of the X-specific region corresponding to that of ePAR, a vcf file of all variants of the US population of European ancestry (Utah residents with northern and western European Ancestry - CEU) data from the International Genome Sample Resource (IGSR), deriving from the 1000 Genomes Project (1000 GP) [1000 GP], was created using the Data Slicer tool in the Ensembl website [Ensembl Data Slicer]. Reported SNPs across the human distal X chromosome, coordinates 2,699,521-2,808,548 (hg19), were extracted from the source database FTP file [1000 GP III FTP] in a vcf format containing 3,107 SNPs per individual in a total 2,504 individuals of both sexes from 29 populations. Raw data were then manipulated in Microsoft Excel 2013. Lists of sample identification, separated by sex and populations were obtained from 1000GP. The CEU data set which comprises 98 unrelated individuals, 49 males and 49 females, was chosen as the X chromosomes of the two ePAR semen donors are also likely to be of European origin. For each SNP allele frequencies and a chi-square statistic to test for Hardy-Weinberg equilibrium (HWE) were calculated using genotypes from both sexes. The method used to treat biallelic markers on the human X chromosome was introduced by Graffelman and Weir [Graffelman and Weir, 2016]. Comparison between the classical and new mathematical approaches is in Table 4.2.

Table 4.2 Comparison between two HWE tests for a biallelic maker of X chromosome

	Classical	Graffelman & Weir (2016)
1. Sex taken into account	only female	both
2. Frequency of allele A (p_A)	$\frac{2n_{AA}+n_{AB}}{2n_f}$	$\frac{m_A+2n_{AA}+n_{AB}}{n_m+2n_f}$
3. Frequency of allele B (p_B)	$1-p_A$	$1-p_A$
4. Chi-square test of HWE	$\frac{(n_{AA}-n_f p_A^2)^2}{n_f p_A^2}+\frac{((n_{AB}-2n_f p_A(1-p_A))^2}{2n_f p_A(1-p_A)}$ $+\frac{((n_{BB}-n_f(1-p_A)^2)^2}{n_f(1-p_A)^2}$	$\frac{(m_A-n_m p_A^2)^2}{n_m p_A}+\frac{((m_B-n_m(1-p_A))^2}{n_m(1-p_A)}$ $+\frac{(n_{AA}-n_f p_A^2)^2}{n_f p_A^2}+\frac{((n_{AB}-2n_f p_A(1-p_A))^2}{2n_f p_A(1-p_A)}$ $+\frac{((n_{BB}-n_f(1-p_A)^2)^2}{n_f(1-p_A)^2}$
5. Degree of freedom (df)	1	2

n_{AA}, n_{AB}, n_{BB}=observed number of *AA*, *AB* and *BB* females; m_A, m_B=observed number of *A* and *B* males; n_m, n_f= total number of males and females

The number of SNPs filtered by different criteria are presented in Figure 4.2. As shown in the Venn diagram, 388 SNPs passed HWE when using the classical calculation method while when using the new one the number increased to 555 SNPs. The number of SNPs with MAF ≥ 0.15 is 213 while that with MAF ≥ 0.20 is 157. The joint number of SNPs both passing HWE by the new approach, and with MAF ≥ 0.15, is 132, while the number with the same criteria but MAF ≥ 0.20 is 103. On the other hand, when using the classical approach for HWE test, the number of SNPs with MAF ≥ 0.15 and with MAF ≥ 0.20 decreased to 72 and 54 respectively.

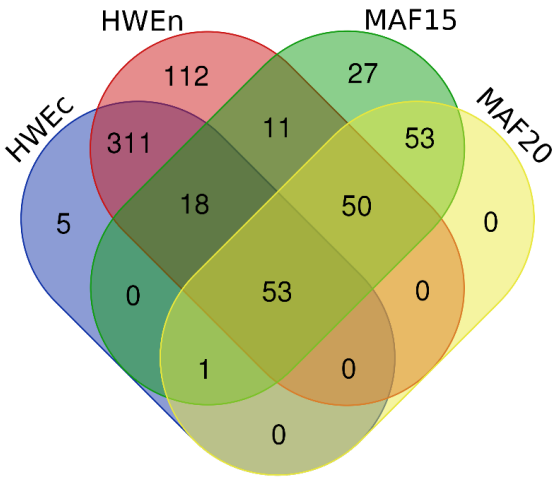


Figure 4.2 Venn diagram showing a number of SNPs filtered by different criteria in CEU population. HWEc = SNPs passing HWE by classical calculation,

HWEn = SNPs passing HWE by new calculation [Graffelman and Weir, 2016], MAF15 = SNPs with MAF ≥ 0.15, MAF20 = SNPs with MAF ≥ 0.20.

Finally, two sets of 132 and 103 SNPs which were HWE-tested by the Graffelman & Weir method and with MAF ≥ 0.15 and 0.20, respectively, were chosen for the female CEU dataset (see list of SNPs presented in the Appendix). Data were then converted into the right input format for a program analysing LD and LOD score which was written and run in True Basic® v.4.1 language software by Prof. Sir Alec Jeffreys and was used in several previous articles e.g. [Holloway *et al.*, 2006, Jeffreys *et al.*, 2004, Jeffreys *et al.*, 2001, Kauppi *et al.*, 2005, May *et al.*, 2002]. This program is based on maximum likelihood estimation of a number of unphased diploid genotypes to calculate $|D'|$ and uses the same approach as in Equation (4.7), described above, to calculate LOD scores, then plots both values in a heatmap comparing pairwise SNPs. Previously, the criteria to have SNPs input to an LD analysis have favoured SNPs which pass HWE and have MAF ≥ 0.15; however, this study tried two different values for MAF, namely ≥ 0.15 and ≥ 0.20. The results are shown in Figure 4.3.

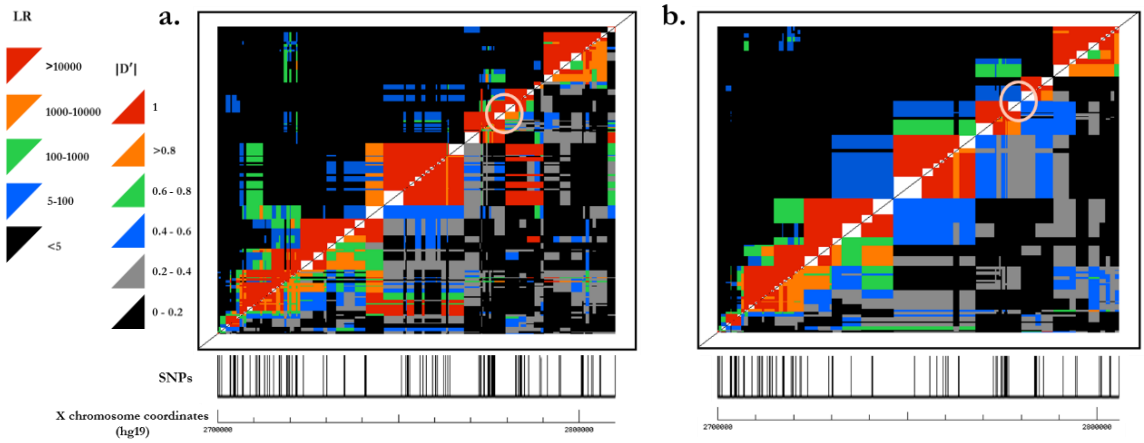


Figure 4.3 Heatmaps presenting LD and *LR* in female CEU samples. The rectangular heatmap segregates into two halves by a diagonal line, such that the upper left half shows LR and the bottom right one shows $|D'|$. **a**. Results for SNPs with MAF ≥ 0.15; **b**. Results for SNPs with MAF ≥ 0.20. The stretches of black bars beneath each heatmap are SNP positions according to hg19 coordinate scales underneath. The pink circles indicate discrepant LD breakdown positions. Strengths of $|D'|$ and LR are indicated by colours, as shown in the key to the left of the figure.

In Figure 4.3, both sets of SNPs revealed very similar LD breakdown positions except for the encircled position where the LOD score, or here presented in term of LR, seemed to be hazier for SNPs with MAF ≥ 0.15. Eventually, the dataset of SNPs with MAF ≥ 0.20 was selected, based on reasoning which will be discussed later, to extract LD breakdown positions as presented in Table 4.3.

Table 4.3 LD breakdown across chromosome X:2,699,521-2,808,548 in 1000GP CEU females

Position	SNP interval (build 150)	Coordinate (hg19)	Length (bp)	$\|D'\|$	LR
1	rs5939320 - rs1486175	2700202 - 2703391	3,189	0.4 - 0.599	100 - 1000
2	rs5939117 - rs71994079	2707060 - 2709009	1,949	0.4 - 0.599	100 - 1000
3	rs3828931 - rs6641657	2740891 - 2751904	11,013	0.4 - 0.599	5 - 100
4	rs5939356 - rs7879795	2763684 - 2772660	8,976	0.4 - 0.599	100 - 1000
5	rs582897 - rs5939137	2776686 - 2783555	6,869	0.4 - 0.599	5 - 100
6	rs6642051 - rs5939139	2786038 - 2791081	5,043	0 - 0.199	<5

The regions of LD breakdown delineate a stretch of LD blocks across this region on the X chromosome as shown in Figure 4.4.

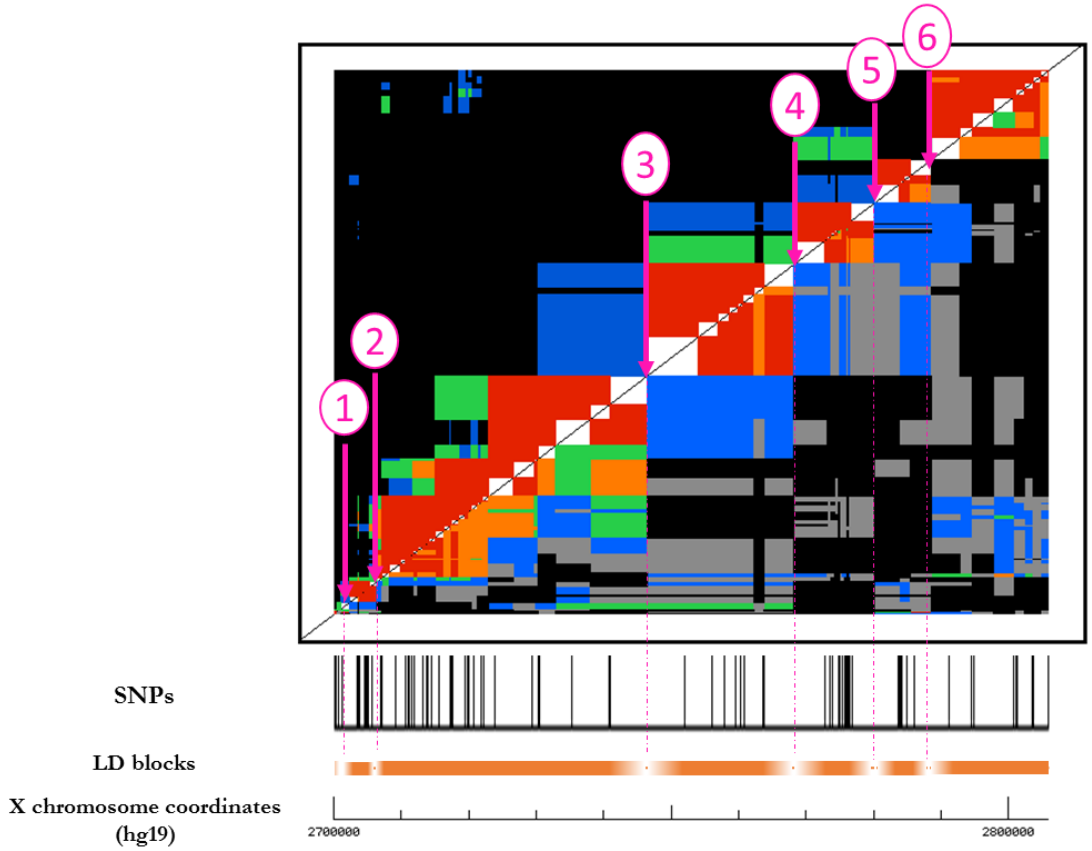


Figure 4.4 Delineating LD blocks from the heatmap. Seven LD blocks in orange are flanked by six LD breakdowns, corresponding to **Table 4.3**, shown as white gradients. The heatmap is based on the SNP dataset with MAF ≥ 0.20.

4.2.3 Identification of candidate recombination hotspots in ePAR for a sperm crossover study

As stated earlier, this study will introduce a novel approach to identify candidate recombination hotspots by using not only LD analysis, but also using ChIP-Seq meiotic DSB data from Pratto *et al.* [Pratto *et al.*, 2014]. The DSB data are derived from experiments performed using testes biopsy material from European males who have PRDM9-*AA*, *AB* and *AC* genotypes so, in this study, only PRDM9*AA* data in the corresponding region of X-specific part in ePAR were taken for analysis. Along the X-specific part of ePAR, seven DSB clusters from two PRDM9*AA* individuals were averaged, and their arbitrary DSB signal strengths, measured by single-stranded DNA (ssDNA) sequence coverage, compared as shown in Table 4.4.

Table 4.4 DSB signal strength in the X-specific region corresponding to ePAR in two PRDM9*AA* males from Pratto *et al.* (2014)

DSB hotspot	Coordinate (hg19)	Interval (bp)	Signal strength in PRDM9-*AA* males		
			Man 1	Man 2	Average
X1	2700356 - 2701802	1,446	63	92	77.5
X2	2719185 - 2720791	1,606	41	37	39
X3	2747022 - 2747827	805	11	11	11
X4	2766587 - 2767405	818	19	23	21
X5	2786474 - 2788752	2,278	319	423	371
X6	2799041 - 2800391	1,350	25	53	39
X7	2806301 - 2809274	2,973	138	176	157

In addition, selection of the most appropriate candidate hotspot regions requires a substantial number of heterozygous SNPs and/or indels in the semen donor who is recruited to the experiment; these must not only cover, but also extend outside a candidate recombination hotspot, and ideally should have evenly-distributed intervals between each other. Therefore, DNAs from Man 20 and 53 were sequenced using a massively-parallel sequencing platform (Ion Torrent PGM^{TM}) over the seven DSB hotspot regions listed in Table 4.4 above. Details of sequencing will be presented in Chapter 6. Heterozygous SNPs for each semen donor, DSB hotspot signals, density of detected Myers' motifs, LD blocks and regions of breakdown from Section 4.2.2, and common SNPs from Section 4.2. were brought together to identify candidate recombination hotspots, as summarised in Figure 4.5.

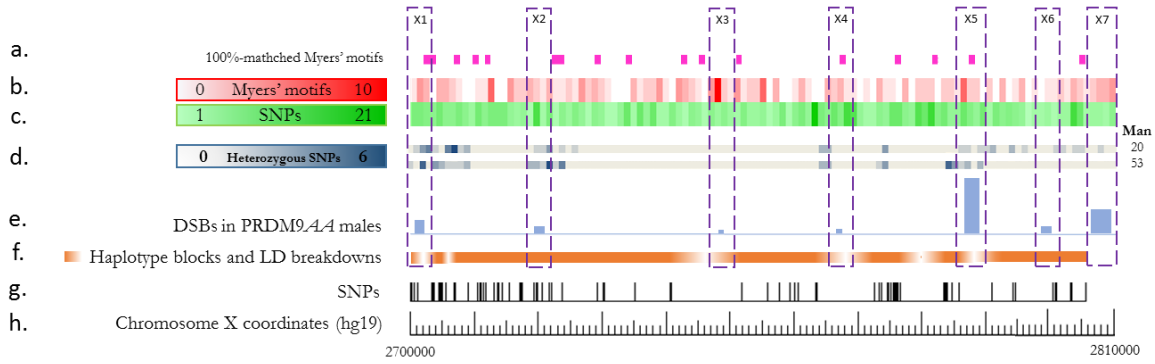


Figure 4.5 Combined diagram to identify candidate hotspots based on different criteria. Rows of heatmaps (**a. - d.**) in 1-kb windows show number and distribution of 7/8 and 8/8-matched Myers' motifs (**b.**) while **a.** indicates

positions of the 8/8-matched motifs, following by all SNPs (**c.**) then the heterozygous SNPs in Man 20 and 53 (**d.**) which exclude some regions flanking DSBs. The blue histogram in **e.** is plotted against arbitrary signal strength of DSB hotspots in PRDM9*AA* males from Pratto *et al.* (2014). The row **f.** underneath lies an orange LD blocks with six LD breakdowns in fading colour as delineated in **Table 4.3** and **Figure 4.4**. The 103-SNP ruler from 1000GP CEU female dataset (**g.**) and the human X-chromosome coordinate scale (**h.**) are placed at the bottom. Seven purple dashed rectangles labelled with X1 to X7 encircle each DSB region. X1-X7 represent the name of DSB hotspots which correspond to **Table 4.4**.

One DSB hotspot, X7, had to be excluded since it was not covered by SNP data. As shown above, two of the remaining six DSB hotspots, i.e. DSB hotspot X2 and X6, did not coincide with regions of LD breakdown. It could be possible that the data from the CEU population in 1000GP possibly contained variation arising as a consequence of PRDM9 motifs that were different from the PRDM9-*AA* individuals studied in Pratto *et al.* [Pratto *et al.*, 2014]. So, taking the same approaches to perform LD analysis as described in Section 4.2.2, and those described in this section for obtaining DSBs data from Pratto *et al.* [Pratto *et al.*, 2014], LD analysis in the African YRI (Yoruba from Ibadan, Nigeria) population from 1000GP [1000 GP], which contains predominantly PRDM9-*C* motifs was carried out. The DSB diagram from PRDM9-*AC* males alongside LD blocks and breakdowns was included into Figure 4.6, which is adapted from Figure 4.5 to broaden the view of recombination hotspots.

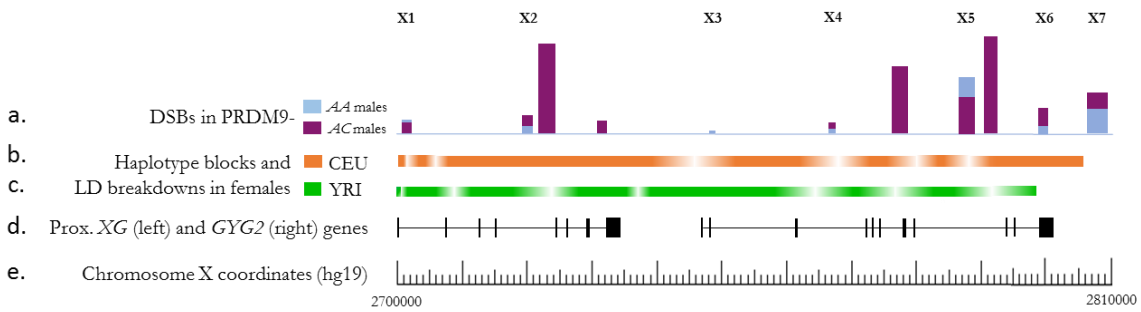


Figure 4.6 Combined DSBs in PRDM9-AA and AC males with LD blocks in CEU and YRI. DSBs were plotted in the top histogram (**a.**) representing PRDM9-A and C hotspots in pale-blue and dark-purple, respectively. The hotspot names (X1-X7) above each bar show the positions mainly of PRDM9-A

hotspots. LD blocks plotted from LD analysis heatmaps of CEU (**b.**) and YRI (**c.**) are shown as similar to Figure 4.3 and 4.4. The blocks of exons of the proximal *XG* and *GYG2* genes lie in **d.** and at the bottom (**e.**) is a chromosome coordinate scale.

After taking the YRI population into consideration, PRDM9-A DSB hotspot X2 then coincided with one LD breakdown hotspot; however, the hotspots X6 and X7 were not covered by the YRI SNP dataset which passed the HWE and MAF criteria so it was not possible to summarize their relation to LD.

Finally, the question of the correlations between Myers' motifs (MM), DSB signal strength in PRDM9*AA* males (DSB), and regions of LD breakdown (LDB) was taken into consideration. Because the LD analysis statistics, i.e. $|D'|$ and LR, relate directly to LD and inversely to LDB, they must be transformed into an inverse ratio to correspond to MM and DSB, which relate directly to LDB. Furthermore, as $|D'|$ is taken into account with LR to assure LD strength, both values should be multiplied. The product was then made an inverse ratio with 1. In this study, the final value is defined as LDB is equals to:

$$LDB = \frac{1}{(|D'| \cdot LR)} \tag{4.9}$$

MM, DSB and LDB data were tallied in 1-kb intervals along the chromosome X length between the coordinates 2,699,149-2,791,148 (hg19) which excluded the most proximal part because of the lack of suitable SNPs in the dataset. Correlation analysis was performed using two approaches, the first analysing the whole region (suffixed with '_all') and the second only the selected regions which contained DSB or LDB (suffixed with '_hotspot'). Software used for this analysis was IBM SPSS Statistics v.24.

MM, DSB and LDB under both approaches were not normally distributed (p-value ≈ 0) by a Shapiro-Wilk test for normality. Then, to select the appropriate non-parametric method between Spearman's rank-order rho (ρ) or Kendall's tau-b (τ_b), scatter plots between each pair of variables were made, as shown in Figure 4.7.

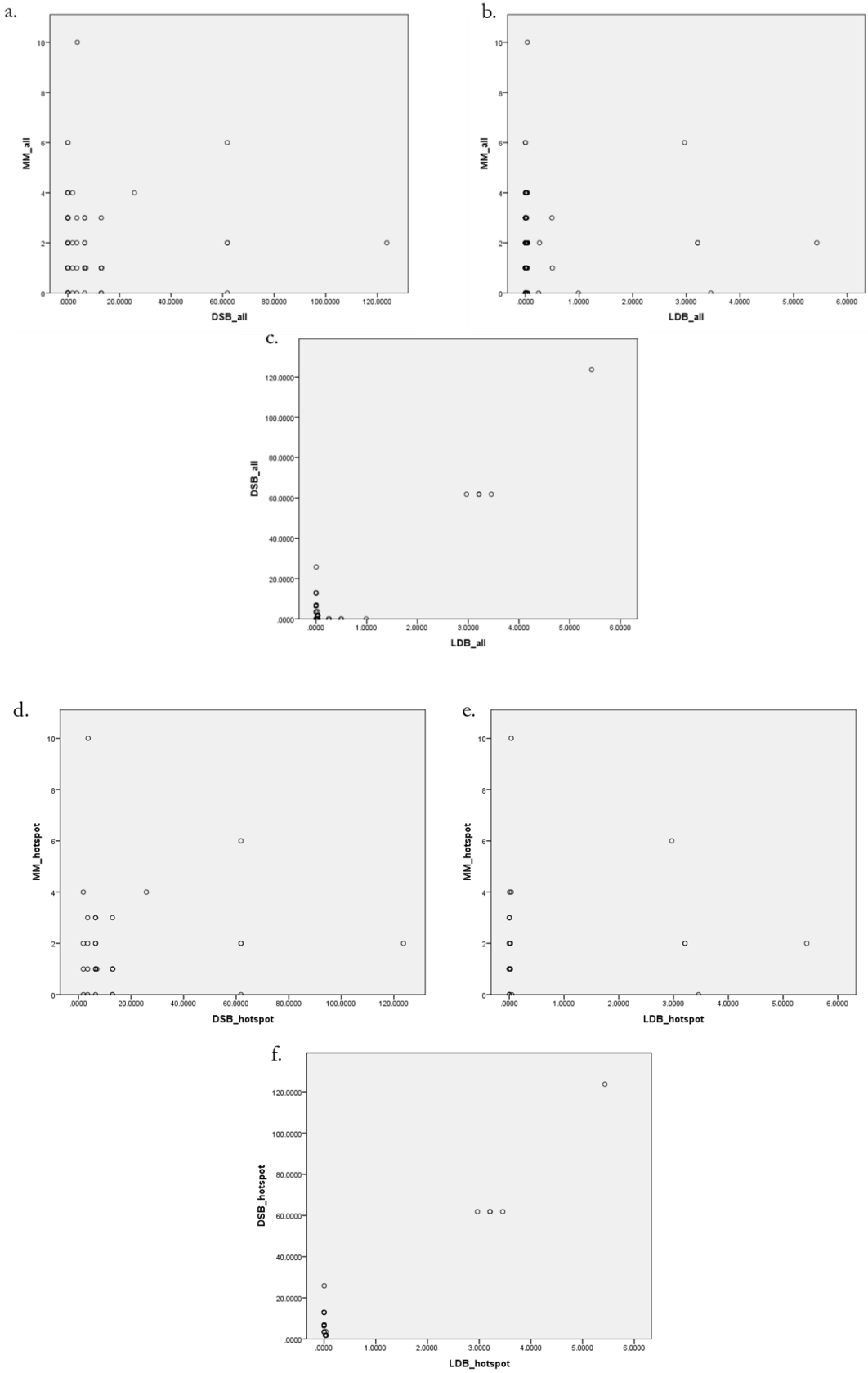


Figure 4.7 Scatter plots between variables. Graphs in the former group (**a**.-**c.**) are MM_all v. DSB_all, MM_all v. LDB_all and DSB_all v. LDB_all, respectively while the latter group (**d.**-**f.**) are similar but for those with "_hotspot" suffixes.

From the plots above, none of relationships was of a monotonic form, so Kendall's τ_b was the most appropriate test; the resulting correlations are shown in Table 4.5 below. This showed very low (only DSB_all *v* LDB_all) to no correlation (the remaining pairwise correlations).

Table 4.5 Pairwise τ_b correlation coefficients and p-values.

	τ_b					
	MM_all	**DSB_all**	**LDB_all**	**MM_hotspot**	**DSB_hotspot**	**LDB_hotspot**
MM_all		0.038	0.036	-	-	-
DSB_all	0.649		0.154*****	-	-	-
LDB_all	0.640	0.055*****		-	-	-
MM_hotspot	-	-	-		0.014	0.149
DSB_hotspot	-	-	-	0.926		0.076
LDB_hotspot	-	-	-	0.302	0.591	
p-value (2-tailed) (* = significant)						

Considering only 8/8-matched Myers' motifs, Figure 4.5 clearly supports the idea that there is no significant correlation between either the number or position of Myers' motifs and the position and strength of PRDM9*AA* DSB hotspots as 8/8-matched motifs are scattered along the entire sequence, and also did not cluster in DSB hotspots. There was also no perfect Myers' motif inside the DSB hotspots X2, 3, 6 and 7.

4.3 Discussion

4.3.1 Significance of HWE and MAF in LD analysis

A test for Hardy-Weinberg proportion is supposed to be a prerequisite before moving to the next step in most statistical calculations in population genetics. Although HWE is based on the traditional assumptions that alleles at a locus should be transmitted via random mating and should not be subject to any evolutionary forces e.g. mutation, migration or genetic drift, passing a HWE test does not guarantee that allele frequencies are always free from those factors. Departure from HWE results in gain or loss of heterozygosity that may accentuate certain haplotypes and either inflate or deflate LD. Common factors which cause departure from HWE and potentially affect LD are genetic drift, population expansion, admixture and migration, population structure, age structure, directional selection, high mutation-rate

SNPs, positive assortative mating, non-random mating, as well as sampling and genotyping errors [Ardlie *et al.*, 2002, Waples, 2015].

Genetic drift tends to inflate LD [Slatkin, 2008], especially when at least one of the linked markers is an advantageous allele (so-called genetic hitchhiking) while a deleterious-allele deflates LD [Dapper and Payseur, 2017, Martin *et al.*, 2006, Ardlie *et al.*, 2002]. This effect will be obvious in a small-sized or finite population [Martin *et al.*, 2006, Felsenstein, 1974, Roze and Barton, 2006] whereas population expansion tends to reduce LD by decreasing drift [Ardlie *et al.*, 2002]. A finite population itself usually results in increasing heterozygotes and production of offspring with random LD [Waples, 2015].

Admixture and migration or gene flow can create false LD especially when populations with different allele frequencies mix; however, by random mating, this effect will abruptly decay in the next generation if there is no real recombination [Ardlie *et al.*, 2002]. In this case, HWE indicates whether the effect of gene flow is resolved by random mating or not. Only directional selection will deviate HWE and might create LD [Korol and Iliadi, 1994].

Population substructure or inbreeding, as well as a bottleneck effect, results in an increase of LD, similar to non-random mating and positive assortative mating. The net outcome is to reduce a number of heterozygotes as well as to increase either of the homozygotes. A similar thing may be caused by the so-called Wahlund effect namely population substructure, as well as age structure which creates a Wahlund-like effect [Slatkin, 2008, Waples, 2015]. SNPs with a high mutation rate such as those within CpG islands often slightly inflate LD though there is no real linkage between a SNP pair [Ardlie *et al.*, 2002].

Finally, high-throughput genotyping, for instance by MPS or SNP microarray, is prone to cause errors. One survey found that genotyping error usually resulted in gain of heterogeneity while loss of heterogeneity favoured evolutionary causes such as population substructure [Chen *et al.*, 2017]. A null allele, which is caused by mutation at a primer or probe-binding site, is an example of a technical error, as well as other methodological errors such as non-random sampling especially in a small population size, which may bias genetic diversity then interfere with LD [Waples, 2015]. These factors can be easily screened in the first place by a HWE test.

To test HWE in the human X chromosome, the classical approach is to use only allele frequencies in females and ignore those in males. Graffelman and Weir [Graffelman and Weir, 2016] found that when HWE is reached, it is able to inferred from the fact that, through time, allele frequencies in both sexes become the same. Therefore the frequencies

from both sexes can be summed up and calculated together. This method was suggested to apply to the X-specific region but not to pseudoautosomal regions. Another issue is that if a HWE test is performed using the popular χ^2, a locus with low MAF (<0.20) tends to exceed the nominal α error (0.05), in other words, it is likely to stay in HWE. By contrast, use of the other test, i.e. the standard exact test, tends to reject HWE. In this study, χ^2 was applied to HWE test and therefore, in the data set containing loci with MAF < 0.20, more SNPs might be expected to pass HWE, and this might affect the result of the LD analysis.

In Figure 4.2, by the new method (HWE_n) compared to the classical one (HWE_c), there were more SNPs passing the HWE test (HWE_n:HWE_c = 555:388). And as stated in the previous paragraph, 123 more SNPs passing HWE with MAF <0.20 were found exclusively by the new method against only 5 of those of the classical way. However, it increased in a number of SNPs with MAF >0.20 uniquely found in HWE_n compared to just 1 in HWE_c that was consistent with the observations of Graffelman and Weir [Graffelman and Weir, 2016]. This shows an advantageous point that the new method is able to uplift a significant number of informative SNPs that help improve clustering and distribution of markers for an LD analysis.

Additionally, apart from HWE, there are other aspects concerning allele frequency to be taken into consideration, including the minor-allele frequency (MAF) of each SNP and how this matches between each pair of SNPs. In an ideal mathematical sense, allele frequencies of each SNP ought to be in balance as much as possible so that MAF tends towards 0.5, and MAFs between SNPs should match. SNPs with matched-MAFs show significantly stronger LD than unmatched [Eberle *et al.*, 2006]; however, it is sometimes difficult to find such a suitable pair of SNPs that are an appropriate distance apart. So, in a general practice, calculation of an LD statistic requires MAF which is higher than a critical or minimum level. Previous papers have chosen a wide variety of minimum MAF which ranges from ≥0.05 [Kaessmann *et al.*, 2002], ≥0.1 [Eberle *et al.*, 2006], ≥0.15 (which was the most frequently selected) [Holloway *et al.*, 2006, Jeffreys *et al.*, 2001, Kauppi *et al.*, 2005, May *et al.*, 2002], ≥0.2 [Jeffreys and May, 2004] and 0.25 [Jeffreys *et al.*, 2001]. These values are considered to be beyond the level found in rare alleles (MAF ~ 0.01-0.05) and the set of such SNPs are called tag SNPs [The International HapMap Consortium, 2007]. SNPs with MAF 0.1-0.2 or ≤0.15, which are considered as low-MAF SNPs, are likely to produce longer LD blocks and less clear breakpoints than those with high MAF (MAF 0.4-0.5), and also present low LR in LD analysis [Eberle *et al.*, 2006, Jeffreys *et al.*, 2001]. For this reason, SNPs with MAF ≥ 0.2 were preferred in this study since low-MAF SNPs potentially create false LD with high uncertainty

that may obscure adjacent genuine LD blocks and provide a blurred picture of LD. As in Figure 4.3, the LD plots analysed from the dataset of 103 SNPs with MAF ≥0.20 (MAF20 group) showed clearer LD and regions of definitive breakdown than the dataset of 132 SNPs with MAF ≥0.15 (MAF15 group), though it contained ~22% less SNPs. It also showed ambiguous LD breakdown positions such that the Position 5 in Figure 4.3 b. was replaced by a probably false LD in Figure 4.3 a. and there was an additional unclear breakdown distal to the Position 6. Therefore the LD plot from MAF20 dataset was eventually chosen to create the LD block diagram in Figure 4.4, to define the breakdown intervals in Table 4.3, and for other downstream analyses.

4.3.2 Coincidence of DSBs and LD breakdown in the distal X-specific region

PRDM9 protein plays a major role in formation of DSBs of which ~10% are repaired by recombination. The predominant type of PRDM9 is produced from the A allele and its zinc fingers are found to likely bind the hotspot-associated DNA sequence, i.e. Myers' motif, that is associated with ~40% of human recombination hotspots [Myers *et al.*, 2008]. For this reason, to locate Myers' motifs in genomic regions of interest may help identify potential recombination hotspots especially in individuals who have the PRDM9-*AA* phenotype. Protein produced from the *PRDM9*-A allele mostly binds to the 8/8-matched Myers' motif; however, it is also able to activate hotspots that do not contain 8/8-matched motif at their centres [Berg *et al.*, 2010]. Similarly, PRDM9 from alleles other than A can bind to several forms of motif with different affinities even outside hotspots [Altemose *et al.*, 2017]. In this study, 7/8- and 8/8-matched Myers' motifs were located in order to see how they clustered in relation to either DSBs or regions of LD breakdown. However, no apparent correlation was found as the motifs were scattered almost evenly along the entire region without a significant cluster corresponding to DSB or LD hotspots. Altemose *et al.* addressed some issues by questioning PRDM9's binding properties, which PRDM9-binding sites go on to become recombination hotspots, and other associated factors, e.g. chromatin features in *cis* [Altemose *et al.*, 2017]. Although in their study, PRDM9-B, a minority allele in the European population [Berg *et al.*, 2010, Baudat *et al.*, 2010], was used as it was the allele possessed by the individual who contributed to the human reference genome sequence for *PRDM9*, 88% of recombination hotspots in *AA* individuals were found to be shared with those in *AB* individuals [Pratto *et al.*, 2014], so the results and conclusion could be applied and might be able to explain what had been found above.

Not all Myers' motifs are bound by PRDM9 with several possible explanations. First, based on a bioinformatics method, it was found that at least 17 types of motifs were plausible to be bound by PRDM9 protein and seven distinct non-degenerate motifs, i.e. a conservative Myers' sequence, overlapped with 53% of all PRDM9-binding hotspots [Altemose *et al.*, 2017]. However, even the top-scoring motif-matched regions which accounted for 0.1% of all motifs had approximately a 50% chance of PRDM9 binding as detected by ChIP-seq genome-wide [Altemose *et al.*, 2017]. Second, of 13 PRDM9 ZFs carried by alleles A and B, only ZFs 7 to 11 are already known to bind Myers' motif, Altemose *et al.* found that the first six ZFs were able to bind other sequences and it was assumed to be a factor that might enhance DNA binding of PRDM9 [Altemose *et al.*, 2017]. In conclusion, a very small number of the overall predicted motifs will be actual PRDM9-binding sites.

Besides the motifs which serve as a PRDM9-binding region, there were motifs that were discovered to influence the formation of recombination hotspots, which were called 'non-PRDM9 recombination-influencing motifs'. One of the strongest influencing motifs is ATCCATG, which presented within PRDM9-binding regions and surrounded PRDM9-binding motifs in *cis* [Altemose *et al.*, 2017]. In THE1B repeats which were retrotransposons with >20,000 positions distributing across human genome and were identified as one of the strongest recombination hotspots in human [Myers *et al.*, 2008], the ATCCATG motif showed ~2.5-fold reduction of average recombination rate [Altemose *et al.*, 2017]. Moreover this motif, alongside 17 other motifs were also found in a strong association with the histone modification H3K9me3 as well as H3K4me3, which was believed to be the strongest predictor of a heterochromatic condition, and also correlated with meiotic recombination reduction [Altemose *et al.*, 2017]. In addition, PRDM9 binding showed no association with molecules linked to meiotic DSBs such as DMC1, which was more strongly correlated to recombination rates at a genome-wide scale.

In summary, the presence of a Myers' motif which is the strongest association with recombination hotspots, does not guarantee binding of PRDM9, however absence of Myers' motifs or other PRDM9-binding motifs is very likely to decrease the chance of DSB induction and ultimately recombination. Moreover, apart from PRDM9-binding motifs, there is also another motif which affects the binding of PRDM9 and contributing to DSB formation, therefore only looking to Myers' motifs is not a good predictor of a recombination hotspot and reflects no association to LD breakdowns. However, as described elsewhere, crossover occurs from ~10% of DSB repair, so the result in Table 4.5 above that showed ~15% correlation between DSBs and LD breakdowns, despite a border-

line statistical significance ($p = 0.055$), is more or less consistent with the general observations.

4.3.3 Candidate recombination hotspots for a sperm crossover study

LD represents the outcome of historical recombination hotspots and is calculated from population data that might be biased by *PRDM9*-allele polymorphism and mutation in PRDM9-binding sites. This study chose the CEU population which is best matched to the semen donors who carry PRDM9-*AA* genotype because the A allele is found in >90% of the population; however, changes in gene equilibrium through time, for instance genetic drift or gene flow, might distort historical LD breakdowns. Moreover, recombination hotspots have various life cycles which may arise or decay over time [Jeffreys and Neumann, 2009]. Therefore using a cross-sectional and very recent data such as the meiotic DSB map [Pratto *et al.*, 2014] in this study might help ensure an identification of recombination hotspot and reduce experimental failure and waste of time if an LD breakdown is not active in a sperm crossover assay, as has been found previously [Kauppi *et al.*, 2005]. In addition, the DSB data also showed a quantitative strength of signal which could be compared with a recombination rate from a sperm crossover study.

The strategy for selecting a candidate recombination hotspot for a sperm crossover study which will be presented in the next chapter is to focus at the hotspots that have region of LD breakdown coinciding with DSBs. In Figures 4.5 and 4.6 above, Regions X1 and X5 were chosen because they match with the abovementioned criteria and also have high strength of DSB signals. Moreover, Hotspot X1 and X5 have distinct DSB strengths which might make it interesting to explore a correlation between DSBs and recombination rates.

4.4 Conclusion

Target regions for the first candidate fine-scale recombination-hotspot sperm study of the ePAR were identified here substantially by LD analysis and DSB mapping. First of all, Myers' motifs were identified by a bioinformatics method and also SNPs from the database were plotted across the region of interest in the X-specific portion of ePAR. Then LD analysis from the CEU population, matched to that of the semen donors, was created by selecting an appropriate set of informative SNPs using the novel approach of HWE testing in bi-allelic markers of the X chromosome described by Graffelman and Weir (2016) and setting

a minimum threshold for MAF as >0.20 to reach the best result. Coincidence of DSB hotspots and regions of LD breakdown was a criterion. Despite a low correlation between LD breakdowns and DSBs, the correlation coefficient was very close to the observed ratio of meiotic DSBs repaired by crossover, additionally, DSBs might also better reflect recent activity of putative recombination hotspots. So consideration of both as a novel approach to identify putative recombination hotspots in this study should be better than using LD alone as previously done. Finally, the hotspots where DSBs coincided with LD breakdown alongside DSB signal strength were taken into account, and as consequence Regions X1 and X5 were selected to be subsequently analysed via a sperm crossover assay according to the criteria and their differential DSB signal strength.

Chapter 5 Recombination Hotspots in the ePAR characterized by sperm-based crossover analysis

5.1 Introduction

Localisation of recombination events can be determined via population-based LD or pedigree-based analyses however the resolution is generally coarse (>10kb) [Kauppi *et al.*, 2004] though recently, by using a large dataset of pedigree-based meioses, up to 30% of events could be mapped to a resolution of <10kb resolution, with the finest resolution being ~2kb [Bhérer *et al.*, 2017]. Routine fine-resolution characterization of based on the principle of sperm typing was developed in the late 1980s [Jeffreys *et al.*, 1998] and this not only to detected clustering of events into so-called recombination hotspots to intervals ~<1-2 kb [Kauppi *et al.*, 2004], but also allowed exploration of other aspects of meiotic recombination dynamics such as noncrossover gene conversion and transmission distortion (TD) or reciprocal asymmetry within a hotspot [Kauppi *et al.*, 2009, Jeffreys and Neumann, 2002, Sarbajna *et al.*, 2012, Odenthal-Hesse *et al.*, 2014]. As discussed in Chapter 4, DSB mapping data [Pratto *et al.*, 2014] have been explored in this study to identify potential hotspots of interest alongside conventional LD analysis; using sperm typing approaches to detect *de novo* recombinant events would be the gold-standard approach to not only confirm hotspot location and properties, but also to compare the recombination activities between the two ePAR-carrying men and make comparisons with the published DSB strengths.

5.1.1 Sperm crossover assays to study the dynamics of recombination hotspots

Due to a lower resolution of LD breakdown and in some cases DSB hotspots (ranging 800 bp – 9.2 kb in X chromosome), sperm crossover experiments may reveal hidden recombination hotspots within one region of LD breakdown, for example, crossover activity indicated two hotspots 1.4-kb-apart coincided with the LD breakdown associated with the *NID*2 gene, a part of the class 2 of human major histocompatibility complex (MHC class II) [Jeffreys *et al.*, 2005]. In addition, though all hotspots analysed to date via sperm assays have very similar hotspot widths (~1-2 kb wide), other properties, for instance, variation in recombination frequencies between men, reciprocity of transmission between haplotypes and the centre of peak crossover activity can be explored by this approach [Jeffreys and

Neumann, 2002, Jeffreys and Neumann, 2005]. This information can further our understanding of the molecular mechanisms involved or can be used to predict a fate or history of each hotspot [Jeffreys *et al.*, 2004, Jeffreys and Neumann, 2009, Jeffreys *et al.*, 2005].

Methods to conduct the sperm experiments were presented in Chapter 2 however the underlying principles and flow of procedures are described here. Once a candidate recombination hotspot is identified, at least two heterozygous SNPs, so-called selector sites, must be found upstream and downstream of the proposed region. To achieve this, the entire candidate region including the extended flanking regions at both ends should be amplified by a long-range PCR, and sequenced to identify every heterozygous marker in the semen donors of interest. In this study, the putative hotspots were selected based on meiotic DSB clusters, so the selector sites were chosen to flank these clusters. In order to successfully carry out a crossover assay, all the heterozygous markers must be phased to establish the linkage haplotypes. In this study, haplotype separation was achieved using allele-specific primers (ASPs) designed for the selector sites and linkage phasing was achieved by allele-specific oligonucleotide (ASO) hybridization with radioactive labelling (γ-^{32}P-ATP) against dotblots of relevant PCR amplicons. Once phase is established, two rounds of nested allele-specific PCR (AS-PCR) using a forward ASP from one parental haplotype in conjunction with a reverse ASP of the opposite haplotype in each round is performed to selectively amplify *de novo* recombinant sperms from hundreds of gametes in each of dozens of separate reactions [Kauppi *et al.*, 2009]. The resulting PCRs are dotblotted onto a nylon membrane, and the heterozygous markers (SNPs or indels) inside the putative hotspot region are then typed by ASO-hybridization [Kauppi *et al.*, 2009]. Theoretically only recombinants should amplify and the crossover breakpoint intervals can be identified via a switch between parental haplotypes [Kauppi *et al.*, 2009]. However AS-PCR may not always produce a perfect allele-specific amplification and therefore can on occasion amplify parental haplotypes, a phenomenon known as bleed-through that is easily detected when typing internal markers. A summary of a sperm crossover experiment is presented in Figure 5.1.

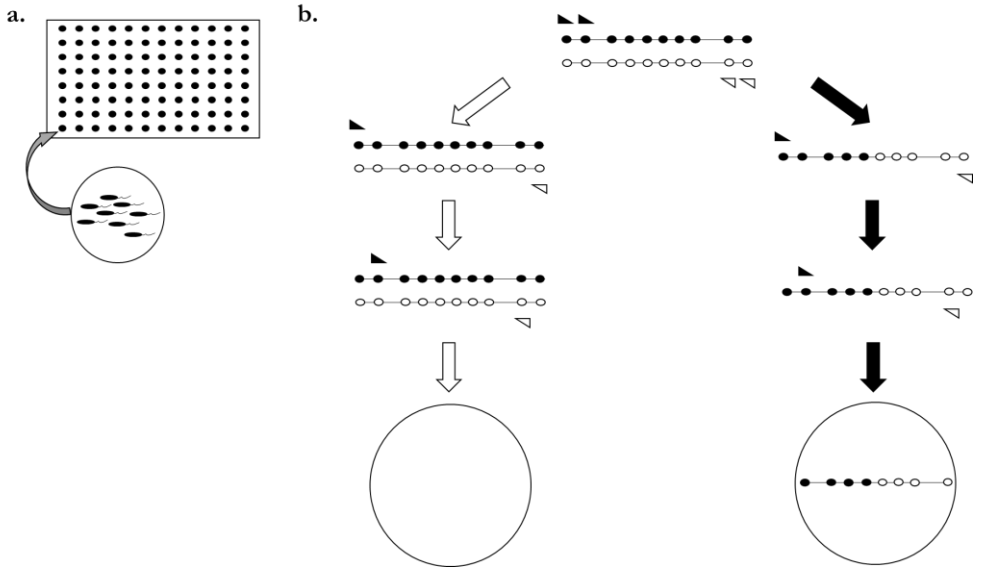


Figure 5.1 **AS-PCR in a sperm-crossover assay. a).** shows an equal pool size of hundreds of DNA molecules per reaction and is performed in a well plate. **b).** shows AS-PCR to detect recombinant molecules. ASPs are represented by black and white isosceles triangles. The upper two strings of black and white dots represent distinct haplotypes with indicating the position of heterozygous markers. The left column (white arrows) illustrates a PCR reaction that does not contain a recombinant and the right column (black arrows) depicts a reaction containing a recombinant. If there is no recombinant, theoretically, no PCR product will be generated in either the first or second round whereas in the presence of a recombinant, amplicons will be produced in the first round and will be amplified by nested-PCR in the second round to ensure that the products are from genuine recombinant molecules rather than from a mis-primed parental haplotype. Typing of the internal informative markers will identify the crossover breakpoint interval. In general, a recombinant event is rare so only one event usually presents in a given PCR reaction, however occasionally more than one event will be present in which case a reaction may give a "mixed" signal with ASO hybridization (i.e. both ASOs will hybridize).

5.1.2 Half-crossover assay allowing study of noncrossover gene conversion

Similar to a crossover assay, a half-crossover assay has two rounds of nested PCR and uses ASPs but the only one at one end in conjunction with an universal primer at the other end of each round [Kauppi *et al.*, 2009]. The outcome PCR product, in contrast to a full-

crossover experiment, is a mixture of one parental haplotype amplicons with, if present, a PCR product of any recombinant sperm. Since recombination generates either crossovers (COs) and non-crossovers (NCOs) including gene conversions, by this method, gene conversion events can also be detected.

Since a half-crossover assay will always amplify one parental haplotype, the number of input DNA molecules per reaction should be much lower than a full-crossover assay – of the order of tens, typically not over 40, instead of hundreds [Kauppi *et al.*, 2009]. For this same reason, only ASOs corresponding to the alleles of the opposite haplotype to that intentionally amplified are used for mapping events. Both CO and NCO rates as well as breakpoints and other recombination features including TD can also be detected via a half-crossover assay but a drawback of this method is that it much more labour-intensive when the recombination rate is very low, especially lower than 0.02% since a typical maximum input molecules are 3,840 per plate. A summary of this method is shown in Figure 5.2.

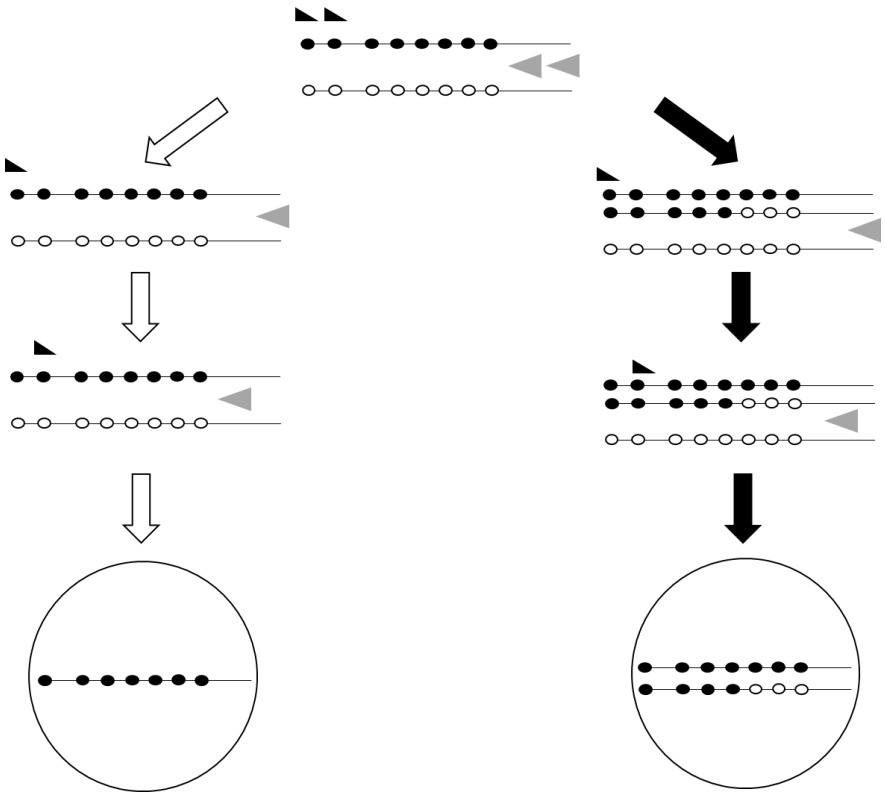


Figure 5.2 **Sperm half-crossover assay.** Two rounds of nested PCR are performed with one ASP (black scalene triangles) at one end with a universal primer (grey isosceles triangles) at the other end. The upper two strings of dots represent distinct parental haplotypes. The left side shows amplifications in

cases where a PCR contains only parental molecules (black string of dots). In contrast, the right side shows a PCR reaction containing a recombinant molecule among several parental molecules.

5.1.3 Survey of informative markers in the ePAR

As described above, the most important components of a crossover (or half-crossover) study are heterozygous markers. Those heterozygous, or informative, markers acting as ASP-binding sites and breakpoint indicators should extend over an interval that is within the capability of long-PCR, <20 kb, and ideally internal markers should be numerous and evenly spaced over the interval, e.g every 200 bp to 1 kb. To survey informative markers, the candidate regions with extension of a few kilobases at each end should sequenced. As the candidate regions in this study, Hotspot X1 and X5, were selected based on meiotic DSB maps as described in Chapter 4, each region including its extension is approximately 5-6 kb long. However, since it was impossible to predict a number of informative markers for each of the ePAR-carrying semen donors in each of these two candidate regions, all seven DSB hotspot intervals were sequenced. Given the overall length of each region, massively-parallel sequencing (MPS) as opposed to Sanger sequencing was used. In fact, the entire 110-kb X-specific portion of the ePAR was sequenced in this way and the full data set will be discussed in the next Chapter.

5.2 Results

5.2.1 MPS resequencing of the meiotic DSB hotspot regions within ePAR

Both semen donors, Man 20 and 53, were subjected to MPS-resequencing for seven DSB regions ranging from 5.3 to 11.6 kb in length. The MPS platform used in this study was Ion Torrent™ which through the nature of the sequencing chemistry does not tackle long homopolymeric regions very well, however it was cost-effective considering the scale of the project which contained 20 individuals and also provided a degree of flexibility as the machine was located in the department. Each region was divided into two shorter overlapping amplicons of 3.3-6.5 kb long. Details of sequencing results covering the entire X-portion of ePAR will be presented in the next chapter however data related to the DSB hotspots especially the two candidate regions, X1 and X5, are shown in Table 5.1 (see

primers and PCR conditions in Appendix). In brief, the sequencing results revealed the median of depth of coverage of ~300× and ~250-bp of mean read length. However, it was found that ~19% of reads were unable to map to the reference sequence of the X portion but they were able to align with the proximal Yq instead. No matter where alternative primers were designed for this region testing with monochromosomal hybrid cell DNA demonstrated that they amplified in favour of the portion of Y chromosome.

Table 5.1 Summary of Ion Torrent™ resequencing of seven ePAR-DSB hotspots

Hotspot	chrX (hg 19) coordinates	Informative markers		Remarks
		Man 20	Man 53	
X1	2697499 - 2707585	22 SNPs	16 SNPs	Distribution covering target sites
X2	2716242 - 2724983	7 SNPs	16 SNPs	No selector SNP at one end
X3	2742080 - 2752829	-	-	Mispriming due to homology on Yq*
X4	2762180 - 2773240	7 SNPs	8 SNPs	Too few internal SNPs
X5	2782027 - 2793594	8 SNPs	13 SNPs	No selector SNP at one end
X6	2794166 - 2805090	5 SNPs	0	No informative marker in one man
X7	2803155 - 2808515	1 SNPs	0	No informative marker in one man

*Details and discussion will be presented in Chapter 6.

The sequencing results above showed that the two candidate hotspots identified in Chapter 3 were suitable for sperm recombination analysis for both Man 20 and Man 53. However, for Hotspot X5 given that there are limited informative markers for use as selector sites for each semen donor at one or other end of the interval, it was clear that it would be necessary to develop half-crossover assays. Of the other five DSB hotspots, Hotspot X2, which also showed LD breakdown, might also have been a good target for sperm crossover analysis but lacked appropriate selector SNPs at one end. Had time permitted this may also have been explored by a half crossover. In summary, the candidate hotspots selected from meiotic DSB mapping in conjunction with an LD analysis were found by resequencing of all the DSB cluster regions to be most suited for sperm recombination analysis using the two identified ePAR-carrying semen donors. Unambiguous sequencing of Hotspot X3 proved difficult despite designing several different primer pairs for this region. Tests using monochromosomal hybrid cell DNA revealed co-amplification of the long arm of the Y chromosome. This technical issue reflects the fact that the region of the X chromosome that

transferred to form the ePAR shares a common evolutionary origin with this proximal portion of Yq, dating back ~30 Mya [Ross *et al.*, 2005].

5.2.2 Developing the crossover assays for the distal hotspot X1

Hotspot X1 contained sufficient informative markers for crossover selection and mapping in both semen donors such that a full-crossover sperm assay could be performed. Although this type of assay only detects COs, it is the most efficient means of establishing recombination rate, especially if it is low, since thousands of DNA molecules can be screened in a single experiment.

This assay requires two pairs of nested ASPs based on the principle that allele-specificity relies only on the last 3'-end nucleotide of each primer [Kauppi *et al.*, 2009]. To evaluate the specificity and efficiency of the designed ASPs, it was necessary to identify individuals homozygous for each of the allelic variants at each selector SNP. A panel of 96 DNAs was used for this purpose. SNP genotyping of the panel was performed by dotblot ASO-hybridization against two PCR amplicons generated with universal primers that covered the selector sites at each of the 5' and 3' ends. Each ASP in conjunction with a universal primer was tested with DNA from both homozygotes to find the right T_m. In some cases, a touchdown or semi-touchdown PCR was applied to gain the most specificity as well as high yield. An example of using a touchdown PCR is shown in Figure 5.3 and details of the ASPs and universal primers are presented in Appendix.

In the actual crossover assay, forward ASPs from one haplotype were coupled with reverse ASPs from the other haplotype as shown in Figure 5.4. The two different reciprocal recombinant orientations can be analysed in this way. For Man 53 it was necessary to split the primary AS-PCR into two steps since the T_m of the forward and reverse ASPs were too dissimilar (58ºC and 56ºC). For this reason, each round of the primary PCR contained one ASP with one universal primer (Figure 5.4 b). Subsequently in both cases, the AS-PCR products were dotblotted onto a nylon membrane and the internal markers were typed by ASO hybridization followed by autoradiography on X-ray film. Finally the autoradiographic data were used to map the crossover breakpoints [Kauppi *et al.*, 2009].

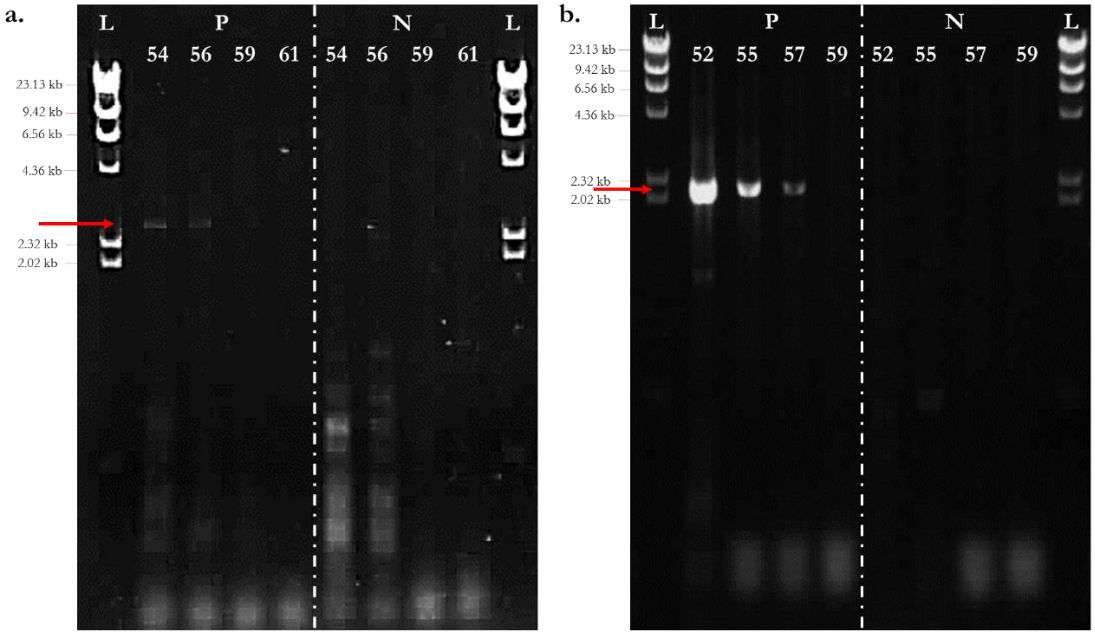


Figure 5.3 **Gel photo of the 1º-ASP testing for Man 20 in Hotspot X1.** Abbreviations stand for: L = λ (HindIII digest), P and N = allele-identical (positive) and allele-opposite (negative) DNA, the numbers corresponding to each lane indicated the annealing temperatures tested (ºC) and the red arrows indicate the expected sizes of amplicons. **a.)** trial of the forward 1º-ASP (namely 9.5F C) indicates its ability to amplify at annealing temperatures from 54 – 59 ºC, though the yield is very low at 59 ºC and short non-specific amplification is seen at both 54 and 56 ºC. **b.)** trial of the reverse 1º-ASP (namely 14.8R C) reveals amplification with annealing temperatures from 52 - 57ºC without generation of non-specific products. As a result, combination of both ASPs for the optimal result should not go beyond 57ºC, the upper limit of the reverse ASP, however, to gain a higher yield a touchdown PCR technique is applied by using an initial annealing temperature of 57 ºC for 3 cycles then dropping the temperature to 56 and again to 55 ºC for 5 and 18 cycles respectively.

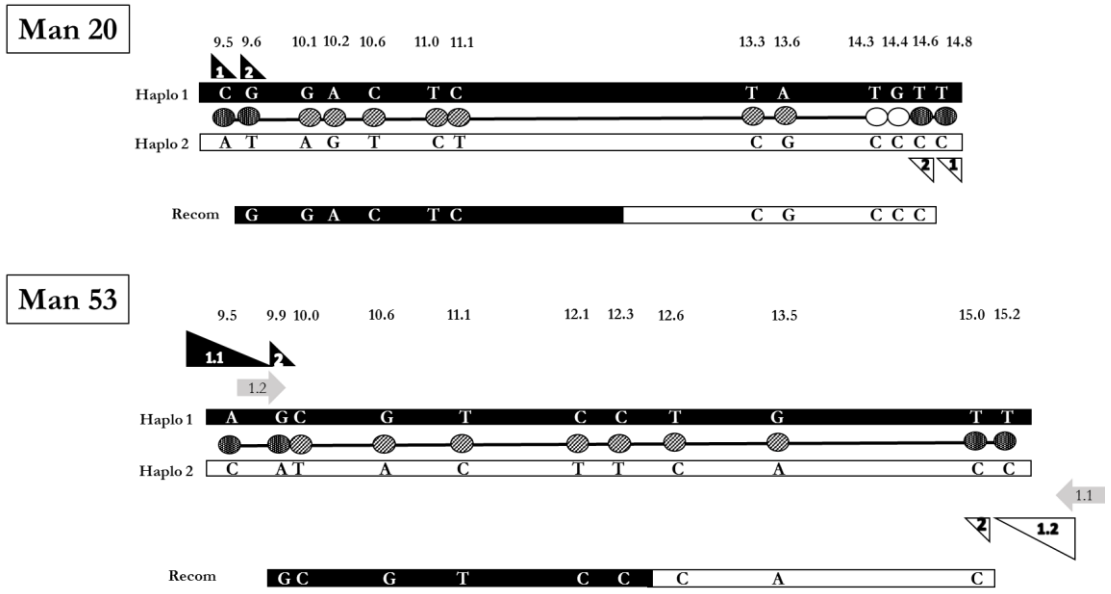


Figure 5.4 **Diagrams showing crossover assays in both semen donors.** Both panels demonstrate the assay in one orientation running from the 5' black ASPs of one haplotype to the 3' white ASPs of the other corresponding to the haplotypes of the same colours. Ovals-on-string represents positions of the informative markers where the internal markers (highlighted with diagonal lines) flanked by two selector sites (dark grain) at both ends. The upper panel presents the assay of Man 20 which uses two rounds of nested AS-PCR. ASPs are shown by the black scalene triangles for the haplotype 1 (black rod with allelic indication) and the white scalene triangles for the haplotype 2 (white rod with allelic indication), with numbers indicating the corresponding PCR round. The name of each marker is shown by a number on the top row. The lower panel shows the assay of Man 53; in this case the 1º AS-PCR is split into two nested rounds by using the forward ASP 1.1 in conjunction with the universal reverse primer 1.1 (grey arrow on the right) followed by the universal forward primer 1.2 (grey arrow on the left) with the reverse ASP 1.2; the 2º AS-PCR is done as usual. The product of an amplified recombinant molecule is shown beneath each panel.

5.2.3 Crossover analysis in Hotspot X1

Reciprocal assays were performed for both semen donor DNAs with pool sizes varying from 400, 500 and 600 input molecules per reaction. In total, 92,000 and 76,800 molecules were screened in Man 20 and Man 53 respectively. Collectively, a raw number of CO events of 40

and 131 were detected in Man 20 and Man 53. By performing Poisson correction (see the equation in Appendix) because the positive events are always under-counted [Jeffreys *et al.*, 1998, Kauppi *et al.*, 2009], the corrected number of *de novo* COs were ~42 and ~158 in Man 20 and Man 53, respectively, to make ~200 recombinants in a total of 168,800 sperms; thus showed ~4-fold significant difference in the RFs with p-value (2-tailed goodness-of-fit test) < 0.0001 whereby Man 20 RF was 0.05% (95% CI 0.03 - 0.06%) against that of Man 53 as 0.21% (95% CI 0.18 – 0.24%).

As the underlying morphology of hotspot CO activity was found to be approximated to a normal distribution [Jeffreys *et al.*, 2001, Jeffreys and Neumann, 2002, Kauppi *et al.*, 2004], the least-squares best-fit normal distribution was found for the combined crossover data from both Man 20 and Man 53 using the program written by Prof. Sir Alec Jeffreys and run in True Basic® v.4.1 language software [Kauppi *et al.*, 2009]. This is shown in Figure 5.5 alongside the histograms showing recombination activities for each of the assayed intervals for each of the two semen donors. Despite significant difference in the RFs, the centre points of the distributions of the donors were offset by just 18 bp and 95% of the CO events were found to cluster within an ~1.3-kb interval (Table 5.2). Furthermore, the combined sperm CO activity represented by the normally-distributed curve showed a very good correspondence to the DSB cluster (Figure 5.5).

Table 5.2 Peak and vicinity of least-squares best-fit normal distribution CO cluster in the distal Hotspot X1 (hg19)

Semen donor	Centre	95% CI of the activity cluster	Peak (cM/Mb)
Man 20	2701041	2703073 - 2701709	54
Man 53	2701059	2700444 - 2701674	262
Combined	2701047	2700402 - 2701691	150

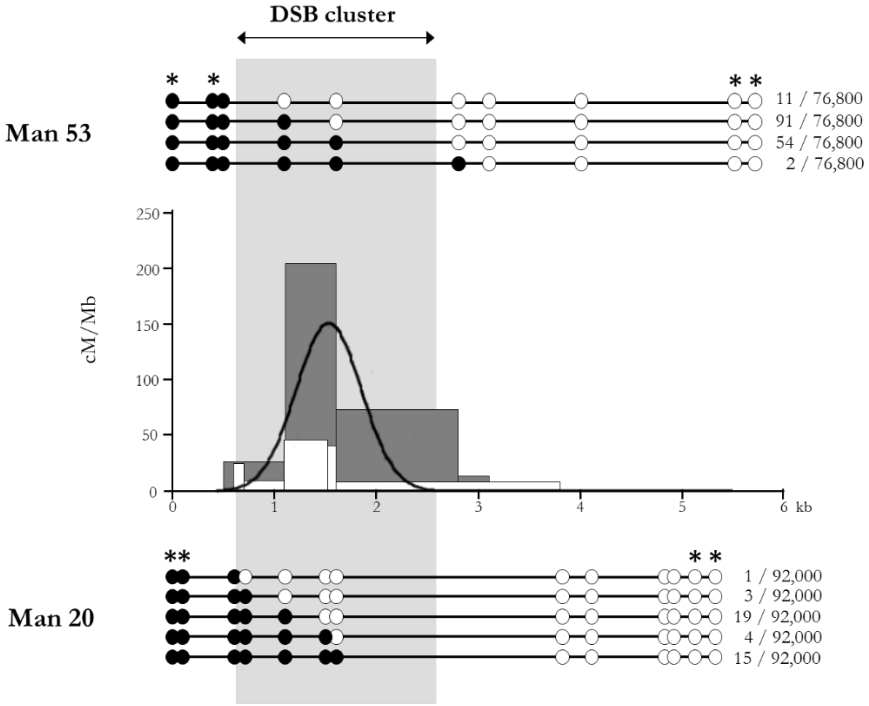


Figure 5.5 ***De novo*** **sperm crossover activity in the distal Hotspot X1.** Recombination activity is represented by histograms and normal-distribution curve in cM/Mb along the assayed intervals. Man 20's activity is shown by the white histogram while Man 53's activity is in the dark grey histogram. The combined least-squares best-fit normal-distribution of both donors is presented by the black curve. Schematics of the assays with locations of informative markers and recovered CO events are shown by a string of black and white circles on rod. The asterisks marked over adjacent markers at both ends represent the selector sites. The types and numbers of recovered recombinant molecules are shown for each man. The pale grey block lying over the assay diagrams and the CO-activity graphs shows the span of DSB cluster mapped by Pratto *et al.* (2014) which coincides with the sperm CO activity in this study.

Generally, COs should produce a 50:50 ratio of reciprocal events at each informative marker, however, many CO hotspots have been shown to exhibit significantly skewed transmission, which is called transmission distortion (TD), between alleles of the markers clustering adjacent to the centre of hotspot [Jeffreys and Neumann, 2002, Jeffreys and Neumann, 2005, Jeffreys and Neumann, 2009, Sarbajna *et al.*, 2012, Webb *et al.*, 2008]. In this study, it was found that TD occurred in Man 53 at rs1970797 (C/T), equivalent to the marker 11.1 in

Figure 5.4. This marker was found to exhibit an over transmission of the T over the C allele (0.62 *cf.* expected 0.50, p-value of 2-tailed exact binomial test = 0.008) and was equal to a gametic ratio of 50.024:49.976. The marker is also the closest to the combined centre of the hotspot, being displaced 126-bp downstream. Although Man 20 showed a 60:40 transmission ratio at rs4892890 (marker 10.6 in Figure 5.4) and this is close to rs1970797, a binomial test revealed this is not significant (p-value of 0.42). This lack of a significant departure from 50:50 is probably due to the small number of COs detected for this donor. Details of reciprocal transmission ratios across the internal markers with a demonstration of the TD marker are in Figure 5.6 below.

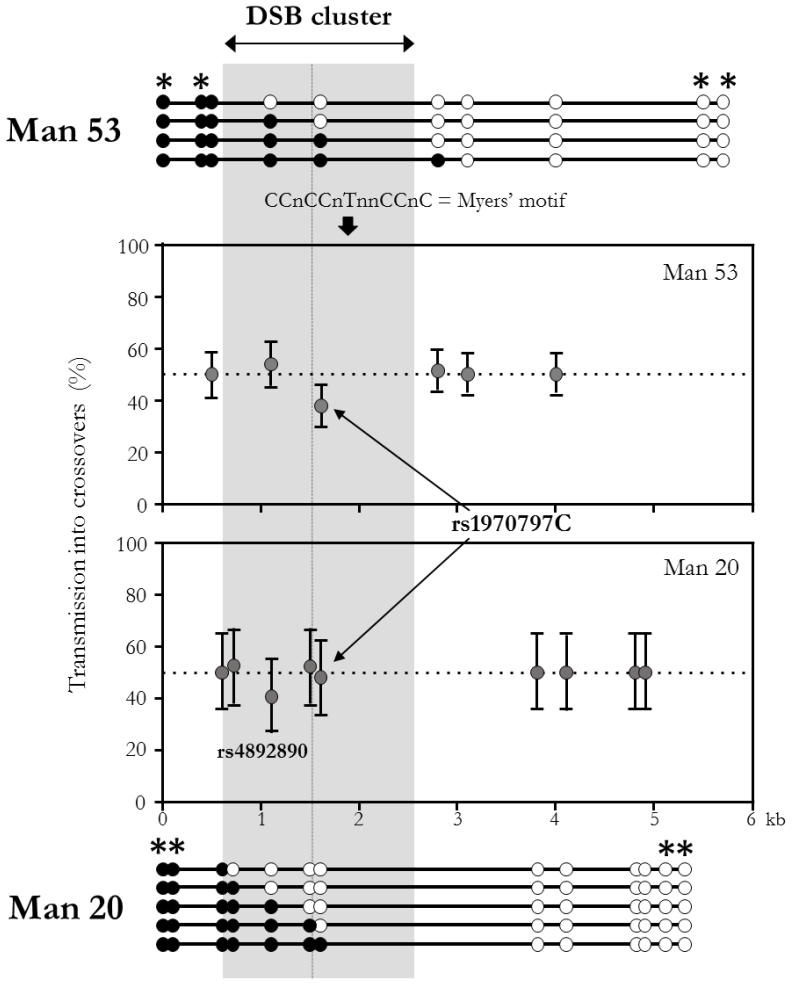


Figure 5.6 Transmission ratio of reciprocal CO event in the distal Hotspot X1. Transmission ratio between both haplotypes of each marker is determined by Bayesian analysis plotted as a dark grey dot with 95% CI whiskers. The horizontal black dotted line in the mid of each graph panel shows the expected 50% transmission of alleles into COs. The upper panel refers to

Man 53 and the lower one to Man 20. TD was found at rs1970797 in Man 53 indicated by the black arrow but not for same SNP in Man 20. The other informative markers showed ratios consistent with an expected 50:50 reciprocal transmission. rs1970797 lies 126-bp proximal to the estimated combined centre of hotspot presented as a vertical grey dotted-line and is located 210-bp distal to the closest Myers' motif as indicated by the black arrow on the top of the upper panel. SNP rs4892890 in Man 20 shows a gametic transmission ratio of ~60:40 similar to the rs1970797 in Man 53 but not statistically significant ($p = 0.43$, 2-tailed binomial exact) probably as a consequence of the small number of COs.

5.2.4 Developing the half-crossover assays for the proximal DSB Hotspot X5

A limited number of informative markers for both men to use as selector sites at one or other end prevented full-crossover assays being developed over the X5 DSB hotspot; to do so would have required >20-kb amplicons, which at best amplify relatively poorly and may be significantly influenced by the quality (molecular weight) of the different sperm DNAs. The only option instead was to design half-crossover assays in which ASPs are used in conjunction with universal primers to amplify the recombinant product from one haplotype together with the parental haplotype. The recombinant haplotype will be detected by the presence of non-amplified-haplotype alleles and it is able to be mapped a breakpoint as similar as a full-crossover assay. Moreover, this type of assay provides an advantage that NCO gene conversion could also be detected. However, as detection of recombination events in a half-crossover assay depends on hybridization, this method is less efficient as inputs per PCR reaction are of the order of tens of sperm DNA molecules (usually 20-40 molecules) [Kauppi *et al.*, 2009]. This approach is very labour-intensive work if the RF is low. In addition, interpretation of the ASO-hybridization data can be tricky since there can be a high level of background noise.

Designing a half-crossover assay follows a similar process as previously described of the assays in the distal Hotspot X1 but allows two rounds of nested PCR by using ASPs as selector sites at one end only and in combination with universal primers at the other end. In this hotspot, Man 20 has the informative markers that could be used for selector sites at the 3' end whereas the selector sites of Man 53 were located at the 5' end. The same DNA panel of 96 individuals was used to identify homozygotes for each of the allelic variants at each

proposed selector SNP by dotblot ASO-hybridization. Subsequently, each ASP in conjunction with a universal primer was tested with DNA from both homozygotes to find the optimal annealing temperature.

To obtain the maximum yields and specificity, pseudo-touchdown PCR conditions were applied to both rounds of one haplotype orientation and to the first round of the other for Man 20. In Man 53, touchdown PCR conditions were only performed in the first round of PCR for both haplotypes. The PCR conditions for both donors are presented in Table 5.3. And as opposed to the full-crossover assays done in Hotspot X1, only 30 sperm DNA molecules per PCR were the input number.

Table 5.3 T_m in AS-PCR for each donor in the Hotspot X5

Donor	AS-PCR round	T_m (cycles)
Man 20	1st	60ºC (5×) → 58ºC (23×)
	2nd *	a.) 61ºC (10×) → 60ºC (18×) b.) 59ºC (28×)
Man 53	1st	60ºC (7×) → 59 ºC (8×) → 58 ºC (10×)
	2nd	62 ºC (25×)

* Two orientations used different conditions

To detect the recombinants, dotblots of the PCR products were subsequently hybridized with ASOs from the opposite haplotype to that which was amplified. In addition, PCR products containing the non-selected parental haplotype were used to make positive control series to determine allelic specificity of ASO hybridization as described in Chapter 2 [Kauppi *et al.*, 2009]. COs can be evident from the presence of positive internal markers of at least starting from the second one of the ASPs side and running continuously towards the terminal marker adjacent to the universal primers of the assayed interval while NCO gene conversion could be detected for one or two positive adjacent internal markers in between but not continuously running towards the terminal one [Kauppi *et al.*, 2009]. In some cases, presence of the only positive terminal marker is difficult to be distinguished between CO and NCO and might need to make an assumption. Diagrams of the half-crossover assays of both men are in Figure 5.7 below.

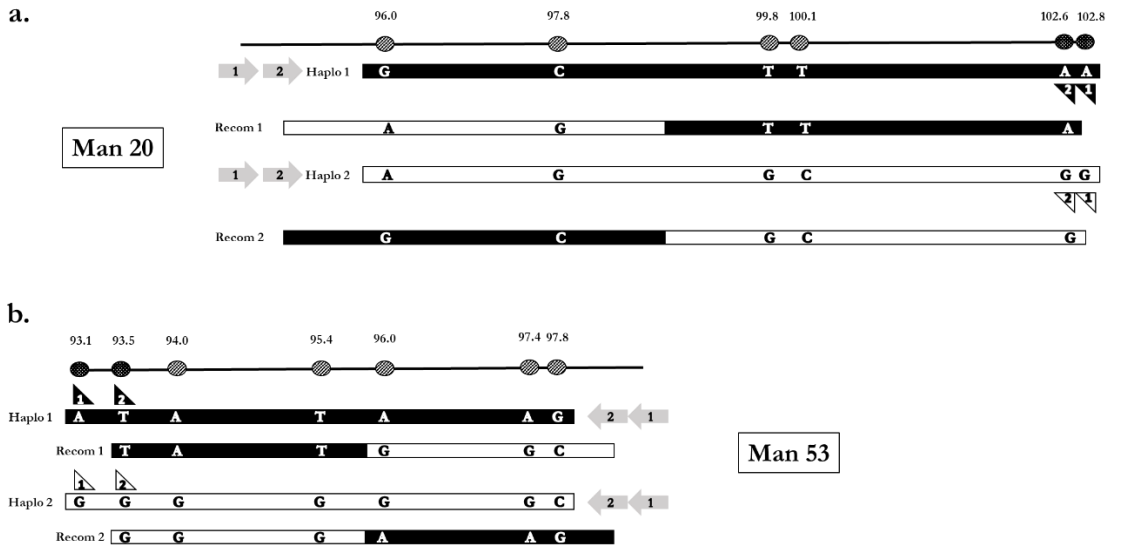


Figure 5.7 **Diagrams showing half-crossover assays in both semen donors.** Panel **a)** and **b)** demonstrate the assay of both haplotype orientations shown in black and white rods. Ovals-on-string represents the informative markers with their names on the top row: two selector sites (darker ovals) are at either end and the rest (lighter ovals) are internal markers. The universal primers used in conjunction with ASPs are shown as grey arrows with numbers corresponding to their ASP partner and PCR round. The upper panel **a)** presents the assays designed for Man 20 which use two rounds of nested PCR of which the reverse ASPs are shown by the black and white scalene triangles corresponding to each haplotype, all are indicated with numbers correlating to each PCR round. The lower panel **b)** shows the assays designed for Man 53. Each assay shows a single representative recombinant molecule for each orientation underneath the haplotype diagram.

5.2.5 Analysis of crossovers and noncrossover gene conversions in Hotspot X5

The Poisson-corrected number of recombinants per total screened molecules was 59/10,650 for Man 20 and 62/11,040 for Man 53, generating comparable recombination fractions for the two: 0.60% (95% CI 0.47 – 0.77%) *cf.* 0.51% (95% CI 0.39 – 0.66%) (p-value > 0.05, 2-tailed goodness-of-fit test) for Man 20 and Man 53, respectively. After taking definite NCOs out, COs including combined events, undistinguished COs and NCOs, at the terminal markers declined to 53/10,650 and 35/11,040 in each man, however, >90% of the events observed for Man 20 involved a switch only at the terminal marker; for Man 53 these kind

of events constituted about one-third of the total. In such cases, it is impossible to distinguish between COs and NCOs, therefore, in this study, about half of such events were arbitrarily assigned as COs in order to be able to analyse the hotspot morphology and estimate the RF. By this approach, the Poisson corrected number of the exclusive CO events were 32/10,650 for Man 20 *cf.* 28/11,040 for Man 53, thus, generating comparable RFs of 0.26% (95%CI 0.18 – 0.38%) *cf.* 0.25% (95%CI 0.18 – 0.37%), respectively.

As for Hotspot X1, the least-squares best-fit normal distribution curve was subsequently generated using recombination frequencies of each assayed interval as shown in Figure 5.7 which included only designated COs at the terminal markers. The CO activity curve also coincided with the DSB cluster as in Pratto *et al.* (2014) as was noted for Hotspot X1. Both men shared the estimated centre points of hotspot by which 95% of the CO event spanned within ~1-kb interval (Table 5.4) and there was good correspondence to the DSB cluster (Figure 5.8). Interestingly, the peak activity was estimated to be ~385 cM/Mb. However, if using the data containing the entire events without manipulation, the hotspot would reduce its span by 250 bp, the centre point would be shifted proximally by 116 bp and the peak activity would inflate to ~830 cM/Mb.

Table 5.4 Peak and vicinity of least-squares best-fit normal distribution CO cluster in the proximal Hotspot X5 (hg19)

Semen donor	Centre	95% CI of the activity cluster	Peak (cM/Mb)
Man 20	2787084	2786106 - 2788061	224
Man 53	2787492	2787335 - 2787710	842
Combined	2787386	2786920 - 2787853	385

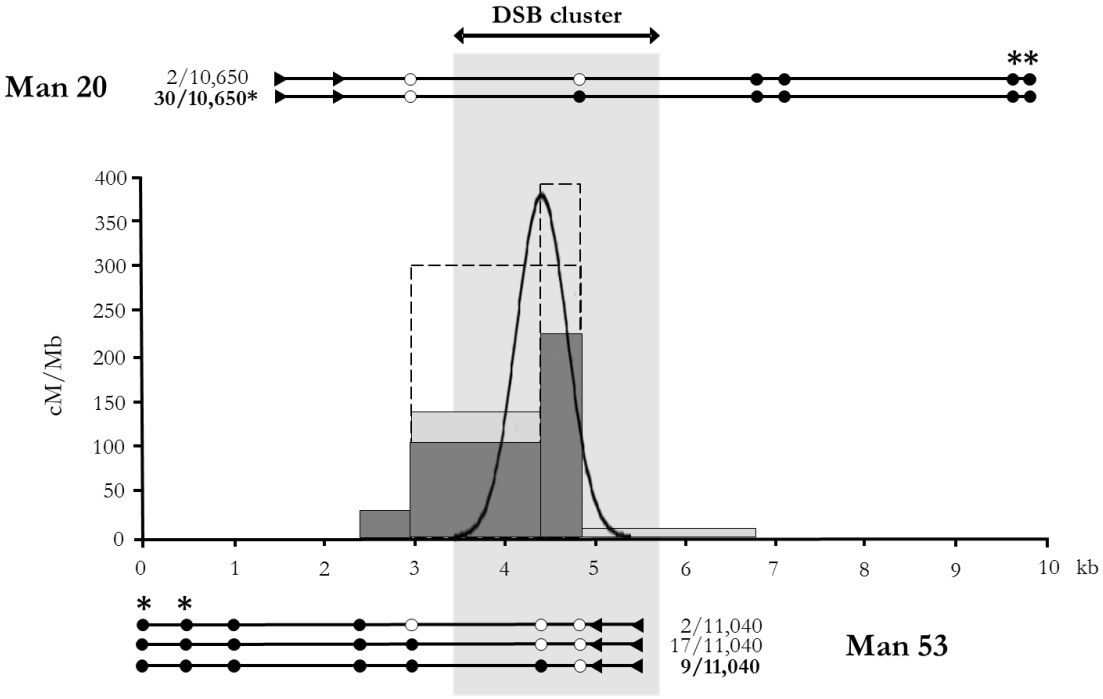


Figure 5.8 ***De novo* sperm crossover activity in the proximal Hotspot X5.** A pooled activity of reciprocal assays of each interval is presented in a histogram in relation to the informative marker intervals (white and black circles on strings) of each man; the upper panel is for Man 20 and the lower one is for Man 53. The light grey histograms show the recombination activities of Man 20 while the dark grey histograms represent those of Man 53. The activities mentioned here are based on a number of total COs which include half of the ambiguous events that occur between the terminal marker and the secondary primer site. The dashed histograms show how the activities would be if including all such events were classified as COs. The recombinant structures depicted by black (parental haplotype) and white (a switch to the opposite haplotype) circles on strings as described above show the ASP selector sites marked by asterisks over the black circles at one end together with the universal primers in black triangles. The frequency of each recombinant type is indicated; frequencies in bold are assigned COs for the events exclusively presenting at the terminal markers. The combined least-squares best-fit normal distribution showed by the black curve is generated from the net COs which include only assigned half of the events at the terminal markers and shows the peak activity of ~385 cM/Mb (if all such events were classified as COs, the peak activity would inflate to ~830 cM/Mb and the width of the hotspot would become X kb). The DSB cluster in Pratto

et al. (2014) is shown by a labelled pale-grey rectangle in the back ground which coincides with the combined hotspot distribution curve.

Twenty one unambiguous NCOs were observed; all involved just a single informative marker. However, assignment of half the terminal events as NCOs gave an aggregated NCO frequency of 27/10,650 (0.25%, 95%CI 0.17 – 0.37%) for Man 20, and 26/11,040 (0.24%, 95%CI 0.16 – 0.34%) for Man 53. Comparing the frequencies of COs versus NCOs, ratios of 1.04 and 1.08 were obtained for Man 20 and Man 53 respectively.

Collectively, the maximal conversion tracts ranged from 1853 to 2812 bp. For Man 53, the peak numbers of NCOs were detected at the SNP marker 97.4 which is the closest to the predicted centre point of hotspot, lying 97bp proximally. Interestingly, gene conversion occurred exclusively at this SNP without any co-conversion at the closest adjacent marker though it lies only 413 bp away.

As the assays were performed in reciprocal orientations for both men, TD was also tested. No TD was observed amongst the COs of either man, in contrast to what was found in the distal hotspot X1 above, however a statistically significant distortion (p-value = 0.011, one-tailed exact binomial test) was observed among NCOs just for Man 53 at the central-most SNP 97.4 (rs186282195). Nine out of ten events that encompassed SNP 97.4 were found to carry the G rather than an A allele, indicating a favourable repair of two-H-bond or 'weak' to a three-H-bond or 'strong' base pair. The NCO events and TD for Man 53 are displayed in Figure 5.9 below.

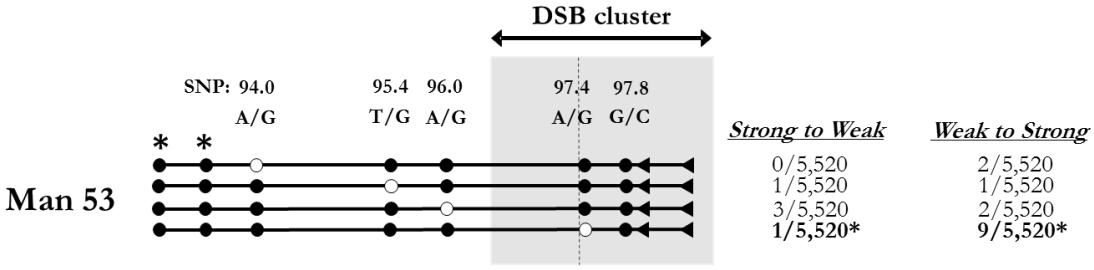


Figure 5.9 **NCO frequencies for Man 53 showing TD.** Testing for a biased transmission amongst NCOs for the informative markers of Man 53 that are consist of weak and strong alleles. SNP 97.4 (rs186282195) shows significantly over-transmission of the strong 'G' allele (9 NCOs) in relation to the weak 'A' allele (1 NCOs). This SNP is located 97 bp proximal to the

centre of the recombination hotspot indicated by the vertical dotted line. Irrespective of the actual number of NCOs at the terminal SNP 97.8, both allelic variants are strong base pairs and there is also no evidence of disparity between the orientations assuming half the events are NCOs, i.e. 4 against 3.

5.2.6 Comparison of ePAR Hotspots X1 and X5 to previous sperm-based autosomal and pseudoautosomal hotspot studies

Previously forty-three sperm-based recombination hotspots across 15 human chromosomes including PARs in the sex chromosomes have been analysed in Leicester. The recombination rates have been found to vary from 0.005 – 1.3125% with median RF = 0.096%, and the intervals of activity have been found to range from 0.94 – 2.5 kb, with a mean = ~1.5 kb [Berg *et al.*, 2010, Holloway *et al.*, 2006, Jeffreys *et al.*, 2001, Jeffreys and May, 2004, Jeffreys *et al.*, 1998, Jeffreys and Neumann, 2005, Jeffreys and Neumann, 2009, Kauppi *et al.*, 2005, May *et al.*, 2002, Odenthal-Hesse *et al.*, 2014, Webb *et al.*, 2008]. Both hotspot, X1 and X5 in this study, show RFs above the median of these previous studies but have comparable width as shown in Figure 5.10 which is plotted alongside DSB activities [Pratto *et al.*, 2014].

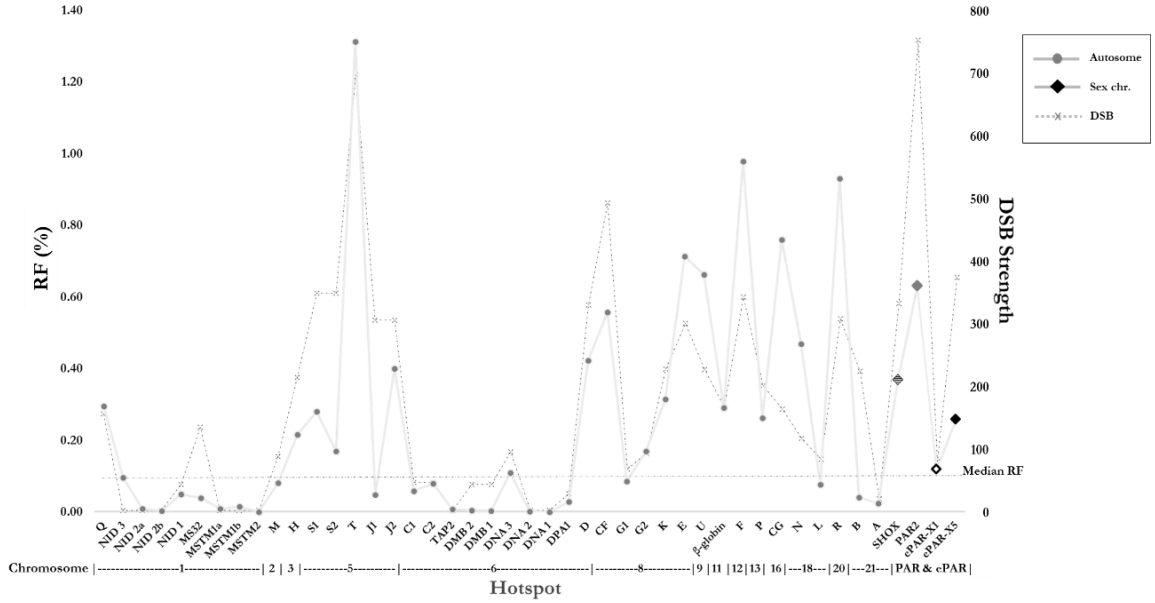


Figure 5.10 Sperm recombination hotspot activities across human genome. Scale on the left Y-axis shows RF in percent from previous sperm DNA recombination studies and the right Y-axis shows arbitrary DSB strength as ascertained from testes biopsy material [Pratto *et al.*, 2014]. The X-axis shows

the names of each hotspot and the hotspots are presented according to their chromosomal locations. The median RF of all previous sperm studies (0.096%) is represented by a horizontal dashed line: after including the RFs of this study, the median of RF slightly increases to 0.11%. Graph of the RF is in the solid grey line while that of DSB activity is in dotted line. The rates of autosomal and pseudoautosomal (including ePAR) hotspots are distinguished by rounds and diamonds, respectively.

Taking the DSB strength data from Pratto *et al.* (2014) which corresponds to each sperm-based hotspot into account, trends of the RFs and DSB strengths seemed to be concordant to each other. To calculate the correlation, both data were tested for normality and they were not in a normal distribution (Shapiro-Wilk's p-value < 0.001) then a scatter plot was generated and both looked monotonic as shown in Figure 5.11 below. Therefore, the non-parametric correlation of choice is Spearman's rho and it showed the significant correlation of 0.823 across the entire set (p-value < 0.001).

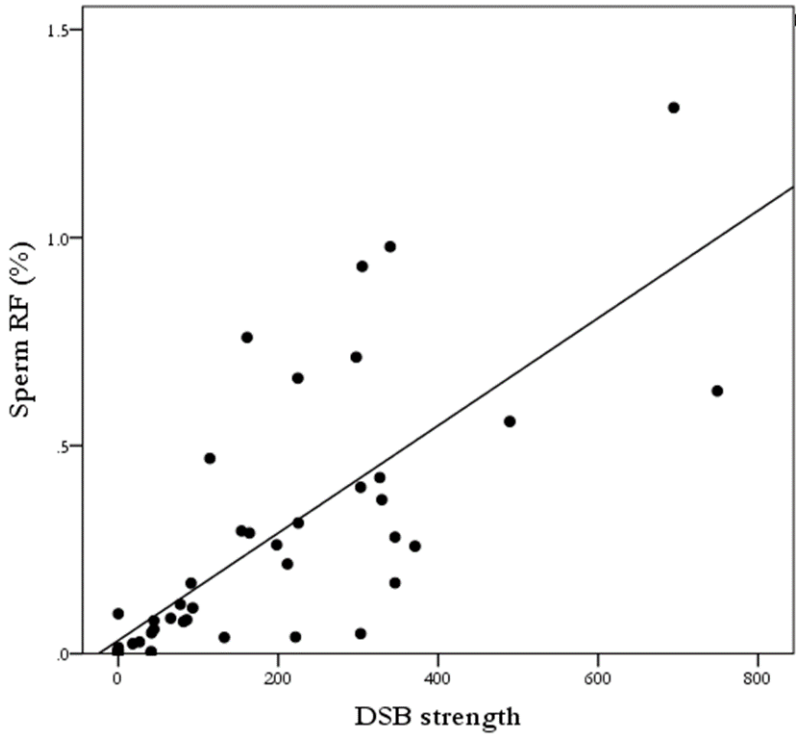


Figure 5.11 Scatter plot between all-hotspot sperm RFs and DSB strength. The graph shows a monotonic correlation of the datasets. The non-parametric Spearman's ρ shows the significant correlation of 0.82 ($p < 0.001$).

5.3 Discussion

Because of conservation of homology in PARs, ~3 Mb (4.6% of the Y chromosome [Otto *et al.*, 2011]) of collective length of both regions play an important role in the sex-chromosome pairing and genetic information exchange during male meiosis I. Obligatory recombination in the male PAR1 [Ross *et al.*, 2005] is also very crucial in maintaining homology between the two heterogametic sex chromosomes and therefore shows a recombination rate ~17-20× the autosome average and ~3-fold higher than that of the male PAR2 [Flaquer *et al.*, 2008, Hinch *et al.*, 2014]. Failure of the pairing may lead to paternal-caused aneuploidy and male infertility in humans [Burgoyne *et al.*, 2009, Gabriel-Robez *et al.*, 1990, Hall *et al.*, 2006, Mohandas *et al.*, 1992, Shi *et al.*, 2001] and also cause apoptosis in mouse [Faisal and Kauppi, 2016].

To succeed in pairing, the length of sequence homology is very important since the disruption of homology across the mouse PAR has been found to be associated with male infertility [Dumont, 2017]. PAR1 recombination activity has been found to peak at the sub-telomeric region and yet still maintain a high rate throughout the region until an abrupt drop in crossover activity occurs at PAB [Hinch *et al.*, 2014]. The recent discovery of the ePAR, that gives rise to variation in length of the human male PAR1 at the population level [Pratto *et al.*, 2014], opened up the possibility of directly examining the recombination behaviour of the 110-kb extension of the X-derived segment proximal to this canonical PAB.

The sperm recombination approaches have given invaluable insight into the recombination dynamics at the sub-kilobase scale, ranging from inter-individual differences in activity [Berg *et al.*, 2010, Berg *et al.*, 2011], through to haplotype-specific effects for a given man [Jeffreys and Neumann, 2002, Sarbajna *et al.*, 2012]. Using sperm assays one can gain a detailed insight into the dynamics of recombination when only one or a few men are available and one can efficiently detect even very low RFs, down to 0.0004% [Kauppi *et al.*, 2009].This approach has been used here to study two DSB hotspots flanking distally and proximally the X-derived portion of the ePAR has been studied in two available semen donors.

5.3.1 ePAR sperm crossovers amongst genome-wide recombination hotspots in relation to DSB maps

Both candidate recombination hotspots showed similar characters to those hotspots previously described, namely that crossovers are not randomly distributed but cluster in to a classical interval of 1-2 kb wide [Jeffreys *et al.*, 2004, Jeffreys *et al.*, 2001, Kauppi *et al.*, 2004, May *et al.*, 2002]. However, the distal hotspot X1 revealed variation in recombination rate

between the two men of around four-fold, while the two men exhibited comparable RFs with modest variability of at most 1.2-fold difference at the proximal hotspot X5. Variation in RFs among semen donors has been observed in several works and is also consistent within the range controlling for both *PRDM9* genotype and *cis*-effects influencing DSB initiation [Berg *et al.*, 2010] and is able to reach up to as much as 100-fold [Sarbajna *et al.*, 2012]. In contrast to the ePAR sperm studies reported here, the DSB strengths among five men studied by Pratto *et al.* showed a greater degree of variation, with a ~30-fold and ~7-fold range for the distal and proximal targeted hotspots, respectively [Pratto *et al.*, 2014]. In this work, the combined sperm crossover rates observed at the proximal hotspot were higher than those for distal cluster though at best ~2-fold difference compared with ~5-fold in the mean DSB strength. These differences may be a consequence of DSB repair, in that only the repair using inter-homologue exchange can be detected via the sperm assay while NCOs which do not encompass informative SNPs or sister-chromatid exchange (SCE) will go undetected.

Taking other sperm-based recombination hotspots into account, the two ePAR intervals show RFs above the median as noted for other pseudoautosomal recombination hotspots, though they show the lowest rates amongst the PAR hotspots. Also, they contribute concordant trends between RF and DSB strength showing stronger relative relationship than that between DSB and LD described in Chapter 4. This finding might suggest sex-specific differences in DSB initiation though the fine-scale study design can only limit to males. As mentioned by Pratto *et al.*, it is possible that DSB cluster could reflect activity that has yet to make an impact at a population level and also, LD-only hotspots could be the result of rare or lower-frequency *PRDM9* alleles not included in their study [Pratto *et al.*, 2014]. However, these two *de novo* recombination hotspots were entirely consistent, in terms of frequencies, distributions and characters, with classic hotspots shaping the landscape of the ePAR.

5.3.2 NCOs as observed in ePAR compared to other recombination hotspots

In this study NCO events were studied only at Hotspot X5, even then there was difficulty in unambiguously identifying such events since many recombinants implicated the terminal interval for both men. For the purposes of analysis, half of such events were arbitrarily assigned as COs and the other NCOs; this approach is likely to only significantly affect the estimated number for Man 20 as most of the recombinants were exclusively found in the terminal marker interval. Nonetheless, both men showed similar NCO:CO ratio of ~1:1.

CO and NCO have been noted to co-localise, indicating that they are different outcomes of the same initiating DSBs, though the proportions vary from hotspot to hotspot and can even differ between men [Jeffreys and May, 2004, Kauppi *et al.*, 2004, Sarbajna *et al.*, 2012] by NCO:CO ratios ranging from CO dominating NCO (<1:12) [Holloway *et al.*, 2006] to dominating by NCO (35:1) [Odenthal-Hesse *et al.*, 2014, Sarbajna *et al.*, 2012]. However, detection of NCO, in contrast to CO, highly depends on the distribution of informative markers and may account for some of the variation observed in many cases but not all.

Most of the NCO events were found to encompass a single informative marker usually close to the estimated centre of recombination consistent with previous evidences that the conversion co-localises with the centre of CO and its tract is usually short with the mean length of ~55 – 299 bp (the possible overall range: 1 - 2.9 kb) [Jeffreys and May, 2004, Sarbajna *et al.*, 2012].

5.3.3 Asymmetrical recombination and transmission bias

Asymmetrical transmission of markers into COs was found in the distal hotspot X1 where at the SNP marker 11.1 (rs1970797) the weak base-pair T was found to be significantly over-transmitted compared with the allelic strong base-pair C. The over-transmitted T allele has been hypothesized to suppress initiation of DSBs on this copy of the homologous sequence but act as the template for repair of induced breaks occurring preferentially on the opposite C-bearing haplotype [Jeffreys and Neumann, 2002, Jeffreys and Neumann, 2005, Jeffreys and Neumann, 2009, Sarbajna *et al.*, 2012, Webb *et al.*, 2008]. This phenomenon is not uncommon having been seen at many hotspots, with the markers showing TD being located very close, e.g. <100 bp, to the centre point of recombination [Jeffreys and Neumann, 2005, Webb *et al.*, 2008, Sarbajna *et al.*, 2012]. Moreover, in some cases these markers have been found to coincide with the Myers' motif, suggesting that they prevent PRDM9 from efficiently binding and triggering the events that lead to DSB induction by Spo11 [Sarbajna *et al.*, 2012]. In other cases, the markers do not obviously coincide with the degenerate Myers' motif, but it is noteworthy that this motif is associated with only ~40% of European LD hotspots [Myers *et al.*, 2008] and hotspots lacking the motif have been shown to be PRDM9-regulated [Berg *et al.*, 2010], indicating that we have yet to fully appreciate the events leading to DSB induction. Regardless of the finer details, it is has been observed that men who are homozygous for recombination-initiating alleles have high RFs and those that are homozygous for recombination-suppressing alleles have very low RFs [Jeffreys and Neumann, 2002, Jeffreys and Neumann, 2005, Sarbajna *et al.*, 2012, Webb *et al.*, 2008]. In

this study, TD was observed in the CO events at the marker 11.1 (rs197097) for Man 53 could not be directly explained by the PRDM9-blockade mechanism because the SNP is hundreds-bp away to the closest Myers' motifs either downstream (~315 bp) or upstream (~388 bp) and, at both motifs, both men carry identical homologous sequences though their RFs vary ~4-folds, however this does not rule out that this SNP might affect PRDM9 binding in *cis* at some point. Therefore, TD of COs at this hotspot might be only an indicator of recombination suppression or promotion to other men who have homozygosity of one of the SNP variants and should observe in a larger sample size. However, like some general hotspots previously observed [Jeffreys and Neumann, 2002, Jeffreys and Neumann, 2009], TD demonstrates that the hotspot is, as a form of meiotic drive, eventually doomed to its demise [Jeffreys and Neumann, 2009].

TD was also found exclusively in NCOs at the marker 97.4 (rs186282195) for the proximal hotspot X5 in Man 53, again this marker is some-hundreds base pairs away from any 7/8 or 8/8 matches to Myers' motif. Interestingly, over-transmission prefers the strong over the -weak base pair conversion as noted in other studies [Arbeithuber *et al.*, 2015]. Previously, TD confined to NCOs has been reported at two autosomal hotspots demonstrating differences in CO and NCO heteroduplex formation and/or mismatch repair: both observations would deserve noting that they revealed a significant GC bias [Arbeithuber *et al.*, 2015, Odenthal-Hesse *et al.*, 2014].

5.4 Conclusion

DSB maps derived from human males who were inferred not to carry ePAR have been shown in two sperm ePAR donors to result in *de novo* recombinant gametes. This indicates that the extension of ~110-kp of homology to the tips of the short arms of the X and Y chromosomes is sufficient to support genetic exchange between the sex chromosomes beyond the canonical PAR1 boundary. Both the distal and proximal DSB clusters examined here display the consistent recombination rates in a similar trend to the others of pseudoautosomal regions which are above the median of those in autosomes. The *de novo* recombinants isolated in this study also demonstrate classical hallmarks of recombination hotspots, in terms of distribution (hotspot width), transmission distortion and therefore meiotic drive, gene conversion tract length and GC-biased transmission distortion restricted to the NCO class of recombinant. The activities are concordant with observed DSB

strengths obtained from ChIP-seq ssDNA coverage as well as a strong correlation between other sperm-based hotspots and DSBs. Given the coincidence of historical LD-based hotspots derived from the population study to both DSB clusters, this sperm-based recombination study helps suggest that largely speaking this region is similarly primed for crossover in both male and female germlines, even though sex-specific differences may also exist.

Chapter 6 Sequencing of the X-portion of ePAR to infer past recombination history

6.1 Introduction

Indirect evidence was presented that ePAR might actually function pseudoautosomally since there were at least two haplotypes sequenced by PacBio MPS among the Y-Hg I2a men in the original study [Mensah *et al.*, 2014]. Since these ePAR-carrying men were predicted to be of the same Y-Hg I2a lineage, and therefore most likely inherited the ePAR from a common ancestor, the polymorphism found in the translocated region was deemed to have resulted from recombination between X and the ePAR rather than mutation accumulation since they differed by twelve SNP variants all of which are located on X chromosomes [Mensah *et al.*, 2014]. The region analysed however was <5% of the total length the the ePAR.

6.1.1 Homology of the Xp portion in ePAR on Yq

Most the human Xp consists of an evolutionary stratum called XAR as it was added from an autosome and has been evolving around 130 – 80 Mya since the divergence of eutherians and marsupials until before the radiation of eutherians [Graves *et al.*, 1998]. In parallel, the human Y chromosome effectively gained the homologous chromosome to the XAR making the YAR, though this later became largely eroded [Ross *et al.*, 2005]. At present, the remaining XAR-YAR homology, apart from PAR1, is present on the X chromosome as a ~6-Mb continuous region just proximal to PABX, while on the Y chromosome it is considerably more fragmented and rearranged consisting on both on the proximal Yq and the mid part of Yp as shown in Figure 6.1 a [Ross *et al.*, 2005]. The translocated X-portion in ePAR, equivalent to the Block 1 of XAR-YAR homology, is located at the very proximal part of Yq as shown in Figure 6.1 b has been found to have ≥70% of sequence similarity [Ross *et al.*, 2005].

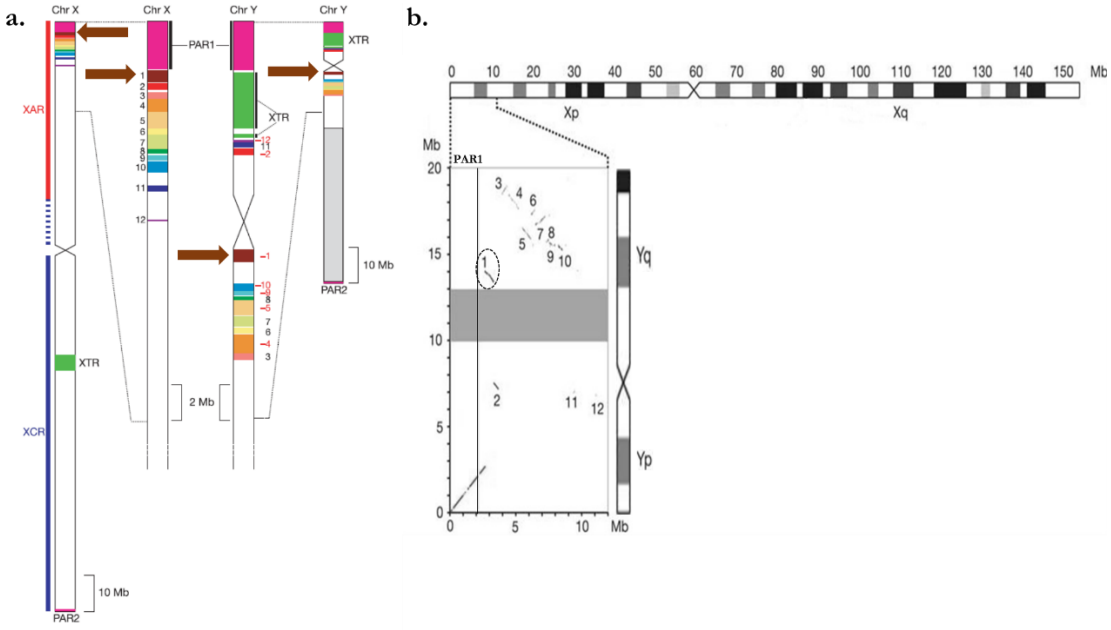


Figure 6.1 **Major homologous blocks between human sex chromosomes and focusing on X-portion of ePAR. a)** Chromosome X and Y in a full scale are shown by the left- and right-most schematics with expanded sections shown in the centre. The coloured blocks show corresponding homologies indicated with number; minus means an inversion of sequence. The X-portion of ePAR is a part of the brown Block 1 on Xp (indicated with brown arrow) which is homologous to the brown Block -1 on the proximal Yq (indicated with the same coloured arrow). The other Blocks (2-12) show fragmentation, shuffling and rearrangement of XAR homology on the YAR. **b)** Expansion of BLASTN alignment between the distal 0-12 Mb of Xp and the entire Y chromosome shows the major homologous blocks between XAR and YAR. The homology Block 1 (encircled with dotted eclipse), demonstrates a cluster of sequence similarity (≥70%) of ~1 Mb between the Xp just proximal to PABX and the proximal Yq. (Adapted from [Ross *et al.*, 2005])

6.1.2 High genetic diversity in PAR1 resulted from recombination

PARs have an larger effective population size than the X-linked (non-PAR) sequence because there are two copies functioning in all individuals for the former while the latter has two copies in females and only one in males: this is also accompanied by greater frequency of recombination, at least in the male germline, which eradicates the effects of background selection and genetic hitchhiking due to linked haplotypes and recombination itself increases

the mutation rates, therefore PARs are found to contain higher genetic diversity than the strictly X-linked region [Cotter *et al.*, 2016]. The ~3-times higher genetic diversity in PAR1 than PAR2 observed recently [Cotter *et al.*, 2016] is consistent with the obligatory recombination behaviour in the male PAR1 [Flaquer *et al.*, 2008, Kauppi *et al.*, 2012, Ross *et al.*, 2005] with ~4-times higher average recombination rate than that of PAR2 [Filatov and Gerrard, 2003, Lien *et al.*, 2000]. As a result, genetic diversity between X and Y chromosome should be expected to abruptly drop beyond the PAB, instead, it was noted recently that it gradually declined and appeared to be high in some regions with high homology such as between XTR and YTR [Cotter *et al.*, 2016]. Thus it has been hypothesized that gene conversion may function in these regions [Cotter *et al.*, 2016, Trombetta *et al.*, 2014].

The observations above accompanying with the fact in Chapter 5 that the recombination hotspots in the X-portion of ePAR are active with similar rates to those previously observed in both PARs, prompted the question how does the entire ePAR recombination rate compare with the autosome average. To address this, resequencing the whole region was carried out using MPS to explore the genetic diversity of all haplotypes and infer recombination behaviour of the full ePAR interval.

6.1.3 Y-Hgs identified in the first-reported ePAR males need confirmation

During the end of the last century until the beginning of this millennium, the analysis of paternal lineage using biallelic or binary polymorphic markers, i.e. SNPs, on the non-recombining region of the Y chromosome experienced considerable progress and as a result a system of grouping Y-chromosome lineages into so called 'haplogroups' was developed [Jobling *et al.*, 1997, Underhill *et al.*, 1997]. With the progress of technology, a vast number of Y-linked SNPs were increasingly found and these improved the resolution of the human Y-phylogenetic tree. Consequently, a hierarchical nomenclature system was introduced which allowed the global picture of population distribution and male migration since Out-of-Africa to be appreciated [Y Chromosome Consortium, 2002]. This system is widely used in a whole host of biogeographic studies including population, medical and evolutionary genetic studies, as well as in forensic applications [Jobling and Tyler-Smith, 2003, Jobling and Tyler-Smith, 2017, Kayser, 2017, Karafet *et al.*, 2008]. The Y-Hg tree topology has been repeatedly revised as more and more mutations were discovered with the tree structure becoming more complex, however the nomenclature rule is still conserved [Karafet *et al.*, 2008]. In brief, the major clades are named with the capital letters beginning with the deepest

root to the most recent clade (A-T) and followed by the alternated numbers and alphabets called the alphanumerical system for further branching subclades such as I2a1 [Y Chromosome Consortium, 2002, Karafet *et al.*, 2008]. As each branch is defined by a SNP, the long alphanumerical name can be replaced by only the main-clade name in one capital letter then hyphen and the name of terminal mutation, for example, Y-Hg I2a1 is represented by the mutant SNP L460, so it can be written as I-L460 [Y Chromosome Consortium, 2002] or at most with the first three letters and the terminal SNP as I2a-L460.

The binary marker system indicating the human Y-Hgs is more robust than using the Y-STR haplotype because SNPs are more stable than STRs as their mean mutation rates are much lower, 2.2×10^{-8} per base per 25 years for SNPs [Jobling and Tyler-Smith, 2017] *cf.* 6.9×10^{-4} per locus per 25 years for STRs excluding rapid-mutating loci [Ballantyne *et al.*, 2012, Zhivotovsky *et al.*, 2004], so it can estimate Time to the Most Recent Common Ancestor (TMRCA) more precisely for the deep-rooted clades and can also increase the precision of lineage discrimination [Jobling and Tyler-Smith, 2003]. However, there is an abundance of Hg-related SNPs that are not robust to type. In practice, a set of Y-STR profiles is often typed and used to predicted the Y-Hg using a widely-accepted approach which calculates a goodness-of-fit score from the allele frequency of each marker and performing the Bayesian probability by comparing with the scores from the model Y-STR profiles of each haplogroup [Athey, 2005, Athey, 2006]. To confirm the terminal SNPs of sub-Hgs, SNP typing by various methods should be subsequently performed, especially in cases where it is necessary to distinguish lineages more precisely.

6.2 Results

6.2.1 Confirmation of the Y-Hgs of ePAR males and estimation of TMRCA

In this chapter, five of the originally-reported ePAR families [Mensah *et al.*, 2014] supplied by Dr. Larmuseau mentioned in Chapter 3, plus five additional ePAR cases, (including one within a CEPH pedigree) who were predicted by 23 Y-STR profiles to have the Y-Hg I2a-L233 or another close lineage were SNP-typed for their terminal sub-Hg marker to aid subsequent inference of past recombination events. Also, three European DNAs who do not carry ePAR but were predicted to have the Y-Hg I2a-P37.2 were included. As in the Figure 3.4 in Chapter 3 showing the hierarchical tree of the Y-Hg I2a, to get to the terminal

marker L233, nine amplicons covering the SNP set were designed and a 9-plex PCR followed by SNaPshot® typing was developed. SNaPshot® is based on a single fluorescent ddNTP primer-extension principle with the resulting products run on a capillary electrophoresis machine as described in Chapter 2. The electropherogram of the developed I2a assay is shown in Figure 6.2 below.

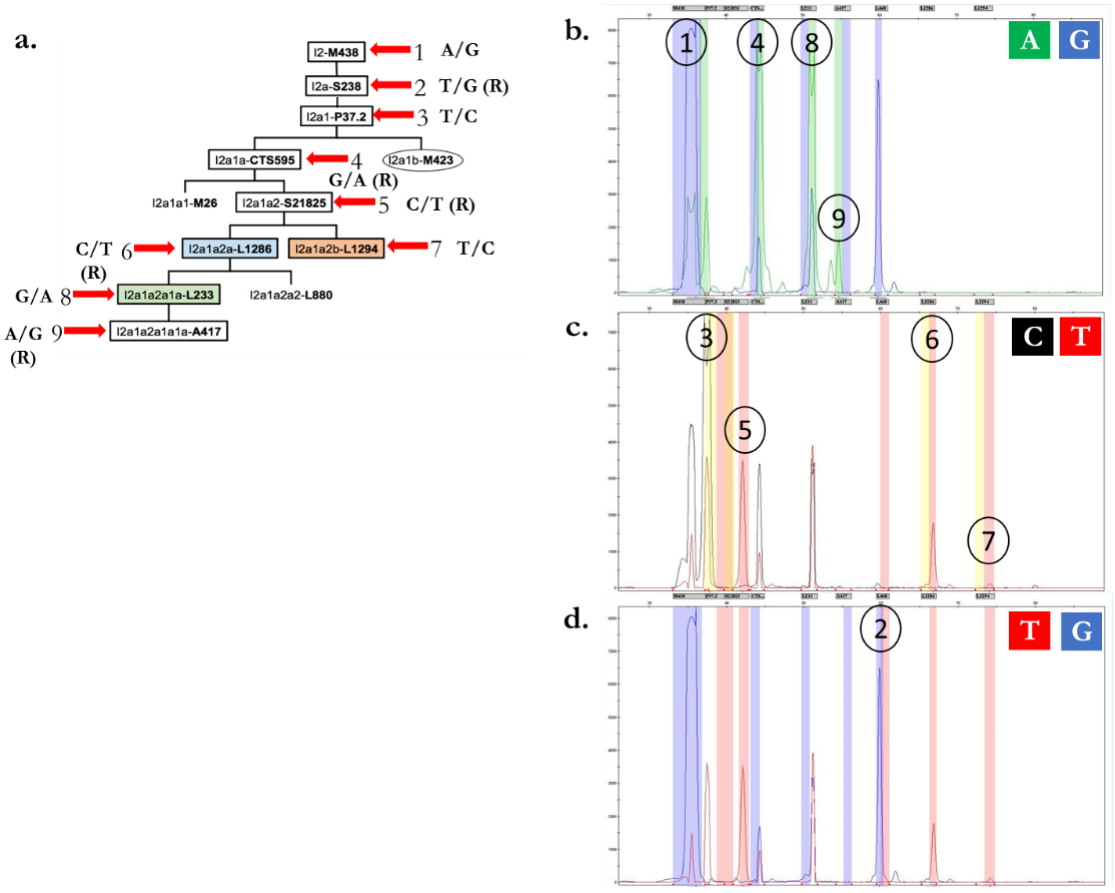


Figure 6.2 **SNP typing to identify the terminal Y-Hg according to hierarchical I2a tree. a)** The Y-Hg I2a tree as presented in Figure 3.4 but with numbers assigned to each SNP in the assay which also corresponding to the electropherograms in panels **b**, **c** and **d**. Ancestral and derived allele status of the SNPs are presented with the number and the allele stages of some SNPs using SNaPshot primers on reverse strands are indicated with (R). Panel **b** to **d** show the screenshot of electropherograms analysed by GeneMapper® Software 3.3 in three dye combinations, green/blue in **b**, black/red in **c** and red/blue in **d**; each colour represents a specific base as indicated (green = A, blue = G, black = C and red = T) with the vertical shading representing bins being set to approximately mark the position of each marker. Because

artefactual peaks may occur and obscure the expected peaks, **b - d** display only two-dye combination for each panel that correspond to the allelic variants of each marker as indicated. In cases where both alleles appear for a given male, status can be inferred from the next marker down the line because Y-hg lineages are determined by hierarchy; for example, the marker 4 is inferred to be A (derived) since the marker 5 is clearly T (derived).

The Y-Hg results of all tested individuals are presented in Table 6.1. Although some individuals had missing markers their terminal SNPs were typable and in several cases it was possible to also compare results with their paternal siblings and/or father.

Table 6.1 SNaPshot® assay determining I2a sub-Hg in ePAR individuals

	M348	P37.2	S21825*	CTS595*	L233	A417*	S238*	L1286*	L1294	Y-Hg
NA12146	G	C	T	A	A	A	G	T	T	L233
Man 20	G	C	T	A	A	A	G	T	T	L233
Man 53	G	C	T	A	A	A	G	T	T	L233
333	G	C	T	A	A	A	G	T	T	L233
6689_01	-	C	-	A	A	-	-	-	-	L233
P1	G	C	T	A	G	A	G	T	T	L1286[a]
F1	G	C	T	A	G	A	G	T	T	L1286[a]
B1	G	C	T	A	G	A	G	T	T	L1286[a]
P2	G	C	T	A	A	A	G	T	T	L233
F2	G	C	T	A	A	A	G	T	T	L233
P5	G	C	T	A	G	-	G	C	C	L1294
F5	G	C	T	A	G	A	G	C	C	L1294
P3	G	C	T	A	A	A	G	T	T	L233
F3	G	C	T	A	A	A	G	T	T	L233
P6	G	C	T	A	A	A	G	T	T	L233
F6	G	C	T	A	A	A	G	T	T	L233

* markers in reverse strands; [a] Terminal marker ending at L1286 is likely derived to Hg L880 line.

Although the first ePAR report stated that this event is likely to recur, the recombinant junction diversity data of the Hg-I2a ePAR carriers, namely Junction 1 and Junction 2, could be interpreted as representing a single crossover breakpoint, in that Junction 2 might simply be the result of subsequent gene conversion [Mensah *et al.*, 2014]. Indeed, the ePAR carriers from Y-Hg I2a do appear to have all acquired their ePAR once in time based on the Y-Hg SNP typing results presented here; although there are three terminal haplogroups, L233, L1286 and L1294, bifurcating to different sub-branches, they are all rooted from the common node, I2a-S21825 as in Figure 6.2 a.

Given that all the males above are implied to have ePAR originated from the common ancestor, the TMRCA could be calculated. The common ancestor node in this case is very young (< 5 kya), so estimation by using Y-STR profiles should be more practical than using Y-SNPs [Balanovsky, 2017]. The most popular approach for using this kind of markers is the average squared difference (ASD) [Goldstein *et al.*, 1995a] because it has been thought to be better than the other two approaches; firstly, the rho estimator (ρ) [Forster *et al.*, 1996] which requires to generate a haplotype network and choose the most common node to be a model haplotype [Balanovsky, 2017]; and secondly, the Bayesian approach, e.g. BATWING Software [Wilson *et al.*, 2003], which requires an appropriate past demographic model that is usually unknown [Wang *et al.*, 2014, Balanovsky, 2017]. The ASD is calculated using the whole data set assuming the modal allele length for each STR to be that of the founder haplotype (ASD_0) the difference in number of repeats of each given STR locus from those of the modal alleles is used to calculate the average. Furthermore, since it is the squared value, the number is large which has the advantage of overcoming the effect of back-mutations compared with ρ [Balanovsky, 2017].

In this study, the ASD approach was selected in conjunction with using the generation time 31 years for males [Fenner, 2005]. Given that the Y-Hg I-L233, I-L1294 and I-L1286 ePAR carriers were inherited from the common ancestor, I-L21825, the ten haplotypes were estimated, using the published mutation rates of Y-STRs [Goldstein *et al.*, 1995a, Goldstein *et al.*, 1995b], to have TMRCA of 3,877 ± 779 yrs, equivalent to 125 male generations.

6.2.2 Manipulation of ePAR resequencing data generating genotype calls and exploration of the homologous translocated X to Yp topology

MPS resequencing using the Ion Torrent™ platform was performed on a total of 20 individuals. To aid with subsequent phasing of alleles, family members were included; there were five paternally-lineal families appearing in Mensah *et al.* [Mensah *et al.*, 2014] (one father-son-brother trio and four farther-son duos: 11 individuals) of Y-Hg I2a and one father-son family of Y-Hg R1b making a total of 13 Belgian and French ePAR individuals together with one CEPH family consisting of an ePAR father (Y-Hg I2a), a mother and a daughter. The remaining four were singleton ePAR individuals from Leicester collections including two semen donors and all of them were Y-Hg I2a.

Thirty-four overlapping amplicons ranging from 1.2 – 6.5 kb were designed to cover the entire 110-kb translocated portion of the X chromosome and subsequently sequenced in two runs. Initial data analysis was performed by the Torrent Suite™ Software 5.0.2. Raw data was processed through the pipeline in the program including quality control, removal of primer and barcoding sequences, and was mapped to the hg19-version of the human reference sequence of the X chromosome corresponding to the translocated X region (X:2697499-2808535). The final format of each sequence data was a bam file. The bam files and their index files were visualized via IGV Software 2.3 [Thorvaldsdóttir *et al.*, 2013]. Together with the statistics from Torrent Suite™ Software, the mean read depth was 300× and the mean number of mapped reads was 308,842 (range: 117,148 – 753,664). However, ~8-19% of the reads per man were found to be unaligned (Figure 6.3 a). The cause of unalignment was due to the same amplicon in every sample which was unintentionally amplified from the proximal of Yq and it was confirmed by counter-aligning against the proximal Yq reference sequence and found matched in one region of ~75% the whole length of the amplicon (Figure 6.3 b).

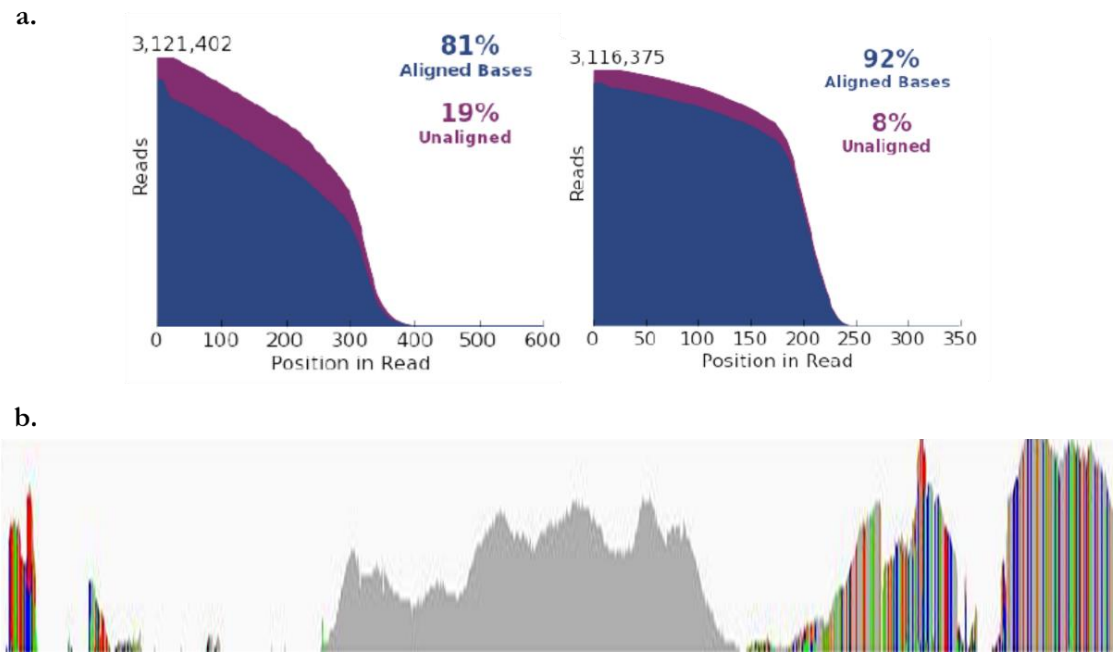


Figure 6.3 **Ion Torrent™ sequencing report of alignment. a)** Two sequencing curves generated from Torrent Suite™ Software 5.0.2 including the problematic amplicon shows high percent of unaligned reads (magenta area) compared to the aligned (blue are). **b)** Screenshot from IGV Software 2.3 [Thorvaldsdóttir *et al.*, 2013] showing the alignment of the translocated X of

ePAR against the proximal Yq displays the well-aligned region (grey patch) of approximately 4.5 kb to the reference sequence located around Y: 14533425-14537470 (hg19): The well-aligned sequence corresponds to the unaligned reads against the X reference sequence from the mis-primed amplicon.

To explore the homology structure of the translocated X portion and the proximal Yq, the human reference sequence of the X-portion of ePAR was used as query in BLAT on the UCSC Genome Browser website [UCSC Genome Browser]. The results showed that collectively the X-derived region of the ePAR (hg19) X:2699736-2808547 (108,812 bp) aligned against to the minus strand of Yq region coordinates Y:14507120-14569262 (62,143 bp) in multiple blocks with duplication and degeneration as shown in Figure 6.4 and Table 6.2 below. The problematic amplicon (X:2742080-2748092, hg19) is a part of the lower Block 1 demonstrated below.

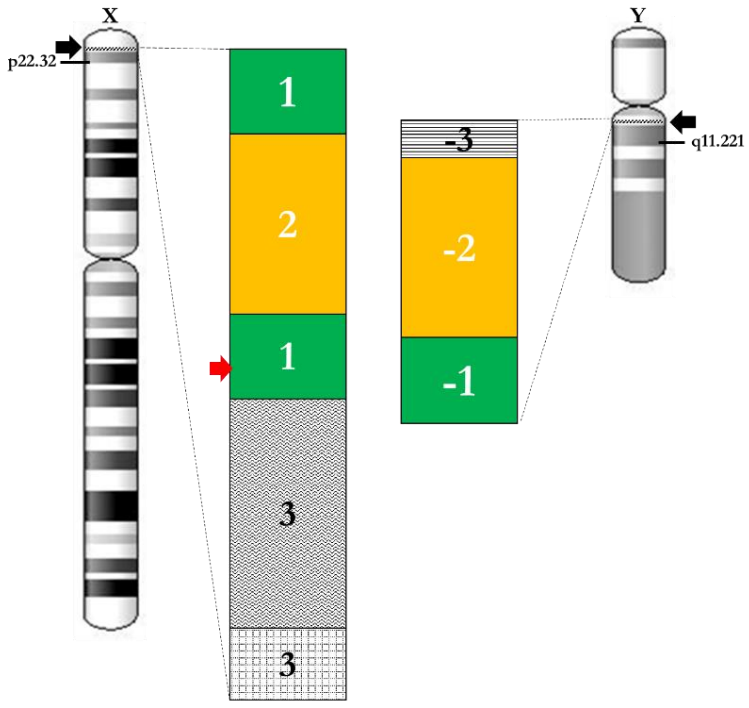


Figure 6.4 **Homology blocks between translocated X and proximal Yq.** The corresponding homologues on each sex chromosome ideogram on both sides are indicated in oblique line strips indicated by black arrows. Both regions correspond to the Block 1 in Figure 6.1 above. Details of the

homologies are demonstrated in coloured blocks in between the ideograms where those of the translocated X line on the forward strand while those of the Yq line inversely on the minus strand and indicated with minus signs. The homology blocks of the X reveal duplication of Block 1 and 3. The mis-primed amplicon is located in the lower Block 1 indicated by the red arrow. For Block 3, the homology sequence in Yq discontinue and looks less homologous because it is alternately inserted by non-homologous sequences in patches. The homology patches of Block 3 on the translocated X are disrupted by non-homologous sequences with varied lengths. The total lengths including inserted sequences of the translocated X homologous is ~2.2 and 6.7× of the length of Yq Block 3. Pictures of both sex chromosome ideograms are taken from https://www.ncbi.nlm.nih.gov/genome/tools/gdp/.

Table 6.2 Coordinates of homology blocks between translocated X and proximal Yq (hg19)

Translocated X	Block	Yq	Similarity	Remarks
X:2699736-2739981 (40,246 bp)	1 + 2	Y:14525425-14539580 (14,156 bp) + Y:14539581-14569262 (29,682 bp)	94.8%	-
X:2739982-2754428 (14,447 bp)	1	Y:14525425-14539580 (14,156 bp)	93.7%	The problematic amplicon lies on this region
X:2754429-2794676 (40,248 bp)	3	Y:1457120-14513353 (6,234 bp)	95.0%	Alternatively inserted by non-homologous sequences with varied length
X:2794677-2808547 (13,871 bp)	3	Y:1457120-14513353 (6,234 bp)	95.0%	

Unmapped sequences were removed before generating the final base calls. These calls were measured using the base quality score or Phred score (Q), named after the computer program, which is defined as the logarithmically related probability of base calling error (p) and can be calculated as [Ewing and Green, 1998]:

$$Q = -10\, log_{10}\, p$$

Therefore, Q = 20, widely written as Q20, equates the error calling = 1% (1 in 100), in the other words, this is equal to 99% of accuracy. Generally, this level is widely accepted as a rule of thumb because it is nearly equivalent to the error rates in Sanger-sequencing which

usually contains $10 \leq Q \leq 20$ [Ewing *et al.*, 1998]. In this study, the mean number of Q20 bases called per individual across two runs was 69,342,818 (range: 24,051,513 – 134,452,533). Summary of statistics for each individual are shown in Appendix. Thus, downstream variant calling was performed with the SAMtools software package 1.3.2 using the bam files and filtering out the reads with a $Q < 20$ and a mapping quality score (mapQ), or a probability of error mapping calculated in the same way as Q, lower than 50.

The genotype calls from both Ion Torrent runs of each individual were merged and indels and the variants located in tandemly repetitive sequences were removed. The diploid genotypes were then validated by comparing the genotype calls for the Belgian and French samples with those established by Mensah *et al.* [Mensah *et al.*, 2014] over the region that had previously been subjected to PacBio sequencing (<5% of the total ePAR). Both datasets were perfectly concordant at the fourteen variants that map uniquely to the translocated X across the thirteen ePAR males. Another five additional sites previously reported to map to SINEs could not be evaluated because they are not present in the hg19 X-chromosome reference sequence used to align the reads. Validation also included direct comparison of the Ion Torrent data with data from the 1000 Genomes Project [1000 Genomes Project Consortium, 2012] for the daughter (NA10847) of the ePAR individual from the CEPH pedigree 1334. Concordance was observed >99.6%.

Haplotypes were determined using PHASE Software 2.1.1 [Stephens and Donnelly, 2003, Stephens *et al.*, 2001] by taking all individual genotypes to phase with the genotypes or haplotypes for both female-only and male-only groups of the CEU (Utah Residents with Northern and Western European Ancestry or CEPH) and GBR (British in England and Scotland) populations from the 1000 Genome Projects Phase III data sets [1000 GP III FTP] and assorted with various combinations. All the ePAR genotypes were phased at least 5 times in different assorted groupings according to PHASE Software instruction. Where appropriate, the family relationships were used to determine which haplotype most likely corresponded to the ePAR. However, two of the ePAR men were without first-degree relatives to compare with but shared the same rare British family name (~0.0027% of population in 1996 electoral registration data) indicative of shared ancestry [King *et al.*, 2006]; genealogical records suggest a likely common ancestor more than five generations ago and Y-STR profiles also indicate the close relatedness in their paternal lines. Furthermore, one of them contains homozygous genotype of >80% of the markers which helps resolve ambiguous loci. This was taken into account when assigning the ePAR haplotypes for the whole data set too.

Finally, ambiguous heterozygous variants which remained unresolved were checked by looking through the raw read haplotypes via IGV Software. The final unresolved variants (mean = 7.14 ± 4.40%) were then resolved parsimoniously by comparing with the resolved variants and homozygous genotypes mainly among the ePAR individuals especially from the same Y-Hg.

Perfect concordance was observed in ePAR assignment for the eleven cases of the Y-Hg I2a where deduction was possible in the previous study [Mensah *et al.*, 2014]. Finally, the predicted and empirically derived haplotypes for each of the two semen donors were compared over each of the recombination assay intervals which had been phased. Man 20 showed complete concordance over the ten informative sites in the distal assay region X1, and his haplotypes matched at six of the seven such sites in the proximal region X5. The corresponding data for Man 53 were 6/8 and 6/7 informative sites, respectively. These lower concordances might be expected as there were no first-degree relatives available to help resolve the phasing in either of these instances. The final set of markers was made by taking CEU and GBR SNP-haplotype of the phase-known male X-chromosomes and selecting only the shared SNPs. In aggregation, the number of excluded sequence was ~9kb (8%) in average of the whole length.

6.2.3 Inferring past recombination events throughout ePAR

Of the total twenty individuals who were sequenced for the entire 110-kb X-translocated part of ePAR, there was one haplotype in the Y-Hg R1b lineage from Mensah *et al.* (P2/F2) [Mensah *et al.*, 2014] and the rest were in the Y-Hg I2a. However, at the beginning of this section, the Hg-R1b ePAR haplotype (P2/F2) from Mensah *et al.* [Mensah *et al.*, 2014] was also included to visualize the whole picture of ePAR in selected populations (CEU and GBR) [1000 Genomes Project Consortium, 2012] that contain wide variety of Y haplogroups.

To gain a comprehensive insight of the recombination history of ePAR, the final phased haplotypes as described above exclusively from the Y-Hg I2a males which composed of five Belgian and French lineages from Mensah *et al.* (2014) (B1/P1/F1, P3/F3, P4/F4, P5/F5 and P6/F6) alongside one ePAR male from the CEPH pedigree (NA12146) and the other four ePAR singletons of British origin including two semen donors, named as in Table 6.1 (Man 20, Man 53, 333 and 6889_01), were compared. To aid the interpretation, only haplotype blocks outside of the DSB-clusters were considered, so the signature of CO could most easily be detected by new combinations of pre-existing and well-defined haplotype

blocks. This generated a core set of 213 SNPs falling into nine blocks which ranged from the length of 558 to 16,143 bp as shown in Figure 6.5. These SNPs overlapped with those from CEU and GBR of the 1000 Genome Project Phase III, with the reasoning that the CEU X chromosomes were most relevant for understanding the history of ePARs identified from the same geographical region (Figure 6.5). A comparison of the ePAR haplotypes in the X-translocated region with those of CEU and GBR men is shown in Table 6.3.

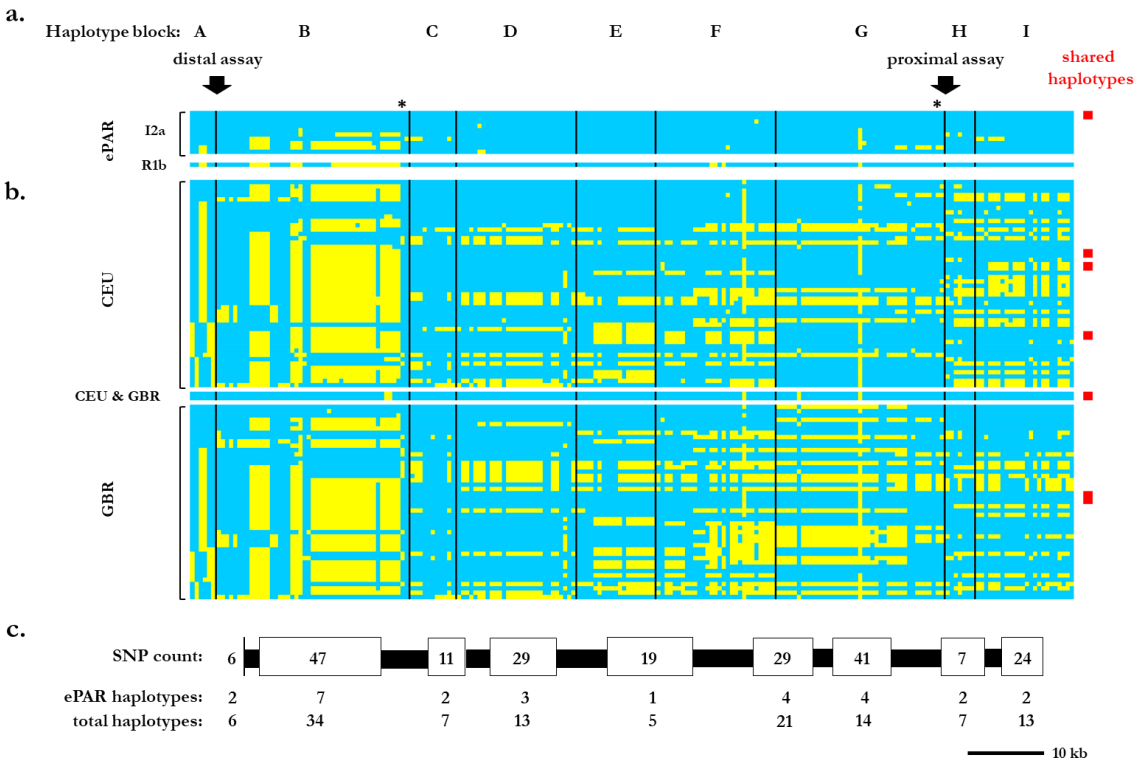


Figure 6.5 **Comparison of the inferred ePAR haplotypes with phase-known haplotypes from the corresponding region of the translocated X chromosome. a)** SNP haplotypes from each of the eleven unrelated sampled ePARs (ten from the Y-Hg I2a and one from the R1b) are displayed in rows and clustered based on the distal haplotype block. Individual 6889_01 is presented at the top with all his alleles colour-coded in blue while yellow denotes the alternative SNP alleles not carried by this man. Black vertical lines correspond to the relative location of mapped *PRDM9* A and C DSB clusters [Mensah *et al.*, 2014]. Two asterisks mark the instances that an A and C cluster lie in very close proximity. Arrows indicate the distal X1 and proximal X5 regions for sperm recombination assays. The top red box at the most right-hand side indicates the second ePAR haplotype which is identical

to that of 6899_01. In total, nine of the ten I2a ePAR haplotypes are unique to this dataset. **b)** showing phase-known X haplotypes from the 1000 Genome Project Phase III [1000 Genomes Project Consortium, 2012] for 49 CEU and 46 GBR males. One shared haplotype between two data sets is displayed separately in between and indicated by the red box. In addition, three pairs of identical X haplotypes were noted among the CEU and one haplotype was found to be carried by three different GBR men (red boxes). In the other words, 42 of the 46 CEU and 43 of the 44 GBR haplotypes are unique. None of these CEU and GBR X haplotypes matches any of the ePAR haplotypes. **c)** Relative scaling of the regions depicted together with summary count of the number of SNPs, number of ePAR haplotypes and the corresponding total number of haplotypes observed among the ePAR, CEU and GBR datasets per SNP block.

Table 6.3 Summary of statistics for a comparison of ePAR haplotype structures with phase-known X chromosomes

Block	**A**	**B**	**C**	**D**	**E**	**F**	**G**	**H**	**I**
No. SNPs:	6	47	11	29	19	29	41	7	24
hg19 (chr X) start:	2699645	2702047	2724389	2732634	2748292	2767637	2778322	2792662	2800624
hg19 (chr X) end:	2700202	2718189	2729346	2741406	2759615	2775601	2786038	2798480	2806196
No. different haplotypes amongst:									
I2a ePAR (n=10)	2	7	2	3	1	4	4	2	2
R1b ePAR (n=1)	1	1	1	1	1	1	1	1	1
CEU (n=49)	5	20	5	9	4	13	11	6	9
GBR (n=46)	6	17	5	6	4	11	9	5	7
No. unique haplotypes amongst:									
I2a ePAR (n=10)	0	4	0	2	0	2	1	0	1
R1b ePAR (n=1)	0	1	0	0	0	1	0	0	0
CEU (n=49)	0	12	2	5	0	6	4	3	5
GBR (n=46)	1	9	2	2	1	5	2	5	3
% I2a ePAR shared with:									
CEU	100	54.5	100	81.8	100	72.7	90.9	100	90.9
GBR	100	54.5	100	81.8	100	72.7	90.9	100	90.9
% CEU haplotypes shared with:									
all ePARs	69.4	49.0	89.8	55.1	67.3	16.3	69.4	55.1	46.9
GBR	100	71.4	91.8	89.8	100	81.6	91.8	83.7	79.6
% GBR haplotypes shared with:									
all ePARs	73.9	45.7	89.1	45.7	60.9	17.4	50.0	80.4	63.0
CEU	97.8	76.1	91.3	93.5	97.8	84.8	84.8	91.3	82.6

Adjacent blocks of markers are separated by clusters of mapped meiotic DSBs as reported in [Pratto *et al.*, 2014]

Of those ten unrelated I2a ePAR lineages, only two were found to have the same haplotype which was designated as the consensus. The remaining eight were different by up to four of

the nine blocks (mode and median = 2). Changes from the consensus were found to range from 1 – 29 SNP(s) per block (mode = 1 and median = 2). In comparison with phased-established X chromosomes from either CEU or GBR males, no complete match was detected, though matches at the level of individual blocks were observed. The results are presented in Figure 6.6.

Figure 6.6 **Simple interpretation of the I2a ePARs. a)** Schematic of consensus X-derived portion of the ePAR carried by individual 6889_01 and Man 20. Green boxes with black outlines display the shared haplotype at each of the nine blocks of SNPs whereas the intervening black bars coincide with mapped *PRDM9* A and C DSB clusters [Mensah *et al.*, 2014]. The width of each box/bar is in proportion to its length. The black triangle to the left points towards the canonical 2.7-Mb PAR1 and ultimately the Yp telomere; the beginning of the I2a Y-specific portion is shown to the right. The frequency of phase-known X haplotypes among 95 CEU plus GBR males [1000 Genomes Project Consortium, 2012] that match the consensus ePAR haplotype for each SNP block are shown in black and the frequencies of singleton haplotypes among the same population are shown in green. **b)** The

remaining eight I2a ePARs, by assumption that they are the result of a single crossover between the consensus and an incoming X-linked haplotype depicted by yellow boxes where the crossover interval is presented with a purple cross. The yellow boxes with red outlines demonstrate the haplotype blocks that differ from the consensus with the number of base pair changes shown in red. Black numbers beneath boxes indicate the observed frequencies of the non-consensus haplotype among the 95 phase-known X haplotypes from the aggregation of CEU and GBR men. Total SNP counts per block are shown in italic number at the bottom.

6.3 Discussion

6.3.1 Common root of sub-Hgs I2a amongst ePAR carriers to infer the recombination history of the X-translocated region

The diversity of the I2a ePAR men shown in Figure 6.6 would be simplest assumed that each unique haplotype is a product of a single different CO event occurring between the X portion of ePAR and the homologous region of the X chromosome. By looking for matches to those phase-known X- haplotypes of CEU and GBR, none was identified while only five compound haplotypes of CEU/GBR were found more than once which is consistent with the previous report of high diversity in this region [Cotter *et al.*, 2016] so it is possible that single exchanges including unsampled X chromosomes could account for the observed diversity of ePAR haplotypes.

The first-reported ePAR carriers in I2a Y-Hg showed their recombination junction diversity in two types [Mensah *et al.*, 2014] but was unclear what the actual mechanism was and could be most simply assumed that they originated from the same event before one junction type underwent further recombination by gene conversion. Also their exact sub-Hgs have not been identified until this study that showed that the Junction 1 occupied by most of them were found in the Hg I-L233 and I-L1286 while the other Hg I-L1294 carried Junction 2. As three Hgs are rooted from the same node, I2a-L21825, the estimated TMRCA using the Y-STR profiles of ten sampled I2a ePAR lineages were 125 generations as presented above. Of the ten I2a ePAR lineages in conjunction with the assumption of the diversity of eight unique sampled I2a-ePAR X-portions resulted from single CO events, it was therefore able

to calculate a minimum recombination rate through the 110-kb region of 0.64% (i.e. $\frac{8}{125 \times 10}$). The calculation schematic is displayed in Figure 6.7 below.

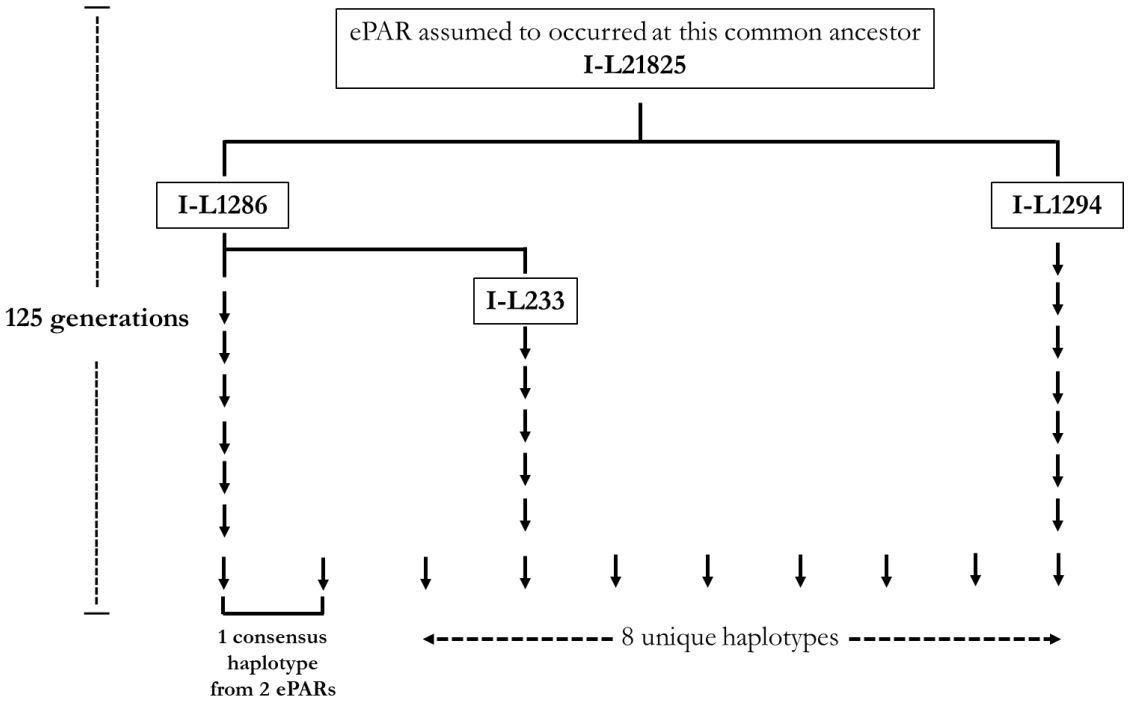


Figure 6.7 Schematic estimation of minimum recombination rate across X-portion of ePAR. Given that the consensus compound haplotype does not recombine and each of eight other haplotypes underwent a single recombination, calculation of a minimum recombination rate across this region is taken 8 CO lineages out of 10 lineages of 125 generations.

The recombination rate is likely to be underestimated of the actual rate for two reasons: first, not all ten sequenced lineages radiated in one line per generation directly form the common ancestor that there are more events potentially identified. Second, there is no way to definitively estimate multiple recombination events from these data. Interestingly, the variation from the consensus extends close to the proximal boundary, so it is definitely possible that these ePARs have experienced additional distal recombination events.

6.3.2 Estimated recombination rate in comparison to the genome average and the PARs

The estimated minimum recombination rate of the entire 110-kb X-portion of ePAR can be converted into ~5.8 cM/Mb that suggests to be higher of at least six times the genome average, compared with that in male recombination rate of at most 0.9 cM/Mb [Yu *et al.*, 2001]. This is compatible with the average RFs observed in sperm CO data for the two targeted intervals compared with the median of autosomes. The canonical PAR1 supports a male CO rate 17× of the genome average and 4× of the next most recombinogenic region of comparable physical length [Hinch *et al.*, 2014], therefore the data supports that ePAR is an active recombination hotspot in male meiosis.

6.4 Conclusion

A cross-sectional observation from sequenced males assumed to have inherited the ePAR from the same ancestral root reveals a wider dynamic picture of recombination events across the region. The data demonstrates high diversity of ePAR comparable to the corresponding strictly X-linked region, and the minimum recombination rate of the entire ePAR is higher than the genome-average suggesting that the recombination behaviour of the ePAR tends to be similar to the hallmark of the canonical PAR1 in males. These estimated data are nonetheless all derived from events occurring in one Y haplogroup and it might be interesting if a wider survey of ePAR chromosomes were considered.

Chapter 7 Surveying ePAR diversity in a large population sample

7.1 Introduction

As has been described in previous chapters, the ePAR owes its origin to non-allelic homologous recombination (NAHR) between LTR elements [Mensah *et al.*, 2014], and has occurred at least twice independently, producing ePARs associated with haplogroups I2 and R1b. The study that first reported the existence of ePAR was restricted in its initial survey population to ~2600 male developmental-disorder patients of Belgian and French origins [Mensah *et al.*, 2014], which present a limited spectrum of haplogroups. This raises the question of whether additional independent incidences of ePAR would be discovered if a larger and more diverse sample of Y-chromosomal lineages could be surveyed. To address this question, two different datasets, the Database of Genomic Variants [DGV] [MacDonald *et al.*, 2014] and UK Biobank [Sudlow *et al.*, 2015], were surveyed for evidence of ePAR based on copy-number variation of the relevant genomic regions in males. Since DNAs were available in the case of UK Biobank, any ePARs detected would also be able to be characterised further via PCR and sequencing approaches.

7.1.1 Database of Genomic Variants

DGV is an online summarized catalogue of human genomic structural variations (SVs) which are defined as genomic alterations involving >50-bp (up to >1 Mb) segments of DNA including indels, inversions, duplications or complex events [MacDonald *et al.*, 2014]. The database has been continuously updated and, to date, contains almost 7,000,000 variant regions across the genome from 72 published resources including peer-reviewed publications and three phases of the 1000 Genomes Project Consortium studies that cover ~70,000 globally distributed individuals [DGV]. The database provides a search platform as an online user-friendly browser that allows researchers to define regions of interest and to visualise types of SV by specifically coloured blocks as shown in Figure 7.1.

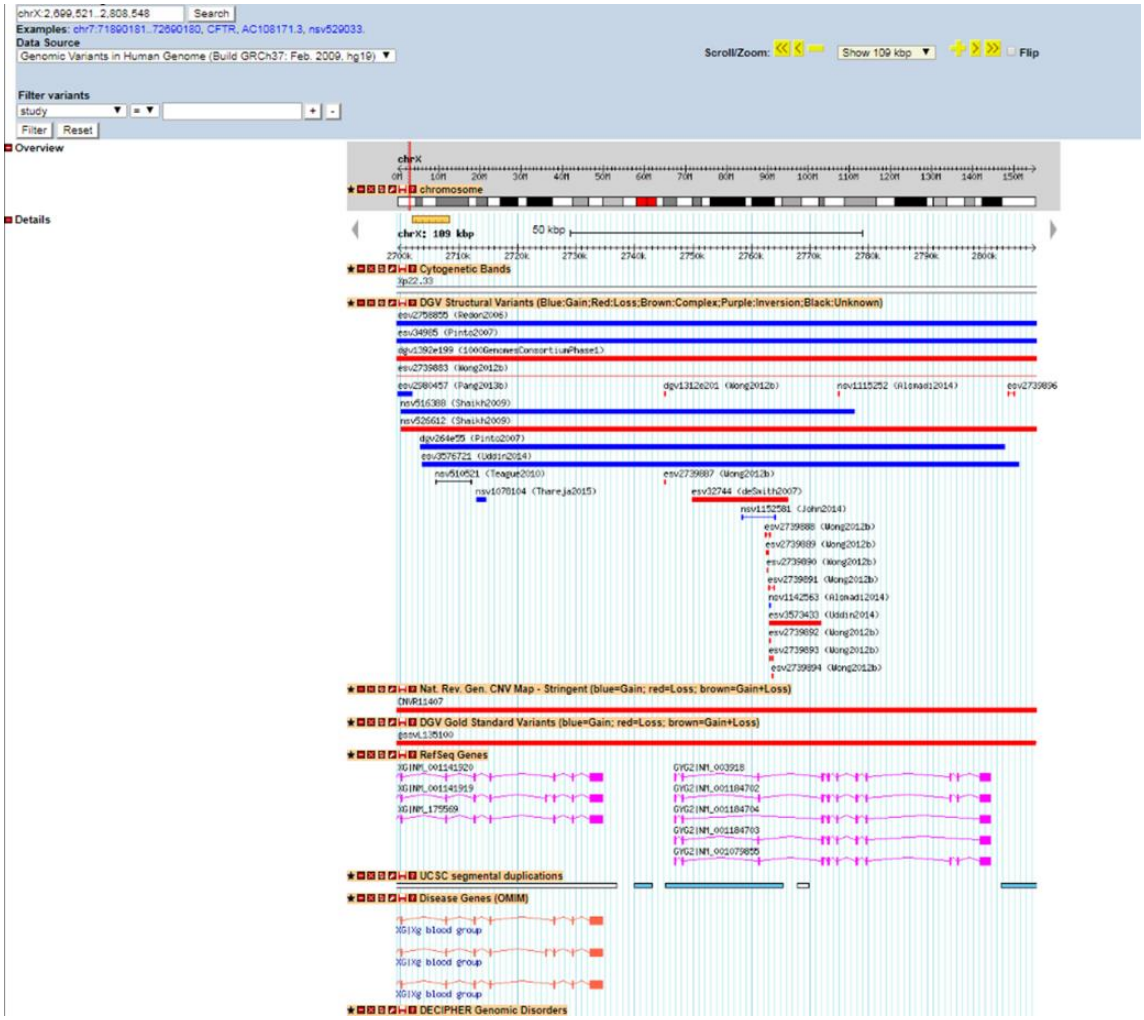


Figure 7.1 **Screenshot of DGV Genome Browser.** The screenshot shows an example based on hg19 coordinates with the search limited to the region of the X chromosome corresponding to that of ePAR. Blue blocks depict gain SVs including insertions and duplications, and red blocks indicate loss SVs. Each block is labelled with the study name and is clickable to view details. Magenta and orange boxes and lines below indicate genes in RefSeq or OMIM databases.

7.1.2 UK Biobank

UK Biobank is a very large 20-year cohort study based on >500,000 UK individuals aged 40-69 years recruited during 2006-2010; about half declared themselves as males [Sudlow *et al.*, 2015]. It collected detailed information on various aspects of individuals, including socio-demographic data, phenotypes including physical measures, and collection of blood, urine and saliva which subsequently underwent multiple biochemical tests [Sudlow *et al.*, 2015]. DNA was extracted from blood and genotyped using the Biobank's second-generation dense

genotyping chip (UK Biobank Axiom®Array) containing over 820,000 markers including ~350,000 common and ~280,000 low-frequency genome-wide SNPs, plus variants specific for CNVs, protein-truncating and other rare coding variants, other previously-identified common and rare disease-association risk variants, Neanderthal-ancestry and exome markers [Sudlow *et al.*, 2015]; therefore research on the cohort is able to span various aspects such as human evolution [Dannemann and Kelso, 2017], common and rare diseases, or CNV across the genome [Loh *et al.*, 2018, Wright *et al.*, 2017]. The microarray also covers the sex chromosomes, including >20,000 X-chromosomal SNPs (incorporating PAR1 and PAR2) and 807 Y-specific SNPs defining the main and some sub-branches of the Y phylogeny. Finally, the array includes >300 mtDNA SNPs (http://www.ukbiobank.ac.uk).

The recruited population is not only of indigenous British origin but also includes individuals from the UK's minority ethnic groups. Self-declared ethnicities of the UK Biobank cohort are shown in Figure 7.2.

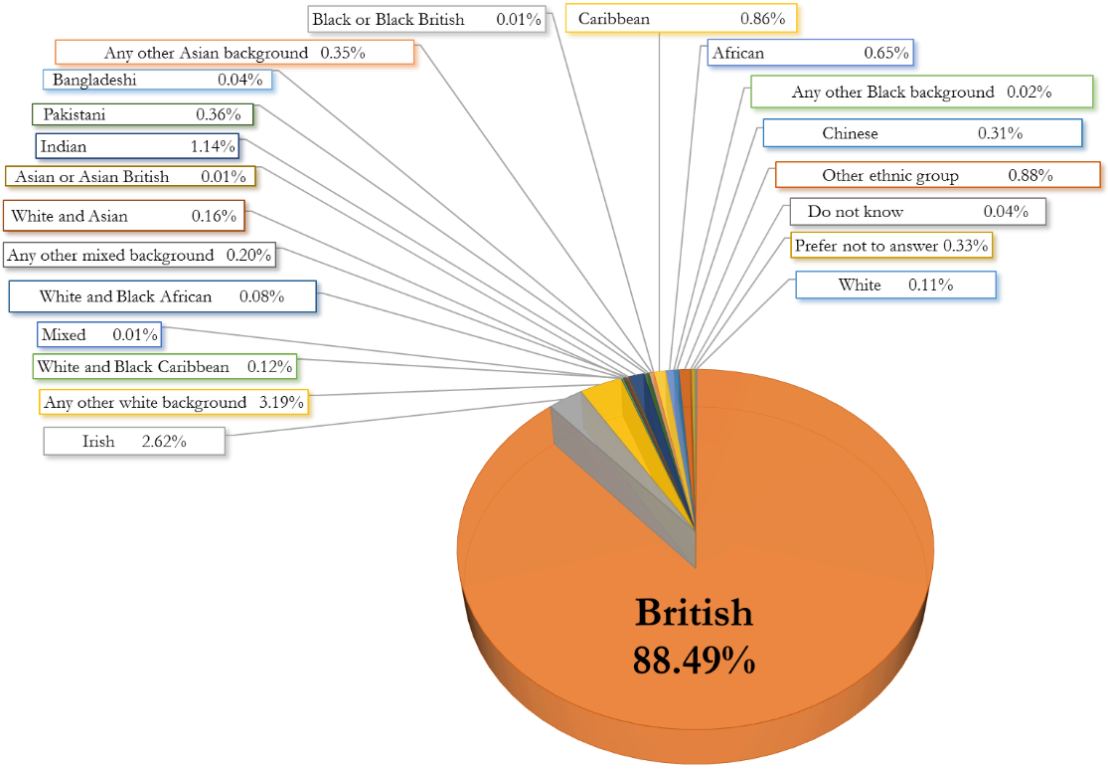


Figure 7.2 **Self-declared ethnicities of UK Biobank individuals.** Most individuals are Europeans with minorities defining as South and East Asians, Caribbeans, Africans and some other undefined ethnic groups (source of data: https://biobank.ctsu.ox.ac.uk/crystal/field.cgi?id=21000).

This Chapter describes a survey for ePARs in a large number of diverse individuals via the two databases described above. For UK Biobank, it also describes the definition of haplogroups, the use of the ePAR confirming Junction-PCR assay, and Sanger sequencing to define sequence diversity of the recombinant junctions in the discovered ePARs.

7.2 Results

7.2.1 Surveying DGV and using phylogenetic approach for potential new ePAR carriers

DGV was used to search for candidate ePAR cases using the information available on copy number variation. Populations in the database are distributed widely including Northern and Western Europeans from Utah US (CEU) and from Ontario Canada and Western European of unspecified milieu in US; Yoruba from Ibadan Nigeria (YRI); Han Chinese from Beijing China (CHB); and Japanese from Tokyo Japan (JPT) [Redon *et al.*, 2006, Shaikh *et al.*, 2009, Levy *et al.*, 2007, Pinto *et al.*, 2007, Uddin *et al.*, 2014, Pang *et al.*, 2013]. The total number of individuals reported specifically to this region was 3,677 and thirteen of them were found to carry an apparent duplication of X-linked SNPs in the vicinity of the ePAR (hg19 chrX:2694151-2808548). Among the 13 X-duplicated cases, 12 were Europeans and one was Japanese (JPT) (Table 7.1).

Table 7.1 Detected duplications potentially representing ePAR from DGV

Published source	Proportion of duplicated cases	List of positive variants	Population
Redon *et al.* (2006)	2/270	esv2758855, esv2758561, essv2294, esv2756781, essv3621, essv21575	CEU, JPT
Shaikh *et al.* (2009)	5/2026	nsv516388, nssv692885, nssv670793, nssv679790, nssv669823, nssv667956	Unspecified ancestry
Levy *et al.* (2007)	1/2	esv29994	Western-European American
Pinto *et al.* (2007)	4/771	dgv264e55, esv2752318, esv34985, essv6987024, essv6980174, esv2752319, essv6983877, essv6988725, esv34451, essv6978703, essv6986694	Europeans of unspecified ancestry, CEU, JPT
Uddin *et al.* (2014)	3/873	esv3576721, essv9822774, essv9822773, essv9822772	European Canadian
Pang *et al.* (2013)	1/1	essv7250024, esv2980457	Western-European American

Note: CEU = Utah Residents (CEPH) with Northern and Western Ancestry, JPT = Japanese in Tokyo, Japan; some papers shared the same individuals. Abbreviation suffixes "sv" is for "structural variant" and "ssv" is for "supporting structural variant" whereas prefix "e" is from European

Bioinformatics Institute (EBI), "n" is from National Center for Biotechnology Information [dbSNP] dbVAR (https://www.ncbi.nlm.nih.gov/dbvar/). DGV also merged various variants independently and prefixed with "dgv".

From DGV, only one dataset [Redon *et al.*, 2006] could be used for Y-Hg prediction since it was the only one for which Y-STR data were available. Based on Y-STR profiles which were provided by Dr. Chris Tyler-Smith (unpublished), 142 individuals (99 Y-lineages in 4 populations) were predicted by the online software provided by NevGen (http://www.NevGen.org/), using a Bayesian-Allele-Frequency Approach [Athey, 2006] (Table 7.2). The above-mentioned CEU and JPT samples were predicted to belong to Y-Hgs I2a-P37.2* and D1b*, respectively.

Table 7.2 Distribution of Y-Hgs in major populations and X-duplicated cases

Population	Total number		No. of Y-Hg by lineage	No. of X-duplicated case to Y-Hg
	Individual	Patrilineage		
CEU	44	26	I1 (6, 23.08%); I2a1 (1, 3.85%); I2a2a (1, 3.85%); R1a (1, 3.85%); R1b (17, 65.38%)	I2a-P37.2* (1)
CHB	22	21	C2e (2, 9.52%); D1b (1, 4.76%); J1a (1, 4.76%); O1 (2, 9.52%); O1b2 (2, 9.52%); O2a1b (12, 57.14%); N1c1 (1, 4.76%)	-
JPT	23	23	C1 (1, 4.35%); D1a (1, 4.35%); D1b (10, 43.48%); O1b2 (8, 34.78%); O2a1b (3, 13.04%)	D1b* (1)
YRI	53	29	E1a (2, 6.9%); E1b1a (27, 93.1%)	-

Note: CEU = Utah Residents (CEPH) with Northern and Western Ancestry; CHB = Han Chinese in Beijing, China; JPT = Japanese in Tokyo, Japan; YRI = Yoruba in Ibadan, Nigeria. *: this predicted haplogroup is not definite and is able to be typed to further sub-Hg subdivisions.

However, only the abovementioned CEU and JPT cases were available for DNA analysis, as they are in the HapMap panel [International HapMap Consortium, 2007] available in the laboratory, and could subsequently be analysed alongside the other DNAs in our collections. Additionally, the CEU case was re-predicted for Y-Hg using Dr. Jon Wetton's software, which gave the result I2a-L233. In summary, the survey from DGV indicated duplicated X-chromosome segments corresponding to the region of ePAR across wider population groups

(frequency ~0.35%); however, only two men, in the CEU and JPT populations, were able to be assessed further via DNA analysis.

As the DNA-available putative-ePAR individuals searched from DGV consist of only two cases, an attempt was made to add extra individuals using a phylogenetic approach, by taking males with Y chromosomes known to belong to the two previously ePAR-reported-haplogroups [Mensah *et al.*, 2014]. In the available sample collections, eleven unrelated cases, including two shared-surname males 333 and 6689_01 previously mentioned in Chapter 6, were additionally subjected to testing for the ePAR. Ten cases were predicted by Dr. Wetton's software to belong to the I2a* Y-Hg which was composed of two sub-Hgs, i.e. I2a-L233 and I2a-M423. The other one belonged to Hg R1b*.

Subsequently, two DGV-surveyed cases, one in the CEU which was case NA12146 from a CEPH family and one in JPT which was case NA18966, were added to those eleven putative cases from the available collection in the Junction-PCR experiments, making a total number of cases for testing to 13. The assay showed that eight were ePAR-positive, all of which belonged to Y-Hg I2a-L233 (Table 7.3). In addition, ePAR cases were Sanger-sequenced to identify their junction diversity as shown in Table 7.3 below.

Table 7.3 Potential DNA candidates detected for the ePAR

Case	Ethnic group	Predicted Y-Hg§	ePAR PCR assay	Junction diversity
NA12146	European American	I2a-L233	Positive	Junc1
Dan_con 24	Danish	I2a-L233	Positive	Junc1
333	English	I2a-L233	Positive	Junc1
6689_01	English	I2a-L233	Positive	Junc1
MT 5D4	Frisian	I2a-L233	Positive	Junc1
MT 8F5	English	I2a-L233	Positive	Junc1
MT 9D5	English	I2a-L233	Positive	Junc1
S_96	Irish	I2a-L233	Positive	Junc1
Hu_01	Hungarian	I2a-M423	Negative	-
Hu_02	Hungarian	I2a-M423	Negative	-
Hu_03	Hungarian	I2a-M423	Negative	-
S132	English	R1b*	Negative	-
NA18966	Japanese	D1b*	Negative	-

Note: Positive/Negative = the case carries/does not carry ePAR. Junc1 = the recombinant junctional sequence according to [Mensah *et al.*, 2014]. §Y-Hgs were predicted from Y-STRs by Dr. Wetton's in-house program except the last one, NA18966, which was predicted by NevGen (http://www.NevGen.org/).

The recombinant junction sequence diversity was found to be consistent with Mensah *et al.* since all ePAR-carriers were predicted to belong to Y-Hg I2a-L233, a sub-Hg of I2a-P37.2 [Mensah *et al.*, 2014] as described in Chapter 6, and all have the same type of recombinant junction, Junction 1, a product of crossover between LTR6B-X and LTR6B-YPAR1. The data so far still support the idea that all Hg I–L233 ePAR cases most probably result from a single rearrangement event and have been inherited down the generations, i.e. identical-by-descent. Even though the I-L1294 samples in Mensah *et al.* carried Junction 2 [Mensah *et al.*, 2014] which is explained via a gene conversion, the data they reported was from only one pedigree. More evidence is required to confirm the relation between I-L1294 and this recombinant mechanism.

The ePAR junction PCR assay was applied to the two DGV samples described above. Unsurprisingly, the CEU sample predicted to belong to Y-Hg I2a-L233 gave a positive result in the PCR test, and subsequent Sanger sequencing showed this individual to carry Junction 1 as defined by Mensah et al. (2014). However, the JPT sample failed to give a positive ePAR junction result, so likely carries a different kind of rearrangement. Time was not available to investigate this case further.

7.2.2 Searching the UK Biobank database for candidate ePAR-bearing Y chromosomes

The original identification of ePAR relied upon inference of copy-number increase of an X-chromosomal segment in males, based on SNP intensity information from a genome-wide SNP chip. A similar approach was taken to the UK Biobank cohort males, a sample which is some hundred times larger than the sample set screened previously [Mensah *et al.*, 2014].

Work on UK Biobank data was undertaken as a collaboration with Prof. Maciej Tomaszewski accompanied by Dr. James Eales and Dr. Xiaoguang (Allan) Xue at the University of Manchester, as part of UK Biobank Main Application 15915 'Paternal lineages of the Y chromosome and men's health and disease'. The Manchester collaborators were provided with SNP coordinates for the region of interest, and carried out bioinformatic analysis of UK Biobank data to screen for duplication of the translocated-X portion of the ePAR to identify putative ePAR carriers. Access to DNA samples was agreed after application to UK Biobank: these samples were held under UK Biobank Main Application 6077 (which aims to measure telomere lengths in all UK Biobank participants), led by Prof. Sir Nilesh Samani and Dr. Veryan Codd at the Department of Cardiovascular Sciences,

Glenfield Hospital, University of Leicester. These Leicester collaborators kindly provided access to the putative ePAR-carrying DNA samples from UK Biobank for experimental analysis. Samples were extracted from the plate-sets held at Glenfield Hospital by the author, and brought to the Department of Genetics & Genome Biology for further analysis.

UK Biobank released full SNP genotyping results on all samples by July 2017, allowing *in silico* screening for ePAR to be undertaken. DNAs from all UK Biobank samples were not delivered to Glenfield Hospital until the third trimester of 2018; because of reasons of timing, therefore, this study could not have access to DNA samples from the entire sample set. Nonetheless, access was gained to >90% of samples.

Over 800,000 SNPs and indels were genotyped by the UK Biobank Axiom® Array typed via fluorescent signal intensity, which allows increases or decreases of copy number, as well as hetero- or homozygosity, to be detected. The SNP list for the array was accessed via http://www.ukbiobank.ac.uk/scientists-3/uk-biobank-axiom-array/, allowing putative ePAR SNPs to be identified: 64 SNPs located in the distal X-specific segment, as a part of ePAR, were screened together with flanking SNPs proximally and distally (details of SNPs are in Appendix). Intensity of the X-SNPs is predicted to be doubled, equivalent to values seen in normal women, in men who carry ePAR compared to those who do not, as demonstrated in a simple schematic diagram in Figure 7.3. Three proximal SNPs were also expected to be present in a triple dose, but given this very small number, and natural variation in SNP intensity, this increase was not screened for.

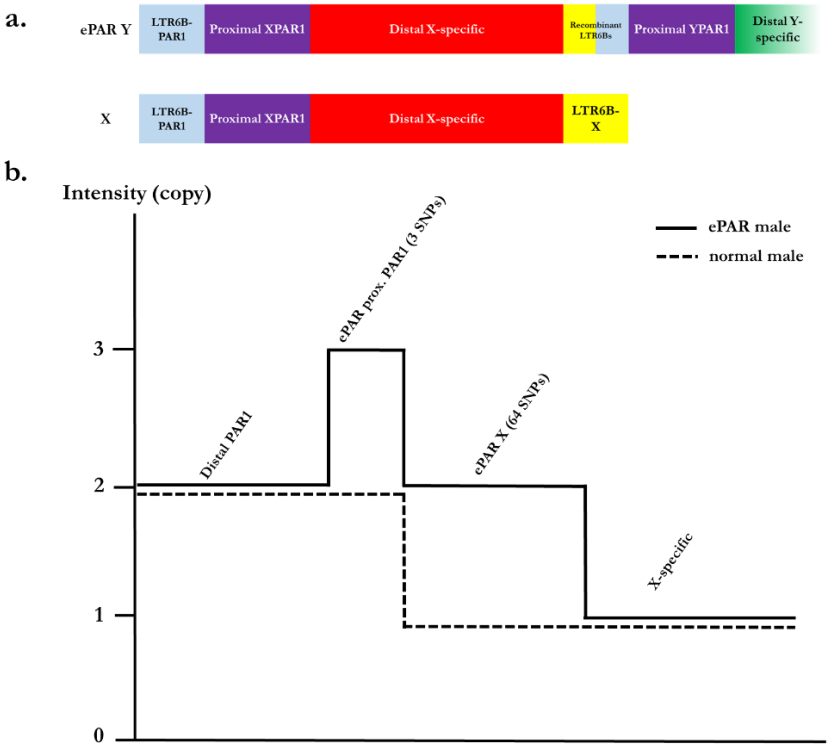


Figure 7.3 **Simplified schematic of rationale behind approach to screen ePAR from UK Biobank SNP microarray data. a)** Colour diagrams show the diploid sequence structures in men who carry the ePAR Y and a normal X chromosome: between them, the two sex chromosomes contain three copies of the proximal PAR1, two copies of the distal X on ePAR, two copies of the rest of PAR1 and one copy of the rest of the X-specific region. **b)** Simple schema demonstrates a comparison of fluorescent intensity between an ePAR carrier and a normal man for the relevant regions, annotated with the number of SNPs lying in relevant regions on the chip.

The Manchester collaborators performed ePAR screening in 223,605 UK Biobank men (~45.8% of the entire available population size) discarding a batch of 7486 individuals who failed an overall signal-intensity QC. Then median $\log_2$ ratio of the QC-passed markers on the X and Y chromosomes, excluding all PARs, was measured and all the QC-passed samples were checked to ascertain whether they were normal 46,XY males by determining the presence of one X and one Y chromosome. This step removed examples of various Y aneuploidies, the total number of male samples decreasing to 222,988. Y-Hgs were generated for all remaining male samples using the yHaplo® Software [Martiniano *et al.*, 2017, Poznik, 2016], which bases haplogroup decisions on presence of ancestral and derived states of

markers, checked against a reference Y phylogeny. After this, each SNP signal intensity was calculated as median of deviation from the median in females, on the other words, a normalised median intensity. Boxplots for SNPs of the ePAR X-portion (as per the simple schema in Figure 7.3) were generated and displayed by haplogroup, as shown in Figure 7.4.

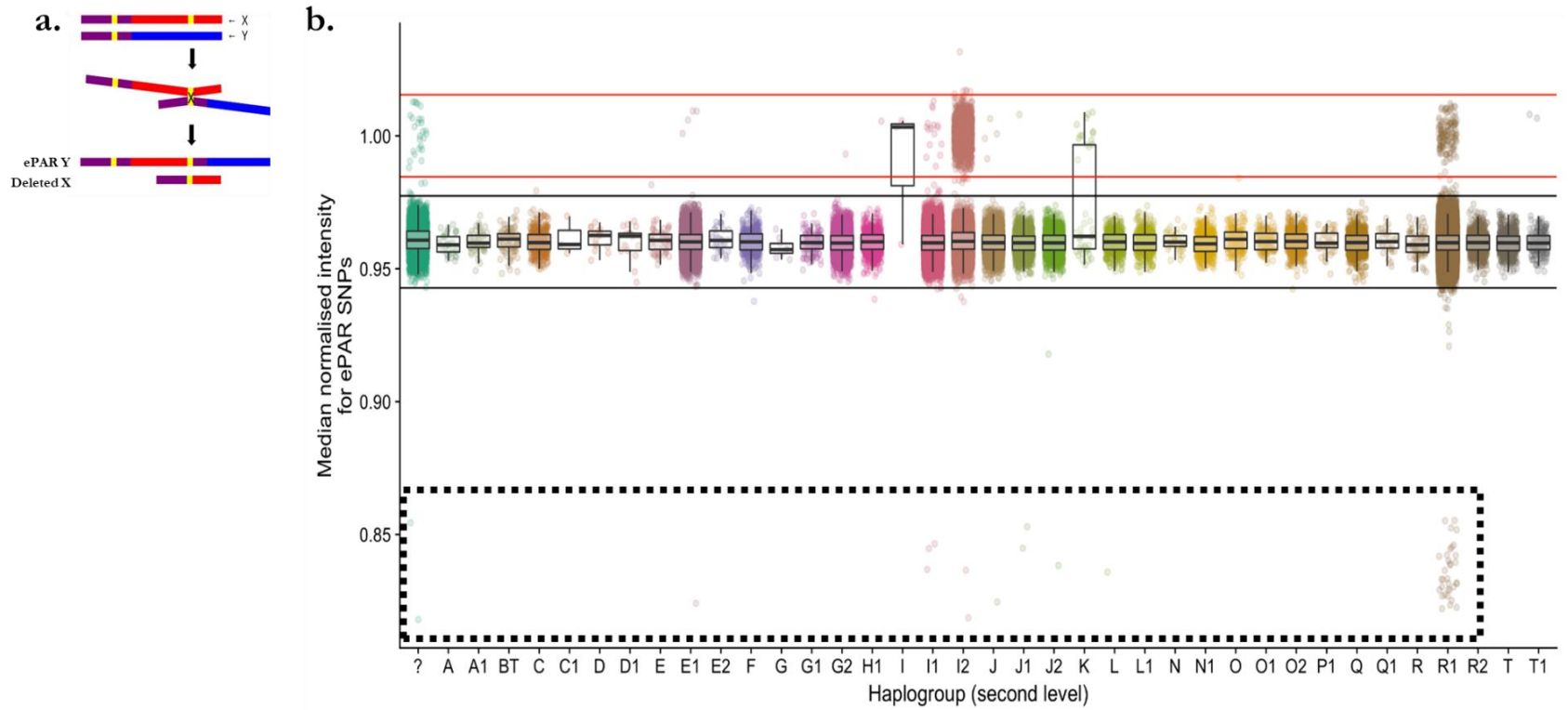


Figure 7.4 Boxplot showing median normalised SNP intensities for 64 ePAR SNPs by Y haplogroup. a) Diagram shows the mechanism of ePAR generation and the end products, including one insertion Y chromosome (ePAR Y) and one reciprocal deletion X chromosome (adapted from [Mensah *et al.*, 2014]). **b)** Median normalised SNP intensity data categorised by Y-Hg which was identified by Y-SNPs on the UK Biobank Axiom® Array (undertaken by collaborators from the University of Manchester, unpublished). The Y-Hg "?" indicates Hg-unidentified cases in which Y- SNPs are of poor quality but sufficient other SNPs pass QC. Note that haplogroups which failed to be typed for the downstream SNPs are categorized as a higher level Hg. For instance, men in the "I" Hg (equivalent to the extremely rare Hg I*) could be in either I1 or I2 but downstream SNPs may be of poor quality. Individuals whose intensities are within the median ± 95%CI of the intensity of the haploid X, i.e. similar to normal men, are plotted between two black horizontal lines whilst those whose intensities are similar to those of the diploid X, i.e. similar to normal women, are plotted between two pink horizontal lines of the median ± 95%CI of those of female X chromosomes. Men whose ePAR-related X chromosome intensities fall in the diploid level are identified as putative ePAR cases. In this Figure, the putative ePAR men apparently predominate in the Hg I2 followed by R1 which is consistent with the reported data [Mensah *et al.*, 2014]; however, there are many other Hgs which contain small numbers of putative ePARs in this study. Interestingly, there are also examples of sporadic cases showing SNP intensities well below the haploid zone (in black dashed rectangle); these are likely to carry a deleted X chromosome, and therefore may represent examples of the reciprocal X deletion of the ePAR shown in the lower section of part (a).

The final total number of men screened for ePAR according to Figure 7.4 b was 216,119. Men whose haplogroup is shown as "?", i.e. unidentified, have poor-quality Y-SNP calls, and for simplicity these 5086 individuals were excluded from further analysis. Therefore, the net number of the UK Biobank male population analysed is 211,033. The total number of putative ePARs was 1,676 or ~0.79% of the sample, equating to a frequency of approximately 1 in 126 among the UK Biobank male population. In addition, the number of putative X deletion men, described in Figure 7.4, was 40. Since the latter are not Y-linked, there is no expectation of over-representation in any particular haplogroup (*cf.* ePAR in Y-Hg I2), and indeed this is observed in Figure 7.4b, where X deletion frequency appears to be distributed proportionally to haplogroup frequency. Summary data of the putative ePAR men by preliminary Hg is presented in Table 7.4.

Table 7.4 Putative ePAR by preliminary high-level Hg

Haplogroup	No. of total indiv.	% in pop.	No. of pos.	% of pos. per Hg	% of pos. per pop.	Odds (1 in ?)
I	3	0.0014	2	66.6667	0.0009	2
K	43	0.0204	14	32.5581	0.0066	3
I2	16026	7.5941	1496	9.3348	0.7089	11
T1	196	0.0929	2	1.0204	0.0009	98
H1	665	0.3151	1	0.1504	0.0005	665
J	4008	1.8992	4	0.0998	0.0019	1002
R1	149973	71.0661	135	0.0900	0.0640	1111
J1	1538	0.7288	1	0.0650	0.0005	1538
I1	25266	11.9725	16	0.0633	0.0076	1579
E1	9329	4.4206	4	0.0429	0.0019	2332
G2	3986	1.8888	1	0.0251	0.0005	3986

indiv. = individuals; pop. = total population, pos. = putative ePAR cases; this data was obtained from the Manchester collaborators.

As expected, the majority of observed apparent ePAR samples belong to Y-Hg I2 – this, in effect, acts as a positive control, supporting the idea that the survey is indeed detecting ePAR Y chromosomes. However, in addition, there is a set of non-I2 chromosomes, belonging to a range of haplogroups, that could represent independent ePAR occurrences. Before proceeding further, it was necessary to confirm the ePAR status of these candidates using the ePAR junction PCR assay.

7.2.3 Confirmation of the ePAR status of putative cases, and haplogroup distribution based on Axiom SNP data

Since the vast majority of putative ePARs belonged to Y-Hg I2 (1496 men, ~9% of the Hg) and were believed to be Hg I2a, and therefore most likely were identical by descent with each other and with the Y-Hg I2a samples tested previously in this project, it was decided not to validate all of these, but instead to analyse a random subset of 100. By contrast, all 91 non-I2 putative ePARs for which DNAs were available were chosen for analysis, and DNA aliquots were collected from Glenfield Hospital. The 24 available DNA samples representing putative X-deletions were also taken, but due to time restrictions these have not been analysed as part of this thesis.

To confirm their ePAR status, the 191 total putative ePAR samples were subjected to the Junction-PCR assay (see Chapter 3). All 100 Y-Hg I2 samples gave positive results, confirming that they indeed carried ePAR, whilst the non-I2 groups of 91 samples showed 41 samples (~45%) positive. The non-ePAR duplication cases might carry other rearrangements which contain larger extended segments, or could have X segments translocated to other parts of the genome different from the position of ePAR. Time was not available to further investigate these samples as part of this thesis.

A summary of the results of ePAR testing, showing SNP-based haplogroup designations, is given in Table 7.5. This indicates that, as well as the expected Hg I2a lineage, there might be other independent instances of ePAR within haplogroups K-M9, and E1a-M132. In contrast, Y-Hg R1b-P311 which covers the previously reported ePARs in Hg R1b-P312 [Mensah *et al.*, 2014], -U152 and -CTS3655 shows separated events of ePAR rather than identical by descent from the common root.

Table 7.5 Results of Junction PCR ePAR test in putative ePAR carriers

Preliminary Hg	No. of sample tested	No. of positive	% of positive
I2-M438	100	100	100.00
I2a-M223	3	1	33.33
I-M170	1	1	100.00
I1-M253	8	5	62.50
I1a-P109	1	0	0.00
R1b-P311	49	23	46.94
R1b-U152	5	2	20.00
R1b-CTS3655	1	1	100.00
R1b-M269	1	0	0.00

R1b-M467	2	0	0.00
R1b-L45	1	0	0.00
R1b-M222	2	0	0.00
R1b-CTS2501	1	0	0.00
R1b-L1065	1	0	0.00
R1b-CTS4314.3	1	0	0.00
R1a-M417	1	0	0.00
K-M9	7	7	100.00
E1a-M132	1	1	100.00
E1b-V13	1	0	0.00
G2a-P15	1	0	0.00
H1a-M52	1	0	0.00
J-M304	2	0	0.00

Note: R1b-P311 is the most common root of the other R1b sub-Hgs in this table. Mensah et al. mentioned two ePAR men as the haplogroup R1b-P312 [Mensah *et al.*, 2014] which is the common root of R1b-U152 and-CTS3655 while the others are separate sub-Hgs.

7.2.4 Haplogroup prediction via Y-STR typing and confirmation of Hg I2a lineages in ePAR carriers

To confirm, and possibly refine, haplogroup designations all the tested samples, ePAR positive or not, underwent Hg prediction following Y-STR profiling. The 191 DNA samples were typed for 23 Y-STRs using the PowerPlex® Y23 System (PPY23) as described in Chapter 2, and subsequently their Hgs were predicted from haplotypes using the batch capability function in the desktop version of NevGen Genealogy Tools Software v1.1 (http://www.nevgen.org/). The prediction results resolved some Hgs to further derived sub-lineages, left some Hgs at the same resolution, and were unable to reliably predict two particular sets of samples belonging to array-defined Hgs K-M9 and E1a-M132, possibly because these are rare in Europeans, and therefore not represented in the reference datasets used by the NevGen predictor (Table 7.6). However, apart from these two unpredicted Hgs, the remainder showed 100% concordance of Hg identification between the two approaches (array-based SNPs and Y-STR-based prediction).

Table 7.6 A comparison between Hg SNP-array typing and Hg prediction by Y-STRs

Hg by Axiom (no.)	Hg by Y-STRs prediction (no.)	Remark
I2-M438 (100)	I2a-L233 (98), I2a-L1294 (2)	Resolved to more derived lineages; all are ePARs
I2a-M223 (3)	I2a-Y10626 (1), I2a-S7753 (1), I2a-L623 (1)	Only one I2a-Y10626 is ePAR
I-M170 (1)	I2a-L233 (1)	Resolved to a more derived lineage; ePAR case
I1-M253 (8)	I1-M253 (1), I1a-Z58 (4), I1a-L338 (1), I1c-Z17926 (1), I1a-S4795 (1)	Unresolved I1-M253; one I1a-Z58 and one I1a-S4795 = non-ePAR
I1a-P109 (1)	I1-M253 (1)	Unable to predict to a deeper lineage; non-ePAR
R1b-P311 (49)	R1b-M343 (49)	Unable to predict to a deeper lineage; 23/49 are ePARs
R1b-U152 (5)	R1b-M343 (5)	Unable to predict to a deeper lineage; 2/5 are ePARs
R1b-CTS3655 (1)	R1b-M343 (1)	Unable to predict to a deeper lineage; ePAR
R1b-M269 (1)	R1b-M343 (1)	Unable to predict to a deeper lineage; non-ePAR
R1b-M467 (2)	R1b-M343 (2)	Unable to predict to a deeper lineage; all are non-ePARs
R1b-L45 (1)	R1b-M343 (1)	Unable to predict to a deeper lineage; non-ePAR
R1b-M222 (2)	R1b-M343 (2)	Unable to predict to a deeper lineage; all are non-ePARs
R1b-CTS2501 (1)	R1b-M343 (1)	Unable to predict to a deeper lineage; non-ePAR
R1b-L1065 (1)	R1b-M343 (1)	Unable to predict to a deeper lineage; non-ePAR
R1b-CTS4314.3 (1)	R1b-M343 (1)	Unable to predict to a deeper lineage; non-ePAR
R1a-M417 (1)	R1a-L146 (1)	Unable to predict to a deeper lineage; non-ePAR
K-M9 (7)	Unable to predict (7)	Pakistani; all are ePARs
E1a-M132 (1)	Unable to predict (1)	Caribean; ePAR
E1b-V13 (1)	E1b-V13 (1)	Unable to predict to a deeper lineage; self-declared as British but this Hg is prominent in Near East, South Europe and North Africa; non-ePAR
G2a-P15 (1)	G2a-L13 (1)	Resolved to a more derived lineage; self-declared as White and Asian, also this Hg is rare in Europe; non-ePAR
H1a-M52 (1)	H1a-M82 (1)	Resolved to a more derived lineage; Pakistani; non-ePAR
J-M304 (2)	J2a-L26 (2)	Resolved to a more derived lineage; self-declared as British but this Hg is low-frequent amongst western Europeans; all are non-ePARs

Because the major ePAR-linked Hg is I2, and, from the previous evidence presented in Chapter 6 that studied Hg I2 ePARs are identical by descent, confirmation of Y-Hgs in these UK Biobank men should help elucidate the relationship among ePAR and non-ePAR individuals within Hg I. To start with, Y-STR data on all Hg I men, no matter their sub-haplogroup or ePAR status, were combined together to generate a median-joining network (MJN), using Network 5 together with Network Publisher Software (http://www.fluxus-engineering.com/) [Bandelt *et al.*, 1999] as presented in Figure 7.5.

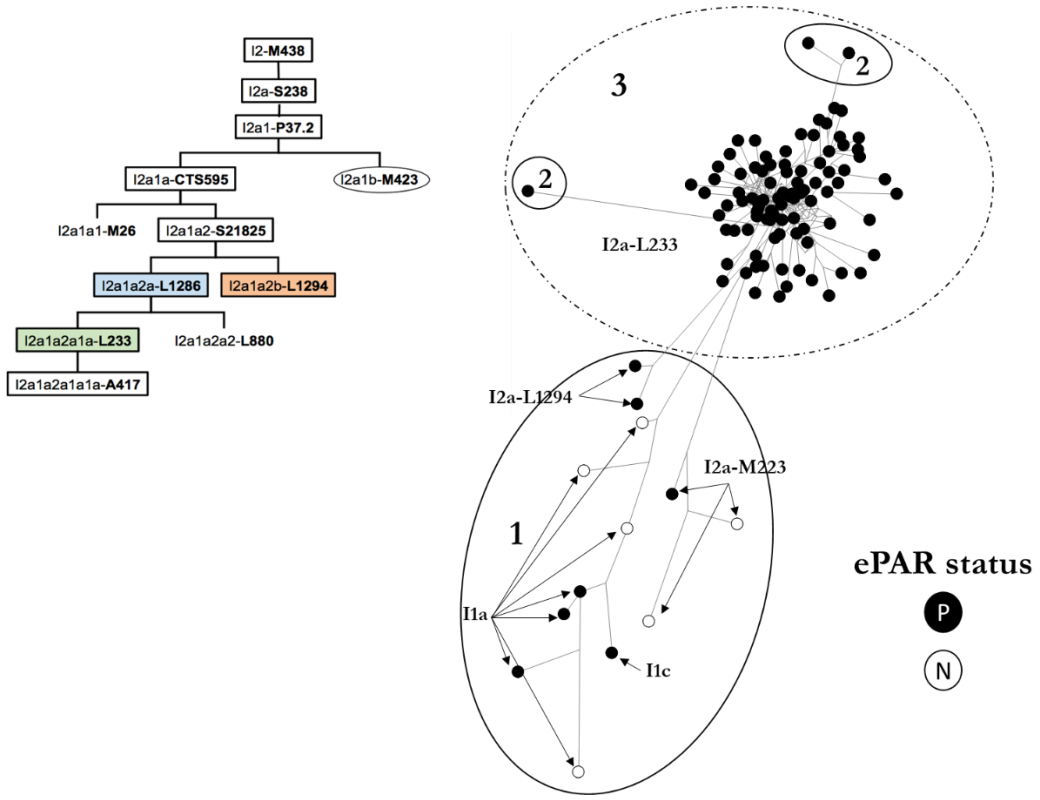


Figure 7.5 Median-joining network (MJN) generated from Y-STR haplotypes in all Hg I men. The MJN was constructed from Y-STR haplotypes determined for all putative ePAR-carrying men typed by the Axiom® Array to be Hg I1 and I2 as described above. Black (positive) and white (negative) circles represent ePAR and non-ePAR haplotypes, respectively. All I-L233 haplotypes are encircled by the dashed ellipse labelled '3' while the rest marked with black-solid ellipse labelled '1' are the other Hg I cases. Twenty-two Hg I men were selected for Hg I2a Y-SNP sub-typing by taking all 13 samples in the black ellipse '1', three I-L233 outliers encircled in black labelled '2'; and six randomly selected I-L233 examples from the core cluster. A simplified Y-Hg phylogenetic tree representing SNPs typed is shown on the upper left: Y-Hg I1 and I2a-M233 are not included in the tree.

Twenty-two UK-Biobank Hg-I men were selected for Hg I2a SNP sub-typing via a SNaPshot® assay as previously presented in Chapter 6. The genotyping results are shown in Table 7.7.

Table 7.7 SNaPshot® assay determining I2a sub-Hgs in UK Biobank Hg-I men

	M348	P37.2	S21825*	CTS595*	L233	A417*	S238*	L1286*	L1294	Y-Hg
I1-A2	G	C	T	A	A	A	G	T	T	L233
I1-E2	G	C	T	A	A	A	G	T	T	L233
I1-A5	G	C	T	A	A	A	G	T	T	L233
I1-A12	G	C	T	A	A	A	G	T	T	L233
I1-G7	G	C	T	A	A	A	G	T	T	L233
I1-G9	G	C	T	A	A	A	G	T	T	L233
I1-B10[a]	G	C	T	A	A	A	G	T	T	L233
I1-C7[a]	G	C	T	A	A	A	G	T	T	L233
I1-D1[a]	G	C	T	A	A	A	G	T	T	L233
I1-G4[$]	G	C	T	A	G	A	G	C	C	L1294
I1-H6[$]	G	C	T	A	G	A	G	C	C	L1294
O1-D2[$]	G	T	C	G	G	A	G	C	T	S238
O1-F7[$]	A	T	C	G	G	A	T	C	T	Not I2
O1-C8[$]	A	T	C	G	G	A	T	C	T	Not I2
O2-A2[$]	A	T	C	G	G	A	T	C	T	Not I2
O2-B2[$]	A	T	C	G	G	A	T	C	T	Not I2
O1-F2[$b]	A	T	C	G	G	A	T	C	T	Not I2
O1-F4[$b]	G	T	C	G	G	A	G	C	T	S238
O1-E5[$b]	G	T	C	G	G	A	G	C	T	S238
O1-H5[$b]	A	T	C	G	G	A	T	C	T	Not I2
O1-F8[$b]	A	T	C	G	G	A	T	C	T	Not I2
O2-A1[$b]	A	T	C	G	G	A	T	C	T	Not I2

* markers in reverse strands; [a] I-L233 outliers; [$] men from non I-L233 cluster (Ellipse 1 in Figure 7.5); [b] (*italic*) non-ePAR men

All Hg-predicted I-L233 cases, including outliers, were confirmed to be genuine Hg I-L233; also, two samples predicted Hg I-L1294 were confirmed by SNP typing. It was found that one striking outlier in MJN of the I-L233 cluster (the longest branch in the left black ellipse '2' in Figure 7.5) is caused by this individual carrying a null allele for Y-STR DYS570, which otherwise has the modal allele 17 (i.e. 17 tandem repeats). The results therefore show that every predicted I-L233 chromosome is indeed Hg I-L233; in addition, ePAR association with Hgs I-L233 and I-L1294 is consistent with the I2a ePAR association previously reported [Mensah *et al.*, 2014] which was confirmed in Chapter 6. Furthermore, all predicted I1a and I1c cases were confirmed for not being I2 haplogroups. However, three I2a-M223 cases predicted to be sub-Hgs I2a-Y10626, I2a-S7753, I2a-L623 were confirmed by the SNP typing as I2a-S238 and did not show further derived alleles down the tree shown above: thus, the predicted sub-Hgs conformed to the SNP typing results.

7.2.5 ePAR junction diversity among UK Biobank samples

To help ascertain the history of ePAR formation, NAHR junction sequence diversity should be taken into account. As previously performed for the ePAR samples in laboratory collections (Chapter 6), Sanger-sequencing was applied to a total of 97 ePAR samples from UK Biobank; these consist of a sub-set of 55 samples from Hg I2a-L233, chosen to maximise diversity by taking all the outlying branches and a few of the inner core from the network, plus the other 42 ePAR-carrying samples not in Hg I2a-L233; including I2a-L1294 (2), I1a-Z58 (3), I1a-L338 (1), I1c-Z17926 (1), I2a-Y10626 (1), R1b-P311 (23), R1b-U152 (2), R1b-

CTS3655 (1), K-M9 (7) and E-M132 (1) (numerals in parenthesis refer to a number of samples).

Surprisingly, the results uncovered eleven distinct sequences including the two types described in the original paper [Mensah *et al.*, 2014]. By aligning all eleven types together, the junction sequences can be sorted into two groups based on Junction 1 and 2 described in Mensah *et al.* (2014) and the other three different types and one group. The nomenclature used to identify each type of junction sequence is based on: 1) the possible recombination interval which is most parsimoniously inferred from the LTR6B-specific alleles flanking the interval, 2) variants found on each sequence and 3) the length of recombinant LTR6B junction.

As described in Chapter 3 and in the original paper [Mensah *et al.*, 2014], Junction 1 is 559 bp while Junction 2 is 551 bp in length. However, to specify the possible recombinant interval in each type is not straightforward because there are many possible approaches able to give different outcomes. To resolve in the most parsimonious way, comes to the question of whether using the reference LTR6B sequences is appropriate, or if allele frequencies of each variant should be taken into account, to modify the reference sequences of both donor and acceptor LTR6Bs. Because the standard reference sequence for each LTR6B [UCSC Genome Browser] was obtained from only one human, that sequence may carry rare alleles at some sites; however, the ePAR individuals could instead be assumed to carry only major-frequency alleles in all variants, either SNPs or indels, to make a more parsimonious solution. Figure 3.6 shows alignment of the Junction 1 against the original reference sequence [UCSC Genome Browser] while Figure 7.6 below is presented with a modified reference sequence taking allele frequencies into account such that there are four SNPs on the LTR6B on YPAR1, changing genotypes.

a.


5 15 25 35 45
Junction 1 TGTGTTGTAC CCGAGCGAGT TAGAAAAACG CCACACTTTG AGACGATTTA
LTR6B-X TNTGTTGTAC CNGAGCGAGT TAGAAAAACG CCACACTTTG AGACGATTTA
LTR6B-YPAR1 TNTGTTGTAC CNGAGCGAGT TAGAAAAACG C-----TTTG AGACGATTTA
305 315 325 335 345
Junction 1 CTAATCAGAC GAATTCCCGG GAACTGCGGA TGTAGCTCGC CACAGTATCT
LTR6B-X CTAATCAGAC GAATTCCCGG GAACTGCGGA TGTAGCTCGC CACAGTATCT
LTR6B-YPAR1 CTAATCAGAC GAATTCCCGG GAACTGCGGA TGTAGCTCGC CACAGTATCT
55 65 75 85 95
Junction 1 AGAGTCCTTT ATTAGCCGGC GACCGAGAGA CGGCTAACGC TCAAAATTCT
LTR6B-X AGAGTCCTTT ATNAGCCNGC GACCGAGAGA CGGCTAANGC TNAAAATTCT
LTR6B-YPAR1 AGAGTCCTTT ATNAGCCNGC GACCGAGAGA CGGCTAANGC TNAAAATTCT
355 365 375 385 395
Junction 1 TATCAGTTAA CTGCATTCTT GGATGTGCTG GGAGTCAGCC TGCACGAGTT
LTR6B-X TATCAGTTAA CTGCATTCTT GGATGTGCTG GGAGTCAGCC TGCACGAGTT
LTR6B-YPAR1 TATCAGTTAA CTGCATTCTT GGATGTGCTG GGAGTCAGCC TGCACGAGTT
105 115 125 135 145
Junction 1 CTCGGCCCCG AGGAAGGGGC TTGATTAACT TTTAGATCTT GGTTTAGGAA
LTR6B-X CTCGGNCCCG AGGAAGGGGC TTGATTAACT TTTAGATCTT GGTTTAGGAA
LTR6B-YPAR1 CTCGGNCCCG AGGAAGGGGC TTGATTAACT TTTAGATCTT GGTTTAGGAA
405 415 425 435 445
Junction 1 CAGTCCTTGA GGAAGGGGCT GCCAGTGAAA GAGCCAAGGT GGAGTCTGGC
LTR6B-X CAGTCCTTGA GGAAGGGGCT GCCAGTGAAA GAGCCAAGGT GGAGTCTGGC
LTR6B-YPAR1 CAGTCCTTGA GGAAGGGGCT GCCAGTGAAA GAGCCAAGGT GGAGTCTGGC
155 165 175 185 195
Junction 1 GGGGAGGGCG GGGGGTCTAG TGAAAACCAT TTTACAGAAG TAAAGTAGGC
LTR6B-X GGGGAGGGCN GGGGGTCTAG TGAAAACCAT TTTACAGAAG TAAAGTAGGC
LTR6B-YPAR1 GGGGAGGGCN GGGGGTCTAG TGAAAACCAT TTTACAGAAG TAAAGTAGGC
455 465 475 485 495
Junction 1 NGGCTCTCTT AGCTAAGGGA GAGTCCATTC AGGTGGAAAG AAGGCTAGGT
LTR6B-X TGGCTCTCTT AGCTAAGGGA GAGTCCATTC AGGTGGAAAG AAGGCTAGGT
LTR6B-YPAR1 TGGCTCTCTT AGCTAAGGGA GAGTCCATTC AGGTGGAAAG AAGGCTAGGT
205 215 225 235 245
Junction 1 AAAAAGTTAA AAGGATAAAT GGTTGCAGGA AAGTAAACAG TTCCAGGTGC
LTR6B-X AAAAAGTTAA AAGGATAAAT GGTTGCAGGA AAGTAAACAG TTCCAGGTGC
LTR6B-YPAR1 AAAAAGTTAA AAGGATAAAT GGTTGCAGGA AAGTAAACAG TTCCAGGTGC
505 515 525 535 545
Junction 1 GAGTAGAGGA AAAGGGAGAG TCTAAAAACA GGTTAGTAAA AACCAGGTTG
LTR6B-X GAGTAGAGGA AAAGGGAGAG TCTAAAAACA GGTTAGTAAA AACCAGGTTG
LTR6B-YPAR1 GAGTAGAGGA AAAGGGAGAG TCTAAAAACA GGTTAGTAAA AACCAGGTTG
255 265 275 285 295
Junction 1 AGGGGCTTTA AGACTATTAC AAGGTGATAG ACGCGGGGCT TTGGGCGTTA
LTR6B-X AGGGGCTTTA AGACTATTA---------TAG ACGCGNGGCT TTGGGCGTTA
LTR6B-YPAR1 AGGGGCTTTA AGACTATTAC AAGGTGATAG ACGCGNGGCT TTGGGCGTTA
555
Junction 1 GGCATTACA
LTR6B-X GGCATTACA
LTR6B-YPAR1 GGCATTACA

b.


5 15 25 35 45
Junction 1 TGTGTTGTAC CCGAGCGAGT TAGAAAAACG CCACACTTTG AGACGATTTA
LTR6B-X TNTGTTGTAC CNGAGCGAGT TAGAAAAACG CCACACTTTG AGACGATTTA
LTR6B-YPAR1 TNTGTTGTAC CNGAGCGAGT TAGAAAAACG C-----TTTG AGACGATTTA
305 315 325 335 345
Junction 1 CTAATCAGAC GAATTCCCGG GAACTGCGGA TGTAGCTCGC CACAGTATCT
LTR6B-X CTAATCAGAC GAATTCCCGG GAACTGCGGA TGTAGCTCGC CACAGTATCT
LTR6B-YPAR1 CTAATCAGAC GAATTCCCGG GAACTGCGGA TGTAGCTCGC CACAGTATCT
55 65 75 85 95
Junction 1 AGAGTCCTTT ATTAGCCGGC GACCGAGAGA CGGCTAACGC TCAAAATTCT
LTR6B-X AGAGTCCTTT ATNAGCCNGC GACCGAGAGA CGGCTAANGC TNAAAATTCT
LTR6B-YPAR1 AGAGTCCTTT ATNAGCCNGC GACCGAGAGA CGGCTAANGC TNAAAATTCT
355 365 375 385 395
Junction 1 TATCAGTTAA CTGCATTCTT GGATGTGCTG GGAGTCAGCC TGCACGAGTT
LTR6B-X TATCAGTTAA CTGCATTCTT GGATGTGCTG GGAGTCAGCC TGCACGAGTT
LTR6B-YPAR1 TATCAGTTAA CTGCATTCTT GGATGTGCTG GGAGTCAGCC TGCACGAGTT
105 115 125 135 145
Junction 1 CTCGGCCCCG AGGAAGGGGC TTGATTAACT TTTAGATCTT GGTTTAGGAA
LTR6B-X CTCGGNCCCG AGGAAGGGGC TTGATTAACT TTTAGATCTT GGTTTAGGAA
LTR6B-YPAR1 CTCGGNCCCG AGGAAGGGGC TTGATTAACT TTTAGATCTT GGTTTAGGAA
405 415 425 435 445
Junction 1 CAGTCCTTGA GGAAGGGGCT GCCAGTGAAA GAGCCAAGGT GGAGTCTGGC
LTR6B-X CAGTCCTTGA GGAAGGGGCT GCCAGTGAAA GAGCCAAGGT GGAGTCTGGC
LTR6B-YPAR1 CAGTCCTTGA GGAAGGGGCT GCCAGTGAAA GAGCCAAGGT GGAGTCTGGC
155 165 175 185 195
Junction 1 GGGGAGGGCG GGGGGTCTAG TGAAAACCAT TTTACAGAAG TAAAGTAGGC
LTR6B-X GGGGAGGGCN GGGGGTCTAG TGAAAACCAT TTTACAGAAG TAAAGTAGGC
LTR6B-YPAR1 GGGGAGGGCN GGGGGTCTAG TGAAAACCAT TTTACAGAAG TAAAGTAGGC
455 465 475 485 495
Junction 1 GGGCTCTCTT AGCTAAGGGA GAGTCCATTC AGGTGGAAAG AAGGCTAGGT
LTR6B-X NGGCTCTCTT AGCTAAGGGA GAGTCCATTC AGGTGGAAAG AAGGCTAGGT
LTR6B-YPAR1 NGGCTCTCTT AGCTAAGGGA GAGTCCATTC AGGTGGAAAG AAGGCTAGGT
205 215 225 235 245
Junction 1 AAAAAGTTAA AAGGATAAAT GGTTGCAGGA AAGTAAACAG TTCCAGGTGC
LTR6B-X AAAAAGTTAA AAGGATAAAT GGTTGCAGGA AAGTAAACAG TTCCAGGTGC
LTR6B-YPAR1 AAAAAGTTAA AAGGATAAAT GGTTGCAGGA AAGTAAACAG TTCCAGGTGC
505 515 525 535 545
Junction 1 GAGTAGAGGA AAAGGGAGAG TCTAAAAACA GGTTAGTAAA AACCAGGTTG
LTR6B-X GAGTAGAGGA AAAGGGAGAG TCTAAAAACA GGTTAGTAAA AACCAGGTTG
LTR6B-YPAR1 GAGTAGAGGA AAAGGGAGAG TCTAAAAACA GGTTAGTAAA AACCAGGTTG
255 265 275 285 295
Junction 1 AGGGGCTTTA AGACTATTAC AAGGTGATAG ACGCGNGGCT TTGGGCGTTA
LTR6B-X AGGGGCTTTA AGACTATTA---------TAG ACGCGNGGCT TTGGGCGTTA
LTR6B-YPAR1 AGGGGCTTTA AGACTATTAC AAGGTGATAG ACGCGNGGCT TTGGGCGTTA
555
Junction 1 GGCATTACA
LTR6B-X GGCATTACA
LTR6B-YPAR1 GGCATTACA

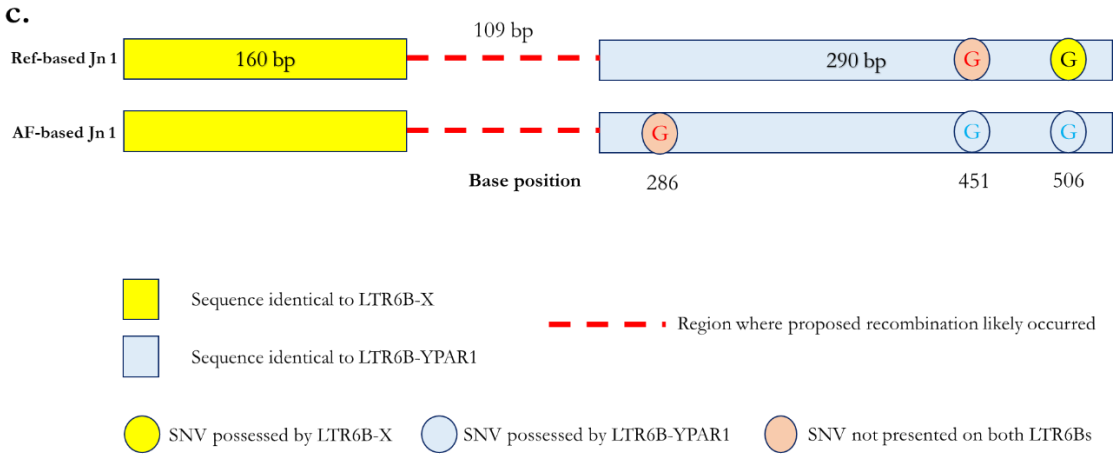


Figure 7.6 **Alignment of Junction 1 against reference sequences of both ePAR recombination-initiating LTR6Bs.** Panel **a.** and **b.** show different versions of the reference sequences such that **a)** is the original version and **b)** is the modified version taking allele frequencies into account. The reference sequence heading of LTR6B-X is highlighted with yellow while that of LTR6B-YPAR1 is highlighted with blue. Base position is based on end-to-end alignment and is placed over the sequence box. Different base pairs between both LTR6B sequences are highlighted in green and are used as a signature of each LTR6B to propose the region where crossing-over has taken place. Proposed recombinant sequence is highlighted on the base position rows, in which yellow is for the sequence identical to LTR6BX, blue is for the sequence identical to LTR6B-YPAR1, and red is for the putative recombination region. **a)** With the original reference sequence, Junction 1 shows two different variants (in red boxes), positions 451 and 506, to the reference sequence of LTR6B-YPAR1 in that the position 451 belongs to neither of the LTR6Bs whilst position 506 is identical to that of LTR6B-X. **b)** The modified reference sequences resolves two inconsistent base positions, 451 and 506, to the sequence of LTR6B-YPAR1 but presents a new different variant at the position 286 (in red box) which is identical to neither LTR6B. **c)** The simplified diagrams show a comparison of two recombinant junction structures between the alignment version against the original reference sequences and against the modified ones.

Using both versions of reference sequences of LTR6Bs, alignment against the modified version which takes allele frequencies of SNPs and indels into account shows the most parsimonious mechanism because this model can explain that recombination has taken place through only one crossover, where the different base pair at position 286 might pre-exist, whilst the model aligned against the original reference version indicates a more complicated series of events by implying double recombination events, a crossover followed by a gene conversion to change base position 506 to the same state as in LTR6B-X.

Similarly, Junction 2 was approached by using either original or modified reference sequences for both LTR6Bs. The results are presented in Figure 7.7 and it is found that using different reference sequences gives a different proposed crossover region and also different variants. Alignment against the original reference version shows the narrower possible recombinant region but leaving one possible gene conversion variant at position 506, while alignment against the modified version, though lengthening the putative crossover range, eradicates the variant.

Figure 7.7 **Alignment of Junction 2 against reference sequences of both ePAR recombination-initiating LTR6Bs.** Panel **a.**, **b.** and **c.** are presented in a similar way to Figure 7.6 but the actual total length of the Junction 2 is shorter (551 *cf.* 559 bp). **a)** shows the alignment between Junction 2 sequence against the original reference sequences that displays the variant at position 506 (in red box) identical to that of LTR6B-X. **b)** shows the alignment against the modified reference sequences and leaves no variant. Both alignments imply different putative recombinant regions with respect to presence of the signature variants of each LTR6B to flank the proposed recombinant part.

c) shows simplified diagrams between two different recombination structures by different alignments.

Using the modified versions of reference sequences proved to be the most parsimonious approach to resolve the recombination models of Junction 1 and 2. To infer the structures of the other junctions, the same approach was applied and the results are shown in Figure 7.8 and 7.9 below (details of sequences are presented in Appendix); there are six main types sharing the same putative recombinant junction while there are only two junctional length as same as either Junction 1 or Junction 2. For Junction 1, there are other three variant types, namely Junction 1.1, 1.2 and 1.3 which are different by single-nucleotide variants. Three other main types, namely Junction 3, 4 and 5 (including 5.1 – a variant of Junction 5), though sharing the same junctional length as Junction 1, are inferred to have different recombination regions: while Junction 2 has one closely-similar sequence, namely Junction 2.1, and the other one, Junction 6, is different in the recombinant region, though has the same junctional length.

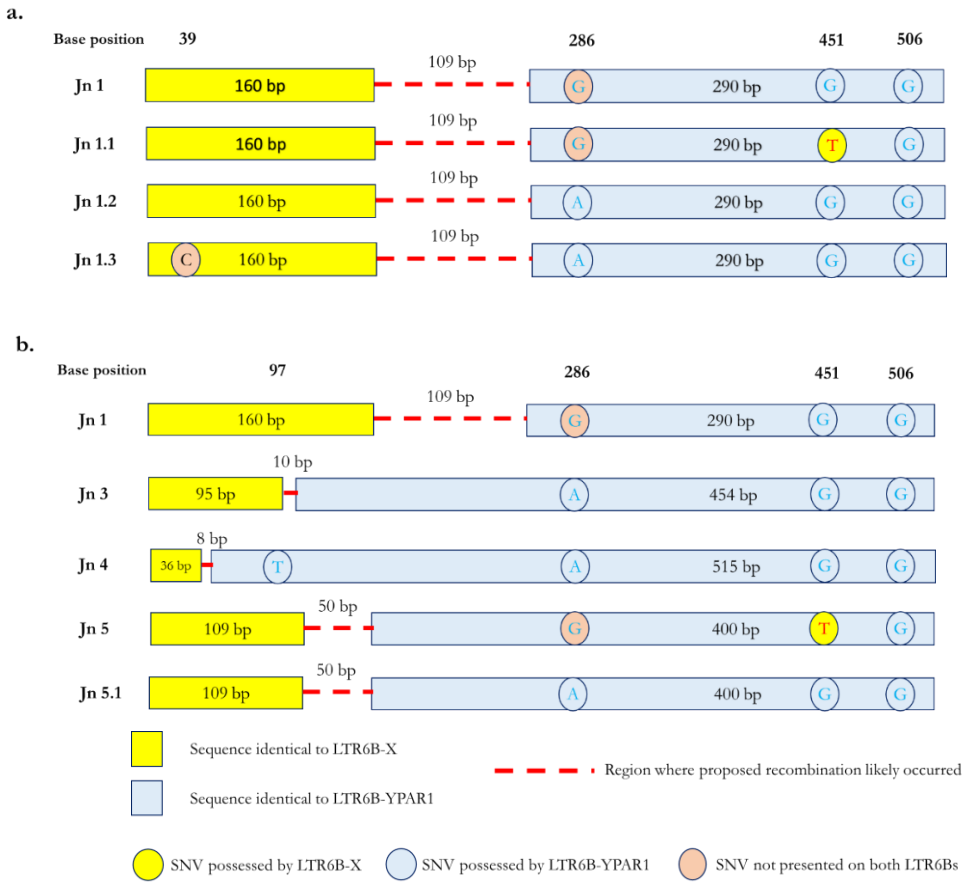


Figure 7.8 Simplified sequence structures of Junction 1 and the other types of the same length. a) shows variations resembling Junction 1 but SNPs which

define Junction 1.1, 1.2 and 1.3. **b)** shows the structures Junction 3, 4, 5 and 5.1 that are apparently different to Junction 1 in their putative recombination regions.

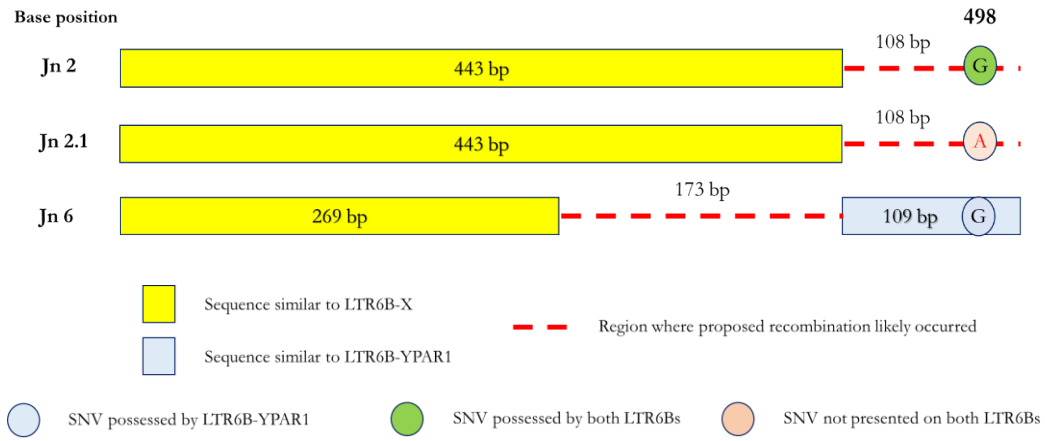


Figure 7.9 **Simplified sequence structures of Junction 2 and the other types of the same length.** Junction 2.1 is almost the same as Junction 2 except for a SNP at position 498 while Junction 6 are inferred to contain a different putative recombination region. The G allele of the SNP position 498 is depicted in different colours between Junction 2 and Junction 6 because it lies, for Junction 2, within the assumed region of recombination which is commonplace between both LTR6Bs; while in Junction 7, the region in which it is located is inferred to be the sequence of LTR6B-YPAR1.

Different recombination junction types imply independent origins of ePARs. To infer the history of ePAR origins, 165 unrelated lineages extracted from the total 172 identified-ePAR samples so far, including the ePARs previously identified from the first report [Mensah *et al.*, 2014] and from a search in the DNA collection described earlier in this thesis, were aggregated and are presented, with both recombination junction and Y-Hg information in Table 7.8. Furthermore, a MJN generated from Y-STRs for all 165 ePAR lineages is shown in Figure 7.10, correlated to the junction sequence types.

Table 7.8 Relation between Y-Hgs and recombination junction types

Hg	Junction type	No. of ePAR unrelated lineages	
		UK Biobank	Mensah *et al.* (2014) + DNA collection
I2a-L233	1	100 (inferred from sampled 55 men)	19 (from 22 men)
I2a-L1294	2	2	2 (from 3 men)
I2a-L1286*	1	0	1 (from 3 men)
I2a-Y10626	6	1	0
I1a-Z58	1.1	3	0
I1a-L338	1.1	1	0
I1c-Z17926	5	1	0
E1a-M132	1	1	0
K-M9	2.1	7	0
R1b-CTS3655	5.1	1	0
R1b-U152	1	1	0
	1.1	1	0
R1b-P311	1	3	1 (from 2 men in sub-Hg R1b-D27)
	1.2	5	0
	1.3	4	0
	2	1	0
	3	6	0
	4	1	0
	5.1	1	0
	6	2	0

* This Y-Hg may derived to a deeper branch via SNP L880.

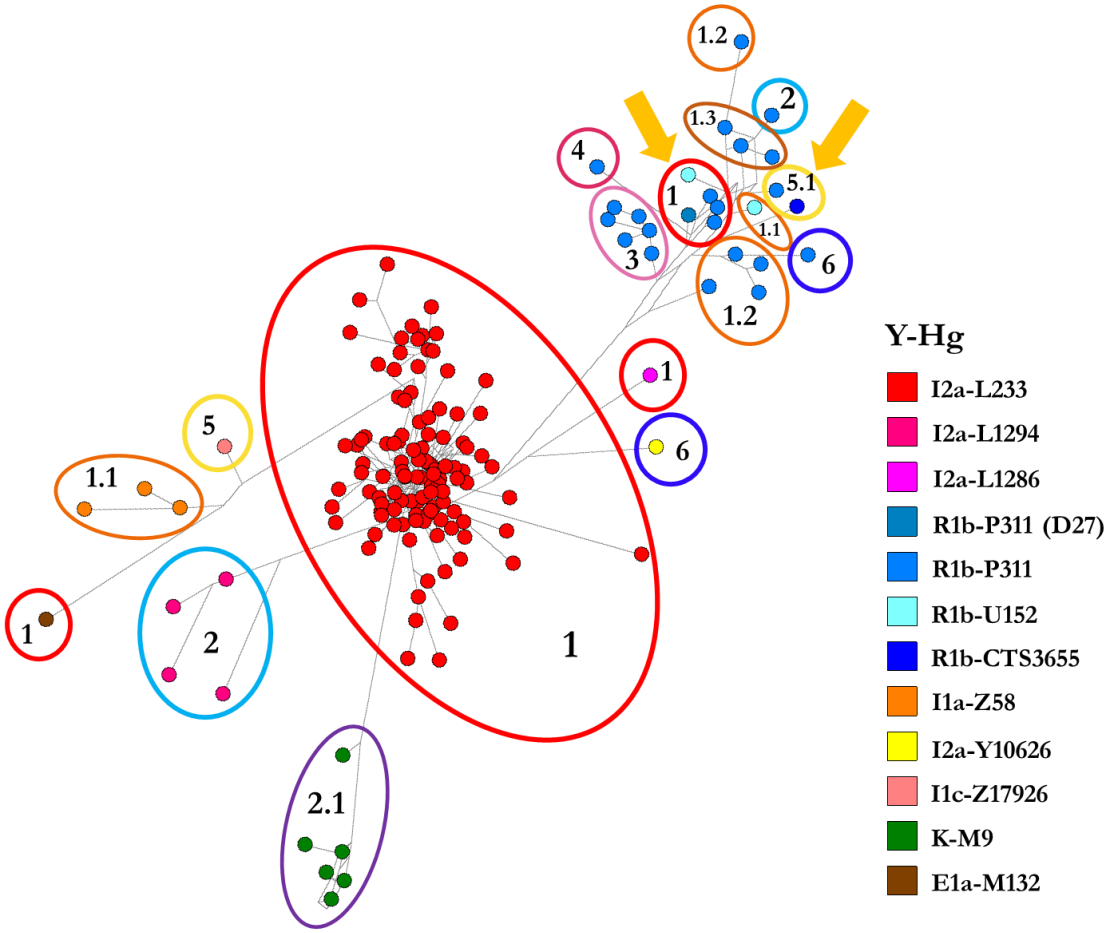


Figure 7.10 **MJN amongst ePAR men showing recombination junction diversity by Y-Hgs.** Junctional types are indicated in numbers inside lines encircled men carrying the same types. Each circle comprises of the ePAR men from same sub-Hg that can be inferred to be separated event except the Y-Hg I2a-L233 and I2a-L1286 which share the same junctional type (Junction 1) and are rooted from the same common ancestor. Two orange arrows indicate the circles corresponding to the same junctional types, each of which contains two distinct Y-Hgs. By this reasoning, considering the junctional types together with the Y-Hgs and the branches in the network, there are at least 19 independent ePAR events.

7.3 Discussion

In the initial discovery of a *de novo* rearrangement generating ePAR on the human Y chromosome [Mensah *et al.*, 2014] the survey was restricted to ~2600 males from Belgium. A small number of ePAR men (~0.5% of the sampled men) from the pioneer article [Mensah *et al.*, 2014] shows that the incidence is low, though, by virtue of observing ePARs in at least two Y-Hgs far apart in the phylogenetic tree, the generation of ePAR was shown to be recurrent, as expected for an event due to NAHR [Liu *et al.*, 2012, Ou *et al.*, 2011]. The limited number of individuals surveyed, and their limited haplogroup diversity given the population studied, means that other ePARs may have been missed. This study attempted to broaden the scope of the study in the incidence of ePAR to a larger population size, and one containing a broader range of ancestries, and hence haplogroups.

7.3.1 Incidence of ePAR in extended datasets

In the data of Mensah *et al.* (2014), ePAR was observed in only two Y-Hgs, I2a-P37.2* and R1b*, which are lineages found among Europeans. The initial attempt in this study to survey globally-reported genomic structural variants via DGV [DGV, MacDonald *et al.*, 2014] identified X-duplications, corresponding to that of the ePAR region, in men from several geographical areas (overall incidence ~0.35%). With a limited access to DNA, ePAR was only confirmed in one of the two DNA-available men who is of European origin and has Y-Hg I2a-L233, a sub-branch of I2a-P37.2*, and therefore likely identical by descent with the previously discovered major ePAR lineage. No novel ePARs were discovered. Extension of a survey to the much larger dataset of UK Biobank males revealed ePARs associated with a greater variety of Y-Hgs. Following both median SNP intensity analysis and confirmatory PCR junction testing, ePARs were observed not only in the major European Hg I2 and minor R1b lineages, but also in Hg I1, several independent sub-haplogroups of R1b, a rare Australasian (K) lineage [Karafet *et al.*, 2014], here as self-declared as Pakistanis, and an E1 lineage typical of West Africa and Caribbean. The results affirm the hypothesis of ePAR recurrence.

This study tested all of the non-I2 ePAR putative individuals for confirmation of ePAR, but only a subset of the I2 group. All I2 samples tested were confirmed as ePAR-positive. Assuming that all of the remaining I2 group are positive too, and given that about 45% of those in non-I2 Hg are the genuine carriers, the overall incidence of ePAR among UK population is approximately 0.75% (1,577/211,033). However, Y-Hg I2 constitutes almost 90% of cases.

7.3.2 History of recurrent *de novo* generation of ePAR

The presence of ePARs in diverse Hgs shows that there have been at least 19 recurrent NAHR events. In the non-I2 group, ePAR cases are found in much lower proportion per Hg (<0.05%) compared to those in the I2 group (~0.71%). This might support the hypothesis presented in Chapter 6 that the ePAR men of the Hg I2a-L233, -L1286* and –L1294 (see the tree in Figure 7.5 above) which are derived from the same node defined by S21825 might inherit their ePAR from the same origin because the I2 ePAR men were resolved to the actual Hgs by Y-STR profiles together with SNP typing to be only I-L233 (~98%) and I-L1294 (~2%). Even though the number of ePAR individuals in Hg R1b were found to be the second most frequent type, they were resolved into more sub-Hgs, at least three, carrying nine different junctional types: this suggests that R1b ePARs likely occurred independently within multiple lineages as shown in Figure 7.10 above.

Focusing on the ePAR carriers in the three I2a sub-Hgs mentioned above, it has been found that they have two junctional types, Junction 1 for I2a-L233 and I2a-L1286*, and Junction 2 for I2a-1294, as previously reported [Mensah *et al.*, 2014]. Mensah *et al.* (2014) proposed the recombination mechanisms of both cases to be the same except that a gene conversion happened later to give rise to the Junction 2 sequence; here, by taking allele frequencies of all variants on both LTR6Bs into account, it was found that both junctions might likely occur with different breakpoints, nonetheless. The most parsimonious approach could resolve Junction 2, after modifying the reference sequences of LTR6Bs based on allele frequencies, to have the recombination breakpoint proximal to that of Junction 1 and without a gene conversion occurring later.

As the ePAR men in I2a-L233 and I2a-L1286* (which is likely I2a-L880) carry the same Junction 1 and both haplogroups are rooted from the same common ancestor, they might carry the same ePAR event identical by descent; or at least, the ePAR carriers in the Hg I2a-L233 could have inherited their ePAR from the same ancestor. One supporting piece of evidence is that there are the maximum number of ePARs in this haplogroup, even though its proportion in the population is much lower than R1b. In addition, the clustering of ePAR individuals in I2a-L233 might reflect the earliest time of ePAR occurring amongst other haplogroups to allow an accumulation of the maximum number of ePAR population.

In summary, the oldest ePAR Hg so far might be Y-Hgs, I2a-L233 and perhaps including I2a-L1286* (expected to be –L880), that descend from the ancestral node I2a-L1286. If this is true, the estimated TMRCA in Chapter 6 should be recalculated; and by using the same methodology as in Chapter 6 [Fenner, 2005, Goldstein *et al.*, 1995a, Goldstein *et al.*, 1995b],

the new estimated TMRCA is 2242 ± 450 years, which equates to ~73 generations (assuming 31 yrs. per generation).

7.3.3 Other putative ePARs proven to be not authentic ePARs might represent other rearrangements

About half of the putative ePARs from the non-I2 group were shown not to carry the same chromosome rearrangement as ePAR. One explanation is that these cases may carry X-translocations larger than that of the canonical ePAR. Since the canonical ePAR is mediated via NAHR between LTR6Bs as described previously, it is possible that the translocation of X segments to create larger ePARs might occur via NAHR between different pairs of LTR6Bs as presented in the next section.

The next step in understanding the basis of these other rearrangements would be a careful analysis of intensities all X-chromosomal SNPs in order to determine the extent of translocated material, and to suggest the locations of breakpoints. In silico analysis of repeated sequences in these breakpoint intervals could suggest candidate repeats as sponsoring elements in NAHR.

7.3.4 Role of LTR6Bs in promoting NAHR in the creation of ePAR and other rearrangements resembling ePAR

Based on a BLAT genome alignment available via the UCSC Genome Browser [UCSC Genome Browser], LTR6B-X and LTR6B-PAR1 share ~97% similarity despite 15 distinguishing SNVs and two indels based on the reference sequence as previously described in Chapter 3. It is interesting to ask if other LTR6Bs or non-LTR6Bs in the genome which have similar lengths and sequences would be able to pair and potentially generate NAHR products. To answer this, the total of 499 of the transcription-factor binding-sites of LTR6Bs (TFBS-LTR6Bs) in the human genome were taken from the dbHERVs-REs database [dbHERV-REs, Ito *et al.*, 2017] representing all known LTR6Bs and two sets of sequences which have high similarity to LTR6B-X, the so-called DistX, and to LTR6B–PAR1, so-called PAR1, via BLAT [UCSC Genome Browser]. Each dataset was selected for sequences >520 bp long and a further selection applied to the other two sets, DistX and PAR1, such that the selected sequences must have >90% similarity to the LTR6Bs. The final number of the selected sequences was 81 for the TFBS and 54 for each of the DistX and the PAR1 set. Of the three datasets, 40 LTR6Bs with 90.00 - 96.60% similarity and ranging from 530 - 560 bp

was found to be commonplace among three datasets and distributed in 19 chromosomes including autosomes and sex chromosomes. The rest of sequences in each set were found to overlap to either of the other sets or only specific to their own sets. Interestingly, among the four specific LTR6B-similar sequences in the PAR1 set, two of them were LTR6A, not LTR6B. Moreover, there were 25 LTR6Bs in the TFBS set which were found not to match with either DistX or PAR1. The results are summarised in Figure 7.11 and their names are listed in Appendix.

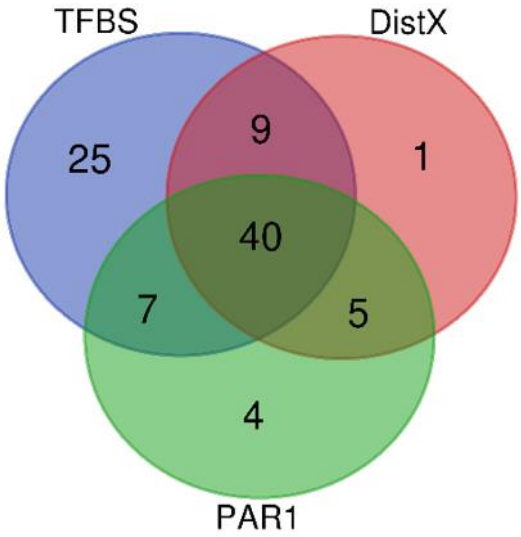


Figure 7.11 Venn diagram showing >520-bp sequences across human genome matching with LTR6B-X and PAR1. LTR6Bs were taken from transcription factor binding sites (TFBS). The sequences, either LTR6Bs or not, which had >90% similarity to LTR6B-X (DistX) or LTR6B-PAR1 (PAR1) were taken to overlie TFBS as described in text.

The results above suggest that LTR6Bs contributing to the generation of ePAR could potentially find homologies at several loci across the genome. As a result, they might induce NAHR and cause the same distal X-chromosome segment as is found on ePAR to be translocated to other genome loci, creating different recombinant junctions to the ePAR. This could be an explanation of some of the X-duplication individuals that show negative ePAR Junction-PCR results. Nonetheless, given the proximity of the two sex chromosomes due to pairing and synapsis in PAR1, it seems likely that a proportion of the X-duplication

cases could represent larger ePAR extensions. Investigating the incidence and recombination behaviour of these would be worthwhile and interesting.

7.3.5 X-deletion corresponding to the ePAR segment might be evidence of reciprocal NAHR

In the original work, the mechanism creating ePAR was described as NAHR by reciprocal translocation [Mensah *et al.*, 2014] generating one "gain" product of the ePAR Y chromosome and one "loss" product of the deleted X chromosome. Although the "loss" outcome has been evidenced, this was seen in just one event from one family [Mensah *et al.*, 2014]. In this study, the survey of UK Biobank found 40 men carrying potential X-deletions in the region corresponding to that of ePAR. Though these male carriers were found to cluster mostly in Hg R1b, there is no causal link between Y-Hg and the event since the X chromosome is independently inherited; Hg R1b is the most common Hg amongst Europeans. As described above, DNAs were collected for this group of individuals, however, since time period was limited, these individuals will be analysed in the future. If these males indeed carry deletions, they are expected to be null for two genes, *GYG2*, encoding glycogenin-2, the predominant glycogenin isoform in the liver, and *XG*, encoding a protein expressed on the red-blood-cell surface. It would be interesting to ask from clinical data available in UK Biobank if men with such deletion chromosomes shared any phenotypic features that could be ascribed to gene loss. However, the more recent report of two unrelated North European male lineages with a rare 102-kb deletion of the X chromosome (chrX:2703633–2805652, hg19) which corresponds to the X-translocated segment in ePAR and the deletion involves *XG* and *GYG2* genes: with focusing on *GYG2* gene expression, they concluded that there is no clear association with diabetes but showed a little rise of plasma glucose while there was no significant change in liver cell morphology [Irgens *et al.*, 2015].

7.4 Conclusion

NAHR generating ePAR is observed to be an active process, since it is observed in diverse Y-Hgs showing that the event is able to recur. Though clustering amongst Europeans thanks to the founder lineage in Hg I2, ePAR can be found across various ethnicities. Clustering of ePAR carriers in the Hg I2a-L233 and probably its close cousin Hg I2a-L1286*, inferred to

have derived to -L880, suggests that these Hgs might be the longest associated with ePAR, descendin from a most recent common ancestor which is ancient enough to constitute the major ePAR population today. The number of ePAR men depends mainly on the number of Y-Hg I2a-L233 (and also –L880) which is estimated to represesent ~0.5% of the UK population. The survey in UK Biobank also found a number of men carrying the X-deletion which might be reciprocal to ePAR and this might help confirm the previously-proposed mechanism of generation of the ePAR. Furthermore, a set of men in non-I2 Hgs who carry X-duplications suggestive of ePAR but shown not to have a genuine ePAR, might carry other translocations of the X onto Y which might be larger in size (including even larger ePARs) or might have different NAHR junctions. The possible recombination breakpoints might occur at LTR6Bs which are found to contain high sequence similarity in several loci across genome. However, neither the non-ePAR X-duplications nor the putative ePAR-reciprocal X-deletions have not yet been researched in depth due to time restrictions.

Chapter 8 Final discussion and future directions

In spite of occupying less than 5% of the length of the human Y chromosome, PAR1 contributes a crucial role during male meiosis as an obligatory pairing and crossover site to promote the correct segregation of the X and Y chromosomes. Failure to pair and exchange genetic material within this region is not only associated with paternally generated sex-chromosome aneuploidy but also relates to male infertility [Burgoyne *et al.*, 2009, Hall *et al.*, 2006, Shi *et al.*, 2001]. Decreased fertility in male mice is associated with disruption of sequence homology across the mouse PAR [Dumont, 2017], a good demonstration of the importance of the length sequence identity for successful X-Y paring. Given the recent discovery of that human PAR1 is polymorphic in length [Mensah *et al.*, 2014], this study sought to examine the recombination behaviour of the proximally extended 110-kb X-derived ePAR segment, and also to ask if the ePAR is recurrent and how frequent it is in the population.

8.1 Approaches to identifying recombination hotspots

A study of recombination behaviour can observe activities, locations, characters and dynamics of targeted hotspots, or survey a whole region of interest. Recombination can be characterised via linkage approaches using pedigrees or population data , or via a sperm study [Kauppi *et al.*, 2004]. Human pedigrees are generally too small to provide useful information about the fine-scale nature of recombination, while population-based linkage disequilibrium (LD) analysis reflects historical patterns of recombination rather than focusing on current activity. Sperm-based studies can illuminate the current dynamics of a hotspot at a fine scale, and observe individual-specific effects; however, until now LD information has been essential to identify a candidate target hotspot for performing sperm typing [Kauppi *et al.*, 2009]. The publication of DSB mapping data [Pratto *et al.*, 2014], which point to current regions of recombination activity, have recently provided an alternative approach.

This study is, to the author's knowledge, the first to introduce the DSB mapping approach to identify putative recombination hotspots for subsequent sperm typing. Chapter 4 shows a low correlation between regions of LD breakdown and DSB hotspots; the two sources of data were analysed in different sexes so this low correlation might be indicative of sex-specific differences in DSB induction, or alternatively the result of low-frequency *PRDM9*

alleles acting at LD-only hotspots, as acknowledged in the DSB mapping study [Pratto *et al.*, 2014]. Moreover, in refined sex-specific genetic maps fewer sex-specific recombination hotspots have been observed in comparison to the total predicted hotspots derived from LD analysis [Bhérer *et al.*, 2017].

In this study, two recombination hotspots were identified within ePAR from the overlapping of LD together with DSB hotspots, and following analysis showed concordant hotspot activities of the two targets between the sperm DNA assay and DSB maps, although they differ in magnitude. It could be speculated that DSBs may be resolved by other repair mechanisms than a crossover, which are undetected by the assay. Sperm recombination activities show high association with regions of extreme LD breakdown [Webb *et al.*, 2008], the results in this study indicate a similar association with mapped DSBs from Pratto *et al.* (2014). The higher DSB strength in the proximal region X5 also correlates with a higher observed recombination rate in this study.

Although it seems that DSB maps are useful to locate recombination hotspots, this study also reveals transmission distortion (TD) in one man, either in COs or NCOs. TD is a feature of hotspot dynamics signalling that a hotspot is doomed to disappear [Jeffreys and Neumann, 2009]. Therefore, DSB maps may correspond well to active hotspots at present but they might not always promise long-term hotspot viability.

DSBs also show a high concordance with the sperm-based recombination hotspots characterised in Chapter 5; however, we cannot definitely conclude that using the DSB-map approach to locate recombination hotspots will always precisely predict useful regions for study because there are DSB hotspots that are not matched to sperm-based hotspots, as has been observed in using the LD-based approach [Kauppi *et al.*, 2005]. It could be suggested that using both LD-based together with DSB-mapping hotspots, especially in co-localisation of both hotspots, for performing sperm-based assays would increase the success of characterising activities and dynamics of recombination hotspots, and would also be useful in comparing individual DSB hotspot activity to that observed recombination directly in sperm DNA.

8.2 Recombination activities of ePAR hotspots in the context of the fine-scale and the entire region

The fine-scale recombination assays in both ePAR hotspots analysed this study show several common characteristics of recombination hotspots such as inter-individual variation in CO

activities, transmission distortion, CO activity clustering within a 1-2 kb interval, NCO activity co-localizing with a CO cluster in very close to the centre of the hotspot, as previously described elsewhere [Berg *et al.*, 2010, Holloway *et al.*, 2006, Sarbajna *et al.*, 2012, Jeffreys and May, 2004, Jeffreys and Neumann, 2002]. Each of the hotspots in this study also shows a recombination rate higher than the median of autosomal activity, in the same way as the other sperm-based hotspots previously analysed in both PARs [Berg *et al.*, 2010, Holloway *et al.*, 2006, Jeffreys *et al.*, 2001, Jeffreys and May, 2004, Jeffreys *et al.*, 1998, Jeffreys and Neumann, 2005, Jeffreys and Neumann, 2009, Kauppi *et al.*, 2005, May *et al.*, 2002, Odenthal-Hesse *et al.*, 2014, Webb *et al.*, 2008]. However, the overall recombination rate of the X-derived part of ePAR is required to be recalculated because of the change in the inferred evolutionary history of ePAR in the Y-Hg I2a.

Chapter 6 shows the overall rate by deducing that three co-ancestral I2a sub-Hgs, i.e. I2a-L1286* (assumed to derive into –L880), I2a-L233 and I2a-L1294, though they carry different junction sequences, share the same inherited ePAR event. After re-evaluating the junction sequences in Chapter 7, it is likely that Junction 1 and 2 have different recombination breakpoints and therefore originate independently. Following TMRCA recalculation based on the hypothesis that only Y-Hg I2a-L1286* and I2a-L233 carry identical-by-descent ePARs, the minimum number of crossover events inferred from the sequences of 10 ePAR haplotypes should be 7 out of 10 after removing one I2a-L1294 haplotype (P5/F5). Therefore, the minimum recombination rate through the 110-kb region, considering 72 generations to the most recent common ancestor, is ~0.97% (i.e. $\frac{7}{72 \times 10}$), or 8.8 cM/Mb, compared to the previously estimated 0.64% or 5.8 cM/Mb. The overall rate is therefore ~8× the genome average and somewhat closer to that of the canonical PAR1 (17× genome average) [Hinch *et al.*, 2014].

8.3 Prevalence and evolutionary history of ePAR

Chapter 7 describes the results of an extended survey for ePARs from various data sources. It might be legitimate to say that ePARs in I2a-L233 and I2a-L880 (inferred to be derived from I2a-L1286*) originated from the common ancestor, so taking a phylogenetic approach to a male individual, if his Y-chromosome is in any of those Hgs, it is highly likely to carry the ePAR. The other Hgs, including the second most commonly asscociated with ePAR, R1b, appear to contain sporadic events that do not cluster by descent. Hg I-L233 and –L880 contribute ~90% of the total observed ePARs and, together with the other Hgs, take the

ePAR prevalence to ~0.75% of the UK population. It is not actually known how many men in UK are Hg I2a-L233 and –L880 but from the empirical data surveyed from UK Biobank in this study, they would be extrapolated to be <0.7% of UK population.

The accumulation of ePAR men within those two Hgs suggests that the ePAR event occurred at the most recent common ancestor of both Hgs at the earliest. The haplogroup found to have the second greatest number of ePARs, R1b, is the most frequent major lineage in the UK. Considering the sub-haplogroups of ePARs within R1b, the low within-sub-haplogroup incidence of each ePAR, and the junction sequence types, there is no evidence of any identity by descent between sub-haplogroups within R1b.

The survey of UK Biobank also confirms that ePAR is recurrent, with a minimum of 19 independent occurrences, distinguished by junctional types, phylogenetic relation and Y-Hgs, identified in total. In principle, the rate of formation of ePAR could be estimated from this minimum number of events and the time encompassed in Y phylogeny of all surveyed chromosomes.

8.4 Factors underlying generation of ePAR could generate other chromosome rearrangements

About half of the non-I2 Hgs in the UK Biobank population include individuals carrying X segmental duplications corresponding to that of the ePAR region, but these have later proven (through a junction PCR assay) not to be genuine ePAR carriers. A possible explanation could be that the duplications encompass larger segments (i.e. longer ePARs), or that the corresponding X-derived segment is translocated to another region of genome.

One possible underlying mechanism for the suggested hypothesis that the X segment is translocated to a region on another human chromosome, is NAHR promoted by LTR6Bs which flank the X-derived segment of ePAR. This is because LTR6Bs flanking the ePAR region match to several highly-similar DNA sequences across the genome. However, it is not simple to examine each of these X-duplicated individuals as the exact regions where the rearrangement occurs are not known. However, a possible initial step is to define exact flanking SNPs to see if the duplication segments are about the same or different size to that of ePAR. To locate their regions of rearrangement, fluorescent in-situ hybridization (FISH) study could be performed.

8.5 Future directions

Owing to time limitations as well as some unavailable DNA samples, there is some work to be continued from this thesis.

In terms of DSB maps and recombination activities in different regions, the other DSB clusters [Pratto *et al.*, 2014] as well as the LD-only hotspots in ePAR should be analysed via a sperm-based recombination study to ask which type shows a better correlation to the sperm recombination hotspots. Moreover, sperm crossover activities could be compared to DSB strengths between individuals and among hotspots. If it is possible to recruit more ePAR semen donors, especially carrying different *PRDM9* alleles, this would enhance the scope of the study.

Obtaining the remaining DNA samples identified to carry either the X duplications suspected to be ePAR, or the X deletions suspected to represent the reciprocal event would be useful, in order to further extend understanding of the landscape of ePAR.

On the issue of X duplication, re-evaluation of SNP intensity data of each positive individual might help confirm the actual sizes and to see if any fall into specific categories based on the extent of X duplication. Given the proximity and propensity of exchange of the X and Y chromosomes, it seems likely that at least some of these are translocations of X material onto the Y, and may represent larger ePARs that will reward future recombination analysis. As suggested above, if any X-duplications are not in this category, FISH might be the assay of choice to see the locations of these duplications in the genome.

Last but not least, X-deletions found among UK Biobank males could be examined to ask whether they represent the putative reciprocal event to ePAR. To do this, a Junction-PCR assay might be designed and if it works, Sanger-sequencing could be performed to identify the diversity of the events.

Bibliography

1000 Genomes Project Consortium 2010. A map of human genome variation from population-scale sequencing. *Nature,* 467**,** 1061-1073.

1000 Genomes Project Consortium 2012. An integrated map of genetic variation from 1,092 human genomes. *Nature,* 491**,** 56-65.

1000 GP. *1000 Genome Project* [Online]. Available: http://www.internationalgenome.org/.

1000 GP III FTP. *FTP data in Phase 3 release of the human X chromosome from 1000 Genome Project* [Online]. Available: ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/ALL.chrX.phase3_shapeit2_mvncall_integrated_v1b.20130502.genotypes.vcf.gz.

1000GP_Phase_3_FTP_X. *FTP data in Phase 3 release of the human X chromosome from 1000 Genome Project* [Online]. Available: ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/ALL.chrX.phase3_shapeit2_mvncall_integrated_v1b.20130502.genotypes.vcf.gz.

Aitken, R. J. & Marshall Graves, J. A. 2002. The future of sex. *Nature,* 415**,** 963.

Altemose, N., Noor, N., Bitoun, E., Tumian, A., Imbeault, M., Chapman, J. R., Aricescu, A. R. & Myers, S. R. 2017. A map of human PRDM9 binding provides evidence for novel behaviors of PRDM9 and other zinc-finger proteins in meiosis. *eLife,* 6**,** e28383.

Arbeithuber, B., Betancourt, A. J., Ebner, T. & Tiemann-Boege, I. 2015. Crossovers are associated with mutation and biased gene conversion at recombination hotspots. *Proc Natl Acad Sci USA,* 112**,** 2109-2114.

Ardlie, K. G., Kruglyak, L. & Seielstad, M. 2002. Patterns of linkage disequilibrium in the human genome. *Nat Rev Genet,* 3**,** 299-309.

Athey Haplogroup Predictor. *Y-Haplogroup Predictor* [Online]. Available: http://www.hprg.com/hapest5/.

Athey, T. W. 2005. Haplogroup Prediction from Y-STR Values Using an Allele Frequency Approach. *J Genet Geneal,* 1**,** 1-7.

Athey, T. W. 2006. Haplogroup prediction from Y-STR values using a Bayesian-allele-frequency approach. *J Genet Geneal,* 2**,** 34-39.

Bailey, J. A., Carrel, L., Chakravarti, A. & Eichler, E. E. 2000. Molecular evidence for a relationship between LINE-1 elements and X chromosome inactivation: the Lyon repeat hypothesis. *Proc Natl Acad Sci USA,* 97**,** 6634-6639.

Balanovsky, O. 2017. Toward a consensus on SNP and STR mutation rates on the human Y-chromosome. *Hum Genet,* 136**,** 575-590.

Balaton, B. P. & Brown, C. J. 2016. Escape artists of the X chromosome. *Trends Genet,* 32**,** 348-359.

Ballantyne, K. N., Keerl, V., Wollstein, A., Choi, Y., Zuniga, S. B., Ralf, A., Vermeulen, M., de Knijff, P. & Kayser, M. 2012. A new future of forensic Y-chromosome analysis: Rapidly mutating Y-STRs for differentiating male relatives and paternal lineages. *Forensic Science International: Genetics,* 6**,** 208-218.

Bandelt, H. J., Forster, P. & Röhl, A. 1999. Median-joining networks for inferring intraspecific phylogenies. *Mol Biol Evol,* 16**,** 37-48.

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., Lopez de Munain, A., Lopez-Parra, A. M., Loutradis, A., Milasin, J., Novelletto, A., Pamjav, H., Sajantila, A., Tolun, A., Winney, B. & Jobling, M. A. 2015. Large-scale recent

expansion of European patrilineages shown by population resequencing. *Nat Commun,* 6**,** 7152.

Baudat, F., Buard, J., Grey, C., Fledel-Alon, A., Ober, C., Przeworski, M., Coop, G. & de Massy, B. 2010. PRDM9 is a major determinant of meiotic recombination hotspots in humans and mice. *Science,* 327**,** 836-840.

Baudat, F., Imai, Y. & de Massy, B. 2013. Meiotic recombination in mammals: localization and regulation. *Nat Rev Genet,* 14**,** 794-806.

Berg, I. L., Neumann, R., Lam, K. W., Sarbajna, S., Odenthal-Hesse, L., May, C. A. & Jeffreys, A. J. 2010. PRDM9 variation strongly influences recombination hot-spot activity and meiotic instability in humans. *Nat Genet,* 42**,** 859-863.

Berg, I. L., Neumann, R., Sarbajna, S., Odenthal-Hesse, L., Butler, N. J. & Jeffreys, A. J. 2011. Variants of the protein PRDM9 differentially regulate a set of human meiotic recombination hotspots highly active in African populations. *Proc Natl Acad Sci USA,* 108**,** 12378–12383.

Berletch, J. B., Yang, F., Xu, J., Carrel, L. & Disteche, C. M. 2011. Genes that escape from X inactivation. *Hum Genet,* 130**,** 237-245.

Bhérer, C., Campbell, C. L. & Auton, A. 2017. Refined genetic maps reveal sexual dimorphism in human meiotic recombination at multiple scales. *Nat Commun,* 8**,** 14994.

Bickmore, W. A. & Cooke, H. J. 1987. Evolution of homologous sequences on the human X and Y chromosomes, outside of the meiotic pairing segment. *Nucleic Acids Res,* 15**,** 6261-6271.

Blanco, P., Shlumukova, M., Sargent, C. A., Jobling, M. A., Affara, N. & Hurles, M. E. 2000. Divergent outcomes of intra-chromosomal recombination on the human Y chromosome: male infertility and recurrent polymorphism. *J Med Genet,* 37**,** 752-758.

Bosch, E., Hurles, M. E., Navarro, A. & Jobling, M. A. 2004. Dynamics of a human interparalog gene conversion hotspot. *Genome Res,* 14**,** 835-844.

Bosch, E. & Jobling, M. A. 2003. Duplications of the *AZF*a region of the human Y chromosome are mediated by homologous recombination between HERVs and are compatible with male fertility. *Hum Mol Genet,* 12**,** 341-347.

Briggs, S. F. & Reijo Pera, R. A. 2014. X chromosome inactivation: recent advances and a look forward. *Curr Opin Genet Dev,* 28**,** 78-82.

Brown, C. J., Ballabio, A., Rupert, J. L., Lafreniere, R. G., Grompe, M., Tonlorenzi, R. & Willard, H. F. 1991. A gene from the region of the human X inactivation centre is expressed exclusively from the inactive X chromosome. *Nature,* 349**,** 38-44.

Brown, W. R. 1988. A physical map of the human pseudoautosomal region. *EMBO J,* 7**,** 2377-2385.

Brunschwig, H., Levi, L., Ben-David, E., Williams, R. W., Yakir, B. & Shifman, S. 2012. Fine-scale maps of recombination rates and hotspots in the mouse genome. *Genetics,* 191**,** 757-764.

Buckle, V., Mondella, C., Darling, S., I.W., C. & P.N., G. 1985. Homologous expressed genes in the human sex chromosome pairing region. *Nature,* 317**,** 739-741.

Burgoyne, P. S. 1982. Genetic homology and crossing over in the X and Y chromosomes of Mammals. *Hum Genet,* 61**,** 85-90.

Burgoyne, P. S., Mahadevaiah, S. K. & Turner, J. M. A. 2009. The consequences of asynapsis for mammalian meiosis. *Nat Rev Genet,* 10**,** 207-216.

Chen, B., Cole, J. W. & Grond-Ginsbach, C. 2017. Departure from Hardy Weinberg Equilibrium and Genotyping Error. *Front Genet,* 8**,** 167.

Ciccodicola, A., D'Esposito, M., Esposito, T., Gianfrancesco, F., Migliaccio, C., Miano, M. G., Matarazzo, M. R., Vacca, M., Franzè, A., Cuccurese, M., Cocchia, M., Curci, A., Terracciano, A., Torino, A., Cocchia, S., Mercadante, G., Pannone, E., Archidiacono, N., Rocchi, M., Schlessinger, D. & D'Urso, M. 2000. Differentially regulated and

evolved genes in the fully sequenced Xq/Yq pseudoautosomal region. *Hum Mol Genet,* 9**,** 395-401.

Cooke, H. J., Brown, W. R. & Rappold, G. A. 1985. Hypervariable telomeric sequences from the human sex chromosomes are pseudoautosomal. *Nature,* 317**,** 687-692.

Cordaux, R. & Batzer, M. A. 2009. The impact of retrotransposons on human genome evolution. *Nat Rev Genet,* 10**,** 691-703.

Cotter, D. J., Brotman, S. M. & Wilson Sayres, M. A. 2016. Genetic Diversity on the Human X Chromosome Does Not Support a Strict Pseudoautosomal Boundary. *Genetics,* 203**,** 485-492.

Daniel, G. 2013. *Human Blood Groups,* West Sussex, UK, John Wiley & Sons.

Dannemann, M. & Kelso, J. 2017. The Contribution of Neanderthals to Phenotypic Variation in Modern Humans. *Am J Hum Genet,* 101**,** 578-589.

Dapper, A. L. & Payseur, B. A. 2017. Connecting theory and data to understand recombination rate evolution. *Philos Trans R Soc Lond B Biol Sci,* 372**,** 20160469.

Davies, B., Hatton, E., Altemose, N., Hussin, J. G., Pratto, F., Zhang, G., Hinch, A. G., Moralli, D., Biggs, D., Diaz, R., Preece, C., Li, R., Bitoun, E., Brick, K., Green, C. M., Camerini-Otero, R. D., Myers, S. R. & Donnelly, P. 2016. Re-engineering the zinc fingers of PRDM9 reverses hybrid sterility in mice. *Nature,* 530**,** 171-176.

dbHERV-REs. *HERV/LTR regulatory element database* [Online]. Available: http://herv-tfbs.com.

de Massy, B. 2013. Initiation of meiotic recombination: how and where? Conservation and specificities among eukaryotes. *Annu Rev Genet,* 47**,** 563-599.

DGV. *Database of Genomic Variants* [Online]. Available: http://dgv.tcag.ca/dgv/app/home.

Dumont, B. L. 2017. Meiotic Consequences of Genetic Divergence Across the Murine Pseudoautosomal Region. *Genetics,* 205**,** 1089-1100.

Eberle, M. A., Rieder, M. J., Kruglyak, L. & Nickerson, D. A. 2006. Allele frequency matching between SNPs reveals an excess of linkage disequilibrium in genic regions of the human genome. *PLoS Genet,* 2**,** e142.

Ellis, N. A., Goodfellow, P. J., Pym, B., Smith, M., Palmer, M., Frischauf, A. M. & Goodfellow, P. N. 1989. The pseudoautosomal boundary in man is defined by an Alu repeat sequence inserted on the Y chromosome. *Nature,* 337**,** 81-84.

Ellis, N. A., Ye, T. Z., Patton, S., German, J., Goodfellow, P. N. & Weller, P. 1994. Cloning of PBDX, an MIC2-related gene that spans the pseudoautosomal boundary on chromosome Xp. *Nat Genet,* 6**,** 394-400.

Ensembl Data Slicer. *Ensembl GRCh37 Genome Browser* [Online]. Available: http://grch37.ensembl.org/Homo_sapiens/Tools/DataSlicer?db=core.

Ewing, B. & Green, P. 1998. Base-calling of automated sequencer traces using phred. II. Error probabilities. *Genome Res,* 8**,** 186-194.

Ewing, B., Hillier, L., Wendl, M. C. & Green, P. 1998. Base-calling of automated sequencer traces using phred. I. Accuracy assessment. *Genome Res,* 8**,** 175-185.

Faisal, I. & Kauppi, L. 2016. Sex chromosome recombination failure, apoptosis, and fertility in male mice. *Chromosoma,* 125**,** 227-235.

Felsenstein, J. 1974. The evolutionary advantage of recombination. *Genetics,* 78**,** 737-756.

Fenner, J. N. 2005. Cross-cultural estimation of the human generation interval for use in genetics-based population divergence studies. *Am J Phys Anthropol,* 128**,** 415-423.

Ferguson-Smith, M. A., Sanger, R., Tippett, P., Aitken, D. A. & Boyd, E. 1982. A familial t(X;Y) translocation which assigns the Xg blood group locus to the region Xp22.3->pter. *Cytogenet Cell Genet,* 32**,** 273-274.

Filatov, D. A. & Gerrard, D. T. 2003. High mutation rates in human and ape pseudoautosomal genes. *Gene,* 317**,** 67-77.

Flaquer, A., Rappold, G. A., Wienker, T. F. & Fischer, C. 2008. The human pseudoautosomal regions: a review for genetic epidemiologists. *Eur J Hum Genet,* 16**,** 771-779.

Forster, P., Harding, R., Torroni, A. & Bandelt, H. 1996. Origin and evolution of native American mtDNA variation : a reappraisal. *Am J Hum Genet,* 59**,** 935-945.

Gabriel-Robez, O., Rumpler, Y., Ratomponirina, C., Petit, C., Levilliers, J., Croquette, M. F. & Couturier, J. 1990. Deletion of the pseudoautosomal region and lack of sex-chromosome pairing at pachytene in two infertile men carrying an X;Y translocation. *Cytogenet Cell Genet,* 54**,** 38-42.

Galupa, R. & Heard, E. 2015. X-chromosome inactivation: new insights into cis and trans regulation. *Curr Opin Genet Dev,* 31**,** 57-66.

Gläser, B., Grützner, F., Taylor, K., Schiebel, K., Meroni, G., Tsioupra, K., Pasantes, J., Rietschel, W., Toder, R., Willmann, U., Zeitler, S., Yen, P., Ballabio, A., Rappold, G. & Schempp, W. 1997. Comparative mapping of Xp22 genes in hominoids – evolutionary linear instability of their Y homologues. *Chromosome Res,* 5**,** 167-176.

Goldstein, D. B., Ruiz Linares, A., Cavalli-Sforza, L. L. & Feldman, M. W. 1995a. An evaluation of genetic distances for use with microsatellite loci. *Genetics,* 139**,** 463-471.

Goldstein, D. B., Ruiz Linares, A., Cavalli-Sforza, L. L. & Feldman, M. W. 1995b. Genetic absolute dating based on microsatellites and the origin of modern humans. *Proc Natl Acad Sci USA,* 92**,** 6723–6727.

Graffelman, J. & Weir, B. S. 2016. Testing for Hardy-Weinberg equilibrium at biallelic genetic markers on the X chromosome. *Heredity,* 116**,** 558-568.

Graves, J. A., Koina, E. & Sankovic, N. 2006. How the gene content of human sex chromosomes evolved. *Curr Opin Genet Dev,* 16**,** 219-224.

Graves, J. A., Wakefield, M. J. & Toder, R. 1998. The origin and evolution of the pseudoautosomal regions of human sex chromosomes. *Hum Mol Genet,* 7**,** 1991-1996.

Hall, H., Hunt, P. & Hassold, T. 2006. Meiosis and sex chromosome aneuploidy: how meiotic errors cause aneuploidy; how aneuploidy causes meiotic errors. *Curr Opin Genet Dev,* 16**,** 323-329.

Hedrick, P. W. 1987. Gametic disequilibrium measures: proceed with caution. *Genetics,* 117**,** 331-341.

Hinch, A. G., Altemose, N., Noor, N., Donnelly, P. & Myers, S. R. 2014. Recombination in the human Pseudoautosomal region PAR1. *PLoS Genet,* 10**,** e1004503.

Holloway, K., Lawson, V. E. & Jeffreys, A. J. 2006. Allelic recombination and de novo deletions in sperm in the human beta-globin gene region. *Hum Mol Genet,* 15**,** 1099-1111.

Hughes, J. F. & Coffin, J. M. 2004. Human endogenous retrovirus K solo-LTR formation and insertional polymorphisms: implications for human and viral evolution. *Proc Natl Acad Sci U S A,* 101**,** 1668-1672.

Hughes, J. F. & Coffin, J. M. 2005. Human endogenous retroviral elements as indicators of ectopic recombination events in the primate genome. *Genetics,* 171**,** 1183-1194.

International HapMap Consortium 2007. A second generation human haplotype map of over 3.1 million SNPs. *Nature,* 449**,** 851-861.

Irgens, H. U., Fjeld, K., Johansson, B. B., Ringdal, M., Immervoll, H., Leh, S., Søvik, O., Johansson, S., Molven, A. & Njølstad, P. R. 2015. Glycogenin-2 Is Dispensable for Liver Glycogen Synthesis and Glucagon-Stimulated Glucose Release. *J Clin Endocrinol Metab,* 100**,** E767-E775.

ISOGG. *ISOGG Y-DNA* [Online]. Available: http://www.isogg.org/tree/index.html.

Ito, J., Sugimoto, R., Nakaoka, H., Yamada, S., Kimura, T., Hayano, T. & Inoue, I. 2017. Systematic identification and characterization of regulatory elements derived from human endogenous retroviruses. *PLoS Genet,* 13**,** e1006883.

Jeffreys, A. J., Cotton, V. E., Neumann, R. & Lam, K. W. G. 2013. Recombination regulator PRDM9 influences the instability of its own coding sequence in humans. *Proc Natl Acad Sci USA,* 110**,** 600–605.

Jeffreys, A. J., Holloway, J. K., Kauppi, L., May, C. A., Neumann, R., Slingsby, M. T. & Webb, A. J. 2004. Meiotic recombination hot spots and human DNA diversity. *Philos Trans R Soc Lond B Biol Sci,* 359**,** 141-152.

Jeffreys, A. J., Kauppi, L. & Neumann, R. 2001. Intensely punctate meiotic recombination in the class II region of the major histocompatibility complex. *Nat Genet,* 29**,** 217-222.

Jeffreys, A. J. & May, C. A. 2004. Intense and highly localized gene conversion activity in human meiotic crossover hot spots. *Nat Genet,* 36**,** 151-156.

Jeffreys, A. J., Murray, J. & Neumann, R. 1998. High-resolution mapping of crossovers in human sperm defines a minisatellite-associated recombination hotspot. *Mol Cell,* 2**,** 267-273.

Jeffreys, A. J. & Neumann, R. 2002. Reciprocal crossover asymmetry and meiotic drive in a human recombination hot spot. *Nat Genet,* 31**,** 267-271.

Jeffreys, A. J. & Neumann, R. 2005. Factors influencing recombination frequency and distribution in a human meiotic crossover hotspot. *Hum Mol Genet,* 14**,** 2277-2287.

Jeffreys, A. J. & Neumann, R. 2009. The rise and fall of a human recombination hot spot. *Nat Genet,* 41**,** 625-629.

Jeffreys, A. J., Neumann, R., Panayi, M., Myers, S. & Donnelly, P. 2005. Human recombination hot spots hidden in regions of strong marker association. *Nat Genet,* 37**,** 601-6.

Jern, P. & Coffin, J. M. 2008. Effects of retroviruses on host genome function. *Annu Rev Genet,* 42**,** 709-732.

Jobling, M. A., Pandya, A. & Tyler-Smith, C. 1997. The Y chromosome in forensic analysis and paternity testing. *Int J Legal Med,* 110**,** 118-124.

Jobling, M. A. & Tyler-Smith, C. 2003. The human Y chromosome: an evolutionary marker comes of age. *Nat Rev Genet,* 4**,** 598-612.

Jobling, M. A. & Tyler-Smith, C. 2017. Human Y-chromosome variation in the genome-sequencing era. *Nat Rev Genet,* 18**,** 485-497.

Kaessmann, H., Zöllner, S., Gustafsson, A. C., Wiebe, V., Laan, M., Lundeberg, J., Uhlén, M. & Pääbo, S. 2002. Extensive linkage disequilibrium in small human populations in Eurasia. *Am J Hum Genet,* 70**,** 673-685.

Kamp, C., Hirschmann, P., Voss, H., Huellen, K. & Vogt, P. H. 2000. Two long homologous retroviral sequence blocks in proximal Yq11 cause AZFa microdeletions as a result of intrachromosomal recombination events. *Hum Mol Genet,* 9**,** 2563-2572.

Karafet, T. M., Mendez, F. L., Meilerman, M. B., Underhill, P. A., Zegura, S. L. & Hammer, M. F. 2008. New binary polymorphisms reshape and increase resolution of the human Y chromosomal haplogroup tree. *Genome Res,* 18**,** 830-838.

Karafet, T. M., Mendez, F. L., Sudoyo, H., Lansing, J. S. & Hammer, M. F. 2014. Improved phylogenetic resolution and rapid diversification of Y-chromosome haplogroup K-M526 in Southeast Asia. *Eur J Hum Genet,* 23**,** 369-373.

Kauppi, L., Jasin, M. & Keeney, S. 2012. The tricky path to recombining X and Y chromosomes in meiosis. *Ann NY Acad Sci,* 1267**,** 18-23.

Kauppi, L., Jeffreys, A. J. & Keeney, S. 2004. Where the crossovers are: recombination distributions in mammals. *Nat Rev Genet,* 5**,** 413-424.

Kauppi, L., May, C. A. & Jeffreys, A. J. 2009. Analysis of meiotic recombination products from human sperm. *In:* KEENEY, S. (ed.) *Meiosis, Volume 1, Molecular and Genetic Methods.* 1st ed. New York, USA, Humana Press.

Kauppi, L., Stumpf, M. P. & Jeffreys, A. J. 2005. Localized breakdown in linkage disequilibrium does not always predict sperm crossover hot spots in the human MHC class II region. *Genomics,* 86**,** 13-24.

Kayser, M. 2017. Forensic use of Y-chromosome DNA: a general overview. *Hum Genet,* 136**,** 621-635.

King, T. E., Ballereau, S. J., Schürer, K. E. & Jobling, M. A. 2006. Genetic signatures of coancestry within surnames. *Curr Biol,* 16**,** 384-388.

Kojima, K. K. 2018. Human transposable elements in Repbase: genomic footprints from fish to humans. *Mob DNA,* 9**,** 2.

Korol, A. B. & Iliadi, K. G. 1994. Increased recombination frequencies resulting from directional selection for geotaxis in Drosophila. *Heredity,* 72**,** 64-68.

Krag, T. O., Ruiz-Ruiz, C. & Vissing, J. 2017. Glycogen Synthesis in Glycogenin 1–Deficient Patients: A Role for Glycogenin 2 in Muscle. *J Clin Endocrinol Metab,* 102**,** 2690-2700.

Kvaløy, K., Galvagni, F. & Brown, W. R. 1994. The sequence organization of the long arm pseudoautosomal region of the human sex chromosomes. *Hum Mol Genet,* 3**,** 771-778.

Laan, M., Wiebe, V., Khusnutdinova, E., Remm, M. & Paabo, S. 2005. X-chromosome as a marker for population history: linkage disequilibrium and haplotype study in Eurasian populations. *Eur J Hum Genet,* 13**,** 452-462.

Lahn, B. T. & Page, D. C. 1999. Four Evolutionary Strata on the Human X Chromosome. *Science,* 286**,** 964-967.

Lange, J., Yamada, S., Tischfield, S. E., Pan, J., Kim, S., Zhu, X., Socci, N. D., Jasin, M. & Keeney, S. 2016. The Landscape of Mouse Meiotic Double-Strand Break Formation, Processing, and Repair. *Cell,* 167**,** 695-708 e16.

Levy, S., Sutton, G., Ng, P. C., Feuk, L., Halpern, A. L., Walenz, B. P., Axelrod, N., Huang, J., Kirkness, E. F., Denisov, G., Lin, Y., MacDonald, J. R., Pang, A. W., Shago, M., Stockwell, T. B., Tsiamouri, A., Bafna, V., Bansal, V., Kravitz, S. A., Busam, D. A., Beeson, K. Y., McIntosh, T. C., Remington, K. A., Abril, J. F., Gill, J., Borman, J., Rogers, Y., Frazier, M. E., Scherer, S. W., Strausberg, R. L. & Venter, J. C. 2007. The Diploid Genome Sequence of an Individual Human. *PLos Biol,* 5**,** e254.

Lewontin, R. C. 1964. The interaction of selection and linkage. I. General considerations; Heterotic models. *Genetics,* 49**,** 49-67.

Lien, S., Szyda, J., Schechinger, B., Rappold, G. & N., A. 2000. Evidence for heterogeneity in recombination in the human pseudoautosomal region: high resolution analysis by sperm typing and radiation-hybrid mapping. *Am J Hum Genet,* 66**,** 557-566.

Liu, P., Carvalho, C. M. B., Hastings, P. J. & Lupski, J. R. 2012. Mechanisms for recurrent and complex human genomic rearrangements. *Curr Opin Genet Dev,* 22**,** 211-220.

Loh, P. R., Genovese, G., Handsaker, R. E., Finucane, H. K., Reshef, Y. A., Palamara, P. F., Birmann, B. M., Talkowski, M. E., Bakhoum, S. F., McCarroll, S. A. & Price, A. L. 2018. Insights into clonal haematopoiesis from 8,342 mosaic chromosomal alterations. *Nature,* 559**,** 350-355.

Lyon, M. F. 1999. X-chromosome inactivation. *Curr Biol,* 9**,** R235-237.

MacDonald, J. R., Ziman, R., Yuen, R. K. C., Feuk, L. & Scherer, S. W. 2014. The Database of Genomic Variants: a curated collection of structural variation in the human genome. *Nucleic Acids Res,* 42**,** D986-D992.

Martin, G., Otto, S. P. & Lenormand, T. 2006. Selection for recombination in structured populations. *Genetics,* 172**,** 593-609.

Martiniano, R., Cassidy, L. M., Ó'Maoldúin, R., McLaughlin, R., Silva, N. M., Manco, L., Fidalgo, D., Pereira, T., Coelho, M. J., Serra, M., Burger, J., Parreira, R., Moran, E., Valera, A. C., Porfirio, E., Boaventura, R., Silva, A. M. & Bradley, D. G. 2017. The population genomics of archaeological transition in west Iberia: Investigation of ancient substructure using imputation and haplotype-based methods. *PLoS Genet,* 13**,** e1006852.

May, C. A., Shone, A. C., Kalaydjieva, L., Sajantila, A. & Jeffreys, A. J. 2002. Crossover clustering and rapid decay of linkage disequilibrium in the Xp/Yp pseudoautosomal gene SHOX. *Nat Genet,* 31**,** 272-275.

Mensah, M. A., Hestand, M. S., Larmuseau, M. H., Isrie, M., Vanderheyden, N., Declercq, M., Souche, E. L., Van Houdt, J., Stoeva, R., Van Esch, H., Devriendt, K., Voet, T., Decorte, R., Robinson, P. N. & Vermeesch, J. R. 2014. Pseudoautosomal region 1 length polymorphism in the human population. *PLoS Genet,* 10**,** e1004578.

Mhlanga, M. M., Malmberg, L. 2001. Using molecular beacons to detect single-nucleotide polymorphisms with real-time PCR. *Methods,* 25**,** 463-471.

Mohandas, T. K., Speed, R. M., Passage, M. B., Yen, P. H., Chandley, A. C. & Shapiro, L. J. 1992. Role of the pseudoautosomal region in sex-chromosome pairing during male meiosis: meiotic studies in a man with a deletion of distal Xp. *Am J Hum Genet,* 51**,** 526-533.

Morris, B. J. & Mangs, A. H. 2007. The human pseudoautosomal region (PAR): origin, function and future. *Curr Genomics,* 8**,** 129-136.

Morton, N. E. 1955. Sequential tests for the detection of linkage. *Am J Hum Genet,* 7**,** 277-318.

Mu, J., Skurat, A. V. & Roach, P. J. 1997. Glycogenin-2, a Novel Self-glucosylating Protein Involved in Liver Glycogen Biosynthesis. *J Biol Chem,* 272**,** 27589-27597.

Myers, S., Bottolo, L., Freeman, C., McVean, G. & Donnelly, P. 2005. A fine-scale map of recombination rates and hotspots across the human genome. *Science,* 310**,** 321-324.

Myers, S., Bowden, R., Tumian, A., Bontrop, R. E., Freeman, C., MacFie, T. S., McVean, G. & Donnelly, P. 2010. Drive against hotspot motifs in primates implicates the PRDM9 gene in meiotic recombination. *Science,* 327**,** 876-879.

Myers, S., Freeman, C., Auton, A., Donnelly, P. & McVean, G. 2008. A common sequence motif associated with recombination hot spots and genome instability in humans. *Nat Genet,* 40**,** 1124-1129.

NCBIdbSNP. *NCBI dbSNP* [Online]. Available: http://www.ncbi.nlm.nih.gov/SNP/.

Nelson, P. N., Carnegie, P. R., Martin, J., Ejtehadi, H. D., Hooley, P., Roden, D., Rowland-Jones, S., Warren, P., Astley, J. & Murray, P. G. 2003. Demystified . . . Human endogenous retroviruses. *J Clin Pathol: Mol Pathol,* 56**,** 11-18.

Odenthal-Hesse, L., Berg, I. L., Veselis, A., Jeffreys, A. J. & May, C. A. 2014. Transmission distortion affecting human noncrossover but not crossover recombination: a hidden source of meiotic drive. *PLoS Genet,* 10**,** e1004106.

Ohno, S. 1967. *Sex Chromosomes and Sex-linked Genes,* New York, USA, Springer-Verlag.

Otto, S. P., Pannell, J. R., Peichel, C. L., Ashman, T. L., Charlesworth, D., Chippindale, A. K., Delph, L. F., Guerrero, R. F., Scarpino, S. V. & McAllister, B. F. 2011. About PAR: the distinct evolutionary dynamics of the pseudoautosomal region. *Trends Genet,* 27**,** 358-367.

Ou, Z., Stankiewicz, P., Xia, Z., Breman, A. M., Dawson, B., Wiszniewska, J., Szafranski, P., Cooper, M. L., Rao, M., Shao, L., South, S. T., Coleman, K., Fernhoff, P. M., Deray, M. J., Rosengren, S., Roeder, E. R., Enciso, V. B., Chinault, A. C., Patel, A., Kang, S. L., Shaw, C. A., Lupski, J. R. & Cheung, S. W. 2011. Observation and prediction of recurrent human translocations mediated by NAHR between nonhomologous chromosomes. *Genome Res,* 21**,** 33-46.

Pang, A. W., Migita, O., MacDonald, J. R., Feuk, L. & Scherer, S. W. 2013. Mechanisms of Formation of Structural Variation in a Fully Sequenced Human Genome. *Human Mutation,* 34**,** 345-354.

Parvanov, E. D., Petkov, P. M. & Paigen, K. 2010. PRDM9 controls activation of mammalian recombination hotspots. *Science,* 327**,** 835.

Pinto, D., Marshall, C., Feuk, L. & Scherer, S. W. 2007. Copy-number variation in control population cohorts. *Hum Mol Genet,* 16**,** R168-R173.

Poznik, G. D. 2016. Identifying Y-chromosome haplogroups in arbitrarily large samples of sequenced or genotyped men. *bioRxiv***,** 088716.

Pratto, F., Brick, K., Khil, P., Smagulova, F., Petukhova, G. V. & Camerini-Otero, R. D. 2014. DNA recombination. Recombination initiation maps of individual human genomes. *Science,* 346**,** 1256442.

Rappold, G. A. 1993. The pseudoautosomal regions of the human sex chromosomes. *Hum Genet,* 92**,** 315-324.

Redon, R., Ishikawa, S., Fitch, K. R., Feuk, L., Perry, G. H., Andrews, T. D., Fiegler, H., Shapero, M. H., Carson, A. R., Chen, W., Cho, E. K., Dallaire, S., Freeman, J. L., Gonzalez, J. R., Gratacos, M., Huang, J., Kalaitzopoulos, D., Komura, D., MacDonald, J. R., Marshall, C. R., Mei, R., Montgomery, L., Nishimura, K., Okamura, K., Shen, F., Somerville, M. J., Tchinda, J., Valsesia, A., Woodwark, C., Yang, F., Zhang, J., Zerjal, T., Zhang, J., Armengol, L., Conrad, D. F., Estivill, X., Tyler-Smith, C., Carter, N. P., Aburatani, H., Lee, C., Jones, K. W., Scherer, S. W. & Hurles, M. E. 2006. Global variation in copy number in the human genome. *Nature,* 444**,** 444-454.

Repbase. *GIRI Repbase Update* [Online]. Available: https://www.girinst.org/protected/repbase_extract.php?access=LTR6B&format=EMBL.

Ross, M. T., Grafham, D. V., Coffey, A. J., Scherer, S., McLay, K., Muzny, D. & et al. 2005. The DNA sequence of the human X chromosome. *Nature,* 434**,** 325–337.

Roze, D. & Barton, N. H. 2006. The Hill-Robertson effect and the evolution of recombination. *Genetics,* 173**,** 1793-1811.

Sanger, R. & Race, R. 1975. *Blood Groups in Man,* Oxford, UK, Wiley-Blackwell.

Sarbajna, S., Denniff, M., Jeffreys, A. J., Neumann, R., Soler Artigas, M., Veselis, A. & May, C. A. 2012. A major recombination hotspot in the XqYq pseudoautosomal region gives new insight into processing of human gene conversion events. *Hum Mol Genet,* 21**,** 2029-2038.

Shaikh, T. H., Gai, X., Perin, J. C., Glessner, J. T., Xie, H., Murphy, K., O'Hara, R., Casalunovo, T., Conlin, L. K., D'Arcy, M., Frackelton, E. C., Geiger, E. A., Haldeman-Englert, C., Imielinski, M., Kim, C. E., Medne, L., Annaiah, K., Bradfield, J. P., Dabaghyan, E., Eckert, A., Onyiah, C. C., Ostapenko, S., Otieno, F. G., Santa, E., Shaner, J. L., Skraban, R., Smith, R. M., Elia, J., Goldmuntz, E., Spinner, N. B., Zackai, E. H., Chiavacci, R. M., Grundmeier, R., Rappaport, E. F., Grant, S. F. A., White, P. S. & Hakonarson, H. 2009. High-resolution mapping and analysis of copy number variations in the human genome: A data resource for clinical and research applications. *Genome Res,* 19**,** 1682-1690.

Shi, Q., Spriggs, E., Field, L. L., Ko, E., Barclay, L. & Martin, R. H. 2001. Single sperm typing demonstrates that reduced recombination is associated with the production of aneuploid 24,XY human sperm. *Am J Med Genet,* 99**,** 34-38.

Sinclair, A. H., Berta, P., Palmer, M. S., Hawkins, J. R., Griffiths, B. L., Smith, M. J., Foster, J. W., Frischauf, A. M., Lovell-Badge, R. & Goodfellow, P. N. 1990. A gene from the human sex-determining region encodes a protein with homology to a conserved DNA-binding motif. *Nature,* 346**,** 240-244.

Skaletsky, H., Kuroda-Kawaguchi, T., Minx, P. J., Cordum, H. S., Hillier, L., Brown, L. G. & , e. a. 2003. The male-specific region of the human Y chromosome is a mosaic of discrete sequence classes. *Nature,* 423**,** 825-837.

Slatkin, M. 2008. Linkage disequilibrium--understanding the evolutionary past and mapping the medical future. *Nat Rev Genet,* 9**,** 477-485.

Stabellini, R., Vasques, L. R., de Mello, J. C., Hernandes, L. M. & Pereira, L. V. 2009. MAOA and GYG2 are submitted to X chromosome inactivation in human fibroblasts. *Epigenetics,* 4**,** 388-393.

Stephens, M. & Donnelly, P. 2003. A comparison of Bayesian methods for haplotype reconstruction from population genotype data. *Am J Hum Genet,* 73**,** 1162-1169.

Stephens, M., Smith, N. J. & Donnelly, P. 2001. A new statistical method for haplotype reconstruction from population data. *Am J Hum Genet,* 68**,** 978-989.

Sudlow, C., Gallacher, J., Allen, N., Beral, V., Burton, P., Danesh, J., Downey, P., Elliott, P., Green, J., Landray, M., Liu, B., Matthews, P., Ong, G., Pell, J., Silman, A., Young, A., Sprosen, T., Peakman, T. & Collins, R. 2015. UK Biobank: An Open Access Resource for Identifying the Causes of a Wide Range of Complex Diseases of Middle and Old Age. *PLoS Med,* 12**,** e1001779.

Sun, C., Skaletsky, H., Rozen, S., Gromoll, J., Nieschlag, E., Oates, R. & Page, D. C. 2000. Deletion of azoospermia factor a *(AZF*a) region of human Y chromosome caused by recombination between HERV15 proviruses. *Hum Mol Genet,* 9**,** 2291-2296.

Tenesa, A., Navarro, P., Hayes, B. J., Duffy, D. L., Clarke, G. M., Goddard, M. E. & Visscher, P. M. 2007. Recent human effective population size estimated from linkage disequilibrium. *Genome Res,* 17**,** 520-526.

The International HapMap Consortium 2007. A second generation human haplotype map of over 3.1 million SNPs. *Nature,* 449**,** 851-861.

Thorvaldsdóttir, H., Robinson, J. T. & Mesirov, J. P. 2013. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Brief Bioinform,* 14**,** 178-192.

Tristem, M. 2000. Identification and characterization of novel human endogenous retrovirus families by phylogenetic screening of the human genome mapping project database. *J Virol,* 74**,** 3715–3730.

Trombetta, B., D'Atanasio, E. & Cruciani, F. 2017. Patterns of Inter-Chromosomal Gene Conversion on the Male-Specific Region of the Human Y Chromosome. *Front Genet,* 8**,** 54.

Trombetta, B., Fantini, G., D'Atanasio, E., Sellitto, D. & Cruciani, F. 2016. Evidence of extensive non-allelic gene conversion among LTR elements in the human genome. *Sci Rep,* 6**,** 28710.

Trombetta, B., Sellitto, D., Scozzari, R. & Cruciani, F. 2014. Inter- and intraspecies phylogenetic analyses reveal extensive X-Y gene conversion in the evolution of gametologous sequences of human sex chromosomes. *Mol Biol Evol,* 31**,** 2108-2123.

UCSC Genome Browser. *UCSC Genome Browser* [Online]. Available: https://genome.ucsc.edu/.

Uddin, M., Thiruvahindrapuram, B., Walker, S., Wang, Z., Hu, P., Lamoureux, S., Wei, J., MacDonald, J. R., Pellecchia, G., Lu, C., Lionel, A. C., Gazzellone, M. J., McLaughlin, J. R., Brown, C., Andrulis, I. L., Knight, J. A., Herbrick, J., Wintle, R. F., Ray, P., Stavropoulos, D. J., Marshall, C. R. & Scherer, S. W. 2014. A high-resolution copy-number variation resource for clinical and population genetics. *Genet Med,* 17**,** 747-752.

Underhill, P. A., Jin, L., Lin, A. A., Mehdi, S. Q., Jenkins, T., Vollrath, D., Davis, R. W., Cavalli-Sforza, L. L. & Oefner, P. J. 1997. Detection of numerous Y chromosome biallelic polymorphisms by denaturing high-performance liquid chromatography. *Genome Res,* 7**,** 996-1005.

Van Ooijen, J. W. & Jansen, J. 2013. *Genetic mapping in experimental populations*.

Vitte, C. & Panaud, O. 2003. Formation of solo-LTRs through unequal homologous recombination counterbalances amplifications of LTR retrotransposons in rice Oryza sativa L. *Mol Biol Evol,* 20**,** 528-540.

Wang, C. C., Gilbert, M. T. P., Jin, L. & Li, H. 2014. Evaluating the Y chromosomal timescale in human demographic and lineage dating. *Investig Genet,* 5**,** 12.

Waples, R. S. 2015. Testing for Hardy-Weinberg proportions: have we lost the plot? *J Hered,* 106**,** 1-19.

Webb, A. J., Berg, I. L. & Jeffreys, A. J. 2008. Sperm cross-over activity in regions of the human genome showing extreme breakdown of marker association. *Proc Natl Acad Sci USA,* 105**,** 10471-10476.

Weller, P. A., Critcher, R., Goodfellow, P. N., German, J. & Ellis, N. A. 1995. The human Y chromosome homologue of XG: transcription of a naturally truncated gene. *Hum Mol Genet,* 4**,** 859-868.

Wilson, I. J., Weale, M. E. & Balding, D. J. 2003. Inferences from DNA data: population histories, evolutionary processes and forensic match probabilities. *J Roy Stat Soc Ser A,* 166**,** 155-188.

Wright, D. J., Day, F. R., Kerrison, N. D., Zink, F., Cardona, A., Sulem, P., Thompson, D. J., Sigurjonsdottir, S., Gudbjartsson, D. F., Helgason, A., Chapman, J. R., Jackson, S. P., Langenberg, C., Wareham, N. J., Scott, R. A., Thorsteindottir, U., Ong, K. K., Stefansson, K. & Perry, J. R. B. 2017. Genetic variants associated with mosaic Y chromosome loss highlight cell cycle genes and overlap with cancer susceptibility. *Nat Genet,* 49**,** 674-679.

Y Chromosome Consortium 2002. Systematic identification of novel protein domain families associated with nuclear functions. *Genome Res,* 12**,** 47-56.

Yi, J. M., Kim, T. H., Huh, J. W., Park, K. S., Jang, S. B., Kim, H. M. & Kim, H. S. 2004. Human endogenous retroviral elements belonging to the HERV-S family from human tissues, cancer cells, and primates: expression, structure, phylogeny and evolution. *Gene,* 342**,** 283-292.

Yu, A., Zhao, C., Fan, Y., Jang, W., Mungall, A. J., Deloukas, P., Olsen, A., Doggett, N. A., Ghebranious, N., Broman, K. W. & Weber, J. L. 2001. Comparison of human genetic and sequence-based physical maps. *Nature,* 409**,** 951-953.

Zeqiraj, E. & Sicheri, F. 2015. Getting a handle on glycogen synthase – Its interaction with glycogenin. *Mol Aspects Med,* 46**,** 63-69.

Zhai, L., Mu, J., Zong, H., DePaoli-Roach, A. A. & Roach, P. J. 2000. Structure and chromosomal localization of the human glycogenin-2 gene GYG2. *Gene,* 242**,** 229-235.

Zhivotovsky, L. A., Underhill, P. A., Cinnioğlu, C., Kayser, M., Morar, B., Kivisild, T., Scozzari, R., Cruciani, F., Destro-Bisol, G., Spedini, G., Chambers, G. K., Herrera, R. J., Yong, K. K., Gresham, D., Tournev, I., Feldman, M. W. & Kalaydjieva, L. 2004. The Effective mutation rate at Y chromosome short tandem repeats, with application to human population-divergence time. *Am J Hum Genet,* 74**,** 50-61.

Ziegler, A. & König, I. R. 2010. *A statistical approach to genetic epidemiology*. 2nd ed. Darmstadt, Germany, Wiley-VCH Verlag.

Appendix

A.1 Coordinates along the extended PAR1 sequence

Portions	hg19 coordinates
LTR6B-X	chrX:2,694,151-2,694,702 (552 bp)
Proximal PAR1 on X	chrX:2,694,703-2,699,520 (4,818 bp)
Distal X-Specific	chrX:2,699,521-2,808,548 (109,028 bp)
Recombinant LTR6B (X+YPAR1) junction*	chrX:2,808,549-2,809,097 (549 bp) chrY:2,644,151-2,644,702 (552 bp)
Proximal PAR1 on Y	chrY:2,644,703-2,649,520 (4,818 bp)

*This part is rearranged depending on junction sequences.

A.2 Allele-Specific Oligonucleotide probe (ASO) sequences

Assay	Man	ASO name	5' to 3' sequence	Location of *SNP* in hg19 (chrX)
Distal	20, 53	9.5 A/C	CACACACATAT*A*/*C*TCCACA	2699555
Distal	53	9.9 G/A	CAGACAATGGC*G*/*A*ATCTAT	2699968
Distal	53	10.0 C/T	TAGTCACGC*C*/*T*GATGAGCT	2700027
Distal	20	10.1 G/A	CGAGAATCCC*G*/*A*ACAGCGG	2700157
Distal	20	10.2 A/G	CTCTTTCTCA*A*/*G*TCTGGTT	2700202
Distal	20	10.6 C/T	GGAGTGCTGAC*C*/*T*GGTCAG	2700608
Distal	53	10.6a A/G	ACCGGTC*A*/*G*GGTTGGAAAG	2700613
Distal	20	11.0 T/C	GGTCTGACCTC*T*/*C*TTCACT	2701073
Distal	20, 53	11.1 C/T	CAAGTCCTTT*C*/*T*TCTCACT	2701185
Distal	53	12.1 C/T	AATCCCAAA*C*/*T*ACCACCCA	2702143
Distal	53	12.3 C/T	CTTAAAATG*C*/*T*GTGGCTGG	2702339
Distal	53	12.6 T/C	GAACACTCAG*T*/*C*CCCTCCC	2702698
Distal	20	13.3 T/C	GATGAGCTGT*T*/*C*GGTGTAC	2703391
Distal	53	13.5 G/A	GGGAGGCAG*G*/*A*TCTGACTA	2703544
Distal	20	13.6 A/G	CAATTGAAC*A*/*G*TCAGAACA	2703633
Distal	20	14.3 C/G	TCTCACT*C*/*G*TATTGCTCAG	2704335
Distal	20	14.4 T/C	CACCACACC*T*/*C*GGCTAATT	2704469
Distal	20	14.6 T/C	GTGCCTGTCC*T*/*C*CATGTTG	2704609
Distal	20	14.8 T/C	CCGGATCCAA*T*/*C*AGGACTA	2704808
Distal	53	15.0 T/C	AGCCACCCAGT*T*/*C*TATGGT	2705011
Distal	53	15.2 T/C	CTGTGATAG*T*/*C*AGCCTGAA	2705265

Proximal	53	93.1 A/G	GCTATTAAAA*A*/*G*CATTCTT	2783107
Proximal	53	93.5 T/G	CGTGTTTG*T*/*G*CCGTCCGTG	2783555
Proximal	53	94.0 A/G	GATGAATGG*A*/*G*TAAAGAAA	2784051
Proximal	53	95.4 T/G	CAAGGGAGG*T*/*G*GAAAACAG	2785428
Proximal	20, 53	96.0 G/A	GGAGGCC*G*/*A*AAGCAGGAAA	2786038
Proximal	53	97.4 A/G	GAGTCGT*A*/*G*CAACATCACT	2787485
Proximal	20, 53	97.8 C/G	CAGGCTA*C*/*G*TCTTGAATTC	2787898
Proximal	20	99.8 T/G	CAACAGCT*T*/*G*CACCATTTG	2789848
Proximal	20	100.1 T/C	CAGCCTCT*T*/*C*TCCAATGAC	2790148
Proximal	20	102.6 A/G	GTGCAGATC*A*/*G*TATCTGTG	2792662
Proximal	20	102.8 A/G	CAGCTGAC*A*/*G*TCACTCAAA	2792838

*The alternative SNPs of allele-specific oligonucleotide probes (ASOs) are shown in *italic*.

A.3 Y-chromosome haplogrouping using a SNaPshot® single-base extension assay

A) Primer sequences and annealing temperature for haplogroup-I2a multiplex

Primer name	5' to 3' sequence	Location of 5' nt in hg19 (chrY)	% in primer mix	T_m (°C)
I2F1	CTAGTGGCTGCATAAGGTAG	16638691	6.12	59
I2R1	GTCAGTCAAAAGCTCCTGTC	16638914		
I2F2	GGAGTCACATCACTTAGGTG	7879298	8.16	
I2R2	CTTTTCTAGGACCAGACCAC	7879507		
I2F3	GAGACAAGCATAGTGATAGGG	14491650	4.08	
I2R3	GGCATGCCCAACTCCTCTTC	14491815		
I2F4/3	CTGAAGCCTGGGTGTGACTTG	6873997	6.12	
I2R4/2	GTTCTGGGCTGATATTTCTGC	6874152		
I2F5	GGGGTGTAGTTTAGACATC	19126608	14.29	
I2R5	GGCTGAGATTCTATCCTGAG	19126738		
I2F6	CAGCACTTGTCTTCTGTTTGC	21778604	10.21	
I2R6	GTGGAGATGGTAAGTTGTCC	21778760		
I2F8	GGTGATTGATAGCTAGACAGC	14487267	14.29	

I2R8	GGATGACAGACTCTATGTG	14487496	
I2F9	CATGGGGAACAGGCAGTGAA	8846879	8.16
I2R9	GAACTTCCTTCCTCTCCAAC	8847138	
I2F10/2	CTATCAGGTAGGCAGAGTG	2887182	28.57
I2R10	TCCTAAAGAGTGCAGAGCTG	2887454	

B) Extension primer sequences for haplogroup-I2a SNaPshot® assay

SNP	hg19 (chrY)	change	YCC 2008 Hg	ISOGG 2018 Hg	5' to 3' primer sequence	% in mix
M438, P215	16,638,804	A to G	I2	I2	AAAGCCTGGAATGT AGACTAATGGT	5.36
S238, L460	7,879,415	A to C	*not listed*	I2a1	AAAAAAAAAAAAAA AAAAAAAAAAAAAA AAAAAAAAAAAAAA CTTGGCTCTGCCTAC AGAG	14.29
P37.2	14,491,684	T to C	I2a	I2a1a	AAAAAAAAAAAGA TAGGGTGGGATTGG TTCA	5.36
CTS595	6,874,115	C to T	*not listed*	I2a1a1	AAAAAAAAAAAAAA AAAAAAAAACTGCC AACAAATTAAATACC	5.36
S21825	19,126,655	G to A	*not listed*	I2a1a1b	AAAAAAAAAAAAAA AAAAAATGGACTTA GGGAACTTCC	10.71
L1286	21,778,662	G to A	*not listed*	I2a1a1b1	AAAAAAAAAAAAAA AAAAAAAAAAAAAA AAAAAAAAAAAAAA AAAAAAGTCTGAAT GTCTCAAACTC	7.14
L233	14,487,362	G to A	*not listed*	I2a1a1b1 a1	AAAAAAAAAAAAAA AAAAAAAAAAAAAA	16.07

					AAAGCTGATTGATA GCTAGAT	
A417	8,846,964	T to C	*not listed*	I2a1a1b1 a1a1	AAAAAAAAAAAAAAA AAAAAAAAAAAAAAA AAAAAAAAATTTTTT ATACCAAAGCTT	14.29
L1294	2,887,401	T to C	*not listed*	I2a1a1b2	AAAAAAAAAAAAAAA AAAAAAAAAAAAAAA AAAAAAAAAAAAAAA AAAAAAAAAAAAAAA ATTAGCCATTTTTTG TG	21.42

A.4 Primer sequences for sperm recombination analysis

Primer name	**5' to 3' sequence**	**Location of 5' nt in hg19 (chrX)**
X25/56F	GGAAGGTATATAAGCTCTGG	2699568
9.5F C/A	GATAAATGCACACACATAT*C/A*	2699536
X9.6F G/T	CCTTCTCCCTGTAACAG*G/T*	2699628
9.9F A/G	TTTTTAACCAGACAATGGC*A/G*	2699949
X9980F	ACACCGTGGGCAGCAAATCA	2699979
X11131F	GGACACTTTGCATTATCATC	2701130
X2.77R	AACATAGCCATGATCCGCAG	2702217
X12654F	GCGAGATCCTTTAAGATGGG	2702653
X13778R	CTGGGCAAGAACACTGGTTA	2703777
X14598R	GTGGCTCACACCTGTAATCCC	2704597
14.6R T/C	TTACTTTCAAGACAACATG*A/G*	2704628
X14763R	TTCTTAAAGGCCCCATCTCC	2704762

14.8R T/C	TTATTAGGATACTAGTCCT*A*/*G*	2704827
15.0R C/T	GCTGTCACAGAACACCATA*G*/*A*	2705030
15.2R C/T	CTTAGTCCATTTCAGGCT*G*/*A*	2705283
X15423R	AACCTCCAACTGCATTGATTC	2705422
93.1F A/G	CCATGCCTGGCTATTAAAA*A*/*G*	2783088
X93213F	GAGGGATTTCCACGGATTG	2783212
93.5F G/T	GATATGTAGCTTCGTGTTTG*T*/*G*	2783535
X94556F	GAGACTAGGAAAGGTGGG	2784555
X95211F	GAGCGTACTTTGTTTAGGGT	2785210
X95737R	CTCCCATTCCTGAATCTCTC	2785736
X98092R	ACTAGAAGACAAAACGCTGG	2788091
X98641R	GTAGACTGAAGTATGGGGAG	2788640
X99925F	CACCACCTCCAGCAGTAATAA	2789924
X100924R	GTCCTGGATTTGAATACTG	2790923
102.6R A/G	CAGTGGCCTACACAGATA*T*/*C*	2792680
X102760R	TCCAGTTCATTCTCCTAATC	2792759
102.8R A/G	ATCCCAGAGATTTTGAGTGA*T*/*C*	2792858

The alternative 3' nucleotides of allele-specific primers (ASPs) are shown in *italic*.

A.5 Primer combinations and annealing temperatures used for sperm recombination analysis

Assay	Man	Primer combination	T_m (°C)	Purpose
Distal	20	9.5F C + X2.77R 9.5F A + X2.77R	58 56	phasing of 5' selector sites
Distal	20	X11131F + 14.8R C X11131F + 14.8R T	54 54	phasing of 3' selector sites

Distal	20	9.5F C + 14.8R C	57-55	crossover detection (1° PCR) orientation A
Distal	20	X9.6F G + 14.6R C	54	crossover detection (2° PCR) orientation A
Distal	20	9.5F A + 14.8R T	56-54	crossover detection (1° PCR) orientation B
Distal	20	X9.6F T + 14.6R T	54	crossover detection (2° PCR) orientation B
Distal	20	X9980F + X14598R	64	crossover detection (3° PCR) orientation A & B
Distal	53	9.5F C + X13778R 9.5F A + X13778R	59 58, 57	phasing of 5' selector sites
Distal	53	X12654F + 15.2R C X12654F + 15.2R T	59 59	phasing of 3' selector sites
Distal	53	9.5F A + X15423R	56	crossover detection (1° PCR #1) orientation A
Distal	53	X25/56F + 15.2R C	58	crossover detection (1° PCR #2) orientation A
Distal	53	9.9F G + 15.0R C	61	crossover detection (2° PCR) orientation A
Distal	53	9.5F C + X15423R	56	crossover detection (1° PCR #1) orientation B
Distal	53	X25/56F + 15.2R T	58	crossover detection (1° PCR #2) orientation B
Distal	53	9.9F A + 15.0R T	61	crossover detection (2° PCR) orientation B
Distal	53	X9980F + X13778R	61	crossover detection (3° PCR) orientation A & B
Proximal	20	X94556F + 102.8R A X94556F + 102.8R G	60, 58	phasing (1° PCR)
Proximal	20	X95211F + X100924R	57	phasing (2° PCR 5' amplicon)

Proximal	20	X99925F + X102760R	58	phasing (2° PCR 3' amplicon)
Proximal	20	X94556F + 102.8R A	60, 58	recombinant detection (1° PCR) orientation A
Proximal	20	X95211F + 102.6R A	61, 60	recombinant detection (2° PCR) orientation A
Proximal	20	X94556F + 102.8R G	60, 58	recombinant detection (1° PCR) orientation B
Proximal	20	X95211F + 102.6R G	59	recombinant detection (2° PCR) orientation B
Proximal	53	93.1F A + X98641R 93.1F G + X98641R	60-58	phasing (1° PCR)
Proximal	53	X93213F + X95737R	58	phasing (2° PCR 5' amplicon)
Proximal	53	X95211F + X98092R	58	phasing (2° PCR 3' amplicon)
Proximal	53	93.1F A + X98641R	60-58	recombinant detection (1° PCR) orientation A
Proximal	53	93.5F T + X98092R	62	recombinant detection (2° PCR) orientation A
Proximal	53	93.1F G + X98641R	60-58	recombinant detection (1° PCR) orientation B
Proximal	53	93.5F G + X98092R	62	recombinant detection (2° PCR) orientation B

A.6 Primer sequences for Ion Torrent sequencing templates

Primer name	5' to 3' sequence	Location of 5' nt in hg19 (chrX)	T_m (°C)
PAR7500F	GAAGAGGCTCCTTACTGCTC	2697499	59
X13778R	CTGGGCAAGAACACTGGTTA	2703777	
X11629F	ACACGCACATCCAGATGGAC	2701628	59

X17586R	GCCCATTGCTTTTACGGAGC	2707585	
X17401F	CAGGGAAAGGTCAACAACGC	2707422	60
X19951R	GCCTTTTGAGGAGGAAGAAC	2709953	
X19801F	ATGCTTGAGTCCCAACAGGC	2709822	64
X22701R	CCCCAACCTCAGCTCTCTCA	2712741	
X22451F3	TGTGGGCGTCCTGCAAAGAG	2712438	64
X25501R3	CTCAGAAGACCAGGCACCAG	2715545	
X25151F	GGGAAGAGGAAGCTTTGGT	2715163	51
X26301R	CAACCTACAGAGTGTTTTG	2716326	
X26243F	CTCAGATCCTCTCCGGTTTG	2716242	59
X32676R	GTGTCCCGAATAGCTCCTTA	2722675	
X28523F	TGTCCGTCATCTTCGTTACC	2718522	59
X34984R	GAAATGAAGTCCTCAGGGCA	2724983	
X34801F	CACACTGTCTGCACTGATGG	2724847	62
X37251R	GTTGCATTCCACATCCTGGC	2727281	
X36801F	CCCCAAGGCTCTTGTGAGTT	2726822	60
X39401R	CCACGGATACGATGGGAGAC	2729418	
X38251F	ATGCATGGAAAGGGTGGCTA	2728271	64
X41101R	CGTGCTTCCTTCCACCTCAG	2731130	
X40151F	TTCGTCACAGGCGAGTCAGG	2730148	64
X42851R	TTAGCACAGGTCCAGCATC	2732876	
X42651F	TTGCACAAAGAGACTGTGGC	2732677	62
X45251R	CTTTCCAAAACTGGTGCTTC	2735298	
X44151F	ACCTGGGTGACAAGGCTAAC	2734155	62
X46701R	CTGGCAGTAGGTCTCCCTAA	2736746	
X45701F	GTGTGGTGAGGACCATGGATT	2735708	64
X48501R	CCTCCCTTTGGTCTACAGCC	2738531	

X48151F	ATGGCCTCCATCTAAGCTGC	2738174	64
X50701R	GGGCTGTGTGTAAATCAGGG	2740714	
X50351aF	GTCTCCCTCCCTGTCTGACC	2740345	64
X52351R	AGACGTTGGCGAAATTGCAG	2742384	
X56941F	TTCCGGACAGAGGCCAATCG	2746940	59
X62830R	AGCTGCACACATAAGGGCAT	2752829	
X62001F2	GCTCCTCATAAGCCTCAGGA	2752017	61
X64351R2	CACAGCTTCCCTCCTTTGCA	2754381	
X64301F	GTTTTTAAGGACCGAGGAGAC	2754302	59
X67251R	GCAAAATCAGATGATAGGATC	2757256	
X67151F	AGCTTATGCTATAGGTTGG	2757155	51
X70051R	TCATTAGGTAGAATGGATTC	2760081	
X69951F	CATCAGTCTTGTGCACAGAG	2759973	60
X72301R	TCTGAGGACACACTGGTTGG	2762309	
X72181F	GGAGGCATTTAATATGGTAG	2762180	57
X78205R	ACTGCTTCCTGTGTGCTCTC	2768204	
X77081F	CAAGCGTGTTATGGAGTGAT	2767080	59
X83241R	TGTGTGCAACAGAGACCCTA	2773240	
X83001F	CCAGGTGCTGTCCAATGTCG	2773030	64
X85751R	GTCGTAGAGCCTTGGCTAGC	2775804	
X85751F	TCTCGTTTCCTGCTTGCTTG	2775750	64
X88551R	AGCACACTGCAGATACTCCAC	2778580	
X88001aF	CTTCATCTCTGGTGGACGGTC	2778014	64
X90651aR	ACTCGGCCTCAGCAGACACC	2780678	
X89501F	AGCAGGTGAATTGAGGAGAC	2779505	62
X92151R	GGCTAGGAGTTTGAGCATTC	2782159	
X92028F	GAGTATCCAAGAATGCTGTG	2782027	59

X98092R	ACTAGAAGACAAAACGCTGG	2788091	
X97080F	GCTGTCCATACGGTATCTGA	2787079	59
X103595R	TGAGTGAATGTGAGGAGGTG	2793594	
X104167	GGATAAGGAATCTGGCTGTC	2794166	57
X110158R	TTTCCCGTCTCATAAAACTG	2800157	
X109074F	CCACTGACTGTCTTCCCTAG	2799073	59
X115091R	TAAGACCTTGGAGAAATGCC	2805090	
X113156F	GCACGAAACAGGTCCTAGTG	2803155	60
X116806R	TGGCCTGAACTGTACTCTTGA	2806805	
MX115182F	AAATGAACGCTAAGCCCCAC	2805181	60
Xdup118516R	CGTCTCCAGTAACATTGCCA	2808515	

A.7 Nested-PCR primer sequences used in the first long-range PCR to confirm two semen donors

Primer	**Forward (location of 5' in hg19)**	**Reverse (location of 5' in hg19)**	**Tm (°C)**
Big amplicon: 102.95F + YPABR (12 kb)	CGATCTTGCAGGATCTCTTTAG (chrX: 2802402)	CTTAATGGCCTTGAGCCTTCTC (chrX: 2650116)	57
Nested amplicon: M4X115939F + BAC35R (3.8 kb)	ACATGGTAGACGCCTGTTCA (chrX: 2805938)	CCACTAGTTCAGGAGTTATATG (chrX: 2695417)	

A.8 Sanger-sequencing primers covering 3.8-kb sequence across the recombinant junction after the first long-range PCR

Primer name	Sequence (location of 5' in hg19)	Direction
X116043R	CGTCTGGTGAGGATCTGCTG (chrX: 2806042)	R
M4X115939F	ACATGGTAGACGCCTGTTCA (chrX: 2805938)	F
BAC35R	CCACTAGTTCAGGAGTTATATG (chrX: 2695417)	R
RealDF	CAGAGAGGATAAGAAAGGAAGAAC (chrX: 2695297)	F
M5X115182F	AAATGAACGCTAAGCCCCAC (chrX: 2805181)	F
LTR118536R	ACAACACAAGAACCCTCTCT (chrX: 2808554)	R
X116717F	AAACTGGGGTCTCTCTATGT (chrX: 2806716)	F
X115543F	CAGTAGAGACAGATTTCGCCA (chrX: 2805542)	F
X115521R	CTGCGCATGGTGGTGCAAGT (chrX: 2805520)	R
X115894R	CTCTGTTGCGTAGGCTGGCT (chrX: 2805893)	R
X117177F	CAAGGCAAGTTTCAAAGCAG (chrX: 2807176)	F
X117883R	TGGTCTTCACTGCTACACTC (chrX: 2807882)	R
X116806R	TGGCCTGAACTGTACTCTTGA (chrX: 2806805)	R

A.9 Duplex Junction-PCR primers for detection of ePAR and used for Sanger-sequencing

Primer name	Forward (location of 5' in hg19)	Reverse (location of 5' in hg19)	Tm (°C)
Testing: XdupF + PAR119344R (848 bp)	TGGCAATGTTACTGGAGACG (chrX: 2808496)	CAAGGAGTCTGCTGGAAGTC (chrY: 2644940 or chrX: 2694940)	60
Control (1,551 bp)	GGGGTCCCGAGATTTATGTT (chrY: 2649035 or chrX: 2699035)	GCTAGAACAAGTTACCCCTC (chrY: 2650585)	

A.10 Distinguishing variants between LTR6B-X and –YPAR1

LTR6B-X		LTR6B-YPAR1		Ref. sequence genotype (X/Y)	Allele Freq. X/Y (%)	Modified genotype (X/Y)
hg19 (chrX)	dbSNP# (build 150)	hg19 (chrY)	dbSNP# (build 150)			
2808548	-	2644150	747251216	G/A	-/99.9	same
2808558	-	2644160	772998225	C/T	-/99.9	same
2808578	-	2644179	112316188	CACAC/-	-/51.3	same or CACAC/CACAC
2808591	-	2644188	-	G/A	-/-	same
2808593	-	2644190	-	T/A	-/-	same
2808609	111810772	2644206	759059762	T/A	94.9/99.7	same
2808614	112993272	2644211	766987125	G/A	94.7/99.7	same
2808634	758242398	2644231	-	C/T	98.9/-	same
2808638	766275661	2644235	755830067	C/T	98.9/99.8	same
2808641	-	2644238	764491368	A/T	-/99.8	same
2808643	-	2644240	311161	T/G	-/16.0	T/T
2808652	751567160	2644249	-	C/T	98.8/-	same
2808655	754499191	2644252	112750833	C/T	98.8/98.8	same
2808706	2316284	2644303	187149430	G/T	96.9/93.8	same
2808815	4061829	2644413	-	-/CAAGGTGA	94.0/-	same
2808824	12843082	2644429	2534625	A/G	73.1/25.0	A/A
2808989	2316283	2644594	2534626	T/T	96.1/33.3	T/G
2809044	139126456	2644649	311162	G/A	97.9/14.8	G/G

A.11 Mid positions (hg19) of detected Myers' motifs across chrX:2699521-2808548

7/8 CnnCCnTnnC CnC	7/8 nCnCCnTnnC CnC	7/8 CCnCCnTnn CCnn	7/8 CCnCnnTnnC CnC	8/8 CCnCCnTnn CCnC	7/8 CCnCCnTnnn CnC	7/8 CCnnCnTnnC CnC	7/8 CCnCCnnnnC CnC	7/8 CCnCCnTnn CnnC
2704425	2701407	2701877	2702751	2701395	2700678	2712110	2701782	2712399
2710692	2712097	2702350	2704719	2704560	2702155	2736276	2707173	2726560
2712212	2712154	2705809	2723278	2706662	2705433	2738478	2711196	2727955
2716496	2712799	2705985	2741655	2717473	2715192	2754476	2715360	2735129
2719538	2726087	2706622	2747722	2718795	2715436	2756470	2715839	2735145
2721607	2738017	2716838	2748324	2724321	2717288	2768599	2719999	2735198
2722507	2738703	2718755	2783800	2729888	2719066	2785295	2720312	2774719
2723222	2747487	2727114	2791583	2739045	2721686	2785551	2721309	2785743
2726213	2747906	2727500	2796804	2741510	2726183	2799450	2724518	2787974
2729208	2748397	2728281	2806108	2747381	2734464	2802236	2737080	2789561
2730937	2757053	2728818	2806636	2764549	2738286	2807236	2737227	2794819
2741796	2761199	2728887		2773695	2744472		2739996	
2747220	2766700	2741836		2779861	2746236		2742256	
2747270	2776692	2742338		2785421	2747611		2742560	
2747715	2778958	2743677		2803382	2749847		2746768	
2750499	2786730	2743944			2750484		2747020	
2750827	2802259	2747443			2753020		2747085	
2751194	2806849	2750732			2780240		2747227	
2754614	2808309	2759698			2782273		2747515	
2756482		2764195			2782501		2747985	
2760320		2765456			2785306		2753139	
2765679		2770159			2802736		2753662	
2766990		2785625					2754028	
2781034		2786275					2754780	
2783915		2800564					2754909	
2786725		2801146					2755019	
2787771		2806691					2755108	
2789496		2807395					2760971	
2793914							2761105	
2795075							2764359	
2798881							2767040	
2802011							2768059	
2804933							2772233	
							2772379	
							2772874	
							2776268	
							2776408	
							2779022	
							2779493	
							2784971	
							2791562	
							2791592	
							2794756	
							2803796	
							2805613	
							2808091	
							2808318	

A.12 SNPs with MAF >0.2 from CEU population (1000 Genomes) for LD analysis

SNP ID	hg19 (chr X)	SNP ID	hg19 (chr X)	SNP ID	hg19 (chr X)	SNP ID	hg19 (chr X)	SNP ID	hg19 (chr X)
rs2306737	2699968	rs311168	2710840	rs1683999	2719840	rs5939356	2763684	rs34599585	2776127
rs2306736	2700027	rs311169	2710995	rs311194	2720702	rs7879795	2772660	rs5939372	2776153
rs5939320	2700202	rs311170	2711429	rs311195	2721646	rs7882856	2772694	rs56077633	2776318
rs4892890	2700608	rs4892892	2711722	rs311196	2721788	rs4892899	2773521	rs5939373	2776372
rs1970796	2701073	rs311174	2713012	rs311197	2722056	rs4892900	2773780	rs5982897	2776686
rs1486175	2703391	rs1639329	2713069	rs6642024	2723666	rs5939133	2774700	rs5939137	2783555
rs11413215	2703500	rs1639330	2713073	rs5939336	2729281	-	2774760	rs5939379	2783776
rs1419931	2703633	rs311175	2713089	-	2730119	rs140413741	2774837	rs74672236	2783915
rs6641645	2704335	rs311177	2713698	-	2730120	rs150374646	2774838	rs5939380	2784051
rs5939322	2704469	rs311179	2713763	rs62582275	2730289	rs7473759	2774990	rs61155915	2784877
rs6641647	2704556	rs311180	2713807	rs138736942	2735159	rs200685855	2775000	rs6642051	2786038
rs5939323	2704609	rs311182	2714277	rs6641652	2740681	rs7473760	2775002	rs5939139	2791081
rs6642018	2704808	rs4484858	2715425	rs3828931	2740891	rs58163192	2775211	rs6567674	2794461
rs6642019	2704989	rs2316291	2717033	rs6641657	2751904	rs141405035	2775259	rs7065465	2794858
rs59808240	2705464	rs311189	2717067	rs4892894	2755925	rs59741043	2775330	rs1269	2800624
rs58649691	2706896	rs5939327	2717234	rs4892896	2757783	rs9781871	2775356	rs1268	2800677
rs5939117	2707060	rs311190	2717538	rs67976306	2759405	rs905375	2775601	rs12844789	2801000
rs71994079	2709008	rs140400664	2719242	rs374688504	2760226	rs5939369	2775691	rs12859113	2801026
rs311167	2710506	rs5939330	2719298	rs55921940	2760784	rs5939370	2775876	rs211664	2801155
rs146747084	2710586	rs311192	2719780	rs202192548	2763555	rs5939371	2775998	rs5939141	2803387
								rs5939386	2803689
								rs211654	2805786
								rs10625423	2805907

A.13 Poisson correction formula

Given P = Poisson correction of number of recombinant event(s)

n = number of observed recombinant event(s) per reaction

N = total number of reactions

n_a = total number of unresolved mixed reaction(s) (ambiguous reaction)

Then $$P = (-\ln\left(1 - \frac{n}{N-n_a}\right))(N - n_a)$$

A.14 Summary statistics for Ion Torrent sequencing across the ePAR

Individual			bases	>=Q20 bases	>=Q20 bases, (%)	reads	mapped reads	% of reads mapped
NA10847	*female*	da	43,979,848	39,307,954	89.38	193,337	179,808	93.00

NA12146	*ePAR*	fa	71,728,327	64,684,810	90.18	290,574	271,962	93.59
NA12239	*female*	mo	86,728,887	77,742,730	89.64	359,398	334,770	93.15
man 20	*ePAR*		78,320,031	70,679,322	90.24	316,557	306,178	96.72
man 53	*ePAR*		72,216,306	65,159,010	90.23	284,570	279,360	98.17
B1 *	*ePAR*	br	66,763,804	59,929,318	89.76	264,729	232,026	87.65
P1 *	*ePAR*	pa	66,218,114	59,721,814	90.19	262,450	235,900	89.88
F1 *	*ePAR*	fa	80,785,440	72,804,184	90.12	336,278	307,895	91.56
P2 *	*ePAR*	pa	78,977,048	70,732,103	89.56	325,234	310,205	95.38
F2 *	*ePAR*	fa	36,711,977	32,891,346	89.59	163,180	157,325	96.41
P3 *	*ePAR*	pa	149,637,283	134,452,533	89.85	810,574	753,664	92.98
F3 *	*ePAR*	fa	84,790,584	75,859,154	89.47	345,952	320,717	92.71
P4 *	*ePAR*	pa	85,253,719	76,863,868	90.16	355,542	347,721	97.80
F4 *	*ePAR*	fa	85,659,925	77,183,601	90.10	343,893	335,654	97.60
P5 *	*ePAR*	pa	26,747,753	24,051,513	89.92	118,408	117,148	98.94
F5 *	*ePAR*	fa	85,472,838	76,732,964	89.77	346,425	321,536	92.82
P6 *	*ePAR*	pa	113,795,421	101,829,085	89.48	476,539	457,852	96.08
F6 *	*ePAR*	fa	75,312,769	66,838,973	88.75	300,658	276,314	91.90
6689_01	*ePAR*		83,255,363	75,124,261	90.23	333,594	321,682	96.43
333	*ePAR*		72,691,290	64,267,808	88.41	320,268	309,115	96.52
		min	**26,747,753**	**24,051,513**	**88.41**	**118,408**	**117,148**	**87.65**
		max	**149,637,283**	**134,452,533**	**90.24**	**810,574**	**753,664**	**98.94**
		mean	**77,252,336**	**69,342,818**	**89.75**	**327,408**	**308,842**	**94.46**

*from [Mensah *et al.*, 2014]. da = daughter, fa=father, mo=mother, br=brother, pa=patient

A.15 Comparison of ePAR haplotype structures with phase-known X chromosomes – SNP markers

SNP ID	hg19 (chr X)	SNP ID	hg19 (chr X)	SNP ID	hg19 (chr X)	SNP ID	hg19 (chr X)	SNP ID	hg19 (chr X)
rs28579419	2699645	rs311182	2714277	rs5982871	2739032	rs142753289	2772820	rs5939380	2784051
rs60075487	2699676	rs5939326	2714531	rs6641652	2740681	rs4892899	2773521	rs200259684	2784879
rs2306737	2699968	rs141584360	2714584	rs3828931	2740891	rs4892900	2773780	rs6642050	2785428
rs2306736	2700027	rs4484858	2715425	rs3828930	2740892	rs5939133	2774700	rs62582323	2785712
rs5939319	2700157	rs2316291	2717033	rs73190969	2741078	rs4993468	2774793	rs62582325	2785737
rs5939320	2700202	rs311189	2717067	rs141205366	2741322	rs7473759	2774990	rs62582327	2785740
rs5982852	2702047	rs5939327	2717234	rs185534666	2741406	rs7473760	2775002	rs62582329	2785749
rs5982853	2702143	rs311190	2717538	rs6641656	2748292	rs148979483	2775148	rs190308716	2785936
rs111595179	2702339	rs5939329	2718135	rs5939348	2748421	rs58163192	2775211	rs12558009	2786029
rs5982854	2702568	rs752782071	2718189	rs5939350	2749521	rs141405035	2775259	rs6642051	2786038
rs73433431	2702698	rs11152547	2724389	rs5939351	2750122	rs9781871	2775356	rs73435434	2792662
rs5982855	2702799	rs7058222	2724429	rs2873118	2750768	rs905375	2775601	rs75069845	2792838
rs5982856	2702946	rs7062707	2724448	rs6641657	2751904	rs5939134	2778322	rs6567674	2794461
rs5982584	2703354	rs5939335	2725660	rs12396748	2752070	rs2316285	2778433	rs7065465	2794858
rs1486175	2703391	rs189287974	2727265	rs12387509	2752333	rs1905995	2778526	rs5982604	2797795
rs1419931	2703633	rs6567647	2728221	rs5982590	2752707	rs5939374	2778546	rs5982605	2797895
rs6641645	2704335	rs12010750	2728516	rs113160927	2753182	rs1905996	2778587	rs12388511	2798480
rs6642018	2704808	rs141066369	2728872	rs7057656	2753527	rs5939375	2778715	rs1269	2800624
rs6642019	2704989	rs5982866	2729011	rs138650022	2753713	rs5982900	2778796	rs1268	2800677
rs113922957	2705011	rs5939336	2729281	rs7892329	2754793	rs1905997	2778832	rs9355	2800788
rs112589751	2705265	rs5982867	2729346	rs5982876	2756413	rs10871869	2778982	rs12844789	2801000
rs5982858	2705462	rs12851007	2732634	rs5982877	2757233	rs10871870	2778996	rs12859113	2801026
rs5982859	2706340	rs5939124	2733109	rs5982878	2757399	rs11152551	2779171	rs211664	2801155
rs5982860	2706487	rs5939125	2733210	rs5982879	2757754	rs11152552	2779211	rs60445074	2801158
rs5939117	2707060	rs3672	2733641	rs4892896	2757783	rs10871871	2779330	rs12391908	2802182
rs5939324	2707142	rs3671	2733668	rs11152548	2759615	rs2306735	2779570	rs62582348	2802356
rs311166	2707978	rs5939340	2734838	rs62582317	2767637	rs4892901	2780265	rs211665	2802368
rs145903180	2709600	rs5939341	2734930	rs145268586	2767957	rs4892902	2780319	rs1637783	2802396
rs112470161	2710504	rs140285516	2735299	rs2316287	2768611	rs5939135	2780533	rs1637784	2802397
rs311167	2710506	rs5939342	2735539	rs5939358	2768713	rs6642045	2780747	rs211668	2802726
rs311168	2710840	rs5939343	2735621	rs9320050	2768851	rs6642046	2780826	rs76495523	2802915
rs311169	2710995	rs5939344	2735741	rs9320051	2768864	rs6642047	2780829	rs59272600	2802991
rs311170	2711429	rs146579399	2735895	rs6641662	2769103	rs149865569	2780978	rs5939140	2803276
rs4892892	2711722	rs6642031	2735926	rs62582319	2769284	rs12391316	2781220	rs56019734	2803303
rs2291380	2712283	rs5939126	2736196	rs5939360	2769334	rs7057853	2781260	rs5939141	2803387
rs311174	2713012	rs1809566	2736301	rs7058332	2771314	rs55776760	2781357	rs5939142	2803469
rs1639329	2713069	rs2316290	2736503	rs5939361	2771315	rs5939376	2782305	rs5939143	2803688
rs1639330	2713073	rs6567653	2737097	rs5939362	2771540	rs11152553	2782384	rs5939386	2803689
rs311175	2713089	rs2018620	2737149	rs4892898	2772275	rs5939377	2782455	rs5018317	2804185
rs7061550	2713211	rs5982869	2737194	rs7879795	2772660	rs5939136	2782633	rs211654	2805786
rs311177	2713698	rs901321	2737282	rs7882856	2772694	rs5939378	2783107	rs211655	2806196
rs311179	2713763	rs5939128	2737851	rs6567664	2772772	rs5939137	2783555		
rs311180	2713807	rs5982870	2738809	rs6567665	2772792	rs5939379	2783776		

A.16 64 SNPs in UK Biobank Axiom®Array corresponding to the X-translocated part of ePAR for screening for putative ePARs in the UK Biobank database

SNP ID	hg19 (chr X)	SNP ID	hg19 (chr X)	SNP ID	hg19 (chr X)	SNP ID	hg19 (chr X)	SNP ID	hg19 (chr X)
rs1419931	2703633	rs2306736	2700027	rs5939362	2771540	rs76495523	2802915	rs377738756	2778108
rs17330993	2779749	rs3749988	2724760	rs5982897	2776686	rs311191	2719111	-	2778106
rs1970797	2701185	rs5939125	2733210	rs6642045	2780747	rs5982872	2744765	-	2779768
rs2306734	2777985	rs62582317	2767637	rs62582329	2785749	rs17090628	2774609	-	2779568
rs5939137	2783555	rs211660	2800295	rs12008127	2791604	rs5982603	2788707	-	2700151
rs5939139	2791081	rs73433431	2702698	rs62582348	2802356	rs41311459	2793670	-	2707747
rs5939320	2700202	rs5982859	2706340	rs56019734	2803303	rs2306735	2779570	-	2715401
rs5939384	2789848	rs4892892	2711722	rs146462965	2700972	rs145587108	2723651	-	2729474
rs5982891	2767186	rs311196	2721788	rs145903180	2709600	rs200824650	2773196	-	2700151
rs12010750	2728516	rs7058222	2724429	rs140532845	2711779	rs138618142	2715401	-	2707747
rs12396748	2752070	rs5939350	2749521	rs140285516	2735299	rs148872483	2793951	-	2715401
rs5982588	2743627	rs11152548	2759615	rs141205366	2741322	rs141116662	2799114	-	2729474
rs60075487	2699676	rs5982593	2765319	rs147225766	2742170	-	2772157		

A.17 Locations (hg19) at 5' of overlapping sequences amongst LTR6B-X (DistX), LTR6B-YPAR1 (PAR1) and TFBS-LTR6Bs (TFBS) with >520-bp in length and 90% similarity

A) Overlapping datasets

Overlapping datasets	no	Location [Y Chromosome Consortium]	Overlapping datasets	no	Location [Y Chromosome Consortium]	Overlapping datasets	no	Location [Y Chromosome Consortium]	Overlapping datasets	no	Location [Y Chromosome Consortium]
DistX-PAR1-TFBS	40	1:33109510 10:100890241 12:96139770 13:82954205 14:101561236 14:102531063 14:44850310 14:86722121 14:88280563 19:23362546 19:54277666 19:8267906 2:155434267 2:201990761 20:31562916 21:44356555 3:100320708 3:125557645 3:189851525 3:45727313 4:296625 4:8430908 5:146421392 5:21165034 5:95043041 6:153003239 6:27015285 7:138776613 7:153128513 7:38371844 8:138031465 8:139094676 8:29177067 8:39801881 8:75518113 X:102639964 X:2694151 X:2808549 X:437735 X:55892938	**DistX-TFBS**	9	11:67649537 8:17029316 9:100461867 11:86134540 3:186728304 2:212456644 21:43497403 5:55295766 14:35817116	**PAR1-TFBS**	7	3:183152034 3:129895527 4:4031968 3:146122793 15:29509086 11:3530578 19:36821387	**DistX-PAR1**	5	Y:387734 19:38045648 Y:2644151 5:154318691 2:240966265

B) Non-overlapping datasets

Non-overlapping datasets	No.	Location [Y Chromosome Consortium]	Non-overlapping datasets	No.	Location [Y Chromosome Consortium]	Non-overlapping datasets	No.	Location [Y Chromosome Consortium]
TFBS	25	Y:7319952 8:145920901 19:23369079 17:8778213 X:16028891 2:124448573 21:41949475 7:7027990	**DistX**	1	6:161311417	**PAR1**	4	11:55823600***** 7:151119837 19:36815778 6:2990914*****

		12:8470084 4:9608319 11:71417032 1:152994433 11:71423923 1:146710445 3:129902406 6:87190855 3:75556067 4:9601453 11:3537465 3:146129393 Y:7131000 13:38450041 2:124439219 3:125550783 7:32642837						

*LTR6A

A.18 Intervals of Myers's motifs and DSB cluster across LTR6B-X and LTR6B-YPAR1 (hg19)

	Myers' motif interval	DSB cluster
LTR6B-X	chrX:2808695-2808707 chrX:2809044-2809056	chrX:2806301-2809274
LTR6B-YPAR1	chrY:2644292-2644304 chrY:2644649-2644661	-

A.19 Junction diversity of ePAR

Each Junction prototype (Junction 1 and 2) is presented together with the reference sequences of LTR6B-X and –YPAR1, which are modified according to the most frequent allele of each SNP or indel, and are placed in the box on top of the variants. Distinct SNPs or indels between two reference LTR6Bs are highlighted in green. Sequences specific to LTR6B-X are highlighted in yellow and those specific to LTR6B-YPAR1 are highlighted in blue. For the sequences which are identical to both reference LTR6Bs and inferred to be the putative recombinant intervals are left unhighlighted.

A) Junction 1 and variants

```
                    ....|....| ....|....| ....|....| ....|....| ....|....|
                         5         15         25         35         45
Junction 1          TGTGTTGTAC CCGAGCGAGT TAGAAAAACG CCACACTTTG AGACGATTTA
LTR6B-X             TGTGTTGTAC CCGAGCGAGT TAGAAAAACG CCACACTTTG AGACGATTTA
LTR6B-YPAR1         TATGTTGTAC CTGAGCGAGT TAGAAAAACG C-----TTTG AGACAAATTA
Junction 1.1        TGTGTTGTAC CCGAGCGAGT TAGAAAAACG CCACACTTTG AGACGATTTA
Junction 1.2        TGTGTTGTAC CCGAGCGAGT TAGAAAAACG CCACACTTTG AGACGATTTA
Junction 1.3        TGTGTTGTAC CCGAGCGAGT TAGAAAAACG CCACACTTCG AGACGATTTA
Junction 3          TGTGTTGTAC CCGAGCGAGT TAGAAAAACG CCACACTTTG AGACGATTTA
Junction 4          TGTGTTGTAC CCGAGCGAGT TAGAAAAACG CCACACTTTG AGACAAATTA
Junction 5          TGTGTTGTAC CCGAGCGAGT TAGAAAAACG CCACACTTTG AGACGATTTA
Junction 5.1        TGTGTTGTAC CCGAGCGAGT TAGAAAAACG CCACACTTTG AGACGATTTA

                    ....|....| ....|....| ....|....| ....|....| ....|....|
                        55         65         75         85         95
Junction 1          AGAGTCCTTT ATTAGCCGGC GACCGAGAGA CGGCTAACGC TCAAAATTCT
LTR6B-X             AGAGTCCTTT ATTAGCCGGC GACCGAGAGA CGGCTAACGC TCAAAATTCT
LTR6B-YPAR1         AGAGTCCTTT ATAAGCCAGC GACCGAGAGA CGGCTAATGC TTAATATTCT
Junction 1.1        AGAGTCCTTT ATTAGCCGGC GACCGAGAGA CGGCTAACGC TCAAAATTCT
Junction 1.2        AGAGTCCTTT ATTAGCCGGC GACCGAGAGA CGGCTAACGC TCAAAATTCT
Junction 1.3        AGAGTCCTTT ATTAGCCGGC GACCGAGAGA CGGCTAACGC TCAAAATTCT
Junction 3          AGAGTCCTTT ATTAGCCGGC GACCGAGAGA CGGCTAACGC TCAAAATTCT
Junction 4          AGAGTCCTTT ATAAGCCAGC GACCGAGAGA CGGCTAATGC TTAATATTCT
Junction 5          AGAGTCCTTT ATTAGCCGGC GACCGAGAGA CGGCTAACGC TCAAAATTCT
Junction 5.1        AGAGTCCTTT ATTAGCCGGC GACCGAGAGA CGGCTAACGC TCAAAATTCT

                    ....|....| ....|....| ....|....| ....|....| ....|....|
                       105        115        125        135        145
Junction 1          CTCGGCCCCG AGGAAGGGGC TTGATTAACT TTTAGATCTT GGTTTAGGAA
LTR6B-X             CTCGGCCCCG AGGAAGGGGC TTGATTAACT TTTAGATCTT GGTTTAGGAA
LTR6B-YPAR1         CTCGGTCCTG AGGAAGGGGC TTGATTAACT TTTAGATCTT GGTTTAGGAA
Junction 1.1        CTCGGCCCCG AGGAAGGGGC TTGATTAACT TTTAGATCTT GGTTTAGGAA
Junction 1.2        CTCGGCCCCG AGGAAGGGGC TTGATTAACT TTTAGATCTT GGTTTAGGAA
Junction 1.3        CTCGGCCCCG AGGAAGGGGC TTGATTAACT TTTAGATCTT GGTTTAGGAA
Junction 3          CTCGGTCCTG AGGAAGGGGC TTGATTAACT TTTAGATCTT GGTTTAGGAA
Junction 4          CTCGGTCCTG AGGAAGGGGC TTGATTAACT TTTAGATCTT GGTTTAGGAA
Junction 5          CTCGGCCCCG AGGAAGGGGC TTGATTAACT TTTAGATCTT GGTTTAGGAA
Junction 5.1        CTCGGCCCCG AGGAAGGGGC TTGATTAACT TTTAGATCTT GGTTTAGGAA
```

```
                    ....|....| ....|....| ....|....| ....|....| ....|....|
                       155        165        175        185        195

Junction 1          GGGGAGGGCG GGGGGTCTAG TGAAAACCAT TTTACAGAAG TAAAGTAGGC
LTR6B-X             GGGGAGGGCG GGGGGTCTAG TGAAAACCAT TTTACAGAAG TAAAGTAGGC
LTR6B-YPAR1         GGGGAGGGCT GGGGGTCTAG TGAAAACCAT TTTACAGAAG TAAAGTAGGC
Junction 1.1        GGGGAGGGCG GGGGGTCTAG TGAAAACCAT TTTACAGAAG TAAAGTAGGC
Junction 1.2        GGGGAGGGCG GGGGGTCTAG TGAAAACCAT TTTACAGAAG TAAAGTAGGC
Junction 1.3        GGGGAGGGCG GGGGGTCTAG TGAAAACCAT TTTACAGAAG TAAAGTAGGC
Junction 3          GGGGAGGGCT GGGGGTCTAG TGAAAACCAT TTTACAGAAG TAAAGTAGGC
Junction 4          GGGGAGGGCT GGGGGTCTAG TGAAAACCAT TTTACAGAAG TAAAGTAGGC
Junction 5          GGGGAGGGCT GGGGGTCTAG TGAAAACCAT TTTACAGAAG TAAAGTAGGC
Junction 5.1        GGGGAGGGCT GGGGGTCTAG TGAAAACCAT TTTACAGAAG TAAAGTAGGC

                    ....|....| ....|....| ....|....| ....|....| ....|....|
                       205        215        225        235        245

Junction 1          AAAAAGTTAA AAGGATAAAT GGTTGCAGGA AAGTAAACAG TTCCAGGTGC
LTR6B-X             AAAAAGTTAA AAGGATAAAT GGTTGCAGGA AAGTAAACAG TTCCAGGTGC
LTR6B-YPAR1         AAAAAGTTAA AAGGATAAAT GGTTGCAGGA AAGTAAACAG TTCCAGGTGC
Junction 1.1        AAAAAGTTAA AAGGATAAAT GGTTGCAGGA AAGTAAACAG TTCCAGGTGC
Junction 1.2        AAAAAGTTAA AAGGATAAAT GGTTGCAGGA AAGTAAACAG TTCCAGGTGC
Junction 1.3        AAAAAGTTAA AAGGATAAAT GGTTGCAGGA AAGTAAACAG TTCCAGGTGC
Junction 3          AAAAAGTTAA AAGGATAAAT GGTTGCAGGA AAGTAAACAG TTCCAGGTGC
Junction 4          AAAAAGTTAA AAGGATAAAT GGTTGCAGGA AAGTAAACAG TTCCAGGTGC
Junction 5          AAAAAGTTAA AAGGATAAAT GGTTGCAGGA AAGTAAACAG TTCCAGGTGC
Junction 5.1        AAAAAGTTAA AAGGATAAAT GGTTGCAGGA AAGTAAACAG TTCCAGGTGC

                    ....|....| ....|....| ....|....| ....|....| ....|....|
                       255        265        275        285        295

Junction 1          AGGGGCTTTA AGACTATTAC AAGGTGATAG ACGCGGGGCT TTGGGCGTTA
LTR6B-X             AGGGGCTTTA AGACTATTA- -------TAG ACGCGAGGCT TTGGGCGTTA
LTR6B-YPAR1         AGGGGCTTTA AGACTATTAC AAGGTGATAG ACGCGAGGCT TTGGGCGTTA
Junction 1.1        AGGGGCTTTA AGACTATTAC AAGGTGATAG ACGCGGGGCT TTGGGCGTTA
Junction 1.2        AGGGGCTTTA AGACTATTAC AAGGTGATAG ACGCGAGGCT TTGGGCGTTA
Junction 1.3        AGGGGCTTTA AGACTATTAC AAGGTGATAG ACGCGAGGCT TTGGGCGTTA
Junction 3          AGGGGCTTTA AGACTATTAC AAGGTGATAG ACGCGAGGCT TTGGGCGTTA
Junction 4          AGGGGCTTTA AGACTATTAC AAGGTGATAG ACGCGAGGCT TTGGGCGTTA
Junction 5          AGGGGCTTTA AGACTATTAC AAGGTGATAG ACGCGGGGCT TTGGGCGTTA
Junction 5.1        AGGGGCTTTA AGACTATTAC AAGGTGATAG ACGCGAGGCT TTGGGCGTTA
```

```
                    ....|....| ....|....| ....|....| ....|....| ....|....|
                       305        315        325        335        345
Junction 1          CTAATCAGAC GAATTCCCGG GAACTGCGGA TGTAGCTCGC CACAGTATCT
LTR6B-X             CTAATCAGAC GAATTCCCGG GAACTGCGGA TGTAGCTCGC CACAGTATCT
LTR6B-YPAR1         CTAATCAGAC GAATTCCCGG GAACTGCGGA TGTAGCTCGC CACAGTATCT
Junction 1.1        CTAATCAGAC GAATTCCCGG GAACTGCGGA TGTAGCTCGC CACAGTATCT
Junction 1.2        CTAATCAGAC GAATTCCCGG GAACTGCGGA TGTAGCTCGC CACAGTATCT
Junction 1.3        CTAATCAGAC GAATTCCCGG GAACTGCGGA TGTAGCTCGC CACAGTATCT
Junction 3          CTAATCAGAC GAATTCCCGG GAACTGCGGA TGTAGCTCGC CACAGTATCT
Junction 4          CTAATCAGAC GAATTCCCGG GAACTGCGGA TGTAGCTCGC CACAGTATCT
Junction 5          CTAATCAGAC GAATTCCCGG GAACTGCGGA TGTAGCTCGC CACAGTATCT
Junction 5.1        CTAATCAGAC GAATTCCCGG GAACTGCGGA TGTAGCTCGC CACAGTATCT

                    ....|....| ....|....| ....|....| ....|....| ....|....|
                       355        365        375        385        395
Junction 1          TATCAGTTAA CTGCATTCTT GGATGTGCTG GGAGTCAGCC TGCACGAGTT
LTR6B-X             TATCAGTTAA CTGCATTCTT GGATGTGCTG GGAGTCAGCC TGCACGAGTT
LTR6B-YPAR1         TATCAGTTAA CTGCATTCTT GGATGTGCTG GGAGTCAGCC TGCACGAGTT
Junction 1.1        TATCAGTTAA CTGCATTCTT GGATGTGCTG GGAGTCAGCC TGCACGAGTT
Junction 1.2        TATCAGTTAA CTGCATTCTT GGATGTGCTG GGAGTCAGCC TGCACGAGTT
Junction 1.3        TATCAGTTAA CTGCATTCTT GGATGTGCTG GGAGTCAGCC TGCACGAGTT
Junction 3          TATCAGTTAA CTGCATTCTT GGATGTGCTG GGAGTCAGCC TGCACGAGTT
Junction 4          TATCAGTTAA CTGCATTCTT GGATGTGCTG GGAGTCAGCC TGCACGAGTT
Junction 5          TATCAGTTAA CTGCATTCTT GGATGTGCTG GGAGTCAGCC TGCACGAGTT
Junction 5.1        TATCAGTTAA CTGCATTCTT GGATGTGCTG GGAGTCAGCC TGCACGAGTT

                    ....|....| ....|....| ....|....| ....|....| ....|....|
                       405        415        425        435        445
Junction 1          CAGTCCTTGA GGAAGGGGCT GCCAGTGAAA GAGCCAAGGT GGAGTCTGGC
LTR6B-X             CAGTCCTTGA GGAAGGGGCT GCCAGTGAAA GAGCCAAGGT GGAGTCTGGC
LTR6B-YPAR1         CAGTCCTTGA GGAAGGGGCT GCCAGTGAAA GAGCCAAGGT GGAGTCTGGC
Junction 1.1        CAGTCCTTGA GGAAGGGGCT GCCAGTGAAA GAGCCAAGGT GGAGTCTGGC
Junction 1.2        CAGTCCTTGA GGAAGGGGCT GCCAGTGAAA GAGCCAAGGT GGAGTCTGGC
Junction 1.3        CAGTCCTTGA GGAAGGGGCT GCCAGTGAAA GAGCCAAGGT GGAGTCTGGC
Junction 3          CAGTCCTTGA GGAAGGGGCT GCCAGTGAAA GAGCCAAGGT GGAGTCTGGC
Junction 4          CAGTCCTTGA GGAAGGGGCT GCCAGTGAAA GAGCCAAGGT GGAGTCTGGC
Junction 5          CAGTCCTTGA GGAAGGGGCT GCCAGTGAAA GAGCCAAGGT GGAGTCTGGC
Junction 5.1        CAGTCCTTGA GGAAGGGGCT GCCAGTGAAA GAGCCAAGGT GGAGTCTGGC
```

```
                    ....|....| ....|....| ....|....| ....|....| ....|....|
                       455        465        475        485        495
Junction 1          GGGCTCTCTT AGCTAAGGGA GAGTCCATTC AGGTGGAAAG AAGGCTAGGT
LTR6B-X             TGGCTCTCTT AGCTAAGGGA GAGTCCATTC AGGTGGAAAG AAGGCTAGGT
LTR6B-YPAR1         GGGCTCTCTT AGCTAAGGGA GAGTCCATTC AGGTGGAAAG AAGGCTAGGT
Junction 1.1        TGGCTCTCTT AGCTAAGGGA GAGTCCATTC AGGTGGAAAG AAGGCTAGGT
Junction 1.2        GGGCTCTCTT AGCTAAGGGA GAGTCCATTC AGGTGGAAAG AAGGCTAGGT
Junction 1.3        GGGCTCTCTT AGCTAAGGGA GAGTCCATTC AGGTGGAAAG AAGGCTAGGT
Junction 3          GGGCTCTCTT AGCTAAGGGA GAGTCCATTC AGGTGGAAAG AAGGCTAGGT
Junction 4          GGGCTCTCTT AGCTAAGGGA GAGTCCATTC AGGTGGAAAG AAGGCTAGGT
Junction 5          TGGCTCTCTT AGCTAAGGGA GAGTCCATTC AGGTGGAAAG AAGGCTAGGT
Junction 5.1        GGGCTCTCTT AGCTAAGGGA GAGTCCATTC AGGTGGAAAG AAGGCTAGGT

                    ....|....| ....|....| ....|....| ....|....| ....|....|
                       505        515        525        535        545
Junction 1          GAGTAGAGGA AAAGGGAGAG TCTAAAAACA GGTTAGTAAA AACCAGGTTG
LTR6B-X             GAGTAGAGGA AAAGGGAGAG TCTAAAAACA GGTTAGTAAA AACCAGGTTG
LTR6B-YPAR1         GAGTAGAGGA AAAGGGAGAG TCTAAAAACA GGTTAGTAAA AACCAGGTTG
Junction 1.1        GAGTAGAGGA AAAGGGAGAG TCTAAAAACA GGTTAGTAAA AACCAGGTTG
Junction 1.2        GAGTAGAGGA AAAGGGAGAG TCTAAAAACA GGTTAGTAAA AACCAGGTTG
Junction 1.3        GAGTAGAGGA AAAGGGAGAG TCTAAAAACA GGTTAGTAAA AACCAGGTTG
Junction 3          GAGTAGAGGA AAAGGGAGAG TCTAAAAACA GGTTAGTAAA AACCAGGTTG
Junction 4          GAGTAGAGGA AAAGGGAGAG TCTAAAAACA GGTTAGTAAA AACCAGGTTG
Junction 5          GAGTAGAGGA AAAGGGAGAG TCTAAAAACA GGTTAGTAAA AACCAGGTTG
Junction 5.1        GAGTAGAGGA AAAGGGAGAG TCTAAAAACA GGTTAGTAAA AACCAGGTTG

                    ....|....
                       555
Junction 1          GGCATTACA
LTR6B-X             GGCATTACA
LTR6B-YPAR1         GGCATTACA
Junction 1.1        GGCATTACA
Junction 1.2        GGCATTACA
Junction 1.3        GGCATTACA
Junction 3          GGCATTACA
Junction 4          GGCATTACA
Junction 5          GGCATTACA
Junction 5.1        GGCATTACA
```

B) Junction 2 and variants

```
                         ....|....| ....|....| ....|....| ....|....| ....|....|
                             5          15         25         35         45
Junction 2               TGTGTTGTAC CCGAGCGAGT TAGAAAAACG CCACACTTTG AGACGATTTA
LTR6B-X                  TGTGTTGTAC CCGAGCGAGT TAGAAAAACG CCACACTTTG AGACGATTTA
LTR6B-YPAR1              TATGTTGTAC CTGAGCGAGT TAGAAAAACG C-----TTTG AGACAAATTA
Junction 2.1             TGTGTTGTAC CCGAGCGAGT TAGAAAAACG CCACACTTTG AGACGATTTA
Junction 6               TGTGTTGTAC CCGAGCGAGT TAGAAAAACG CCACACTTTG AGACGATTTA

                         ....|....| ....|....| ....|....| ....|....| ....|....|
                             55         65         75         85         95
Junction 2               AGAGTCCTTT ATTAGCCGGC GACCGAGAGA CGGCTAACGC TCAAAATTCT
LTR6B-X                  AGAGTCCTTT ATTAGCCGGC GACCGAGAGA CGGCTAACGC TCAAAATTCT
LTR6B-YPAR1              AGAGTCCTTT ATAAGCCAGC GACCGAGAGA CGGCTAATGC TTAATATTCT
Junction 2.1             AGAGTCCTTT ATTAGCCGGC GACCGAGAGA CGGCTAACGC TCAAAATTCT
Junction 6               AGAGTCCTTT ATTAGCCGGC GACCGAGAGA CGGCTAACGC TCAAAATTCT

                         ....|....| ....|....| ....|....| ....|....| ....|....|
                            105        115        125        135        145
Junction 2               CTCGGCCCCG AGGAAGGGGC TTGATTAACT TTTAGATCTT GGTTTAGGAA
LTR6B-X                  CTCGGCCCCG AGGAAGGGGC TTGATTAACT TTTAGATCTT GGTTTAGGAA
LTR6B-YPAR1              CTCGGTCCTG AGGAAGGGGC TTGATTAACT TTTAGATCTT GGTTTAGGAA
Junction 2.1             CTCGGCCCCG AGGAAGGGGC TTGATTAACT TTTAGATCTT GGTTTAGGAA
Junction 6               CTCGGCCCCG AGGAAGGGGC TTGATTAACT TTTAGATCTT GGTTTAGGAA

                         ....|....| ....|....| ....|....| ....|....| ....|....|
                            155        165        175        185        195
Junction 2               GGGGAGGGCG GGGGGTCTAG TGAAAACCAT TTTACAGAAG TAAAGTAGGC
LTR6B-X                  GGGGAGGGCG GGGGGTCTAG TGAAAACCAT TTTACAGAAG TAAAGTAGGC
LTR6B-YPAR1              GGGGAGGGCT GGGGGTCTAG TGAAAACCAT TTTACAGAAG TAAAGTAGGC
Junction 2.1             GGGGAGGGCG GGGGGTCTAG TGAAAACCAT TTTACAGAAG TAAAGTAGGC
Junction 6               GGGGAGGGCG GGGGGTCTAG TGAAAACCAT TTTACAGAAG TAAAGTAGGC
```

```
                    ....|....| ....|....| ....|....| ....|....| ....|....|
                      205        215        225        235        245
Junction 2          AAAAAGTTAA AAGGATAAAT GGTTGCAGGA AAGTAAACAG TTCCAGGTGC
LTR6B-X             AAAAAGTTAA AAGGATAAAT GGTTGCAGGA AAGTAAACAG TTCCAGGTGC
LTR6B-YPAR1         AAAAAGTTAA AAGGATAAAT GGTTGCAGGA AAGTAAACAG TTCCAGGTGC
Junction 2.1        AAAAAGTTAA AAGGATAAAT GGTTGCAGGA AAGTAAACAG TTCCAGGTGC
Junction 6          AAAAAGTTAA AAGGATAAAT GGTTGCAGGA AAGTAAACAG TTCCAGGTGC

                    ....|....| ....|....| ....|....| ....|....| ....|....|
                      255        265        275        285        295
Junction 2          AGGGGCTTTA AGACTATTA- -------TAG ACGCGAGGCT TTGGGCGTTA
LTR6B-X             AGGGGCTTTA AGACTATTA- -------TAG ACGCGAGGCT TTGGGCGTTA
LTR6B-YPAR1         AGGGGCTTTA AGACTATTAC AAGGTGATAG ACGCGAGGCT TTGGGCGTTA
Junction 2.1        AGGGGCTTTA AGACTATTA- -------TAG ACGCGAGGCT TTGGGCGTTA
Junction 6          AGGGGCTTTA AGACTATTA- -------TAG ACGCGAGGCT TTGGGCGTTA

                    ....|....| ....|....| ....|....| ....|....| ....|....|
                      305        315        325        335        345
Junction 2          CTAATCAGAC GAATTCCCGG GAACTGCGGA TGTAGCTCGC CACAGTATCT
LTR6B-X             CTAATCAGAC GAATTCCCGG GAACTGCGGA TGTAGCTCGC CACAGTATCT
LTR6B-YPAR1         CTAATCAGAC GAATTCCCGG GAACTGCGGA TGTAGCTCGC CACAGTATCT
Junction 2.1        CTAATCAGAC GAATTCCCGG GAACTGCGGA TGTAGCTCGC CACAGTATCT
Junction 6          CTAATCAGAC GAATTCCCGG GAACTGCGGA TGTAGCTCGC CACAGTATCT

                    ....|....| ....|....| ....|....| ....|....| ....|....|
                      355        365        375        385        395
Junction 2          TATCAGTTAA CTGCATTCTT GGATGTGCTG GGAGTCAGCC TGCACGAGTT
LTR6B-X             TATCAGTTAA CTGCATTCTT GGATGTGCTG GGAGTCAGCC TGCACGAGTT
LTR6B-YPAR1         TATCAGTTAA CTGCATTCTT GGATGTGCTG GGAGTCAGCC TGCACGAGTT
Junction 2.1        TATCAGTTAA CTGCATTCTT GGATGTGCTG GGAGTCAGCC TGCACGAGTT
Junction 6          TATCAGTTAA CTGCATTCTT GGATGTGCTG GGAGTCAGCC TGCACGAGTT
```

```
                    ....|....| ....|....| ....|....| ....|....| ....|....|
                       405        415        425        435        445
Junction 2          CAGTCCTTGA GGAAGGGGCT GCCAGTGAAA GAGCCAAGGT GGAGTCTGGC
LTR6B-X             CAGTCCTTGA GGAAGGGGCT GCCAGTGAAA GAGCCAAGGT GGAGTCTGGC
LTR6B-YPAR1         CAGTCCTTGA GGAAGGGGCT GCCAGTGAAA GAGCCAAGGT GGAGTCTGGC
Junction 2.1        CAGTCCTTGA GGAAGGGGCT GCCAGTGAAA GAGCCAAGGT GGAGTCTGGC
Junction 6          CAGTCCTTGA GGAAGGGGCT GCCAGTGAAA GAGCCAAGGT GGAGTCTGGC

                    ....|....| ....|....| ....|....| ....|....| ....|....|
                       455        465        475        485        495
Junction 2          TGGCTCTCTT AGCTAAGGGA GAGTCCATTC AGGTGGAAAG AAGGCTAGGT
LTR6B-X             TGGCTCTCTT AGCTAAGGGA GAGTCCATTC AGGTGGAAAG AAGGCTAGGT
LTR6B-YPAR1         GGGCTCTCTT AGCTAAGGGA GAGTCCATTC AGGTGGAAAG AAGGCTAGGT
Junction 2.1        TGGCTCTCTT AGCTAAGGGA GAGTCCATTC AGGTGGAAAG AAGGCTAGGT
Junction 6          GGGCTCTCTT AGCTAAGGGA GAGTCCATTC AGGTGGAAAG AAGGCTAGGT

                    ....|....| ....|....| ....|....| ....|....| ....|....|
                       505        515        525        535        545
Junction 2          GAGTAGAGGA AAAGGGAGAG TCTAAAAACA GGTTAGTAAA AACCAGGTTG
LTR6B-X             GAGTAGAGGA AAAGGGAGAG TCTAAAAACA GGTTAGTAAA AACCAGGTTG
LTR6B-YPAR1         GAGTAGAGGA AAAGGGAGAG TCTAAAAACA GGTTAGTAAA AACCAGGTTG
Junction 2.1        GAGTAAAGGA AAAGGGAGAG TCTAAAAACA GGTTAGTAAA AACCAGGTTG
Junction 6          GAGTAGAGGA AAAGGGAGAG TCTAAAAACA GGTTAGTAAA AACCAGGTTG

                    ....|....
                       555
Junction 2          GGCATTACA
LTR6B-X             GGCATTACA
LTR6B-YPAR1         GGCATTACA
Junction 2.1        GGCATTACA
Junction 6          GGCATTACA
```

Publication Arising From This Study

PLOS GENETICS

RESEARCH ARTICLE

Recombination hotspots in an extended human pseudoautosomal domain predicted from double-strand break maps and characterized by sperm-based crossover analysis

Nitikorn Poriswanish[1,2], Rita Neumann[1], Jon H. Wetton[1], John Wagstaff[1], Maarten H. D. Larmuseau[3], Mark A. Jobling[1], Celia A. May[1]*

1 Department of Genetics & Genome Biology, University of Leicester, Leicester, United Kingdom, **2** Department of Forensic Medicine, Faculty of Medicine Siriraj Hospital, Mahidol University, Bangkok, Thailand, **3** Laboratory of Forensic Genetics and Molecular Archaeology, Department of Imaging and Pathology, KU Leuven, Belgium

* cam5@leicester.ac.uk

OPEN ACCESS

Citation: Poriswanish N, Neumann R, Wetton JH, Wagstaff J, Larmuseau MHD, Jobling MA, et al. (2018) Recombination hotspots in an extended human pseudoautosomal domain predicted from double-strand break maps and characterized by sperm-based crossover analysis. PLoS Genet 14(10): e1007680. https://doi.org/10.1371/journal.pgen.1007680

Editor: Galina Petukhova, Uniformed Services University of the Health Sciences, UNITED STATES

Received: May 2, 2018

Accepted: September 5, 2018

Published: October 8, 2018

Data Availability Statement: The underlying sequence data are available at https://www.ncbi.nlm.nih.gov/sra/SRP155538. All other relevant data are within the paper and its Supporting Information files.

Funding: NP postgraduate research was supported by the HRH Prince Mahidol Inheritance and Staff Development Funds of the Faculty of Medicine Siriraj Hospital, Mahidol University, Thailand. The

Abstract

The human X and Y chromosomes are heteromorphic but share a region of homology at the tips of their short arms, pseudoautosomal region 1 (PAR1), that supports obligate crossover in male meiosis. Although the boundary between pseudoautosomal and sex-specific DNA has traditionally been regarded as conserved among primates, it was recently discovered that the boundary position varies among human males, due to a translocation of ~110 kb from the X to the Y chromosome that creates an extended PAR1 (ePAR). This event has occurred at least twice in human evolution. So far, only limited evidence has been presented to suggest this extension is recombinationally active. Here, we sought direct proof by examining thousands of gametes from each of two ePAR-carrying men, for two subregions chosen on the basis of previously published male X-chromosomal meiotic double-strand break (DSB) maps. Crossover activity comparable to that seen at autosomal hotspots was observed between the X and the ePAR borne on the Y chromosome both at a distal and a proximal site within the 110-kb extension. Other hallmarks of classic recombination hotspots included evidence of transmission distortion and GC-biased gene conversion. We observed good correspondence between the male DSB clusters and historical recombination activity of this region in the X chromosomes of females, as ascertained from linkage disequilibrium analysis; this suggests that this region is similarly primed for crossover in both male and female germlines, although sex-specific differences may also exist. Extensive resequencing and inference of ePAR haplotypes, placed in the framework of the Y phylogeny as ascertained by both Y microsatellites and single nucleotide polymorphisms, allowed us to estimate a minimum rate of crossover over the entire ePAR region of 6-fold greater than genome average, comparable with pedigree estimates of PAR1 activity generally. We conclude ePAR very likely contributes to the critical crossover function of PAR1.

funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

Author summary

95% of our genome is contained in 22 pairs of chromosomes shared by all humans. However, women and men differ in their sex chromosomes: while women have two X chromosomes, men have an X and a smaller, sex-determining Y chromosome. To ensure correct partition of X and Y into sperm, genetic exchange (crossover) must occur between these very different chromosomes in a short, shared region. The location of the boundary of this region was thought to have been conserved since before the divergence from old world monkeys at least 27 million years ago, but recently it has been shown that some human males carry an extended version on their Y chromosomes, thanks to the transposition of a piece of DNA from the X chromosome. Here, we asked if genetic exchange occurs in this newly extended region. To do this, we used previously published information that signposted the positions within the X chromosome segment which exhibit the hallmarks of crossover initiation. We then sought direct evidence of crossover in the sperm of men carrying the extension. This work showed that the signposts were accurate, pointing to frequent crossover in this novel shared sex-chromosomal domain.

Introduction

The major pseudoautosomal region (PAR1), located at the tips of the short arms of the human sex chromosomes, is a region of interchromosomal homology [1, 2]. In contrast to its smaller counterpart (PAR2) on the long arms of the sex chromosomes [3], PAR1 plays an essential role in male meiosis by supporting pairing and obligatory exchange between the X and Y [4], failure of which can lead to sex-chromosomal aneuploidy such as Klinefelter syndrome (47, XXY), and is associated with increased infertility [5–7]. The human PAR1 is ~2.7 Mb in length, and until recently it was thought to have been stable during most of primate evolution [8]. Indeed, since its initial molecular characterization, it was widely accepted that the boundary was fixed approximately at its present location before the divergence of the old world monkeys and great apes 27–32 million years ago [9] and is delineated by an *Alu* element insertion on the human Y chromosome. However, despite this, there is evidence that its boundary, PAB1, has shifted distally in the past, as the proximal 240 bp of sex-specific DNA shows 77% sequence similarity between the human X and Y [8,10]. More recently, direct evidence of pseudoautosomal region plasticity came from a chance discovery during an aCGH (array comparative genomic hybridization) screen for copy number variation (CNV) in ~4,300 patients with developmental disorders, which showed that a small subset of men carry an extended PAR1 (ePAR): this demonstrates that the PAR1 boundary is not static, but polymorphic in modern humans [11].

The Mensah *et al*. study [11] established that creation of the ePAR involved transfer of ~110 kb of X-chromosomal PAR1-proximal sequence and concomitant duplication of a ~5-kb portion of PAR1 to the Y chromosome. Furthermore, this insertional translocation was deemed most likely to be the result of non-allelic homologous recombination (NAHR) mediated by flanking ~550-bp LTR6B elements, and consistent with this, a family segregating the predicted reciprocal ~115-kb deleted form of the X was also identified [11] (Fig 1A). In contrast to most males but akin to females, men carrying the ePAR have two full-length copies of the apparently clinically irrelevant *XG* blood group gene [12], as well as two copies of the *GYG2* gene, encoding a precursor for glycogen synthase particularly important in the liver [13,14] (Fig 1B). Interestingly, both these genes escape inactivation in females [15]. In the Mensah *et al*. study, the ePAR was observed in 15 independent Belgian and French families: all Belgian ePAR Y

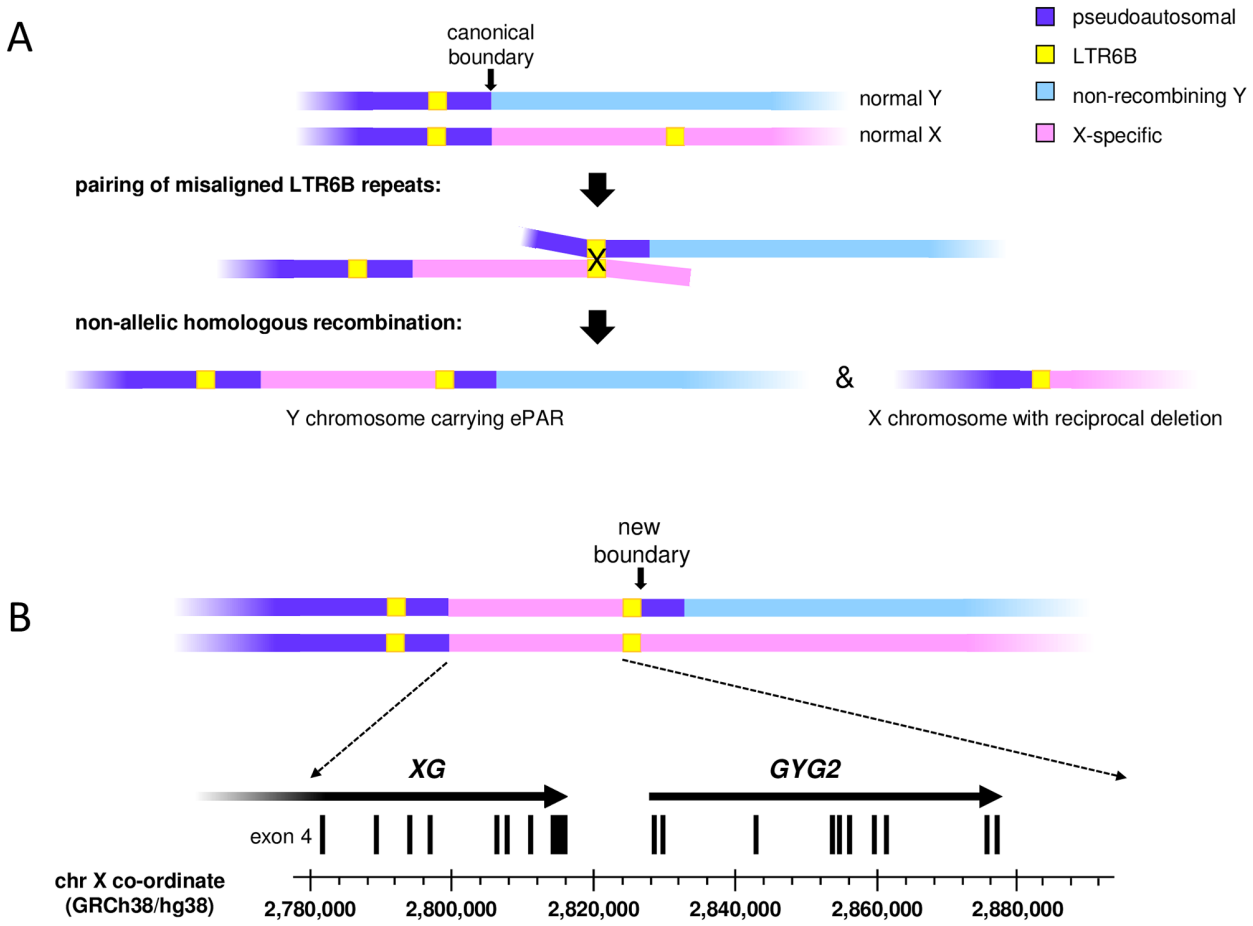


Fig 1. Schematic representation of formation and organization of the ePAR. (A) Normal pairing of the short arms of the X and Y chromosomes in male meiosis is limited to PAR1 (purple) such that homologous recombination can occur up to the canonical boundary as marked. However, mispairing of the PAR1 LTR6B element (yellow box) carried on a Y chromosome (blue) with one located more proximally on an X chromosome (pink) can result in non-allelic homologous recombination and the generation of gametes containing either an ePAR-carrying Y chromosome, or the reciprocal deleted X chromosome. Schematics not to scale. (B) Pairing and therefore homologous exchange between an ePAR Y chromosome and a normal X chromosome is predicted to extend proximally to a new boundary. The first three exons of the *XG* blood group gene fall within PAR1, while the remaining exons are carried on the X chromosome; men carrying the ePAR thus have two full-length XG genes like females. Similarly, whilst most men are hemizygous for the *GYG2* gene, ePAR carriers have two copies of this gene.

https://doi.org/10.1371/journal.pgen.1007680.g001

chromosomes belonged to one sub-haplogroup, I2a (I-P37.2), while those of the two first-degree relative French carriers belonged to a sub-haplogroup within the distantly-related lineage R1b, namely R-P312 [11]. This indicated that the creation of ePAR is recurrent and has occurred at least twice based on the global Y-chromosomal phylogeny [16].

Mensah *et al.* presented indirect evidence that the translocated region within the ePAR actually functions pseudoautosomally. PacBio sequencing of <5% of the total ePAR indicated that at least two haplotypes exist amongst the haplogroup-I2a men; these were interpreted as a consequence of recombination between X and ePAR rather than mutation accumulation, because they differed by twelve single nucleotide polymorphism (SNP) variants all of which are also observed on X chromosomes [11]. More recently a gradual decline in X-chromosome genetic diversity spanning the canonical boundary was noted [17]: this contrasts with the expected abrupt drop at the boundary given the lower effective population size of strictly X-linked sequences (*i.e.* two copies in females but only one in males) compared with a truly

pseudoautosomal sequence (two copies in both sexes) and provides further evidence consistent with the ePAR supporting exchange between the X and Y.

Here we build on these initial studies to seek direct evidence that the ePAR supports meiotic exchange by identifying *de novo* sperm recombinants (crossovers [COs], and noncrossovers [NCOs]) that map to this region from two men carrying ePAR-bearing Y chromosomes belonging to the I2a haplogroup. Since double-strand breaks (DSBs) induced by the protein SPO11 are known to initiate meiotic recombination [18], we target two subregions of the X chromosome involved in the translocation that are known, via single-stranded DNA sequencing (SSDS) data, to support DSBs in the male germline of presumed non-ePAR-carrying individuals [19]. Furthermore, we sequence >90% of the entire translocated region to extend our understanding of the recombinational history of the region as a whole.

Results

Identification of sperm donors for analysis of recombination in ePAR

The ePAR has been found in two Y-chromosome haplogroups, I2a and R1b, that are frequent in Europe [20, 21] so we focused on North European semen donors in our collection. Man 20 had previously been found to have a duplication of at least 17 kb of X-chromosome sequence that spanned PAB1 and was therefore a candidate ePAR-carrier (S1 Fig). A second candidate, man 53, was identified on the basis of a similar Y microsatellite or Short Tandem Repeat (Y-STR) haplotype and therefore predicted to share Y-chromosome haplogroup I2a-L233 with man 20. ePAR status was confirmed by sequencing of the proximal insertion junction; both sperm donors carry the "Junc1" sequence shared by eight of the nine independently sampled haplogroup I2a ePARs studied by Mensah *et al*. [11].

Targeting subregions of ePAR for de novo recombination analysis

Since human recombination events cluster into narrow 1–2 kb-wide hotspots [22–27] we sought to identify potential hotspot sites within the ePAR. Hotspot location is largely determined by PRDM9 [28–30], a presumed chromatin-remodelling protein, which binds DNA via its highly polymorphic zinc finger domain, and thus targets the induction of DSBs to specific locations (for a review see [31]). As both ePAR-carrying sperm donors were known to be homozygous for the common A-type zinc finger allele at the *PRDM9* locus, we considered the distribution and strength of meiotic DSB clusters induced by the PRDM9 A allele on the X chromosome region that makes up the ePAR, as ascertained by read depth in a previous SSDS study using testis biopsy material from presumed normal-PAR1-carrying men [19] (Fig 2A). We also considered the distribution of so-called hotspot motifs believed to be the cognate binding sites for the most common form of PRDM9 [30], as well as the SNP density across the entire region as reported in dbSNP [32], because recombination and DNA diversity often show a positive correlation [33]; neither showed a clear correspondence to the SSDS signals (S2 Fig). Finally, since our ability to detect recombinants is wholly reliant on informative SNPs in the sperm donors under study, we determined the distribution of heterozygous SNPs for >90% of the entire ePAR in both man 20 and man 53 using Ion Torrent sequencing (Fig 2B). These data suggested that recombination assays could be developed for both donors in each of two DSB clusters, as indicated in the figure. The distal assay region was located ~ 2.6 kb proximal to the canonical X-specific PAB1 and coincided with a moderately strong PRDM9-A-induced DSB cluster. The proximal assay region was some 86 kb upstream of this and coincided with the strongest DSB cluster as determined by SSDS read depth.

We also compared the male meiotic DSB data with the pattern of historical female-dominated X-chromosomal recombination activity as determined by linkage disequilibrium (LD),

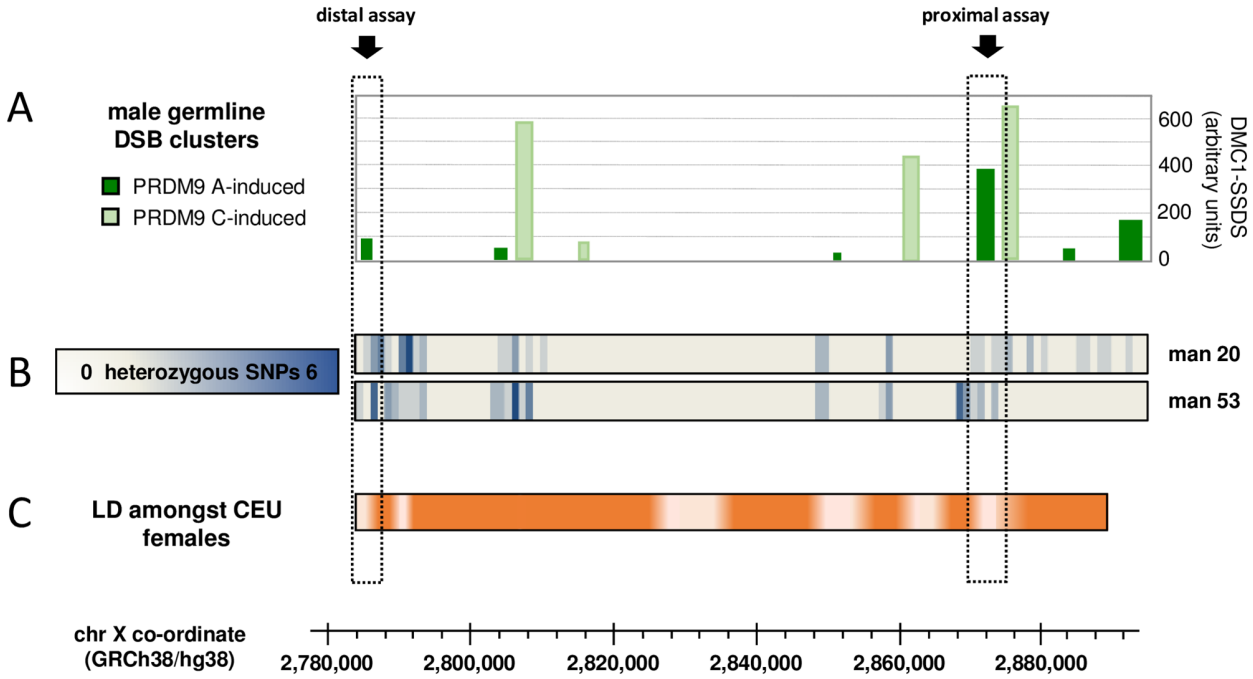


Fig 2. Choosing target regions within ePAR to assay for male germline *de novo* recombinants. (A) Distribution and intensity of DSB clusters that fall within the X-derived portion of the ePAR. Clusters were defined [19] by anti-DMC1 SSDS using testis biopsies from five men and designated as being induced by PRDM9 A (dark green) or PRDM9 C (light green); DSB strength is shown as the mean across relevant individuals using the arbitrary values reported in the original work [19]. (B) Frequency of heterozygous SNP markers per 1-kb interval identified in each of the two ePAR-positive sperm donors (man 20, man 53) as determined by Ion Torrent sequencing. (C) Linkage disequilibrium (LD) heat map derived from the 50 CEU females from the 1000 Genomes Project [34] (more intense orange equates to stronger LD). Data are based on |D´| values determined from SNPs with minor allele frequency >0.2 that passed tests for Hardy-Weinberg equilibrium specifically derived for markers on the X chromosome [72]; these stringent criteria meant that LD corresponding to the most proximal portion of the ePAR interval could not be examined. Scaling is shown with respect to GRCh38/hg38, and the two chosen assay intervals, distal and proximal, are indicated by the dashed boxes and arrows. See also S2 Fig.

https://doi.org/10.1371/journal.pgen.1007680.g002

using SNP data from the 1000 Genomes Project [34] (Fig 2C). Both the distal and proximal assay intervals coincide with regions of LD breakdown, suggesting that these intervals have been active in the female germline too. In fact, five of the six regions of LD breakdown were found to correspond to either PRDM9-A- or PRDM9-C-induced DSB clusters (PRDM9 C and related alleles are known to activate different subsets of hotspots compared with the A allele, and collectively encode the next most common class of PRDM9 protein [35]). Conversely, only six of the ten male DSB clusters coincide with historical recombination activity in the female germline. We found no clear relationship between DSB strength and LD breakdown; of the six PRDM9-A-induced DSB clusters, the two weakest map to regions of historical recombination in the female germline, but the next two weakest do not.

Sperm crossovers in the distal region cluster into a classical hotspot

Each sperm donor was found to be heterozygous for at least two SNPs both upstream and downstream of the DSB cluster in the distal region. This allowed development of a full CO assay for each, whereby forward allele-specific primers (ASPs) from one parental haplotype are used in conjunction with reverse ASPs from the opposite haplotype to selectively amplify *de novo* recombinants from multiple PCR reactions each containing several hundred molecules [36]. This is an efficient means by which to both estimate CO frequencies and to recover CO molecules for breakpoint mapping by subsequent typing of intervening SNPs.

Reciprocal assays were carried out for each of the two men. Collectively, 200 *de novo* COs were isolated and mapped from a total of 168,800 sperm molecules screened. Ninety-five percent of events clustered into a 1.3-kb-wide interval, entirely consistent with both autosomal and pseudoautosomal sperm CO hotspots [22,23,37], and with the peak of CO activity almost exactly mapping to the centre of the DSB cluster (Fig 3A). Despite a shared distribution of events (the inferred centre points of each donor's distribution are estimated to be offset by <10 bp), the two men exhibited a ~4-fold difference in rate (man 53 RF = 0.21% (95% CI 0.18–0.24%), man 20 RF = 0.05% (95% CI 0.03–0.06%), $P << 0.0001$, 2-tailed goodness of fit test). This is within the observed range noted at other characterized sperm CO hotspots, controlling for both *PRDM9* status and *cis*-effects influencing initiation (see below) [38], and is comfortably within the 30-fold range of DSB strength as measured by SSDS read depth across the five men tested over this interval [19].

Ordinarily, reciprocal events should show a 50:50 ratio of alleles at heterozygous SNP sites; however, several CO hotspots have been shown to exhibit significant transmission distortion (TD) between alleles for markers close to the hotspot centre [37,39–42]. This phenomenon is most readily explained by differences in the frequency of recombination-initiating DSBs between the two parental haplotypes, since the repair of such lesions uses the intact homologue which in turn leads to over-transmission of the recombination-suppressing haplotype. TD is also referred to as CO asymmetry because the centre point of events is shifted between the reciprocal orientations even though the rates remain the same. Man 53 showed evidence of TD at the rs1970797 C/T polymorphism, with significant over-transmission of the T-allele (0.62 *cf*. 0.50, $P = 0.008$, two-tailed exact binomial test) and a displacement of the centre points of the reciprocal distributions of 126 bp. This SNP is the closest informative marker to the overall hotspot centre (Fig 3B). Given the CO rate estimate for man 53 and this level of TD observed amongst his COs, this equates to a gametic ratio of 50.024:49.976 and demonstrates that this hotspot, like some autosomal hotspots [39,40], is subject to a form of meiotic drive that will ultimately lead to its demise [43].

Detection of sperm crossover and non-crossover events in the proximal region

The distribution of informative markers for both men in the proximal interval was such that similar CO assays could not be developed without requiring >20 kb amplicons that would at best result in very low PCR efficiencies. Instead, we designed assays in which ASPs are used in conjunction with universal primers to amplify one haplotype, and recombinants are detected by the presence of alleles from the non-amplified haplotype [36]. Since the latter is dependent on hybridization, this approach is less efficient as pool sizes are of the order of tens, not hundreds, of sperm per PCR, but it offers the advantage that both CO and NCO events can be detected.

Across the two men, a total (*i.e*. CO+NCO) of 120 recombinants were detected from 21,690 sperm, and comparable recombination fractions were noted for each (man 20, 0.60% (95% CI 0.47–0.77%) and man 53, 0.51% (95% CI 0.39–0.66%), $P > 0.05$, 2-tailed goodness of fit test). Despite the need to design different assays (Fig 4), in both cases, the most common type of event involved a switch of haplotype only at the terminal marker adjacent to the universal primer. In such cases it is impossible to distinguish COs from NCOs; furthermore, from analysis of other recombination hotspots, both are expected to co-localise, albeit with varying proportions [24,37,44,45]. In order to gain insight into the hotspot morphology, we therefore arbitrarily assigned half of such events as COs. Under this scenario, the proximal DSB cluster encompasses a 1.1-kb-wide hotspot with a peak activity of ~385 cM/Mb (see Fig 4A).

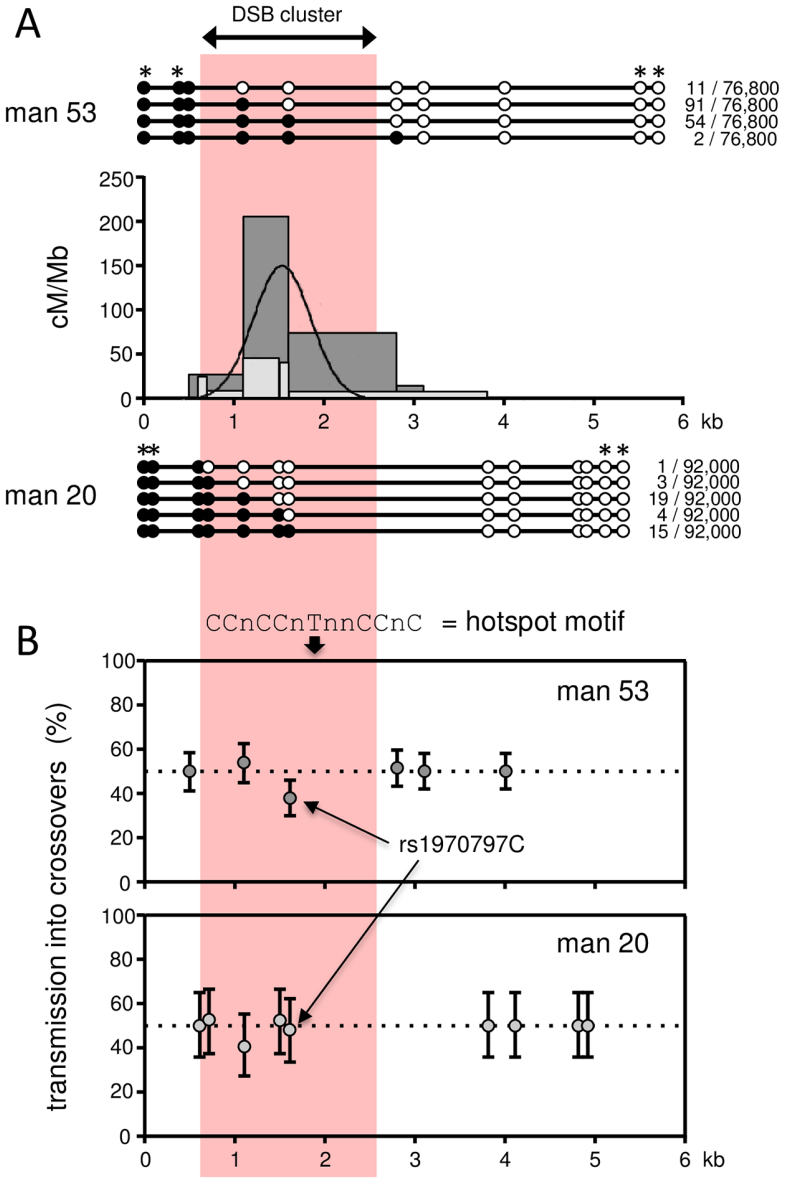


Fig 3. *De novo* sperm crossover activity in the distal region. (A) Sperm CO profiles for each of the two ePAR-carrying men analysed. A total of 158 recombinants were recovered from 76,800 sperm from man 53 using reciprocal crossover assays (*i.e.* Ab plus Ba COs, where AB and ab are the parental haplotypes) compared with 42 from a total of 92,000 sperm from man 20. Recombination activity expressed in cM/Mb along the assayed intervals is shown in the central graphs with the crossover activity of man 53 shown by the dark grey histogram and that of man 20 by the light grey histogram. The combined least-squares best-fit normal distribution for both men is shown by the black curve. The recovered CO structures together with their frequencies are shown above (man 53) and below (man 20) with heterozygous SNP locations represented by circles. SNPs marked with an asterisk were exploited for recombinant recovery (see Methods and S6 Table). The pink panel spans the interval in which male meiotic DSBs were previously mapped by DMC1-SSDS [19] and coincides with the peak CO activity determined from *de novo* sperm events in this study. (B) Transmission frequencies of SNP alleles into reciprocal COs with 95% credible intervals determined by Bayesian analysis. Transmission of the 'strong' allele (C or G) is shown for transition polymorphisms and transmission of the purine allele (A or G) is shown for the transversion polymorphisms. The upper panel shows the transmission data from man 53, the lower panel those for man 20. All markers with the exception of rs1970797 in man 53 were consistent with the expected 50:50 transmission of the two alleles into reciprocal events. This polymorphism lies 126 bp proximal to the predicted hotspot centre and 210 bp distal of the closest hotspot motif [73], as indicated at the top of the upper panel. CO asymmetry has previously been noted at hotspots that do not contain obvious matches to this motif, yet are nonetheless specifically activated by PRDM9 A [38,42]. Note that our failure to observe asymmetry for man 20 may simply be a consequence of the small number of COs detected for this donor.

https://doi.org/10.1371/journal.pgen.1007680.g003

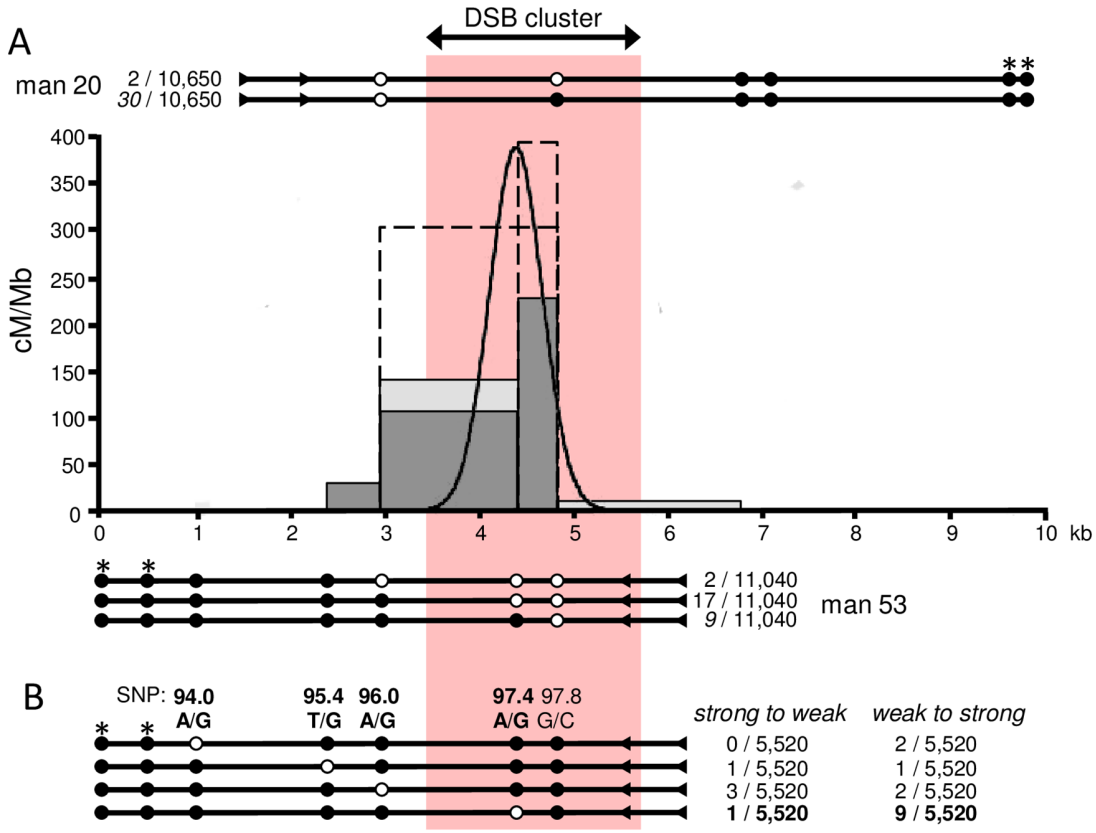


Fig 4. *De novo* recombination in the proximal region. (A) Sperm CO profiles in relation to the proximal DSB cluster as determined by DMC1-SSDS [19] (pink panel) for man 20 (light grey histogram) and man 53 (dark grey histogram), with the combined least-squares best-fit normal distribution shown by the black curve. As for Fig 3, data from reciprocal assays have been pooled and the recovered structures and their frequencies for each man are shown above and below the histograms with informative SNPs represented by circles. In these assays, ASPs were designed against the SNPs marked with asterisks and were used in conjunction with universal primers (triangles) to selectively amplify each parental haplotype; recombinants were then detected by probing for the alleles of the opposite haplotype represented here by white circles (see Methods, S6 and S7 Tables). Note that CO events involving only the terminal marker closest to the universal primers are indistinguishable from NCO events in this assay, so we arbitrarily designated half of such events as COs in these cases (numbers given in italics) but indicate with dashed boxes in the graph how the profiles would appear should all such events actually be COs. In the latter case, the hotspot width would be reduced by 250 bp, the centre point would be shifted proximally by 116 bp, and the peak activity would be ~ 830 cM/Mb. (B) Testing for GC bias amongst NCOs. Of the four informative SNPs for man 53 that carry a 'weak' and 'strong' allele, SNP 97.4 shows over-transmission into NCOs of the 'strong' allele G relative to the 'weak' allele A ($P = 0.011$, one-tailed binomial exact test). This SNP lies 97 bp proximal to the hotspot centre as shown by the black curve in (A). Whilst we cannot be sure of the number of NCOs involving the terminal marker SNP 97.8, both alleles at this SNP base pair with three hydrogen bonds (*i.e.* are 'strong') and there is no evidence of disparity between the orientations assuming at least half the terminal recombinants are NCOs (*i.e.* 9). For man 20, terminal marker SNP 96.0 recombinants were recovered in the two orientations with similar frequencies, again suggesting an absence of TD. Note a further two NCOs each affecting a different single site (SNP 99.8 or SNP 100.1) were also recovered for this man but are not depicted in this figure.

https://doi.org/10.1371/journal.pgen.1007680.g004

Sperm recombination data from the two assay intervals show comparable trends to those observed by Pratto *et al.* for the two DSB clusters. However measured, the proximal region shows more modest variability in recombination (at most a 1.2-fold difference between the sperm donors), compared with the distal region where a 4-fold difference in CO was noted, whilst DSB strength differed ~7- and ~30-fold respectively amongst the four men analysed by Pratto *et al.* [19]. Similarly, overall higher rates of recombination are observed in the proximal than distal region, though at best there is only ~12-fold difference compared with ~50-fold

noted in mean DSB strength. Of course, only DSBs repaired using the homologue can be identified in our assays, and NCO events that do not encompass informative SNPs will go undetected but still contribute to the single-stranded DNA signal used to generate the DSB maps.

Unsurprisingly, given the distribution of markers, all twenty-one events that could be scored unambiguously as NCOs encompassed just a single polymorphic site with maximal conversion tract lengths ranging from 1853–2812 bp. Nineteen of these were observed for man 53 with peak numbers seen at SNP 97.4, the marker that lies nearest to the predicted hotspot centre (Fig 4B), entirely consistent with previous characterization of human meiotic NCOs [37,44,45]. Indeed, the closest adjacent marker to SNP 97.4 lies just 413 bp away, yet no co-conversions were observed suggesting, as seen in other studies, that most of the NCO tracts not only occur at the centre of the CO hotspot but are in fact short [44].

Assays were carried out in both orientations so it was possible to also test for TD in the proximal region. In contrast to the distal assay, none was observed amongst the COs for either man; however, significant bias was observed amongst the NCOs for the central-most SNP, 97.4, for man 53. Nine of the ten NCOs spanning SNP 97.4 contained the G- rather than A-allele indicating a preferential repair of 'weak' to 'strong' base pairs (Fig 4B) as noted in other studies [46]. TD confined to NCOs has previously been noted at two autosomal hotspots indicating differences in CO and NCO heteroduplex formation and/or mismatch repair; it is noteworthy that in both of these cases there was also a significant GC bias [45].

Inferring past recombination events throughout ePAR

To gain a comprehensive understanding of the recombination history of ePAR we set out to sequence the entire region for the two sperm donors, six of the originally reported families of the Mensah *et al*. study [11], plus a further three carriers including one who is part of a CEPH pedigree (see Methods). Including family members to aid with subsequent phasing of alleles, this equated to twenty individuals, and ten independent I2a ePARs plus one R1b ePAR (S1 Table). We designed overlapping amplicons spanning the 110-kb transferred region of the X and sequenced them on an Ion Torrent platform to a mean read depth of 300x. We observed some unintended amplification from the long arm of the male-specific region of the Y (see Methods); this technical issue reflects the fact that the region of the X chromosome that transferred to form the ePAR shares a common evolutionary origin with this proximal portion of Yq, dating back ~30 Mya [47]. We therefore excluded approximately 9 kb from further analysis and determined SNP haplotypes for the remaining ~92% of the ePAR using the program PHASE [48,49]. We made use of family relationships where appropriate to determine which haplotype most likely corresponded to the ePAR. Two of the ePAR men for whom there were no first-degree relatives to analyse shared the same uncommon British surname indicative of shared ancestry (~1000 carriers in Great Britain in the year 1998) [50]. Whilst genealogical records suggest a putative common ancestor more than five generations ago, Y-STR profiling provides evidence of close paternal line relatedness of these two men (S4 Table); we took this into account when assigning their ePAR haplotypes.

We focused on SNPs that overlap with those in the 1000 Genomes Project dataset for CEU (Utah Residents [CEPH] with Northern and Western European Ancestry) and GBR (British in England and Scotland), reasoning that Western European X chromosomes were most relevant for understanding the history of ePARs identified in the same geographical region (Fig 5). To aid interpretation, we focussed on SNPs that fall outside of the DSB clusters, as the signature of CO is most easily detected by new combinations of pre-existing and well-defined flanking LD haplotype blocks. This left a core set of 213 markers (S2 Table) split across nine regions or "blocks", ranging in size from 558 to 16,143 bp. Of the ten independently sampled I2a ePARs,

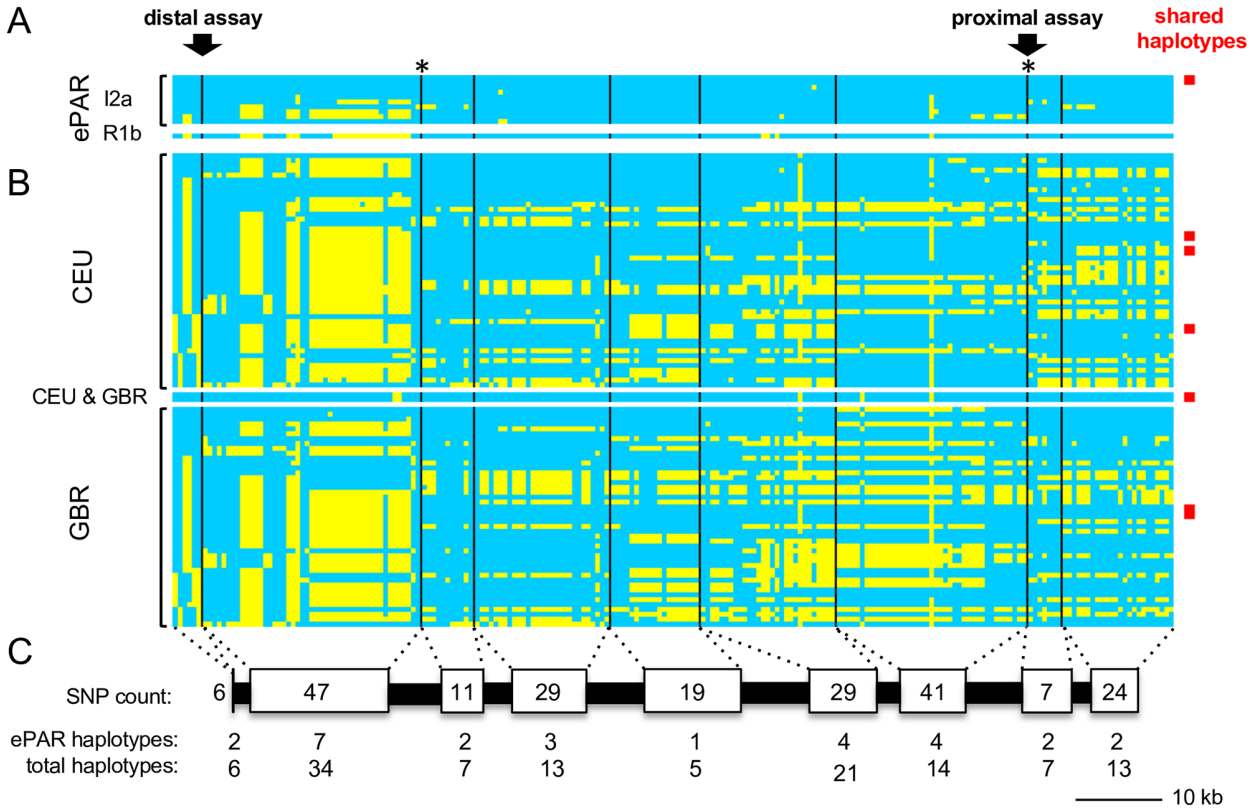


Fig 5. Comparison of inferred ePAR haplotypes with phase-known haplotypes from the corresponding region of the X chromosome. (A) SNP haplotypes from each of the eleven independently sampled ePARs (ten from the haplogroup I2a Y chromosome lineage and one from the R1b Y lineage) are shown in rows, clustered according to the distal haplotype block. Individual 6889_01 is shown at the top with all his alleles colour-coded blue; yellow denotes the alternative SNP alleles not carried by this man. Black vertical lines correspond to the relative locations of mapped PRDM9 A and C DSB clusters; in two instances, marked by asterisks, an A and C cluster lie in very close proximity. Arrows indicate the distal and proximal sperm recombination assay regions and the red box indicates a second ePAR that is identical to that of 6889_01. In total, nine of the ten ePAR haplotypes are unique to this dataset. (B) Phased X haplotypes from the 1000 Genomes Project [34] for the 49 CEU males and 46 GBR males. One haplotype is shared between the two sample sets as indicated. In addition, three pairs of identical X haplotypes were noted among the CEU and one X haplotype was found to be carried by three different GBR men (red boxes). In total, 42 of the 46 CEU haplotypes and 43 of the 44 GBR haplotypes are unique. None of these X haplotypes matches any of the ePAR haplotypes. (C) Relative scaling of the regions depicted together with summary count of the number of SNPs, number of ePAR haplotypes and the corresponding total number of haplotypes seen amongst the ePAR, CEU and GBR datasets per block of SNPs.

https://doi.org/10.1371/journal.pgen.1007680.g005

only two were found to have the same compound haplotype which was designated as the consensus (Fig 6A). The eight remaining I2a ePARs differed by up to four of the nine blocks (mode and median = 2) with changes from the consensus ranging from 1 to 29 SNP sites per block (mode = 1, median = 2). No complete matches were observed with phase-known X chromosomes from either the CEU or GBR males, though matches at the level of individual blocks were observed (9/19 that differ from the consensus, Fig 6B, S3 Table).

The simplest explanation of the diversity of the I2a ePARs would be that each unique haplotype is the outcome of a single, different CO event with an X chromosome (Fig 6). We therefore looked for matches for the predicted incoming (i.e. strictly X-linked) haplotype among the 95 phase-known CEU/GBR male X-haplotypes, but failed to identify any. Only five different compound haplotypes were seen more than once amongst this data set, consistent with high diversity in this region [17] and so it is possible that single exchanges involving unsampled X chromosomes could account for our observed ePAR haplotypes. Using published mutation rates for 23 Y-STRs we estimated the time to most recent common ancestor

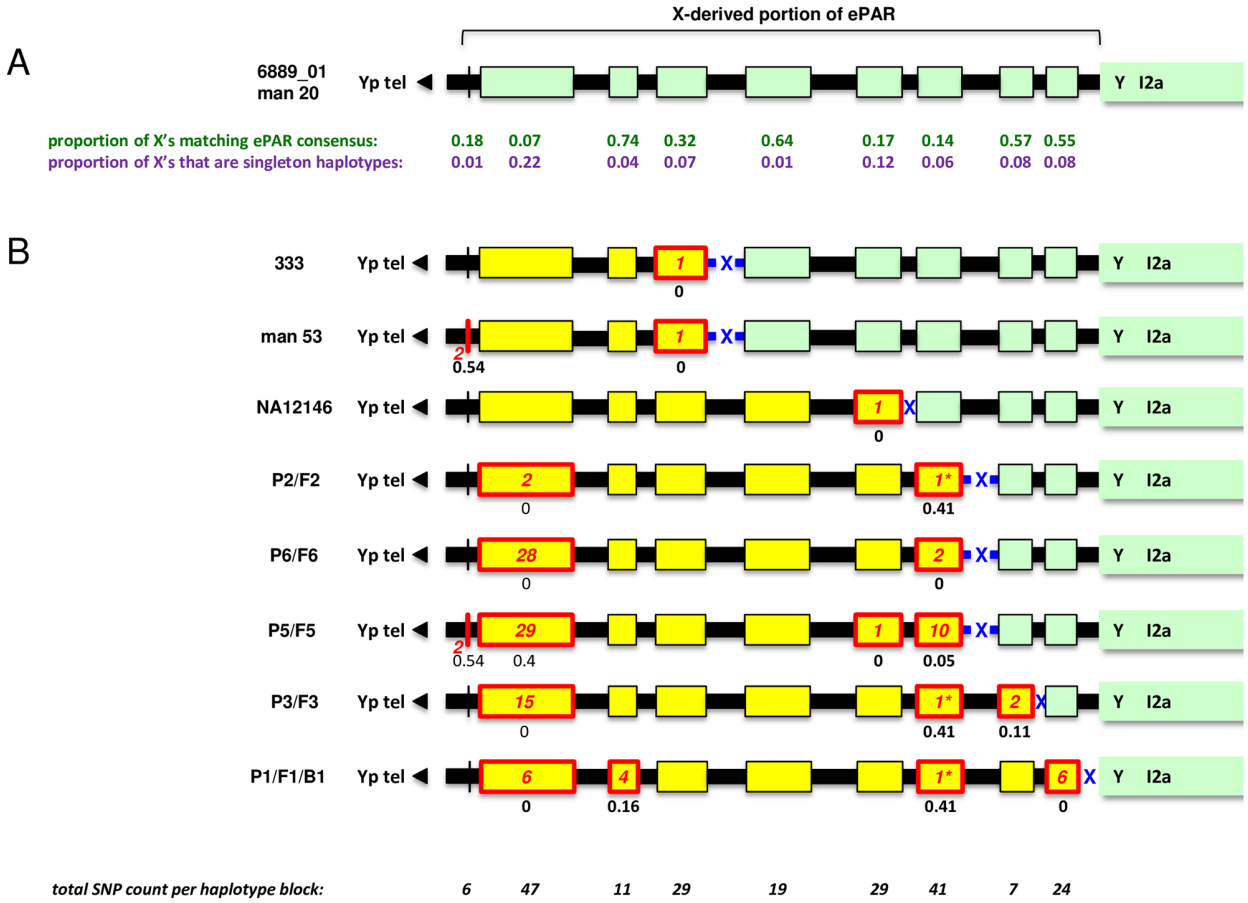


Fig 6. Simple interpretation of the I2a ePARs. (A) Schematic of consensus X-derived portion of ePAR carried by individuals 6889_01 and man 20. Green boxes with black outlines represent the shared haplotypes at each of the nine blocks of SNP markers whilst the intervening black boxes coincide with mapped PRDM9 A and C DSB clusters [19]; widths of all boxes are proportional to their length. The black triangle to the left points towards the canonical 2.7-Mb PAR1 and ultimately to the Yp telomere; the start of Y-specific I2a sequence is shown to the right. The frequency of phase-known X haplotypes among the 95 CEU+GBR males that match the modal ePAR haplotype for each block of SNPs are shown in green; the frequencies of singleton haplotypes amongst the same are shown in purple. (B) The remaining eight I2a ePARs, assuming they are the result of a single crossover between the consensus and an incoming X-linked haplotype depicted by yellow boxes (crossover interval shown with blue cross). Boxes with red outlines show haplotype blocks that differ from the consensus with the number of SNP changes shown in red; asterisks identify three haplotype blocks that differ from the consensus by the same single base pair change. Black numbers beneath boxes indicate the observed frequency of the non-consensus haplotype amongst the 95 phase-known X haplotypes from the CEU+GBR males. Total SNP counts per block are shown in italics at the bottom.

https://doi.org/10.1371/journal.pgen.1007680.g006

(TMRCA) for the I2a ePARs of our ten sequenced lineages at 3,877 ± 779 yrs, (S4 Table) [51–53], equating to 125 generations averaging 31 years. Assuming a minimum of eight recombination events to account for the nine extant I2a ePAR variants amongst the ten lineages examined, we thus obtain a minimum recombination rate of 0.64% (i.e. 8/(125 x10)).

This recombination rate is likely an underestimate of the true rate for two reasons; not all ten sequenced lineages radiated in one generation directly from the common ancestor (S3 Fig), and we have no way of definitively identifying multiple recombination events in these data. Interestingly, the two most diverged I2a sub-haplogroups also carry the most differentiated ePARs and importantly the variation from the consensus extends close to the proximal boundary, so it is entirely possible that these ePARs have experienced additional distal recombination events. Conversely, although the 23-Y-STR haplotypes of two of the lineages differ by a single repeat at just one STR, suggesting very recent shared ancestry, their respective ePAR

haplotypes differ greatly, implying that a recent recombination has occurred close to the new boundary (P2/F2 and P3/F3 in Fig 6 and S3 Fig). This recent shared ancestry is also confirmed by the fact that both families have an identical surname that has a low frequency in Belgium (*ca*. 550 carriers in 2008) suggesting a close genealogical relatedness in the patrilineal line [54]. Nonetheless, our minimum recombination estimate is compatible with the sperm CO data for the two intervals surveyed, and suggests that the entire ePAR has a recombination rate of at least six times genome-average (~5.8 cM/Mb, compared with a genome-average male recombination rate of at most 0.9 cM/Mb [55]). The canonical PAR1 supports a male crossover rate seventeen times higher than genome-average and four times greater than the next most recombinogenic region of comparable physical length [56]; our data therefore demonstrate that the ePAR is an active, recombinationally-hot domain in the male germline.

Discussion

Despite comprising less than 5% of the human Y chromosome, PAR1 plays a fundamental role during male meiosis. Indeed, failure of the human X and Y chromosomes to pair and exchange genetic information within this region is not only associated with paternal inheritance of sex-chromosomal aneuploidy but also intimately linked with male fertility *per se* [5–7, 57, 58]. Our appreciation of the latter has been furthered by mouse studies demonstrating that high levels of achiasmate X and Y trigger a spindle assembly checkpoint resulting in an apoptotic response [59]. Increased infertility in male mice has also been linked with disruption of sequence homology across the mouse PAR [60], demonstrating the importance of the length of sequence identity for successful X-Y pairing. Given the recent discovery that the human PAR1 varies in length among humans [11], we therefore sought to examine the recombination behaviour of this proximally extended 110-kb X-derived ePAR.

We measured recombination activity in the ePAR by directly examining gametic DNA from appropriate sperm donors. Since thousands of sperm can be screened per donor, this approach not only allows efficient estimation of rates (down to 0.0004%, [36]) but can give detailed insight into the dynamics of recombination, even when only one or two men are available for study. Such analyses have been instrumental in establishing that human meiotic recombination, including that in PAR1, is not randomly distributed, but clusters into narrow 1-2-kb-wide intervals, or hotspots [61, 62]. However, this approach, which is based on long PCR, is not easily scalable to even modestly-sized genomic regions such as the ePAR, so here we exploited published human male meiotic DSB maps [19] in order to target tractable sub-regions for bulk sperm analysis. *De novo* recombinants were detected in both sub-regions analysed, and their frequencies, distributions and characteristics were entirely consistent with classic hotspots shaping the recombination landscape of the ePAR. We complemented these sperm data by examining ePAR diversity amongst men of the I2a Y sublineage, estimating that the entire region has a historical recombination frequency of at least six times the male genome average, and thus we conclude that the ePAR very likely contributes to the critical crossover function attributed to the canonical PAR1. Whether this expansion leads to a selective advantage, as proposed for rearrangements altering the mouse PAR (see [60]), remains to be seen.

Sperm DNA approaches have given unprecedented insight into the dynamics of recombination at the sub-kilobase scale, ranging from inter-individual differences in activity [35, 38] through to haplotype-specific differences for a given man [37, 39] but have traditionally relied on pedigree or LD analysis to identify suitable target regions [22,42]. Here, for the first time, we primarily made use of recombination initiation maps to guide our efforts. As noted on a genome-wide scale, the male-specific DSB clusters on the X chromosome relating to the ePAR show reasonable correspondence with LD-based hotspot prediction (6/10 [60%] DSB clusters

map to LD hotspots, *cf*. 73% genome-wide, whereas 5/6 [83%] LD hotspots in the region map to DSB clusters, *cf*. 68% genome-wide [19]). Since the LD landscape in this region is dominated by female recombination, this indicates that the chromatin structure of this portion of the X chromosome during prophase I in most males must be very similar to that in females, though of course repair of such DSBs in these non-ePAR carriers must be via the sister chromatid. Since we observe NCOs in both orientations, it seems this 110-kb region probably experiences the same clustering of initiating lesions when embedded on the Y chromosome, and that subsequent spreading of the synaptonemal complex from the canonical PAR1 ensures engagement and repair with whichever homologue is intact.

Although we observed reasonable correspondence with LD hotspot locations, there were some exceptions and it is tempting to speculate that these may be indicative of sex-specific differences in DSB induction. However, as acknowledged by Pratto *et al*., LD-only hotspots could be the consequence of lower-frequency PRDM9 alleles not assessed in their study, and it is possible that DSB clusters could reflect activity that has yet to make an impact at the population level [19]. Alternatively, repair of DSBs to give rise to NCOs exclusively would have extremely localised effects on haplotype diversity and may even go undetected in the absence of suitably located polymorphisms. Recent refined sex-specific genetic maps derived from >100,000 meioses in pedigrees indicate that there are in fact only a few hundred female- or male-specific recombination hotspots throughout the autosomes in comparison to the tens of thousands of total hotspots predicted by LD [63]. On the other hand, sexually dimorphic regions, *i.e*. 10-kb intervals with significant sex differences in rate, are observed to be more common by an order of magnitude.

Overall, population-based methods are generally good at predicting hotspot location, as noted here and elsewhere [42], but they do not perform so well in predicting hotspot activity. Certainly there is no consistent relationship between LD breakdown and DSB strength (i.e. DMC1-SSDS signal) in our data, though the latter were ascertained in men unlikely to be ePAR carriers and may therefore be particularly influenced by the lifetime of ssDNA intermediates [64] and/or differences in DMC1 loading [65] since SSDS signal on the strictly sex-specific portions of the X and Y is 3-7x higher than on the autosomes [19]. Our sperm data show comparable rates to those observed at autosomal and PAR1 hotspots and although limited to just two intervals, nonetheless show the expected relative relationship with DSB strength. Future sperm CO+NCO analyses might therefore specifically target the strongest DSB clusters reported by Pratto *et al*. [19] to see if they manifest as hotter than characterized sperm hotspots within the autosomes. Such hotspots would offer the opportunity to recover efficiently even atypical events that might provide further mechanistic insight into human meiotic recombination.

Our study suggests that the haplogroup I2a-associated ePAR is likely to have a more geographically restricted distribution than originally proposed [11]. In the course of identifying carriers we established by junction PCR that the ePAR was present within the two sister I2a sub-lineages I-L1286 and I-L1294, both of which occur predominantly within Northwestern Europe, but was absent from two Hungarian males within the I-M423 sub-haplogroup as determined by resequencing of 3.7 Mb of Y-specific DNA [66] (see S3B Fig). The majority of I-P37.2 men belong to the sub-lineage I-M423, which is predominantly found within Southeastern Europe, and rarely encountered in Northwestern Europe [67], hence its probable absence from the dataset tested by Mensah *et al*. So, whilst we would expect to find haplogroup-I2a ePAR carriers at a frequency of approximately 1% among Northwest European men as originally reported [11], we would expect only a minority of I2a men in Southwest Europe to be carriers of the ePAR.

Breakpoint sequence analysis [11] has shown that the ePAR owes its origin to NAHR between repeated sequences (LTR6B elements), so it is inherently likely to be recurrent. Indeed, its presence in two distinct Y haplogroups shows that it has occurred at least twice. The increasing size of population-based genome-wide SNP datasets, (*e.g.* [68]), may allow further examples of the ePAR, or, indeed, other PAR1 extensions, to be identified and characterized. With sufficient numbers of independent occurrences in hand, the influence of sequence diversity of the mediating LTR6B sequences will be able to be understood in detail.

Methods

Samples and ethical approvals

North European semen samples were collected with written informed consent, and ethical approval for their use in recombination studies has been granted to CAM by NRES-East Midlands (REC ref. 6659). Sperm DNA was prepared as described in [36]. Additional DNA samples were also collected with written informed consent following University of Leicester ethical review (refs.: maj4-46d9 and maj4-cb66). Blood DNA samples originally analysed in [11] were part of an institutional genome-wide CNV study that was approved by KU Leuven review board (protocol number S55513). Lymphoblastoid cell-line DNA from CEPH family 1334 is available from the Coriell Institute (https://www.coriell.org/).

Identification of potential ePAR sperm donors and other ePAR carriers

One sperm donor (man 20) was previously identified as carrying a duplication of the X chromosome that encompassed the canonical PAR1 boundary and extended at least 12 kb proximal to this (S1 Fig). Twenty-three Y-STRs were typed in 81 donors, including man 20, using the PowerPlex Y23 kit (Promega). Y-chromosome haplogroups were predicted from the resulting STR haplotypes using a Bayesian Allele Frequency approach (http://www.nevgen.org/). Man 20 and man 53 were predicted to carry the haplogroup I2a-L233 sublineage. Two further unrelated ePAR carriers were found by surveying PowerPlex Y23 data to predict haplogroup I2a Y chromosomes among laboratory collections of DNA samples. A first-generation male from CEPH family 1334 (NA12146) was identified as another candidate carrier; he was reported to have an apparent duplication of X-linked SNPs in the vicinity of the ePAR1 (hg19 chrX:2694151–2808548; hg38 chrX:2776110–2890507) in DGV (http://dgv.tcag.ca/dgv/app/home), and predicted to belong to the same I2a sub-haplogroup based on his Y-STR profile (data kindly provided by C.Tyler-Smith, Wellcome Trust Sanger Institute). We also typed two Hungarian males known from sequencing of 3.4Mb of their male specific Y to have the most distantly related I2a sublineage (I2a-M423) [66] to determine whether all males within I2a possessed an ePAR.

Confirmation of ePAR status

A duplex PCR consisting of a 848-bp fragment spanning the ePAR junction (*i.e.* distal X-specific LTR6B and proximal PAR1-specific LTR6B) together with a 1551-bp control fragment from the *SRY* gene was used to verify the ePAR rearrangement. PCRs were carried out in the buffer described in [69] using primers ePARjunc-F (5′-TGGCAATGTTACTGGAGACG), ePARjunc-R (5′-CAAGGAGTCTGCTGGAAGTC), SRY-F (5′-GGGGTCCCGAGATTTATGTT) and SRY-R (5′-GCTAGAACAAGTTACCCCTC), with an annealing temperature of 60˚C and extension temperature of 65˚C.

Confirmation of Y-chromosome haplogroup

A multiplex PCR encompassing nine haplogroup-identifying SNPs within I2a was developed with an annealing temperature of 59˚C and extension temperature of 65˚C (S5A Table). The resulting products were used in a SNaPshot single-base extension assay using the primer mix detailed in S5B Table according to the manufacturer's instructions (Thermo Fisher Scientific). The phylogenetic relationships of the haplogroups detected by the SNaPshot assay are shown in S3B Fig.

Detection of sperm *de novo* recombinants

Assays capable of detecting *de novo* reciprocal crossovers spanning the most distal DSB cluster were designed for each sperm donor following the guidelines in [36]. Similarly, assays able to simultaneously detect reciprocal *de novo* crossovers as well as non-crossover gene conversion events were designed for the proximal target region [36]. Details of the allele-specific primers (ASPs) directed against SNP variants used for recombinant selection are given in S6 and S7 Tables. Phasing of these markers was established empirically using ASP-derived amplicons as templates for allele specific oligonucleotide (ASO) typing [36]. *De novo* recombinants were also characterized by the same method. Details of ASOs are given in S8 Table.

Sequence analysis of the ePAR

Overlapping long-PCR amplicons were designed to cover the ePAR region (details of the primer pairs are given in S9 Table). The amplicons were pooled equimolar for each individual in two sets, cleaned with Agencourt AMPure XP beads (Beckman Coulter) and used to make individual-specific libraries using the Ion Xpress Library kit and barcodes (Thermo Fisher Scientific) according to the manufacturer's instructions for 400-bp sequencing. Libraries were size-selected on 1.8% LE agarose, gel-purified using a Zymoclean DNA Recovery kit (Zymo Research), quantified using an Agilent 2100 Bioanalyzer and pooled equimolar. Sequencing templates were prepared using the Ion PGM HI-Q OT2 Kit and sequencing was performed according to manufacturer's instructions in two runs on an Ion Torrent PGM using the Ion PGM HI-Q Sequencing Kit and 316v2 Chips (Thermo Fisher Scientific). Reads were mapped to the human reference sequence (hg19) using the Torrent Suite Software 5.0.2. The mean number of Q20 bases called per individual across the two runs was 69,342,818 (range: 24,051,513–134,452,533) and mean number of mapped reads was 308,842 (range: 117,148–753,664). Summary statistics for each individual sequenced are shown in S1 Table. See S1 Text for details of validation. The fastq files can be accessed at https://www.ncbi.nlm.nih.gov/sra/SRP155538.

Phasing of the ePAR

Variant calls were generated by SAMtools 1.3.2 using the bam files and selecting only reads with a minimum mapping quality of 50 and a minimum base quality of 20. The variant calls from the two runs were merged for each individual. Inclusion of female samples and appropriate monochromosomal hybrid cell-line DNA controls (https://www.coriell.org/0/Sections/Collections/NIGMS/Map02.aspx?PgId=496) at the template preparation stage indicated that despite careful design of primer pairs, it was impossible to prevent amplification of portions of Yq11.2; genotype calls for these regions were therefore excluded from further analysis along with Indels and markers mapping to tandemly repetitive sequences. Haplotypes were derived using the program PHASE (http://stephenslab.uchicago.edu/phase/download.html) [48,49], checked for compatibility amongst families and in cases of remaining ambiguity resolved

parsimoniously (mean = 7.14 ± 4.40%). See S1 Text for details of validation. Phased X haplotypes over the interval involved in the ePAR translocation were obtained from the CEU and GBR males from the 1000 Genomes Project [34] for comparison.

TMRCA of haplogroup I2a ePARs

A median-joining Y-STR network of the haplogroup I2a ePARs was constructed using the Network software from Fluxus Engineering [70] and all 23 Y-STRs of the PowerPlex Y23 kit; the bilocal DYS385a,b was included because these Y chromosomes are closely related and the 'phasing' issue can be ignored. The TMRCA was estimated from the 23 Y-STR data using the ASD method [51,52] as described in [53], assuming a generation time of 31 yrs [71].

Supporting information

S1 Fig. Duplication of the X chromosome in a North European sperm donor.
(PDF)

S2 Fig. Features considered when choosing intervals for sperm recombination analysis.
(PDF)

S3 Fig. Median-joining network of I2a ePAR-carrying males.
(PDF)

S1 Table. Summary statistics for Ion Torrent sequencing across the ePAR.
(PDF)

S2 Table. Comparison of ePAR haplotype structures with phase known X chromosomes—SNP markers.
(PDF)

S3 Table. Comparison of ePAR haplotype structures with phase known X chromosomes—Summary data.
(PDF)

S4 Table. PowerPlex Y 23 haplotypes for haplogroup I2a ePARs.
(PDF)

S5 Table. Y-chromosome haplogrouping using a SNaPshot single-base extension assay.
(PDF)

S6 Table. Primer sequences for sperm recombination analysis.
(PDF)

S7 Table. Primer combinations and annealing temperatures used for sperm recombination analysis.
(PDF)

S8 Table. Allele-specific oligonucleotide probe (ASO) sequences.
(PDF)

S9 Table. Primer sequences for Ion Torrent sequencing templates.
(PDF)

S1 Text. Validation of Ion Torrent data.
(PDF)

Acknowledgments

We thank anonymous DNA donors for their contributions to this work, Chris Tyler-Smith for providing Y-STR data for the CEU males, Joris Vermeesch for giving permission to analyse the Belgian/French samples, Matthew Hestand for providing PacBio data, Toby Evans and Poonam Thakkar for help with the preparation of long-PCR sequencing templates and Gurdeep Lall for advice with the SNaPshot assay.

Author Contributions

Conceptualization: Celia A. May.

Formal analysis: Nitikorn Poriswanish, Jon H. Wetton, Celia A. May.

Investigation: Nitikorn Poriswanish, Rita Neumann, Jon H. Wetton, Celia A. May.

Methodology: Nitikorn Poriswanish, Rita Neumann, Celia A. May.

Project administration: Celia A. May.

Resources: Jon H. Wetton, Maarten H. D. Larmuseau, Mark A. Jobling, Celia A. May.

Software: John Wagstaff.

Visualization: Nitikorn Poriswanish, Jon H. Wetton, Celia A. May.

Writing – original draft: Celia A. May.

Writing – review & editing: Nitikorn Poriswanish, Rita Neumann, Jon H. Wetton, Maarten H. D. Larmuseau, Mark A. Jobling.

References

1. Cooke HJ, Brown WR and Rappold GA (1985) Hypervariable telomeric sequences from the human sex chromosomes are pseudoautosomal. Nature 317: 687–692. PMID: 2997619
2. Simmler M-C, Rouyer F, Vergnaud G, Nyström-Lahti M, Ngo KY, de La Chapelle A et al. (1985) Pseudoautosomal DNA sequences in the pairing region of the human sex chromosomes. Nature 317: 692–697. PMID: 2997620
3. Bickmore WA and Cooke HJ (1987) Evolution of homologous sequences on the human X and Y chromosomes, outside of the meiotic pairing segment. Nucleic acids research. 15:6261–71. PMID: 3502702
4. Rouyer F, Simmler M-C, Johnsson C, Vergnaud G, Cooke HJ and Weissenbach J. (1986) A gradient of sex linkage in the pseudoautosomal region of the human sex chromosomes. Nature 319: 291–295. https://doi.org/10.1038/319291a0 PMID: 3941746
5. Hall H, Hunt P and Hassold T (2006) Meiosis and sex chromosome aneuploidy: How meiotic errors cause aneuploidy. Current Opin Genetics Dev 16: 323–329.
6. Shi Q, Spriggs E, Field LL, Ko E, Barclay L and Martin RH (2001) Single sperm typing demonstrates that reduced recombination is associated with the production of aneuploid 24, XY human sperm. Am J Med Genet 99: 34–38. PMID: 11170091
7. Mohandas T, Speed R, Passage M, Yen P, Chandley A and Shapiro L (1992) Role of the pseudoautosomal region in sex-chromosome pairing during male meiosis: Meiotic studies in a man with a deletion of distal Xp. Am J Hum Genet 51: 526–533. PMID: 1496984
8. Ellis N, Yen P, Neiswanger K, Shapiro LJ and Goodfellow PN (1990) Evolution of the pseudoautosomal boundary in old world monkeys and great apes. Cell 63: 977–986. PMID: 2124175
9. Hughes JF, Skaletsky H, Brown LG, Pyntikova T, Graves T, Fulton RS, Dugan S, et al. (2012) Strict evolutionary conservation followed rapid gene loss on human and rhesus Y chromosomes. Nature 483: 82–86. https://doi.org/10.1038/nature10843 PMID: 22367542
10. Ellis NA, Goodfellow PJ, Pym B, Smith M, Palmer M, Frischauf A-M et al. (1989) The pseudoautosomal boundary in man is defined by an *Alu* repeat sequence inserted on the Y chromosome. Nature 337: 81–84. https://doi.org/10.1038/337081a0 PMID: 2909893

11. Mensah MA, Hestand MS, Larmuseau MHD, Isrie M, Vanderheyden N, Declercq M et al. (2014) Pseudoautosomal region 1 length polymorphism in the human population. PLoS Genet 10: e1004578. https://doi.org/10.1371/journal.pgen.1004578 PMID: 25375121
12. Johnson NC (2011) XG: the forgotten blood group system. Immunohematology 27: 68–71. PMID: 22356523
13. Mu J, Skurat AV and Roach PJ (1997) Glycogenin-2, a novel self-glucosylating protein involved in liver glycogen biosynthesis. J Biol Chem 272: 27589–27597. PMID: 9346895
14. Mu J and Roach PJ (1998) Characterization of human glycogenin-2, a self-glucosylating initiator of liver glycogen metabolism. J Biol Chem 273: 34850–34856. PMID: 9857012
15. Carrel L and Willard HF (2005) X-inactivation profile reveals extensive variability in X-linked gene expression in females. Nature 434: 400–404. https://doi.org/10.1038/nature03479 PMID: 15772666
16. Oven M, Geystele A, Kayser M, Decorte R and Larmuseau M. H. (2014) Seeing the wood for the trees: a minimal reference phylogeny for the human Y chromosome. Hum Mutat 35: 187–191. https://doi.org/10.1002/humu.22468 PMID: 24166809
17. Cotter DJ, Brotman SM and Sayres MAW (2016) Genetic diversity on the human X chromosome does not support a strict pseudoautosomal boundary. Genetics. 203. 485–492. https://doi.org/10.1534/genetics.114.172692 PMID: 27010023
18. Keeney S (2001) Mechanism and control of meiotic recombination initiation. Curr Top Dev Biol 52: 1–53. PMID: 11529427
19. Pratto F, Brick K, Khil P, Smagulova F, Petukhova GV and Camerini-Otero RD (2014) Recombination initiation maps of individual human genomes. Science 346: 1256442. https://doi.org/10.1126/science.1256442 PMID: 25395542
20. Rootsi S, Kivisild T, Benuzzi G, Bermisheva M, Kutuev I, Barać L, et al. (2004) Phylogeography of Y-chromosome haplogroup I reveals distinct domains of prehistoric gene flow in Europe. Am J Hum Genet 75: 128–137. https://doi.org/10.1086/422196 PMID: 15162323
21. Balaresque P, Bowden GR, Adams SM, Leung H-Y, King TE, Rosser ZH et al. (2010) A predominantly neolithic origin for European paternal lineages. PLoS Biol 8: e1000285. https://doi.org/10.1371/journal.pbio.1000285 PMID: 20087410
22. Jeffreys AJ, Kauppi L and Neumann R (2001) Intensely punctate meiotic recombination in the class II region of the major histocompatibility complex. Nature Genet 29: 217–222. https://doi.org/10.1038/ng1001-217 PMID: 11586303
23. May CA, Shone AC, Kalaydjieva L, Sajantila A and Jeffreys AJ (2002). Crossover clustering and rapid decay of linkage disequilibrium in the Xp/Yp pseudoautosomal gene *SHOX*. Nature Genet 31: 272–275. https://doi.org/10.1038/ng918 PMID: 12089524
24. Holloway K, Lawson VE and Jeffreys AJ (2006) Allelic recombination and de novo deletions in sperm in the human β-globin gene region. Hum Mol Genet 15: 1099–1111. https://doi.org/10.1093/hmg/ddl025 PMID: 16501000
25. Jeffreys AJ, Murray J and Neumann R (1998) High-resolution mapping of crossovers in human sperm defines a minisatellite-associated recombination hotspot. Mol Cell 2: 267–273. PMID: 9734365
26. Jeffreys AJ, Neumann R, Panayi M, Myers S and Donnelly P (2005) Human recombination hot spots hidden in regions of strong marker association. Nature Genet 37: 601–606. https://doi.org/10.1038/ng1565 PMID: 15880103
27. Tiemann-Boege I, Calabrese P, Cochran DM, Sokol R and Arnheim N (2006) High-resolution recombination patterns in a region of human chromosome 21 measured by sperm typing. PLoS Genet 2: e70. https://doi.org/10.1371/journal.pgen.0020070 PMID: 16680198
28. Hayashi K, Yoshida K and Matsui Y (2005) A histone H3 methyltransferase controls epigenetic events required for meiotic prophase. Nature 438: 374–378. https://doi.org/10.1038/nature04112 PMID: 16292313
29. Baudat F, Buard J, Grey C, Fledel-Alon A, Ober C, Przeworski M et al. (2010) PRDM9 is a major determinant of meiotic recombination hotspots in humans and mice. Science 327: 836–840. https://doi.org/10.1126/science.1183439 PMID: 20044539
30. Myers S, Bowden R, Tumian A, Bontrop RE, Freeman C, MacFie TS et al. (2010) Drive against hotspot motifs in primates implicates the *PRDM9* gene in meiotic recombination. Science 327: 876–879. https://doi.org/10.1126/science.1182363 PMID: 20044541
31. Paigen K and Petkov PM (2018) PRDM9 and its role in genetic recombination. Trends in Genet 34: 291–300.
32. Sherry ST, Ward M-H, Kholodov M, Baker J, Phan L, Smigielski EM et al. (2001) dbSNP: The NCBI database of genetic variation. Nucleic Acids Res 29: 308–311. PMID: 11125122

33. Hellmann I, Ebersberger I, Ptak SE, Pääbo S and Przeworski M (2003) A neutral explanation for the correlation of diversity with recombination rates in humans. Am J Hum Genet 72: 1527–1535. https://doi.org/10.1086/375657 PMID: 12740762

34. Consortium GP (2012) An integrated map of genetic variation from 1,092 human genomes. Nature 491: 56–65. https://doi.org/10.1038/nature11632 PMID: 23128226

35. Berg IL, Neumann R, Sarbajna S, Odenthal-Hesse L, Butler NJ and Jeffreys AJ (2011) Variants of the protein PRDM9 differentially regulate a set of human meiotic recombination hotspots highly active in African populations. Proc Natl Acad Sci USA 108: 12378–12383. https://doi.org/10.1073/pnas.1109531108 PMID: 21750151

36. Kauppi L, May CA and Jeffreys AJ (2009) Analysis of meiotic recombination products from human sperm. In Meiosis (pp. 323–355). Humana Press.

37. Sarbajna S, Denniff M, Jeffreys AJ, Neumann R, Soler Artigas M, Veselis A et al. (2012) A major recombination hotspot in the XqYq pseudoautosomal region gives new insight into processing of human gene conversion events. Hum Mol Genet 21: 2029–2038. https://doi.org/10.1093/hmg/dds019 PMID: 22291443

38. Berg IL, Neumann R, Lam K-WG, Sarbajna S, Odenthal-Hesse L, May CA et al. (2010) *PRDM9* variation strongly influences recombination hot-spot activity and meiotic instability in humans. Nature Genet 42: 859–863. https://doi.org/10.1038/ng.658 PMID: 20818382

39. Jeffreys AJ and Neumann R (2002) Reciprocal crossover asymmetry and meiotic drive in a human recombination hot spot. Nature Genet 31: 267–271. https://doi.org/10.1038/ng910 PMID: 12089523

40. Jeffreys AJ and Neumann R (2009) The rise and fall of a human recombination hot spot. Nature Genet 41: 625–629. https://doi.org/10.1038/ng.346 PMID: 19349985

41. Jeffreys AJ and Neumann R (2005) Factors influencing recombination frequency and distribution in a human meiotic crossover hotspot. Hum Mol Genet 14: 2277–2287. https://doi.org/10.1093/hmg/ddi232 PMID: 15987698

42. Webb AJ, Berg IL and Jeffreys A (2008) Sperm cross-over activity in regions of the human genome showing extreme breakdown of marker association. Proc Natl Acad Sci USA 105:10471–10476. https://doi.org/10.1073/pnas.0804933105 PMID: 18650392

43. Coop G and Myers SR (2007) Live hot, die young: Transmission distortion in recombination hotspots. PLoS Genet 3: e35. https://doi.org/10.1371/journal.pgen.0030035 PMID: 17352536

44. Jeffreys AJ and May CA (2004) Intense and highly localized gene conversion activity in human meiotic crossover hot spots. Nature Genet 36:151–156. https://doi.org/10.1038/ng1287 PMID: 14704667

45. Odenthal-Hesse L, Berg IL, Veselis A, Jeffreys AJ and May CA (2014) Transmission distortion affecting human noncrossover but not crossover recombination: A hidden source of meiotic drive. PLoS Genet 10: e1004106. https://doi.org/10.1371/journal.pgen.1004106 PMID: 24516398

46. Arbeithuber B, Betancourt AJ, Ebner T and Tiemann-Boege I (2015) Crossovers are associated with mutation and biased gene conversion at recombination hotspots. Proc Natl Acad Sci USA 112: 2109–2114. https://doi.org/10.1073/pnas.1416622112 PMID: 25646453

47. Ross MT, Grafham DV, Coffey AJ, Scherer S, McLay K, Muzny D, et al. (2005) The DNA sequence of the human X chromosome. Nature 434: 325–337. https://doi.org/10.1038/nature03440 PMID: 15772651

48. Stephens M, Smith NJ and Donnelly P (2001) A new statistical method for haplotype reconstruction from population data. Am J Hum Genet 68: 978–989. https://doi.org/10.1086/319501 PMID: 11254454

49. Stephens M and Scheet P (2005) Accounting for decay of linkage disequilibrium in haplotype inference and missing-data imputation. Am J Hum Genet 76: 449–462. https://doi.org/10.1086/428594 PMID: 15700229

50. King TE, Ballereau SJ, Schürer KE and Jobling MA (2006) Genetic signatures of coancestry within surnames. Current Biology 16: 384–388. https://doi.org/10.1016/j.cub.2005.12.048 PMID: 16488872

51. Goldstein DB, Linares AR, Cavalli-Sforza LL and Feldman MW (1995) An evaluation of genetic distances for use with microsatellite loci. Genetics 139: 463–471. PMID: 7705647

52. Goldstein DB, Linares AR, Cavalli-Sforza LL and Feldman MW (1995) Genetic absolute dating based on microsatellites and the origin of modern humans. Proc Natl Acad Sci USA 92: 6723–6727. PMID: 7624310

53. Khubrani YM, Wetton JH and Jobling MA (2018) Extensive geographical and social structure in the paternal lineages of Saudi Arabia revealed by analysis of 27 Y-STRs. Forensic Sci Int: Genet 33:98–105.

54. King TE and Jobling MA (2009) Founders, drift, and infidelity: the relationship between Y chromosome diversity and patrilineal surnames. Mol Biol Evol 26: 1093–1102. https://doi.org/10.1093/molbev/msp022 PMID: 19204044

55. Yu A, Zhao C, Fan Y, Jang W, Mungall AJ, Deloukas P, Olsen A, Doggett NA, Ghebranious N, Broman KW and Weber JL (2001) Comparison of human genetic and sequence-based physical maps. Nature 409: 951. https://doi.org/10.1038/35057185 PMID: 11237020

56. Hinch AG, Altemose N, Noor N, Donnelly P and Myers SR (2014) Recombination in the human pseudoautosomal region PAR1. PLoS Genet 10: e1004503. https://doi.org/10.1371/journal.pgen.1004503 PMID: 25033397

57. Burgoyne PS, Mahadevaiah SK and Turner JM (2009) The consequences of asynapsis for mammalian meiosis. Nature Rev Genet 10: 207–216. https://doi.org/10.1038/nrg2505 PMID: 19188923

58. Gabriel-Robez O, Rumpler Y, Ratomponirina C, Petit C, Levilliers J, Croquette M et al. (1990) Deletion of the pseudoautosomal region and lack of sex-chromosome pairing at pachytene in two infertile men carrying an X;Y translocation. Cytogenet Genome Res 54: 38–42.

59. Faisal I and Kauppi L (2016). Sex chromosome recombination failure, apoptosis, and fertility in male mice. Chromosoma 125: 227–235. https://doi.org/10.1007/s00412-015-0542-9 PMID: 26440410

60. Dumont BL (2017). Meiotic consequences of genetic divergence across the murine pseudoautosomal region. Genetics. 205. 1089–100. https://doi.org/10.1534/genetics.116.189092 PMID: 28100589

61. Jeffreys AJ, Holloway JK, Kauppi L, May CA, Neumann R, Slingsby MT et al. (2004) Meiotic recombination hot spots and human DNA diversity. Phil Trans Roy Soc B: Biol Sci 359:141–152.

62. Kauppi L, Jeffreys AJ and Keeney S (2004) Where the crossovers are: Recombination distributions in mammals. Nature Rev Genet 5: 413–424. https://doi.org/10.1038/nrg1346 PMID: 15153994

63. Bhérer C, Campbell CL and Auton A (2017) Refined genetic maps reveal sexual dimorphism in human meiotic recombination at multiple scales. Nature Comm 8:14994.

64. Lange J, Yamada S, Tischfield SE, Pan J, Kim S, Zhu X, Socci ND, Jasin M and Keeney S (2006) The landscape of mouse meiotic double-strand break formation, processing, and repair. Cell 167: 695–708.

65. Kauppi L, Barchi M, Baudat F, Romanienko PJ, Keeney S and Jasin M (2011) Distinct properties of the XY pseudoautosomal region crucial for male meiosis. Science 331: 916–920. https://doi.org/10.1126/science.1195774 PMID: 21330546

66. Hallast P, Batini C, Zadik D, Maisano Delser P, Wetton JH, Arroyo-Pardo E et al. (2014) The Y-chromosome tree bursts into leaf: 13,000 high-confidence SNPs covering the majority of known clades. Mol Biol Evol 32: 661–673. https://doi.org/10.1093/molbev/msu327 PMID: 25468874

67. Šarac J, Šarić T, Havaš Auguštin D, Novokmet N, Vekarić N, Mustać M et al. (2016) Genetic heritage of Croatians in the Southeastern European gene pool—Y chromosome analysis of the Croatian continental and island population. Am J Hum Biol 28: 837–845. https://doi.org/10.1002/ajhb.22876 PMID: 27279290

68. Sudlow C, Gallacher J, Allen N, Beral V, Burton P, Danesh J et al. (2015) UK Biobank: An open access resource for identifying the causes of a wide range of complex diseases of middle and old age. PLoS Med 12: e1001779. https://doi.org/10.1371/journal.pmed.1001779 PMID: 25826379

69. Jeffreys AJ, Wilson V, Neumann R and Keyte J (1988) Amplification of human minisatellites by the polymerase chain reaction: Towards DNA fingerprinting of single cells. Nucleic Acids Res 16:10953–10971. PMID: 3205737

70. Bandelt H-J, Forster P and Röhl A (1999) Median-joining networks for inferring intraspecific phylogenies. Mol Biol Evol 16: 37–48. https://doi.org/10.1093/oxfordjournals.molbev.a026036 PMID: 10331250

71. Fenner JN (2005) Cross-cultural estimation of the human generation interval for use in genetics-based population divergence studies. Am J Phys Anthropol 128:415–423. https://doi.org/10.1002/ajpa.20188 PMID: 15795887

72. Graffelman J and Weir B (2016) Testing for Hardy–Weinberg equilibrium at biallelic genetic markers on the X chromosome. Heredity 116: 558–568. https://doi.org/10.1038/hdy.2016.20 PMID: 27071844

73. 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 Genet 40:1124–1129. https://doi.org/10.1038/ng.213 PMID: 19165926